



vera.ai: VERification Assisted by Artificial Intelligence

D3.3 - Cross-modal and user feedback enhanced verification tools

Project Title	vera.ai
Contract No.	101070093
Instrument	HORIZON-RIA
Thematic Priority	CL4-2021-HUMAN-01-27
Start of Project	15 September 2022
Duration	36 months



vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.

The content of this document is © of the author(s) and respective referenced sources. For further information, visit veraai.eu.

Deliverable title	Cross-modal and user feedback enhanced verification tools
Deliverable number	D3.3
Deliverable version	V1.0
Previous version(s)	N/A
Contractual Date of delivery	14.07.2025
Actual Date of delivery	21.07.2025
Nature of deliverable	Report
Dissemination level	Public
Partner Responsible	USFD
Author(s)	Apostolidis Konstantinos, Vasileios Mezaris, Nikolaos Kaparinos, Ieremias Omiros Boutsios, Stefanos Papadopoulos, Olga Papadopoulou, Symeon Papadopoulos (CERTH), Mehdi Boussâa, Miguel Colom (ENS), Olesya Razuvayevskaya, Siqi Sun (USFD)
Reviewer(s)	Ivan Srba, Martin Hyben (KInIT), Danae Tsaouraki (ATC)
EC Project Officer	Peter Friess

Abstract	Deliverable 3.3 builds upon the progress reported in D3.1, further advancing the work in Tasks 3.1, 3.2, and 3.4, while introducing novel AI methods and experiments carried out in Tasks 3.5 and 3.6. This document presents enhanced methodologies for multilingual credibility assessment, audiovisual content analysis, the extraction of verification clues from visual content, detection of decontextualised content and the fact-checker in the loop approach. These advancements are accompanied by reference software implementations and services
Keywords	AI, verification, credibility signals, text misinformation analysis, audiovisual content analysis, verification cues, trustworthy AI, decontextualization

Copyright

© Copyright 2025 vera.ai Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the vera.ai Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.

Revision History

Version	Date	Modified by	Comments
V0.1	20/01/2025	Olga Papadopoulou (CERTH)	Draft ToC
V0.2	02/06/2025	Siqi Sun (USFD), Olga Papadopoulou (CERTH)	Initial draft
V0.3	16/06/2025	All authors involved	Added content in Sections 3-8
V0.4	20/06/2025	All authors involved	Finalization of Sections 3-8
V0.5	23/06/2025	Siqi Sun (USFD), Olga Papadopoulou (CERTH)	Authored Executive Summary,
V0.6	26/06/2025	Siqi Sun (USFD), Olga Papadopoulou (CERTH)	Introduction and Sections 2 and Conclusions and Challenges Ahead
V0.7	04/07/2025	Ivan Srba (KInIT), Martin Hyben (KInIT), Danae Tsabouraki (ATC)	First draft ready for review
V0.8	10/07/2025	All authors involved	Addressing feedback, applying final corrections
V0.9	14/07/2025	Olga Papadopoulou, Symeon Papadopoulos (CERTH)	Ultimate formatting and consistency checks. Preparing for submission to EC
V1.0	21/07/2025	Olga Papadopoulou, Symeon Papadopoulos (CERTH)	Deliverable sent to EC

Glossary

Abbreviation	Meaning
AGT	Acoustic Geo-tagging
AI	Artificial Intelligence
AR	Arabic
ASC	Audio Scene Classification
ASR	Automatic Speech Recognition
AVSC	Audio and Visual Scene Classification
B-FPGM	Bayesian-optimized Filter Pruning via Geometric Median
CASA	Computational Auditory Scene Analysis
CBMI	Content-Based Multimedia Indexing
CCN	Consistency Checking Network
CER	Character Error Rate
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CRNN	Convolutional-Recurrent Neural Network
DCASE	Detection and Classification of Acoustic Scenes and Events
DE	German
ECENet	Explainable and Context-Enhanced Network
EN	English
ES	Spanish
FPGM	Filter Pruning via Geometric Median
FR	French
GCD	Great Circle Distance
HMM	Hidden Markov Models
IT	Italian
IPF	Iterative Proportional Fitting
KL	Kullback-Leibler
KSE	Keyframes Selection and Enhancement
LAMAR	Latent Multimodal Reconstruction
LLM	Large Language Model
LSTM	Long Short-Term Memory
mAP	Mean Average Precision
MC	Miscaptioned Images
MMD	Multimodal Misinformation Detection
MUSE	MULTImodal SimilaritiEs
NEM	Named-Entity Manipulation
NLP	Natural Language Processing
OCR	Optical Character Recognition
OOB	Out-of-Context
PSNR	Peak Signal-to-Noise Ratio
RAG	Retrieval Augmented Generation
RED	Relevant Evidence Detection
RED-DOT	Relevant Evidence Detection Directed Transformer

SED	Sound Event Detection
SEN	Stance Extraction Network
SFP	Soft Filter Pruning
SSDL	Self-Supervised Distilled Learning
SwC	Search within Cell
TPE	Tree-structured Parzen Estimator
VADD	Visual-Audio Discrepancy Detection
VBS	Video Browser Showdown
ViT	Vision Transformer
WER	Word Error Rate
WP	Work Package
XAI	Explainable AI

Table of Contents

Revision History	4
Glossary	5
Index of Tables	11
Index of Figures	13
Executive Summary	16
1 Introduction.....	17
2 Overview of Research Contributions, Impact, and Integration	19
2.1 Alignment of User Requirements.....	19
2.2 Research Contributions.....	22
2.3 Key Performance Indicators (KPIs)	23
2.4 Integration into vera.ai User Facing Tools	25
2.5 List of Publications and Preprints.....	26
3 Text Misinformation Analysis and Verification	28
3.1 Methodology.....	29
3.1.1 Use of Persuasion Techniques in Online Disinformation.....	29
3.1.2 Authorship Identification.....	30
3.1.3 Targeted Audience Estimation.....	31
3.2 Evaluation	31
3.2.1 Use of Persuasion Techniques in Online Disinformation.....	32
3.2.2 Authorship Identification.....	33
3.2.3 Targeted Audience Estimation.....	35
3.3 Implementation and Integration.....	37
3.3.1 Credibility Signals in the Assistant Tool	37
3.3.2 Authorship Identification.....	39
3.3.3 Targeted Audience Estimation.....	40
3.4 Concluding Remarks.....	41
4 Audiovisual Content Analysis and Enhancement	43
4.1 Background	43
4.1.1 Audio Analysis.....	43
4.1.2 Evaluation, Integration, and Outreach of the KSE Service.....	45
4.2 Methodology.....	47

4.2.1	Sound Event Detection & Siren Sound Classification.....	47
4.2.2	Acoustic Geo-Tagging using Location-Specific Sounds	49
4.2.3	Audio-Visual Scene Classification	50
4.2.4	User-Centered Evaluation and Integration of the KSE Service	50
4.3	Evaluation	58
4.3.1	Siren Sound Classification	58
4.3.2	Acoustic Geo-Tagging using Location-Specific Sounds	59
4.3.3	Audio-Visual Scene Classification	60
4.3.4	Text Detection	60
4.3.5	Face Detection	61
4.3.6	Face Enhancement	63
4.3.7	KSE Performance Evaluation.....	65
4.3.8	KSE Accuracy Evaluation	66
4.4	Implementation and Integration.....	67
4.5	Concluding Remarks.....	68
5	Extraction of Text and Geolocation from Images.....	69
5.1	Background	69
5.1.1	Optical Character Recognition (OCR).....	69
5.1.2	Geolocation	70
5.1.3	Geolocation Interpretability	71
5.2	Methodology.....	72
5.2.1	Text Extraction and Language Identification	73
5.2.2	Geolocation from a Single Image.....	74
5.2.3	Geolocation Interpretability	79
5.2.4	Sanitization of Dataset.....	80
5.3	Evaluation	81
5.3.1	Text Extraction and Language Identification	81
5.3.2	Geolocation Evaluation.....	86
5.3.3	Geolocation Interpretability with Combi-CAM	90
5.3.4	Sanitization of Dataset.....	91
5.4	Implementation and Integration.....	94
5.4.1	Text Extraction and Language Identification	94
5.4.2	Geolocation from a Single Image.....	94

5.4.3	Geolocation Explainability with Combi-CAM.....	96
5.5	Concluding Remarks.....	97
6	Detection of Decontextualised Content.....	98
6.1	Background	100
6.2	Methodology.....	103
6.2.1	Synthetic Misinformers.....	104
6.2.2	Addressing Unimodal Biases.....	104
6.2.3	Relevant Evidence Detection.....	105
6.2.4	Illusory Progress due to Dataset-specific Biases.....	106
6.2.5	Evidence Verification and Filtering	107
6.2.6	Latent Multimodal Reconstruction.....	108
6.2.7	Framework for Audio-Text Decontextualization Detection.....	109
6.3	Evaluation	113
6.3.1	Image-text Decontextualization	113
6.3.2	Audio-text Decontextualization.....	114
6.4	Implementation and Integration.....	117
6.5	Concluding Remarks.....	118
7	Fact-checker-in-the-loop Approach.....	119
7.1	Background	119
7.1.1	The Role of Conversational Agents in Improving User Experience and Collecting Feedback 119	
7.1.2	User Intent Detection	120
7.1.3	Retrieval Augmented Generation.....	121
7.2	Methodology.....	122
7.2.1	Intent Detection	122
7.2.2	Retrieval Augmented Generation.....	123
7.3	Evaluation	126
7.4	Implementation and Integration.....	128
7.4.1	Integration of the Synthetic Image Chatbot	129
7.4.2	Full Integration	130
7.5	Concluding Remarks.....	132
8	Efficient Annotator Reliability Assessment	133
8.1	Background	133

8.2	EffiARA Python Package	134
8.3	EffiARA Webtool	139
8.4	Evaluation	140
8.4.1	Case Studies	140
8.4.2	Load Testing	140
8.5	Concluding Remarks	141
9	Conclusions and Challenges Ahead	142
	References	145
	Annex I: Text Misinformation Analysis and Verification	166
	Annex II: Audiovisual Content Analysis and Enhancement	169
	Annex III: Extraction of Text and Geolocation from Images	175
	III.I OCR	175
	III.II Geolocation	176

Index of Tables

Table 1 Tool requirements identified for the explainable AI methods for analysis and verification of text, audio, image & video misinformation	19
Table 2 Design requirements identified for the explainable AI methods for analysis and verification of text, audio, image & video misinformation	21
Table 3 Summary of achieved key performance indicators (KPIs) across all detection tasks, compared to initial baseline values defined at the project's start. Relative improvements are shown in percentage terms, highlighting progress made during the project lifecycle	23
Table 4 Summary statistics of disinformation datasets analyzed for persuasion techniques	30
Table 5 Odd ratios of occurrence of persuasion techniques in one domain vs. the average proportion in the others. Statistically significant ratios are underlined (with $p < 0.05$ — Fisher exact test). A → Islamic issues, B → COVID-19, C → Climate change, D → Russ	33
Table 6 Comparison of Authorship Attribution Accuracy by Language and Model for Trained and Unseen Authors	34
Table 7 The R2 results for individual features	36
Table 8 Distribution of articles in the training set for persuasion technique detection task	38
Table 9 Indicative examples of User Feedback, Source, and Actions Taken for the KSE Service	47
Table 10 Processing time distribution across individual stages of the KSE pipeline	51
Table 11 Comparison of text detection methods on the Total-Text dataset, highlighting their F1 scores and average inference times	61
Table 12 Comparative results (mAP) on the WIDER FACE using the EResFD model between uniform FPGM [12] and the proposed B-FPGM. T is the target pruning rate	62
Table 13 Comparison of face detection methods on the WIDER FACE dataset (hard subset), highlighting their F1 scores and average inference times	63
Table 14 Relative processing time overhead (%) of the KSE pipeline with and without deblurring across multiple test videos	63
Table 15 Comparison of selected face super-resolution methods: model size, inference time, quantitative metrics, and visual inspection scores	64
Table 16 Processing time as a percentage of video duration for successive KSE optimizations	66
Table 17 Quantitative evaluation of enhancement accuracy using PSNR and SSIM for optimized vs. unoptimized KSE pipelines under various resolution and processing configurations. Higher scores reflect closer alignment with the original 4K reference, demonstrating	67
Table 18 OCR benchmark results on English Memes dataset. The best results are in bold	83
Table 19 OCR Benchmark Results on HierText Dataset (Original with 4 Decimal Places)	85
Table 20 Accuracy (%) on five granularity ranges of the proposed method compared to state-of-the-art methods on the Im2GPS evaluation set	89
Table 21 Accuracy (%) on five granularity ranges of the proposed method compared to state-of-the-art methods on the Im2GPS3k evaluation set	89
Table 22 Accuracy (%) on five granularity ranges of the proposed method compared to state-of-the-art methods on the YFCC4k evaluation set	90
Table 23 Accuracy (%) and count at five granularity ranges of the model performance at inference for all 10,000 images, 4830 localizable images and 5170 non localizable images of the MP16 dataset	92

Table 24 Accuracy (%) and count at five granularity ranges of the two models' performance, after being trained only on localizable images (partial model) and all the images (full model).....	92
Table 25 Comparative analysis of models with and without external evidence for detecting out-of-context (OOC) image-text pairs, trained on the NewsCLIPpings dataset and evaluated on both its test set and the VERITE benchmark	114
Table 26 Accuracy of each E5 model variant, ordered from smallest to largest, in predicting the most relevant text content from the claim extracted from the audio transcription by Qwen 2.5 7b (Qwen Team, 2024)	115
Table 27 Average similarity score between claims extracted from the audio transcription by Qwen 2.5 7b and original/manipulated texts across E5 models	115
Table 28 Average accuracy, precision and recall for ImageBind and two variations of our proposed method	116
Table 29 F-score on multilingual data for various types of user intents. Boldfaced scores demonstrate statistically significant better overall performance of one model compared to the other for a certain language	127
Table 30 Retriever performance in Terms of Recall@k and MRR for different number of retrieved documents and for three types of document search.....	127
Table 31 Performance of various LLMs on the standalone question generation task based on BLEU, METEOR and ROUGE-2 metrics.....	128
Table 32 Performance of the model on various types of the test sets based on RAGAs, BERTScore and Latency	128
Table 33 Processing time for each stage at varying dataset sizes. Tests conducted running the webtool locally on a laptop with 16GB RAM and an i7-6600U @ 3.400GHz.....	141
Table AII- 1 Overview of User Feedback, Source, and Actions Taken for the KSE Service	169
Table AII- 2 Key Parameters for Temporal Segmentation and Keyframe Selection	173
Table AIII- 1 OCR benchmark results on Arabic Memes dataset. The best results are in bold	175
Table AIII- 2 OCR benchmark results on Bulgarian Memes dataset. The best results are in bold.....	175
Table AIII- 3 OCR benchmark results on North Macedonian Memes dataset. The best results are in bold	176
Table AIII- 4 Accuracy (%) on five granularity ranges of experimental methods on the Im2GPS evaluation set.....	180
Table AIII- 5 Accuracy (%) on five granularity ranges of experimental methods on the Im2GPS3k evaluation set.....	180
Table AIII- 6 Accuracy (%) on five granularity ranges of experimental methods on the Im2GPS3k evaluation set.....	180

Index of Figures

Figure 1 Proportion of persuasion techniques in the different disinformation narratives	32
Figure 2 Impact of Number of Reference Articles on Accuracy for Unseen Authors	34
Figure 3 KL Divergence between Synthetic and Reference Distributions	36
Figure 4 User Interface example for Persuasion Techniques	39
Figure 5 User Interface example for Authorship Attribution	40
Figure 6 User Interface for Audience Profile Estimation	41
Figure 7 Random selection of siren recordings taken from the RSC database. Mel spectrogram illustrations highlight characteristic stable, alternating and sweep-like pitch contours.....	48
Figure 8 The position of the ZIPs download links under the “Fragmentation & Keyframes” panel.....	55
Figure 9 When hovering over a keyframe, its position is highlighted in the timeline using an orange glow	55
Figure 10 The function “Open keyframe in viewer” link, shown in the popup dialog when hovering a keyframe	55
Figure 11 Position of the “Show Advanced Options” label (top panel) and the expanded view it reveals when clicked (bottom panel)	56
Figure 12 Asynchronous provision of keyframes - The “Fragmentation & Keyframes” panel is shown with the results of the video fragmentation while the analysis continues to the subsequent stages.....	57
Figure 13 Clickable URLs for the session identifier and the submitted video source.....	57
Figure 14 Confusion matrix obtained for regional siren classification using a vision-based classification approach.	59
Figure 15 Selection of indicator sounds for three acoustic scene and city classes obtained with both proposed retrieval approaches. The font size correlates with indicator sound relevance	60
Figure 16 A sample image used as input (left) and the corresponding ground-truth (right) for evaluating the face enhancement methods.	65
Figure 17 Earth partitioning. Left: Google S2 library (Greece). Each blue cell contains at least one image based on its coordinates. Right: GADM database (Greece). The partition reflects the administrative levels defined by GADM. Some cells may not contain images.....	75
Figure 18 The Earth is partitioned into continent and country levels using the GADM database. The image displays the coarser administrative boundaries of these regions.....	76
Figure 19 Distribution of MP-16 images across the world. The higher density of blue cells in Europe and North America reflects the training dataset’s bias, with far fewer images in Asia, Africa, and Oceania ...	77
Figure 20 Efficiency analysis of OCR models. The Y-axis represents efficiency, defined as case-insensitive character accuracy divided by the inference time per image	86
Figure 21 Example of non-localizable images from the Im2GPS evaluation dataset. These images could have been taken in many different locations in the world, as they lack distinctive features that would help narrow down their location	87
Figure 22 Example of non-localizable images from the Im2GPS evaluation dataset. These images could have been taken in many different locations in the world, as they lack distinctive features that would help narrow down their location	88
Figure 23 Activation maps obtained with different methods for the two different scenes, one the Opera of Sydney, and the other a view of Paris. The geolocation tool is able to correctly locate the both scenes,	

and thanks to Combi-CAM we can observe that it looked mainly at the tops and facades of buildings in the case of Paris, and at the Opera itself in the image of Sydney. This gives insights about which elements were used by the network to make the final decision	91
Figure 24 Bound on the complexity of the full and partial model. The Y-axis represents the value of the generalization bound, which can be likened to model complexity, the X-axis is the Epoch number, i.e., training time.....	93
Figure 25 API response of geolocation estimation REST API.....	95
Figure 26 Predicted most probable location with associated confidence score output by the model.....	95
Figure 27 Relevant images retrieved from our database corresponding to the estimated location	96
Figure 28 Online demo for Combi-CAM	97
Figure 29 Examples of decontextualized image-caption pairs and a high-level overview of the detection framework. Both examples are sourced from Reuters.com and have been verified as false or misleading	99
Figure 30 This figure illustrates two cases of audio decontextualization in which an excerpt from a longer speech is isolated and circulated out of context. Without modifying the content, this change in context creates a misleading impression. and its use in disinformation raises serious ethical and societal concerns	100
Figure 31 Edited version of an original audio recording (left) disseminated on social media, with the original source (right). Two reused segments are highlighted in green and purple; the omitted segment between them is clearly visible	102
Figure 32 Overview of our proposed workflow: We generate OOC and NEM samples from truthful image-caption pairs using one or more “Synthetic Misinformers”, train a detection model on the synthetic training data, and evaluate it on real-world data.	104
Figure 33 Examples of multimodal misinformation from VERITE and asymmetric multimodal cases from the COSMOS test set	105
Figure 34 An image-text claim is assessed using external information, including both images and text, collected from the web. The system retrieves and re-ranks this information, while RED-DOT identifies the most relevant items to evaluate the claim's validity.....	106
Figure 35 Two examples from the NewsCLIPPings dataset, one truthful (top) and one out-of-context (bottom) alongside their retrieved and re-ranked external evidence, along with computed multimodal similarity scores.....	107
Figure 36 High-level overview of the proposed automated fact-checking pipeline, using Evidence Verification Network to identify unreliable and leaked information, thus ensuring credible evidence for accurate verification.....	108
Figure 37 Workflow of audio-text context analysis framework	110
Figure 38 Example of problematic output from WhisperX and Faster-Whisper when transcribing audio containing only environmental sounds.....	111
Figure 39 Entailment Score distribution using the Minicheck Model with various LLMs.....	112
Figure 40 Chatbot Workflow	123
Figure 41 The Chatbot being used to collect feedback of an image of a cat generated by Google Gemini	129
Figure 42 Post on the feedback Slack channel with the user’s email address, feedback message, and the image.....	130

Figure 43 JSON messages to and from the chatbot	132
Figure 44 An overview of the EffiARA annotation pipeline, covering sample distribution, annotation, label generation, agreement calculation, reliability estimation, and dataset compilation	135
Figure 45 Example agreement visualisations as (A) a 2D graph, (B) a heatmap, and (C) a 3D graph for six annotators (annotations were synthetically generated).....	137
Figure AI- 1 Top 10 most correlated LIWC features for each of the four domain-specific PTs in climate change compared to other domains. Statistically significant ($p < 0.05$) coefficients are indicated with an asterisk (*).	167
Figure AIII- 1 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	176
Figure AIII- 2 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	177
Figure AIII- 3 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	177
Figure AIII- 4 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	177
Figure AIII- 5 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	178
Figure AIII- 6 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	179
Figure AIII- 7 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	179
Figure AIII- 8 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location.....	179

Executive Summary

Generative AI and the rapid evolution of disinformation techniques pose significant risks to information integrity and public discourse, making it easier to create and disseminate synthetic or misleading content across multiple modalities. While these technologies offer innovative opportunities for content creation, they also introduce new challenges for verifying the authenticity and context of information.

This deliverable reports on the results and developments from the following tasks:

- T3.1 - **Multilingual credibility assessment and evidence retrieval,**
- T3.2 - **Audiovisual content analysis and enhancement,**
- T3.4 - **Extraction of verification clues from visual content,**
- T3.5 - **Multimodal fact-checking,**
- T3.6 - **Fact-checker-in-the-loop approach**

This deliverable systematically covers all modalities relevant to multimodal disinformation detection, presenting advancements across text, audio, image, and video analysis. It introduces novel AI-based methods for credibility assessment, audiovisual content analysis, geolocation and text extraction from images, detection of decontextualised content, human-in-the-loop verification and annotator reliability assessment.

Crucially, almost all methods that demonstrated satisfactory performance were successfully deployed and integrated into practical tools, such as the Verification Plugin, Truly Media, and the Assistant chatbot. These tools are actively co-developed with fact-checkers, tested in realistic scenarios, and made available as APIs or user interfaces, ensuring real-world applicability. In alignment with the principles of open science, we have published the associated source code, datasets, and documentation to support transparency and foster further research in this area.

The overarching aspects of our approach are:

- **Multilinguality:** Leveraging language-agnostic models to support tasks such as source credibility classification, persuasion detection, and OCR in multilingual settings.
- **Trustworthiness:** Incorporating confidence scoring and user override functions in the Verification Plugin to support human decision-making.
- **Interpretability:** Integrating explainable AI methods, including Combi-CAM for geolocation justification and token-level attention for textual analysis.
- **Real-world applicability:** Co-developing tools with fact-checkers, validating them in realistic use-case scenarios, and deploying them as APIs, user interfaces, and components of the Verification Plugin, Truly Media and Assistant chatbot.

To further support transparency and usability, special attention was paid to **explainability** and **user feedback integration**. Several of the developed methods have been deployed and integrated into user-facing tools (See Implementation and Integration of each section). All developments are aligned with real-world requirements identified in WP2, ensuring practical relevance and seamless integration into verification workflows.

1 Introduction

Deliverable 3.3 continues the work initiated in D3.1 as part of WP3, Trustworthy AI for Multimodal Content Verification, within the scope of the vera.ai project. The goal of this deliverable is to present the latest research and development outcomes in enhancing the trustworthiness and effectiveness of AI-driven content verification methods. Specifically, it reports on the progress made in Tasks 3.1, 3.2, and 3.4, while also introducing new research, methodologies, and experiments conducted under Tasks 3.5 and 3.6.

This work is motivated by the increasing sophistication of disinformation techniques that exploit both textual and audiovisual modalities, often in multiple languages and across diverse platforms. In response, Deliverable 3.3 introduces advanced AI models and methodologies to:

- Detect persuasion techniques and assess credibility signals in multilingual texts.
- Analyze audiovisual content to extract representative keyframes and meaningful segments, helping users quickly grasp the structure and key moments of a video, additionally detecting contextual clues - including faces, text, and significant sound events - and finally enhancing key visual elements to support clearer interpretation.
- Extract and verify text from complex images using OCR and evaluate performance across multiple languages as well as extract the location depicted in an image using solely visual content.
- Identify decontextualised or misleading content, such as images taken out of their original context or being miscaptioned, and audio-text pairs where the spoken content has been repurposed or misrepresented through unrelated captions.
- Incorporate fact-checkers in the loop and assess annotator reliability to strengthen human-AI collaboration.

These contributions are supported by reference implementations, APIs, and integration into user-facing tools that enhance transparency, usability, and integration into verification workflows. The developments presented in this deliverable align closely with the overarching goals of vera.ai, which seeks to build transparent, robust, and user-centered AI solutions for the fight against disinformation. Particular emphasis is placed on interpretability, efficiency, and multimodal support, ensuring that the tools produced are not only powerful, but also trustworthy and user-friendly.

The document is organized into nine main sections:

- The deliverables begin with Section 2, titled ‘Overview of Developments and Achievements.’ This section serves as a summary of the main research and development carried out in the project. More detailed descriptions of the research and development work are provided in the individual sections that follow. We begin by outlining the advancements specific to the project, describe the user requirements identified in WP2 during the requirements gathering process for vera.ai, and explain how the tools developed in this task address those requirements. Then we outline the context and challenges posed by generative AI technologies, followed by a summary of the project-specific developments. Finally, we conclude with a list of Key Performance Indicators (KPIs), including their target values, discussing the KPIs that have been achieved as well as explaining those that have not. The section concludes by listing 19 publications that have been published in top-tier conferences and journals.

- Section 3 details advances in text misinformation detection, including persuasion technique analysis, authorship identification, and targeted audience identification. This section presents multilingual and explainable AI models capable of extracting fine-grained credibility signals from text, allowing fact-checkers to move beyond binary classifications toward deeper, context-sensitive verification.
- Section 4 details enhancements to the Keyframe Selection and Enhancement (KSE) service based on user feedback, focusing on speed, accuracy, and new features like sound event detection. It also describes the development of novel audio analysis methods for tasks like acoustic scene classification, siren sound classification (for geolocation), and detecting audio-visual inconsistencies
- Section 5 discusses geolocation from a single image and extraction of text through OCR. It highlights tool enhancements that improve robustness across languages and domains, introduces visual explainability methods such as Combi-CAM, and presents real-world evaluations across multiple benchmark datasets, reinforcing the practical value of these tools in verifying visual clues.
- Section 6 introduces methodologies for detecting decontextualised content and evaluating the contextual integrity of media, including the use of synthetic training data, building real-world evaluation sets, addressing unimodal and data-specific biases, and applying evidence-based detection techniques.
- Section 7 describes the integration of human feedback through a fact-checker-in-the-loop framework. This includes intent detection and Retrieval-Augmented Generation (RAG) pipelines designed to support real-time interaction between users and AI systems. By incorporating user feedback directly into the verification process, this section demonstrates how conversational agents can enhance transparency, refine system outputs, and improve overall user trust. Integration plans with tools like the Assistant chatbot are also discussed, alongside performance benchmarks and planned user-centric evaluations.
- Section 8 proposes methods for efficient annotator reliability estimation. This section presents both a Python package and a web-based tool that enable fast estimation of annotation quality in collaborative environments. Through case studies and load testing, it demonstrates how reliability scores can inform downstream tasks such as training data selection or model evaluation, offering practical value in large-scale annotation workflows.
- Section 9 concludes the main contributions of each section, challenges ahead and potential research directions.

Together, these sections reflect an integrated research and development effort that translates theoretical advances into deployable tools for fighting disinformation in a wide variety of real-world contexts.

2 Overview of Research Contributions, Impact, and Integration

This section provides a consolidated overview of the research outputs delivered under WP3 during the M18–M36 period. It summarises the main technical contributions across all tasks, highlights their relevance to user requirements defined in WP2, and outlines progress toward the Key Performance Indicators (KPIs).

2.1 Alignment of User Requirements

Table 1 aligns the user requirements identified in WP2 during the vera.ai requirements gathering process with the approaches investigated in WP3 and the respective section in this deliverable. The user requirements are detailed in D2.1 - AI against Disinformation: Use Cases and Requirements.

Table 1 Tool requirements identified for the explainable AI methods for analysis and verification of text, audio, image & video misinformation

#	User need	Tools requirements	Related Section
#18	Deblur, enhance image	A tool that will help to deblur images to improve clarity, enhance image quality through attribute adjustments, and perform object detection for recognizing and tracking specific objects in images.	4
#19	Lack of audio-based factchecking tools for podcasts	Tools for analysing audio for podcasts	5
#28	Geolocation from image, audio and video	Tool for detecting geographic location where a particular video or image was recorded. It should utilise various methods and data sources to extract relevant information from the videos and images, such as geotagged frames, visual cues, or audio signals that can be correlated with known geographic features.	5
#34	Detecting events with sound	A tool that can tag relevant sound events (sirens, screams, explosions) in long footage, to quickly seek to the most semantically relevant timestamps	4
#36	Detect faces in keyframes	Improve the detection of keyframes containing faces in fragmented videos. This includes extracting images of faces, persons, and text labels.	4
#37	Detect persons in keyframes	Keyframe selection service improvement in InVid-WeVerify plugin	4
#38	Source evaluation	A tool to assess the reliability, credibility, and trustworthiness of digital sources of information.	3
#39	Propaganda detector	Automatic detector or detection methods for persuasive communication or propaganda online.	3
#40	Detect language, keywords and social-media native linguistic	Database of linguistic cues or markers that can help distinguish between reliable and unreliable	3

	markers used for spreading disinformation	information, dogwhistling, as well as respectful and hate speech.	
--	---	---	--

Note: The numbers of the user needs refer to a requirements table that is maintained by the use case partners and accompanies D2.1 - AI against Disinformation: Use Cases and Requirements.

User Need #18 (Deblur, enhance image): To address this need, we developed adaptive image restoration methods that enhance the clarity of blurred or low-quality video frames. These improvements not only support better visual inspection but also significantly improve the performance of downstream face and object detection algorithms. The image enhancement pipeline is integrated into the Keyframe Selection and Enhancement (KSE) tool and made accessible through RESTful APIs, supporting practical use in professional fact-checking workflows.

User Need #19 (Lack of audio-based factchecking tools for podcasts): While our tools were not designed exclusively for podcast analysis, we developed audio classification models, such as siren sound recognition and audio scene classification, that form a solid foundation for audio-based verification. These models can be adapted to detect content manipulations in podcast-style audio and are accessible via API for integration into broader workflows.

User Need #28 (Geolocation from image, audio, and video): This need was addressed through the development of a hybrid geolocation system that combines visual classification and retrieval, powered by CLIP embeddings. The system predicts likely regions and then refines location estimates via similarity search among geotagged images. To improve interpretability, we introduced Combi-CAM, a visual explanation technique that shows which parts of the image influenced the model's decision. For audio, siren recognition and ambient sound classification contribute contextual location clues.

User Need #34 (Detecting events with sound): We developed models for acoustic scene classification and siren detection that can tag meaningful sound events, such as alarms, explosions, or crowd noise. These models were trained using urban-specific indicator sounds and improved through contrastive learning techniques. Their outputs help fact-checkers identify relevant timestamps in long videos for faster content assessment.

User Need #36 (Detect faces in keyframes): The KSE service incorporates improved face detection algorithms that benefit directly from our image enhancement pipeline. These enhancements enable more accurate extraction of faces from keyframes, supporting identity verification and source validation tasks.

User Need #37 (Detect persons in keyframes): Building on the improvements described above, the updated KSE tool enhances the selection and annotation of keyframes with recognizable persons. This directly supports fact-checkers working within tools like the InVID-WeVerify plugin.

User Need #38 (Source evaluation): We developed a suite of explainable text classifiers targeting credibility signals such as genre, framing, subjectivity, and persuasion techniques. These are integrated into the Assistant chatbot within the Verification Plugin, enabling real-time evaluation of a text's credibility dimensions.

User Need #39 (Propaganda detector): Our persuasion technique classifier, which achieved top performance on the SemEval-2023 benchmark, detects a wide range of rhetorical strategies often used in propaganda. This tool enables nuanced, explainable analysis of content manipulation and has been applied across multiple disinformation domains.

User Need #40 (Detect linguistic manipulation markers): The project addressed this through classifiers that detect subjectivity, framing, and persuasive language patterns across multiple languages. Together, these tools form a linguistic profile that helps distinguish credible from manipulative content and supports deeper semantic analysis of online disinformation.

Besides the functional tool requirements presented in Table 1, the vera.ai requirements analysis delivered a set of user needs that refer to processes of discovery, analysis and presentation of results, and focuses on usability, explainability, sharing and integration. In the context of WP3 and more specifically T3.1, T3.2 and T3.4, the design requirements are addressed as listed in Table 2.

Table 2 Design requirements identified for the explainable AI methods for analysis and verification of text, audio, image & video misinformation

#	User need	Design requirements	Related Section
#1	Assessing credibility remains problematic	Indicate results with credibility indicators	3
#2	Video analysis visualisation needs improving	Results of video analysis process should be displayed side-by-side with the target file	4
#3	Need for human verification and trust in results: Understand how the tool came to its conclusion for user cross-examination	The tools need to provide comprehensive reports that outline the fact-checking process, identify misinformation techniques, collect evidence, explain the methodologies used, provide visual and textual descriptions of the drawn conclusions, and offer cross-examination access points within documented process.	3, 5
#4	Understand the language of the tool: jargon/vocabulary and process description, as no common terminology between users and developers	The terms used in describing the process and results (vocabulary) need to be in a language understandable to the users, not just the developers: in plain English, and using fact-checkers vocabulary for specialised words	3
#5	High level of granularity needed	Implement a high level of granularity in defining credibility signals	3
#6	Signals differ in terms of what they covers, whether it is background, webdesign, or other areas	The tool should rely on standardised use of semantic technology language and linguistic signals for assessing credibility.	3
#7	Need for multilingual capabilities	Make the credibility signal tool multilingual	3
#8	Large variety of signals needed	Metrics should include: the source's authenticity, reputation, expertise, bias, and potential motives, as well as accuracy of the information it provides.	3

#9	Visualising results	Visualize results in a way that users can easily interpret them by matching their own visual language	3, 4
#10	Understand how to use the tools	Provide sufficient guiding information to trust & use the tools correctly: Users need tools to be accessible, i.e. (self-)explanatory, with limited and basic training.	3, 4
#21	Transparency about uncertainties	Provide error rate, probability of false positives/negatives	5

Note: The numbers of the user needs refer to a requirements table that is maintained by the use case partners and accompanies D2.1 - AI against Disinformation: Use Cases and Requirements.

2.2 Research Contributions

The increasing sophistication and scale of multimodal disinformation campaigns demand trustworthy AI methods that go beyond unimodal content analysis, addressing the interplay between text, image, audio, and video in both multilingual and cross-platform contexts. Today, generative models can fabricate content that mimics authentic style and context with high fidelity. For example, misinformation is often framed using persuasive linguistic devices, embedded within convincing visuals, and sometimes accompanied by location-mismatched or decontextualized audio. Existing detection tools, particularly those based on supervised learning, struggle to generalize across domains, modalities, or evolving narrative tactics.

To address this, we applied methods combining semantic, stylistic, and contextual signals across modalities through Tasks T3.1 and T3.2. Persuasion detection, authorship identification and audience profiling offer novel perspectives on source characteristics and targeting strategies in disinformation.

These models were built with scalability in mind, integrating into real-time APIs and end-user tools, such as the Assistant chatbot within the Verification Plugin.

In the audiovisual domain, we developed methods to detect mismatches between visual scenes and their accompanying audio environments using joint audio-visual scene classification. We also created acoustic geo-tagging techniques that leverage location-specific "indicator sounds" and country-distinctive siren classification based on pitch contours. These tools support contextual analysis for content verification.

Further, D3.3 introduces novel capabilities for image geolocation and multilingual OCR, addressing user needs for extracting verification cues from images, even in visually complex or text-rich scenarios. We also provide a new technique for the interpretation of the geolocation results. These services are now benchmarked across real-world datasets and exposed via APIs.

Recognizing that disinformation often results not only from manipulation but also from decontextualization, we proposed new detection strategies to flag content used out of its original narrative frame. In addition, we implemented a fact-checker-in-the-loop framework that gathers user feedback to refine system output, closing the loop between AI inference and expert judgment. Regarding annotator reliability assessment, the EffiARA framework for fast and scalable reliability estimation is developed and exposed via Python packages and webtools.

Finally, our research outcomes are not only academic: 19 peer-reviewed publications, benchmarked systems, and operational tools have emerged from this work. These contributions directly align with WP2 user needs, such as source reliability estimation, cross-platform campaign detection, and visual-audio inconsistency analysis, and demonstrate the project’s commitment to deployable, explainable, and multilingual verification solutions.

2.3 Key Performance Indicators (KPIs)

A key performance indicator (KPI) defined for this project is **achieving a relative performance improvement of over 20% in detection accuracy** compared to the baseline established at project initiation. This KPI reflects our ambition to significantly enhance the effectiveness of synthetic content detection across all modalities.

In the tables below, we provide a summary of the KPIs, while detailed results and methodology for each KPI are presented in the corresponding sections of the deliverable. Specifically, Table 3 presents the achieved KPI values along with their relative improvement compared to the baseline established at the beginning of the project. In Table 4, we extend this evaluation by comparing our current results against the most recent state-of-the-art (SOTA) methods available at the time of writing.

Table 3 Summary of achieved key performance indicators (KPIs) across all detection tasks, compared to initial baseline values defined at the project’s start. Relative improvements are shown in percentage terms, highlighting progress made during the project lifecycle

Method	State-of-the-art value	Current value of vera.ai method	Improvement (metric)	Section
Persuasion detection / A fine-tuned RoBERTa-Large model (Razuvayevskaya et al., 2024)	22.3 (Liu et al., 2022)	43.0	92.31% (Accuracy)	3
Siren Sound Classification / Fine-tuning of pre-trained MobileNetV2 architecture on synthetic dataset of generated pitch contours	0.38 (SWIPE pitch tracking + CNN with learnable front-end)	0.72	89.5% (F-score in multiclass classification (9 classes))	4
Acoustic Geotagging / Sound Event Detection	0.59	0.66	11.86% (Macro-weighted F-score on USMv3 dataset (25 classes))	4
Audio-Visual Scene Classification / (Apostolidis et al., 2024)	95.1 (Wang et al., 2021)	97.24	2.25% (Classification accuracy on the provided TAU Urban Acoustic Scenes 2020 Mobile challenge dataset)	4

Keyframe Selection and Enhancement (KSE) service Efficiency	1036.5 % (mean baseline processing time as % of video duration for the KSE baseline as reported in D3.1)	63.1 %	93.9 % reduction ($\approx 16.4x$ faster)	4
Keyframe Selection and Enhancement (KSE) service Quality	18.215 (PSNR), 0.399 (SSIM) (KSE baseline as reported in D3.1)	25.128 (PSNR), 0.691 (SSIM)	38% PSNR, 73% SSIM (Reconstruction accuracy vs. 4K reference)	4
Location estimation / Google S2 35K (Im2GPS3k dataset)	8.4 (Pramanick et al., 2022)	16.2	92.9% (Accuracy)	5
OCR (CRAFT+CRNN based architecture)	42.17 (Google Vision API)	47.39	12.4% (Character Accuracy)	5
Image-text decontextualization (no external evidence)/ LAMAR	60.2 (Luo et al., 2021)	84.8	40.8% (Accuracy for binary classification)	6
Image-text decontextualization (evidence-based) / AITR	84.7 (Abdelnabi et al., 2022)	93.3	10.15% (Accuracy for binary classification)	6
Fact-checker-in-the-loop approach /User intent detection	85.00 (WebApp corpus), 94.00 (Ask Ubuntu corpus) (Orhan et al., 2022)	95.27	12% (WebApp), 1% (Ask Ubuntu) (Accuracy of intent detection)	7
Efficient Annotator Reliability Assessment/ RUC-MCD	0.691 (Cook et al., 2024)	0.740	7.09% (F1-macro score)	8

While the majority of methods met or exceeded the KPI threshold of 20% improvement over baseline, there are several cases where the target was not fully reached. As shown in Table 3, OCR performance did not reach the targeted 20% improvement, but achieved a sufficient gain of 12.4%. This result reflects the already high state of the art in the field, as well as the inherent trade-offs between speed and robustness that vary depending on content complexity and language. Our location estimation model showed strong performance on YFCC4k dataset, where we achieved the target value of 20% relative improvement (compared to the state of the art at the beginning of the project) across all five granularity ranges with a notable improvement of 92,9% for 1km. In contrast, on the Im2GPS3k dataset, the target was nearly achieved at the 25km granularity, while performance at the other granularities fell short. The Im2GPS dataset results were overall lower than expected, which we attribute to its limited size and higher levels of annotation noise, negatively affecting model performance. Annotator reliability assessment and audio-

visual scene classification achieved solid performance in absolute terms but did not exceed the relative threshold.

The KPIs for the fact-checker-in-the-loop approach and intent detection are user-centric and include a qualitative assessment of (a) user satisfaction with the feedback collection process, (b) the total number of collected data points, and the overall number of users involved in both (c) controlled and (d) "in-the-wild" evaluation activities. Since the chatbot has not yet been integrated into the user-facing plugin tool, we are currently reporting only the quantitative performance of the user intent detection component. This includes the recognition of both feedback and explanation intents. Due to the highly tool-specific nature of our dataset, many of the existing state-of-the-art (SOTA) benchmarks are not directly comparable to our approach. Instead, we provide performance results on two benchmarking datasets with a comparable training size and number of intents: *WebbApp* (8 intents, 89 utterances), and *Ask Ubuntu* (5 intents, 168 utterances). To ensure comparability, we report our results in terms of accuracy for English only, aligning with the metrics used in the benchmarking datasets, while F1-scores can be found in Section 7. However, we would like to stress that despite the similarity in terms of the dataset size, language, and metrics, our results are not directly comparable to those in Table 3, as our dataset was specifically developed within the project to reflect the unique functionality of the tools we target. The qualitative, user-centric KPIs will be presented in the D2.2, following the completion of the user evaluation activities. To calculate the KPIs for the fact-checker-in-the-loop approach we will use data collected from the user questionnaire administered during the final evaluation cycle. Therefore, the results will be reported in D2.2 due M36.

2.4 Integration into vera.ai User Facing Tools

Significant advancements were achieved in bridging the gap between research prototypes and end user accessibility through systematic integration of developed methods into the vera.ai user-facing tools. The modular and API-driven design of the project enabled deployment of services developed under WP3 into operational environments, including the Verification Plugin, Truly Media, and the vera.ai Assistant.

In the domain of multimodal content verification, a wide range of components—including multilingual persuasion technique classifiers, genre and source veracity predictors, audio scene classifiers, and image geolocation modules—were exposed through REST APIs, allowing dynamic interaction with real-world content. These services have been made available through the central vera.ai API gateway and are actively integrated into verification workflows in WP5.

Notably, the newly developed credibility analysis components were embedded in the Verification Plugin Assistant chatbot, enabling fact-checkers and journalists to receive on-the-fly insights into content framing, writing style, and potential manipulation. Similarly, the keyframe extraction and enhancement service was incorporated into the plugin to support efficient visual evidence selection from videos, combining face and text detection with audio-based indicators.

Furthermore, user-centric explainability and feedback mechanisms, such as the “fact-checker-in-the-loop” interface, were designed to be interoperable with the plugin architecture, ensuring that both expert and non-expert users can engage with AI-generated assessments. Multilingual OCR, geolocation and

cross-modal inconsistency detectors have also been prepared for gradual deployment into the vera.ai ecosystem.

Overall, these integration efforts ensure that cutting-edge AI capabilities developed in WP3 directly serve the practical needs identified by end users in WP2, advancing vera.ai's mission of making trustworthy content verification tools usable, scalable, and aligned with real-world challenges.

2.5 List of Publications and Preprints

Text Misinformation Analysis and Verification (Section 3):

- Leite, J. A., Razuvayevskaya, O., Scarton, C., & Bontcheva, K. (2025, May). A cross-domain study of the use of persuasion techniques in online disinformation. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 1100-1103)
- Hromadka, T., Smolen, T., Remis, T., Pecher, B., & Srba, I. (2023). KInITVeraAI at SemEval-2023 Task 3: Simple yet Powerful Multilingual Fine-Tuning for Persuasion Techniques Detection. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 629–637, Toronto, Canada. Association for Computational Linguistics.
- Razuvayevskaya, O., Wu, B., Leite, J. A., Heppell, F., Srba, I., Scarton, C., ... & Song, X. (2024). Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. Plos one, 19(5), e0301738.
- Leite, J. A., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2025). Weakly supervised veracity classification with LLM-predicted credibility signals. EPJ Data Science, 14(1), 16.
- Wu, B., Razuvayevskaya, O., Heppell, F., Leite, J. A., Scarton, C., Bontcheva, K., & Song, X. (2023, July). SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 1995-2008).
- Leite, J. A., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2024, October). EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 5380-5384).

Audiovisual Content Analysis and Enhancement (Section 4):

- Abeßer, J., Schwär, S., & Müller, M. (2025). Pitch contour exploration across audio domains: A vision-based transfer learning approach. arXiv. <https://arxiv.org/abs/2503.19161> (submitted to the IEEE/ACM Transaction on Audio, Speech, and Language Processing)
- Abeßer, J., Rodríguez Mejía, J. M., Cuccovillo, L., & Aichroth, P. (2024). Siren sounds as acoustic landmarks for content verification. In Proceedings of the Annual Meeting on Acoustics (DAGA), Hannover, Germany.
- Abeßer, J. (2025). Automatic retrieval of indicator sounds for acoustic geo-tagging. In Late-Poster at the Annual Meeting on Acoustics (DAS/DAGA 2025), Copenhagen, Denmark.

Geolocation from a Single Image and Extraction of Text (Section 5):

- Singh, I., Colom, M., & Bontcheva, K. (2025). A Comparative Analysis of OCR Models on Diverse Datasets: Insights from Memes and Hiertext Dataset. In Proceedings of the Winter Conference on Applications of Computer Vision (pp. 1343-1353).
- David Faget, Jose-Luis Lisani, Miguel Colom. Enhancing Aerial Geolocalization Explainability With a Novel Grad-CAM Approach. TechRxiv. October 07, 2024. (Preprint).

Detection of Decontextulised Content (Section 6):

- Apostolidis, K., Abeßer, J., Cuccovillo, L., & Mezaris, V. (2024). Visual and audio scene classification for detecting discrepancies in video: A baseline method and experimental protocol. In Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, MAD '24 (pp. 30–36).
- Chrysidis, Z., Papadopoulos, S. I., Papadopoulos, S., & Petrantonakis, P. (2024, June). Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. In Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation (pp. 73-81).
- Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. (2023, June). Synthetic misinformers: Generating and combating multimodal misinformation. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation (pp. 36-44).
- Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2024). VEIRTE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. International Journal of Multimedia Information Retrieval, 13(1), 4.
- Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025). Red-dot: Multimodal fact-checking via relevant evidence detection. IEEE Transactions on Computational Social Systems.
- Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025, February). Similarity over Factuality: Are we making progress on multimodal out-of-context misinformation detection?. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 5041-5050). IEEE.
- Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025, April). Latent Multimodal Reconstruction for Misinformation Detection. arXiv preprint arXiv:2504.06010.

3 Text Misinformation Analysis and Verification

This section documents the progress of the research outcomes achieved within the vera.ai project related to textual content analysis and classification that have been carried out in T3.1 *Multilingual credibility assessment and evidence retrieval*.

The main aim of T3.1 is to provide support to media professionals through *automatic and explainable extraction of textual credibility signals*. Our goal in this task was to research novel AI-based methods which can automatically detect the presence of various credibility signals that can help media professionals to speed up content analysis. The overarching aspect of methods, which we have developed, is *multilinguality*. It represents not only an interesting and, in many cases so far under-researched potential, but also a high added value when the researched models are being deployed and used by media professionals, who naturally and inevitably work in a highly multilingual environment.

During the course of the task, we have researched, evaluated and deployed automatic extraction of 6 credibility signals, namely detection/classification of: genre, framing, persuasion techniques, subjectivity, authorship and targeted audience.

The prototypes of the first four of them (genre, framing, persuasion techniques, subjectivity) have been introduced in the deliverable D3.1 and in the respective research papers ([Wu et al., 2023a](#); [Hromadka et al., 2023](#); [Schlicht et al., 2023](#)). In the follow-up work, we focused on improving accuracy as well as computational efficiency of training multilingual language models, while comparing the full fine-tuning and parameter-efficient fine-tuning ([Razuvayevskaya et al., 2024](#)). This research resulted in the state-of-the-art persuasion classifier, which we further employed to conduct a so-far missing cross-domain study of the use of persuasion techniques in online disinformation. Furthermore, we introduced two additional credibility signals based on automatic detection of authorship and targeted audience.

Thanks to the successful deployment and integration of these AI-based credibility signals detection methods, altogether with their built-in explainability, we enabled media professionals with an option to provide a textual document (e.g., a news article, a blog post, a social media post, etc.) and obtain an in-depth analysis of its credibility (user need #38, #39, #40). Such analysis complements existing services, as well as provide a more granular and more explainable analysis in comparison with so-far predominant but simplistic fake news detection (indirectly addressing user needs #1 to #10 that focus on usability, explainability, sharing and integration), i.e., typically a binary classification whether a piece of content is true or false, which lacks a necessary transparency required by media professionals.

Finally, we utilized our obtained experience with research on automatic detection of credibility signals as a part of systematic and comprehensive literature review of 175 research papers while focusing on textual credibility signals and Natural Language Processing (NLP) ([Srba et al., 2024](#)).

In the following subsections, we extend the previous deliverable D3.1 and provide more in-depth information on our results on 1) the study of use of persuasion techniques in online disinformation, 2) authorship identification, and 3) targeted audience estimation.

3.1 Methodology

In this section, we present a three-part methodological framework to analyze online disinformation. First (**Subsection 3.1.1**), we examine the use of persuasion techniques by applying a state-of-the-art classifier that identifies rhetorical strategies commonly used to manipulate readers. Second (**Subsection 3.1.2**), we perform authorship identification using a multilingual transformer model to assess whether stylistic patterns can help attribute disinformation to specific individuals or writing profiles. Third (**Subsection 3.1.3**), we estimate the likely socio-demographic profile of a text's audience based on linguistic features, using a pipeline informed by audience marketing data and media source characteristics. Together, these methods provide a comprehensive view of how disinformation is rhetorically constructed, attributed, and targeted.

3.1.1 Use of Persuasion Techniques in Online Disinformation

To investigate how persuasive language is used in disinformation, we apply a state-of-the-art persuasion technique classifier across multiple datasets. In this section, we first describe the classifier and the specific techniques it targets. We then present the disinformation datasets used in our analysis, which span several application domains.

Identifying Persuasion Techniques:

In this study, we leverage the state-of-the-art persuasion classifier introduced in [Razuvayevskaya et al. \(2024\)](#), which uses a RoBERTa-Large pretrained model fine-tuned jointly on several different languages translated into English. The persuasion classifier was trained using the largest dataset of human annotated persuasion techniques, which was introduced in SemEval-2023 ([Piskorski et al., 2023](#)). The data consists of news articles in 9 languages, collected from a variety of mainstream and alternative sources. It is annotated with 23 persuasion techniques at the sentence level, with the task framed as a multi-label classification problem, allowing multiple techniques to be tagged simultaneously within each sentence. Further details about the classifier can be found in [Razuvayevskaya et al. \(2024\)](#). The classifier is currently ranked first on the SemEval task's post-competition leaderboard and is available through an API¹.

The study is focused on the following 16 persuasion techniques: *Loaded Language* (22% of instances in the SemEval training data), *Name Calling-Labeling* (16%), *Doubt* (13%), *Questioning the Reputation* (7%), *Exaggeration-Minimisation* (6%), *Appeal to Fear-Prejudice* (5%), *Repetition* (3%), *Appeal to Authority* (3%), *Slogans* (3%), *Conversation Killer* (3%), *Appeal to Hypocrisy* (3%), *Guilt by Association* (2%), *Appeal to Values* (2%), *False Dilemma-No Choice* (2%), *Flag Waving* (2%), *Causal Oversimplification* (2%).

To improve the reliability of our analysis, we discard the following seven persuasion techniques produced by the SemEval-2023 classifier, due to their low frequency in the training set (below 2%): *Causal Oversimplification* (1%), *Appeal to Time* (1%), *Straw Man* (1%), *Appeal to Popularity* (1%), *Obfuscation-Vagueness-Confusion* (1%), *Red Herring* (1%), *Whataboutism* (1%).

¹ <https://cloud.gate.ac.uk/shopfront/displayItem/persuasion-classifier>

Disinformation Domains:

We study four disinformation datasets from diverse domains (see Table 4): CIDII (Islamic issues) ([Hamed et al., 2023](#)), COVID-19 ([Patwa et al., 2021](#)), Climate Fever (climate change) ([Diggelmann et al., 2021](#)), and EUvsDisinfo (Russo-Ukrainian war) ([Leite et al., 2024](#)). All texts are in English, except for EUvsDisinfo, which spans 41 additional languages where we translated the non-English texts to English using GPT4o. EUvsDisinfo also includes topics beyond the Russo-Ukrainian war, which we filtered out. Specifically, the topics are already annotated in the original dataset under the ‘TAGS’ field of each EUvsDisinfo article. We used these tags to filter out articles unrelated to the Russo-Ukrainian war, retaining only those explicitly labeled with relevant tags. The articles in all these datasets are human labelled. In this study specifically, we analyse articles human labelled as disinformation and exclude all trustworthy articles across all datasets. The texts are split into sentences using NLTK, and the persuasion classifier is used to identify persuasion techniques within each sentence.

Table 4 Summary statistics of disinformation datasets analyzed for persuasion techniques

Domain	Islamic Issues	COVID-19	Climate Change	Russo-Ukrainian War
# Sentences	1,687	7,369	254	10,240
Avg. Sent. Length	102.3	95.9	109.0	150.7
Avg. PTs per Sent.	1.1	1.1	1.2	1.4

3.1.2 Authorship Identification

Attributing a text to its true author is a valuable tool in combating disinformation, particularly when misleading content is falsely associated with credible figures to enhance its perceived legitimacy. These tactics erode public trust and complicate verification efforts. To address this challenge, we present a multilingual authorship identification system designed to analyze writing patterns in news articles and detect stylistic similarities between authors.

We collected a multilingual dataset of news articles in English, French, German, and Italian, focusing on mainstream sources with high readership to ensure diversity and relevance. Articles were retrieved using the Media Cloud API² and a custom web scraper to expand source coverage. From this dataset, we selected a subset of authors with sufficient and diverse content to support effective model training and evaluation. Specific authors were chosen to represent a wide range of outlets and topics to increase stylistic diversity and minimise source specific bias.

To help the model focus mainly on underlying writing style, we applied several preprocessing steps aimed at minimising bias. This included filtering and cleaning the text, anonymizing it by removing references to media outlets, author names and linguistic artifacts that could risk leaking source identity. To further reduce reliance on topical, contextual or time sensitive signals, we replaced named entities such as people, places and organisations with generic labels.

² <https://www.mediacloud.org/>

We fine-tuned XLM-RoBERTa, a multilingual transformer model using a triplet loss objective. This training strategy pushes the model to map articles from the same author closer together in the embedding space while pushing apart those from different authors especially in challenging cases (for example authors writing about similar topics or for the same media outlet). This results in compact representations of authorship profiles. At inference time, new texts are encoded and compared to reference embeddings to identify the most similar authors. This allows the system to incorporate new unseen authors by adding reference articles without having to retrain the underlying model.

3.1.3 Targeted Audience Estimation

Disinformation often targets specific demographic groups, exploiting their characteristics to increase its persuasive impact. Identifying the likely audience of a given text helps to uncover the strategic intent behind such narratives and informs mitigation efforts. We designed a pipeline that predicts a reader's sociodemographic profile, including age, gender, income, education, and urbanity, based solely on article content. This approach uses audience marketing data from approximately 200 major media outlets across five European countries (France, Italy, Spain, UK, Germany)

To address overlap in audience demographics across sources, we generated synthetic reader populations using a Bayesian sampling method based on joint probability distributions derived via iterative proportional fitting. This allowed us to create realistic reader profiles aligned with observed audience characteristics. We then clustered these synthetic individuals into distinct socio-demographic groups and assigned each cluster to its most representative media sources. This process amplifies distinctions between sources and strengthens the link between articles and their likely target audience.

We then trained a multi-output regressor on EuroBERT embeddings to infer the socio-demographic characteristics of article audiences. By modeling the targets jointly, the system captures underlying correlations between variables such as income, education, and urbanity, leading to more coherent and robust predictions. This approach demonstrates that meaningful audience profiles can be extracted directly from text, relying solely on linguistic patterns.

3.2 Evaluation

This section presents evaluation outcomes from three application scenarios in disinformation analysis, covering the use of persuasion techniques across domains (3.2.1), multilingual authorship identification (3.2.2), and audience profiling based on socio-demographic inference (3.2.3). In Subsection 3.2.1, we provide a comparative analysis of persuasion techniques across four disinformation domains (Islam, COVID-19, climate change, and the Russo-Ukrainian war), highlighting both their prevalence and rhetorical adaptations through statistical and linguistic indicators. In Subsection 3.2.2, we evaluate a multilingual authorship attribution pipeline, showing significant improvements over semantic baselines in identifying author-specific stylistic patterns across English, French, German, and Italian. In Subsection 3.2.3, we present a two-part approach to audience profiling, combining synthetic population modeling with multilingual text regression to infer socio-demographic traits from news content, achieving high accuracy for attributes such as gender, income, and urbanity. Collectively, these evaluations demonstrate the robustness and applicability of the developed methods for real-world disinformation detection and analysis.

3.2.1 Use of Persuasion Techniques in Online Disinformation

Figure 1 shows the proportion of persuasion techniques in each dataset, offering a visual comparison. To complement this and enable a more detailed quantitative analysis, we calculate the odds ratios (ORs) for each technique between datasets in Table 5. We calculate ORs by comparing the odds of a technique appearing in one domain to its average proportion in the other three domains. Throughout our analysis, we only discuss statistically significant ORs, based on Fisher's exact test with $p < 0.05$.

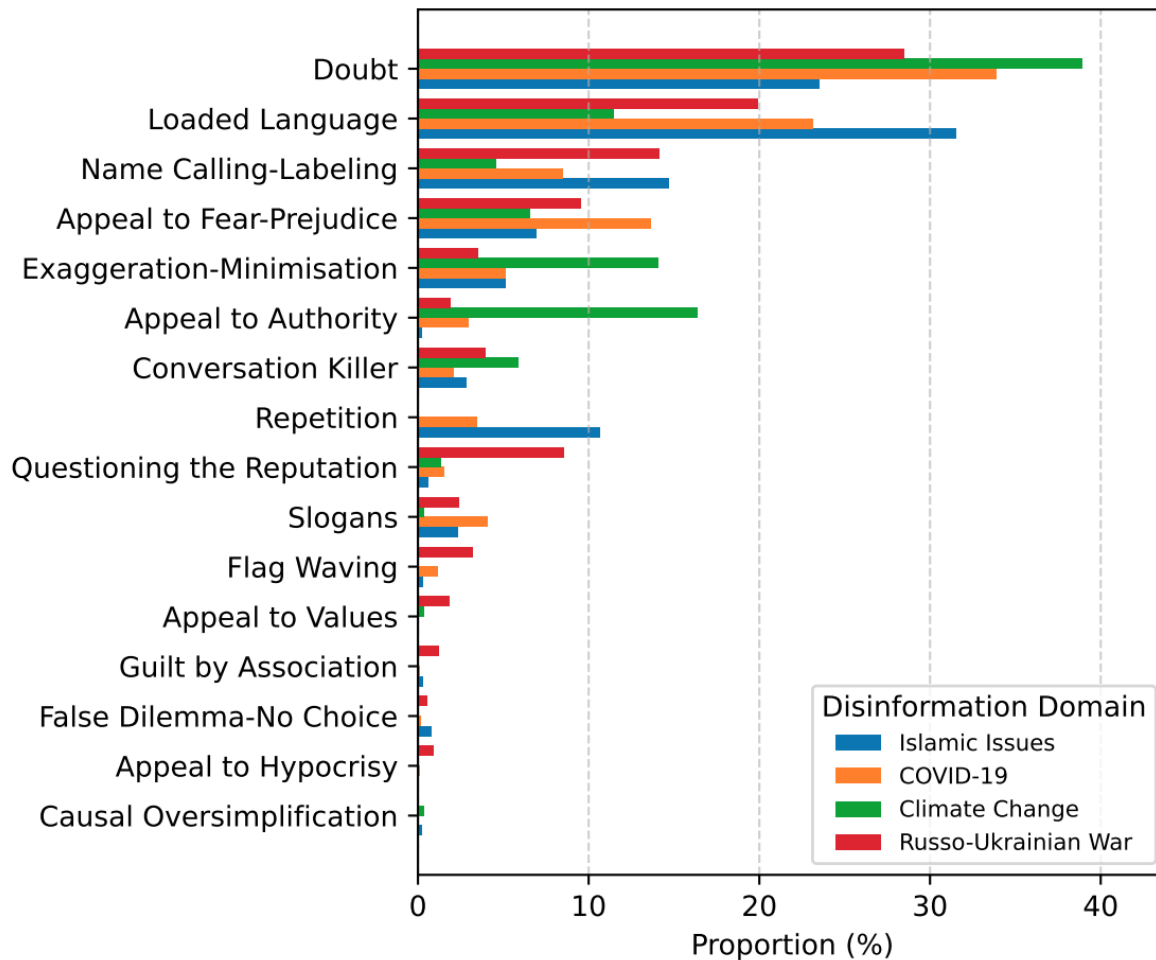


Figure 1 Proportion of persuasion techniques in the different disinformation narratives

We observe that *Loaded Language* and *Doubt* are used ubiquitously across all disinformation domains, comprising more than 20% of the techniques in each, except *Loaded Language* in climate change (11%). Specifically *Doubt* attempts to undermine trust in credible sources or established facts, creating confusion and making audiences more susceptible to accepting misleading or false claims ([Proctor & Schiebinger, 2008](#)). *Loaded Language* aims to evoke strong emotional responses, such as fear or anger, which can override rational analysis and lead individuals to accept false information without critical scrutiny ([Gennaro & Ash, 2022](#)).

Annex I discuss the persuasion techniques with odds ratios (ORs) greater than 1 and statistical significance, indicating that their usage is more prevalent in one domain compared to the others.

Table 5 Odd ratios of occurrence of persuasion techniques in one domain vs. the average proportion in the others. Statistically significant ratios are underlined (with $p < 0.05$ — Fisher exact test). A → Islamic issues, B → COVID-19, C → Climate change, D → Russ

Persuasion Technique	A	B	C	D
Doubt	0.60	1.18	1.59	0.84
Loaded Language	2.07	1.14	0.39	0.88
Name Calling-Labelling	1.72	0.74	0.34	1.61
Appeal to Fear-Prejudice	0.67	1.90	0.63	1.06
Exaggeration-Minimisation	0.66	0.66	3.39	0.41
Appeal to Authority	0.03	0.46	11.37	0.28
Conversation Killer	0.70	0.49	2.06	1.10
Repetition	10.15	0.98	0.01	0.01
Questioning the Reputation	0.15	0.43	0.36	8.06
Slogans	1.04	2.47	0.11	1.06
Flag Waving	0.19	0.98	0.01	6.99
Appeal to Values	0.06	0.06	0.51	13.26
False Dilemma-No Choice	3.33	0.37	0.01	1.72
Guilt by Association	0.64	0.22	0.01	9.45
Appeal to Hypocrisy	0.14	0.30	0.01	19.28
Causal Oversimplification	1.97	0.13	3.88	0.01

3.2.2 Authorship Identification

We evaluated the model on a balanced multilingual test set composed of authors writing in English, French, German, and Italian. Performance was measured using top-1 and top-3 accuracy, both by language and in aggregate (see Table 6). To assess whether the model identifies stylistic features of authorship rather than relying on semantic and topical similarity, we compared it to a multilingual semantic embedding model used in BERTopic and optimized for similarity search. As shown in the following table, there was a consistent improvement across all languages, indicating that the fine-tuned pipeline is more sensitive to authorship related patterns.

We further evaluated the model's generalization capabilities by testing on unseen authors excluded from embedding fine-tuning. This setting assesses the pipeline ability to attribute authorship based on similarity in the learned embedding space, without retraining.

Table 6 Comparison of Authorship Attribution Accuracy by Language and Model for Trained and Unseen Authors

	Top-1 Accuracy (%)		Top-3 Accuracy (%)	
	Semantic Model	Authorship Model	Semantic Model	Authorship Model
English	56.5	91.1	72.9	96.8
French	26.4	78.2	48.2	95.3
German	30.0	78.6	55.3	95.3
Italian	46.9	91.5	68.5	96.8
Total	40.0	84.8	61.2	96.0
Unseen Authors	40.1	63.8	61.1	93.3

To estimate the role of reference content, we also varied the number of reference articles used for unseen authors. Figure 2 shows that accuracy improves with more reference texts, though gains tend to stabilise after a certain point, suggesting diminishing returns beyond a basic stylistic representation.

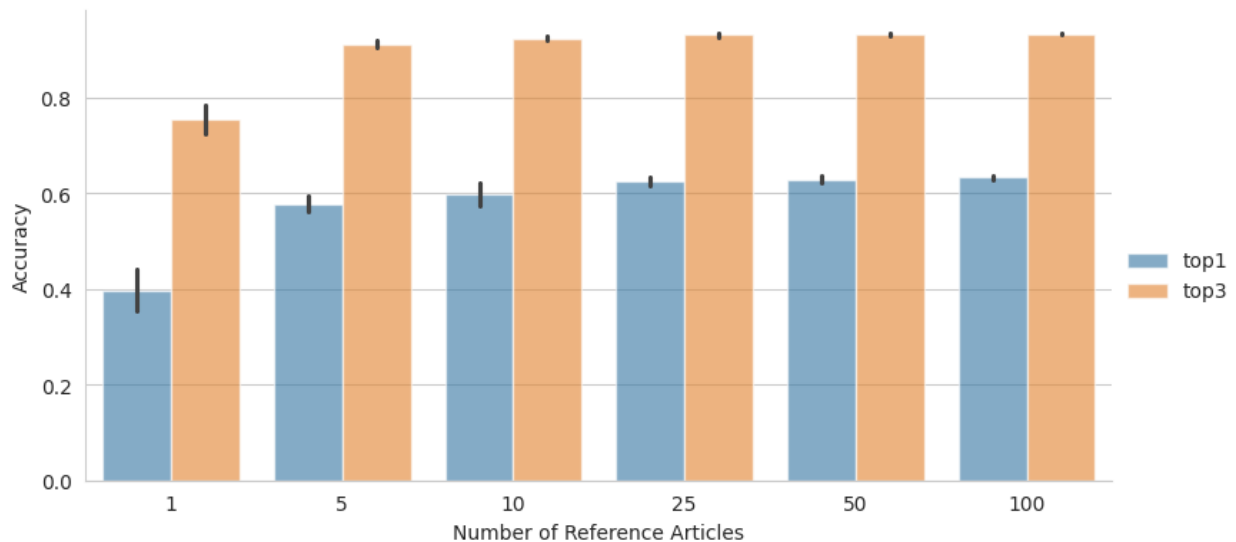


Figure 2 Impact of Number of Reference Articles on Accuracy for Unseen Authors

To further validate these findings, ongoing work involves expanding the number of authors and languages included in the study. These extensions aim to assess the robustness of the approach and explore its applicability to a wider range of real-world authorship attribution scenarios.

3.2.3 Targeted Audience Estimation

To investigate the targeted audience, we developed a methodology for estimating the likely socio-demographic profile of a news article's readership. This approach involves two key components: the construction of synthetic population distributions and the use of a multilingual transformer-based regressor.

Synthetic Population Distributions

We present a methodology for estimating the socio-demographic profile of a news article's likely readership based exclusively on its textual content. The pipeline predicts five core audience traits: age, gender, income tier, education level, and urbanity, using a combination of multilingual language embeddings, structured regression models, and a synthetic training dataset derived from real-world audience statistics.

The foundation of the synthetic data is GWI Core Plus, a large-scale, continuously updated research panel that provides marginal distributions of socio-demographic attributes for over 200 mainstream newspapers and magazines across five European countries: France, Italy, Spain, Germany, and the United Kingdom.

Because GWI only provides marginal (univariate) distributions per outlet, it does not capture inter-variable dependencies such as those between income and education or between urbanity and age. To address this, we introduced a two-step population synthesis approach. First, we reconstructed joint distributions using Iterative Proportional Fitting (IPF) applied to pairs of features. Second, we used Bayesian sampling over these estimated joint distributions to generate complete, individual-level reader profiles. This procedure produces high-resolution synthetic populations that align with observed marginals while preserving the structure of conditional relationships across variables. To evaluate the quality of the synthetic data, we computed Kullback-Leibler (KL) divergence scores between the synthetic and reference marginal distributions for each socio-demographic feature, at the level of individual media brands (Figure 3).

As illustration for French journal/magazine the maximum KL divergence observed across all features and sources is 0.00048, computed over 20,000 synthetic readers per journal or magazine.

Overall, the results confirm that our synthetic population generation process maintains close alignment with known statistical properties of media audiences, enabling the training of supervised models with confidence that no substantial demographic bias is introduced.

Text-Based Audience Estimation using Multi-Output Regression

After cleaning the text from provenance metadata, authorship cues, disclaimers, and other non-linguistic elements, articles were encoded using EuroBERT, a recent transformer-based language model optimized for high-fidelity representation of European-language content. Each article was mapped to a 768-dimensional embedding capturing both semantic and stylistic attributes. We trained a multi-output regressor on these embeddings to predict the socio-demographic characteristics of the audience, leveraging synthetic profiles generated via IPF and Bayesian sampling. By modeling the joint structure of the targets, the model captures correlations across traits, leading to substantial gains in accuracy.

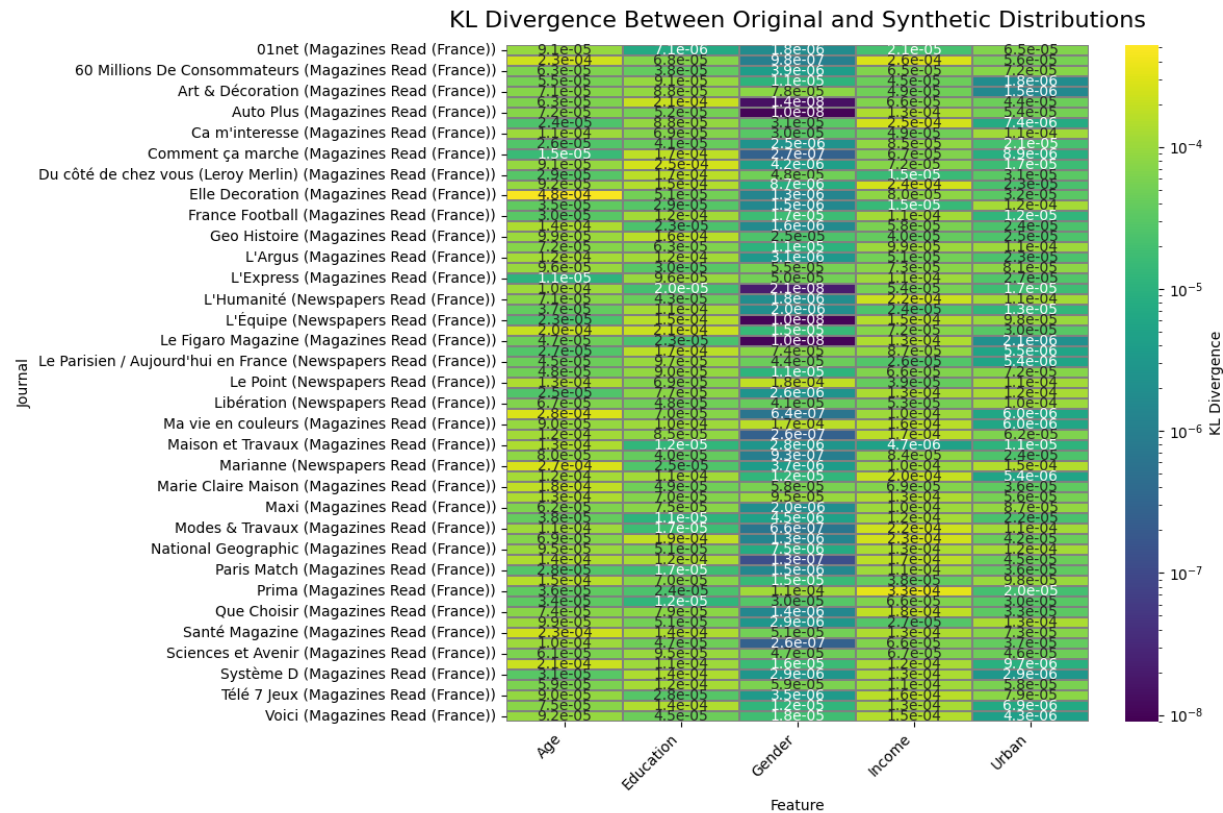


Figure 3 KL Divergence between Synthetic and Reference Distributions

The results (Table 7) show good performance, particularly for gender ($R^2 = 0.790$), income ($R^2 = 0.734$), and urbanity ($R^2 = 0.683$), indicating strong signal in the text for these attributes. Education ($R^2 = 0.489$) and age ($R^2 = 0.294$) are more challenging but still yield meaningful predictions. These outcomes confirm that socio-demographic profiling from content is feasible and effective when grounded in linguistic modeling and statistically aligned synthetic data.

Table 7 The R2 results for individual features

Feature	R ²
Age	0.294
Income	0.734
Education	0.489
Urbanity	0.683
Gender	0.790

3.3 Implementation and Integration

This section presents the practical integration of three key components into real-world applications: credibility signal detection (3.3.1), authorship identification (3.3.2) and targeted audience estimation (3.3.3). Each component was developed with a focus on efficiency, scalability, and user accessibility, leveraging a shared infrastructure and modular RESTful APIs. These systems have been incorporated into the Verification Plugin and its associated interfaces, enabling both technical and non-technical users to perform advanced content analysis. The following subsections detail the implementation strategies, model choices, and interface designs for each functionality.

3.3.1 Credibility Signals in the Assistant Tool

To make the credibility signal detectors usable in real-world scenarios, we prioritized efficiency and responsiveness. Due to the limitations of running multiple language-specific classifiers in real-time, we opted for single multilingual models as the final deployed solution. Specifically, we used mBERT for genre and persuasion detection and distilled mBERT for framing detection.

This integration was realized in the *Assistant* feature of the Verification Plugin, a widely adopted tool supporting both media professionals and the general public. The plugin provides a suite of AI-based tools for verifying online content, including deepfake detection, image manipulation analysis, and fact-check retrieval. The *Assistant* module focuses specifically on the analysis of textual content.

Multiple credibility signals were incorporated: news framing, news genre, persuasion techniques, subjectivity. The classifiers are hosted on remote servers and exposed via RESTful API endpoints. When a user submits a URL or text to the *Assistant*, the extracted content is automatically sent to each classifier through GET or POST requests. The system retrieves the results asynchronously, displaying them in individual tabs within the interface. Tabs remain greyed out and unclickable until results are fully received. In case of processing failures, the system notifies the user and disables the affected tabs.

The user interface presents the analysis in a visually interpretable form. For each classifier, key sentences are highlighted within the original text based on their contribution to the classification, with color gradients indicating importance or confidence. They are displayed in separate tabs within the section named “Extracted text”. A separate “Raw text” tab provides the unannotated version of the input.

- **News framing** is handled by a multi-label classifier predicting up to nine frame categories. The interface displays the extracted text on the left with highlighted sentences, and a list of predicted frames on the right. Sentence highlights can be toggled on or off, and a confidence-based color scale helps users interpret the results. If no frames are detected, a “No detected topics” message is shown.
- **News genre** prediction involves one of three labels: objective reporting, opinionated news, or satire. Sentence-level highlights illustrate the basis for the prediction, while a single class label is displayed with an accompanying confidence indicator.
- **Persuasion techniques** involve 23 possible classes, and a single sentence may contain multiple techniques. The 23 possible classes of persuasion technique are listed in Table 8. Relevant sentences are highlighted, and hovering reveals a pop-up with the detected strategies and their confidence values. On the right, users can see the list of predicted classes, and clicking on one, it

filters the highlights to sentences associated with that class. Clicking the class again resets the view. An example of an interface is provided in Figure 4.

- **Subjectivity detection** highlights only those sentences identified as subjective. Confidence values are again shown through a color gradient. If no subjective content is found, the interface displays “No detected sentences.”

Table 8 Distribution of articles in the training set for persuasion technique detection task

Persuasion Technique	Number of articles						Total
	EN	FR	IT	RU	PL	DE	
Doubt	518	327	882	509	295	288	2819
Repetition	544	92	22	69	13	8	748
Appeal_to_Fear-Prejudice	310	210	285	54	108	182	1149
Appeal_to_Authority	154	76	70	10	41	225	576
Slogans	153	149	54	72	36	87	551
Loaded_Language	1809	944	903	641	310	242	4849
False_Dilemma-No_Choice	122	73	61	28	12	41	337
Flag_Waving	287	37	35	42	68	65	534
Name_Calling-Labeling	979	428	566	253	475	743	3444
Causal_Oversimplification	213	125	50	39	12	33	472
Appeal_to_Hypocrisy	40	134	82	103	162	136	657
Obfuscation-Vagueness-Confusion	18	113	21	19	36	62	269
Exaggeration-Minimisation	466	258	143	131	111	157	1266
Red_Herring	44	55	23	2	12	30	166
Guilt_by_Association	59	130	53	24	94	122	482
Conversation_Killer	91	170	178	88	50	121	698
Appeal_to_Popularity	15	82	37	8	30	63	235
Straw_Man	15	135	51	21	15	15	252
Questioning_the_Reputation	0	348	383	303	164	310	1508
Appeal_to_Time	0	41	27	28	14	11	121
Whataboutism	16	62	8	7	8	13	114
Appeal_to_Values	0	100	131	48	101	73	453
Total	5,853	4,089	4,065	2,499	2,167	3,027	21,700

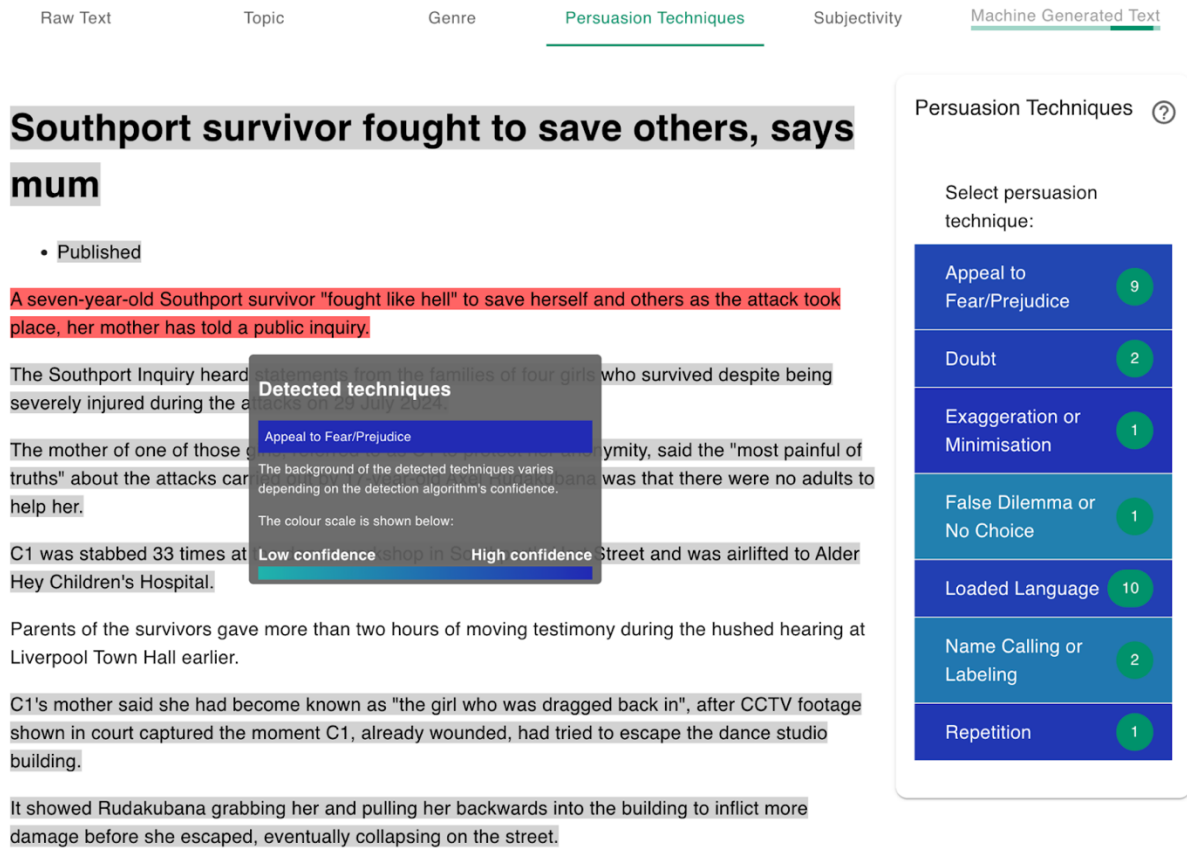


Figure 4 User Interface example for Persuasion Techniques

This practical integration demonstrates how research-grade credibility signals can be adapted into real-time, user-facing tools. The combination of backend classifier infrastructure, API-driven deployment, and a carefully designed UI allows for detailed content analysis that remains accessible and informative to end users.

3.3.2 Authorship Identification

The system is accessible via a REST API and integrated into a user-friendly interface, enabling real-time authorship attribution for use in journalism, content moderation, and forensic analysis.

Two implementation considerations influenced our design and model training to prepare the system for practical integration and deployment. We aimed to keep the model size reasonable while choosing a multilingual model not only to leverage cross-lingual transfer capabilities but also to enable deployment of a single model across all supported languages. This approach allows us to simplify deployment and reduce resource requirements.

The model is exposed through a REST API which offers two ways of interaction. Users can either submit a URL of a news article, in which case the article content is automatically extracted, or directly provide the text of the article. The system returns the top five authors from the reference set whose text content and writing styles are closest to the submitted article, along with a similarity score. A higher score means greater proximity and potentially, higher likelihood of authorship.

The API also supports the possibility of adding new authors at inference time, without retraining the embedding model. Users can include the name of the new author and a list of reference article URLs as part of the request payload. The more reference articles provided, the more reliable the attribution will be.

To make the system accessible to non technical users, we developed a standalone demonstration interface (Figure 5). This web based tool allows users to submit an article URL or text and view the top 5 closest authors. The results are displayed with a color scale to help interpret the similarity scores. The interface also displays the list of available authors in the system and provides a form to add new ones with their reference articles to extend the author pool interactively.

Authorship Attribution

Input article URL Input text article

URL

Enter an article URL:

<https://www.theguardian.com/uk-news/2025/jan/27/hoard-of-coins-dating-from-roman-conquest-of-britain-found-near-utrecht>

Submit URL

Ancient British coins found in Dutch field likely to be spoils of Roman conquest

A hoard of British coins bearing the inscription of King Cunobelin and found in a Dutch field have been identified as very likely to be the spoils of war of a Roman soldier from the conquest of Britain.

The 44 gold coins, known as staters, were discovered alongside 360 Roman coins, by two amateur archaeologists with metal detectors in a field in Bunnik, near Utrecht. The coins are believed to have been given as military pay.

The staters bear the name of the British Celtic king also known as Cunobelinus, immortalised by Shakespeare as Cymbeline in the play of that name, who reigned between AD5 and AD40 in the south-east of Britain.

Analysis of what is the first mixed composition collection found on mainland Europe suggests the coins were deliberately buried in a shallow pit and stored in a cloth or leather pouch.

The coins, which were found less than 30cm below the surface of the soil, are said to amount to what would have been 11 years in wages for an ordinary Roman soldier.

Four of the British staters are regarded as posthumous issues, probably struck by Cunobelinus's successors as ruler of the Catuvellauni tribe, the brothers Togodumnus and Caratacus, at approximately AD43.

Authorship Proximity

Author

Daniel Boffey

Maxime Poul

Julie Cloris

Florian A. Lehmann

Andreas Frei

Colour scale

low proximity

high proximity

Figure 5 User Interface example for Authorship Attribution

3.3.3 Targeted Audience Estimation

The system is accessible via a REST API and integrated into an interactive interface (Figure 6), enabling users to analyze content, explore predicted audience traits, and monitor demographic targeting in real time.

Similar to the authorship identification system, the audience profile estimation system was designed with deployment and integration in mind. The model is shared through the same REST API structure, accepting either a news article URL or a raw text input. If a URL is submitted the system automatically extracts and processes the article content. Otherwise, users can submit a title and full text directly. This unified structure ensures seamless integration for both tasks.

The API returns predictions for the following five audience dimensions: Age, Gender, Income, Urban context, and Education level. All five are modelled as continuous values, allowing for more nuanced estimations. For example, Age is predicted as a float between 16 and 64, and Gender is positioned along a spectrum from female to male. Income, Urban context, and Education are also output as continuous

values, which can be interpreted using predefined intervals that correspond to familiar audience categories (low to highest income, urban, suburban and rural environments, or education levels from lower secondary to postgraduate). These quantized interpretations help normalize for demographic differences across countries, making the system adaptable to diverse contexts.

The model is integrated into the same application used for authorship identification. Users can input an article and receive a visual overview of the estimated audience. Each dimension is displayed with a progress bar and position indicator, helping to interpret where the content likely falls within demographic ranges. This setup allows for a quick evaluation of content orientation and provides a general audience profile without requiring technical expertise.



Figure 6 User Interface for Audience Profile Estimation

3.4 Concluding Remarks

This section presented an in-depth empirical analysis of textual disinformation, focusing on three complementary dimensions: persuasion techniques, authorship, and target audience. First, we conducted a large-scale, cross-domain study of sixteen persuasion techniques in disinformation narratives related to COVID-19, climate change, Islamic issues, and the Russo-Ukrainian war. While techniques such as *Doubt* and *Loaded Language* were prevalent across all domains, others showed strong domain specificity. For example, *Repetition* and *False Dilemma* in Islamic narratives, and *Appeal to Hypocrisy* and *Guilt by Association* in the war context. Our results also highlight how linguistic style adapts to domain: in climate change narratives, *Appeal to Authority* tends to be more formal and analytic, while *Exaggeration-Minimisation* relies on cultural and moral framing.

Second, we evaluated a multilingual authorship attribution model, demonstrating that it effectively captures stylistic signatures across English, French, German, and Italian. The model significantly outperforms semantic similarity baselines and generalises well to unseen authors when provided with sufficient reference material. These results show the feasibility of attributing disinformation content even when explicit metadata is unavailable.

Third, we estimated the socio-demographic profiles of likely audiences using a regression model trained on synthetic populations. The model achieved strong predictive performance on variables such as gender, income, and urbanity, suggesting that disinformation content carries implicit signals about its intended audience. Notably, this audience-targeting analysis complements our rhetorical findings, indicating that both form and content are adapted to fit the perceived worldview of specific groups.

These findings provide a multi-layered understanding of how disinformation is constructed, stylised, and aimed at influencing particular users. Our work demonstrates the analytical value of applying state-of-the-art models to diverse and multilingual datasets, offering insights that can inform detection tools, media literacy initiatives, and future research on disinformation dynamics.

4 Audiovisual Content Analysis and Enhancement

This section describes the work conducted as part of T3.2: Audiovisual Content Analysis and Enhancement. Informed by the user needs defined in WP2, the overall goal includes capabilities such as detecting faces (user need #36), identifying individuals (#37), recognizing events associated with sound (#34), enhancing or deblurring images (#18), and extracting geolocation-relevant audio cues (#28).

All non-audio user needs were addressed by the Keyframe Selection and Enhancement (KSE) service described in D3.1, where we implemented and validated a suite of visual analysis tasks, i.e. face and text regions detection, image enhancement.

Building on that foundation, the current focus is twofold: refining the visual methods in KSE based on feedback from multiple participatory evaluations (PE), WP2 assessment cycles, and service integrations over the past year, and developing AI-based audio processing methods to meet the remaining audio-related requirements. To broaden the KSE service's utility, we have integrated specialized sound event detectors into the KSE pipeline, enabling keyframe selection not only based on visual information but also on the timing of salient sound events. This allows us to select a wider set of keyframes, including moments that coincide with detected sound events (e.g., sirens, crowd applause).

To meet the audio-related needs, we developed a suite of specialized audio processing methods. These can operate either as standalone tools, providing modality-specific signals to support complex tasks such as geolocation, or as audio feature extractors integrated into multimodal analysis pipelines alongside image-based methods such as the KSE.

In parallel, we systematically addressed the user feedback collected from participatory evaluations (PE), WP2 assessment cycles, and KSE service integrations throughout the past year. This feedback was organized to define priorities and guide iterative improvements to the service.

The remainder of this section first situates our work within the current state of the art in audio analysis, followed by a detailed description of the developed methods, including salient sound event detection (4.2.1), acoustic scene classification and audio-based geolocation cues (4.2.2 and 4.2.3). Finally, we present the methodology for collecting and organizing received feedback (4.2.4), along with the specific corrective actions taken (subsections 4.2.4.1 through 4.2.4.10).

4.1 Background

This section establishes the technical foundations for audio analysis relevant to multimedia services like KSE (4.1.1) and documents the evaluation and deployment journey of the KSE service itself (4.1.2).

4.1.1 Audio Analysis

The audio analysis methods relevant to this work build upon a wide range of research domains and techniques, some of which directly inform the development and improvement of the KSE service. These include Detection and Classification of Acoustic Scenes and Events (DCASE), Computational Auditory Scene Analysis (CASA), and soundscape research.

The methods for analyzing audio data described in this section have been influenced by various research disciplines, such as Detection and Classification of Acoustic Scenes and Events (DCASE), Computational Auditory Scene Analysis (CASA), and soundscape research. In the last 10 years, there have been enormous technical advances in the detection of individual sound events (sound event detection), which have been particularly stimulated by the adaptation of deep learning methods from other areas such as computer vision. As summarized in ([Mesaros et al., 2021](#)), the greatest challenges for SED lie in the large number of real-world sound classes and the complexity of real sound scenes, which are characterized by the overlay of many simultaneously audible sounds. Methodologically, convolutional recurrent neural networks (CRNNs) and transformer-based models are particularly noteworthy, as they are used for the temporal detection and modeling of sound events in spectrogram representations of audio signals.

Siren sound detection and classification has been approached especially in the context of smart city research. For instance, [Damiano et al. \(2024\)](#) introduce a deep learning model to classify emergency siren sounds, focusing on better generalization through data augmentation. Due to the challenge of acquiring diverse siren datasets, they create synthetic data by simulating moving siren sources with acoustic effects and blend them with actual urban noise. They train several CNN classifiers using spectrograms from these synthetic audio scenes and test them on unseen real-world recordings. This approach of spectrogram analysis combined with extensive data augmentation led their models to generalize better than those trained on limited real recordings. [Marchegiani and Newman \(2022\)](#) developed a deep learning system for identifying and classifying emergency sirens (and alarms like car horns) amid noisy city sounds, while also pinpointing their source. Their method transforms stereo audio into spectrograms and uses a U-Net style CNN to semantically segment the spectrogram, isolating siren/horn sounds from background noise. This multi-task framework classifies the alarm type and cleans the audio by extracting the alert signal. The cleaned audio helps the system determine the sound's direction-of-arrival, integrating geo-location by estimating the siren's origin. This spectrogram-based CNN method achieved about 94% accuracy in correctly identifying siren/horn sounds and accurately localizing them despite challenging urban noise.

In the context of acoustic geo-tagging (AGT), a first study on categorizing cities using acoustic scene recordings ([Bear et al., 2019](#)) assessed different modeling paradigms for both individual and combined acoustic scene classification (ASC) and AGT. Utilizing the DCASE 2018 ASC dataset with recordings from six European cities, the study showed that a multitask learning approach combining both tasks achieved a city classification accuracy of around 57 %. The exploration of multimodal location estimation in videos was a key focus of the "Placing Task," conducted as part of the MediaEval Benchmark³ from 2012 to 2016. In the study by [Lei et al. \(2012\)](#), a system for classifying cities was introduced that utilized acoustic feature vectors comprising Mel Frequency Cepstral Coefficients (MFCC), employing a hybrid of Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). [Elizalde et al. \(2016\)](#) applied these same acoustic feature vectors to estimate video audio tracks as a weighted combination of urban sound-derived acoustic feature vectors. This approach highlighted distinct sound patterns among 18 prominent cities worldwide. [Kumar et al. \(2017\)](#) extended the concept of segmenting audio clips into distinct sounds by using a Non-Negative Matrix Factorization (NMF) variant, deconstructing acoustic feature vectors that included MFCC features as well as their first-order and second-order derivatives.

³ <http://www.multimediaeval.org/>

Finally, Acoustic Scene Classification (ASC) deals with recognizing the environment of an audio recording (e.g. park, metro station, bus) from its sound. In recent years (2019–2024), ASC research has rapidly evolved, driven largely by deep learning and benchmark challenges like DCASE. In the last decade, ASC has matured from basic CNN classifiers to sophisticated systems employing hybrid and attention-based neural networks, self-supervised pretraining, extensive augmentation, and meta-learning for adaptation. These developments, nurtured by competitive evaluations like DCASE, have significantly improved ASC accuracy and robustness on real-world audio. The field continues to evolve with interest in making models more efficient and generalizable, ensuring that ASC technology can be deployed in diverse devices and environments. These advances in audio analysis directly enable multimodal services like KSE to incorporate sophisticated sound recognition capabilities alongside visual analysis.

4.1.2 Evaluation, Integration, and Outreach of the KSE Service

While the previous section provided an overview of foundational methods and research that inspire or inform features relevant to the KSE system, this section shifts focus to practical aspects of the KSE service itself. It covers how the service was evaluated, integrated, and disseminated across platforms and partner environments. It documents the evaluation feedback, integration efforts, and dissemination activities for the KSE service, covering insights from PE sessions, WP2 evaluations and implementation in verification tools.

Feedback from Participatory Evaluation Sessions

The KSE participated in the Participatory Evaluation (PE) procedure in late 2024. PE employs a flexible, user-centric approach that prioritises real workflows over technical metrics. Participants evaluated feature usefulness, text/face detection relevance, speed-quality trade-offs, version improvements, usability, and integration potential. They provided written feedback by January 10, 2025. Key feedback included requests for clearer error messages, progress indicators, timeline decluttering, keyframe zoom, shareable URLs, downloadable ZIP archives, expanded format support, and mobile upload improvements. Feature suggestions included configurable processing profiles, integration with geolocation, reverse image search, and fact-checking databases, summary generation from annotations, detection of anomalies such as disappearing fingers, blood detection, language and transcript analysis, and partial video processing via timestamps. Users also requested linking detected faces and text to timeline positions, and batch video analysis. Face detection feedback raised concerns about adversarial misuse, called for clear communication of its purpose, and suggested improvements in profile view detection and confidence score presentation. Opinions on the timeline interface varied, with some users appreciating shot change indications, while others preferred improvements.

Feedback from WP2 Evaluation Cycles

The KSE service was evaluated during WP2's second (Feb 2nd half - March 31, 2024) and fifth (Feb 2nd half - March 31, 2025) assessment cycles. Major technical issues identified included facial hallucination artifacts, fragmented text region detection, delays in keyframe extraction, and requests for denser keyframe output. Minor interface issues were also noted, such as non-functional viewer links on macOS browsers that required manual correction. The detection of subtle visual anomalies like transient object disappearance was suggested as a new feature. A non-optimal communication between two python packages manifested as degraded output resolution relative to source video frames. Additionally, the

omission of certain video segments was traced to overly sensitive thresholds in the blur detection algorithm, which was corrected through parameter tuning.

To improve the quality of selected keyframes, we incorporated a mechanism that evaluates the sharpness of each frame using the variance-of-Laplacian-based blur detection algorithm (see Section 3.2 of D3.1). This step helps ensure that frames with extremely low visual clarity, such as those transitioning from color-uniform scenes to heavy motion-induced blur, are excluded from keyframe selection. However, early testing revealed that the algorithm’s sensitivity occasionally led to the unintended omission of useful, mildly blurred frames. Notably, some of these omissions were reported by AFP integrators, as detailed in Table 10. The issue was traced back to overly strict threshold settings. By carefully tuning the parameters, we achieved a more balanced configuration that preserves meaningful content while still filtering out severely blurred frames.

Integration into the Verification Plugin

KSE replaced the previous Video Fragmentation service in the Verification Plugin, developed by AFP, significantly improving video fragmentation, keyframe extraction, and introducing face/text enhancement features. Integration was supported through OpenAPI v3 documentation and adjustments to the JSON output format, developed in close cooperation with integrators. Continuous technical support was provided during integration. Feedback from the main integrators (AFP), noted suboptimal keyframe quality, minor bugs in the frontend’s operation, and requested backend support for local file uploads.

Integration into Truly Media

Truly Media, co-developed by ATC and Deutsche Welle, is widely used by journalists, fact-checkers, and investigators from organizations such as EDMO, ZDF, Reuters, Amnesty International, and FactCheck Mongolia. We worked closely with ATC to ensure the stable integration of KSE into the platform. While the feedback collected by ATC focused on the UI of the integrated version in Truly Media, it also gave ideas for improvements to the standalone UI.

Key suggestions included adding shot change indicators to the timeline and enabling manual keyframe selection during video playback. Users requested email notifications with access links, persistent dashboards for resuming projects, and timeline linking for detected faces and text. Additional requests included clearer progress indicators during processing and informative error messages if processing stalled. Users also proposed deeper investigation tools combining KSE with reverse image search, geolocation, synthetic media detection, transcription, translation, audio analysis, and manual note-taking.

Gathering and analysing feedback

Table 9 presents indicative examples of this feedback while Table AII- 1 in Annex II illustrates the whole list of User Feedback, Source, and Actions Taken for the KSE Service in detail. The tables are organized into four columns: Issue Category, User Feedback, Source, and Action Taken. The User Feedback column includes users’ own words quoted wherever possible, providing direct insight into their experience. The table covers a wide range of topics, grouping the feedback into categories such as UI/UX, upload and format support, backend integration, keyframe extraction, face detection and enhancement, and new feature requests. For each item, the source of the feedback and the corresponding actions taken are also listed. The full description of the measures implemented to address the feedback is available in Section 4.2.4.

Table 9 Indicative examples of User Feedback, Source, and Actions Taken for the KSE Service

Issue Category	User Feedback	Source	Action Taken
UI / UX	Split opinions on timeline: some find it “good” (especially that yellow bar highlights change in shots/cuts) others find it to be “improvable”, with comments similar to “Timeline not clear”. “The timeline should indicate shot changes”.	PE, ATC	Redesigned timeline for clarity; made hovering over a keyframe highlights its position in the timeline more clear - see also Section 4.2.4.7.
Upload & Format Support	Issues with video uploads.	PE	Improved feedback and error logging during upload/caching - also see, Section 4.2.4.10.

In the UI/UX domain, users asked for clearer timeline visuals (including shot-change indicators, hover-highlights, and more informative status and error messages) alongside features like downloadable ZIP archives for keyframes, faces, and text, and improved link behavior across browsers. These requests led to a redesigned timeline, animated progress indicators, friendlier error logs, added download buttons, and timeline enhancements. Upload and format support were expanded through server-side format conversion with FFMPEG and mobile-friendly upload processes. Backend integration improvements included new session-monitoring endpoints, support for local-file uploads, and caching mechanisms to optimize repeated requests. Keyframe extraction was accelerated and quality-enhanced via threshold tuning and error handling for cases such as YouTube video retrieval failures and missing frames. Face detection was improved through more robust models, clearer feature explanations to prevent misuse, and options to disable face detection if desired. Finally, several exploratory feature requests, such as manual keyframe selection, OCR integration, sound event analysis, and anomaly detection, were either addressed via partner integrations or considered outside the core scope of the KSE service.

4.2 Methodology

This section presents recent advancements in sound event detection, acoustic geo-tagging, and audio-visual scene classification, as well as the integration of user feedback into the KSE service.

4.2.1 Sound Event Detection & Siren Sound Classification

As a first component for understanding the acoustic scene in the background of a given audio recording, an efficient CNN model for sound event detection (SED) is designed to detect the most prominent sound events within an analysis window of five seconds. The “VGG-like” SED model with 222k parameters, which has been shown to be effective in ([Fonseca et al., 2020](#)), was trained using the USM V2 dataset ([Abeßer, 2022](#)). This dataset features synthetic five-second soundscapes with focus on 26 sound classes from urban environments, including airplane, alarm, birds, bus, car, and more, encompassing human-made, vehicle, construction, and security-related sounds. The dataset features different sound polyphony levels between

2 to 6 sounds at random levels and stereo placements. The audio clips, derived from the FSD50k dataset (Fonseca et al., 2020), comply with commercial use licenses.

Among the array of potential "sound indicators" useful for geo-tagging, the sirens of emergency vehicles stand out due to their distinctively loud characteristics, ensuring their audibility in even the most noise-filled urban settings. The divergence in siren sounds from country to country stems from historical biases and different national standards, arguably making them quite unique across geographic origins.

Initial research results on the automatic classification of the country of origin of emergency vehicle siren sounds were already reported in D3.1. The experiments were based on the Regional Siren Classification (RSC) dataset⁴ with 270 audio recordings of sirens from nine countries (Canada, China, France, Germany, India, Italy, Japan, Spain, and the USA) and three siren types (police, ambulance, and firefighters) (Abeßer et al., 2024).

Figure 7 exemplifies the considerable diversity in pitch contours from this dataset: Through an acoustic examination of the siren sounds, distinctive frequency patterns or pitch contours are evident, varying from periodic modulation, exemplified by American firefighters, to gradual sweeps seen in Canadian firefighters, and stable pitch sequences as demonstrated by the German police.

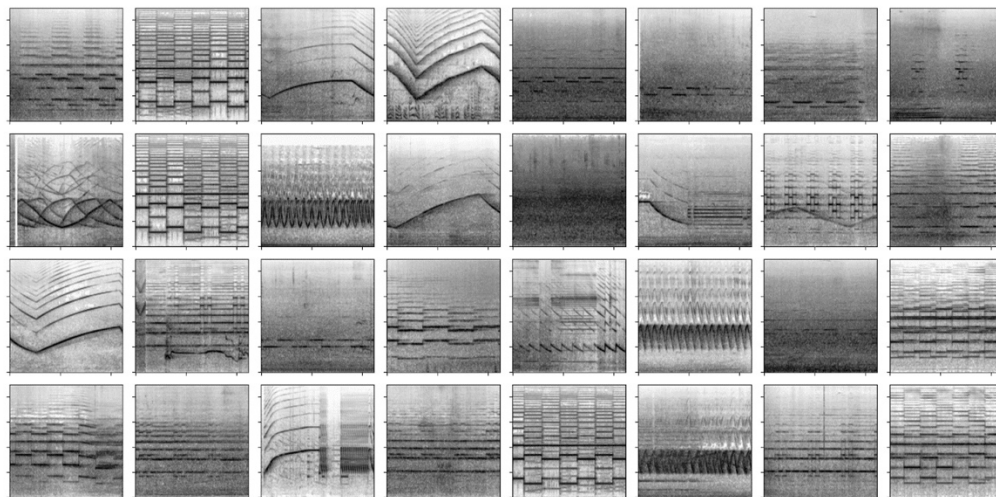


Figure 7 Random selection of siren recordings taken from the RSC database. Mel spectrogram illustrations highlight characteristic stable, alternating and sweep-like pitch contours.

In a recent study by Abeßer et al. (2025), a novel vision-based approach for pitch contour classification was introduced based on the MobileNetV2 model. Traditional methods for pitch contour analysis require a dedicated pitch tracking step, which is often error-prone due to tracking errors and diverse parameter ranges for fundamental frequency and modulation frequency of pitch contours in different audio domains such as music, speech, environmental sounds, and bioacoustic.

The novel approach avoids explicit pitch tracking and processes instead audio files converted to spectrograms using a MobileNetV2 model. The model is pre-trained successively using two tasks, object detection in natural images followed by pitch contour classification in spectrograms. For the latter task,

⁴ https://github.com/jakobabesser/regional_siren_classification_dataset

the novel Synthetic Pitch Contour (SPC) dataset was generated including 3,500 one-second long audio clips covering the seven pitch contour classes stable, alternating, glissando, vibrato, bend, triangle, and sawtooth. The pre-trained model was evaluated based on eight downstream tasks from the audio domains speech, music, bioacoustics, and everyday sounds.

4.2.2 Acoustic Geo-Tagging using Location-Specific Sounds

In addition to siren sounds, other types of sound such as church bells, car engines, or bird singing can provide location-specific auditory cues for acoustic geo-tagging (AGT) to identify the geographic location of an audio recording. Soundscape researchers coined the term “soundmarks” ([Schafer, 1994](#)) to describe distinctive acoustic landmark sounds which are unique for a specific location. One famous example of such a soundmark is the chimes of Big Ben which can only be heard in proximity of the Palace of Westminster in London, UK.

While soundmarks would allow for very accurate AGT, their practical relevance is very limited due to their rare presence in user-generated recordings. As a novel concept, a recent study ([Abeßer, 2025](#)) introduces the concept of “indicator sounds”, which are sounds characteristic of a particular location but not necessarily exclusive to it. As an example, car engine noise is more commonly associated with street traffic than with natural settings, thus it acts as an indicator sound for urban outdoor areas. The premise of the concept of indicator sounds is that once being identified, they could inform successive more detailed AGT initiatives for instance to accurately identify the city where a recording was made.

The less strict definition of indicator sounds offers a significant advantage: they can be automatically retrieved from a dataset of representative audio recordings for different geoentity classes such as cities or acoustic scenes. In ([Abeßer, 2025](#)), two data-driven approaches for indicator sound retrieval were introduced. Given an audio clip, the “Pretrained Audio Neural Network for Sound Event Detection” (PANN) model ([Kong et al., 2021](#)) for sound event detection (SED) is used to extract frame-level probabilities for 527 sound classes, as predefined in the AudioSet dataset ([Gemmeke et al., 2017](#)), which are then averaged over the clip’s duration to indicate the overall presence of different types of sounds in an audio clip.

In the first approach, a larger audio dataset is subdivided into multiple 1-vs-N partitions, allowing audio recordings from one geoentity class to be compared against those from all other categories. Using the clip-level sound class probabilities as feature vectors, a Random Forest classifier is trained for each dataset partition. Based on its overall contribution to all decision trees, an “importance value” is assigned to each sound class which measures how relevant it is for the current geoentity class, especially in comparison to the other classes. Notably, the significance of distinct sound events for identifying locations varied across different acoustic scenes and cities. Each location was associated with certain events of high relevance, potentially forming the “indicator sounds” that the study suggests could be leveraged for advanced research.

In the second approach, two MobileNetV2 networks ([Sandler et al., 2018](#)) pre-trained for object recognition on natural images were adapted to the tasks of acoustic scene classification and city classification by fine-tuning them on Mel spectrograms derived from audio recordings. Following the training of both networks, the post-hoc explainability tool known as Layerwise Relevance Propagation (LRP) ([Bach et al., 2015](#)) was employed to generate relevance maps for specific input spectrograms. These

time-frequency maps were then consolidated across frequencies to produce a relevance function that highlights temporal regions in an audio clip important for the classifier’s decision. As a final step, the correlation between this temporal relevance function and the frame-level sound class probabilities obtained using the PANN model provides a relevance score indicating how characteristic each type of sound is for the geoentity class of the current audio clip.

4.2.3 Audio-Visual Scene Classification

Along with geo-tagging and country-of-origin detection, Acoustic Scene Classification (ASC) is a task with potential for contextual analysis, involving the automated identification of the environment type of an audio recording, such as a street, park, subway, or indoor area, based on its acoustic features. Over the last decade, the Detection and Classification of Acoustic Scenes and Events (DCASE)⁵ research community has thoroughly explored this task ([Abeßer, 2020](#)). However, the role of ASC in decontextualization detection remains mostly unexplored. Thus, in this section, we highlight another study ([Apostolidis et al., 2024](#)), which introduces a novel method to verify audio-visual content by performing joint audio and visual scene classification (AVSC). The underlying hybrid deep neural network model allows identifying discrepancies between the geographic scenes identified in audio and visual streams. To aid future research on this emerging topic, the study introduces a new benchmark dataset and an experimental protocol for comparability and reproducibility.

For the ASC task, three types of embeddings are obtained from the audio stream of a video using pre-trained deep neural networks. OpenL3 embeddings ([Cramer et al., 2019](#)) come from a convolutional neural network (CNN) that was trained in a self-supervised way with an audio-visual correspondence task. The above-mentioned PANN embeddings form the second embedding. The third embedding is extracted using a ResNet model inspired from ([Grollmisch & Cano, 2021](#)), pre-trained for a simplified 3-class ASC task: indoor-outdoor-vehicle classification. In addition to audio analysis, three vision-based embeddings are derived from the video stream using the vision transformer (ViT) ([Dosovitskiy et al., 2021](#)), the CLIP (Contrastive Language-Image Pretraining) model ([Radford et al., 2021](#)), and a pre-trained ResNet50 model on the Places365 dataset. The hybrid model features two input branches for extracting single-modality embeddings, multiple layers for modality fusion, and a final classification head consisting of two dense layers with dropout for overfitting prevention.

4.2.4 User-Centered Evaluation and Integration of the KSE Service

Building on the observations and feedback collected during the evaluation sessions outlined in Section 4.1.2, we undertook a series of targeted improvements to enhance the efficiency, usability, and flexibility of the KSE pipeline. These enhancements addressed specific bottlenecks, user interface shortcomings, and operational limitations identified during testing. This section presents the technical updates and optimizations made across key components of the system—including processing speed, segmentation quality, keyframe selection, detection accuracy, and backend functionality—with the goal of aligning the KSE more closely with the needs of real-world users such as journalists and verification professionals.

⁵ <http://dcase.community/>

Faster Analysis

To accelerate the KSE pipeline, we first analyzed its processing stages to identify performance bottlenecks. Logs from 25 successfully completed KSE sessions were examined, and for each stage we calculated both the average processing time and its standard deviation. Table 10 summarizes these findings, providing a clear view of how computational effort is distributed across the pipeline.

Table 10 Processing time distribution across individual stages of the KSE pipeline

Stage Name	Stage Processing Time as Percentage of Total Processing Time (%)	
	Average	Standard Deviation
Video reading for temporal segmentation	14	±1
Video temporal segmentation to shots	7	±1
Video temporal segmentation to sub-shots	6	±1
Video temporal segmentation to VSSs	3	±1
Keyframe selection	9	±1
Text detection	21	±9
Face detection	16	±8
Multi-object tracking	4	±1
Text enhancement	9	±5
Face Enhancement	8	±6
Writing keyframes	4	±2
Writing results (JSON files, logs, HTML files)	1	±1

The analysis showed that text detection accounted for the largest share of total processing time, followed by face detection. Both stages also exhibited relatively high variability across different runs, as indicated by their higher standard deviation values, suggesting sensitivity to input video content. As a result, our optimization efforts prioritized these two stages. For text detection, the original MixNet model ([Zeng et al., 2023](#)), while accurate, proved computationally demanding and was therefore replaced with the significantly faster FAST model ([Chen et al., 2021](#)), which offers comparable accuracy. The efficiency improvements resulting from this change are detailed in Section 4.3.4. Face detection optimizations, likewise aimed at improving both speed and stability, are discussed separately in the following with corresponding evaluations presented in Section 4.3.5.

Improvements on Video Temporal Segmentation and Keyframe Selection

We enhanced the segmentation module responsible for identifying keyframes by refining video temporal segmentation and keyframe selection. This module was externally benchmarked in the CBMI and VBS competitions, where performance was assessed by expert human participants. Their feedback informed

adjustments to segmentation thresholds, helping to strike a better balance between avoiding over-segmentation and accurately capturing shot transitions.

To ensure robustness across a wide range of video types, including those with minimal visual variation and those with rapid content changes, we introduced a fallback mechanism that guarantees meaningful segmentation into subshots and visually similar sequences (VSSs). This mechanism enforces minimum thresholds for the number of subshots and VSSs. If these criteria are not met, the segmentation is repeated with gradually relaxed parameters, up to a fixed number of attempts to prevent excessive retries. The final set of parameter values fine-tuned through extensive visual inspection of their impact on the dataset of CBMI and VBS competitions are summarized in Table AII- 2 (Annex II).

For keyframe selection, professional feedback from journalists involved in media verification, strongly favored generating more extensive keyframe sets. They emphasized that missing important frames poses a greater risk than dealing with some redundancy. As a result, we adjusted our strategy to prioritize broad content coverage while keeping duplication within reasonable limits. To achieve this, we implemented improved visual similarity assessments using CNN features extracted from middle-to-early layers of pretrained ImageNet models, offering finer-grained similarity detection compared to simple visual descriptors. Based on these features, we generate two keyframe sets: 1) a strict set with a high similarity threshold (for reducing redundancy) and, 2) a broad set with a lower threshold, offering a richer frame selection for detailed inspection. The frontend was updated to present these additional keyframes in a compact, non-intrusive interface, allowing users to switch between concise and exhaustive keyframe views as needed.

We also refined the keyframe extraction quality by improving the video download process. Earlier, yt-dlp occasionally selected lower-quality video streams even when higher-quality versions were available, particularly compatible streams could not be merged, when ffmpeg could not be found. Ensuring ffmpeg availability allows yt-dlp to correctly merge high-quality separate video and audio streams. Additionally, we avoided forced re-encoding during merges, which can degrade quality, by “remuxing” into MKV containers. This method does not re-encode video or audio, avoiding quality loss; it simply combines existing streams into a single container without altering their content. This ensures that extracted keyframes originate from the highest available source quality.

Improvements on Face detection

To enhance the parameter efficiency of detection models, we adopted a more elaborate strategy based on model pruning to reduce computational demands while preserving or even improving accuracy. We developed a hybrid pruning framework, which we refer to as B-FPGM (Bayesian-optimized Filter Pruning via Geometric Median) - a lightweight face detection pruning framework that combines a) Soft Filter Pruning (SFP), b) Filter Pruning via Geometric Median (FPGM), and c) Bayesian optimization to achieve efficient and adaptive model compression. It improves upon traditional methods by dividing the network into layer groups and using Bayesian optimization to assign optimal, group-specific pruning rates. This two-stage pruning approach (first soft, then hard) allows the model to adaptively recover performance during training before permanently removing redundant filters. Extensive experiments demonstrate that B-FPGM significantly reduces model size and computation while maintaining competitive detection accuracy on benchmark datasets.

This approach combines three complementary techniques: Filter Pruning via Geometric Median (FPGM), Soft Filter Pruning (SFP), and Bayesian optimization for fine-grained pruning control. The pruning process begins by selecting a compact face detection network, such as EResFD, as the initial backbone model. This network is divided into multiple layer groups, allowing for differentiated pruning strategies across the network's architecture. Bayesian optimization is then used to automatically determine the optimal pruning rate for each layer group, enabling a highly customized reduction in model complexity. This avoids the pitfalls of uniform or manually tuned pruning, which often result in suboptimal trade-offs between size and accuracy. This is an iterative two-stage pruning approach (first iterative soft-pruning, then hard-pruning). The complete pruning methodology consists of 4 stages. Initially the face detection network is pre-trained normally. Bayesian optimization is then employed to identify the optimal pruning rate for each layer group. Subsequently, the network is iteratively soft-pruned, using the optimal pruning rates, and re-trained for a few epochs. Finally, the network is hard-pruned and fine-tuned. A complete evaluation of the performance of the pruned model is presented in Section 4.3.5.

Dealing with Face Hallucinations

To address the problem of face hallucination artifacts (i.e., exaggerated or artificial details introduced by aggressive enhancement models) we explored three strategies: first, the integration of enhancement tasks such as super-resolution and deblurring into a unified module; second, a rigorous re-evaluation of face super-resolution models focusing on face hallucination artifacts; and third, a reconsideration of the image enlargement scale to reduce hallucination risks.

The first strategy involved adaptive enhancement, where context-aware rules decide whether to apply super-resolution or deblurring based on the visual content of each region. Specifically, we used the variance of the Laplacian metric, originally developed for temporal segmentation into visually similar sequences (see Section 3.2 of D3.1), to measure the sharpness of detected face regions. If the sharpness falls between two predefined thresholds, deblurring is applied before super-resolution. For deblurring specifically, we incorporated Google's MAXIM model ([Tu et al., 2022](#)), a state-of-the-art architecture that combines local and global processing using a multi-axis MLP within a UNet-like structure. MAXIM delivers excellent deblurring accuracy, achieving around 32 dB PSNR on the GoPro benchmark. However, its significant computational cost (22 million parameters and roughly 339 billion FLOPs for a simple 256x256 image) makes it likely impractical for use in the KSE.

The second strategy involved a re-evaluation of the face hallucination models initially examined in D3.1. This time, we applied a more stringent evaluation protocol, testing models on facial images completely outside their original training data and thoroughly visually inspecting the results for visual artifacts. Based on these tests, we identified the GCFSR model as the most robust, consistently producing natural-looking results with minimal hallucinations. A complete description of the evaluation process and findings is provided in Section 4.3.6.

Third, we reduced maximum upscaling from 4x to 2x. While 4x can recover finer details, it significantly increases hallucination risks in low-quality inputs by requiring more speculative high-frequency generation. The lower 2x scaling provides safer enhancement while accelerating processing - a practical compromise for diverse real-world content.

Sound Event Detection

A sound event detection module was seamlessly integrated into the KSE pipeline via a remote backend service provided by IDMT. The detected sound events include: 1) airplane, 2) alarm, 3) birds, 4) bus, 5) electric car, 6) regular car, 7) church bell, 8) crowd, 9) dogs, 10) drilling, 11) glass breaking, 12) gunshot, 13) hammer, 14) helicopter, 15) jackhammer, 16) lawn mower, 17) motorcycle, 18) sawing, 19) scream, 20) siren, 21) speech, 22) thunderstorm, 23) tram/train, and 24) truck. The technical details of the sound event detection method are provided in Section 4.2.1. By feeding audio-derived event timestamps into the existing keyframe selection logic, this addition broadens the pool of candidate frames beyond purely visual salience, ensuring that moments coinciding with significant sound events are also considered. In practice, the module flags time intervals for detected sound events, and the KSE algorithm then selects corresponding video frames, thereby enriching the final set of keyframes with audio-driven highlights.

The process begins by extracting the audio track from the video using ffmpeg, which converts it into an AAC file compatible with the backend service. This audio file is then uploaded to the backend server via an HTTP POST request. After a successful upload, the backend responds with a unique file identifier (file ID), which is used in a second POST request to initiate the audio analysis. The system then enters a polling phase, repeatedly checking the backend for the processing status. Once the analysis is complete, a result identifier (result ID) is returned, signaling that the output is ready. A final GET request retrieves the analysis results in JSON format. These results are saved locally and integrated into the session's directory structure within the KSE pipeline, making them easily accessible to downstream modules and user interfaces.

Process Level Profiles

The KSE service processes videos through a multi-stage pipeline: splitting content into shots/subshots, extracting keyframes, detecting faces/text, grouping identical elements, and enhancing selected items. While effective, user feedback revealed two limitations: 1) inflexibility for time-sensitive tasks where full processing is unnecessary, and 2) a likely user behavior where when given the option to select among different enhancement quality levels, users tend to run the full enhancement pipeline multiple times in order to compare outputs across these levels.

To address this, we implemented configurable processing profiles that skip entire pipeline stages. Unlike parameter tuning—which yielded marginal speed gains—this approach lets users bypass functional blocks entirely. Available profiles include: a) segmentation-only (shot segmentation and keyframes), b) segmentation with detection (adds face/text detection without enhancement), and c) full enhancement (complete pipeline). This balances flexibility with straightforward operation, allowing users to match processing depth to their specific needs while eliminating redundant computations.

UI updates

To improve overall user experience, accommodate the optional sound event detection and the processing profiles, the KSE interface underwent a series of updates. Specifically, the interface was restructured to clearly reflect the processing stages by displaying more user-friendly log messages. We also provide better transparency into ongoing operations by using more detailed step messages.

New download options were introduced, allowing users to retrieve ZIP archives of detected text regions, faces, and selected keyframes directly from the results sections (see Figure 8). The timeline visualization

was widened for improved indication of the current keyframe (see Figure 9), and previously reported viewer link issues were resolved (see Figure 10).

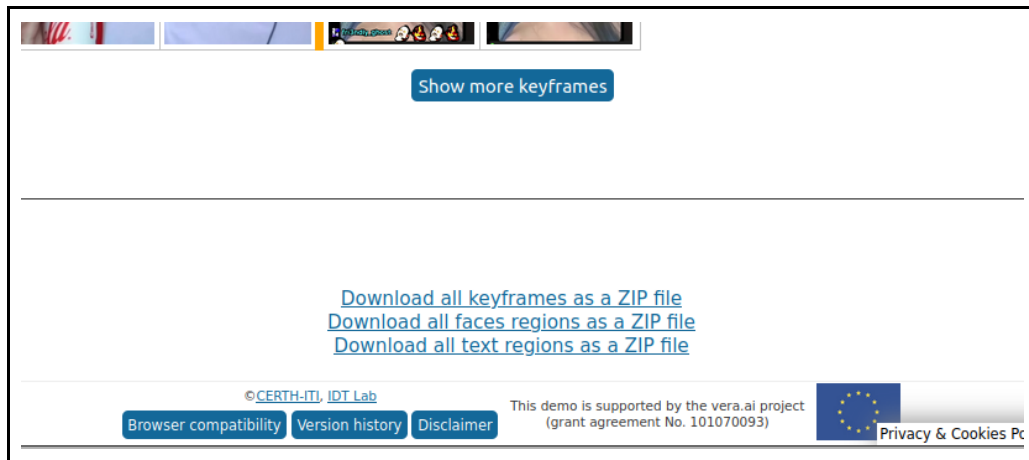


Figure 8 The position of the ZIPs download links under the "Fragmentation & Keyframes" panel

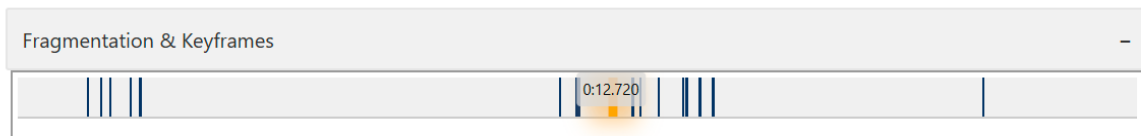


Figure 9 When hovering over a keyframe, its position is highlighted in the timeline using an orange glow

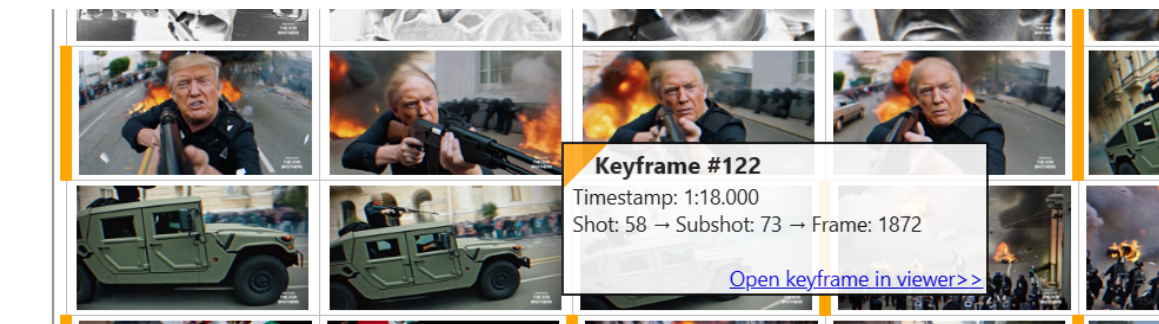


Figure 10 The function "Open keyframe in viewer" link, shown in the popup dialog when hovering a keyframe

The processing profile selection is conveniently located under a collapsible "advanced options" section, ensuring that the main interface remains uncluttered for casual users while still offering full control for advanced users. Similarly, the newly integrated sound event detection feature is also accessible via the "advanced options" panel (see Figure 11). Sound events are displayed in a dedicated table, with keyframes linked to corresponding detected events. Hovering over table entries reveals precise timestamps and shot information, and in an alternative interface version, short video segments can be played directly via HTML5 video elements to further enrich the user experience.

The figure consists of two panels. The top panel shows a form with a file upload button labeled "...OR upload a file (mp4, webm, avi, mov, wmv, ogv, mpg, flv)", an optional email field, a "Usage Instructions" link, a "Submit" button, and a "Hide Advanced Options »" link. The bottom panel shows the expanded view with an "Insert a video URL (supported sources: see Usage Instructions)..." field, the same file upload button, optional email field, "Usage Instructions" link, "Submit" button, and a "Show Advanced Options «" link. The expanded view also includes a "Process Level:" section with "Segment ✓", "Detect ✓", and "Enhance" buttons, and a "Process Audio:" toggle switch.

Figure 11 Position of the “Show Advanced Options” label (top panel) and the expanded view it reveals when clicked (bottom panel)

A significant update involved the asynchronous provision of results. To maximize responsiveness, results are now displayed interactively as they become available. Keyframes and shot segmentation results are presented immediately within collapsible panels, while face and text detections are processed in the background and updated progressively, accompanied by progress bars that inform users of the current processing state (see Figure 12). This approach enables continuous user interaction with partial results, minimizing unnecessary waiting times.

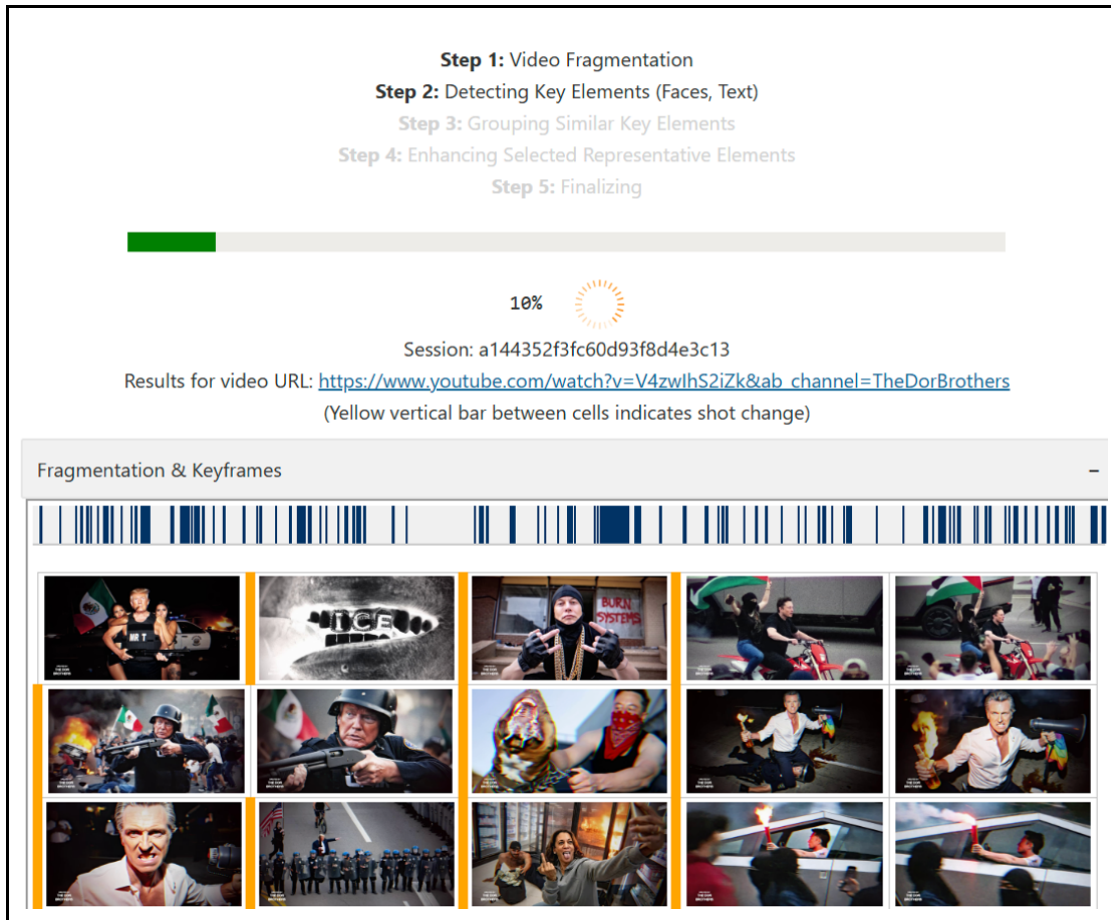


Figure 12 Asynchronous provision of keyframes - The “Fragmentation & Keyframes” panel is shown with the results of the video fragmentation while the analysis continues to the subsequent stages

Additionally, session management was streamlined by transforming session identifiers into clickable URLs, allowing users to directly access and share their analysis results (see Figure 13). Collectively, these UI improvements were designed to enhance usability, support varying user requirements, and fully leverage the flexibility introduced by the process-level profiles.

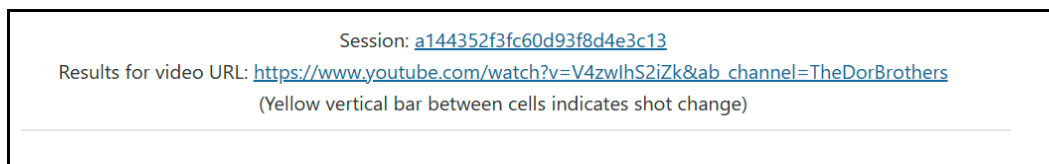


Figure 13 Clickable URLs for the session identifier and the submitted video source

Local File Uploads via Backend Calls

An important limitation in the original KSE API was that it previously only supported video start calls to backend processing via URLs, preventing integrators from submitting video files stored locally. To address this, we enhanced the `/kse/start` endpoint to support both URL-based and file-based submissions in a unified way. The adjusted endpoint automatically detects the type of each incoming request. For standard API clients using `application/json`, it processes the request as before. For clients submitting files via `multipart/form-data`, it extracts metadata from a JSON field and handles file uploads separately. When a

file is uploaded, the system creates a unique session ID, securely stores the file in a dedicated session directory, and performs validation checks, including file size limits and basic video integrity checks, before starting processing. To help partners integrate with the updated API, we provided clear documentation and examples showing how to construct multipart submissions combining JSON metadata with file uploads.

URL Blacklist

In response to repeated spam attempts involving inappropriate content, such as submission of pornographic videos, a URL blacklisting mechanism was developed and integrated into the KSE. These unwanted requests consumed system resources and provided no benefit to the intended users - journalists and fact-checkers. The blacklist mechanism filters out undesirable domains using carefully curated lists that are regularly updated to address new misuse sources. The filtering includes efficient lookup mechanisms to quickly assess whether a submitted URL is blocked.

Troubleshooting Google Bot Flagging

A significant operational issue occurred when Google servers flagged KSE's downloader as automated bot activity, blocking access to many online videos needed for processing. To resolve this, we developed a dedicated media downloader running on a separate server. The downloader offers HTTP API endpoints for video, image, and file retrieval, organizes downloads into session-specific folders, and handles requests asynchronously to ensure high availability. Rate-limiting controls both concurrent and daily downloads per domain, reducing system load.

4.3 Evaluation

This section presents a comprehensive evaluation of multiple system components. It assesses audio processing capabilities, including siren classification by country and acoustic geo-tagging using location-specific sounds. The evaluation further covers multimodal performance through audio-visual scene classification and discrepancy detection. Computer vision components are rigorously tested, including text detection, face detection (with pruning analysis), and enhancement techniques. Finally, the Keyframe Extraction and Enhancement (KSE) service undergoes both performance benchmarking and accuracy assessment. Results demonstrate the effectiveness of the proposed methods through quantitative metrics, qualitative analysis, and expert feedback, validating improvements across detection, classification, and processing efficiency tasks.

4.3.1 Siren Sound Classification

In the domain of everyday sounds, the task of classifying the country of origin of siren recordings was investigated using the RSC dataset. The proposed vision-based approach combined with the two-stage pre-training approach clearly outperformed baseline CNN method to learn feature representations based on pitch contours extracted using the state-of-the-art SWIPE pitch tracking algorithm ([Camacho & Harris, 2008](#)).

Based on the nine countries included in the RSC dataset, the model achieved an F-score of 0.72. The confusion matrix obtained with the best model is illustrated in Figure 14. As a general observation, siren

sounds from France, Canada, Germany, Japan, and the USA possess the most distinct pitch contours. Although there exists some degree of misclassification (e.g., China - Canada, Italy - Germany, France - Spain, Japan - Spain), the results confirm that pitch contour classification can be used for recognizing and classifying siren sounds as acoustic landmarks.

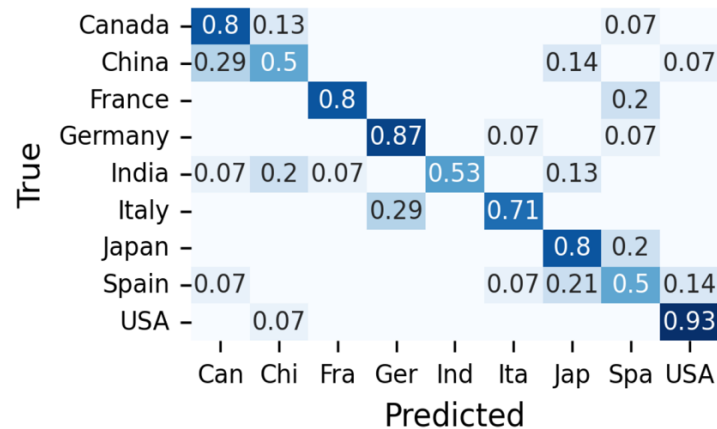


Figure 14 Confusion matrix obtained for regional siren classification using a vision-based classification approach.

4.3.2 Acoustic Geo-Tagging using Location-Specific Sounds

Both methods for indicator sound retrieval were evaluated using the TAU 2020 Mobile Urban Acoustic Scenes dataset ([Heittola, 2020](#)), which includes 64 hours of 10-second audio recordings captured in 10 different European capitals across 10 different acoustic scene classes.

Figure 15 illustrates a wordcloud that highlights the most relevant indicator sound classes retrieved using the first approach (left) and the second approach (right). In this word cloud, the relevance score obtained for each sound class is reflected in the font size. Both methods effectively identify pertinent sound classes necessary to differentiate between indoor, outdoor, and vehicle acoustic scenes. Inevitably, the inclusion of numerous sound categories results in some redundancy due to the presence of similar classes. The first method employs a 1-vs-all strategy, concentrating on detecting sounds particular for certain geoentities compared to others. Conversely, the second method offers a more neutral perspective, highlighting sound classes with the greatest general distinguishability across various similar geoentities. In regard to city classification, the recognized sound classes represent rather specific recording spots than the entire city. For instance, Vienna is recognized for its sirens, Helsinki for the sounds of water, and Barcelona for its street musicians.

These qualitative results highlight the usefulness of "indicator sounds" as significant signals for acoustic geo-tagging, particularly in situations where distinct soundmarks are infrequent or nonexistent. These findings pave the way for audio-based contextual analysis, such as identifying discrepancies between these indicator sounds and the asserted or anticipated location of an audio file being examined.

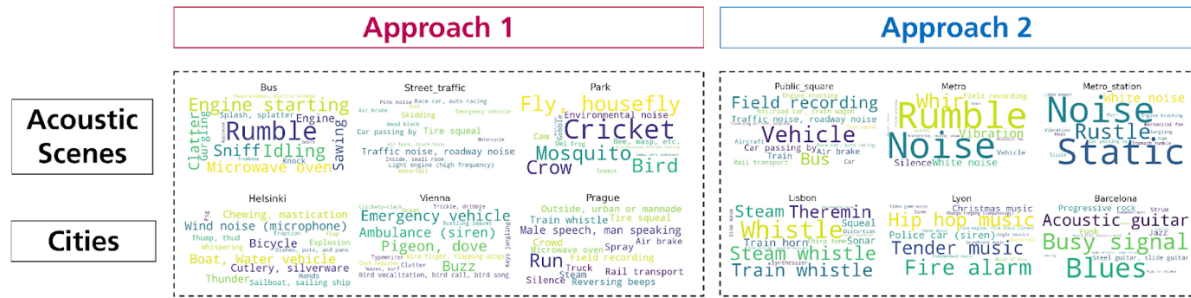


Figure 15 Selection of indicator sounds for three acoustic scene and city classes obtained with both proposed retrieval approaches. The font size correlates with indicator sound relevance

4.3.3 Audio-Visual Scene Classification

This evaluation assesses various modality fusion strategies, utilizing both concatenation and self-attention layers to better evaluate the significance of different elements within audio-visual embedding representations. The TAU Urban Audio-Visual Scenes 2021 dataset ([Wang et al., 2021](#)) comprises videos labeled within ten urban acoustic settings (e.g., shopping mall, park, airport). It first assesses scene classification performance using the individual input modalities and then via the proposed fusion approach that integrates both modalities. Alongside this ten-class taxonomy, a simplified three-class taxonomy is also examined, categorizing videos as indoor, outdoor, or captured in moving vehicles. While the classification accuracy for the ten-class taxonomy is at 97.2 % for using only the video stream and 78.8.% using only the audio stream, the fusion model that combines both modalities achieves a near perfect result of over 99 % accuracy. As expected, confusion occurs between acoustically and visually similar settings—such as a tram versus a bus, or an airport versus a shopping mall. In detecting visual-audio inconsistencies, the best system hits 97.2% accuracy, emphasizing the usefulness of scene classification in spotting decontextualization in multimedia. The research highlights the necessity for alternative feature fusion methods and contrastive learning techniques to enhance performance in complex, real-world cases.

Based on the audio-visual scene classification dataset, a novel dataset for the visual-audio discrepancy detection (VADD) task is proposed, where audio streams are intentionally interchanged between scenes to create decontextualized content. Using a three-class taxonomy, a discrepancy is identified, such as when the audio from an outdoor video is substituted with an indoor recording. An ablation study highlighted the significance of the self-attention layer placement within the network architecture. The top model features a self-attention layer at the conclusion of both the audio and vision input branches after domain-specific embedding vectors have been concatenated. These results show that decontextualized audio-visual content can be identified using deep learning methods for scene classification. Identifying cross-modal inconsistencies is an important tool for automated content verification, deepfake detection, and media forensics.

4.3.4 Text Detection

Table 11 compares state-of-the-art text detection methods on the Total-Text dataset, emphasizing their accuracy and inference speed to guide the selection of the most suitable approach for real-time

applications. We opted for a faster text detection method. In this table, a comprehensive overview of text detection results is provided. The FAST method stands out as the most efficient option, achieving an F-Measure of 81.6% with the shortest inference time at 173 ms. Although MixNet delivers a higher accuracy of 90.5%, its significantly slower execution time (326 ms) may not be ideal for real-time applications. Thus, when balancing accuracy and speed, FAST emerges as the preferable choice for scenarios requiring real-time text detection.

Table 11 Comparison of text detection methods on the Total-Text dataset, highlighting their F1 scores and average inference times

Method	F1 (%) on Total-Text	Average inference time (ms)
MixNet: Toward Accurate Detection of Challenging Scene Text in the Wild (Zeng et al., 2023)	90.5%	326
FAST: Faster Arbitrarily-Shaped Text Detector with Minimalist Kernel Representation (Chen et al., 2021)	81.6%	173

4.3.5 Face Detection

The effectiveness of the pruning framework was validated through extensive experiments on the WIDER FACE dataset, covering the Easy, Medium, and Hard subsets. Looking at Table 12 we observe that our B-FPGM approach offers a clear advantage over uniform FPGM pruning, achieving a more favorable balance between model size and accuracy. Even at higher pruning rates, B-FPGM maintained high mean average precision (mAP), while substantially reducing the number of parameters - reaching up to 60% parameter reduction with notably better performance than uniform pruning at equivalent sparsity levels. Notably, the actual sparsity achieved by B-FPGM varied across layer groups, concentrating more aggressive pruning on certain groups (typically Groups 3 to 6) while preserving critical layers, allowing the model to retain high accuracy despite significant compression. The use of Bayesian optimization further contributed to stable pruning configurations across runs, yielding consistently reliable models.

However, as discussed in earlier sections (see Table AII- 1 Annex II), user feedback revealed suboptimal performance of the EResFD face detection component (regardless of whether it is pruned using B-FPGM or not). Specifically, participants reported the presence of false positives as well as failures in detecting faces viewed in profile. These shortcomings prompted us to re-evaluate our emphasis on parameter efficiency. Instead of retaining a lightweight but less reliable solution, we prioritized detection quality and reverted to the YOLOv7 face detector ([Bousmaha et al., 2023](#)), known for its higher accuracy and robustness.

Table 12 Comparative results (mAP) on the WIDER FACE using the EResFD model between uniform FPGM [12] and the proposed B-FPGM. T is the target pruning rate

Method	T	Easy	Medium	Hard	Actual Sparsity	# of Params
EResFD (orig.) (Jeong et al., 2024)	0%	0.8902	0.8796	0.8041	0%	92,208
EResFD (rep.)	0%	0.8660	0.8555	0.7731	0%	92,208
uniform FPGM (Gkrispanis et al., 2024)	10%	0.8728	0.8582	0.7757	5.25%	87,368
B-FPGM	10%	0.8622	0.8506	0.7636	10.24%	82,765
uniform FPGM	20%	0.8369	0.8201	0.7230	16.84%	76,677
B-FPGM	20%	0.8601	0.8477	0.7475	22.27%	71,673
uniform FPGM	30%	0.8311	0.8160	0.7175	24.36%	69,746
B-FPGM	30%	0.8348	0.8266	0.7231	31.59%	63,079
uniform FPGM	40%	0.8124	0.7952	0.6807	35.95%	57,055
B-FPGM	40%	0.8227	0.8192	0.7205	40.02%	55,306
uniform FPGM	50%	0.7103	0.6830	0.5254	48.72%	47,284
B-FPGM	50%	0.7956	0.7955	0.6993	50.37%	45,578
uniform FPGM	60%	0.5209	0.4566	0.2936	54.05%	42,369
B-FPGM	60%	0.6974	0.6937	0.6051	59.87%	37,030

Rather than simply returning to the original YOLOv7s model used in D3.1, we took the opportunity to upgrade to a more advanced configuration. Specifically, we adopted the yolov7-w6 model with test-time augmentation (TTA), which offers significantly improved detection performance with only a small increase in inference time, i.e., from 18 ms to 21 ms per frame. The YOLOv7 GitHub repository provides a breakdown of pretrained model variants across three complexity categories: a) Lite models (e.g., yolov7-lite-t/s) range from 0.8 to 3.0 billion FLOPs and offer the lowest accuracy (71.5–78.5% on the hard subset); b) Medium models (e.g., yolov7-tiny/s) span 13.2 to 16.8 billion FLOPs and show improved performance (82.1–85.2%); and c) Heavy models (e.g., yolov7-w6 variants) peak at 88.3–90.5% accuracy with 89.0 billion FLOPs. By choosing the heavier yolov7-w6+TTA model, we accepted a slight increase in computational cost in exchange for significantly improved reliability, an essential trade-off for journalistic and forensic use cases where accuracy must not be compromised. Table 13 presents a comparison of face detection methods on the WIDER FACE dataset (hard subset), showing that the B-FPGM (T=10%) model achieved a mean average precision (mAP) of 76.4% at 9 ms per frame, while YOLOv7s reached 85.2% at 18 ms. The upgraded yolov7-w6+TTA model achieved the highest score at 90.5%, demonstrating its superiority in handling difficult detection cases.

Table 13 Comparison of face detection methods on the WIDER FACE dataset (hard subset), highlighting their F1 scores and average inference times

Method	mAP (%) on WIDER FACE (hard subset)	Average inference time (ms)
yolov7s model of YOLOv7 face detector	81.6	18
B-FPGM (T=10%)	76.4	9
yolov7-w6+TTA model of YOLOv7 face detector	90.5	21

4.3.6 Face Enhancement

As reported in Section 4.2.4 ‘Dealing with Face Hallucinations’ we explored three strategies: first, the integration of enhancement tasks such as super-resolution and deblurring into a unified module; second, a rigorous re-evaluation of face super-resolution models focusing on face hallucination artifacts; and third, a reconsideration of the image enlargement scale to reduce hallucination risks.

Regarding employing deblurring in coordination with super-resolution methods for face enhancement, Table 14 illustrates the relative processing time overhead introduced by the KSE pipeline with and without deblurring across a series of test videos. It is evident that enabling deblurring substantially increases computational costs, often by several multiples compared to the baseline pipeline without deblurring. This significant processing burden stems from the inherent complexity of deblurring algorithms, as also noted in prior studies ([Nah et al., 2017](#); [Kupyn et al., 2018](#)), where it is cited that even deblurring deep learning models face challenges in balancing processing speed and restoration accuracy, particularly when handling high-resolution video content.

Table 14 Relative processing time overhead (%) of the KSE pipeline with and without deblurring across multiple test videos

Test video	Test video #1	Test video #2	Test video #3	Test video #4	Test video #5
KSE w/o deblurring (%)	14.9%	45.9%	75.2%	81.3%	98.4%
KSE + deblurring (%)	65.1%	247.4%	374.1%	362.6%	640.2%

Moreover, the perceptual improvement from deblurring in our experiments did not consistently deliver the level of quality enhancement that would justify its high computational cost. Consequently, instead of employing computationally intensive deblurring in the KSE’s pipeline, we opted to focus on selecting a more appropriate enhancement model from the candidate methods reviewed in D3.1. These models were rigorously tested on diverse datasets, including previously unseen inputs, to ensure robust performance.

Table 15 presents a comparative overview of several face super-resolution methods evaluated in our study. The table summarizes the evaluation results of the tested methods across multiple quantitative and qualitative metrics. Alongside the number of parameters and inference time, we report standard image quality metrics such as LPIPS (lower is better), SSIM and PSNR (higher is better), as well as FID (lower is better), which reflects distribution similarity to real images. In addition, a visual inspection score was

assigned. The visual inspection score reflects the perceived quality of generated faces on a continuous scale from 0.0 to 1.0. Scores below 0.1 indicate that no recognizable face structure is present. Values between 0.1 and 0.2 correspond to images where a face is distinguishable, but lacking in clear traits. Scores from 0.2 to 0.4 reflect faces that exhibit basic, recognizable features, though fine details may still be missing or distorted. As the score approaches 0.6, faces become sharper and display more accurate anatomical traits. Scores between 0.6 and 0.8 represent faces that are generally accurate in terms of identity and structure. Finally, values close to 1.0 denote high-fidelity faces with well-preserved details, sharp contours, and realistic textures that closely resemble natural facial appearances.

Table 15 Comparison of selected face super-resolution methods: model size, inference time, quantitative metrics, and visual inspection scores

Method	Preview	No of parameters	Inference time (seconds)	LPIPS score ↓	SSIM score ↑	PSNR score ↑	FID score ↓	Visual inspection score ↑
RestoreFormer+ (Wang et al., 2023)		79,308,423	0.420	0.29	0.74	25.70	347.20	0.80
CodeFormer (Zhou et al., 2022)		59,788,608	0.540	0.31	0.73	25.61	342.99	0.85
VQFR (Gu et al., 2022)		83,486,539	0.400	0.30	0.69	23.31	317.04	0.65
BFRffusion (Chen et al., 2024)		1,284,246,671	18.470	0.25	0.77	27.32	258.06	0.95
GCFSR (He et al., 2022)		88,742,317	0.110	0.27	0.79	27.12	366.51	0.85

This subjective score provides an intuitive measure of how well each method reconstructs visually meaningful and recognizable faces, complementing the objective metrics. The inclusion of the Visual Inspection Score allows for a more holistic assessment, especially in challenging restoration scenarios where numerical scores alone may not fully capture perceived quality. We used image pairs like the one in Figure 16 to render the enhanced face images and test the methods. Although over 30 models were deployed, tested, and visually inspected during the evaluation phase, only a subset of the most promising methods based on visual quality are included here for conciseness. Notably, some diffusion-based approaches, such as BFRdiffusion, deliver highly realistic and visually impressive results; however, their inference times, measured in the order of several seconds, are prohibitively long for practical applications. Consequently, diffusion methods were not considered further in our final selection, though BFRdiffusion is included in the table for completeness.



Figure 16 A sample image used as input (left) and the corresponding ground-truth (right) for evaluating the face enhancement methods.

Among the remaining contenders, RestoreFormer achieves the highest visual quality, demonstrated by a SSIM score of 85.4% on the CelebA dataset, albeit with an inference time of 420 ms. VQFR offers a comparable SSIM of 84.7% but requires slightly more processing time at 540 ms. CodeFormer, while delivering competitive performance with a SSIM of 83.3%, operates faster with an inference time of 260 ms, making it suitable for speed-sensitive scenarios. Of particular interest is GCFSR, which strikes an effective balance between speed and quality, achieving a SSIM of 84.3% while boasting the fastest average inference time of 110 ms.

Given our application context where face regions are only modestly upscaled to minimize hallucination artifacts, we ultimately selected GCFSR due to its fast execution and consistently clear, faithful visual results confirmed through extensive inspection on real-world data. Several other models, despite faithful implementation of official instructions and error-free inference, were excluded due to their tendency to produce distorted or unnatural facial reconstructions.

4.3.7 KSE Performance Evaluation

Recognizing that swift performance is essential for the success of the KSE service, we evaluated the baseline and optimized pipelines on five videos of varying lengths (from 83s to 1193s) and motion complexity. Our key metric is the ratio of processing time to video duration. Video complexity, influenced by factors such as motion and content dynamics, plays a crucial role in the processing load; longer videos

with more complex motion typically require more computational effort. Percentages below 100% indicate faster-than-real-time performance.

Table 16 provides a detailed comparison of processing times across various versions of the KSE service, with each score representing the processing time as a percentage of the video duration. It is shown that the baseline pipeline runs at 500–1600% of real time which makes it impractical for real-time applications. Adding VSS segmentation cuts this to 16–220%. Introducing MOT grouping further halves those numbers, and software-level optimizations bring even the hardest case (Test video #5) to just 98% of real time. This progressive reduction from 1623% down to 98% demonstrates the effectiveness of our optimization actions, from calculating VSS to employing MOT techniques and implementing further software Enhancements.

Table 16 Processing time as a percentage of video duration for successive KSE optimizations

Test video	Test video #1	Test video #2	Test video #3	Test video #4	Test video #5
Video duration (seconds)	991	578	238	1193	83
KSE baseline (%) (as reported in D3.1)	523.6%	612.8%	1023.2%	1400.0%	1622.7%
KSE (VSS) (%)	15.7%	92.0%	156.8%	193.3%	219.3%
KSE (VSS, MOT) (%)	15.8%	51.4%	85.3%	96.7%	106.9%
KSE (VSS, MOT, S/W opt.) (%)	14.9%	45.9%	75.2%	81.3%	98.4%

4.3.8 KSE Accuracy Evaluation

In addition to evaluating the performance of the Keyframe Extraction and Enhancement (KSE) service, it is crucial to rigorously assess its accuracy. This evaluation followed two complementary strategies: one based on expert feedback and another on quantitative analysis mirroring the approach taken for the performance KPI.

To capture the practical utility and real-world effectiveness of KSE, a series of expert sessions were conducted. Over a two-month period, 21 professionals participated in guided evaluations, providing extensive qualitative feedback that was later refined through follow-up remote workshops. The consensus emerging from these sessions was overwhelmingly positive. Participants praised the tool’s capabilities in video verification, text extraction, and its intuitive visualization of shot changes. While constructive suggestions were made, particularly regarding improvements in face detection and timeline representation, all participants confirmed the system’s value and expressed willingness to integrate KSE into their workflows, with indicative comments like “*most important service for video verification*” and “*easy to use as it resembles functionalities from other comparable tools*”.

Alongside the qualitative assessment, a structured quantitative evaluation was carried out to measure how the optimizations introduced to KSE impacted enhancement quality. The goal was to determine whether the refined pipeline not only speeds up processing but also produces clearer, more accurate outputs. The unoptimized KSE version from the project’s first year served as the baseline. The working hypothesis, based on the optimizations outlined in D3.1 and previous sections of this deliverable, was that selectively enhancing only the best-quality detections would lead to better results.

To test this, a controlled experiment was conducted using high-resolution (4K) source videos as ground truth. These videos were first downsampled to simulate real-world low-quality input. We applied a) the unoptimized pipeline applied to downsampled input, and b) the optimized pipeline applied to downsampled input. In each case, the enhanced key elements were compared to the same regions in the original 4K reference. As shown in Table 17, the optimized KSE pipeline achieved a PSNR of 25.128 dB, a notable improvement over the baseline’s 18.215 dB. While this 38% increase might seem modest at first glance, it reflects a much larger reduction in reconstruction error due to the logarithmic nature of the PSNR scale. SSIM also improved significantly, rising from 0.399 to 0.691, i.e., a 73% relative gain in perceived visual similarity.

Table 17 Quantitative evaluation of enhancement accuracy using PSNR and SSIM for optimized vs. unoptimized KSE pipelines under various resolution and processing configurations. Higher scores reflect closer alignment with the original 4K reference, demonstrating

KSE Version	PSNR score (dB) ↑	SSIM score ↑
KSE baseline (as reported in D3.1)	18.215	0.399
KSE (all optimizations reported in D3.1 and D3.3)	25.128	0.691

4.4 Implementation and Integration

All the technologies referenced in Section 4 of this deliverable are fully encapsulated within the KSE server. This encapsulation has been maintained since Deliverable 3.1, where the technologies discussed in Section 3 were initially integrated into the KSE service. Subsequent adjustments, refinements, and improvements derived from user feedback and evaluation processes have been applied consistently to this same core service.

The KSE service is implemented as a Python 3.9 software component using the Flask web framework. The deep learning components rely on both PyTorch and TensorFlow frameworks for model inference. The API is exposed as a RESTful service supporting asynchronous submission, status polling and result retrieval. The analysis requests are validated through a user authentication process and processed on a one-by-one basis based on a queueing mechanism. The service runs on an NVIDIA RTX4080Super GPU and supports faster than real-time analysis.

4.5 Concluding Remarks

In our evaluation of audiovisual content analysis methods, we identified key trade-offs between classification performance, model robustness, and practical deployment in fact-checking workflows. For sound event and siren classification, our fine-tuned models achieved strong performance in identifying region-specific audio cues, offering valuable support for contextual verification and geolocation. However, model accuracy varied across environments, emphasizing the need for domain-specific adaptation and expanded training data.

Our acoustic geo-tagging approach demonstrated effectiveness by leveraging indicator sounds and contrastive learning. In audio-visual scene classification, combining both modalities led to improved classification accuracy, confirming that multimodal fusion offers tangible benefits for media verification tasks. Visual analysis tasks, including face and text detection, benefited significantly from our image enhancement pipeline, with improved clarity enabling better downstream detection results. The KSE tool showed notable improvements in both speed and usability, aligning closely with fact-checker needs for extracting and navigating visual evidence.

Overall, our results emphasize the effectiveness of integrating user feedback into system design, as well as the importance of balancing accuracy, scalability, and explainability in real-world audiovisual verification scenarios.

5 Extraction of Text and Geolocation from Images

For journalists and fact-checkers it is important to retrieve information from images and videos in order to better understand the context of the media and support verification efforts in the fight against disinformation. In T3.4, we focus on extracting visual clues from images to assist in identifying any embedded text and estimating the location depicted in the image that could provide additional verification signals. To this end, we address two challenging problems: i) **text extraction** and **language identification** using an **Optical Character Recognition (OCR) tool**, and ii) **geolocation** based solely on visual features to infer the likely geographic origin of the image.

5.1 Background

Here we outline a brief literature review for the tasks of OCR and the geolocation estimation.

5.1.1 Optical Character Recognition (OCR)

Much of the textual content found in our surroundings contains valuable semantic information, which is essential for various real-world applications. OCR plays a pivotal role in numerous applications such as scene-text reading ([Wang et al., 2011b](#)), document text extraction and visual data analysis ([Mishra et al., 2019](#)). Notable advancements have been made in text detection and recognition, even in challenging situations such as text in video ([Yin et al., 2016](#)), curved format ([Ch'ng & Chan, 2017](#); [Yuliang et al., 2017](#)), or images with multi-lingual content ([Nayef et al., 2019](#); [Nayef et al., 2017](#)).

Recently, the rapid development of Large Language Models (LLMs) like GPT-4 has unlocked remarkable applications in zero-shot learning across a wide range of real-world scenarios. The success of proprietary LLMs has fueled significant interest in open-source alternatives such as LLaMA ([Touvron et al., 2023](#)) and Alpaca ([Taori et al., 2023](#)). This success has also extended into the vision-language domain (Chen et al., 2023), as seen in the rise of large multimodal models like MiniCPM ([Yao et al., 2024](#)), Idefics (Laurençon et al., 2024), and Qwen-VL ([Bai et al., 2023](#)). As shown in prior work ([Liu et al., 2024](#)), these multimodal models demonstrate impressive zero-shot OCR performance, even without specialised training on OCR-specific datasets. Despite notable advancements in OCR technologies, the task remains especially challenging for complex datasets such as multilingual memes and natural scene-text images. These datasets frequently feature text in diverse fonts, sizes, orientations, and languages, presenting a demanding testbed for OCR models. Therefore, identifying these challenges is essential to foster the development of more robust models capable of recognising text in challenging scenarios.

To this end, we evaluate the performance of several state-of-the-art end-to-end OCR models on two challenging datasets: the HierText dataset ([Long et al., 2022](#); [Long et al., 2023](#)), which provides hierarchical annotations from natural scenes and documents, and a multilingual meme dataset from the SemEval 2024 shared task ([Dimitrov et al., 2024](#)). Through comprehensive benchmarking, we highlight the strengths and weaknesses of each model, identifying Google Vision and GPT-4o-mini as top performers across both datasets. Additionally, we observe that open-source models like EasyOCR offer competitive results despite being outperformed by closed-source models. Meanwhile, large multimodal models struggle with blurry and multilingual text, often generating irrelevant text. This underscores a promising opportunity to

enhance large multimodal models' OCR capabilities through domain-specific training and adaptation. This study offers valuable insights into the current strengths and limitations of OCR systems, with a focus on real-world, multilingual, and complex text extraction tasks.

5.1.2 Geolocation

Recent advances in image geolocation have followed three main methodological directions: i) classification-based, ii) retrieval-based, and iii) Retrieval Augmented Generation (RAG) methods.

Classification-based methods represent one of the earliest approaches, an early and intuitive approach to single image geolocation. These techniques operate by discretizing the where the Earth's surface is divided into a discrete number of geographical regions (grid cells). Early works in this area ([Weyand et al., 2016](#); [Seo et al., 2018](#); [Pramanick et al., 2022](#); [Clark et al., 2023](#); [Izbicki et al., 2020](#)) employed traditional computer vision features but with the advancement of deep learning Convolutional Neural Networks (CNNs) became the dominant architecture for this task, achieving significant improvements in localization accuracy by learning hierarchical features directly from image pixels. CNNs automatically learn complex visual patterns from the data, eliminating the need for manual feature engineering. More recently, there has been a shift towards leveraging pre-trained Vision-Language Models (VLMs) ([Liu et al., 2024a](#); [Haas et al., 2023](#)) like Contrastive Language-Image Pre-training (CLIP) ([Radford et al., 2021](#)). CLIP, trained on massive datasets of image-text pairs, learns highly transferable visual representations that have proven particularly effective for various downstream tasks, including geolocation.

Retrieval-based methods frame the geolocation problem as a search task within a database of geo-tagged images. These approaches aim to find the most visually similar image and infer the location of the query image based on the GPS coordinates of the retrieved image. Early retrieval-based methods ([Cao et al., 2020](#); [Vivanco Cepeda et al., 2024](#); [Lee et al., 2022](#); [Shao et al., 2023](#); [Zhu et al., 2022](#)) relied on hand-crafted features, such as color histograms or texture descriptors, to represent image content and employed various distance metrics, such as Euclidean distance or cosine similarity, to measure visual similarity. More recent advancements leverage deep learning for feature extraction, learning powerful and discriminative representations that capture complex visual semantics. Retrieval can be performed either image-to-image, directly finding the most similar image ([Cao et al., 2020](#); [Lee et al., 2022](#); [Shao et al., 2023](#)), or image-to-GPS ([Vivanco Cepeda et al., 2024](#)), where the model directly predicts GPS coordinates based on the retrieved images.

The emergence of Retrieval-Augmented Generation (RAG) methods ([Zhou et al., 2024](#); [Jia et al., 2024](#)) represents a more recent and sophisticated approach to single image geolocation. These methods leverage the power of large multimodal models (LMMs) by incorporating relevant geographical context retrieved from an external knowledge source. Typically, a retrieval module identifies geographically relevant images or textual information, such as descriptions of landmarks, maps, or satellite imagery, which is then fed into the LMM along with the query image. This allows the LMM to reason about the location based on both the visual content and the provided context, leading to state-of-the-art performance.

While RAG methods show significant promise, they also introduce complexities related to the effectiveness of the retrieval component. The performance is dependent on the quality of the retrieved

information, and if the retrieval module fails to find relevant context, the LMM may not be able to make an accurate prediction. Furthermore, the computational demands of large language models can be significant, requiring substantial computational resources and time.

5.1.3 Geolocation Interpretability

Geolocalization from a single image is a challenging and complex problem. Unlike many other vision applications, single-image geolocalization often relies on fine-grained visual cues found in small regions of an image ([Hays et al., 2008](#); [Weyand et al., 2016](#); [Seo et al., 2018](#)). Even for images that appear to depict the same location, the buildings and vegetation in the background are crucial for distinguishing them. Similarly, for most other images, the global context covering the entire image is as significant for geolocalization as individual foreground objects. Moreover, the same location exhibits drastic appearance variations under different daytime or weather conditions.

Because of this complexity, the need for robust explainability tools becomes increasingly important. These tools provide insights into the underlying decision-making process of geolocalization models. By emphasizing fine-grained visual cues (such as the architectural style, vegetation, and any background elements) explainability tools help understand how these subtle differences contribute to model predictions. This is especially critical in scenarios where similar landmarks exist across different locations, as it helps distinguish between visually alike yet geographically distinct places. Moreover, explainability enhances the reliability of geolocalization models by ensuring that their decisions are based on meaningful and contextually relevant features, thereby enabling more transparent and trustworthy applications in real-world settings.

In recent years, Grad-CAM ([Selvaraju et al., 2017](#)), or Gradient-weighted Class Activation Mapping, has emerged as the most widely used gradient-based explainability tool in the field of computer vision and its application has been extensive in geoscience and remote sensing research ([Selvaraju et al., 2017](#), [Yang et al., 2021](#); [Tasneem et al., 2023](#)).

Several variants of Grad-CAM have been proposed. Grad-CAM++ ([Chattopadhyay et al., 2017](#)) improves upon the basic Grad-CAM by addressing its limitations in localization and interpretation. It incorporates higher-order derivatives to better handle cases where multiple regions contribute to the class score, leading to more precise and detailed activation maps. Score-CAM ([Wang et al., 2020](#)) proposes a more accurate method to assess the contribution of each location in the feature map which is not based on gradient information.

Similarly to Grad-CAM, these two variants are applied exclusively to the final convolutional layer of the network, but recent works take into account the activations of shallower layers. Layer-CAM ([Jiang et al., 2021](#)) computes the contribution of each pixel in the prediction as the maximum of the activation maps obtained at different stages of the network architecture. CSG-CAM ([Guo et al., 2023](#)) weights the activation maps using channel saliency and gradients and combines information from shallow and last convolutional layers.

Grad-CAM works by highlighting the regions in an image that are most relevant to the model's predictions, typically by visualizing the gradients at the last convolutional layer of a Convolutional Neural Network (CNN). Researchers often apply Grad-CAM to this final layer as it captures the most abstract and high-

level features of the input image. However, it is well known that in a CNN, each layer plays a relevant role in processing and transforming the input data ([Lecun et al., 2002](#); [Krizhevsky et al., 2012](#); [Simonyan et al., 2014](#)). Earlier layers capture low-level features such as edges and textures, while deeper layers progressively combine these features to encode more complex patterns and shapes. Previous studies ([Hsu and Li, 2023](#)) have observed that, in some cases, Grad-CAM applied to the last layer may not precisely highlight key regions of an image.

5.2 Methodology

For the purpose of extracting relevant visual clues for media verification, we focus on two existing systems:

- An Optical Character Recognition (OCR) system developed by USFD, able to find text in the images, to locate, and to extract the script of the text it detected
- A geolocation system⁶ proposed by CERTH, that given an image is able to provide with a location estimation, as well as a confidence score and a set of similar images in its database

Both tools had been originally developed in the context of the WeVerify⁷ project, and our goal was to perform research to further improve them so they become powerful verification tools. For the evaluation, we refer to the Key Performance Indicators (KPI) and show in detail how they were improved with respect to their initial state.

- For the **OCR tool**, we provide:
 - a **comprehensive benchmark** of modern OCR models across diverse and challenging datasets, including multilingual meme images and complex natural scenes.
 - an **evaluation framework** using character accuracy, word accuracy, and position-independent word accuracy to capture recognition performance at multiple levels.
 - a comparison of traditional OCR systems, open-source models, and large multimodal models, assessing their strengths, weaknesses, and inference efficiency.
 - an **efficiency analysis** highlighting the trade-offs between recognition accuracy and inference time, which is crucial for real-world deployment.
 - a **critical analysis of large multimodal models**, uncovering their tendency to generate irrelevant or hallucinated text and pointing to the need for better instruction tuning and OCR-specific adaptation.
- For the **geolocation tool**, we provide:
 - a thorough **analysis** of the **existing AI model** in order to obtain configurations that provide better performance.
 - A **new method on interpretability**.
 - A method to **sanitize the training dataset**, to reduce its **size** and arrive at a **simpler model**, less prone to mistakes due to overfitting.

⁶ The tool is available at <https://mever.itl.gr/location/> - access to the tool is provided upon request.

⁷ WeVerify was a Horizon 2020 project, funded under grant agreement No 825297. Website: <https://weverify.eu/>

The interpretability of the obtained results is of special interest since they are intended to be used by professional journalists and fact-checkers for verification. Indeed, it is an absolute requirement to trust the methods that they provide insights on the process they followed to arrive at their result or decision. Specifically for geolocation, we developed a variant of the CNN-explainability Grad-CAM that we shall call *Combi-CAM*, that allows us to obtain an improved heatmap of the regions of the image that the network gave the most importance when taking its geolocation decision. As in this case there is no ground-truth available to compare with, the assessment is performed by visual inspection.

5.2.1 Text Extraction and Language Identification

Much of the textual content found in our surroundings contains valuable semantic information, which is essential for various real-world applications. OCR plays a pivotal role in numerous areas such as scene-text reading ([Wang et al., 2011b](#)), document text extraction, and visual data analysis ([Mishra et al., 2019](#)). Notable advancements have been made in text detection and recognition, even under challenging conditions such as video text ([Yin et al., 2016](#)), curved formats ([Ch et al., 2017](#); [Ch'ng et al., 2017](#)), or images with multilingual content ([Nayef et al., 2019](#); [Nayef et al., 2017](#)).

Recently, the rapid development of Large Language Models (LLMs) like GPT-4 (OpenAI, n.d.) has enabled remarkable zero-shot learning applications across a wide range of real-world scenarios. The success of proprietary LLMs has also sparked interest in open-source alternatives such as LLaMA ([Touvron et al., 2023](#)) and Alpaca ([Taori et al., 2023](#)). This progress extends into the vision-language domain (Chen et al., 2023), with the emergence of large multimodal models like MiniCPM (Yao et al., 2024), Idefics ([Laurenccon et al., 2024](#)), and Qwen-VL ([Bai et al., 2023](#)). As shown in prior work, these models demonstrate impressive zero-shot OCR performance even without training on OCR-specific datasets ([Liu et al., 2024](#)). However, OCR remains particularly challenging for complex datasets such as multilingual memes and natural scene-text images. These datasets often contain text in diverse fonts, sizes, orientations, and languages, presenting a demanding testbed for OCR models. Identifying these challenges is essential to guide the development of more robust models for real-world scenarios.

To this end, we evaluate the performance of several state-of-the-art end-to-end OCR models on two challenging datasets: the HierText dataset, which provides hierarchical annotations from natural scenes and documents ([Long et al., 2022](#); [Long et al., 2023](#)), and a multilingual meme dataset from the SemEval 2024 shared task (Dimitrov et al., 2024). Through comprehensive benchmarking, we highlight the strengths and weaknesses of each model, identifying Google Vision and GPT-4o-mini as top performers across both datasets. Additionally, we observe that open-source models like EasyOCR offer competitive results despite being outperformed by closed-source models. Meanwhile, large multimodal models struggle with blurry and multilingual text, often generating irrelevant outputs. This underscores a promising opportunity to improve multimodal OCR capabilities through domain-specific training and adaptation. This study offers valuable insights into the current strengths and limitations of OCR systems, especially in multilingual and complex real-world text extraction tasks.

In this tool, we conduct an extensive zero-shot evaluation of OCR models on the SemEval 2024 Task 4 Meme dataset, as well as the new HierText scene-text dataset. Both datasets present unique challenges i.e. the Meme dataset tests the models' ability to extract text from images with complex layouts, fonts,

and background noise, while the HierText dataset assesses performance on text in real-world environments.

5.2.2 Geolocation from a Single Image

Regarding geolocation, we aim to determine the geographic location of an image using only its visual content. This capability is increasingly vital in the digital age, where manipulated or misleading images spread rapidly through social media and online platforms. Accurate geolocation helps verify the authenticity of visual claims, supporting the efforts to combat misinformation and disinformation.

Inferring location from a single image is highly complex and challenging due to the Earth's vast visual diversity and the existence of visually similar scenes in geographically distant locations. Factors like lighting, weather, season, occlusions, and camera angles further complicate the task by altering or obscuring key visual cues. Our method builds on the work of [Kordopatis-Zilos et al. \(2021\)](#), as a baseline, and we propose a deep learning-based approach that leverages a pre-trained CLIP ViT-L/14 ([Radford et al. 2021](#)) model for feature extraction, a classification module for coarse-grained location prediction, and a retrieval module for fine-grained localization using a Search within Cell (SwC) strategy. To improve geolocation accuracy, we systematically investigated key components of the geolocation pipeline, including earth partitioning strategies, presence of noise in the training dataset, training losses, hyperparameter optimisation, SwC and distance metrics for image similarity.

Our experimental setup is based on MediaEval Placing Task 2016 (MP-16) (Choi et al., 2015) dataset, which provides over four million geotagged images commonly used in large-scale image classification tasks. This dataset, randomly sampled from the YFCC100m collection ([Thomee et al., 2016](#)), offers a large and diverse collection of images with corresponding GPS coordinates, making it suitable for training and evaluating our geolocation model. We employed three evaluation datasets, Im2GPS dataset ([Hays et al., 2008](#)), Im2GPS3k dataset (Vo et al., 2017) and YFCC4k dataset ([Thomee et al., 2016](#)). We first present the implementation details of the classification module, followed by an analysis of how each of these factors impacts model performance and identifying which contribute to the improvements.

Implementation details for classification. We used the CLIP ViT-L/14 model for feature extraction due to its strong performance in capturing semantic information from images. Each image is encoded into a 768-dimensional feature vector which serves both as input to a classification head for coarse location prediction and for similarity search within the cell prediction. The classification head consists of three fully connected layers that map the CLIP embeddings to the output space representing the geographic cells. The number of output neurons in this layer corresponds to the number of cells in the Earth partition, which varies depending on the partitioning scheme and granularity.

Earth partitioning. Recognizing that the geographic distribution of images is highly non-uniform across the Earth's surface, with some regions being densely populated with images while others are sparsely represented, an adaptive partitioning strategy was employed to create a set of discrete cells that serve as the basis for the classification task. This approach is inspired by previous work ([Weyand et al., 2016](#), [Muller-Budack et al., 2018](#) and [Kordopatis-Zilos et al., 2021](#)), but extends it by considering two distinct

partitioning schemes: one based on Google's S2 Geometry Library⁸ and another based on the administrative boundaries defined by the GADM⁹ dataset.

We used the S2 Geometry Library to partition the Earth into hierarchical cells via a quadtree structure. Each training image is assigned to an S2 cell based on its GPS coordinates. As shown in Figure 17 (left), a hierarchical process was followed. The S2-based partitioning process is initialized with a coarse division of the Earth's surface, leveraging the S2 library's hierarchical structure. Experiments were conducted with different Earth partitions, ranging from 35.000 to 5.000 distinct classes, to assess the impact of grid granularity on performance.

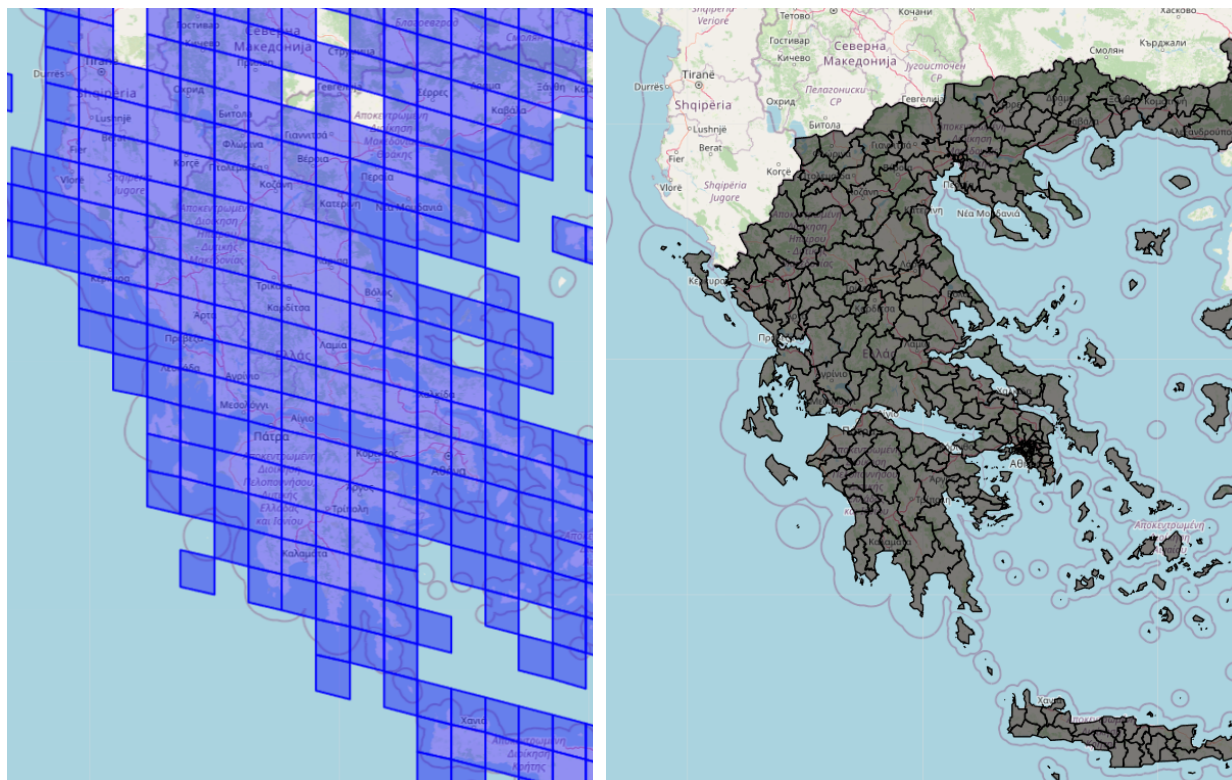


Figure 17 Earth partitioning. Left: Google S2 library (Greece). Each blue cell contains at least one image based on its coordinates. Right: GADM database (Greece). The partition reflects the administrative levels defined by GADM. Some cells may not contain images

In addition to the S2-based partitioning, we explored an alternative scheme based on the GADM dataset, as shown in Figure 17 (right). The GADM dataset defines hierarchical administrative boundaries, such as continents, countries, states, provinces, and cities. For the GADM-based partitioning (Figure 18), six levels were used, starting from continent prediction down to city levels. This allows us to explore the effectiveness of using politically defined boundaries for geolocation, as opposed to the purely geometric approach of S2. The hypothesis is that these semantically meaningful boundaries might align better with

⁸ S2 Geometry Library. <https://github.com/google/s2geometry>

⁹ <https://gadm.org/> - accessed on 4 November 2024

the visual features in images. Each image in the training set is assigned to its corresponding GADM region based on its GPS coordinates.

By employing both S2 and GADM, two distinct geographic partitioning schemes were created, each representing a different way of discretizing the Earth's surface. The S2 approach offers geometric partitioning, while the GADM approach yields a semantic partitioning based on administrative boundaries. To ensure comparability of results and to assess the influence of granularity on performance, experiments were conducted with various Earth partitions, ranging from 35.000 to 5.000 distinct classes for both S2 and GADM.

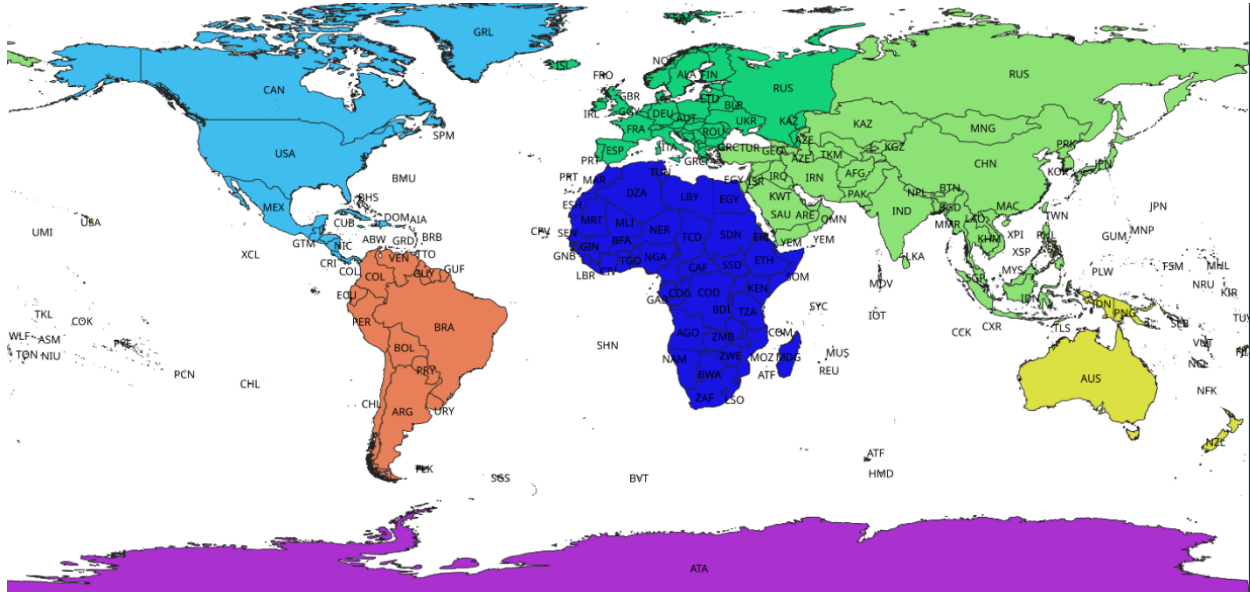


Figure 18 The Earth is partitioned into continent and country levels using the GADM database. The image displays the coarser administrative boundaries of these regions

Training dataset. The MediaEval Placing Task 2016 (MP-16) ([Choi et al., 2015](#)) dataset was selected for this research, providing over four million geotagged images commonly used in large-scale image classification tasks. This dataset, randomly sampled from the YFCC100m collection ([Thomee et al., 2016](#)), offers a large and diverse collection of images with corresponding GPS coordinates, making it suitable for the task of geolocation estimation. A thorough analysis of the training dataset revealed two significant limitations: geographic bias (see Figure 19), and the presence of non-localizable images. Non-localizable images were defined as those lacking distinctive geographic markers or visual cues that could be used to accurately identify a specific location. Examples of non-localizable images include indoor scenes, close-up shots of objects without any geographical context, or generic landscapes that could be found in many different places around the world. These images introduce noise into the training data and can hinder the model's ability to learn meaningful location-specific features.

The geographic bias was analyzed by calculating the distribution of images per continent. The analysis revealed a significant imbalance: 42% of the images originate from Europe, 36% from North America, and 12% from Asia, while the remaining 10% are distributed across South America, Oceania, and Africa. This indicates that 90% of the training data originates from three dominant regions, with South America,

Oceania, and Africa being significantly underrepresented. This geographic bias caused the model to perform better on overrepresented regions while struggling to generalize to underrepresented regions. This poses a significant challenge for any model trained on such data.

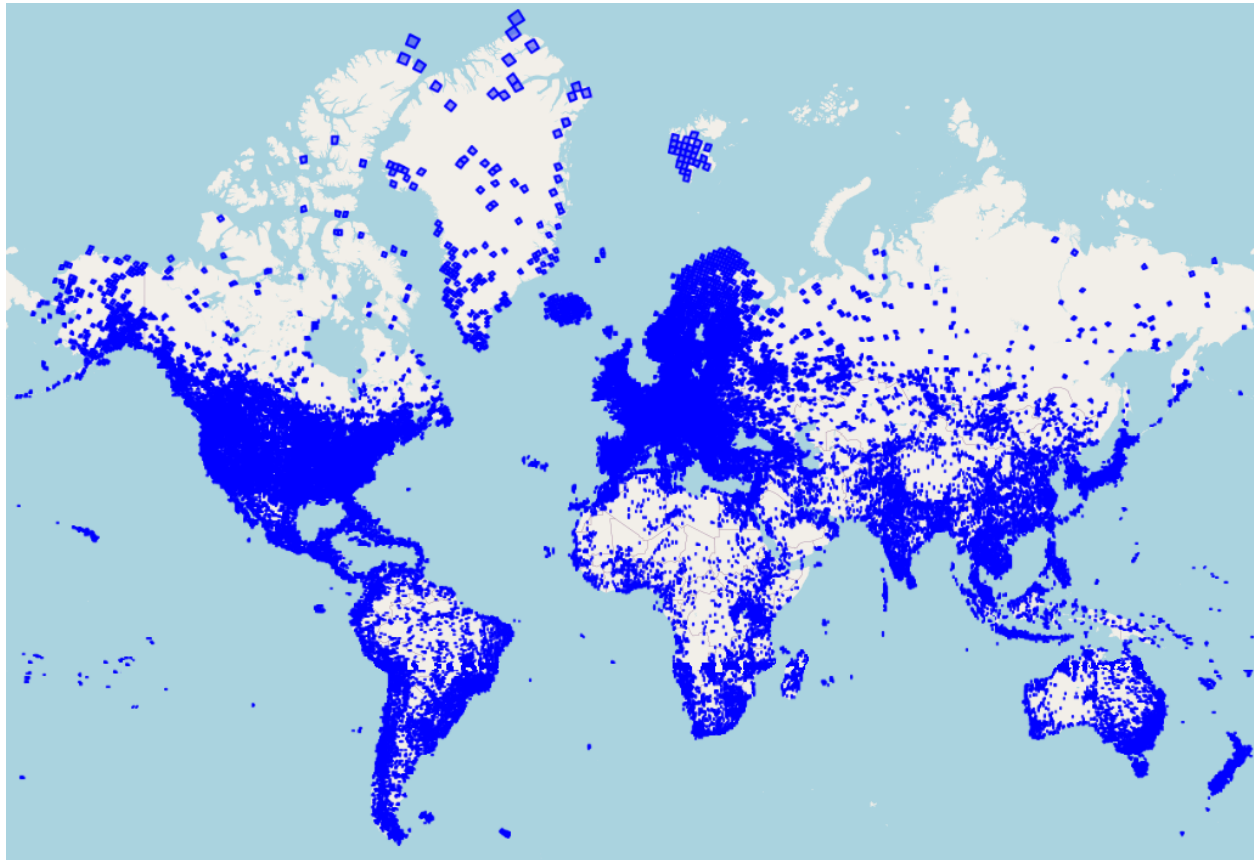


Figure 19 Distribution of MP-16 images across the world. The higher density of blue cells in Europe and North America reflects the training dataset's bias, with far fewer images in Asia, Africa, and Oceania

Training losses. To mitigate the geographic bias of the MP-16 dataset and its impact on model training, we experimented with different loss functions to improve robustness. In addition to the standard cross-entropy loss, which is sensitive to class imbalance, we evaluated label smoothing and focal loss as alternative approaches to handle uneven class distributions effectively.

Label smoothing loss ([Szegedy et al., 2016](#)) is a regularization technique that prevents the model from becoming overconfident in its predictions by slightly relaxing the confidence in the ground truth labels. Instead of using hard targets (e.g., assigning a probability of 1 to the correct class and 0 to all other classes), it assigns a small probability to other classes, encouraging the model to learn a more nuanced representation of the data. This can be particularly beneficial in cases of imbalanced datasets, like MP-16, as it can help prevent the model from overfitting to the majority classes.

Focal loss ([Ross et al., 2017](#)), on the other hand, is designed to address class imbalance by downweighting the loss assigned to well-classified examples and focusing more on hard, misclassified examples. This is achieved by adding a modulating factor to the Cross-Entropy Loss, which reduces the loss contribution from easy examples and increases the contribution from difficult ones. Focal loss can be particularly

effective when dealing with datasets where certain classes are significantly underrepresented, like MP-16.

Hyperparameter optimization. To identify the optimal hyperparameters for the model, the Optuna framework ([Cha, 2007](#)) was employed for automatic hyperparameter optimization. Optuna utilizes efficient search algorithms, such as Tree-structured Parzen Estimator (TPE), to intelligently explore the hyperparameter space and find optimal combinations. This automated approach facilitated a systematic investigation of a wide range of hyperparameter configurations, including the learning rate, learning rate scheduler parameters, step size, batch size, dropout rate, optimizer choice (e.g., Adam, AdamW, SGD), and loss function parameters (e.g., for Focal loss, cross-entropy loss, Label smoothing, etc.). The optimization process was guided by the objective of maximizing the model's geolocation accuracy on a held-out validation set. The Optuna study ran for 50 trials, incorporating early stopping based on the validation loss to prevent overfitting. Using Optuna, the model was fine-tuned to achieve an optimal balance between performance and computational cost, ensuring the robustness and reliability of the experimental results. This automated approach facilitated the identification of the optimal hyperparameter configuration for the model, leading to improved performance.

Retrieval within a cell. To combine the strengths of both the classification and retrieval modules for geolocation, an aggregation scheme was employed called Search within Cell (SwC). The SwC approach operates in two stages.

The classification module predicts the geographic cell of the query image. To reduce the risk of errors, from relying solely on the top prediction, we consider the top-k most probable cells, allowing the exploration of multiple candidate locations. In our experiments, k ranged from 1 to 5 to evaluate the impact of incorporating multiple predictions.

Within each of the top-k predicted cells, a similarity search is performed against the MP-16 database. This involves comparing the feature vector of the query image, extracted using the CLIP ViT-L/14 model, to the feature vectors of all images within the candidate cell. The images with the highest similarity scores are retrieved.

The final geolocation estimate is then derived from the GPS coordinates associated with the most similar image retrieved across all considered cells. By considering the top-k predictions, the SwC method aims to mitigate the risk of classification errors where the correct cell might not be the single most probable one but still ranks among the top few predictions. This approach leverages the complementary nature of classification and retrieval, where the classification module provides a coarse-grained location estimate, narrowing down the search space, and the retrieval module refines it based on fine-grained visual similarity within the candidate cell.

This two-stage approach facilitates the combination of the global context provided by the classification module with the local detail provided by the retrieval module.

Distance metrics. A key component of the retrieval process within the SwC framework is the choice of distance metric used to quantify the similarity between image feature vectors. The distance metric determines how we measure the similarity between two feature vectors in the high-dimensional space. To determine the most suitable metric for this task, several commonly used distance metrics were evaluated: Cosine similarity ([Salton et al., 1975](#)), Euclidean distance ([Boyd et al., 2004](#)), Hamming distance

([Hamming, 1950](#)), Manhattan distance ([Deza et al., 2009](#)), and Minkowski distance (Minkowski, 1910). Each of these metrics provides a different way of measuring the distance or dissimilarity between two vectors in a multi-dimensional space.

While most metrics yielded comparable geolocation accuracy, which indicated that CLIP embeddings are robust to different distance metrics, the cosine similarity consistently performed slightly better likely due to its ability to capture the angular similarity between feature vectors, which may be more relevant than absolute distance in high-dimensional feature spaces. In contrast, the Hamming distance performed poorly reflecting its limitations when applied to continuous data, as it is primarily designed for binary strings ([Norouzi et al., 2012](#)).

Based on these findings, cosine similarity was selected as the primary distance metric for the retrieval component of the SwC method due to its superior performance.

5.2.3 Geolocation Interpretability

We devised an approach that leverages Grad-CAM throughout several layers of the network, integrating the information from each layer into a unified heatmap. By doing so, we capture not only the high-level abstractions but also the fine-grained details and intermediate patterns that contribute to the model's decision-making process. This provides a more complete and explainable visualization of how different regions of an image influence the model's predictions, allowing for a deeper understanding of the network's internal working.

Our proposal is based on ideas similar to Layer-CAM, but the integration of the information of the different network's stages is done differently and the obtained results are more accurate in terms of localization of regions of interest in the images.

In recent years, Grad-CAM (Gradient-weighted Class Activation Mapping), has emerged as the most widely used gradient-based explainability tool in the field of computer vision and its application has been extensive in geoscience and remote sensing research. Grad-CAM works by highlighting the regions in an image that are most relevant to the model's predictions, typically by visualizing the gradients at the last convolutional layer of a Convolutional Neural Network (CNN).

Researchers often apply Grad-CAM to this final layer as it captures the most abstract and high-level features of the input image. However, it is well known that in a CNN, each layer plays a relevant role in processing and transforming the input data. Earlier layers capture low-level features such as edges and textures, while deeper layers progressively combine these features to encode more complex patterns and shapes.

By considering several layers of the network and integrating the information from each layer into a unified heatmap, we capture not only the high-level abstractions but also the fine-grained details and intermediate patterns that contribute to the model's decision-making process. This approach provides a more complete and explainable visualization of how different regions of an image influence the model's predictions, allowing for a deeper understanding of the network's internal working.

To understand Grad-CAM, let A_k^λ represent the activation of the k-th feature map from the convolutional layer λ . For a given class c , the goal is to determine the importance of each feature map A_k^λ in contributing to the class score y^c .

The importance weights α_k^c , which reflect the contribution of the k-th feature map to y^c , are computed by averaging the gradients of y^c with respect to the activations A_k^λ :

$$\alpha_k^{\{\lambda, c\}} = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ki,j}^\lambda},$$

where Z is the number of pixels in each feature map of layer λ and $\partial y^c / \partial A_{ki,j}^\lambda$ denotes the partial derivative of the class score with respect to the activation at spatial location (i, j) in the feature map A_k^λ .

Finally, the class-specific localization map for the layer $L^{(\lambda, c)}$ (or heatmap) is computed by applying a ReLU activation to the weighted sum of the feature maps, using the importance weights $\alpha_k^{(\lambda, c)}$:

$$L^{\{\lambda, c\}} = \text{ReLU} \left(\sum_k \alpha_k^{\{\lambda, c\}} \cdot A_k^\lambda \right).$$

This heatmap $L^{(\lambda, c)}$ highlights the regions in the input image that most strongly influence the prediction of class c , by emphasizing the positive contributions of each of the feature maps.

In Section 5.3 we show actual results and discuss how Combi-CAM as an interpretability method is relevant for interpreting geolocation results during verification.

5.2.4 Sanitization of Dataset

Most geolocation methods rely on large datasets of images for the training phase. However, as previously mentioned, not all images in these datasets contain sufficient geographic cues for the model to extract meaningful information. In fact, incorporating such low-information images can be detrimental: the model may focus on irrelevant or overly specific visual features, leading to overfitting and reduced overall accuracy. Given the scale of typical training datasets, manually inspecting each image is impractical. To address this, we propose an automatic approach to assess whether an image contains enough geographic information, referred to as being *localizable*, to contribute effectively to the model's learning process.

Given an image, we aim to classify it into one of two categories: Localizable or Non-Localizable. To do that we introduce a two-step method combining semantic understanding (via CLIP) and supervised classification (via XGBoost).

First, we analyze the visual content of an input image using the CLIP model, to do that we use a list of 367 location-related keywords derived from the Places365 dataset. These keywords represent various types of environments and landmarks (e.g., "beach," "airport," "library," "mountain," ...). For each image, we compute how closely it matches each of the 367 keywords using CLIP's similarity scores. We then keep

the top 5 most relevant keywords, those that the model considers to best describe the image content. This step effectively summarizes each image with a small, semantically meaningful set of tags that reflect its most prominent visual cues.

Once each image is represented by its top-5 keywords, we convert these keywords into a one-hot encoded feature vector. This vector is passed into an XGBoost classifier, which predicts whether the image is localizable or non-localizable.

To train the classifier, we build a labeled dataset of 10,000 images. To obtain the labels of this dataset, each of these images is passed through a geolocation model to estimate its geographic coordinates. We then compare the predicted location to the ground truth location. If the prediction error is greater than 2500 km, we label the image as non-localizable; otherwise, it is labeled as localizable.

The trained XGBoost model can then be used to filter out non-localizable images in larger training datasets automatically.

5.3 Evaluation

In this section we evaluate the performance of the OCR and geolocation tools with respect to the potential to fight against disinformation.

To evaluate OCR performance, we benchmarked a range of models on two representative datasets: multilingual memes (SemEval 2024) and scene-text documents (HierText). The tested models include both traditional OCR systems and large multimodal models. Results show that multimodal models like GPT-4o-mini and Gemini-1.5-pro consistently achieve higher accuracy, particularly in noisy, low-resolution, or stylized images. In contrast, traditional open-source tools such as EasyOCR and PaddleOCR provide faster inference with moderate accuracy. These findings demonstrate that OCR model selection significantly impacts text extraction quality in disinformation contexts, with trade-offs between speed and robustness depending on content complexity and language.

For geolocation, to benchmark our methodology against the state of the art, we evaluate our model's performance on the same evaluation datasets, enabling a direct assessment of its effectiveness. To assess the impact of using only localizable images, we evaluated geolocation accuracy in two scenarios: one where we only apply the localizability filter at inference time and the other where only localizable images are used for training the geolocation model. Experiments on the MP16 dataset show that filtering at inference time significantly boosts accuracy, acting as a sanity check for our sanitization method. The model trained solely on localizable images maintains performance with reduced computational cost, model complexity and overfitting. These findings highlight the effectiveness of the dataset sanitization method in real world scenarios.

5.3.1 Text Extraction and Language Identification

We tested a wide range of OCR models on the Meme and HierText datasets. These include Tesseract ([Smith, 2007](#)), EasyOCR ([Baek et al., 2019](#); [Shi et al., 2016](#); [Simonyan & Zisserman, 2014](#)), PaddleOCR ([Li et al., 2022](#)), DocTR ([Chen et al., 2021](#); [Shi et al., 2016](#)), KerasOCR ([Baek et al., 2019](#); [Shi et al., 2016](#)), MMOCR ([Kuang et al., 2021](#)) and Google Vision OCR. For MMOCR, we tested detection models like

DBNet++ ([Liao et al., 2022](#)), TextSnake ([Long et al., 2018](#)) and recognition models including ABINet ([Fang et al., 2021](#)), SATRN ([Lee et al., 2020](#)), and MASTER ([Lu et al., 2021](#)).

We further assessed open-source multimodal models: Qwen-VL ([Bai et al., 2023](#)), Monkey ([Li et al., 2024](#)), Idefics2 ([Laurenccon et al., 2024](#)), Phi-3-Vision-128K-Instruct (Abdin et al., 2024), MiniCPM-v2 ([Yao et al., 2024](#)), InternLM-XComposer2-vl-7b ([Dong et al., 2024](#)), LLaVA-v1.6-Mistral-7b ([Liu et al., 2024b](#)), and mPLUG-DocOwl-1.5 ([Hu et al., 2024](#)). These models were prompted with: “Recognise the text in this image and return only the text as a string.” We also evaluated closed-source models like GPT-4o-mini (OpenAI, n.d.) and Gemini-1.5-pro by Google DeepMind.

To evaluate the performance of the OCR models, we use three different metrics to assess the quality of text recognition at both the character and word levels. These metrics are as follows,

Character Accuracy: Character Error Rate (CER) measures the ratio of character-level errors (insertions, deletions, substitutions) to the total characters in the ground truth. It is calculated as:

$$CER = \frac{\text{Character – level Levenshtein Distance}}{\text{Total Characters}}$$

Character Accuracy is 1 - CER, with higher values indicating better character recognition.

Word Accuracy: Word Error Rate (WER) assesses the ratio of word-level errors to the total words in the ground truth, computed as:

$$WER = \frac{\text{Word – level Levenshtein Distance}}{\text{Total Words}}$$

Word Accuracy is 1 - WER, evaluating the overall word recognition performance.

Position Independent Word Accuracy: Position Independent Word Error Rate (PI – WER) measures word error rate without considering word order. Position Independent Word Accuracy is calculated as 1 – PI – WER, which is useful for scenarios where one does not want to penalise wrong word order.

These metrics provide a comprehensive evaluation of character and word recognition. Please note that both character and word accuracy can be negative when the number of errors exceeds the total number of characters or words in the reference, indicating particularly poor recognition performance.

Results and Discussion

Memes Dataset: Table 18 presents the OCR benchmark results for various models tested on the English subset of the Meme dataset. The performance is evaluated using three key metrics: Character Accuracy (Char Acc), Word Accuracy (Word Acc), and Position Independent Word Accuracy (PI-Word Acc), with results shown for both case-sensitive and case-insensitive settings. Additionally, we report the processing time for each model.

The results indicate significant variability in performance across the different models. Text recognition models like ABINet (Dbnetpp+ABINet and TextSnake+ABINet) and SATRN (DBNetpp+SATRN) consistently show lower scores. In contrast, Tesseract achieves a character accuracy of 0.2578 in case-sensitive mode, but its word accuracy drops drastically to 0.0486, highlighting its difficulty in handling complete words.

Modern models like EasyOCR and PaddleOCR perform more consistently across both character and word-level metrics. EasyOCR, for example, achieves a balanced performance with a character accuracy of 0.4615 and a word accuracy of 0.2891 in the case-sensitive setting. These models also tend to have lower processing times, indicating their efficiency in handling OCR tasks on meme-type images.

Table 18 OCR benchmark results on English Memes dataset. The best results are in bold

Model	Char Acc (CS)	Word Acc (CS)	PI-Word Acc (CS)	Char Acc (CI)	Word Acc (CI)	PI-Word Acc (CI)	Time (s)
EasyOCR	0.4615	0.2891	0.2864	0.4739	0.3063	0.3083	0.1938
PaddleOCR	0.4656	0.2736	0.2851	0.4698	0.2777	0.2896	0.8069
Tesseract	0.2578	0.0486	0.0218	0.2734	0.0517	0.0319	0.7330
DocTR	0.2783	-0.0776	-0.0748	0.2867	-0.0679	-0.0600	6.0000
Dbnetpp+ABINet	-0.1798	-0.4171	-0.3660	0.0669	-0.2197	0.0301	0.6000
TextSnake+ABINet	-0.0720	-0.1508	-0.1506	0.2060	-0.1120	-0.1039	0.4400
DBNetpp+SATRN	-0.0159	-0.2170	-0.0069	0.0091	-0.1982	0.0312	3.0000
DBNetpp+MASTER	-0.0029	-0.2174	0.0044	0.0209	-0.1982	0.0418	1.2000
Keras OCR	0.1964	-0.0531	0.2167	0.1964	-0.0531	0.2167	0.6000
mPLUG-DocOwl-1.5	0.2709	0.1977	0.1840	0.3883	0.3393	0.3352	0.9071
Qwen-VL	0.3656	0.2387	0.2071	0.6255	0.5698	0.5674	0.5120
Monkey	0.2952	0.2054	0.1855	0.4852	0.4390	0.4367	0.9100
Idefics2-8b	0.5680	0.4584	0.4393	0.7894	0.7165	0.7193	3.3368
Phi-3-Vision-128K-Instruct	0.5795	0.4319	0.4241	0.6042	0.4643	0.4650	1.8000
MiniCPM-v2	0.5466	0.4489	0.4397	0.6321	0.5526	0.5574	0.7224
InternLM-XComposer2-vl-7b	0.4806	0.3868	0.3734	0.7457	0.7056	0.7104	3.3000
LLaVA-v1.6-Mistral-7b	0.4643	0.3751	0.3571	0.5909	0.5318	0.5324	1.5800
Gemini-1.5-pro	0.5817	0.5768	0.5813	0.5852	0.5820	0.5889	2.2000
GPT-4o-mini	0.6994	0.6790	0.6768	0.7098	0.6926	0.6931	3.6000
Google Vision	0.4175	0.4074	0.4180	0.4217	0.4134	0.4236	0.2776

Open-source large multimodal models, such as Idefics2 and MiniCPM-v2, exhibit strong performance in both case-sensitive and case-insensitive evaluations. Idefics2, for instance, achieves a character accuracy of 0.5680 in case-sensitive mode, which improves to 0.7894 in case-insensitive mode, demonstrating its

ability to adapt to various text formats and cases. However, these models generally have higher processing times, with Idefics2 taking 3.3368 seconds per image. This can be a drawback for large-scale applications.

Case-sensitive and case-insensitive character/word accuracies for six multilingual OCR engines on three challenging meme corpora are reported in Annex III.I. Among the closed-source large multimodal models, GPT-4o-mini shows the highest accuracy across the board, with character and word accuracies of 0.6994 and 0.6790, respectively, in case-sensitive mode. Interestingly, GPT-4o-mini performs even better than the Google Vision API. Their performance further improves in the case-insensitive setting, suggesting that they are capable of handling OCR tasks effectively, even when confronted with diverse text styles and cases.

Overall, whereas OCR models like EasyOCR and PaddleOCR provide decent baseline performance, more modern models and large multimodal systems like GPT-4o-mini and Idefics2 demonstrate superior results, particularly in handling complex, irregular, and multilingual text found in meme-type images. The choice of OCR model depends on the trade-off between accuracy and processing time, with vision-language models offering high accuracy at the cost of increased inference time.

HierText Dataset: The OCR benchmark results on the HierText dataset (Table 19) reveal notable differences in model performance. Google Vision leads the results, achieving the highest character and word accuracies, with a word accuracy of 0.6256 in the case-sensitive setting and 0.6326 in the case-insensitive setting. In contrast, large multimodal models like Monkey, Phi-3-Vision-128K-Instruct, and LLaVA-v1.6-Mistral-7b display negative accuracies, indicating severe underperformance on this scene-text dataset. This underperformance stems from the models generating irrelevant text, such as "Sorry, but I cannot recognise any text due to a blurred image" or "No text is present in the image." Despite attempts to explicitly instruct the models not to generate such text and return an empty string instead, they consistently failed to comply.

Other models, such as Gemini-1.5-pro and EasyOCR, show competitive performance, with Gemini-1.5-pro attaining a word accuracy of 0.3076 and EasyOCR achieving 0.2207 in the case-sensitive setting. Several open-source models, like ABINet, SATRN, and MASTER, struggle significantly, with word accuracies close to zero, despite having relatively high character accuracy. This is because these models tend to correctly recognise individual characters but struggle with word-level accuracy due to poor handling of complex scene text structures or inconsistencies in word segmentation. Additionally, we found that models often merge multiple words into one, ignoring the spaces that separate individual words.

Table 19 OCR Benchmark Results on HierText Dataset (Original with 4 Decimal Places).

Models	Char Acc (CS)	Word Acc (CS)	PI-Word Acc (CS)	Char Acc (CI)	Word Acc (CI)	PI-Word Acc (CI)	Time
EasyOCR	0.4162	0.2207	0.2172	0.4457	0.2412	0.2390	0.0420
PaddleOCR	0.3838	0.1638	0.1542	0.3858	0.1656	0.1576	0.1100
Tesseract	0.3634	0.2262	0.2149	0.3712	0.2318	0.2214	0.1900
ABINet	0.3274	0.0045	0.0045	0.5416	0.0070	0.0070	0.0700
SATRN	0.4862	0.0025	0.0025	0.5479	0.0025	0.0025	0.3600
MASTER	0.4888	0.0005	0.0005	0.5410	0.0005	0.0005	0.2700
KerasOCR	0.3476	0.0649	0.2197	0.3476	0.0649	0.2197	0.3400
Monkey	-0.5251	-0.7472	-0.7513	-0.4356	-0.6932	-0.6976	0.7500
Idefics2-8b	0.2815	0.0331	0.0089	0.3423	0.0956	0.0852	1.0500
Phi-3-Vision-128k-Instruct	-1.2023	-1.1258	-1.1328	-1.0325	-1.0004	-1.0020	1.0000
MiniCPM-v2	0.6677	0.1469	0.1435	0.6828	0.1539	0.1509	0.2550
InternLM-XComposer2-vl-7b	-0.7128	-0.5100	-0.5199	-0.5895	-0.4016	-0.4060	2.7000
LLaVA-v1.6-Mistral-7b	-1.9409	-1.9976	-1.9993	-1.8073	-1.8716	-1.8696	1.0300
Gemini-1.5-pro	0.3456	0.3076	0.3012	0.3674	0.3210	0.3158	2.9000
GPT-4o-mini	0.2655	0.1480	0.1361	0.2969	0.1606	0.1506	1.5800
Google Vision	0.7580	0.6256	0.6197	0.7607	0.6326	0.6287	0.1500

Regarding inference speed, similar to the Meme dataset (Table 19), large multimodal models like Monkey and LLaVA-v1.6-Mistral-7b take significantly longer, with times of 0.7500 and 1.0300 seconds, respectively. In contrast, open-source models such as EasyOCR and PaddleOCR offer faster processing times of 0.0420 and 0.1100 seconds, making them more suitable for time-sensitive applications.

Efficiency Analysis

We assess the efficiency of various OCR models, which are defined as character accuracy (case-insensitive) divided by the inference time taken per image. Higher efficiency values indicate that a model can perform accurate OCR quickly, making it more suitable for real-world applications where both accuracy and speed are critical. All the open-sourced models are tested on NVIDIA A100 cards, except for PaddleOCR, which could not be run on GPU due to the complexity involved. For closed-source models, we used their respective APIs for evaluation.

Figure 20 illustrates the efficiency results for the models evaluated on the English subset of the Meme dataset. It shows that EasyOCR emerges as the most efficient model with a score of 2.4454, followed by Google Vision with an efficiency of 1.5191 and Qwen-VL at 1.2217. These models demonstrate a strong balance between character accuracy and inference speed, making them suitable for high-demand environments. On the other hand, large multimodal models like Qwen-VL and MiniCPM-v2 also show competitive efficiency values of 1.2217 and 0.8750, respectively, indicating their capacity to handle OCR tasks effectively while maintaining moderate speed.

Overall, the efficiency metric offers valuable insights into the trade-offs between speed and accuracy in OCR tasks. The results show that while advanced multimodal models are improving, other OCR models like EasyOCR remain more efficient for text extraction tasks. These findings suggest an opportunity for optimising the speed and accuracy of multimodal models to make them more practical for real-world OCR applications.

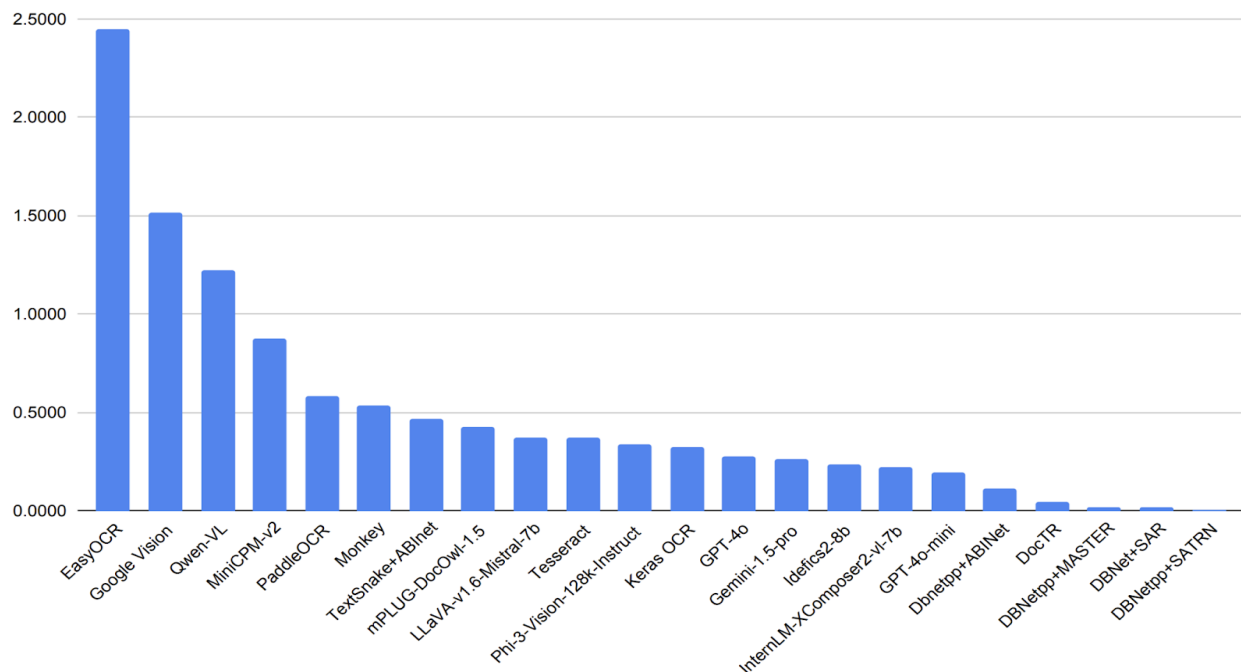


Figure 20 Efficiency analysis of OCR models. The Y-axis represents efficiency, defined as case-insensitive character accuracy divided by the inference time per image

5.3.2 Geolocation Evaluation

Evaluation datasets. To evaluate the performance of the proposed geolocation approach, four established datasets were employed, providing a comprehensive benchmark for comparison against existing methods. The MP-16 dataset served as the source for both training and retrieval. For comparative evaluation against state-of-the-art methods, the Im2GPS dataset ([Hays et al., 2008](#)), Im2GPS3k dataset ([Vo et al., 2017](#)) and YFCC4k dataset ([Thomee et al., 2016](#)) datasets were selected. These datasets offer a diverse range of image characteristics and varying levels of difficulty, allowing for a thorough assessment of the proposed method's performance and its comparison to existing approaches in the field.

Similar to the MP-16 training dataset, the evaluation datasets revealed limitations regarding geographic bias and the presence of non-localizable images. These evaluation sets also disproportionately represent images from Europe, North America, and Asia, with far fewer images from South America, Oceania, and Africa. This geographic bias can affect the generalizability of the models, potentially leading to better performance on images from overrepresented regions. Non-localizable images are those that lack distinctive geographic markers or visual cues that can accurately identify a specific location. These images introduce noise into the evaluation process and can make it difficult to assess the true performance of the models.

To address the second limitation, a manual examination was performed on the Im2GPS dataset, which is the smallest in number of images and is feasible for manual inspection. Rules were established to define non-localizable images based on the definition of 'lacking distinctive geographic markers or visual cues' (see Figure 21 and Figure 22 for examples). Upon examination of the datasets, the presence of noise due to non-localizable images was observed, which could potentially affect the evaluation accuracy. However, to maintain consistency and allow for direct comparisons with existing work, we used these datasets for evaluation purposes despite the noise. This ensures that the method's performance is assessed under the same conditions as in previous studies.

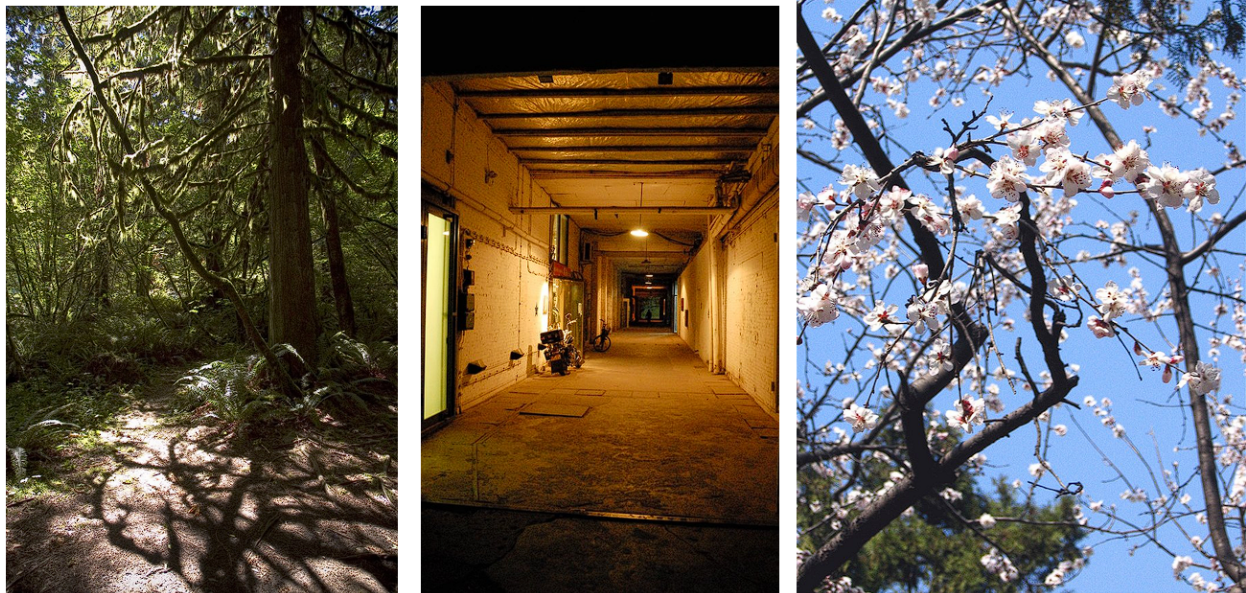


Figure 21 Example of non-localizable images from the Im2GPS evaluation dataset. These images could have been taken in many different locations in the world, as they lack distinctive features that would help narrow down their location



Figure 22 Example of non-localizable images from the Im2GPS evaluation dataset. These images could have been taken in many different locations in the world, as they lack distinctive features that would help narrow down their location

Evaluation metrics. The evaluation of geolocation performance is based on the percentage of images that are placed within a predefined granularity range. An image is considered correctly placed when the haversine distance of the estimated location to the ground truth is lower than the granularity range. The haversine distance is calculated as the Great Circle Distance (GCD) between the two locations, which is the shortest distance between two points on the surface of a sphere. This metric is appropriate for measuring distances on the Earth's surface. Five granularities are considered, including 1km, 25km, 200km, 750km, and 2500km, corresponding roughly to street, city, region, country, and continent granularity levels, respectively. These granularity levels allowed us to assess the performance of the models at different levels of precision, from fine-grained street-level localization to coarse-grained continent-level localization.

In Annex III.II, Table AIII- 4, Table AIII- 5 and Table AIII- 6 present a subset of our experiments along with their results. As can be observed, the Google S2 35k model achieved the highest performance across all evaluation datasets. Therefore, we use this model for comparison against state-of-the-art approaches throughout our analysis.

Comparison with state-of-the-art methods

The proposed method is compared against state-of-the-art approaches, including classification-based, retrieval-based, and Retrieval-Augmented Generation (RAG) methods.

As anticipated, the G3 ([Jia et al., 2024](#)) and Img2Loc ([Zhou et al., 2024](#)) models demonstrate state-of-the-art performance across the evaluation datasets (Table 20, Table 21 and Table 22), reflecting their recent development and unique integration of RAG with, specifically, GPT-4V ([Achiam et al. 2023](#)). The superior performance of these models underscores the significant advantage conferred by leveraging LLMs and external knowledge bases using RAG, a factor that was critical in achieving their high accuracy. While the proposed model does not incorporate RAG, it exhibits competitive performance compared to other non-RAG-based methods. Notably, in certain finer granularities, the proposed method outperforms several established approaches.

Table 20 Accuracy (%) on five granularity ranges of the proposed method compared to state-of-the-art methods on the Im2GPS evaluation set

Im2GPS					
Method	1km	25km	200km	750km	2500km
PlaNet (Weyand et al., 2016)	8.4	24.5	37.6	53.6	71.3
Translocator (Pramanick et al., 2022)	19.9	<u>48.1</u>	<u>64.6</u>	75.6	86.7
GeoGuessNet (Clark et al., 2023)	22.1	50.2	69.0	<u>80.0</u>	89.1
GeoCLIP (Vivanco Cepeda et al., 2024)	13.6	41.3	59.3	76.8	89.4
PIGEON (Hass et al., 2024)	14.8	40.9	63.3	82.3	<u>91.1</u>
Leveraging EfficientNet (Kordopatis-Zilos et al., 2021)	<u>21.9</u>	44.3	55.3	67.5	81.9
Google S2 35K (Our method)	16.4	46.4	62.0	77.7	91.6

Table 21 Accuracy (%) on five granularity ranges of the proposed method compared to state-of-the-art methods on the Im2GPS3k evaluation set

Im2GPS3k					
Method	1km	25km	200km	750km	2500km
PlaNet (Weyand et al., 2016)	8.5	24.8	34.3	48.4	64.6
Translocator (Pramanick et al., 2022)	11.8	31.1	46.7	58.9	80.1
GeoGuessNet (Clark et al., 2023)	12.8	33.5	45.9	61.0	76.1
GeoCLIP (Vivanco Cepeda et al., 2024)	14.1	34.5	50.6	69.7	83.8
PIGEON (Hass et al., 2024)	11.3	36.7	53.8	<u>72.4</u>	85.3
Leveraging EfficientNet (Kordopatis-Zilos et al., 2021)	15.0	30.0	38.0	52.3	67.6
Img2Loc (Zhou et al., 2024)	17.1	45.1	57.9	72.9	<u>84.7</u>
G3 (Jia et al., 2024)	<u>16.6</u>	<u>40.9</u>	<u>55.6</u>	71.2	<u>84.7</u>
Google S2 35K (Our method)	13.5	37.0	50.3	66.9	81.9

Table 22 Accuracy (%) on five granularity ranges of the proposed method compared to state-of-the-art methods on the YFCC4k evaluation set

YFCC4K					
Method	1km	25km	200km	750km	2500km
PlaNet (Weyand et al., 2016)	5.6	14.3	22.2	36.4	55.8
Translocator (Pramanick et al., 2022)	8.4	18.6	27.0	41.1	60.4
GeoGuessNet (Clark et al., 2023)	10.3	24.4	33.9	50.0	68.7
GeoCLIP (Vivanco Cepeda et al., 2024)	9.59	19.3	32.6	55.0	68.7
PIGEON (Hass et al., 2024)	10.4	23.7	40.6	<u>62.2</u>	<u>77.7</u>
Leveraging EfficientNet (Kordopatis-Zilos et al., 2021)	7.9	14.3	21.9	<u>37.4</u>	<u>56.5</u>
Img2Loc (Zhou et al., 2024)	14.1	<u>29.5</u>	<u>41.4</u>	59.3	76.9
G3 (Jia et al., 2024)	24.0	35.9	47.0	64.2	78.1
Google S2 35K (Our method)	<u>16.2</u>	27.2	39.0	59.2	75.3

5.3.3 Geolocation Interpretability with Combi-CAM

In the case of Combi-CAM, we do not have any metric to evaluate the results, as the output of the method is a heatmap that highlights which parts of the image were the most relevant for the geolocation decision of the network. The evaluation is therefore done by visual inspection by experts.

Consider the image in the Opera of Sydney as presented in Figure 23. The geolocation tool correctly located the image, but this should not be enough for a fact-checker to trust the tool. Even though the coordinates of the geolocation tool come with a confidence score, it could be that the elements used to make the decision are not the most adequate. For example, imagine that for the image in the Opera of Sydney the network would have used information mainly from the sea coast, and not from the Opera itself. In that case it could happen that the result is correct, but certainly one should not trust it if no relevant features were taken into account. However, we can observe that this was not the case for the Opera example, since the network actually used proper visual clues, namely the Opera itself, and specific parts of other buildings. This, along with a large confidence score, implies a certain level of trustworthiness.

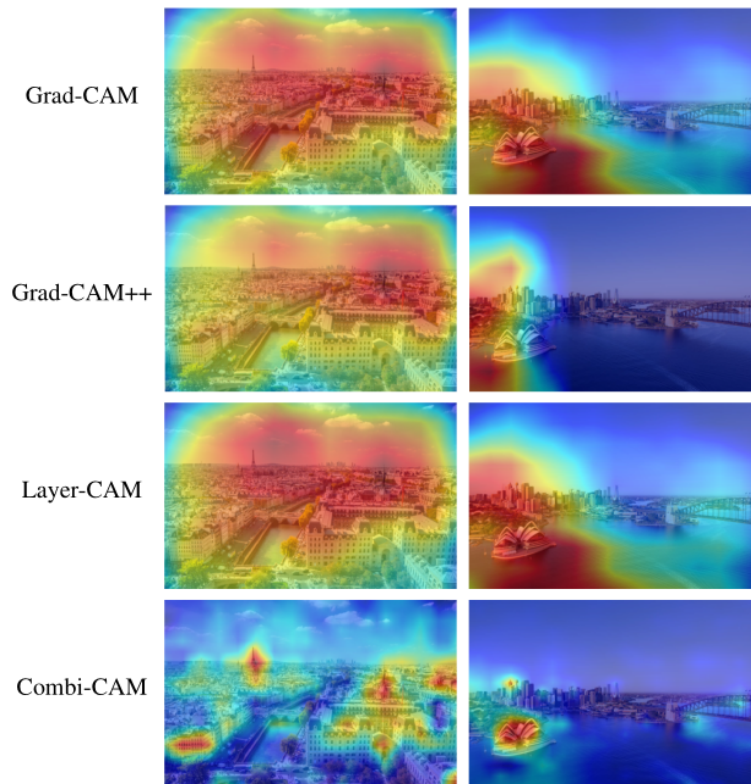


Figure 23 Activation maps obtained with different methods for the two different scenes, one the Opera of Sydney, and the other a view of Paris. The geolocation tool is able to correctly locate the both scenes, and thanks to Combi-CAM we can observe that it looked mainly at the tops and facades of buildings in the case of Paris, and at the Opera itself in the image of Sydney. This gives insights about which elements were used by the network to make the final decision

5.3.4 Sanitization of Dataset

Once the set of localizable images is defined and identified, the next step is to evaluate whether restricting training and inference to only this type of images has a measurable impact on geolocation accuracy. To this purpose, we conducted a benchmarking study across two scenarios:

1. **Inference only** — using the localizability filter at inference time;
2. **Training only** — using only localizable images during model training.

For all experiments, we use a subset of the MP16 dataset, containing a large collection of images including both localizable and non-localizable images.

Accuracy at inference. To assess the impact of filtering based on localizability at inference time, we evaluated geolocation accuracy on a subset of 10,000 images sampled from the MP16 dataset. We retrieved localizable images using the XGBoost classifier, with a confidence threshold of 0.5. This results in two groups, one composed of 4830 localizable images and the other of 5170 non-localizable images. We then run inference using the geolocation model on the entire subset and compare accuracy across various distance thresholds.

These results (Table 23) show a clear performance gap between localizable and non-localizable images, confirming that localizability correlates strongly with geolocation success. While this outcome is expected, since we filtered the dataset based on criteria that imply geolocation relevance, it also confirms that our method for identifying localizable images is meaningful.

Table 23 Accuracy (%) and count at five granularity ranges of the model performance at inference for all 10,000 images, 4830 localizable images and 5170 non localizable images of the MP16 dataset.

	All images		Localizable only		Non Localizable only	
Distance Range	Count	Percentage	Count	Percentage	Count	Percentage
Less than 2500 km	7671	76.71%	4237	87.72%	3434	66.42%
Less than 750 km	6199	61.99%	3758	77.80%	2441	47.21%
Less than 200 km	4638	46.38%	3105	64.28%	1533	29.65%
Less than 25 km	3729	37.29%	2684	55.56%	1045	20.21%
Less than 1 km	3123	31.23%	2201	45.66%	922	17.80%

Accuracy for training only:

To benchmark the improvement that can be achieved using the sanitization of the dataset we test two models:

- **Full model:** The model trained on all the images of the dataset, i.e, the localizable images and non-localizable. It has been trained on the full MP16 dataset, i.e, 4.6 million images, until convergence.
- **Partial model:** The model trained on only the localizable images of the dataset. It has been trained on 2.3 million localizable images from the MP16 dataset, until convergence.

The results from those two models, benchmarked against a training set of 2000 images taken from the im2gps dataset, without sanitizing them beforehand, are found in Table 24.

Table 24 Accuracy (%) and count at five granularity ranges of the two models' performance, after being trained only on localizable images (partial model) and all the images (full model)

	Partial Model		Full Model	
Distance Range	Count	Percentage	Count	Percentage
Less than 2500 km	1506	75.30%	1523	75.15%
Less than 750 km	1195	59.75%	1212	60.60%
Less than 200 km	833	41.65%	853	42.62%
Less than 25 km	599	29.95%	618	30.90%
Less than 1 km	247	12.35%	255	12.75%

These results reflect performance at the end of training, where both models have largely converged. Notably, accuracy remains comparable whether the model is trained on the full dataset or only on the localizable subset. This indicates that using only localizable images, effectively almost halving the dataset, achieves similar performance to training on the entire set. In practice, the **partial model** required approximately **half the computational resources**, and **half the training time** compared to the full model, highlighting the efficiency gains of filtering the training data.

While both networks share the same architecture, and thus the same number of trainable parameters, the effective complexity of the resulting models can vary depending on the training data. A model trained on a cleaner, more geographically informative dataset, i.e., only localizable images, may require less resources to achieve similar performance. This motivates an investigation into how training data quality affects model complexity.

To test this, we used the two trained models, full and partial, and recorded them during the training epochs. Then, we assessed their complexity using the spectral norm-based generalization bound, which involves computing the largest singular value of each weight matrix and their stable rank, then multiplying them across the network. This gives an upper bound on the model's capacity to generalize.

Figure 24 shows how this bound evolves across training epochs. The partial model (blue) consistently maintains a lower generalization bound than the full model (red). This gap is present across all epochs and stabilizes at around a 20% difference. A lower generalization bound implies the model is less complex, which is typically associated with better generalization and lower sensitivity to noise.

These results suggest that training with only localizable images not only maintains accuracy but also leads to a simpler and more efficient learned function. The reduced complexity could stem from:

- Lower variance in the training set;
- Elimination of confusing or uninformative visual features.

In essence, our results show that filtering the training data to include only localizable images promotes better generalization and may reduce the learning burden on the model without compromising performance and furthermore reducing overfitting.

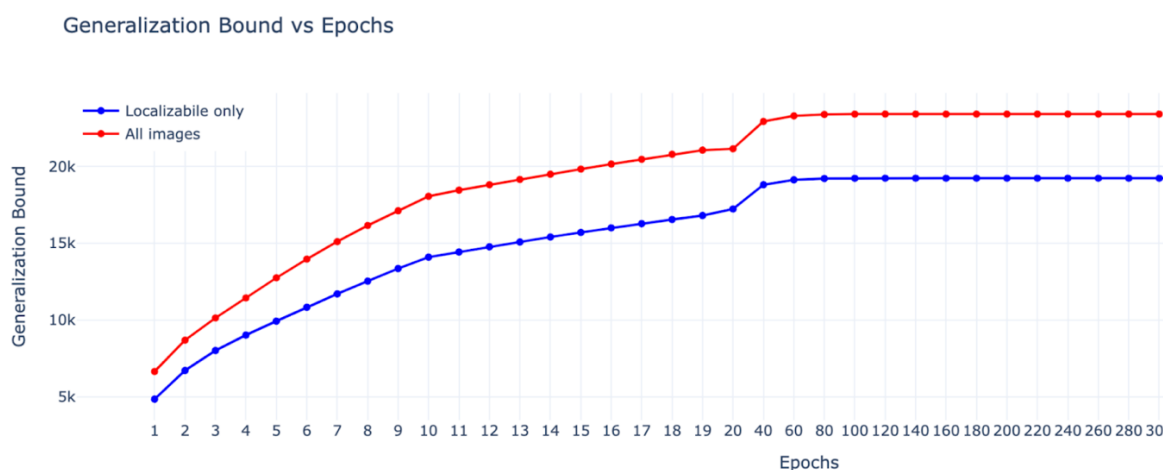


Figure 24 Bound on the complexity of the full and partial model. The Y-axis represents the value of the generalization bound, which can be likened to model complexity, the X-axis is the Epoch number, i.e., training time.

5.4 Implementation and Integration

For all the methods (OCR, geolocation, and interpretability), we expose results in the form of services accessible with their corresponding API to fact-checkers and professional journalists, and online demos for everyone, professionals or the general public, to try.

We analyze the existing tools and compare them in terms of technology and performance with the state of art, as a baseline from which the performed research is aimed at improving the proposed KPIs. Both fundamental research, as well as an exploration of the parameters and the architecture of the corresponding neural networks, are applied. We arrive at new methods and techniques that, once integrated, will improve the existing OCR and Geolocation tools.

5.4.1 Text Extraction and Language Identification

We have made the OCR approach available as a simple to use REST service, accessible via GATE Cloud. There are two ways to call the API. The first allows you to pass in the URL of an image to process, whilst the second allows you to upload an image to the service. Both approaches require calling the following endpoint (calls can be authenticated using a GATE Cloud API Key via Basic Authentication to enable higher request quotas and rate limits): <https://cloud-api.gate.ac.uk/process/ml-ocr/process>

To process an image available at a publicly accessible URL, one can simply make a GET request passing the URL as the url query parameter. If instead one wishes to upload an image for processing, then they can POST the file contents to the endpoint instead.

Regardless of how one calls the service, the response will take the same form. Specifically, the service returns a JSON array, where each element describes a bounding box within the image where text was detected. An example of such an object would be:

```
{
  "text": "Some text extracted from the image",
  "bounding_box": [[18,10],[1038,10],[1038,114],[18,114]],
  "language": {
    "code": "en",
    "name": "English",
    "probability": 0.6472072005271912
  }
}
```

5.4.2 Geolocation from a Single Image

For integration purposes, we leveraged the already existing CERTH's geolocation estimation service, which provides a REST API. The method is accessible via the `/location/v3/geolocate` endpoint, where users submit an image URL. An example API response for a completed job of a submitted image URL is provided in Figure 25 while Figure 26 and Figure 27 illustrate the resulting detection as displayed in a standalone demo developed for demonstration purposes. The result is presented as a pin on a map, providing a human-readable location along with a confidence score reflecting the certainty of the detected

coordinates. Figure 27 additionally shows similar images retrieved from the background dataset, which serve as supporting evidence for the estimated location.



Figure 25 API response of geolocation estimation REST API

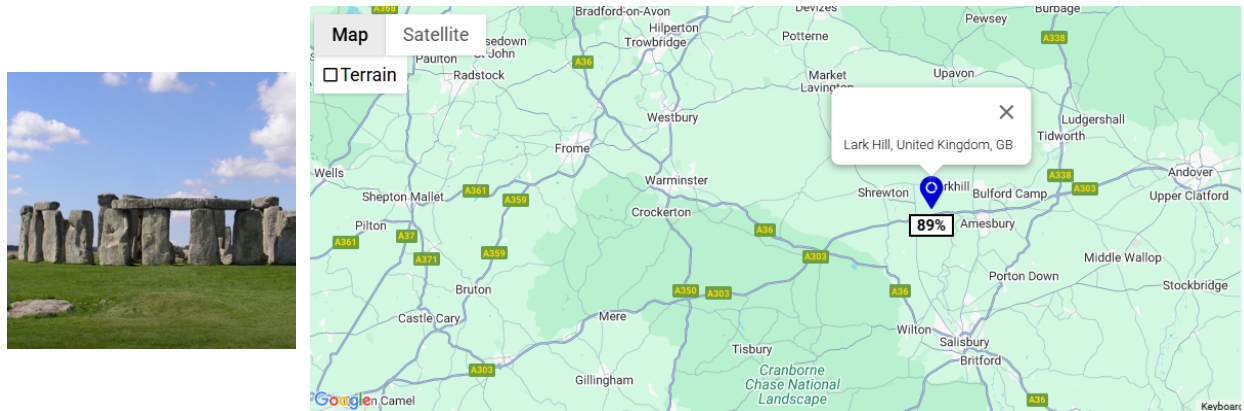


Figure 26 Predicted most probable location with associated confidence score output by the model



Figure 27 Relevant images retrieved from our database corresponding to the estimated location

5.4.3 Geolocation Explainability with Combi-CAM

An online demo for the Combi-CAM method (Figure 28), our new method for geolocation explainability, is available at [IPOL](https://ipolcore.ipol.im/demo/clientApp/demo.html?id=77777000546)¹⁰. The demo allows any user to upload an image and obtain a heatmap that highlights which parts of the image were mainly used to make the final decision. It also allows comparison with other state of the art, including Grad-CAM, Grad-CAM++, Score-CAM, and Layer-CAM. The demo is open to all users, without the need of any registration.

Besides the online demo, the method can be accessed by a REST API, for everyone to test with their own data, on the endpoints <https://integration.ipol.im/vera>.

For transparency and reproducible research purposes, the complete source code of the method is available via [GitHub repository](https://github.com/DavidFaget/Combi-CAM/)¹¹ under a free-software licence, the GPL 3.

Finally, a preprint, entitled '[Enhancing Aerial Geolocalization Explainability With a Novel Grad-CAM Approach](https://www.techrxiv.org/users/836918/articles/1228608-enhancing-aerial-geolocalization-explainability-with-a-novel-grad-cam-approach)¹²' with all the details is available in TechRXiv

¹⁰ <https://ipolcore.ipol.im/demo/clientApp/demo.html?id=77777000546>

¹¹ <https://github.com/DavidFaget/Combi-CAM/>

¹² <https://www.techrxiv.org/users/836918/articles/1228608-enhancing-aerial-geolocalization-explainability-with-a-novel-grad-cam-approach>

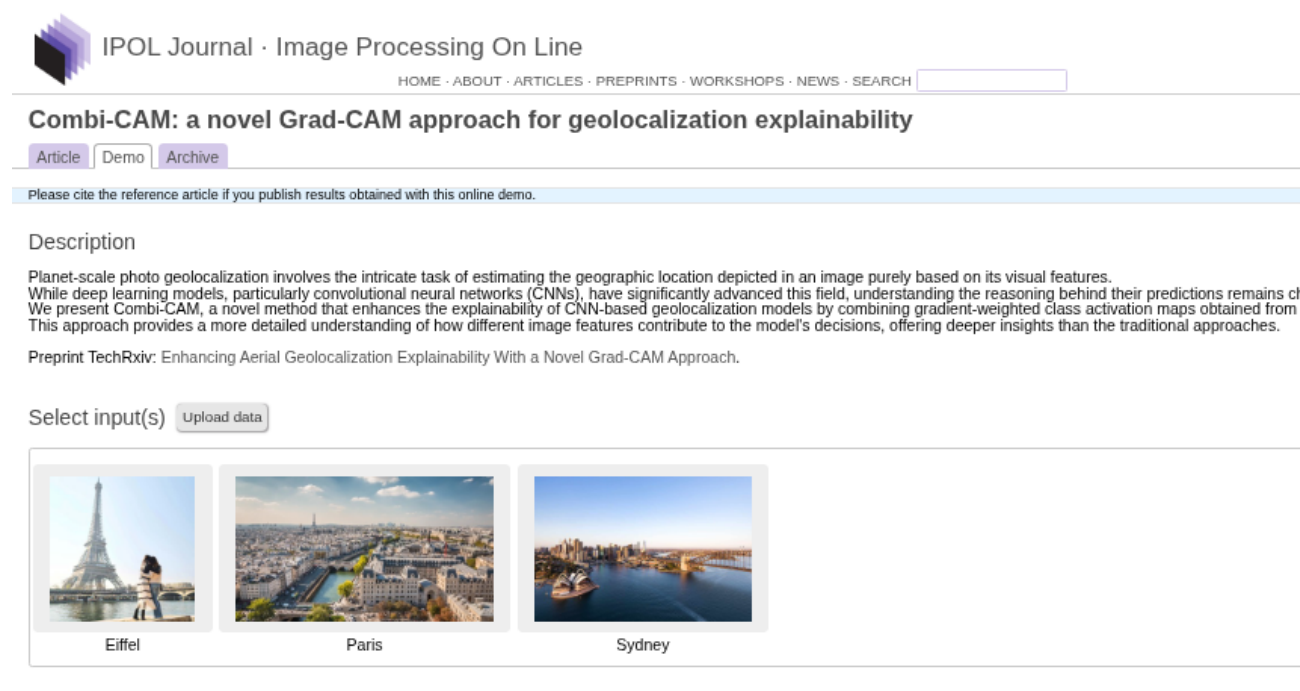


Figure 28 Online demo for Combi-CAM

5.5 Concluding Remarks

In our evaluation of OCR and geolocation tasks, we observed a trade-off between accuracy and practical deployment considerations.

For OCR, we comprehensively evaluated a range of OCR models on two distinct datasets, MemeDataset and HierText, highlighting their strengths and limitations in extracting text from multilingual images and visually diverse document types. While closed-source GPT-4o-mini and Google Vision delivered top accuracy, open-source models such as EasyOCR provided a strong balance between accuracy and speed. This makes EasyOCR more suitable for time-sensitive and cost-effective applications. We also find that the open-sourced large multimodal models struggled with complex text, often generating irrelevant text. Overall, traditional OCR models remain more efficient, while large multimodal models offer potential but need further optimisation for real-world use.

For the task of geolocation, our model demonstrates a strong general ability to predict the correct location. As expected, RAG-based methods outperform our approach, largely due to their access to external knowledge through large language models. However, when compared to other non-RAG-based models, our method outperforms them at several levels of granularity, highlighting its effectiveness within that category.

These findings highlight the need for continued optimization, especially in balancing accuracy, efficiency, and real-world applicability.

6 Detection of Decontextualised Content

Decontextualized content refers to information (textual, visual, or auditory) that has been stripped of or divorced from its original context, leading to distorted or misleading interpretations. A particularly pervasive form of this is multimodal misinformation, where multiple modalities, such as images and accompanying text, are combined in ways that misrepresent the original meaning, source, or timeframe of the content. For example, Figure 29 features a photo of a littered festival ground, misleadingly captioned as having taken place in June 2022 after a Greta Thunberg speech. In reality, the image is from Glastonbury Festival in 2015—years before Thunberg became a public figure¹³. This instance illustrates how altering named entities (e.g., people and dates) can be used to unfairly target individuals or movements. Another example (bottom image) shows a collapsed railway bridge falsely claimed to be from the 2022 Russia-Ukraine conflict in Kursk. While a bridge was indeed damaged in Kursk on that day, the image in question was taken in 2020 in Murmansk, Russia¹⁴.

Decontextualized images pose a significant challenge, particularly given the fact that they are easy to create, requiring no specialized knowledge or software (Aneja et al., 2023) and that visual content is substantially more attention-grabbing and widely shared on social media than plain text, amplifying its potential for rapid dissemination and broader influence (Li et al., 2020). Moreover, when false statements are paired with images, their perceived credibility increases, as visuals tend to convey an implicit sense of authenticity (Newman et al., 2012).

In this task, we focused on detecting multimodal misinformation in decontextualized image-text pairs. We review relevant datasets (manually annotated, weakly labeled, or synthetically generated) as well as detection methods, and techniques for retrieving and filtering external evidence, as well as key challenges in the field, including model generalization, unimodal bias, and the evaluation of system performance.

¹³ <https://www.reuters.com/article/factcheck-glastonbury-greta-idUSL1N2YE1JD>

¹⁴ <https://www.reuters.com/article/factcheck-destroyed-bridge-idUSL2N2WU1CM>

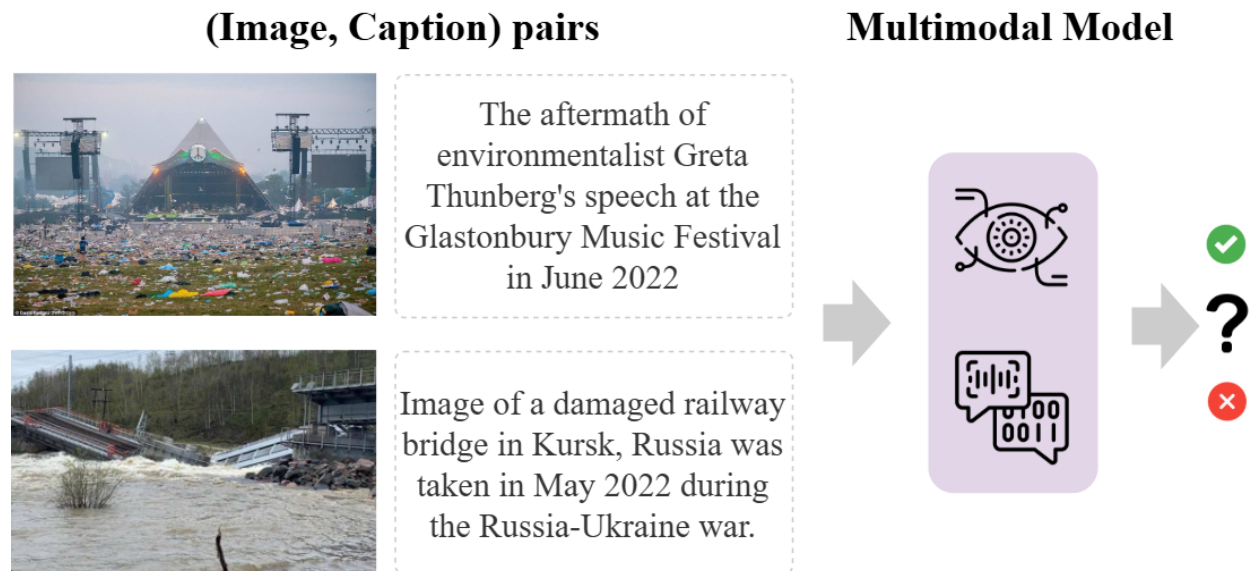

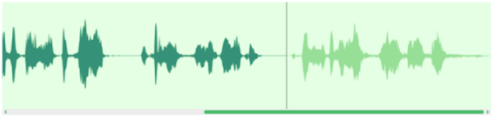


Figure 29 Examples of decontextualized image-caption pairs and a high-level overview of the detection framework. Both examples are sourced from Reuters.com and have been verified as false or misleading

Unlike written text, speech carries not only semantic information but also rich contextual cues—such as speaker identity, emotional tone, and recording conditions. These non-verbal features engage listeners on a more personal and credible level, enhancing both the authenticity and emotional impact of spoken content.


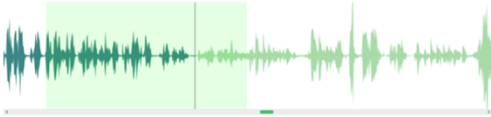
Precisely because of this communicative richness, even minimal edits to audio—such as splicing, adding music, or altering tone—can profoundly influence perception. While such editing techniques are widely used in both professional and user-generated media, they can also be exploited for disinformation. Decontextualization of audio, combined with subtle modifications, can significantly distort meaning while preserving surface-level plausibility (see examples in Figure 30).

Selected File

▶ Play / ⏸ Pause

Detected Source File

▶ Play / ⏸ Pause

Detected Matches

▶ Play Segment 0

Posted message: We cannot win this re-election; we can only re-elect Donald Trump

Real message: We cannot win this re-election [...] if we get engaged in this circular firing squad here — it's got to be a positive campaign [...]

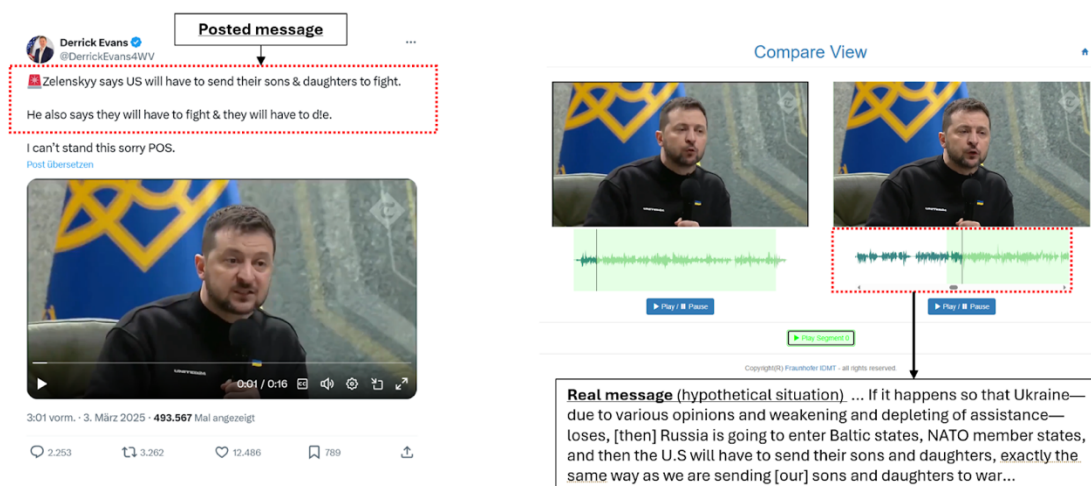


Figure 30 This figure illustrates two cases of audio decontextualization in which an excerpt from a longer speech is isolated and circulated out of context. Without modifying the content, this change in context creates a misleading impression, and its use in disinformation raises serious ethical and societal concerns

In this task, we propose a baseline framework for audio-text decontextualization detection. This framework integrates two key components: a) *Audio provenance analysis*, for detecting reused audio segments and tracing their origins; b) *Context analysis*, for evaluating the logical consistency between audio and its associated text. By combining these two analytical layers, the framework enables robust cross-checking between modalities, not only within a single query but also across source material and provenance chains.

6.1 Background

Multimodal Misinformation Detection (MMD) involves determining whether a combination of modalities, such as an image with accompanying text or an audio clip with its transcript or description, presents truthful or misleading information. This task is typically addressed as a classification problem, where a system must decide whether an image-text pair or text-audio pair is factual or deceptive.

In regards to **image-text** pairs, we discern between two common types of such misinformation: out-of-context (OOC) images, where a real image is paired with unrelated but truthful text taken from another context, and miscaptioned images (MC), where the text has been altered to misrepresent the image's origin or meaning. A notable subtype of miscaptioning is Named-Entity Manipulation (NEM), which involves changing key details like names, dates, or places to mislead viewers. In more advanced settings, external evidence such as related articles or verified images may also be retrieved from the web to help verify claims. This additional information, often gathered through search engine APIs or reverse image search tools, plays a crucial role in evidence-based approaches to misinformation detection.

To build effective image-text detection systems, prior research has explored various training datasets, modality representation techniques, fusion strategies, and model architectures. Existing datasets can be grouped into annotated, weakly annotated, synthetic, and evidence-enhanced types, each with distinct strengths and limitations. Annotated datasets (i.e., Fauxtography (Zlatkova et al., 2019)) provide high-

quality labels from fact-checkers like Snopes and Reuters, but they are typically small and expensive to create, limiting scalability. Weakly annotated datasets (i.e., Twitter MediaEval ([Boididou et al., 2018](#)) and Fakeddit ([Nakamura et al., 2020](#))) use social media metadata or heuristics to scale up data collection, though their labels are often noisier and less reliable. In recent years, there has been an increase in the use of synthetic datasets (i.e., COSMOS ([Aneja et al., 2023](#)), and NewsCLIPpings ([Luo et al., 2021](#))) to simulate misinformation through pairings an image with an out-of-context text (OOC) or by manipulating named entities (NEM), enabling balanced and large-scale training but sometimes lacking real-world complexity. Evidence-enhanced datasets (i.e., NewsCLIPpings+ ([Abdelnabi et al., 2022](#))) incorporate retrieved web content or structured claims to support more realistic, fact-based verification, though they remain limited in number and modality coverage. These dataset types reflect a trade-off between label accuracy, scale, and realism, with hybrid and evidence-based approaches offering promising directions for future development.

In regards to the development and training of detection models for image-text pairs involve three fundamental components: (1) Modality Representation, where text and images are encoded using pre-trained models; (2) Modality Fusion, where representations are integrated to capture cross-modal relationships; and (3) an Optimizable Model, typically a neural network trained to perform the classification task. For modality representation, earlier methods used models like word2vec and VGG, or BERT and ResNet50 for text and visual feature extraction, respectively, but more recent works tend to favor large pre-trained transformers such as CLIP, which supports joint vision-language encoding. Fusion strategies have evolved from simple concatenation of modalities to more advanced techniques such as bilinear pooling, self-attention and co-attention mechanisms, enhancing interaction across modalities. To improve generalization and robustness, models often integrate multi-task learning, contrastive objectives, or leverage external evidence. Recent approaches leverage datasets enhanced with external evidence and attempt to verify internal image-text consistency and external alignment with collected information.

Decontextualization of **audio-text** content pairs often involves both unaltered reuse and content-level manipulations, such as: temporal or spatial cropping, segment reordering and overlaying synthetic speech or unrelated background audio. These manipulations may be designed to enhance narrative coherence, amplify emotional salience, or evade forensic detection tools. Both decontextualization and manipulation are frequently used in disinformation to strengthen misleading narratives.

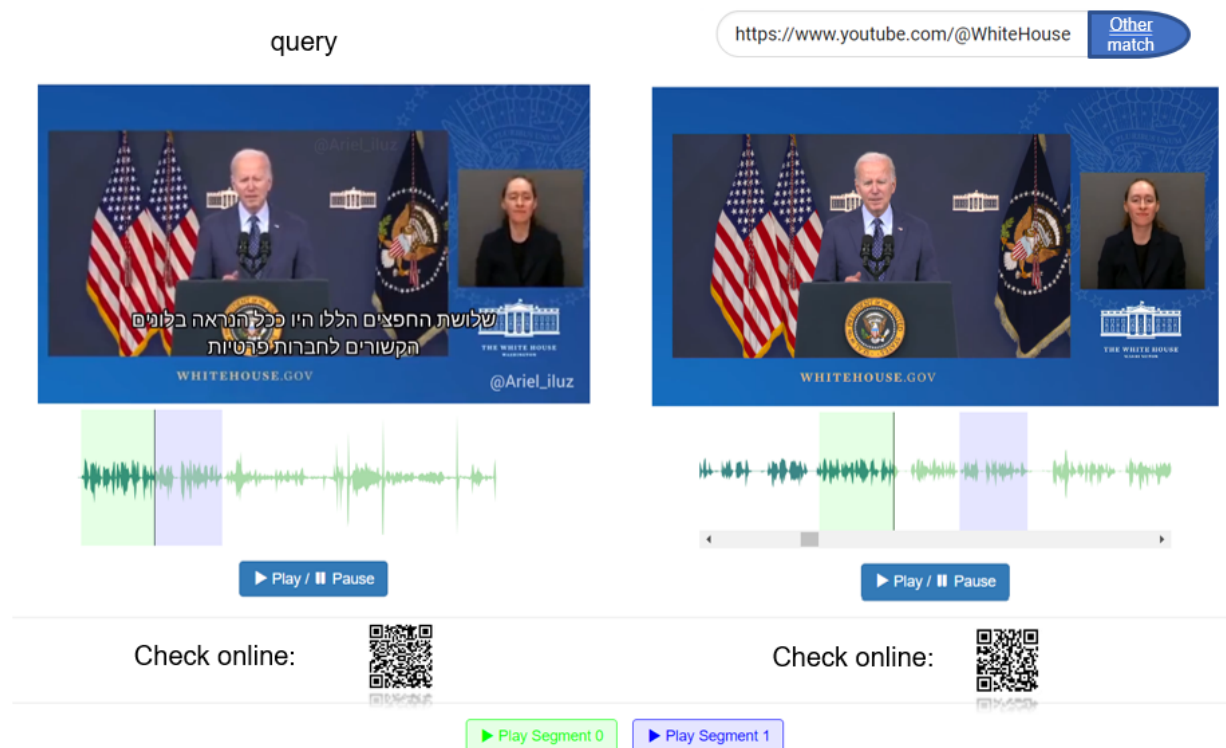


Figure 31 Edited version of an original audio recording (left) disseminated on social media, with the original source (right). Two reused segments are highlighted in green and purple; the omitted segment between them is clearly visible

The real-life cases of audio-text decontextualization (e.g., as in Figure 30 and Figure 31) raise the question: Is it sufficient to treat decontextualization detection as a classification task—i.e., can we judge the alignment of an “audio-text” pair based solely on its internal content? The answer, as with image-text pairs, is only to a limited extent. While it’s possible to detect modality misalignment or named entity inconsistencies (e.g., mismatched time/place), such clear-cut cases are rare. More commonly, decontextualization involves reusing audio content in a new, misleading setting. In these cases, detection requires the search for external evidence.

In image-based scenarios, reverse image search provides crucial external validation. For audio, however, reverse search remains challenging due to the temporal structure of audio, common manipulations (cropping, reordering), platform-dependent transformations (compression, transcoding) and subtle edits (pitch shifting, time-stretching).

To address this, we propose a framework that: a) provides a semi-automatic and effective solution for audio reverse search, and b) provides a contextual alignment score between audio and text, including cross-checking with similar audio-text pairs.

Designing such a framework requires two complementary components: **provenance analysis** and **context analysis**. Provenance analysis focuses on tracing the origin and transformation of all audio files and their segments in an examined set. It identifies whether the audio has been reused, edited, or modified in relation to the identified sources within the examined set. This step is especially important for verifying cases where segments have been selectively extracted or reused, as illustrated in Figure 33. On the other hand, context analysis evaluates whether an audio-text pair is contextually consistent. Even when audio

content remains unaltered, context analysis is vital for flagging instances where it is used misleadingly or presented within a false or deceptive narrative framework.

Used together, these two approaches assess both technical authenticity (via provenance) and semantic plausibility (via context). Neither dimension alone is sufficient. For example, provenance might confirm a clip is authentic, but context analysis could still reveal it's being used deceptively.

Despite growing interest, dedicated audio provenance solutions are lacking. While similarity engines exist, they typically return flat similarity scores, lacking the granularity needed to reconstruct provenance chains or localize reused segments—critical tasks in decontextualization analysis. Our audio provenance approach ([Gerhardt et al., 2024](#)), is the first of its kind for audio and adapts principles from image and video forensics to the specific challenges of the audio domain.

Additionally, there is no standard contextual metric for assessing the alignment between audio and text modalities. Benchmarking work, such as that introducing the M3A dataset ([Xu et al., 2024](#)), has evaluated multimodal misinformation using general-purpose models (e.g., ImageBind from [Girdhar et al., \(2023\)](#)). However, these models are not optimized for audio-text logical alignment.

Our main contributions in this task include the development of a contextual alignment similarity measure specifically designed for evaluating audio-text pairs. In addition, we present a unified framework that integrates this alignment score with the audio provenance analysis previously developed in Work Package 4. This combined approach is tailored to effectively address real-world scenarios involving the decontextualization of audio-text content within the broader context of misinformation detection.

6.2 Methodology

As part of the vera.ai project, we introduced a series of contributions to advance the detection of decontextualized and misleading multimodal content. We began with “Synthetic Misinformers” ([Papadopoulos et al., 2023](#)), the creation and use of synthetic training data for the task, and followed with VERITE ([Papadopoulos et al., 2024](#)), a benchmark designed to mitigate unimodal biases in the task. We then proposed the Relevant Evidence Detection Directed Transformer (RED-DOT) ([Papadopoulos et al., 2025](#)), a framework for filtering and ranking relevant external evidence. In a critical analysis ([Papadopoulos et al., 2025b](#)), we showed how models often exploit spurious dataset artifacts, questioning the validity of current progress in out-of-context detection. To improve reliability in evidence-based systems, Credible, Unreliable or Leaked? ([Chrysidis et al., 2024](#)) focused on verifying the credibility of retrieved evidence and addressing the problem of “leaked” data. Finally, the Latent Multimodal Reconstruction (LAMAR) model ([Papadopoulos et al., 2025c](#)) introduced a reconstruction-based approach using LVLM-generated captions to enhance model generalization and robustness in the task. To address the challenge of detecting audio-text decontextualization, we present a novel framework that combines audio provenance analysis with cross-modal semantic comparison, enabling the identification of both internal inconsistencies and misleading reuse of audio content in new contexts.

6.2.1 Synthetic Misinformers

Our goal was to develop a robust pipeline for training and evaluation of detection models. We observed that many existing approaches, which rely on weakly annotated or synthetically generated training data, often overlook the issue of model generalization, particularly performance on real-world multimodal claims. High accuracy on synthetic benchmarks does not guarantee real-world effectiveness, as these datasets tend to be oversimplified and closely aligned with the training distribution.

To address this, we designed a standardized pipeline (Figure 32) that controls for model architecture and optimization while varying only the data generation strategy, which we term Synthetic Misinformers. These include out-of-context (OOC), named entity manipulation (NEM), and hybrid methods that combine both. The pipeline involves generating synthetic misinformation from truthful image-caption pairs (taken from the VisualNews dataset (Liu et al., 2021)), extract visual and textual features using CLIP (Radford et al., 2021), training a transformer-based classifier (termed DT-Transformer), and evaluating its performance on the COSMOS test set; comprising real-world data collected from [snopes.com](https://www.snopes.com)¹⁵.

Through a comprehensive comparison of existing and novel Synthetic Misinformers, we found that hybrid approaches yielded modest improvements over single-method strategies. However, all models exhibited significant performance drops when tested on real-world data, highlighting the limitations of synthetic training. Surprisingly, unimodal text-based models sometimes outperformed multimodal ones, exposing a bias toward textual cues, particularly in NEM-generated samples. These findings point to the need for more realistic datasets and evaluation frameworks that can reliably measure multimodal reasoning in misinformation detection.

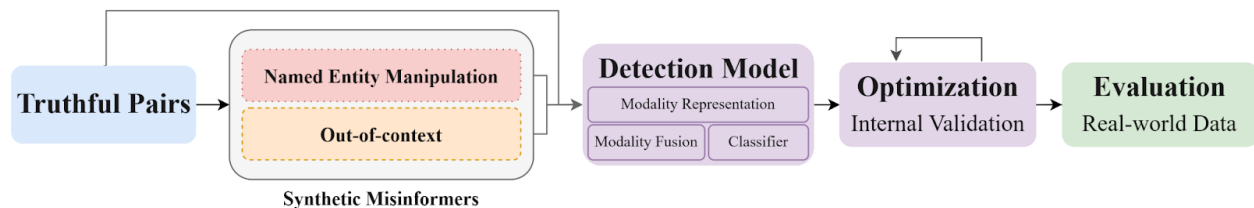


Figure 32 Overview of our proposed workflow: We generate OOC and NEM samples from truthful image-caption pairs using one or more “Synthetic Misinformers”, train a detection model on the synthetic training data, and evaluate it on real-world data.

6.2.2 Addressing Unimodal Biases

In our previous analysis, we observed that unimodal models, leveraging only texts or only images, often performed as well as, or even better than, their multimodal counterparts. For instance, text-only methods achieved strong results on the COSMOS dataset, while image-only methods performed competitively on the Twitter MediaEval benchmark; on a supposedly multimodal task. This raised concerns that these datasets may contain unintended cues or shortcuts that allow models to bypass the need for true multimodal reasoning by relying solely on one modality. This rendered assessment of progress in the field more difficult.

¹⁵ <https://www.snopes.com>

To address the issue of unimodal bias in existing evaluation benchmarks, we developed the VERITE dataset. We hypothesized that a key source of this bias lies in what we define as Asymmetric Multimodal Misinformation (AMM), cases in which the misleading signal is present in only one modality (either image or text), while the other serves merely as a decorative, illustrative, or redundant element. Our analysis estimated that nearly 50% of examples in the COSMOS dataset exhibit this asymmetry, skewing model behavior and inflating performance metrics by allowing models to succeed without truly leveraging cross-modal understanding. Instead, we hypothesized that genuine multimodal misinformation requires a strong semantic interplay between image and text, where neither modality alone suffices to detect the falsehood. Illustrative examples of both multimodal and asymmetric multimodal misinformation are shown in Figure 33. For instance, a claim such as “deceased people turning up to vote” is accompanied by an image that is only tangentially related and was originally included for aesthetic purposes, not as evidentiary support.

For the construction of VERITE, we sourced data from reputable fact-checking organizations, namely Snopes and Reuters, and only included genuinely multimodal misinformation, where the misleading narrative emerges from the combination of image and text, and cannot be identified by analyzing either modality in isolation. VERITE comprises three balanced categories: True, Out-of-Context (OOC), and Miscaptioned (MC). Furthermore, we employed modality balancing, whereby each image is paired with both a truthful and a misleading caption, and each truthful caption is also associated with an out-of-context image. OOC pairs were annotated by retrieving visually similar but contextually divergent images. Through extensive experimentation, we found that multimodal models consistently outperformed their unimodal counterparts on VERITE, supporting our hypothesis that removing AMM and employing modality balancing are effective strategies to mitigate unimodal bias and promote true multimodal reasoning.

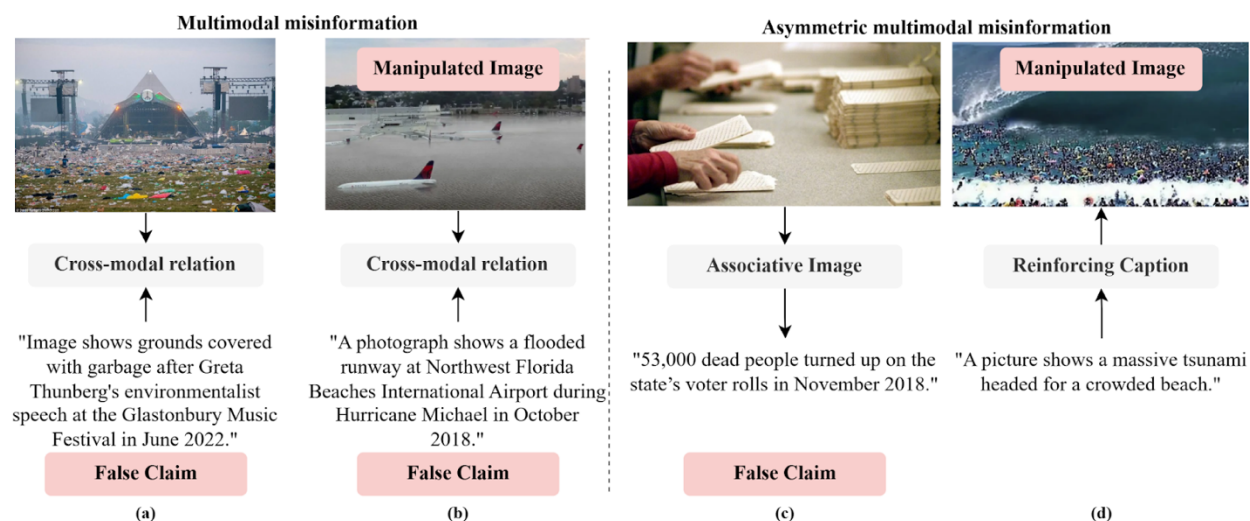


Figure 33 Examples of multimodal misinformation from VERITE and asymmetric multimodal cases from the COSMOS test set

6.2.3 Relevant Evidence Detection

After defining the overall pipeline and developing a more robust evaluation dataset, we extended our work on the evidence-based detection of decontextualized images. Building on recent advances in the

field, we leveraged the synthetic NewsCLIPPings dataset, enhanced with external information sources (Abdelnabi et al., 2022). In this setup, the input text is used to retrieve related images from the web, followed by reverse image search to collect associated textual evidence, such as article titles and image captions. Although prior work has shown promising results, it typically assumes that all retrieved content is relevant. This assumption often breaks down in real-world settings, where retrieved evidence frequently includes irrelevant or misleading information. For instance, Figure 34 shows an image from an anti-surveillance protest in Hong Kong being falsely presented as evidence of public resistance to 5G and COVID-19 conspiracies. We observe that among the retrieved textual evidence, only the top-ranked item is truly relevant, while others refer to unrelated events; underscoring the need for effective relevance filtering in misinformation detection.

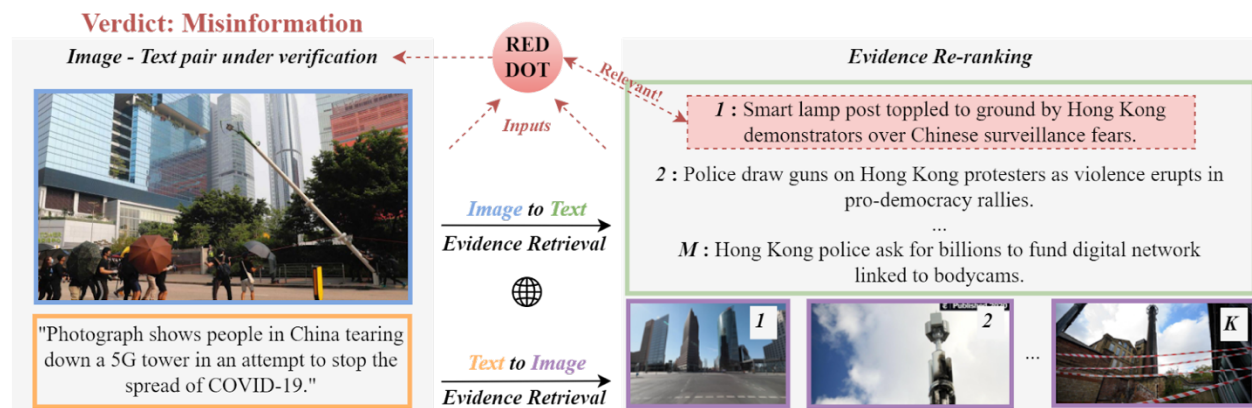


Figure 34 An image-text claim is assessed using external information, including both images and text, collected from the web. The system retrieves and re-ranks this information, while RED-DOT identifies the most relevant items to evaluate the claim's validity

To address this limitation, we introduced Relevant Evidence Detection (RED), a framework that not only evaluates stance or semantic consistency between claims and evidence, but also filters out irrelevant content. Our approach incorporates a re-ranking step based on intra-modal similarity, and includes hard negative samples to train models that can more effectively distinguish true evidence from distractors. Moreover, we introduced RED-DOT (Relevant Evidence Detection Directed Transformer), a multi-task learning model designed to jointly predict evidence relevance and misinformation labels. RED-DOT was trained on the synthetic samples from NewsCLIPPings+ and evaluated on real-world examples from the VERITE benchmark. Our results showed that using a smaller set of well-ranked, highly relevant evidence significantly improves detection accuracy. In contrast, adding more items introduced reduced performance, by letting noisy and loosely related content into the model; highlighting the critical role of evidence filtering in achieving both accuracy and efficiency.

6.2.4 Illusory Progress due to Dataset-specific Biases

To evaluate progress in decontextualized image-text detection and uncover potential shortcuts in existing benchmarks, we introduced MUSE (MULTimodal SimilaritiEs), a simple yet effective baseline that uses CLIP-based similarity scores between a claim and its top-ranked external evidence. MUSE computes a range of intra-modal (e.g., text-text evidence, image-image evidence) and cross-modal similarities (e.g., image-text evidence, text-image evidence) and feeds them into lightweight classifiers like decision trees or MLPs. Despite its simplicity, MUSE achieved accuracy on par with or better than complex architectures

on both NewsCLippings+ and VERITE, all while using less than 1% of the computational resources. Building on this, we proposed AITR (Attentive Intermediate Transformer Representations), a Transformer-based model that captures decision signals from intermediate attention layers. When augmented with MUSE features, AITR's performance improved significantly, highlighting the utility of similarity-based features even within more expressive architectures.

However, MUSE's strong performance also highlights key limitations in current OOC datasets. In NewsCLippings+, out-of-context examples are created by mismatching image-text pairs originally sourced from reputable outlets such as BBC or The Guardian. Consequently, external evidence retrieved from the web often includes the original article or closely related content, making simple similarity scores highly predictive. As illustrated in Figure 35, truthful pairs consistently yield higher similarity scores with their retrieved evidence, an artifact of retrieval surfacing near-identical information across modalities. While effective, this shortcut limits the ability to assess whether models truly reason about factual consistency, and raises concerns about their robustness in detecting novel or emerging misinformation.

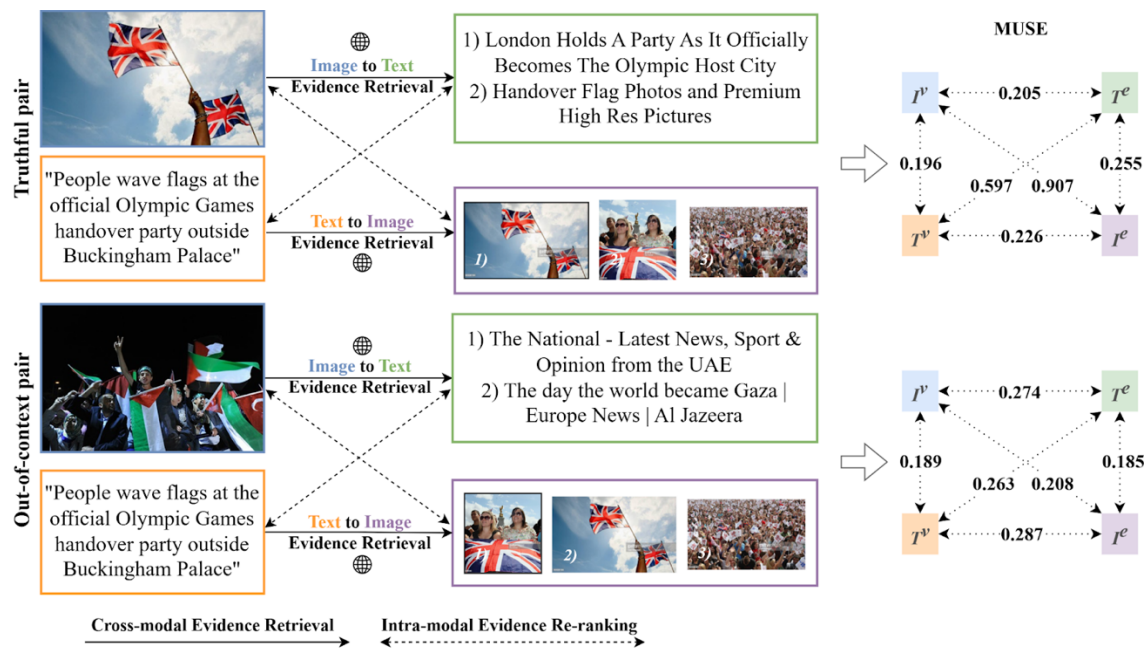


Figure 35 Two examples from the NewsCLippings dataset, one truthful (top) and one out-of-context (bottom) alongside their retrieved and re-ranked external evidence, along with computed multimodal similarity scores

6.2.5 Evidence Verification and Filtering

In addition to analyzing relevance and dataset biases, we addressed two more critical challenges in evidence-based detection, namely, credibility and leaked evidence. Effective detection depends not only on retrieving relevant content but also on ensuring that evidence comes from trustworthy sources and is not already fact-checked. Leaked evidence refers to instances where content from fact-checking articles appears in training data, giving models access to pre-verified conclusions (Glockner et al., 2022). This creates an unrealistic training scenario where models learn to rely on high-quality fact-checked

information that is unavailable when detecting new, unverified misinformation. As a result, they may perform seemingly well on benchmarks but struggle with real-world generalization.

To tackle this, we introduced a supervised evidence verification pipeline, beginning with the creation of CREDULE (CREDible, Unreliable, or LEaked), a dataset of 91,632 samples evenly split across the three classes. CREDULE was constructed by modifying and merging several established datasets (MultiFC, PUBHEALTH, Politifact, NELA-GT, Fake News Corpus, and Getting Real About Fake News), and includes both short and long-form content. We then trained EVVER-Net (EVIDence-VERification Network), a neural classifier that distinguishes between credible, unreliable, and leaked evidence, and experimented with various Transformer backbone encoders (DeBERTa, CLIP, T5, LongT5, and Longformer). Moreover, we experimented with incorporating domain credibility scores from Media Bias/Fact Check¹⁶ (MBFC).

We also applied EVVER-Net to widely used fact-checking datasets, including LIAR-PLUS, FACTIFY, and MOCHEG, and found high rates of leaked evidence. For instance, 98.5% of LIAR-PLUS and 83.6% of MOCHEG samples contained leaked content. While we did not deploy the full pipeline, Figure 36 illustrates a potential end-to-end framework with EVVER-Net serving as a filtering step to identify and remove unreliable or leaked evidence, thus ensuring that only credible and non-leaked sources are used during model training.

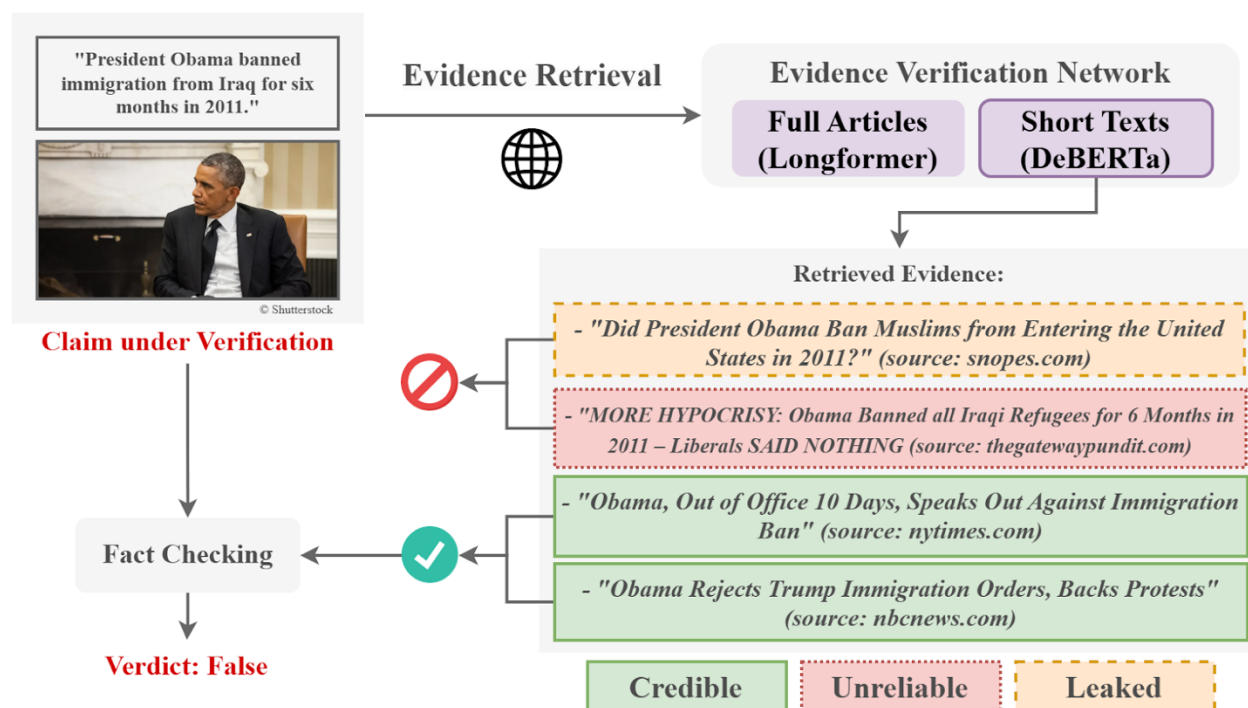


Figure 36 High-level overview of the proposed automated fact-checking pipeline, using Evidence Verification Network to identify unreliable and leaked information, thus ensuring credible evidence for accurate verification

6.2.6 Latent Multimodal Reconstruction

Recent work on miscaptioned images has primarily relied on synthetic training datasets based on either named-entity manipulations or out-of-context pairings. However, these approaches often produce

¹⁶ <https://mediabiasfactcheck.com>

formulaic and overly simplistic examples that fail to reflect the complexity of real-world misinformation. To overcome this limitation, we developed the "MisCaption This!" dataset using the Large Vision-Language Model LLaVA 7B. Instead of rule-based entity swaps, we applied an adversarial prompt selection strategy to guide LLaVA in generating more diverse, naturalistic false captions that subtly misrepresent the accompanying images. This method better captures the nuanced nature of real-world misinformation and provides a richer training signal for multimodal detection models.

In addition, we proposed LAMAR (Latent Multimodal Reconstruction), a Transformer-based architecture that incorporates a reconstruction task inspired by human fact-checking practices. LAMAR is trained to recover the embedding of the original, truthful caption using both the image and the manipulated caption as input. The reconstructed embedding is then integrated with the fused modalities through gating, masking, or self-attention mechanisms. This auxiliary signal enhances the model's ability to capture subtle inconsistencies and improves its overall robustness.

6.2.7 Framework for Audio-Text Decontextualization Detection

The workflow of audio-text decontextualization detection begins with a single "query" audio-text pair, where the primary objective is to assess whether the audio and text modalities are logically and semantically aligned. As discussed in the background section, decontextualization is most effectively detected not by analyzing the query pair in isolation, but by cross-comparing it with the original source(s) of the audio content. This comparison allows us to verify whether the audio has been used consistently with its original context or misrepresented in a new setting.

To support this goal, our framework integrates two core components: *Audio provenance analysis* – to identify the source(s) of the audio in the query pair and *Context analysis* – to evaluate the logical alignment between audio and text, using a robust cross-modal comparison.

Together, these components enable the identification of reused audio content, the verification of whether the audio is used consistently with its original context, and the detection of cross-modal inconsistencies in real-world posts.

Audio Provenance Analysis

Audio provenance tools operate over sets of audio files, supporting applications like content verification and similarity indication for coordinated behavior detection. Developed in Work Package 4 (see Deliverables 4.1 and 4.2), it consists of two main components: Audio Reuse Detection and Audio Phylogeny Analysis. Background and implementation details are documented in Deliverable 4.1, while the full provenance framework is described in Deliverable 4.2.

Context Analysis

The goal of context analysis in the audio-text decontextualization framework is to determine whether the semantic content of an audio signal aligns with the textual message it is paired with. Unlike audio provenance, which focuses on tracing the origin and reuse of audio content, context analysis deals with the logical consistency and semantic fidelity between modalities.

Our objective is to develop a processing framework that can reliably assess whether the audio and text accompanying a media post are logically and contextually aligned. To this end, we employ the M3A

dataset (Xu et al., 2024), the only publicly available dataset containing audio-text pairs for misinformation detection. The dataset serves both for tuning our system and benchmarking it against the approach used in the original M3A study.

The core processing steps of our audio-text context analysis framework are presented below and illustrated in Figure 37.

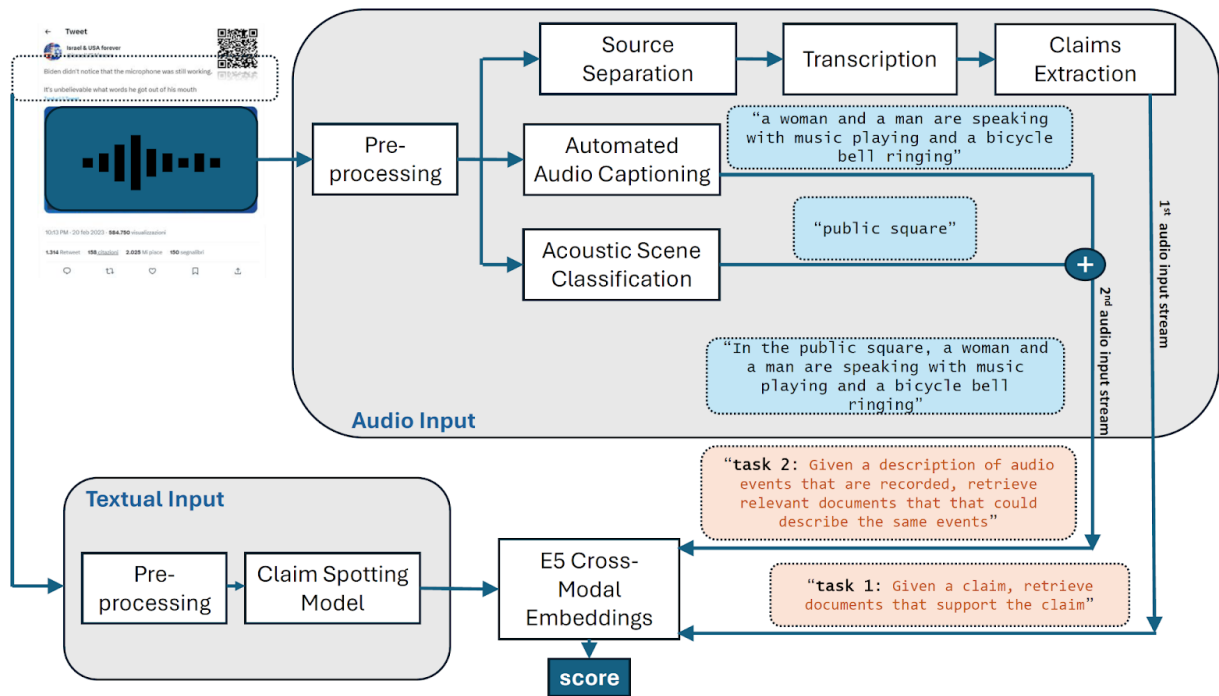


Figure 37 Workflow of audio-text context analysis framework

Inputs to the cross-modal comparison

The M3A dataset comprises data from 60 news outlets. For framework tuning, we selected *apnews*, which includes 679 videos. After removing files containing only digital silence, we retained 644 usable audio files.

The first processing branch applies Automatic Speech Recognition (ASR). We evaluated multiple Whisper-based models, including Whisper Medium (Radford et al., 2023), WhisperX (Bain et al., 2023), and Faster-Whisper. Initial assessments revealed that Whisper Medium struggled in acoustically complex environments, failing to distinguish foreground speech from background noise. To address this, we employed the Demucs source separation model (Rouard et al., 2023) to isolate speech signals before ASR. This preprocessing step enhanced the quality of the input signals, resulting in more accurate downstream transcription.

Following the application of Demucs, WhisperX and Fast-Whisper models were used for transcription of the vocal part of the audio signal. Among these, WhisperX delivered superior performance, characterized by precise word-level alignment, improved handling of overlapping speech, retrieving no speech information and accurate speaker segmentation. Furthermore, WhisperX demonstrated computational efficiency on mid-range GPUs, supporting scalability across varied hardware configurations. Faster-

Whisper, while faster in inference, exhibited slightly reduced transcription quality in scenarios involving multiple speakers or noisy backgrounds. Consequently, WhisperX was selected as the preferred transcription model for subsequent analysis due to its balance of accuracy and efficiency.

The transcriptions produced by WhisperX served as input for a claim extraction task, aimed at identifying news-worthy content. Several large language models (LLMs) were evaluated: Mistral ([Jiang et al., 2023](#)), Qwen ([Qwen Team, 2024](#)), Gemma ([Mesnard et al., 2024](#)), Claude, and fine-tuned variants of Claude and Gemma. A standardized prompt was employed across all models to maintain consistency in input conditions and to enable the detection of hallucinated content.

Qwen and Gemma stood out by generating semantically accurate claims and reliably identifying when no newsworthy content was present (e.g., stating “No news-worthy claims present in this document”). Notably, Qwen produced fully decontextualized claims, eliminating the need for post-processing e.g., by using T5 ([Raffel et al., 2020](#)) for coreference resolution.

We evaluated the quality of extracted claims using entailment-based minicheck scores, calculated with the Bespoke-Minicheck-7B model ([Tang et al., 2024](#)), comparing claims to ground-truth news outlines. The WhisperX-Qwen pipeline yielded the highest entailment scores (see Figure 38).

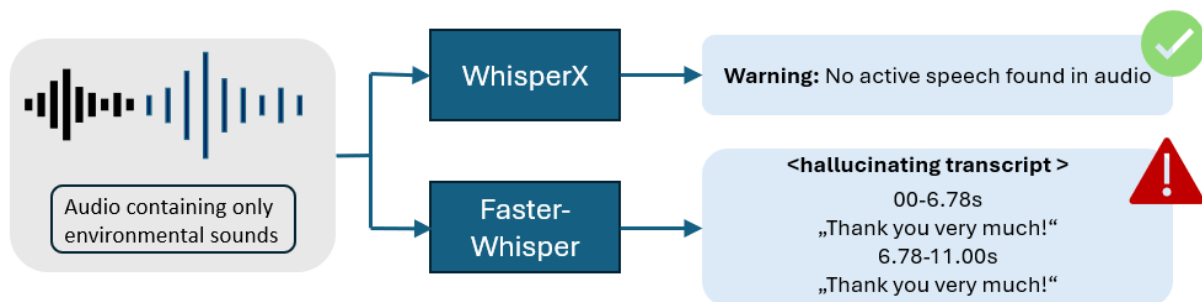


Figure 38 Example of problematic output from WhisperX and Faster-Whisper when transcribing audio containing only environmental sounds

Based on these results, Qwen was chosen for claim extraction from audio transcripts. These claims form one of the audio input streams to the cross-modal comparison module, as shown in Figure 39.

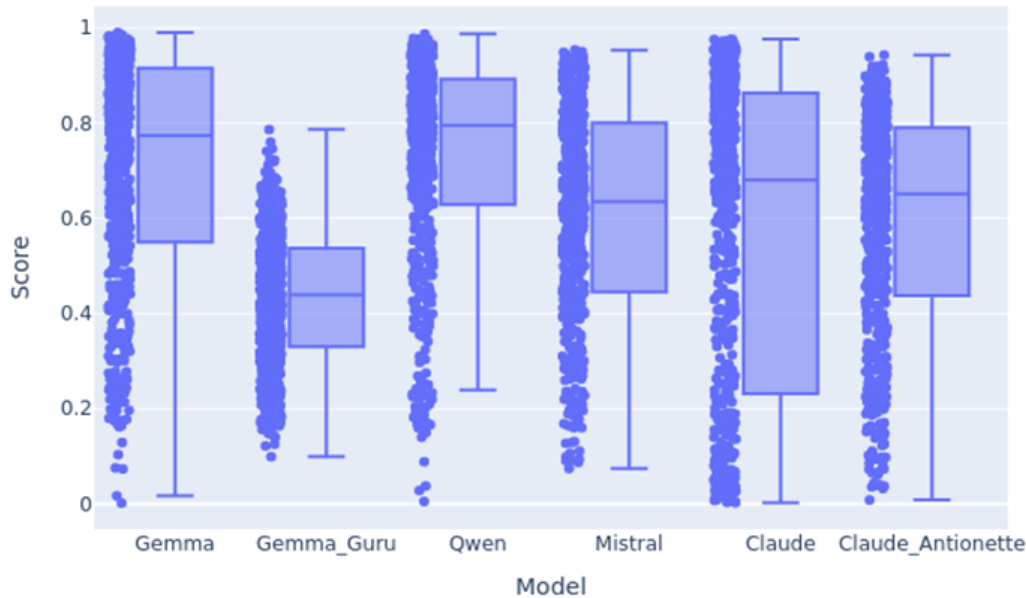


Figure 39 Entailment Score distribution using the Minicheck Model with various LLMs

In parallel, the audio signals are processed using an audio captioning model trained on WavCaps ([Mei et al., 2024](#)), generating a general sentence describing environmental sounds. We enrich this caption with information from an audio scene classification model (see Section 4.2.3), yielding a second audio input stream.

The textual description from the media post serves as the third input. After standard preprocessing (e.g., removing newlines and splitting into sentences), we apply the ClaimBuster model ([Hassan et al., 2017](#)), which categorizes sentences as Non-Factual, Unimportant Factual, or Check-worthy Factual. Only the latter two are retained for comparison.

Cross-Modal Comparison

To compute the similarity between the text and audio modalities, it is necessary to convert each input into a shared embedding space. That is, we need to compute vector representations of the semantics expressed by each modality such that we can measure the similarity or distance between them.

We use a family of state-of-the-art LLM embedding models, E5 ([Wang et al., 2022](#)), to compute vector representations of the audio and text inputs. Due to the way in which E5 models are trained, these vector representations lie in a shared space that represents the semantics of the input. We can compute the cosine similarity between these vector representations to obtain a measure of semantic similarity, which can be used to identify whether the audio and text inputs are aligned.

In theory, the cosine similarity lies in the range $[-1, 1]$. In practice, however, the similarity scores computed from E5 embeddings lie in a narrower range, approximately $[0.7, 1]$. In order to increase the interpretability of the similarity scores and emphasize the differences between very similar and very dissimilar inputs, we rescale the cosine similarity using an empirical lower bound. Specifically, we obtain 100,000 pairs of random sentences from the Common Crawl News dataset ([Mackenzie et al., 2020](#)) and compute the average cosine similarity between these pairs using each E5 model. Because these sentences

are on average completely dissimilar, the average similarity is a reliable lower bound. Across models, we computed an empirical lower bound of approximately 0.74 for all E5 variants except for e5-mistral-7b-instruct which has a lower bound of approximately 0.51. We compute the rescaled similarity score s^* from the raw similarity score s and the empirical lower bound l using the following equation.

$$s^* = \frac{s - l}{1 - l}$$

For each audio input stream, we defined a specific task (see Figure 37). These tasks, along with the extracted claims or audio event descriptions, constitute the audio-side inputs. On the other side, we use the processed textual descriptions of the posts. The E5 model is then used to compute vector representations for all inputs, and the final similarity score is obtained by averaging the cosine similarities between each audio input stream and the corresponding textual embeddings.

Once we define the context alignment measure, the next critical step is to apply it in cross-checking the alignment between the modalities of the query audio-text pair and those of the source(s) identified through audio provenance analysis. The underlying assumption, as outlined in the background section, is that the most effective way to detect decontextualization is not by evaluating the query pair in isolation, but by comparing it to the original context from which the audio was potentially reused. This allows us to detect not only logical misalignment within the query pair but also inconsistencies that emerge when the same audio content is repurposed in a misleading textual setting.

A walkthrough of this framework on a real-world example, along with fine-tuning and benchmarking experiments, is presented in the following evaluation section.

6.3 Evaluation

We describe our evaluation methodologies and results for image-text and audio-text decontextualization tasks. For the image-text task, we report large relative improvements over previous work on two datasets. For audio-text decontextualization, we show that our method outperforms previous work by nearly 200% on M3A.

6.3.1 Image-text Decontextualization

In Table 25, we compare our methods, both with and without external evidence, against the most relevant approaches developed since the start of the project. Specifically, we evaluate against fine-tuned CLIP ([Luo et al., 2021](#)), Self-Supervised Distilled Learning (SSDL) ([Mu et al., 2023](#)), the Consistency Checking Network (CCN) ([Abdelnabi et al., 2022](#)), the Stance Extraction Network (SEN) ([Yuan et al., 2023](#)), the Explainable and Context-Enhanced Network (ECENet) ([Zhang et al., 2023](#)), and SNIFFER ([Qi et al., 2024](#)).

Our models consistently outperform prior methods, with the most recent method, AITR, achieving the highest performance on both NewsCLIPPings and VERITE; both with and without using external evidence. These results represent relative improvements of 39.7% without evidence and 10.2% with evidence on the NewsCLIPPings dataset, compared to methods available at the start of the project.

However, as discussed earlier, much of this apparent progress is influenced by dataset-specific artifacts and shortcuts, rather than true reasoning over the factuality of image-text pairs. This limitation becomes clear when evaluating on the "True vs. Mismatched" subset of VERITE, which requires identifying factual inconsistencies (such as altered locations, dates, or people) between images and text. Here, performance drops to 51.2%, underscoring the difficulty of the task when shortcuts are unavailable, as in OOC pairs. These findings highlight the need for further research, particularly in the design of evaluation benchmarks and the collection of high-quality external evidence.

Table 25 Comparative analysis of models with and without external evidence for detecting out-of-context (OOC) image-text pairs, trained on the NewsCLIPpings dataset and evaluated on both its test set and the VERITE benchmark

Model	External Evidence	NewsCLIPpings	VERITE (True vs OOC)
Fine-tuned CLIP (Luo et al., 2021)	NO	60.2	-
SSDL (Mu et al., 2023)		71.0	-
DT-Transformer (Ours)		77.1	69.4
RED-DOT (Ours)		81.5	73.5
MUSE (Ours)		80.7	70.9
AITR (Ours)		84.1	74.1
LAMAR (Ours)		84.8	76.3
CCN (Abdelnabi et al., 2022)	YES	84.7	-
SEN (Yuan et al., 2023)		87.1	-
ECENet (Zhang et al., 2023)		87.7	-
SNIFER (Qi et al., 2024)		88.4	74.0
RED-DOT (ours)		90.3	76.9
MUSE (ours)		90.0	80.6
AITR (ours)		93.3	82.7

6.3.2 Audio-text Decontextualization

Preliminary Experiments for Framework Calibration

As a proof-of-concept, we conducted experiments on the 644 AP News samples from the M3A dataset. We selected the Named Entity Manipulation (NEM) subset, where the audio remains unchanged, but textual descriptions are modified by swapping named entities (Persons, Organizations, Locations, or All).

We then compute the rescaled cosine similarity between the E5 embeddings of Qwen-extracted claims and both original and manipulated text snippets. This score is used to predict which text snippet is paired to each claim. We report prediction accuracy for different E5 variants (see Table 26).

Table 26 Accuracy of each E5 model variant, ordered from smallest to largest, in predicting the most relevant text content from the claim extracted from the audio transcription by Qwen 2.5 7b ([Qwen Team, 2024](#))

E5 Model	Original Text	Manipulated			
		Organization	Location	Person	All
e5-small-v2	0.670	0.682	0.592	0.508	0.487
e5-base-v2	0.692	0.656	0.614	0.538	0.524
e5-large-v2	0.714	0.726	0.655	0.580	0.528
multilingual-e5-large	0.670	0.662	0.614	0.527	0.502
multilingual-e5-large-instruct	0.710	0.726	0.640	0.592	0.532
e5-mistral-7b-instruct	0.702	0.739	0.640	0.557	0.537

Overall, the results exhibit the trend we expect: the models obtain a higher accuracy using the original text content vs. the manipulated content, and the content with all three named-entity types replaced consistently obtains the lowest accuracy. However, we notice that for four out of six E5 models replacing only the Organization named-entities results in a slightly *higher* accuracy. To investigate this, Table 27 reports the average similarity score computed by each E5 model between the claims and the predicted texts. We see that when it comes to raw scores, the manipulated texts are in every case lower than the original texts, as we expect.

Table 27 Average similarity score between claims extracted from the audio transcription by Qwen 2.5 7b and original/manipulated texts across E5 models

E5 Model	Original Text	Manipulated			
		Organization	Location	Person	Complete
intfloat/e5-small-v2	0.414	0.370	0.367	0.344	0.341
intfloat/e5-base-v2	0.306	0.277	0.275	0.238	0.240
intfloat/e5-large-v2	0.301	0.250	0.250	0.228	0.220
intfloat/multilingual-e5-large	0.414	0.379	0.377	0.350	0.348
intfloat/multilingual-e5-large-instruct	0.447	0.387	0.380	0.358	0.339
intfloat/e5-mistral-7b-instruct	0.472	0.413	0.418	0.394	0.365

As a further investigation, we conducted a manual examination of the original and organization-manipulated texts as well as the extracted claims. We found that the organization named-entity is often the news outlet or communicating body, and this is rarely included in the claim extracted from the audio transcript. It may be that the difference in accuracy is a non-significant side-effect of the variations in semantics computed by E5. A larger evaluation sample size and significance tests are necessary to determine whether this is indeed the case.

Preliminary experiments demonstrated the effectiveness of this approach, particularly in its ability to distinguish between original and manipulated text samples—validating the discriminative power of our method. Encouraged by these results, we selected the multilingual-e5-large-instruct model as the best-performing variant due to its strong balance between accuracy and generalization across languages and contexts. We then proceeded to design comparative experiments to benchmark our approach against the state-of-the-art ImageBind model, aiming to further assess the robustness and utility of our framework in detecting audio-text decontextualization.

Comparative experiments against benchmark

To our knowledge, no existing method specifically targets audio-text decontextualization. The authors of M3A, however, evaluated this task using ImageBind ([Girdhar et al., 2023](#)), a general purpose multimodal model.

We compared our framework (Figure 37) to ImageBind using 25,896 original posts and 87,017 manipulated counterparts (via named entity swapping). We measured accuracy, precision, and recall at Equal Error Rate (EER) points. Two variants of our system were tested: a) Using only claims from transcripts, b) Full system using both audio input streams. The results are shown in Table 28.

Table 28 Average accuracy, precision and recall for ImageBind and two variations of our proposed method

Method	Accuracy	Precision	Recall
ImageBind (Girdhar et al., 2023)	0.385	0.385	0.384
Ours (claims only)	0.642	0.642	0.642
Ours complete	0.746	0.745	0.745

The low performance of ImageBind highlights the difficulty of addressing such a fine-grained task as audio-text decontextualization using general-purpose models not explicitly tuned for this kind of semantic precision. Despite no task-specific training, our method substantially outperformed ImageBind. The inclusion of both audio input streams (claims + audio scene) further improved performance, underscoring the benefit of integrating environmental cues.

These promising results suggest strong potential for future research. Current efforts include experiments across the full M3A dataset and testing on another manipulation type, Multimodal Misalignment, which will complete our planned evaluation.

It is worth noting that current limitations in available datasets —particularly in terms of size, annotation quality, and language coverage— prevent effective task-specific training for this problem. This remains an open challenge, and part of our future work involves addressing these data constraints to enable fine-tuning and broader evaluation of our model. As noted earlier, many original-manipulated pairs within the M3A dataset lack sufficient audio context to conclusively determine alignment. Future work will involve labeling samples as aligned, misaligned, or undetermined, enabling the training of supervised classifiers for decontextualization detection.

Framework walkthrough

After defining a cross-modality context alignment measure capable of quantifying the semantic agreement between audio and text, we pair this methodology with audio provenance analysis (outlined earlier) to address real-world cases of audio-text decontextualization detection.

The process begins with an audio-text query pair whose contextual alignment needs to be verified. The first step typically involves dataset collection, often conducted by investigative journalists. This can be achieved through keyword-based searches over reliable media sources or by aggregating all posts published within a specified time window from targeted accounts or news outlets.

These large and often manually non-searchable datasets serve as input to the audio provenance analysis module, which helps surface potential source audio segments related to the query. Then, leveraging the context alignment functionality, the system computes a similarity score—or a threshold-based verdict—between the source audio modality and the query text modality. This end-to-end flow enables a scalable approach to verify the contextual integrity of audio snippets circulating in online media.

6.4 Implementation and Integration

Although we achieved notable improvements in detecting decontextualized images and met several of our key performance indicators, we ultimately decided not to deploy a tool or service for this specific task. The decision reflects both the current limitations of the technology and broader challenges in the field.

Despite strong model performance on controlled tasks, such as 80% accuracy on the “True vs. Out-of-Context” subset of the VERITE benchmark, one should be cautious of dataset-specific shortcuts and a narrow definition of OOC, where images are mismatched with otherwise truthful captions originally associated with different images. Instead, these models struggled to generalize to more complex cases, such as miscaptioned images, where the accompanying text introduces factual errors (e.g., incorrect dates or locations). Further compounding the issue, our models were primarily trained on curated data from news sources like The Guardian and BBC, featuring concise, structured captions. These do not reflect the style, noise, or variability of social media posts, where much of real-world misinformation now spreads. The main limitation falls at the absence of any large-scale, up-to-date dataset of out-of-context image posts from platforms like Twitter, Facebook, or Instagram to train and evaluate our methods against.

Given these limitations in model robustness, dataset coverage, and evidence reliability, we concluded that deploying such a tool would not offer meaningful or trustworthy support for end users. At this stage, the models are neither reliable enough for the general public nor sufficiently accurate or interpretable to assist professional fact-checkers in a production setting.

Concerning the audio-text decontextualization framework, the audio provenance analysis component has been integrated into the Verification Plugin. Looking forward, once our context alignment framework undergoes further fine-tuning and task-specific training (which is currently limited due to dataset constraints as discussed), we plan to integrate the contextual alignment functionality within the provenance analysis workflow. This will offer a seamless, end-to-end solution for detecting audio-text mismatches in misinformation scenarios.

6.5 Concluding Remarks

Our exploration of decontextualised content detection reveals the growing importance of context-aware and multimodal strategies in combating misinformation. While unimodal models—relying solely on image or text—can detect blatant inconsistencies, they often fall short when content manipulations are subtle or context loss is implicit. By incorporating external evidence and leveraging techniques like Relevant Evidence Detection (RED-DOT) and latent multimodal reconstruction, our models achieved higher accuracy and stronger generalization across diverse datasets.

Synthetic training data, generated using purpose-built "Synthetic Misinformers", played a key role in improving robustness. However, we also observed that such synthetic samples may introduce biases not always reflective of real-world data. This highlights the need for carefully balanced training strategies, particularly when deploying models in production environments.

Our audio-text decontextualization framework introduces a novel approach to detecting misleading audio repurposing, an emerging challenge in the age of AI-generated content. Though still maturing, this cross-modal method demonstrates significant potential for enhancing contextual verification.

The results point to a clear direction: tackling decontextualisation effectively requires moving beyond single-modality pipelines and embracing evidence-rich, explainable approaches that reflect the complexity of real-world disinformation.

7 Fact-checker-in-the-loop Approach

In this section, we explore the common architectures and challenges involved in using conversational agents to facilitate the user experience of verification tools. Specifically, we examine tasks such as user intent detection, dialogue state tracking, and retrieval-augmented generation (RAG) for enhancing the explainability of AI tools and collecting user feedback. A key enhancement to this framework is the incorporation of a Fact-checker-in-the-loop approach, which embeds real-time fact verification capabilities into the agent’s dialogue flow. This allows the system not only to generate informative responses but also to validate the factual accuracy of those responses using structured knowledge bases or external retrieval systems. By integrating fact-checking into the user interaction loop, conversational agents can improve reliability, increase user trust, and provide more transparent explanations for the behavior of complex AI models. Additionally, we discuss the role of Large Language Models in supporting these processes, particularly in the extraction and summarization of relevant information from document collections. Moreover, this section evaluates the effectiveness of various metrics for assessing the quality and factual consistency of retrieval-augmented replies within such systems. Lastly, the integration of the synthetic image chatbot showcases how conversational agents, enhanced by fact-checking and RAG techniques, can support user-centered interaction around visual and textual information.

7.1 Background

Driven by recent advances in large language models, conversational agents have become powerful tools for interacting with users, collecting feedback, and communicating AI decisions. This section provides an overview of key components underpinning these capabilities, including the design of feedback-oriented agents (7.1.1), techniques for detecting user intent (7.1.2), and retrieval-augmented generation methods for explainable responses (7.1.3).

7.1.1 The Role of Conversational Agents in Improving User Experience and Collecting Feedback

Powered by recent advances in artificial intelligence (AI) and natural language processing (NLP), conversational agents—also known as chatbots or virtual assistants—have become widely used in various domains. These include automated query answering ([Cui et al., 2017](#); [Vu et al., 2021](#)), personal task managers ([Toxtli et al., 2018](#); [Prathiksha et al., 2024](#)), healthcare assistants for symptom screening and behavioral interventions ([Li et al., 2019](#); [Pan et al., 2023](#); [Ferreira et al., 2024](#)), mental health support ([Heinz et al., 2025](#); [Coghlan et al., 2023](#)), online education ([Hwang et al., 2023](#); [Baha et al., 2023](#)), hospitality ([Yilmaz & Şahin-Yilmaz, 2024](#)), entertainment ([Garcia et al., 2021](#)), and recruitment ([Albassam et al., 2023](#); [Rukadikar et al., 2024](#)).

One emerging application area focuses on collecting user feedback to improve services (Sachdeva et al., 2024; Sun et al., 2023). These agents support personalized, interactive communication, enabling organizations to obtain real-time insights more effectively than traditional surveys. Unlike static forms, conversational agents engage users dynamically, adapting to responses to elicit detailed and context-specific feedback. This results in higher-quality input and improved user satisfaction.

Recent research supports the use of deep learning for feedback processing. Park et al. developed a hybrid model integrating both explicit and inferred feedback ([Park et al., 2020](#)), while Kachuee et al. proposed a self-supervised method using contrastive learning with BERT encodings, demonstrating the value of leveraging unlabeled data ([Kachuee et al., 2020](#)).

Beyond feedback collection, conversational agents enhance explainability in AI systems. Explainable AI (XAI) aims to make AI decisions understandable, fostering user trust. XAI methods can be global—explaining general model behavior—or local—explaining individual predictions ([Guidotti et al., 2018](#); Ribeiro et al., 2016). Some methods generate post-hoc explanations, while others use self-explanation mechanisms embedded in decision-making, the latter being more interpretable and efficient (Arya et al., 2019; [Danilevsky et al., 2020](#)).

Arya et al. (2019) further distinguish between static and interactive XAI, noting that natural language explanations are often clearer than visual or template-based formats ([Miller & Jing, 2024](#); [Lipton, 2018](#)). [Weitz et al. \(2021\)](#) showed that explainable virtual agents increase user trust in speech recognition models. Similarly, [Joshi et al. \(2024\)](#) found that systems with integrated explanations led to greater user trust and acceptance in a travel booking scenario.

This section focuses on the use of conversational agents in (i) collecting user feedback and (ii) supporting AI explainability. We examine design and implementation strategies, evaluating issues such as engagement, bias, feedback quality, and communicating complex decisions. Conversational agents help organizations understand user preferences and also serve as mediators of AI transparency. We highlight methods for detecting user intent and the use of large language models (LLMs) in generating retrieval-based explanations. Evaluation challenges are discussed using both intrinsic and extrinsic metrics. As a case study, we explore how LLM-powered conversational agents were integrated in the vera.ai project to enhance misinformation detection and verification tools. The section outlines implemented methods, evaluation techniques, and discusses open challenges, including maintaining conversational history and preventing hallucinations. Finally, we discuss future directions in the use of conversational agents in the era of LLMs.

7.1.2 User Intent Detection

The effectiveness of conversational agents largely depends on their ability to understand user input, with intent recognition being a crucial step. Intent recognition classifies user utterances into predefined categories (e.g., booking a hotel, scheduling a meeting), enabling smooth task execution. It is often followed by slot-filling, which extracts relevant details like time, location, or price range to complete the task ([Tur & De Mori, 2011](#)).

Fallback intents handle cases where the agent fails to recognize user input, and multi-intent recognition is gaining attention as users may express multiple intents simultaneously ([Meng et al., 2022](#)). Early intent classification methods relied on supervised learning algorithms like SVMs ([Mendoza & Zamora, 2009](#)) and Hidden Markov Models (HMM) ([Bhargava et al., 2013](#); [Zhang et al., 2009](#)). Slot-filling was typically addressed via sequence labeling using CRFs and their variants ([Sarikaya et al., 2014](#), [Dai et al., 2021](#)). These approaches required significant feature engineering and struggled with generalization.

Recent advances in deep learning—particularly RNNs, LSTMs, and transformer-based models like BERT and GPT—have significantly improved performance in joint intent recognition and slot-filling tasks ([Kurata et al., 2016](#); [Chen et al., 2019](#); [Weld et al., 2022](#)). In user feedback scenarios, intent detection helps identify sentiment and feedback categories ([Moradizyvehi, 2022](#)). Self-supervised learning methods, such as multi-task pretraining ([Zhang et al., 2022](#)), have been used to identify novel intents in unlabeled data.

Slot-filling in feedback systems extracts details like product name or issue type, though unstructured feedback often requires clarification questions to fill missing information ([Zhang et al., 2020](#)). Despite progress, challenges remain: vague feedback, domain adaptation, and maintaining context in multi-turn dialogues ([Sun et al., 2022](#)). Zero-shot and few-shot learning approaches are being explored to handle unseen intents ([Liu et al., 2022](#); [Atuhurra et al., 2024](#)).

Another key issue is real-time processing. Applications like virtual assistants demand fast responses, but deep learning models can be computationally intensive. Techniques such as knowledge distillation, pruning, and quantization help reduce complexity while preserving accuracy ([Vakili et al., 2020](#); [Pawlik, 2025](#)). However, balancing efficiency and accuracy remains a key research challenge.

7.1.3 Retrieval Augmented Generation

The inherent opacity of many AI models has made explainability essential, particularly in user-facing applications. Retrieval-Augmented Generation (RAG) ([Lewis et al., 2020](#)) has emerged as a promising solution by combining retrieval mechanisms with language models to generate contextually grounded and transparent responses. RAG models typically consist of two components: (i) a retrieval module that searches indexed sources for relevant information based on the user's query, and (ii) a language model that uses the retrieved content to generate an informed response. This hybrid architecture enhances both the relevance and explainability of answers by grounding generative outputs in domain-specific evidence.

RAG is particularly useful for question answering tasks requiring sourced and detailed explanations. Unlike standard generative models that rely solely on training data, RAG systems dynamically fetch information from external sources, reducing hallucinations and increasing factual accuracy ([Izacard & Grave, 2021](#)). This capability is especially valuable in high-stakes domains like healthcare ([Al et al., 2023](#)), finance ([Yepes et al., 2024](#)), and legal services ([Wiratunga, 2024](#)), where reliability is critical. By retrieving verifiable data, RAG systems enhance user trust and facilitate informed decision-making.

Another key application is in explainable AI (XAI), where RAG can function as a post-hoc global and local explanation tool. It retrieves relevant documentation or examples to help users understand AI model decisions ([Yuan, 2024](#)). Guttikonda et al. ([Guttikonda et al., 2024](#)) tested RAG-based explanations across diverse applications—predicting manufacturing costs, classifying brain tumors, and assessing loan eligibility—by integrating project documentation as a source for retrieval. For instance, in the healthcare case, users could ask how a diagnosis was made, and the RAG system would surface relevant explanations from clinical documents and model descriptions.

Despite its advantages, implementing RAG for explainability presents challenges. The relevance and quality of retrieved documents are critical—irrelevant or misleading sources can undermine trust and clarity ([Cuconasu et al., 2024](#)). Retrieval quality depends on advanced techniques like query rewriting,

chunking, metadata annotation, re-ranking, and fine-tuning embedding models ([Setty et al., 2024](#)). Some studies suggest that injecting noise or diverse examples into retrieval can improve robustness ([Cuconasu et al., 2024](#)). To handle low-confidence outputs, Chen et al. propose a counterfactual prompting framework, where the model generates alternative answers to validate or challenge its own predictions, prompting it to abstain from low-confidence responses ([Chen et al., 2024b](#)).

Integrating RAG in real-time systems poses additional challenges. Retrieval and generation introduce latency, which is problematic for responsive user experiences ([Hofstatter et al., 2023](#)). While quantized models improve efficiency, they often reduce accuracy ([Zhou et al., 2025](#)). Methods like hierarchical caching, query routing, semantic search, and using cached documents have been proposed to improve response speed without sacrificing quality ([Reynolds, 2024](#); [Chan, 2024](#)).

The explainability of RAG-generated outputs themselves is also under scrutiny. Tools like RAGViz visualize attention across retrieved documents to illustrate how context affects the generated response (Wang, 2024). However, retrieved texts may be too complex for non-expert users, limiting accessibility. Moreover, the multi-component nature of RAG systems—retriever, ranker, generator—introduces new transparency challenges, particularly around misinformation propagation and architectural complexity (Ni, 2025).

Finally, research on RAG explainability remains nascent, especially in terms of standardized evaluation. There is a lack of unified datasets and benchmarks for measuring different facets of explainability, such as relevance, factuality, and interpretability. In the following section, we discuss the goals and obstacles in building evaluation frameworks for RAG-based models.

7.2 Methodology

Figure 40 illustrates the high-level architecture of the conversational chatbot for user feedback collection and XAI generation. As can be seen, it consists of two main steps, (i) user intent detection, and (ii) retrieval-augmented generation for the explanation-seeking intent. The user intents can be broadly grouped into three main categories: (a) feedback collection, (b) explanation generation, and (c) state transition interaction.

7.2.1 Intent Detection

For the user intent detection, we trained a multilingual multilabel model that is able to discriminate between 6 types of user intents. The multilabel definition of the task is essential as users often tend to express several questions as one message.

One of the challenges of training highly-specific intent recognition models is the lack of data. For our task, we manually created a seed dataset of 300 examples of user questions in English along with a sparse vector of associated intents. We split the data into training (80%), validation (10%) and test (10%) sets, stratified by the intent class. We used the validation set to empirically estimate the optimal threshold for predicting individual intents, which in our case was 0.58.

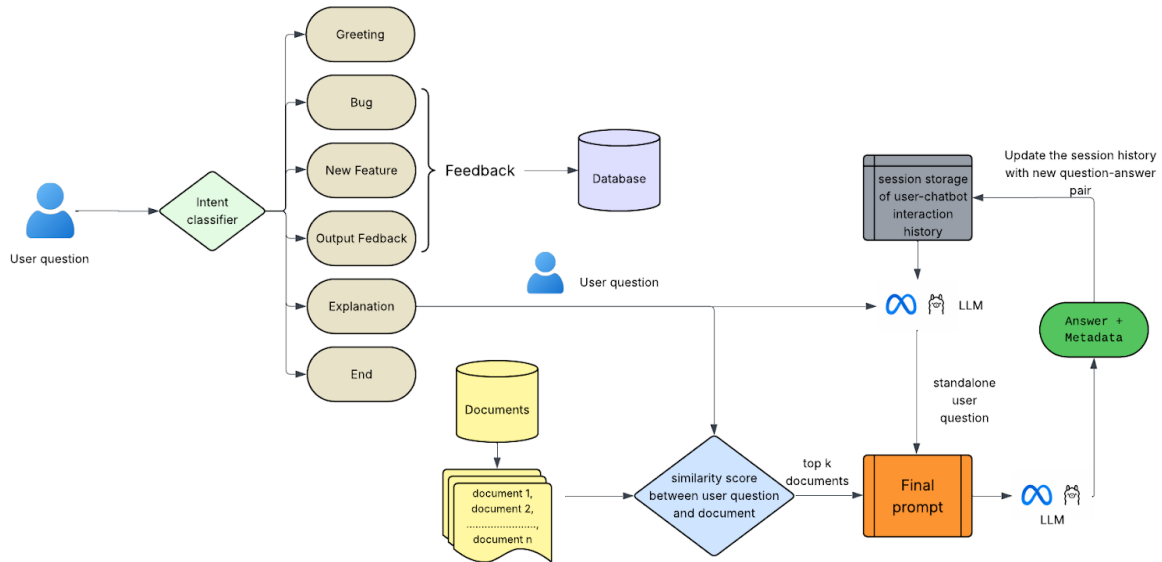


Figure 40 Chatbot Workflow

We experimented with fine-tuning two types of multilingual language models, mBERT-base ([Pires et al., 2019](#)) and XLM-RoBERTa-base ([Liu et al., 2019](#)). We opted for the base versions of the models due to the importance of latency metric in generating replies. Given that the intent detection step is only the first step towards generating an answer, larger models introduced significant delays that could potentially affect users' satisfaction. Each model was trained for 20 iterations. We used with the learning rate of $1e-5$, AdamW optimiser ($=1e-8$) and ReLU activation function in the classifier. For comparison purposes, we trained each model 3 times with various random seeds. Although the training data was only available in English, we benchmarked our models on 6 languages, by translating the test set in English (EN) into German (DE), French (FR), Spanish (ES), Arabic (AR) and Italian (IT) using Google Translate and evaluating how the model performs in zero-shot cross-lingual scenarios.

7.2.2 Retrieval Augmented Generation

To effectively address user requests for explanations, the chatbot employs a specialized retrieval-augmented generation (RAG) framework. This approach combines information retrieval with language generation to produce coherent, relevant, and informative responses. As illustrated in Figure 42, when the intent classifier detects an explanation-seeking query, the system delegates the request to the RAG module. This module operates through a three-stage pipeline: document retrieval, standalone question generation, and final response generation. The remainder of this section delves into each of these stages in detail.

Document retrieval

For this sub-task, we generated an indexed database for storing chunks of the documents related to the models and user interface. This includes but is not limited to the implementation details behind the AI models (model cards) used for synthetic image detection, the types of data the models were trained on, associated publications, interpretation of the output scores and information regarding the user interface.

Metadata annotations were previously shown to be beneficial for retrieval in RAG-based assistants ([Shi et al., 2024](#)). Therefore, instead of splitting documents into same-length chunks, we annotated the document sections based on the types of information they contain and performed the document split based on markdown headers, with the token overlap of 20 between all the splits. This step is implemented by means of a Langchain framework ([Mavroudis, 2024](#)). We then encode the embeddings into the all-mpnet-base-v2 ([Song et al., 2020](#)) representation and store the indexed embeddings of the document chunks in a permanent Chroma database. We also implement the background process that is constantly checking whether some of the source documents were changed or new documents are added. In case of a change, the database is updated with new embeddings and is re-indexed. This allows to ensure the up-to-dateness of the RAG model, one of the critical extrinsic evaluation criteria ([Abbasian et al., 2024](#)). We then use a vector store retriever from Langchain to perform a similarity search between the embedded user question and the database contents. We experimented with three types of document search: similarity-based, mmr, and by calibrating a similarity score threshold. To perform a thorough evaluation of various components of the RAG chatbot, we manually constructed a set of user questions along with up to five document chunks containing the answer to those questions. We then evaluated each of the three search techniques based on the Recall@k and MRR score.

Standalone question generation

Incorporating history into RAG systems provides a robust mechanism for improving the quality and relevance of conversational agents, directly influencing the sensibility within the SSI framework for measuring the extrinsic performance of chatbots. By incorporating the conversational history into the retriever, RAG systems can generate responses that are not only factually grounded but also contextually sensitive and coherent over time.

We implement the conversation history by means of dynamically storing the user-chatbot conversations, accompanied by a unique session identifier, in an SQLite database. At each new user question, we retrieve the prior messages that correspond to the current session, and analyse the current question in the context of prior interaction. To do so, we pass the user question and the retrieved conversation history through the LLM in order to formulate the standalone question. We define a standalone question as a question that can be understood on its own, without additional context. To generate a standalone question, the LLM is provided with the following prompt:

```
Given a chat history and the latest user question which might reference
context in the chat history, formulate a standalone question which can
be understood without the chat history. Do NOT answer the question, just
reformulate it if needed and otherwise return it as is.
```

```
{history}
```

```
{question}
```

Final answer generation

The final stage of the explanation generation process is to generate the final answer given the prompt along with top 3 retrieved documents and a standalone question generated at the previous step.

We performed a thorough prompt engineering to ensure both extrinsic and intrinsic evaluation criteria, such as minimising the amount of ungrounded information in the reply and encouraging the answer conciseness. We observed that the location of the context in the prompt plays pivotal importance in generating the grounded reply, with the context at the start of the prompt resulting in the LLM ignoring the retrieved documents and generating the answers based on its own pre-training information. At the same time, placing the context and the question at the end of the prompt was causing LLM forgetting particular instructions, specifically about the answer length. It also often resulted in model hallucinations in the form of generating follow-up questions on behalf of the user and answering those questions. Below is the optimal prompt format that was used in our final implementation.

You are an assistant chatbot for answering user queries about the model details and interface of the verification plugin for synthetic image detection.

- Answer ONLY the question asked and do NOT generate related questions or additional information.

- If the answer is not found in the provided context, respond with: 'We're sorry, but we couldn't find the answer in our database. Please feel free to ask another question. Thank you!'

- Use three sentences maximum and keep the answer concise.

{question}

{context}

RAGAs framework (James et al., 2024) provides support for generating a synthetic test set for RAG systems given the task-specific documents by means of creating a knowledge graph and generating the context-grounded questions and answers. When generating questions, it is possible to choose the distribution between one-hop and multi-hop specific and abstract query generation. Here, *one-hop* queries can typically be answered by following from one connected node to another, while *multi-hop* queries rely on more complex reasoning with intermediate nodes between the two target nodes. In our use case scenario, multi-hop questions can connect the document chunks describing the user interface and model implementation, such as "*The ProGAN is shown as light green -- are you sure your model is accurate enough?*". Both single- and multi-hop queries can be either abstract or specific. Here, *specific* questions target concrete concepts and facts that can be directly extracted from the context. For example: "*What metrics were used to measure the model performance?*" is a specific question in our use case scenario. *Abstract* questions, on the other hand, are challenging for RAG models as they require a high-level analysis of the context and the retrieval of document metadata. An example of an abstract question in our application can be: "*How do you know it is synthetic?*".

In order to perform a better comparison, we generated four types of test examples, with 100 questions and answers in each set: one-hop specific, one-hop abstract, multi-hop specific, and multi-hop abstract. We later manually inspected each of the generated examples and selected 50 in each category that are best fitted to be included in the test set. This allowed us to compile 4 separate test sets, and one joined test set that contains a mixture of 200 question-answer pairs.

Additionally, we wanted to test the model's robustness in terms of negative rejection. For this purpose, we manually created a set of 50 questions that cannot be answered based on the documentation. As a reference answer to each of these questions, we specified "*I don't know*", as this is what the model was instructed to do in such cases.

7.3 Evaluation

User intent detection

Table 29 represents the performance of multilingual BERT and RoBERTa models on English and 5 unseen languages. Despite being trained on a relatively small dataset and the complexity of the task in terms of the multiclass multilabel scenario, the models are able to achieve a reliable performance even on complex intents, such as explanation generation or feedback provision. As can be seen, XLM-RoBERTa model is on average outperformed by the mBERT model for the majority of languages despite being significantly larger (270 million parameters compared to 110 million). The two zero-shot cross-lingual scenarios where XLM-RoBERTa demonstrates better average performance are the predictions on French and Spanish. Unlike mBERT which performs consistently better on English test data, RoBERTa exhibits better performance on zero-shot languages for certain intent-language pairs. For example, the model's performance in *explanation* intent detection on French and Spanish surpasses that on English despite being trained on English data.

Unsurprisingly, context-independent intents, such as *greeting* and *end* are significantly easier to identify than the task-specific intents that often rely on the specific context or previous interaction history. This is consistent across all the languages and for both models. For example, the user question that states "*This image was downloaded from BBC*" is highly complex in nature due to its implicit intent. It does not explicitly express the dissatisfaction with the models' prediction, but instead mentions the source from which the image is retrieved. To link this statement with the intent to provide feedback about the models' prediction, even a human expert needs to perform a series of inference steps. First, one needs to remember a general knowledge fact suggesting that the BBC is typically associated with publishing trustworthy media. Additionally, the expert needs to have access to the model's output, which flags the image as synthetic in this case. Such complex reasoning steps are known to be challenging for even LLMs, therefore it is not unusual to observe smaller language models struggling to identify implicit dialogue intents.

Table 29 F-score on multilingual data for various types of user intents. Boldfaced scores demonstrate statistically significant better overall performance of one model compared to the other for a certain language

Intent Type	Multilingual BERT						Multilingual RoBERTa					
	EN	DE	FR	ES	AR	IT	EN	DE	FR	ES	AR	IT
Greeting	100±0.0	98±1.1	99±0.4	100±0.0	99±0.4	99±0.7	100±0.0	100±0.0	98±2.2	99±1.1	100±0.0	94±3.6
Explanation	80±4.1	71±1.9	65±2.9	62±1.8	75±1.7	74±1.6	77±1.9	75±3.8	80±2.6	68±4.7	65±2.1	70±3.2
Bug	92±2.4	65±1.7	68±4.2	60±3.5	79±2.5	81±1.4	79±1.4	60±3.2	73±3.7	79±1.3	70±2.0	69±2.8
New Feature	93±1.4	67±2.0	60±1.8	71±1.9	77±2.3	80±1.9	81±1.7	71±1.4	77±2.9	73±1.9	78±3.9	78±3.6
Output Feedback	84±3.7	69±1.5	72±3.3	65±2.6	73±1.7	82±3.5	83±3.3	69±5.1	71±1.0	80±3.0	80±1.2	81±1.7
End	95±1.5	94±2.4	95±1.6	98±1.1	95±1.0	93±2.2	96±3.1	88±3.7	91±4.4	94±2.5	93±1.7	95±1.3
F1-macro	89±4.2	77±3.1	77±4.3	76±2.7	83±2.1	85±2.9	86±3.8	77±5.2	82±4.1	82±3.4	81±2.9	81±3.9

RAG for explanation generation

Document retrieval: Table 30 shows the results of search algorithm and retrieval length calibration. As can be seen, although the recall benefits from retrieving more documents, a more balanced MRR metric shows optimal performance for 3 documents. This balance is particularly important for RAG-based systems, as irrelevant contexts can introduce additional noise and decrease factuality during the subsequent generation steps. We also observed the maximal marginal relevance (MMR) (Guo & Sanner, 2010) algorithm for retrieving relevant documents to yield best results. We therefore proceeded with using 3 document chunks and MMR search type for the retriever in the main implementation.

Table 30 Retriever performance in Terms of Recall@k and MRR for different number of retrieved documents and for three types of document search

Number of documents to return	Similarity-based		MMR		Threshold-based	
	Recall@k	MRR	Recall@k	MRR	Recall@k	MRR
k=5	0.54	0.24	0.56	0.24	0.55	0.25
k=2	0.76	0.26	0.77	0.26	0.76	0.25
k=3	0.81	0.28	0.85	0.30	0.83	0.29
k=4	0.81	0.24	0.87	0.28	0.85	0.29
k=5	0.82	0.24	0.87	0.28	0.86	0.29

Standalone question generation: Table 31 provides an overview of the model's performance on this task. As can be seen, Llama model consistently outperforms other LLMs despite being the smallest one in terms of the number of parameters (1B compared to 7B for the rest). Our qualitative analysis revealed that Qwen model was more prone to keeping the input unchanged, while Falcon added more of the context

that is required to generate a standalone question. All of the models, including Llama, suffered from long history inputs. In cases where the history accounted for more than 5 exchanged question-answer pairs, the models would often either ignore the instruction or failed to account for older history to interpret the question.

Based on our initial evaluation, we chose Llama-3.2-1B-Instruct for both the standalone question and the final answer generation. The choice of a specific LLM was motivated by several factors, such as the performance on the standalone question generation, latency of generation (which was lower due to the smallest size of the model), and the cost of running and maintaining the LLM.

Table 31 Performance of various LLMs on the standalone question generation task based on BLEU, METEOR and ROUGE-2 metrics

LLM	BLEU	METEOR	ROUGE-2
Llama-3.2-1B-Instruct	0.58	0.26	0.72
Falcon3-7B-Instruct	0.54	0.19	0.61
Qwen2.5-7B-Instruct-1M	0.48	0.21	0.58
Mistral-7B-Instruct-v0.2	0.55	0.24	0.72

Final answer generation: We evaluated the performance of the system using RAGAs' framework that takes sample documents, user query, reference answers and the model's outputs as an input. Although factual correctness within RAGAs already captures the similarity between the generated answer and the reference, we additionally evaluated the model based on the conventional semantic similarity between just two components, reference answers and model's outputs, using BERTScore. Finally, we measured the average latency in receiving the model's final output. Table 32 demonstrates the model's performance based on various test sets and metrics.

Table 32 Performance of the model on various types of the test sets based on RAGAs, BERTScore and Latency

LLM	Context Recall	Faithfulness	Factual Correctness	BERTScore	Latency (sec)
One-hop specific	0.92	0.89	0.71	0.94	6.2
One-hop abstract	0.72	0.65	0.63	0.87	8.5
Multi-hop specific	0.70	0.77	0.68	0.86	6.2
Multi-hop abstract	0.67	0.54	0.47	0.82	14.3
Joined test set	0.81	0.69	0.59	0.87	8.8
Negative rejection	0.38	0.25	0.21	0.78	7.4

7.4 Implementation and Integration

We integrated a synthetic image chatbot into the vera.ai synthetic image detection tool to assist users and collect feedback through a familiar instant messaging-style interface. Users can submit images for analysis, receive a score indicating the likelihood of AI generation, and interact with the chatbot for

explanations or to report feedback. The chatbot is fully integrated via a GATE cloud service endpoint, seamlessly connecting the front-end interface with the assistant backend and enabling real-time responses, feedback logging, and conversation management.

7.4.1 Integration of the Synthetic Image Chatbot

We use the chatbot to provide assistance and to collect user feedback for the vera.ai synthetic image detection tool. The interface allows users to submit an image for analysis, and the tool provides a score representing how likely it is that the image was generated by AI. The users can then submit questions and feedback via the chatbot interface illustrated in Figure 41. User inputs and chatbot responses are displayed as “message bubbles”, giving the interface a familiar instant messaging type look.

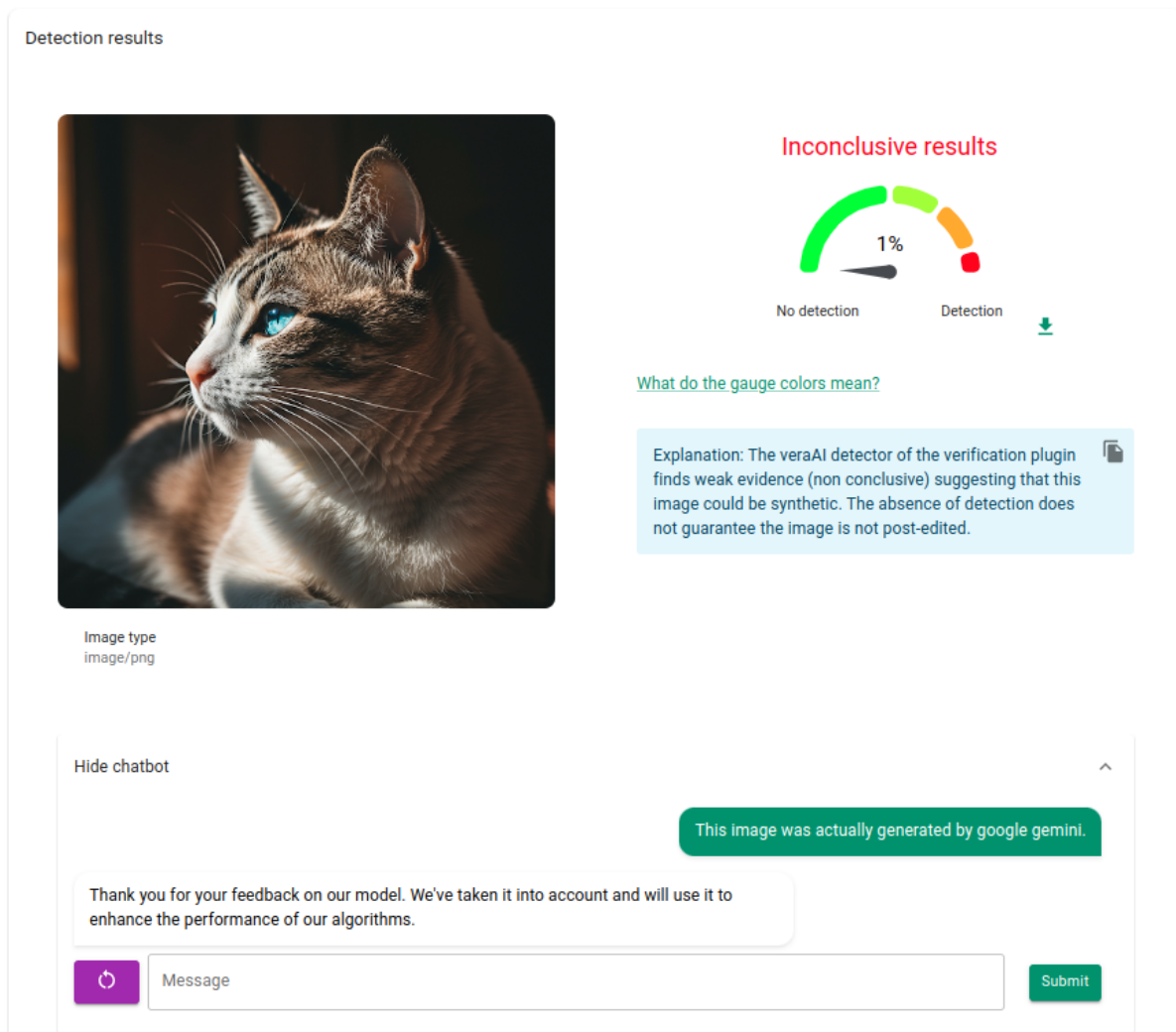


Figure 41 The Chatbot being used to collect feedback of an image of a cat generated by Google Gemini

When a user submits a message, it is sent to the assistant backend, along with the session ID. The request is then passed straight through to the chatbot endpoint. A two-stage process then follows to generate the chatbot response.

IMPROVEMENT

This image was actually generated by google gemini.
[Sent via the chatbot assistant]



<mailto:m.foster@sheffield.ac.uk>

Figure 42 Post on the feedback Slack channel with the user's email address, feedback message, and the image

First, the intent of the user's message is classified. If the intent is feedback collection, a stock response is sent back to the user, depending on whether the feedback is a *bug report*, a *new feature request*, or *output feedback*. For example, in Figure 41, the user has informed the chatbot that the image they submitted has, in fact, been generated by AI. In addition to generating a response to the user, the chatbot automatically posts all feedback messages to the tool's feedback Slack channel, along with the email address of the user and the image related to their feedback, so that reports can be followed up. This is illustrated in Figure 42.

However, if the intent of the user's message is classified as explanation generation, the RAG model is called to generate a suitable answer to the user's query. Once such an answer has been generated, it is passed back to the front-end and displayed to the user.

Since Llama does not explicitly keep track of a session history itself, we need to manage this on our end. To do so, we use a database that indexes conversations by the session ID, which is included as part of each user message. The chatbot can then easily retrieve the conversation history from the database and pass this to Llama to use when generating its response. This poses an interesting question: at what point should we delete the conversation history? This is important, both to keep the conversation history relevant to the user's current task, and to ensure that the database remains a manageable size in terms of both storage and query response time. We keep track of the session history from two perspectives. Firstly, the user can start a new conversation at any time by clicking the "New conversation" button, shown to the left of the message box in Figure 41. This will start a new chatbot session and clear the old session history. Second, if the user provides a new input to the analysis tool, this begins a new session. Here, though, it is less clear whether or not existing sessions should be deleted, since the user may have several browser tabs open. Furthermore, previous chatbot conversations can be used to fine-tune the chatbot model so that its responses get more relevant over time. This is particularly useful for common problems where similar questions are frequently asked. Therefore, at this stage of integration, we retain conversation histories, and delete them after a set time period has elapsed in order to protect data and manage the growth of the database.

7.4.2 Full Integration

To integrate with the assistant, we implemented a GATE cloud service, accessible via the `/gcloud/chatbot` endpoint. Users submit their queries and comments to the chatbot via a familiar instant messaging style interface. These messages are sent to the assistant backend, which forwards them

on to the chatbot itself. The replies are then sent back to the assistant backend, and then passed through to the plugin frontend to be displayed to the user. When a user reports a problem with the plugin or requests a new feature, a message is pushed to the development slack by the plugin containing contact information for the user, their message, and the image that they processed. No personal information is passed to the chatbot, unless the user directly identifies themselves within their message.

Along with user messages, a session ID is also passed to the chatbot so that the conversation history of each user session can be maintained. We also send through the tool name and its result to provide the chatbot with the context of the user's queries. Currently, the chatbot is only integrated into the "syntheticImageDetection" tool, where the `result` is the name, prediction score, and error status of each of the detection models, as shown in the Figure 43 below. In future, the chatbot will be integrated into other tools within the plugin.

```
# Message from user

{
  "message": "Can you explain why this image was not detected as
a
          fake?",
  "sessionID": "5b5766a2-5ec0-48ae-922f-4302f131d09e",
  "tool": "syntheticImageDetection",
  "result": [{
    "apiServiceName": "progan_r50_grip",
    "predictionScore": 0.08768573170527816,
    "isError": false
  }, {
    "apiServiceName": "ldm_r50_grip",
    "predictionScore": 0.12565169017761946,
    "isError": false
  }, {
    "apiServiceName": "gan_r50_mever",
    "predictionScore": 0.12492934474721551,
    "isError": false
  }]
}

# Response from chatbot

{
  "message": "The image was not detected as fake because it was a
picture of a real person and not a model.",
  "status": "success",
  "userMessageClasses": ["EXPLAIN"]
},

# Message from user
```

```

{
  "message": "The image was actually a picture of a cat generated by AI",
  "sessionID": "5b5766a2-5ec0-48ae-922f-4302f131d09e",
  "tool": "syntheticImageDetection",
  "result": [{
    "apiServiceName": "progan_r50_grip",
    "predictionScore": 0.08768573170527816,
    "isError": false
  }, {
    "apiServiceName": "ldm_r50_grip",
    "predictionScore": 0.12565169017761946,
    "isError": false
  }, {
    "apiServiceName": "gan_r50_mever",
    "predictionScore": 0.12492934474721551,
    "isError": false
  }]
}

# Response from chatbot

{
  "message": "Thank you for your feedback on our model. We've taken it into account and will use it to enhance the performance of our algorithms.",
  "status": "success",
  "userMessageClasses": ["IMPROVEMENT"]
}

```

Figure 43 JSON messages to and from the chatbot

7.5 Concluding Remarks

This section explored the integration of retrieval-augmented conversational agents within a fact-checker-in-the-loop framework, aiming to enhance both user feedback collection and explainability in content verification. We detailed the challenges faced by current RAG systems, including hallucination, latency, domain transferability, and the disjoint nature of retriever-generator architectures. Empirical evidence from our experiments highlighted the inadequacy of traditional evaluation metrics like BLEU and BERTScore in capturing retrieval faithfulness, motivating the adoption of emerging intrinsic metrics such as RAGAs. Furthermore, we discussed the limitations of semantic similarity in document retrieval and advocated for future development of re-ranking methods centered on contextual relevance. Despite achieving perfect retrieval in some cases, LLMs still hallucinated retrieved content, emphasizing the need for deeper integration between retrieved evidence and generation mechanisms. Overall, this section demonstrates the need for holistic optimization of RAG pipelines and greater alignment between retrieval fidelity and generative faithfulness, setting the foundation for future research toward more trustworthy, adaptive, and feedback-aware conversational verification tools.

8 Efficient Annotator Reliability Assessment

Labelled data is the foundation of training and evaluating downstream tasks in machine learning models. However, data annotation is often an expensive and time-consuming process, significantly affecting the quality of model training. Obtaining annotations from experts is ideal, but this expertise is often logistically and financially costly.

8.1 Background

Crowd-sourcing platforms such as Amazon's Mechanical Turk and CrowdFlower (now known as **Figure Eight**) provide a cheaper alternative by using non-expert annotators; this generally results in lower quality annotations with higher levels of inter-annotator disagreement ([Nowak & Ruger, 2010](#)). Effectively collecting, evaluating and managing annotator disagreement is essential in addressing these challenges.

We introduce EffiARA (Efficient Annotator Reliability Assessment) framework which supports annotation quality assessment and management throughout the annotation process, allowing users to:

- Distribute data points to annotators;
- Generate labels for each annotator;
- Assess agreement among annotators;
- Assess annotator reliability;
- Redistribute data points to obtain the desired level of agreement;
- Generate aggregated labels at the data point level, taking either a soft- or hard-label approach.

To our knowledge, no existing annotation framework provides systematic support for annotator workload allocation which can then be used to estimate the cost of the annotation project. This, in addition to the set of functionalities surrounding the annotation process, makes the EffiARA annotation framework a unique solution for structuring data annotation and modelling annotators.

Additionally, by aggregating annotators' labels for each data point, tempered by measures of annotator reliability, we can obtain a consensus that better reflects the “true” label distribution. Annotator reliability can also be used to dynamically weight individual data points during model training to ensure that the model prioritises reliable annotations ([Cook et al., 2024](#)).

Annotation Frameworks: There have been many attempts to formalise the annotation process for a number of annotation tasks and a range of tools are available. Many frameworks focus on sequence labelling tasks such as POS tagging and named-entity recognition ([Bird & Liberman, 2021](#); [Cornolti et al., 2013](#); [Bontcheva et al., 2013](#); [Lin et al., 2019](#)). More recently, with the introduction of pre-trained LLMs capable of document-level processing, document annotation tools and frameworks have been created such as GATE Teamware 2 ([Wilby et al., 2023](#)). A number of annotation frameworks are task-specific, aiming to provide a set of guidelines and tools for following them, for example event ordering ([Cassidy et al., 2014](#)), biodiversity information extraction ([Lücking et al., 2022](#)), and surgical video analysis ([Meireles et al., 2021](#)).

Annotator Agreement: Agreement among annotators is often used to assess the quality of a dataset. Commonly used metrics include Scott's Pi ([Scott, 1955](#)), Cohen's Kappa ([Cohen, 1960](#)), Fleiss' Kappa ([Fleiss,](#)

1971), and Krippendorff's alpha (Krippendorff, 1970). For each metric, there are various interpretations and accepted agreement thresholds used to determine the reliability of a dataset (Krippendorff, 2018; Landis, 1977). Obtaining datasets where this agreement threshold is met, particularly in scenarios with non-expert annotators such as crowd-sourcing, is challenging and costly (Hsueh et al., 2009; Nowak & Rüger, 2010).

Annotation Aggregation: Rather than ensuring acceptable levels of agreement, many approaches use disagreement as additional information, utilising it to understand the subjectivity of particular data points or the reliability of annotators.

The soft label approach incorporates a level of subjectivity into aggregated labels for each data point and has been shown to improve both classification performance and model calibration (Wu et al., 2023b; Cook et al., 2024). Popular methods of label aggregation include majority voting (hard-label only), (Dawid & Skene, 1979), GLAD (Whitehill et al., 2009), and MACE (Hovy et al., 2013) for categorical data; these methods have been implemented in Python as part of the Crowd-Kit tool (Ustalov et al., 2021).

Annotator Reliability: Assessing annotator reliability can be used to assess the quality of individual annotators and may be used to understand the quality of data available, remove low-reliability annotators (Cook et al., 2025), inform the training of machine learning models through aggregating soft labels with reliability (Dawid & Skene, 1979; Wu et al., 2023b), or affect the loss function during training (Cao et al., 2023, Cook et al., 2024). There are different approaches to assessing annotator reliability, such as learning through Expectation Maximisation (Cao et al., 2023) or directly inferring the reliability of an annotator from their agreement with others (Inel et al., 2014; Dumitrache et al., 2018; Cook et al., 2024; Cook et al., 2025). Through impacting the generation of soft labels or directly impacting the training loss, more information about which annotations are more trustworthy is provided, leading to more performant and robust models. An alternative approach to assessing annotator reliability involves comparing annotators' annotations to a set of gold-standard labels; this approach is often used to filter out bad annotators. All three approaches have been shown to improve model performance on classification tasks when compared to methods that trust each annotator equally.

8.2 EffiARA Python Package

The EffiARA annotation framework structures the annotation process from start-to-finish. It distributes samples to annotators, generates and aggregates labels, computes inter- and intra-annotator agreement, and assesses annotator reliability. A visual representation of the EffiARA pipeline is provided in Figure 44 and we describe each stage in detail below.

The annotation pipeline is implemented as a set of modular tools in the EffiARA Python package. The source code is available at <https://github.com/MiniEggz/EffiARA> and the package has been released on PyPi for quick installation: <https://pypi.org/project/effiara/>. Documentation is available here: <https://effiara.readthedocs.io>.

The package relies on a number of core Python libraries. Two fundamental libraries required by the EffiARA framework are NumPy (Oliphant et al., 2006; Harris et al., 2020) and pandas (McKinney, 2011) for efficient mathematical operations on arrays and the manipulation of data.

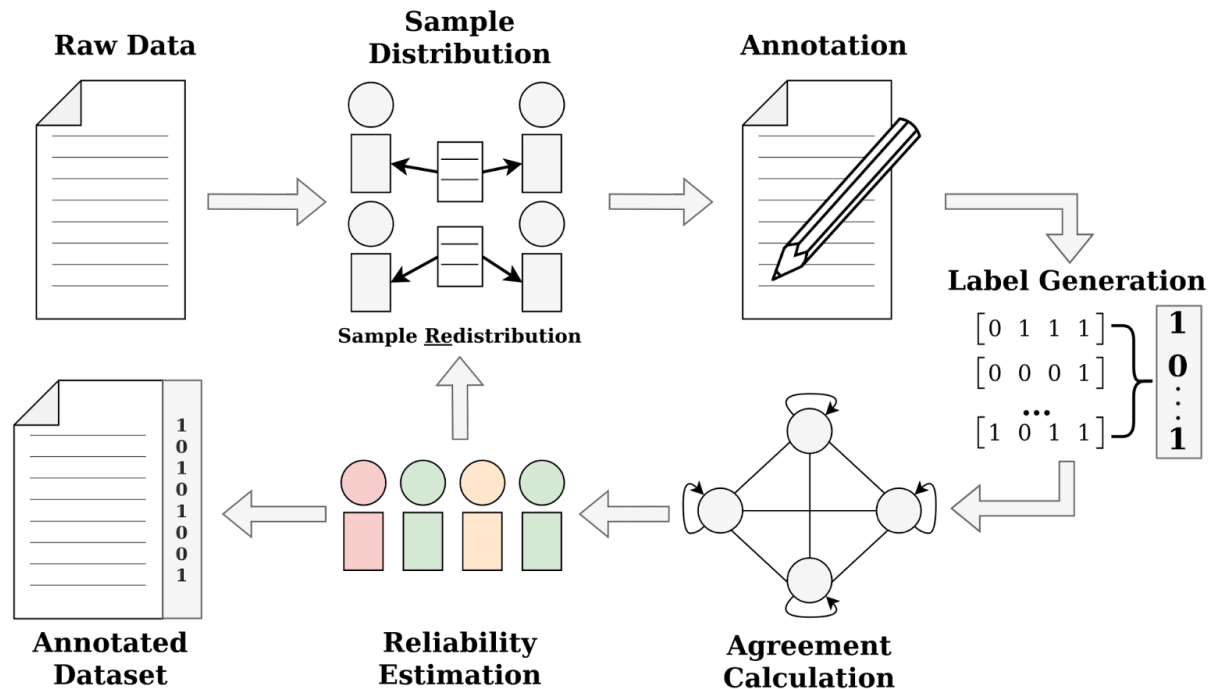


Figure 44 An overview of the EffiARA annotation pipeline, covering sample distribution, annotation, label generation, agreement calculation, reliability estimation, and dataset compilation

Sample Distribution: The first stage in the EffiARA pipeline enables annotation coordinators to estimate resource requirements: how many annotators are needed, how much time is required from each annotator, and how many samples can be produced, given the time and number of annotators. Once resources have been finalised, data points can be distributed among annotators with the EffiARA distribution algorithm, which ensures annotator agreement can be effectively assessed (Cook et al., 2024).

Both of these functionalities are implemented in the *SampleDistributor* class. We first use SymPy (Meurer et al., 2017) to solve for the missing variable in the resource-understanding equation introduced in (Cook et al., 2024) (Algorithm 1). We then use pandas to split the data into separate DataFrames for each annotator, with one DataFrame containing left-over samples that may be used later.

Data Annotation: The sample allocations obtained in the previous step can then be used to assign samples to annotators and complete the annotation process using existing tools such as GATE Teamware 2 or Amazon's Mechanical Turk.

Label Generation: Label generation involves transforming raw annotations obtained from annotators into numerical encodings compatible with annotator agreement metrics (such as Cohen's Kappa, Fleiss' Kappa, Krippendorff's alpha, or cosine similarity) and model training. These transformations may be at the individual annotator level (for example, transforming first- and second-choice annotations into a categorical distribution), or at the data point level (aggregating annotations from multiple annotators).

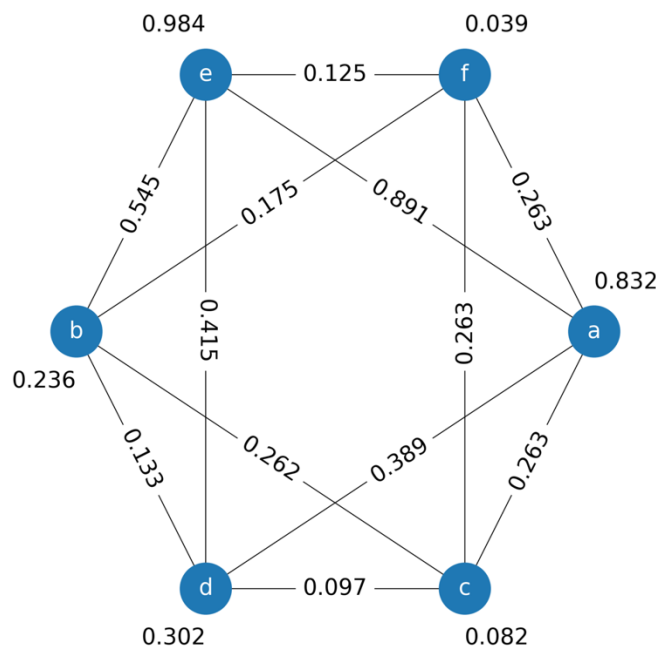
As the exact transformations required are often task-specific, the abstract *LabelGenerator* class guides users to implement their own label generation code with three necessary methods:

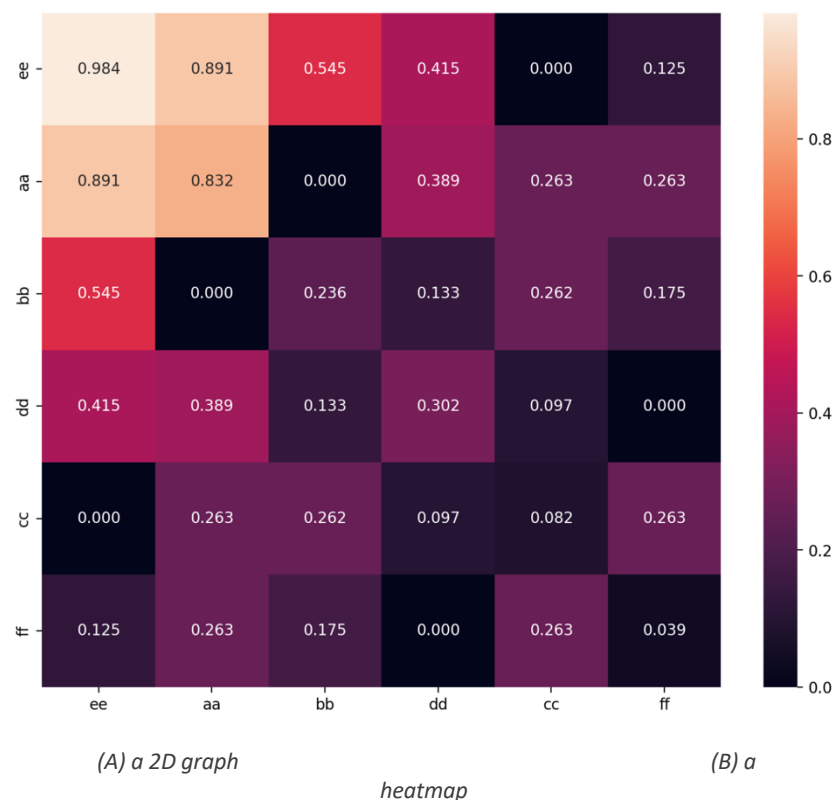
- *add_annotation_prob_labels* is used to represent each individual's raw annotations;
- *add_sample_prob_labels* is used to aggregate labels at the data point level, retaining disagreement in a soft label approach;
- *add_sample_hard_labels* aggregates the annotations into a hard label, through methods such as majority voting or taking the maximum probability label from the aggregated soft label.

For annotator agreement calculations, only *add_annotation_prob_label* must be implemented. To instantiate a class inheriting from *LabelGenerator*, the user must provide a list of annotator names and the label mapping (a dictionary where the key is the value represented in the DataFrame and the value is a numeric representation). This enables the extraction of individual annotations and their representation as a distribution across the available classes.

We provide a number of preset label generators, such as: the *DefaultLabelGenerator*, for the cases in which no special label aggregation is necessary; the *EffiLabelGenerator*, mirroring the label generation and aggregation shown in ([Cook et al., 2024](#)); the *TopicLabelGenerator*, for multi-label tasks such as topic-extraction ([Cook et al., 2025](#)); and the *OrdinalLabelGenerator*, used for ordinal annotation tasks where a number of features are labelled on a scale. With the *label_generator.from_annotations* method, the specific class inheriting from *LabelGenerator* is instantiated from the raw annotations, requiring no additional coding from the user.

Annotator Agreement & Reliability: Once labels for each annotation have been generated, inter- and intra-annotator agreement are calculated using equations introduced in ([Cook et al., 2024](#)). Annotator agreement can then be visualised in a 2D or interactive 3D graph, where each node represents an annotator and edges between annotators represent the pairwise agreement between two annotators, with the value next to each node representing an annotator's agreement with themselves. For cases with many annotators, where a graph could be unwieldy, we also provide a heatmap visualisation, where annotators are ordered by reliability; note that intra-annotator agreement is displayed on the diagonal. Examples of these visualisations are given in Figure 45.





3D Annotator Graph (Reliability as height)

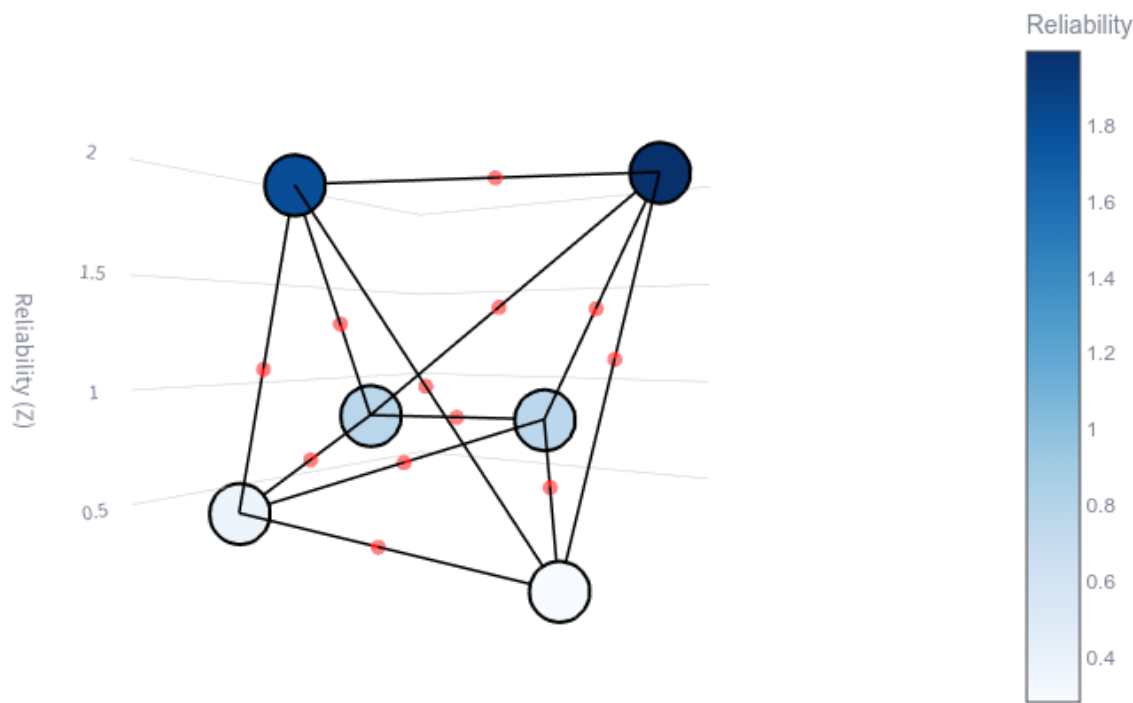


Figure 45 Example agreement visualisations as (A) a 2D graph, (B) a heatmap, and (C) a 3D graph for six annotators (annotations were synthetically generated)

Using these agreement calculations, annotator reliability can then be calculated, using a combination of an annotator's intra-annotator agreement and average inter-annotator agreement, weighted by an α parameter controlling the strength of intra-annotator agreement from 0-1. The resulting agreement values are centered around 1, enabling the recursive inter-annotator agreement calculation from (Cook et al., 2024). The reliability values can then be accessed and utilised, potentially removing certain annotators from the annotation process (Cook et al., 2025). Reliability scores may also be utilised in label aggregation (in a *LabelGenerator*) or used to weight the loss function in model training (Cook et al., 2024).

Annotator agreement and reliability is calculated and stored in the *Annotations* class. The *Annotations* class is instantiated with a pandas DataFrame representation of the dataset, a *LabelGenerator* object (which will be generated using the *LabelGenerator.from_annotations* function if no instance inheriting from *LabelGenerator* is passed), an agreement metric (defaulting to Krippendorff's alpha), an overlap threshold, and the reliability alpha.

On instantiating an *Annotations* class, the annotator graph (supported by the NetworkX library (Hagberg et al., 2008)) is initialised with each annotator equally reliable. Intra-annotator agreement is first calculated for each annotator node with the *calculate_intra_annotator_agreement* instance method, using data points each user has annotated twice themselves. Inter-annotator agreement is then calculated between each user, utilising the *overlap_threshold* to decide whether there is sufficient overlap between the two annotators to assess agreement. Here, the *pairwise_agreement* function is used as a common interface to interface with the implemented pairwise agreement metrics in the agreement module. Python modules used to handle agreement calculations include the Krippendorff library (Castro, 2017) for Krippendorff's alpha and Scikit-Learn (Pedregosa et al., 2011) for Cohen's Kappa and Fleiss' Kappa. NumPy and pandas are also used for vector calculations and manipulation of the data to obtain pair annotations.

Once agreement has been calculated among annotators and with themselves, annotator reliability is calculated with a recursive application of the annotator reliability equation until reliability values converge. To ensure convergence, the calculated reliability values are normalised to have a mean of 1 after each iteration. Annotator reliability values can then be accessed through the *get_user_reliability* and *get_reliability_dict* methods.

Inter- and intra-annotator agreement values can also be easily accessed via the graph itself using the NetworkX API and the *__getitem__* method of the *Annotations* class. The graph and heatmap agreement visualisations shown in Figure 47 utilise Matplotlib (Tosi, 2009) and Seaborn (Waskom, 2021) and they are displayed using the *display_annotator_graph* and *display_agreement_heatmap* methods respectively.

The optional *annotators* and *other_annotators* arguments for the heatmap allow a user to display the agreement between one set of users and another, with the default setting comparing all annotators to one another. This may be useful in cases where you already have a set of reliable annotators or you have a gold-standard set of annotations you would like to compare a set of annotators to.

Sample Redistribution: In cases where a consensus must be reached on a high proportion of data points, samples may be redistributed among annotators to resolve disagreement. The *SampleRedistributor* provides this functionality. It functions very similarly to the *SampleDistributor* with the additional

constraint that an annotator who has already annotated an individual data point will not be reassigned it. Sample redistribution can be done iteratively until the desired level of agreement is reached.

The *SampleRedistributor* inherits from the *SampleDistributor* class, overloading the *distribute_samples* method, applying a round-robin-style allocation using the EffiARA sample distribution variables, ensuring that annotators are not given samples they have already annotated.

Final Dataset: Once the desired level of agreement has been reached, potentially with the aim of generating gold-standard labels in classification tasks, the final dataset is ready, with annotations tied to annotator identities, allowing for training strategies that utilise the expertise and reliability of individual annotators. Users may utilise the *concat_annotations* method in the *data_generation* module for assistance in merging annotations into the final dataset.

8.3 EffiARA Webtool

To make the functionalities of the EffiARA package more accessible and quicker to use, we have also released the webtool at <https://effiara.gate.ac.uk>. The webtool allows non-technical experts to run annotation projects and gain insights into annotator agreement and reliability with ease. Even for those comfortable using the Python package, the webtool provides a convenient interface for performing tasks quickly. A system demonstration is available at <https://www.youtube.com/watch?v=KcmQfPiskcY>.

The webtool supports common tasks within the annotation pipeline (excluding the annotation step itself). Finer-grained control and more advanced functionality may be achieved with the Python package, particularly through customisation of modules like the *LabelGenerator*. As the project is open-sourced, technical users are able to make their own modifications and run them as a local web-application or make a pull request to add their additional use-cases. The webtool source code is available at <https://github.com/MiniEggz/EffiARA-webtool>.

The application contains four main workflows:

- **Sample Distribution.** This workflow handles all aspects of distributing samples from an unannotated dataset, including understanding the resources available. The *sample_id* column is added to each data point to allow recompilation after annotation.
- **Annotation Project.** This workflow is used to generate an annotation project for specific platforms. Currently, project generation for GATE Teamware 2 ([Wilby et al., 2023](#)) is supported. Future iterations may include other platforms but this task is most likely solved to some extent by the individual annotation platforms.
- **Dataset Compilation.** Once data annotation is complete, this workflow allows the user to upload a ZIP file containing all annotation CSV files. It supports users in renaming columns, moving all reannotations under the correct columns (beginning with *re_*) and into the correct row (alongside their original annotation of the data point), and merging the annotations from different annotators to create a final dataset ready for analysis.
- **Annotator Reliability.** With the compiled dataset, users can analyse annotator reliability. The user first selects their label generator and they then have full control over the label mapping or they may choose to generate it automatically using the *LabelGenerator.from_annotations* method. Users then choose the desired output: any combination of outputting annotator reliability, the annotator agreement graph (in 2D or interactive 3D) and an annotator heatmap. The workflow

also offers a number of options for calculating annotator reliability, such as the agreement metric, the reliability alpha, and the overlap threshold (the minimum number of data points annotated by both annotators to enable agreement assessment); the workflow also offers display configurations for the graphs.

The webtool is built upon the EffiARA Python package and shares the same dependencies. It is implemented using Streamlit ([Khorasani et al., 2022](#)) and Plotly ([Sievert, 2020](#)) is used to create the interactive 3D annotator agreement and reliability visualisation. The zipfile and tempfile libraries handle uploads and downloads, ensuring data is deleted once processed.

8.4 Evaluation

Two previous works involving dataset creation have annotated data following the EffiARA methodology, creating RUC-MCD ([Cook et al., 2024](#)) and the Chinese News Framing dataset ([Cook et al., 2025](#)). Both studies provide support for the annotation framework.

8.4.1 Case Studies

RUC-MCD. In the work introducing the EffiARA annotation framework ([Cook et al., 2024](#)), utilising reliability scores in the label generation and model training stages was shown to improve classification performance. Applying a soft-label approach, using TwHIN-BERT-Large, assessing reliability with inter-annotator agreement only, intra-annotator agreement only, and a combination of both all improved classification performance. Classification performance increased an F1-macro score of 0.691 to 0.740 using the EffiARA reliability scores calculated using a reliability alpha of 0.5. The dataset used in this study was of low-to-moderate agreement, highlighting the frameworks utility in datasets containing disagreement.

Chinese News Framing. This work utilises the EffiARA reliability scores to identify unreliable annotators during the annotation process, leading to an increased overall level of agreement among annotators, which is highly indicative of data quality ([Krippendorff, 2018](#)). By removing the low-reliability annotator and replacing them with an existing high-reliability annotator, the average inter-annotator agreement (measured using Krippendorff's alpha) increased from 0.396 to 0.465.

8.4.2 Load Testing

To assess the usability of the application, we also carried out load testing on the web application when hosted locally on a laptop with an Intel i7-6600U @ 3.400GHz and 16GB RAM, meaning upload and download speed were not a factor. Sample distribution remains quick and responsive for a large number of samples, taking less than a second for datasets of 100,000 samples. Dataset compilation and processing both scale roughly linearly with respect to dataset size with the tool requiring significantly longer to process datasets containing as many as 100,000 data points. Datasets containing 10,000 data points and under require less than one minute for dataset compilation and dataset processing (including annotator reliability calculation and visualisation rendering). The time taken for each key action in the webtool can be seen in Table 33. While running tasks that take longer, the web application remains responsive.

Table 33 Processing time for each stage at varying dataset sizes. Tests conducted running the webtool locally on a laptop with 16GB RAM and an i7-6600U @ 3.400GHz

Number of Samples	Sample Distribution	Dataset Compilation	Dataset Processing
500	~0.06s	~3s	~3s
1,000	~0.06s	~6s	~6s
5,000	~0.10s	~30s	~25s
10,000	~0.12s	~1m	~45s
100,000	~0.5s	~10m	~7m 20s

8.5 Concluding Remarks

In this section, we introduced the EffiARA Python package alongside an accessible web application that provides a graphical interface to the EffiARA annotation framework. EffiARA supports the design, compilation, and reliability assessment of annotation projects at the document level. Future development will focus on expanding the range of supported annotation settings, optimising computational performance, and enhancing usability based on user feedback. The package and webtool will be actively maintained, ensuring they remain usable and up-to-date with users' annotation requirements.

9 Conclusions and Challenges Ahead

In the previous sections, we have presented a significant body of work advancing the state of multimodal AI for disinformation detection and verification, building upon the foundations laid in D3.1. This deliverable has presented substantial advancements in developing trustworthy, cross-modal AI systems for misinformation analysis and content verification. Through integrated work on text-based analysis, audiovisual content understanding, image-based clue extraction, contextual integrity assessment, and user-in-the-loop systems, the methods and tools reported in D3.1 contribute to a robust and modular foundation for automated verification pipelines within the vera.ai project.

In the domain of **text-based misinformation**, we conducted a large-scale, multi-domain analysis of persuasive strategies used in disinformation narratives, uncovering both recurring and domain-specific rhetorical techniques. This insight into persuasive framing informs downstream detection models. In parallel, we developed a multilingual authorship attribution system that identifies stylistic markers across languages, showing superior performance especially for unseen authors in low-resource contexts. To better understand the intended impact of misinformation, we also introduced a novel approach for predicting the demographic profile of a news article's likely audience. This pipeline leverages synthetic data and deep multilingual embeddings to estimate user traits like gender, income, and urbanity with high accuracy.

Our efforts in **audiovisual analysis** extended the scope of verification beyond text. By analyzing sirens, indicator sounds, and other contextual audio features, we demonstrated how acoustic signals can support geolocation and content classification. Improvements in geolocation from images were achieved through network refinements, training data sanitation, and orientation normalization, resulting in better accuracy and interpretability. In parallel, we benchmarked OCR models on challenging multilingual datasets, revealing both the strengths of traditional systems and the limitations of large multimodal models, particularly regarding hallucinations and the need for domain-specific fine-tuning.

To tackle the challenge of **decontextualized media**, we developed a comprehensive detection framework that integrates provenance tracking, contextual alignment, and external evidence retrieval. New benchmarks such as VERITE and RED-DOT were introduced to evaluate multimodal claims and evidence quality, addressing prior limitations like unimodal bias. Although our models show promising results on both synthetic and real-world data, concerns around generalizability and coverage have led us to refrain from deploying a public-facing tool at this stage.

We have also made significant advances in **human-AI collaboration**. A multilingual chatbot was designed to combine intent recognition with retrieval-augmented generation, enabling explainable, interactive verification support. Evaluations across six languages showed reliable and accurate performance, while integration with other vera.ai modules, such as synthetic media detectors, highlighted its potential for real-world applications. Meanwhile, the EffiARA framework was introduced to address the critical issue of annotation quality. By offering tools for dynamic task allocation, reliability scoring, and inter-annotator agreement analysis, EffiARA improves both the scalability and reliability of dataset creation.

Despite these achievements, several **challenges** emerged that highlight the complexities of deploying multimodal AI systems in real-world disinformation contexts. One of the most persistent difficulties was

the **limited generalizability of models trained on synthetic datasets**. While such datasets enable large-scale and controlled experimentation, they often lack the linguistic noise, contextual ambiguity, and domain variability found in real-world media. This mismatch was particularly problematic in multimodal tasks, where models trained in idealized settings struggled to perform reliably when confronted with social media content containing informal language, partial or misleading context, and noisy visual or audio signals.

Another challenge arose from **modality imbalance**. Although many systems were designed to operate across multiple modalities, we observed that in practice, models frequently defaulted to exploiting the most predictive unimodal features, typically textual signals, at the expense of integrating complementary cues from images or audio. This modality collapse led to brittle decision-making and reduced the system's ability to reason holistically. We addressed this partially by introducing balanced training strategies and benchmark datasets such as VERITE, but the problem remains open, especially for tasks involving ambiguous or low-quality input signals across modalities.

In the domain of audio provenance, we encountered a **significant gap in tooling and methodologies**. While image-based provenance analysis can leverage mature technologies such as reverse image search and perceptual hashing, audio lacks equivalent infrastructure for tracking reused or manipulated segments. Challenges include segment variability, background noise, and the lack of large, labeled datasets for training and evaluation. Our initial integration of audio provenance components into the verification pipeline revealed both the potential and the fragility of current approaches, emphasizing the need for further foundational research and tool development in this area.

Our efforts to build a fact-checker-in-the-loop system via a multilingual chatbot also revealed key limitations. Although the integration of retrieval-augmented generation (RAG) and user intent detection proved effective for generating context-aware, grounded responses, we found that the system **required continuous fine-tuning to mitigate hallucinations and maintain consistency across languages**. Ensuring that generated outputs remained explainable and traceable demanded careful curation of evidence sources and ongoing system monitoring.

Finally, **ensuring annotation quality at scale proved non-trivial**. Crowdsourced annotation, while cost-effective, introduced variability in label consistency, especially in subjective or multi-label tasks like persuasion detection or evidence relevance. The EffiARA framework helped address this by providing dynamic annotator scoring and real-time visual feedback, but implementing these features required carefully defined agreement metrics and thoughtful interface design to support both annotation efficiency and reliability. Striking the right balance between cost, quality, and speed remains an ongoing concern, particularly for the large-scale dataset construction needed in real-world AI deployments.

Looking ahead, we aim to further embed explainability and ethical safeguards into our AI systems, ensuring transparency not only at the model level but also in user-facing outputs. This includes developing architectures that provide interpretable reasoning paths, as well as user interfaces that clearly communicate model confidence, evidence provenance, and decision rationale. We will also continue to design collaborative environments that allow domain experts—such as journalists and fact-checkers—to interact directly with system outputs, interrogate underlying assumptions, and feed corrections or counter-evidence back into the model lifecycle.

Expanding provenance-aware tools beyond static images to include audio and video is a key research priority. This will require not only new algorithmic techniques but also curated, multimodal benchmark datasets that reflect the complexity of real-world manipulations. In parallel, we plan to scale up dataset creation through dynamic annotation strategies, leveraging real-time feedback on annotator reliability to optimize task routing, reduce redundancy, and improve overall label quality.

While no single method can fully address the multifaceted nature of online disinformation, the integrated methodologies, tools, and insights presented in this deliverable offer a coherent and extensible foundation. By aligning technical innovation with linguistic sensitivity and human-centric design, vera.ai is taking important steps toward building robust, transparent, and socially responsible AI systems for trustworthy information verification.

References

- Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., ... & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine*, 7(1), 82.
- Abdelnabi, S., Hasan, R., & Fritz, M. (2022). Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14940-14949).
- Abdin, Y., Ahmed, R., Khan, M., & Liu, Y. (2024). Phi-3-Vision-128K-Instruct. *arXiv*.
- Abeßer, J. (2020). A review of deep learning based methods for acoustic scene classification. *Applied Sciences*, 10(6). <https://doi.org/10.3390/app10062020>
- Abeßer, J. (2022). Classifying sounds in polyphonic urban sound scenes. In *Proceedings of the 152nd Audio Engineering Society (AES) Convention*.
- Abeßer, J., Rodríguez Mejía, J. M., Cuccovillo, L., & Aichroth, P. (2024). Siren sounds as acoustic landmarks for content verification. In *Proceedings of the Annual Meeting on Acoustics (DAGA)*, Hannover, Germany.
- Abeßer, J., Schwär, S., & Müller, M. (2025). Pitch contour exploration across audio domains: A vision-based transfer learning approach. *arXiv*. <https://arxiv.org/abs/2503.19161> (submitted to the *IEEE/ACM Transaction on Audio, Speech, and Language Processing*)
- Abeßer, J. (2025). Automatic retrieval of indicator sounds for acoustic geo-tagging. In *Late-Poster at the Annual Meeting on Acoustics (DAS/DAGA 2025)*, Copenhagen, Denmark. <https://doi.org/10.24406/publica-4474>
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Albassam, W. A. (2023). The power of artificial intelligence in recruitment: An analytical review of current AI-based recruitment strategies. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, 8(6), 4.
- Al Ghadban, Y., Lu, H., Adavi, U., Sharma, A., Gara, S., Das, N., ... & Hirst, J. E. (2023). Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*, 2023-12.
- Alimoradi, Z., Ohayon, M. M., Griffiths, M. D., Lin, C. Y., & Pakpour, A. H. (2022). Fear of COVID-19 and its association with mental health-related factors: systematic review and meta-analysis. *BJPsych open*, 8(2), e73.

Aneja, S., Bregler, C., & Nießner, M. (2023, June). COSMOS: catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 12, pp. 14084-14092).

Apostolidis, K., Abeßer, J., Cuccovillo, L., & Mezaris, V. (2024). Visual and audio scene classification for detecting discrepancies in video: A baseline method and experimental protocol. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, MAD '24* (pp. 30–36). <https://doi.org/10.1145/3643491.3660287>

Atuhurra, J., Kamigaito, H., Watanabe, T., & Nichols, E. (2024). Domain Adaptation in Intent Classification Systems: A Review. *arXiv preprint arXiv:2404.14415*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7), 1–46. <https://doi.org/10.1371/journal.pone.0130140>

Baek, J., Kim, G., Lee, S., Park, H., & Han, D. (2019). What is wrong with scene text recognition model comparisons? Dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4715–4723).

Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2024). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, 29(8), 10153-10176.

Bai, Y., Zhang, Z., Wang, S., Gao, Y., Zhang, C., Wang, H., ... & Zhou, J. (2023). Qwen-VL: A versatile vision-language model. *arXiv preprint arXiv:2309.13038*.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Bear, H. L., Heittola, T., Mesaros, A., Benetos, E., & Virtanen, T. (2019). City classification from multiple real-world sound scenes. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 11–15). New Platz, New York, USA. <https://doi.org/10.1109/WASPAA.2019.8937271>

Bhargava, A., Celikyilmaz, A., Hakkani-Tür, D., & Sarikaya, R. (2013, May). Easy contextual intent prediction and slot detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8337-8341). IEEE.

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., & Kompatsiaris, Y. (2018). Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1), 71-86.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., & Gorrell, G. (2013). GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47, 1007-1029.

- Bousmaha, R., Laouedj, S., Aggoune, L., & Benslimane, S. M. (2023, October). YOLOv7-Face: A Real-Time Face Detector. In 2023 International Conference on Networking and Advanced Systems (ICNAS) (pp. 1-6). IEEE.
- Boyd, S., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press.
- Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin, 10, 1-47.
- Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. Speech communication, 33(1-2), 23-60.
- Camacho, A. and Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America, 124 (3), no. 3, 1638–1652.
- Campos Ferreira, M., Veloso, M., & Tavares, J. M. R. (2024, May). A comprehensive examination of user experience in AI-based symptom checker chatbots. In International Conference on Decision Support System Technology (pp. 98-108). Cham: Springer Nature Switzerland.
- Cao, B., Araujo, A., & Sim, J. (2020). Unifying deep local and global features for image search. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16 (pp. 726-743). Springer International Publishing.
- Cao, Z., Chen, E., Huang, Y., Shen, S., & Huang, Z. (2023, July). Learning from crowds with annotation reliability. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2103-2107).
- Cassidy, T., McDowell, B., Chambers, N., & Bethard, S. (2014, June). An annotation framework for dense event ordering. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 501-506).
- Castro, S. (2017). Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. GitHub repository.
- Ch, K., & Yuliang, L. (2017). Total-Text: A comprehensive dataset for scene text detection and recognition. In Proceedings of the IEEE International Conference on Document Analysis and Recognition (ICDAR) (pp. 935–942).
- Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. City, 1(2), 1.
- Chan, B. J., Chen, C. T., Cheng, J. H., & Huang, H. H. (2025, May). Don't do rag: When cache-augmented generation is all you need for knowledge tasks. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 893-897).

- Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE winter conference on applications of computer vision (WACV) (pp. 839-847). IEEE.
- Ch'ng, C. K., & Chan, C. S. (2017). Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 935–942). IEEE.
- Chen, Z., Wang, J., Wang, W., Chen, G., Xie, E., Luo, P., & Lu, T. (2021). Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. arXiv preprint arXiv:2111.02394.
- Chen, X., Li, L. H., Yu, C., Kuo, C. H., Huang, J., & Fei-Fei, L. (2023). VLP: Vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1234–1243).
- Chen, X., Tan, J., Wang, T., Zhang, K., Luo, W., & Cao, X. (2024). Towards real-world blind face restoration with generative diffusion prior. IEEE Transactions on Circuits and Systems for Video Technology.
- Chen, L., Zhang, R., Guo, J., Fan, Y., & Cheng, X. (2024b November). Controlling Risk of Retrieval-augmented Generation: A Counterfactual Prompting Framework. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 2380-2393).
- Chrysidis, Z., Papadopoulos, S. I., Papadopoulos, S., & Petrantonakis, P. (2024, June). Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. In Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation (pp. 73-81).
- Choi, J., Hauff, C., Van Laere, O., & Thomee, B. (2015, September). The Placing Task at MediaEval 2015. In MediaEval.
- Clark, B., Kerrigan, A., Kulkarni, P. P., Cepeda, V. V., & Shah, M. (2023). Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23182-23190)
- Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. Digital health, 9, 20552076231183542.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.
- Cook, O., Grimshaw, C., Wu, B., Dillon, S., Hicks, J., Jones, L., ... & Song, X. (2024). Efficient Annotator Reliability Assessment and Sample Weighting for Knowledge-Based Misinformation Detection on Social Media. arXiv preprint arXiv:2410.14515.
- Cook, O., Mu, Y., Yang, X., Song, X., & Bontcheva, K. (2025). A dataset for analysing news framing in Chinese media. arXiv preprint arXiv:2503.04439.

Cornolti, M., Ferragina, P., & Ciaramita, M. (2013, May). A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 249-260).

Cramer, A. L., Wu, H. H., Salamon, J., & Bello, J. P. (2019). Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3852–3856). <https://doi.org/10.1109/ICASSP.2019.8682475>

Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., ... & Silvestri, F. (2024, July). The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 719-729).

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017, July). Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, system demonstrations* (pp. 97-102).

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711*.

Damiano, S., Cramer, B., Guntoro, A., & van Waterschoot, T. (2024). Synthetic data generation techniques for training deep acoustic siren identification networks. *Frontiers in Signal Processing*, 4, 1358532. <https://doi.org/10.3389/frsip.2024.1358532>

Dai, Y., Zhang, Y., Liu, H., Ou, Z., Huang, Y., & Feng, J. (2021). Elastic crfs for open-ontology slot filling. *Applied Sciences*, 11(22), 10675.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 20-28.

Deza, E., Deza, M. M., Deza, M. M., & Deza, E. (2009). *Encyclopedia of distances* (pp. 1-583). Springer Berlin Heidelberg.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2020). Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Dimitrov, D., Alam, F., Hasanain, M., Hasnat, A., Silvestri, F., Nakov, P., & Da San Martino, G. (2024). SemEval-2024 Task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 2009–2026). Association for Computational Linguistics.

Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., ... & Wang, J. (2024). Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Dumitrache, A., Aroyo, L., & Welty, C. (2018). CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 6(1), 20–28.

Elizalde, B., Chao, G.-L., Zeng, M., & Lane, I. (2016). City-identification of Flickr videos using semantic acoustic features. In Proceedings of the IEEE Second International Conference on Multimedia Big Data (BigMM) (pp. 303–306). <https://doi.org/10.1109/BigMM.2016.28>

Fang, S., Xie, H., Wang, Y., Zhan, F., & Lu, J. (2021). Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of experimental psychology: general*, 144(5), 993.

Fonseca, E., Favory, X., Pons, J., Font, F., & Serra, X. (2022). FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 829–852. <https://doi.org/10.1109/TASLP.2021.3134365>

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378.

García-Méndez, S., De Arriba-Perez, F., González-Castaño, F. J., Regueiro-Janeiro, J. A., & Gil-Castiñeira, F. (2021). Entertainment chatbot for the digital inclusion of elderly people without abstraction capabilities. *IEEE Access*, 9, 75878–75891.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 776–780), New Orleans, LA, USA.

Gennaro, G., & Ash, E. (2022). Emotion and reason in political language. *The Economic Journal*, 132(643), 1037–1059.

Gerhardt, M., Cuccovillo, L., & Aichroth, P. (2024). Audio Provenance Analysis in Heterogeneous Media Sets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4387–4396).

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). ImageBind: One Embedding Space To Bind Them All. In CVPR.

Gkrispanis, K., Gkalelis, N., & Mezaris, V. (2024). Filter-pruning of lightweight face detectors using a geometric median criterion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 280–289).

- Glockner, M., Hou, Y., & Gurevych, I. (2022, December). Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5916-5936).
- Grollmisch, S., & Cano, E. (2021). Improving semi-supervised learning for audio classification with FixMatch. *Electronics*, 10(15). <https://doi.org/10.3390/electronics10151807>
- Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., & Cheng, M. M. (2022, October). Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision* (pp. 126-143). Cham: Springer Nature Switzerland.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 52138–52160. Retrieved from <https://doi.org/10.1145/3236009>
- Guo, X., Hou, B., Yang, C., Ma, S., Ren, B., Wang, S., & Jiao, L. (2023). Visual explanations with detailed spatial information for remote sensing image classification via channel saliency. *International Journal of Applied Earth Observation and Geoinformation*, 118, 103244.
- Guo, S., & Sanner, S. (2010). Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 833–834). ACM.
- Guttikonda, D., Indran, D., Narayanan, L., Pasarad, T., & BJ, S. (2025). Explainable AI: A Retrieval-Augmented Generation Based Framework for Model Interpretability. In *Proc. 17th Int. Conf. on Agents and Artificial Intelligence* (Vol. 3, pp. 948-955).
- Haas, L., Skreta, M., Alberti, S., & Finn, C. (2024). Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12893-12902).
- Haas, L., Alberti, S., & Skreta, M. (2023). Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275*
- Hagberg, A. A., Swart, P. J., & Schult, D. A. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)* (pp. 11–15).
- Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). Disinformation detection about Islamic issues on social media using deep learning techniques. *Malaysian Journal of Computer Science*, 36(3), 242-270.
- Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, 29(2), 147-160.
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior*, 16(1), 107-112.

- Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017, August). Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings* (pp. 1803–1812). <https://doi.org/10.1145/3097983.3098131>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.
- Hays, J., & Efros, A. A. (2008, June). Im2gps: estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.
- He, J., Shi, W., Chen, K., Fu, L., & Dong, C. (2022). Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1889-1898).
- Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., ... & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *Nejm Ai*, 2(4), Aloa2400802.
- Heittola, T., Mesaros, A., & Virtanen, T. (2020). TAU Urban Acoustic Scenes 2020 Mobile, Development Dataset. <https://doi.org/10.5281/zenodo.3670167>
- Hofstätter, S., Chen, J., Raman, K., & Zamani, H. (2023, July). Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1437-1447).
- Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., & Hovy, E. (2013, June). Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1120-1130).
- Hsu, C. Y., & Li, W. (2023). Explainable GeoAI: can saliency maps help interpret artificial intelligence's learning process? An empirical study on natural feature detection. *International Journal of Geographical Information Science*, 37(5), 963-987.
- Hsueh, P. Y., Melville, P., & Sindhvani, V. (2009, June). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing* (pp. 27-35).
- Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., ... & Zhou, J. (2024, November). mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 3096-3120).
- Hwang, G. J., & Chang, C. Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 31(7), 4099-4112.

Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., & Aroyo, L. (2014). CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *The Semantic Web – ISWC 2014* (pp. 486–504). Springer.

Izacard, G., & Grave, É. (2021, April). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 874-880).

Izbicki, M., Papalexakis, E. E., & Tsotras, V. J. (2020). Exploiting the earth’s spherical geometry to geolocate images. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II* (pp. 3-19). Springer International Publishing.

Jeong, J., Kim, B., Yu, J., & Yoo, Y. (2024). EResFD: Rediscovery of the effectiveness of standard convolution for lightweight face detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 988-998).

Jia, P., Liu, Y., Li, X., Wang, Y., Du, Y., Han, X., ... & Zhao, X. (2024). G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *arXiv preprint arXiv:2405.14702*.

Jiang, P. T., Zhang, C. B., Hou, Q., Cheng, M. M., & Wei, Y. (2021). Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30, 5875-5888.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Singh Chaplot, D., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). *Mistral 7B* [Preprint]. *arXiv*.

Joshi, R., Graefe, J., Kraus, M., & Bengler, K. (2024). Exploring the impact of explainability on trust and acceptance of conversational agents—A Wizard of Oz study. *International Conference on Human-Computer Interaction*, 199–218.

Kachuee, M., Yuan, H., Kim, Y., & Lee, S. (2020). Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents. *arXiv preprint arXiv:2010.11230*.

Khorasani, M., Abdou, M., & Hernández Fernández, J. (2022). *Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework*. Apress.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1), 61-70.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2021). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2880–2894.

- Kordopatis-Zilos, G., Galopoulos, P., Papadopoulos, S., & Kompatsiaris, I. (2021, August). Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In Proceedings of the 2021 International Conference on Multimedia Retrieval (pp. 155-163).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
- Kuang, Z., Sun, Y., Lin, J., Zhang, R., & Ouyang, W. (2021). MMOCR: Openmmlab text detection, recognition and understanding toolbox.
- Kumar, A., Elizalde, B., & Raj, B. (2017). Audio content based geotagging in multimedia. In Proceedings of the INTERSPEECH Conference (pp. 1874–1878)
- Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., & Matas, J. (2018). Deblurgan: Blind motion deblurring using conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8183-8192).
- Kurata, G., Xiang, B., Zhou, B., & Yu, M. (2016, November). Leveraging Sentence-level Information with Encoder LSTM for Semantic Slot Filling. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2077-2083).
- Laurencon, J., Dupont, Y., & Garcia, R. (2024). Idefics: A multimodal model for understanding images and text. arXiv preprint arXiv:2401.12345.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
- Lei, H., Choi, J., & Friedland, G. (2012). Multimodal city-verification on Flickr videos using acoustic and textual features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2273–2276). <https://doi.org/10.1109/ICASSP.2012.6288367>
- Leite, J. A., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2024, October). EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 5380-5384).
- Lee, J., Park, S., Baek, J., Oh, S. J., Kim, S., & Lee, H. (2020). On recognizing texts of arbitrary shapes with 2D self-attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 546-547).
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

- Lee, S., Seong, H., Lee, S., & Kim, E. (2022). Correlation verification for image retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5374-5384).
- Li, Y., Shibata, H., & Takama, Y. (2019, November). Chatbot-mediated personal daily context modeling upon user story graph. In 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 1-6). IEEE.
- Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of marketing research*, 57(1), 1-19.
- Li, H., Zhang, C., Xu, Y., Cui, L., Wang, X., & Wei, F. (2022). PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system. *arXiv*.
- Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., ... & Bai, X. (2024). Monkey: Image resolution and text label are important things for large multi-modal models. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 26763-26773).
- Liao, M., Zou, Z., Wan, Z., Yao, C., & Bai, X. (2022). Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(1), 919-931.
- Lin, B. Y., Lee, D. H., Xu, F. F., Lan, O., & Ren, X. (2019, January). AlpacaTag: An active learning-based crowd annotation framework for sequence tagging. In Proceedings of the 57th Conference of the Association for Computational Linguistics.
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31-57.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Liu, H., Zhao, S., Zhang, X., Zhang, F., Sun, J., Yu, H., & Zhang, X. (2022, July). A simple meta-learning paradigm for zero-shot intent classification with mixture attention mechanism. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2047-2052).
- Liu, Y., Zhang, X., Wang, Y., & Li, H. (2024). Visual instruction tuning for multimodal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5678–5687).
- Liu, Y., Ding, J., Deng, G., Li, Y., Zhang, T., Sun, W., ... & Liu, Y. (2024a). Image-Based Geolocation Using Large Vision-Language Models. *arXiv preprint arXiv:2408.09474*.
- Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024b). Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 26296-26306).

Long, S., Ruan, J., Zhang, W., He, X., Wu, W., & Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In Proceedings of the European conference on computer vision (ECCV) (pp. 20-36).

Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., & Raptis, M. (2022). Towards end-to-end unified scene text detection and layout analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1049–1059).

Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., & Raptis, M. (2023). ICDAR 2023 competition on hierarchical text detection and recognition. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR).

Lu, S., Xie, H., Zhan, F., & Lu, J. (2021). MASTER: Multi-aspect non-local network for scene text recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Lücking, A., Driller, C., Stoeckel, M., Abrami, G., Pachzelt, A., & Mehler, A. (2022). Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. *Language resources and evaluation*, 56(3), 807-855.

Luo, G., Darrell, T., & Rohrbach, A. (2021, November). NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6801-6817).

Mavroudis, V. (2024). LangChain.

Mackenzie, J., Benham, R., Petri, M., Trippas, J. R., Culpepper, J. S., & Moffat, A. (2020, October). CC-News-En: A large English news corpus. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 3077-3084).

Marchegiani, L., & Newman, P. (2022). Listening for sirens: Locating and classifying acoustic alarms in city scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 17087–17096. <https://doi.org/10.1109/TITS.2022.3158076>

McKinney, W. (2011). pandas: A foundational Python library for data analysis and statistics. In *Python for High Performance and Scientific Computing*, 14(9), 1–9.

Mei, X., Liu, H., Kong, Q., Ko, T., Plumbley, M. D., Zou, Y., & Wang, W. (2024). WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 3339-3354. IEEE.

Meireles, O. R., Rosman, G., Altieri, M. S., Carin, L., Hager, G., Madani, A., ... & SAGES Video Annotation for AI Working Groups. (2021). SAGES consensus recommendations on an annotation framework for surgical video. *Surgical endoscopy*, 35(9), 4918-4929.

- Mendoza, M., & Zamora, J. (2009, August). Identifying the intent of a user query using support vector machines. In *International symposium on string processing and information retrieval* (pp. 131-142). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Meng, H., Xin, Z., Liu, T., Wang, Z., Feng, H., Lin, B., ... & Sui, Z. (2022). Dialogusr: Complex dialogue utterance splitting and reformulation for multiple intent detection. *arXiv preprint arXiv:2210.11279*.
- Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(1), 67–83.
- Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... & Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., ... & Granger, B. E. (2017). SymPy: Symbolic computing in Python. *PeerJ Computer Science*, 3, e103.
- Miller, T., & Jing, Z. (2024). Explanation in artificial intelligence: Insights from the social sciences. *Digital Humanities Research*, 4(2), 90.
- Minkowski, H. (1910). *Geometrie der zahlen* (Vol. 1). BG Teubner.
- Mishra, A., Alahari, K., & Jawahar, C. V. (2019). OCR: A brief overview and future directions. *Pattern Recognition Letters*, 123, 1–10.
- Moradizyveh, S. (2022). Intent recognition in conversational recommender systems. *arXiv preprint arXiv:2212.03721*.
- Mu, M., Das Bhattacharjee, S., & Yuan, J. (2023). Self-supervised distilled learning for multi-modal misinformation identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2819-2828).
- Muller-Budack, E., Pustu-Iren, K., & Ewerth, R. (2018). Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 563-579).
- Nah, S., Hyun Kim, T., & Mu Lee, K. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3883-3891).
- Nakamura, K., Levy, S., & Wang, W. Y. (2020, May). Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6149-6157).
- Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., ... & Chazalon, J. (2017). ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification—RRC-MLT. In **2017 14th*

IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1454–1459). IEEE.

Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khlif, W., ... & Liu, C.-L. (2019). ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1582–1587). IEEE.

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19, 969-974.

Norouzi, M., Fleet, D. J., & Salakhutdinov, R. R. (2012). Hamming distance metric learning. *Advances in neural information processing systems*, 25.

Nowak, S., & Rüger, S. (2010, March). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval* (pp. 557-566).

Nowak, S., & Rüger, S. (2010, March). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval* (pp. 557-566).

Hromadka, T., Smolen, T., Remis, T., Pecher, B., & Srba, I. (2023). KInITVeraAI at SemEval-2023 Task 3: Simple yet Powerful Multilingual Fine-Tuning for Persuasion Techniques Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 629–637, Toronto, Canada. Association for Computational Linguistics.

Oliphant, T. E. (2006). *A Guide to NumPy*. Trelgol Publishing.

Orhan, U., Tosun, E.G. & Ozkaya, O. Intent Detection Using Contextualized Deep SemSpace. *Arab J Sci Eng* 48, 2009–2020 (2023). <https://doi.org/10.1007/s13369-022-07016-9>

Pan, A., Musheyev, D., Bockelman, D., Loeb, S., & Kabarriti, A. E. (2023). Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA oncology*, 9(10), 1437-1440.

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. (2023, June). Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation* (pp. 36-44).

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2024). `VERITE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1), 4.

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025). Red-dot: Multimodal fact-checking via relevant evidence detection. *IEEE Transactions on Computational Social Systems*.

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025b). Similarity over Factuality: Are we making progress on multimodal out-of-context misinformation detection?. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 5041-5050). IEEE.

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025c). Latent Multimodal Reconstruction for Misinformation Detection. arXiv preprint arXiv:2504.06010.

Park, D., Yuan, H., Kim, D., Zhang, Y., Matsoukas, S., & Kim, Y. (2020). Large-scale hybrid approach for predicting user satisfaction with conversational agents. arXiv preprint arXiv:2006.07113.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1 (pp. 21-29). Springer International Publishing.

Pawlik, L. (2025). How the Choice of LLM and Prompt Engineering Affects Chatbot Effectiveness. *Electronics*, 14(5), 888.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Pires, T., Schlinger, E., & Garrette, D. (2019, July). How Multilingual is Multilingual BERT?. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4996-5001).

Piskorski, J., Stefanovitch, N., Da San Martino, G., & Nakov, P. (2023, July). Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 2343-2361).

Pramanick, S., Nowara, E. M., Gleason, J., Castillo, C. D., & Chellappa, R. (2022, October). Where in the world is this image? transformer-based geo-localization in the wild. In European Conference on Computer Vision (pp. 196-215). Cham: Springer Nature Switzerland.

Prathiksha, K., Malar, S. T., Nivedha, P., Divya, D. M., & Srijayanthi, S. (2024, August). IntelliAid: A Personal Assistant Chat-Bot for Enhanced Task Management. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 432-436). IEEE.

Proctor, R. N., & Schiebinger, L. (2008). Agnotology: The making and unmaking of ignorance.

Qi, P., Yan, Z., Hsu, W., & Lee, M. L. (2024). Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13052-13062).

Qwen Team (2024). Qwen2 technical report. arXiv preprint arXiv:2412.15115.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language

supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (pp. 8748–8763).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.

Razuvayevskaya, O., Wu, B., Leite, J. A., Heppell, F., Srba, I., Scarton, C., ... & Song, X. (2024). Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *Plos one*, 19(5), e0301738.

Reynolds, A., & Corrigan, F. (2024). Improving real-time knowledge retrieval in large language models with a dns-style hierarchical query rag. *Authorea Preprints*.

Ross, T. Y., & Dollár, G. K. H. P. (2017, July). Focal loss for dense object detection. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2980-2988).

Rouard, S., Massa, F., & Défossez, A. (2023, June). Hybrid transformers for music source separation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Rukadikar, A., & Khandelwal, K. (2024). Navigating change: a qualitative exploration of chatbot adoption in recruitment. *Cogent Business & Management*, 11(1), 2345759.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4510–4520).

Sarikaya, R., Celikyilmaz, A., Deoras, A., & Jeong, M. (2014, September). Shrinkage based features for slot tagging with conditional random fields. In INTERSPEECH (pp. 268-272).

Seo, P. H., Weyand, T., Sim, J., & Han, B. (2018). Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 536-551).

Setty, S., Thakkar, H., Lee, A., Chung, E., & Vidra, N. (2024). Improving Retrieval for RAG based Question Answering Models on Financial Documents (No. 2404.07221).

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, 321-325.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. CRC Press.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*.

Srba, I., Razuvayevskaya, O., Leite, J., A., Moro, R., Schlicht, I. B., Tonelli, S., García, F. M., Lottmann, S. B., Teyssou, D., Porcellini, V., Scarton, C., Bontcheva, K., Bielikova, M. (2024). A Survey on Automatic Credibility Assessment of Textual Credibility Signals in the Era of Large Language Models. doi:10.48550/arXiv.2410.21360 arXiv:2410.21360.

Schafer, R. M. (1994). *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books.

Schlicht, I. B., Khellaf, L., & Altiok, D. (2023). Dwreco at CheckThat! 2023: enhancing subjectivity detection through style-based data sampling. *arXiv preprint arXiv:2307.03550*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

Shao, S., Chen, K., Karpur, A., Cui, Q., Araujo, A., & Cao, B. (2023). Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 11036-11046).

Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), 2298–2304.

Shi, Y., Zi, X., Shi, Z., Zhang, H., Wu, Q., & Xu, M. (2024). Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems. In *ECAI 2024* (pp. 2258–2265). IOS Press.

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. CRC Press.

Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 2, pp. 629–633). IEEE.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33, 16857-16867.

Sun, X., Zheng, H., & Tang, Z. (2022, March). Historical Information-Based Intent Detection for Multiturn Dialogue. In Proceedings of the 8th International Conference on Computing and Artificial Intelligence (pp. 566-572).

Tang, L., Laban, P., & Durrett, G. (2024). Minicheck: Efficient fact-checking of llms on grounding documents. arXiv preprint arXiv:2404.10774.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... & Hashimoto, T. B. (2023). Stanford alpaca: an instruction-following llama model (2023).

Tasneem, S., & Islam, K. A. (2023, October). Development of Trustable Deep Learning Model in Remote Sensing through Explainable-AI Method Selection. In 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 0261-0267). IEEE.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., ... & Li, L. J. (2016). Yfcc100m: The new data in multimedia research. Communications of the ACM, 59(2), 64-73

Tosi, S. (2009). Matplotlib for Python Developers. Packt Publishing Ltd.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Scao, T. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Toxtli, C., Monroy-Hernández, A., & Cranshaw, J. (2018, April). Understanding chatbot-mediated task management. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-6).

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). Maxim: Multi-axis mlp for image processing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5769-5780).

Tur, G., & De Mori, R. (2011). Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons.

Ustalov, D., Pavlichenko, N., & Tseitlin, B. (2024). Learning from Crowds with Crowd-Kit. Journal of Open Source Software, 9(96), 6227.

Vakili Tahami, A., Ghajar, K., & Shakery, A. (2020, July). Distilling knowledge for fast retrieval-based chatbots. In Proceedings of the 43rd International ACM SIGIR conference on research and development in information retrieval (pp. 2081-2084).

Vivanco Cepeda, V., Nayak, G. K., & Shah, M. (2024). Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems, 36.

Vo, N., Jacobs, N., & Hays, J. (2017). Revisiting im2gps in the deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 2621-2630).

- Vu, T. L., Tun, K. Z., Eng-Siong, C., & Banchs, R. E. (2021, March). Online faq chatbot for customer support. In *Increasing naturalness and flexibility in spoken dialogue interaction: 10th international workshop on spoken dialogue systems* (pp. 251-259). Singapore: Springer Singapore.
- Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 24-25).
- Wang, S., Mesaros, A., Heittola, T., & Virtanen, T. (2021). A curated dataset of urban scenes for audio-visual scene analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/ICASSP39728.2021.9415085>
- Wang, K., Babenko, B., & Belongie, S. (2011b). End-to-end scene text recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1457–1464).
- Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ... & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Wang, Z., Zhang, J., Chen, T., Wang, W., & Luo, P. (2023). RestoreFormer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12), 15462-15476.
- Waskom, M. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). “Let me explain!”: Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 15(2), 87–98.
- Weyand, T., Kostrikov, I., & Philbin, J. (2016). Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14* (pp. 37-55). Springer International Publishing.
- Whitehill, J., Wu, T. F., Bergsma, J., Movellan, J., & Ruvolo, P. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22.
- Wilby, D., Karmakharm, T., Roberts, I., Song, X., & Bontcheva, K. (2023, May). GATE Teamware 2: An open-source tool for collaborative document classification annotation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 145-151).
- Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., ... & Fleisch, B. (2024, June). CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In *International Conference on Case-Based Reasoning* (pp. 445-460). Cham: Springer Nature Switzerland.

- Wu, B., Razuvayevskaya, O., Heppell, F., Leite, J. A., Scarton, C., Bontcheva, K., & Song, X. (2023a). SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1995–2008, Toronto, Canada. Association for Computational Linguistics.
- Wu, B., Li, Y., Mu, Y., Scarton, C., Bontcheva, K., & Song, X. (2023b, December). Don't waste a single annotation: improving single-label classifiers through soft labels. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5347-5355).
- Xu, Q., Chen, H., Du, H., Zhang, H., Łukasik, S., Zhu, T., & Yu, X. (2024). M3A: A multimodal misinformation dataset for media authenticity analysis. *Computer Vision and Image Understanding*, 249, 104205.
- Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., ... & Sun, M. (2024). MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.
- Yang, R., Pu, F., Xu, Z., Ding, C., & Xu, X. (2021). DA2Net: Distraction-attention-driven adversarial network for robust remote sensing image scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5.
- Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024). Financial report chunking for effective retrieval augmented generation. arXiv preprint arXiv:2402.05131.
- Yılmaz, G., & Şahin-Yılmaz, A. (2024). An overview of chatbots in tourism and hospitality using bibliometric and thematic content analysis. *Worldwide Hospitality and Tourism Themes*, 16(2), 232-247.
- Yin, X.-C., Zuo, Z.-Y., Tian, S., & Liu, C.-L. (2016). Text detection, tracking and recognition in video: A comprehensive survey. *IEEE Transactions on Image Processing, 25*(6), 2752–2773.
- Yuan, X., Guo, J., Qiu, W., Huang, Z., & Li, S. (2023, December). Support or Refute: Analyzing the Stance of Evidence to Detect Out-of-Context Mis- and Disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 4268-4280).
- Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. arXiv preprint arXiv:2402.10828.
- Yuliang, L., Lianwen, J., Shuaitao, Z., & Sheng, Z. (2017). Detecting curve text in the wild: New dataset and method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4977–4986).
- Zhang, Q., Man, D., & Yang, W. (2009, November). Using HMM for intent recognition in cyber security situation awareness. In *2009 Second International Symposium on Knowledge Acquisition and Modeling* (Vol. 2, pp. 166-169). IEEE.

Zhang, N., Huang, Q., Xia, X., Zou, Y., Lo, D., & Xing, Z. (2020). Chatbot4qr: Interactive query refinement for technical question retrieval. *IEEE Transactions on Software Engineering*, 48(4), 1185-1211.

Zhang, Y., Zhang, H., Zhan, L. M., Wu, X. M., & Lam, A. (2022, May). New Intent Discovery with Pre-training and Contrastive Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 256-269).

Zhang, F., Liu, J., Zhang, Q., Sun, E., Xie, J., & Zha, Z. J. (2023, October). Ecenet: Explainable and context-enhanced network for multi-modal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 1231-1240).

Zeng, Y. X., Hsieh, J. W., Li, X., & Chang, M. C. (2023). MixNet: toward accurate detection of challenging scene text in the wild. *arXiv preprint arXiv:2308.12817*.

Zhu, S., Shah, M., & Chen, C. (2022). Transgeo: Transformer is all you need for cross-view image geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1162-1171).

Zhou, S., Chan, K., Li, C., & Loy, C. C. (2022). Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35, 30599-30611.

Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., ... & Mai, G. (2024, July). Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2749-2754).

Zhou, C., Han, S., Zhang, S., Zhou, Y., Zhang, W., & Jin, C. (2025). Efficient fine-tuning of quantized models via adaptive rank and bitwidth. *arXiv preprint arXiv:2505.03802*.

Zlatkova, D., Nakov, P., & Koychev, I. (2019, November). Fact-Checking Meets Fauxtography: Verifying Claims About Images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2099-2108).

Annex I: Text Misinformation Analysis and Verification

Domain-Specific Use of Persuasion Techniques (RQ1):

In disinformation on Islamic issues, *Name Calling-Labeling*, *Causal Oversimplification*, *Loaded Language*, *False-Dilemma-No Choice*, and *Repetition* are the most prevalent techniques, with ORs of 1.72, 1.97, 2.07, 3.33, and 10.15 respectively. *Repetition* capitalises on the illusory truth effect, a psychological phenomenon where repeated statements are more likely to be perceived as true ([Hasher & Goldstein, 1977](#); [Fazio et al., 2015](#)). In this domain, **repetition is often used to associate events and/or groups**: “*All Muslim terrorists kill for the same reason the Saudi terrorists did on 9/11, the same reason ISIS killed [...], the same reason Boko Haram is killing [...], the same reason Muhammad waged wars [...]*”. **False Dilemma** forces binary choices to portray Islam as inherently extremist, as in “*There’s no middle ground, nothing like moderate Islam.*”

In the context of COVID-19 disinformation, *Loaded Language*, *Appeal to Fear-Prejudice*, and *Slogans* appear significantly more frequently (ORs of 1.14, 1.90, 2.47, respectively). *Slogans* use short and memorable phrases to convey key ideas, as in “*We need to open up, our lives depend on it*”. Both *Loaded Language* and *Appeal to Fear-Prejudice* are designed to evoke strong emotional responses, which can bypass rational analysis and promote the acceptance of false information ([Gennaro et al., 2022](#)). **Fear is linked to both the virus**—“*#coronavirus this is extremely scary and terrifying*”—**and vaccines**, as in “*According to Bill Gates the COVID-19 RNA vaccine will permanently alter our DNA*”. Notably, fear of COVID-19 has been linked to mental health issues during the pandemic period ([Alimoradi et al., 2022](#)).

Climate change disinformation uses more frequently *Doubt* (1.59), *Conversation Killer* (2.06), *Exaggeration-Minimisation* (3.39), and *Appeal to Authority* (11.37) than the rest. *Appeal to Authority* exploits trust in credible entities like the IPCC: “*Latest IPCC Reports show global temperature forecasts exceeded actual readings.*” References to **NASA** and the **UN** are also common. **Exaggeration-Minimisation** downplays the urgency of climate change: “*In the past, warming has never been a threat to life on Earth*”. **Conversation Killer** dismisses the debate with absolute certainty, leaving no room for debate: “*Sea level rise is not going to happen.*”

Disinformation about the Russo-Ukrainian war prominently features *Flag Waving* (6.99), *Questioning the Reputation* (8.06), *Guilt by Association* (9.45), and *Appeal to Hypocrisy* (19.28). These techniques target Western and Ukrainian credibility. **Appeal to Hypocrisy** highlights perceived inconsistencies, e.g., “*They said one thing and did another.*”—referring to NATO's territorial advancement. **Guilt by Association** links entities to controversial groups, with 28% of tagged sentences referencing the word ‘Nazi’, e.g., “*Nazi symbolism is actively utilised in their daily life*”—referring to Ukrainian soldiers. *Questioning the Reputation* undermines trust in political entities: “*As for Europe, it lost its political independence after World War II*”. **Flag Waving** invokes patriotism to justify Russia's actions, e.g., “*It is against this Evil that our soldiers bravely fight side by side*”.

Contextual Domain Adaptation (RQ2):

To investigate the contextual adaptation of persuasion techniques to specific linguistic, cultural, and psychological patterns across disinformation domains, we compute the correlation between LIWC features ([Boyd et al., 2022](#)) and persuasion techniques. Due to space constraints, we limit our analysis to

the climate change dataset only, focusing on the four persuasion techniques that occur disproportionately in this context: *Appeal to Authority*, *Conversation Killer*, *Doubt*, and *Exaggeration-Minimisation*. Figure 48 presents the ten highest correlations between these techniques and the LIWC features. We compare these correlations across the other three domains to highlight patterns unique to climate change disinformation, focusing on features highly correlated within climate change but not within other domains. We provide the results for the other domains in the supplementary material.

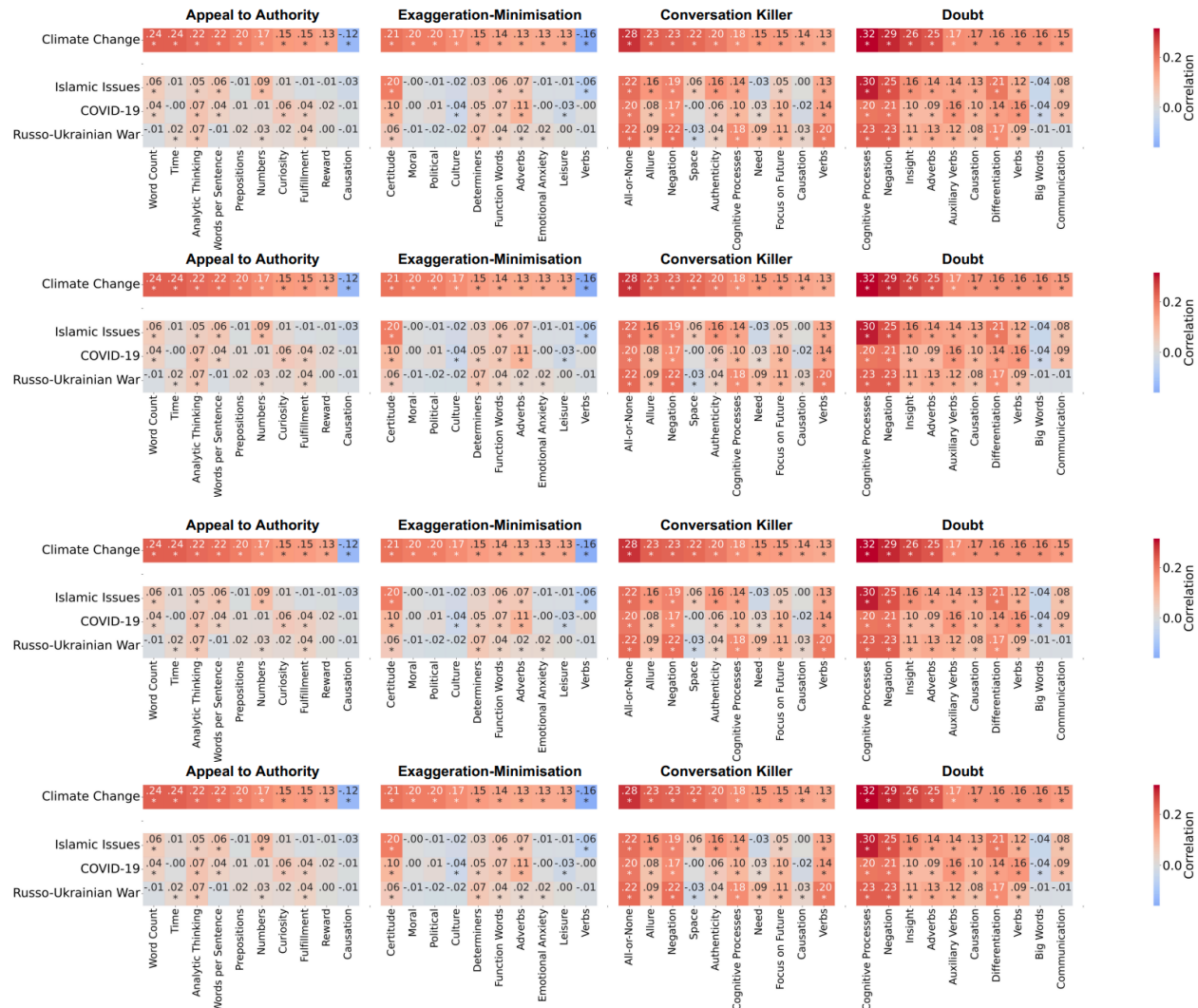


Figure AI- 1 Top 10 most correlated LIWC features for each of the four domain-specific PTs in climate change compared to other domains. Statistically significant ($p < 0.05$) coefficients are indicated with an asterisk (*).

Appeal to Authority in climate change disinformation uses distinct linguistic and psychological features to emphasise credibility and logic. These texts have a higher overall and per-sentence word count, creating longer, more complex sentences that convey authority. A high Analytic score reflects formal, logical language, reinforcing a well-reasoned tone. Frequent use of **prepositions** and **numbers** adds detail to enhance the legitimacy of claims. **Temporal markers** like 'when' and 'now' situate arguments in time, adding urgency or inevitability. Psychologically, this technique appeals to **curiosity** and **reward**, engaging readers intellectually and offering positive outcomes. Terms like 'enough' and 'full' (i.e., **fulfillment**)

suggest solutions framed as complete and authoritative. Notably, these texts avoid **causation** language (e.g., 'because', 'why'), favouring definitive statements over explanations. These patterns align with the rhetorical context of climate change narratives, where the discussion of scientific topics requires a tone of credibility and rigour to persuade audiences.

Exaggeration-Minimisation employs specific linguistic and psychological features to amplify or downplay issues strategically. A high correlation with **certitude** words (e.g., 'really', 'of course') reinforces claims with a tone of absolute confidence, making arguments appear definitive and irrefutable. Terms related to **moral behaviour** (e.g., 'honour', 'deserve') inject ethical undertones, framing the issue as one of right versus wrong. Cultural references, particularly related to **politics** (e.g., 'govern', 'congress') and broader **culture** (e.g., 'american', 'chinese'), ground the discussion in societal and geopolitical contexts, often linking climate change to governance or national responsibility. Linguistically, the technique relies heavily on **determiners** (e.g., 'the', 'that') and **function words** (e.g., 'to', 'I'), creating a conversational and relatable tone. The frequent use of **adverbs** (e.g., 'so', 'just') adds nuance or emphasis to descriptions, subtly influencing how issues are perceived. Emotional cues, especially those tied to **anxiety** (e.g., 'worry', 'fear'), heighten the sense of urgency or concern, tapping into the audience's fears about climate change. Together, these features amplify certainty, invoke morality, embed discussions within cultural and political frameworks, and simplifies complex issues while evoking emotional responses.

Both *Conversation Killer* and *Doubt* exhibit a lesser degree of contextual adaptation, as most of the LIWC features that display high correlation in climate change, are also highly correlated in other domains (e.g., Cognitive Processes, Negation). Nevertheless, some domain-specific features are present. *Conversation Killer* leverages **spatial context (Space)** (e.g., 'in', 'there') to ground arguments in specific locations, **need-related states** (e.g., 'need', 'have to') to emphasise urgency and necessity, and **causation** (e.g., 'because', 'why') provides justifications that reinforce the authority of dismissive rhetoric. Similarly, *Doubt* amplifies skepticism with the prominent use of **long words** to create an impression of sophistication and communication features (e.g., 'say', 'tell') that reference ambiguous sources, subtly eroding trust in credible information.

Annex II: Audiovisual Content Analysis and Enhancement

Table All- 1 Overview of User Feedback, Source, and Actions Taken for the KSE Service

Issue Category	User Feedback	Source	Action Taken
UI / UX	Split opinions on timeline: some find it “good” (especially that yellow bar highlights change in shots/cuts) others find it to be “improvable”, with comments similar to “Timeline not clear”. “The timeline should indicate shot changes”.	PE, ATC	Redesigned timeline for clarity; made hovering over a keyframe highlights its position in the timeline more clear - see also Section 4.2.4.7.
UI / UX	“Not understood error messages”.	PE, ATC	Provided friendlier, descriptive error/log messages.
UI / UX	“Caching message (not understood)”, “stuck on ‘Caching’ message”.	PE	Replaced vague “Caching” messages with more informative status updates.
UI / UX	“Clock saying that it’s working”, “SeeWhen uploading content, indicate how much of the process has been done or is left to do stuck in a specific stage”.	PE, ATC	Applied denser backend status polling and added animated progress indicators. Add percentage indicators during caching.
UI / UX Feature Request	“Can I collectively download the extracted keyframes?”, Request for “Download keyframes as a zip” button, Request for “Download faces as a zip” button, Request for “Download text as a zip” button.	PE	Added buttons for downloading keyframes, faces, text images packaged as ZIP. Also, see Section 4.2.4.7.
UI / UX	“Work with an URL that is send to user by email that can simply be pasted in the search bar and lead to results”, “URL to share the results”, Need for explanation on how to use the 24-digit session ID.	PE	After analysis completion, implemented the display of a link instead of a string - see also Section 4.2.4.7.
UI / UX	“Click on “open keyframe in viewer” does not work on Mac (Chrome and Safari). You can right click, and choose “open link in a new tab” to make it work, but...”.	PE, WP2	Fixed link-opening behavior and tested on various browsers
UI / UX Bug	“In the keyframe viewer, there is a padding-left issue. It’s really on the border of the screen.”	WP2	The underlying problem was identified and corrected.

UI / UX Feature Request	Request to introduce a tool to zoom into the keyframe.	PE, ATC	Already implemented - the keyframe viewer had a problem and was not opening during the testing session.
UI / UX Bug	"...it appears that the text of the logo is too close to the left border of the keyframe". [keyframe ares not visible in the keyframe viewer]	WP2	The underlying problem was identified and corrected.
Upload & Format Support	Issues with video uploads.	PE	Improved feedback and error logging during upload/caching - also see, Section 4.2.4.10.
Upload & Format Support	"I had an issue submitting a local file with *.mov extension" "Support of more extensions"	PE	Added server-side conversion via FFMPEG to MP4 to broaden accepted input formats.
Upload & Format Support	Add a feature for uploads from mobile devices.	PE	Ensured mobile upload works - tested on common mobile browsers to optimize layout and feedback.
Backend integration	Device a convenient way to monitor the session queue and check whether it is stuck.	AFP	Exposed an endpoint for real-time monitoring of session queuing statistics (total processed sessions, average and range of waiting times over intervals expressed in 24, 72, 120, and 240 hours),
Backend integration Feature Request	"Support for local file uploads via backend calls".	AFP	Implemented a backend endpoint for server-side uploads - see Section 4.2.4.8.
Backend integration	Resubmitting a session causes the re-analysis of the video.	AFP	To generate session-specific identifiers, we hash the video source string (either the URL or filename). This allows subsequent requests for the same content to immediately return cached results.
Keyframe extraction	Need for a larger set of keyframes.	AFP	See Section 4.2.4.2.
Keyframe extraction	"speed up the provision of keyframes in the frontend".	AFP	See Section 4.2.4.7.
Keyframe extraction	"suboptimal keyframe quality".	AFP	See Section 4.2.4.2.

Keyframe extraction Bug	Missing shots / frames on KSE.	AFP	Carefully adjusted threshold in the blurry frames detection algorithm.
Bug	All requests for new sessions regarding videos from YouTube failed with “video retrieval failed” message	AFP	See Section 4.2.4.10.
Keyframe extraction	Faster extraction of keyframes.	WP2	See Sections 4.2.4.1, 4.2.4.7.
Keyframe extraction	Continued concerns over processing speed.	AFP	Optimize processing pipeline. Also, see Section 4.2.4.2.
Face Detection and Enhancement	“Face hallucinations on enhanced faces”.	WP2	See Section 4.2.4.4.
Face Detection and Enhancement	“for the face detected, it looks like a more blurred or lower quality face than the ones inside the video itself”.	WP2	See Sections 4.2.4.2, 4.2.4.3.
Face Detection and Enhancement	“FD not working properly”, “need for better facial recognition in profile views”, “Frontal faces better recognized than profile faces”.	PE	Used more robust face-detection models - see Section 4.2.4.3.
Face Detection and Enhancement	“Face detection not helpful with identifying faces” “What exactly can detection of faces be helpful for?”, Worry for adversarial uses of face detection.	PE	Clarified feature scope in the “Usage Instructions” dialog by explaining that detection is for spotting faces, not identification.
Feature Request	“The upload page should include an option to not select face detection as part of the results”.	ATC	See Section 4.2.4.6.
Bugs	“Bug Missing shots / keyframes on KSE” .	WP2	Finetuned threshold for very blurry frames exclusion.
Bugs	Disjoint detection of text regions.	WP2	A connected components algorithm was introduced to merge text areas located close to each other.
Feature Request	“Add option to pick keyframes manually”, “The user could for example tap on the space key to select an image while viewing the video”.	ATC	Rejected - out-of-scope for the KSE’s vision.

Feature Request	“Add option to change keyframe selection criteria”, “Differences, for example changes in image resolution...could be displayed on the timeline as zones or as simple shot changes”.	ATC	Rejected - out-of-scope for the KSE’s vision.
Feature Request	“...the ability to find key frames elsewhere...”, “generate custom keyframes in addition to automatically generated ones.”	ATC	Rejected - out-of-scope for the KSE’s vision.
Feature Request	“Audio analysis within a certain time-window around the keyframe”. “Detect more suspicious elements E.g. gunshots detection”.	PE	Incorporate sound-event detection - see Section 4.2.4.5.
Feature Request	“Add OCR to extracted text”, “Transcription of video next to keyframe”.	PE	Once the integration into verification tools is completed, the users can use the OCR tool of the platform.
Feature Request	“integrating reverse image search for comprehensive fact-checking” “geolocation databases for comprehensive fact-checking” “Specifically, users request combining KSE with the following tools: Reverse image search, Geolocation, DBKF, Synthetic media detection, OCR with transcription and translation”	ATC, PE	The integrators in TrulyMedia and verification plugin are considering this.
Feature Request	“Written summary of video” “Language detection”	PE	Rejected - out-of-scope for the KSE’s vision.
Feature Request	“Could it be possible to extract logos?”	WP2	Rejected - out-of-scope for the KSE’s vision.
Feature Request	“Offer speed options” Two versions: fast (analysis that provides results quickly but maybe less accurate) and deep (analysis takes longer but higher accuracy), “trade-off between speed and quality” “Fast profile for operations” “The upload page should include an option to not select face detection as part of the results”	PE	See Section 4.2.4.6.
Feature Request	“Two keyframes show something weird with the hand but more blurred than	WP2	Research anomaly-detection techniques (e.g., frame-

	explicit...Could we detect those inconsistencies?" "detecting anomalies? 3 seconds...disappearing hand...spotting those anomalies automatically."		difference analysis, temporal inconsistency detection) to flag suspicious edits/artifacts automatically.
Feature Request	"Detect resolution changes as an indicator of tampering/manipulation"	ATC	Rejected - out-of-scope for the KSE's vision.
Feature Request	"Clicking on a detected face or -text region- should highlight where on the timeline it belongs to"	ATC	Timeline highlights the timestamp when hovering over a face or text region.
Feature Request	"Manual note taking"	ATC	Built-in annotation tools are available in Truly Media.
Feature Request	"Display likelihood of two face being the same person"	ATC	Rejected - out-of-scope for the KSE's vision.

Table All- 2 Key Parameters for Temporal Segmentation and Keyframe Selection

Parameter	Description
READ_BATCH_SIZE	Batch size for reading frames. Balances memory usage and efficiency
FRAME_SCALE	Scaling factor applied to frame dimensions after resizing, for computational efficiency
DCT_SELECT	Number of top-left DCT coefficients used in temporal segmentation
DCT_RESIZE	Target resolution for resizing frames before computing DCT features
TRANS_NET_OVERLAP	Number of overlapping frames between batches during TransNetV2 inference
TRANS_NET_BATCH_SIZE	Batch size used during TransNetV2 model inference
TRANS_FRAMES_THRESH	Confidence threshold for detecting transition frames from TransNetV2 predictions
SHOT_THRESH	Threshold used to convert TransNetV2 probability scores into binary shot boundaries
SUBSHOTS_TRIES	Maximum number of retries allowed if the MIN_SUBSHOTS constraint is not met
MIN_SUBSHOTS	Minimum required number of subshots per video; triggers retries if unmet
TPS_IGNORE	Initial threshold used to detect subshot boundaries via derivative analysis
TPS_T_BIAS	Multiplicative factor to relax the TPS_IGNORE threshold in retries
VSS_TRIES	Maximum number of retries allowed if the MIN_VSS constraint is not met
MIN_VSS	Minimum required number of VSS segments; triggers retries if unmet
VSS_SIM_THRESH	Initial cosine similarity threshold for defining VSS segments

VSS_T_BIAS	Multiplicative factor to relax the VSS_SIM_THRESH in retries
LAP_SIZE	Resolution to which frames are resized before computing Laplacian sharpness
DCNN_BATCH_SIZE	Batch size for ResNet feature extraction used in keyframe selection
KEY_MIN_DCNN_D	Euclidean distance threshold for selecting distinct keyframes within subshots
KEY_MIN_DCNN_D_X	Relaxed distance threshold used to allow additional, less distinct keyframes

Annex III: Extraction of Text and Geolocation from Images

The following annex includes all supplementary figures and tables for readability while remaining an integral part of the evaluation results and analysis.

III.I OCR

Memes Dataset:

Table 35, Table 36 and Table 37 present OCR benchmark results for multilingual OCR models across the Arabic, Bulgarian, and North Macedonian meme datasets, respectively. The results show variability in model performance, with GPT-4o-mini consistently outperforming other models across all three languages. For instance, it achieves the highest character and word accuracies, such as a word accuracy of 0.5605 in Arabic and 0.7290 in Bulgarian. This demonstrates its strong ability to handle diverse scripts and complex text styles. Google Vision also performs well, particularly in Bulgarian, but trails behind GPT-4o-mini in most metrics.

Table AIII- 1 OCR benchmark results on Arabic Memes dataset. The best results are in bold

Model	Char Acc (CS)	Word Acc (CS)	PI-Word Acc (CS)	Char Acc (CI)	Word Acc (CI)	PI-Word Acc (CI)
EasyOCR	0.6299	0.3731	0.3590	0.6299	0.3731	0.3590
PaddleOCR	0.0572	-0.2018	-0.1983	0.0572	-0.2018	-0.1983
MiniCPM-v2	0.1197	0.0032	0.0049	0.1206	0.0047	0.0042
Qwen-VL	-0.9445	-0.7711	-0.7711	-0.9445	-0.7711	-0.7711
GPT-4o-mini	0.8110	0.5605	0.5523	0.8110	0.5605	0.5523
Google Vision	0.5911	0.4702	0.4921	0.5911	0.4702	0.4921

Table AIII- 2 OCR benchmark results on Bulgarian Memes dataset. The best results are in bold

Model	Char Acc (CS)	Word Acc (CS)	PI-Word Acc (CS)	Char Acc (CI)	Word Acc (CI)	PI-Word Acc (CI)
EasyOCR	0.4891	0.2737	0.2541	0.5962	0.3514	0.3388
PaddleOCR	0.4820	0.1378	0.0852	0.5782	0.1874	0.1245
MiniCPM-v2	0.1838	0.0471	0.0448	0.2121	0.0483	0.0448
Qwen-VL	-0.0491	-0.1739	-0.1769	-0.0421	-0.1739	-0.1777
GPT-4o-mini	0.8109	0.7290	0.7025	0.8385	0.7671	0.7546
Google Vision	0.6240	0.5750	0.5756	0.6329	0.5947	0.6113

Table AIII- 3 OCR benchmark results on North Macedonian Memes dataset. The best results are in bold

Model	Char (CS)	Acc	Word (CS)	Acc	PI-Word (CS)	Acc	Char (CI)	Acc	Word (CI)	Acc	PI-Word (CI)	Acc
EasyOCR	0.1728		-0.0072		-0.0333		0.3483		0.1167		0.1020	
PaddleOCR	0.1479		-0.0628		-0.0919		0.3446		0.0271		-0.0295	
MiniCPM-v2	0.1759		0.0295		0.0320		0.2147		0.0365		0.0373	
Qwen-VL	-0.5075		-0.0670		-0.0670		-0.4754		-0.0507		-0.0499	
GPT-4o-mini	0.7167		0.6150		0.5868		0.8103		0.7216		0.6989	
Google Vision	0.4232		0.3293		0.3271		0.4338		0.3527		0.3573	

Among open-source models, EasyOCR stands out as the best performer, delivering competitive results across all datasets. For example, on the Arabic dataset, EasyOCR achieves a word accuracy of 0.3731, surpassing other open-source multilingual models like PaddleOCR and MiniCPM-v2. This highlights EasyOCR's efficiency in handling OCR tasks in multilingual, meme-style images, although it still falls short of the closed-source GPT-4o-mini and Google Vision in overall accuracy.

III.II Geolocation

Qualitative results. Figure AIII- 1. Figure AIII- 2, **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** and Figure AIII- 4 showcase examples of successful geolocation predictions. Each figure presents a pair of images: the query image with its ground truth location on the left, and the corresponding retrieved image from the database with its associated location on the right. In all three cases, the model accurately predicted the location of the query image.

Ground truth: Café de Flore (France)



Prediction: Café de Flore (France)



Figure AIII- 1 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Ground truth: Forbidden City (China)**Prediction: Forbidden City (China)**

Figure AIII- 2 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Ground truth: Taj Mahal (India)**Prediction: Taj Mahal (India)**

Figure AIII- 3 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Ground truth: Puno (Peru)**Prediction: Toquile (Peru), (close to Puno)**

Figure AIII- 4 Example of a correctly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Figure AIII- 5, Figure AIII- 6, Figure AIII- 7 and Figure AIII- 8 showcase examples of unsuccessful geolocation predictions. While the predictions are incorrect, they reflect the model’s ability to capture the essence of the images, with predicted locations that are visually or contextually similar to the ground truth.

Ground truth: Rio de Janeiro (Brazil)



Prediction: Buenos Aires (Argentina)

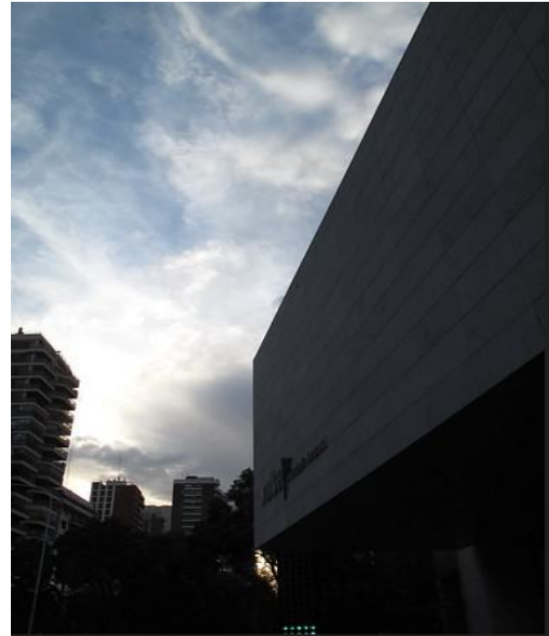


Figure AIII- 5 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Ground truth: Santiago de Chile (Chile)



Prediction: Atlanta (USA)



Figure AIII- 6 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Ground truth: Monumental Stadium, Buenos Aires (Argentina) Prediction: Horsens (Denmark)

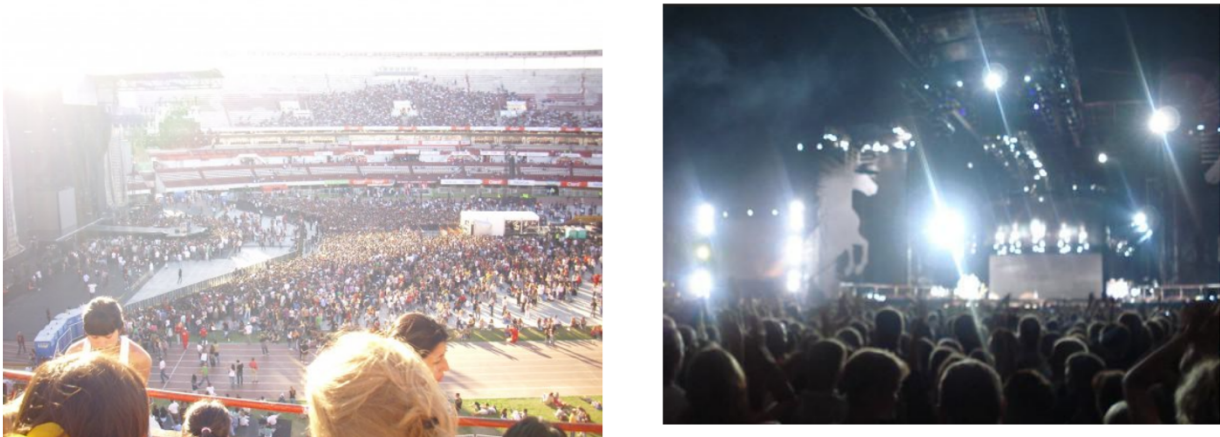


Figure AIII- 7 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Ground truth: ENS Paris Saclay (Gif-sur-Yvette, France) Prediction: Berlin (Germany)



Figure AIII- 8 Example of an incorrectly geolocated query image. Left: Query image and its ground truth location. Right: Retrieved image and its location

Table AIII- 4, Table AIII- 5 and Table AIII- 6 present the results of a series of experiments conducted to identify the best-performing geolocation model. These experiments are designed to evaluate the impact of different Earth partitioning strategies, prediction schemes, and loss functions on overall model performance. For each experimental set up, we detail the specific configuration used, including the partitioning method (e.g., Google S2 or GADM), the number of spatial classes, the type of prediction output (top-1 or top-3), the hierarchical structure (if applicable), and the loss function employed during training. The performance of each variant is compared to highlight the effectiveness of the final selected model, Google S2 with 35K classes and top-3 prediction accuracy, which consistently outperforms the alternatives across the evaluation benchmarks.

Table AIII- 4 Accuracy (%) on five granularity ranges of experimental methods on the Im2GPS evaluation set

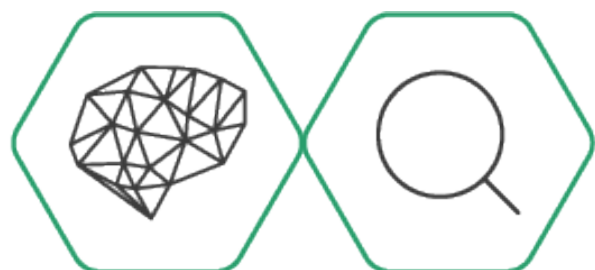
Im2GPS					
Method	1km	25km	200km	750km	2500km
Google S2 35K	16.4	46.4	62.0	77.6	91.6
Hierarchical classification	13.5	40.5	54.4	<u>74.7</u>	88.2
Top-k=1	<u>15.6</u>	<u>44.7</u>	<u>59.9</u>	74.3	<u>90.7</u>
GADM 26K	14.3	43.5	55.7	76.0	87.4
Cross-entropy loss	11.3	39.4	51.7	72.8	85.9

Table AIII- 5 Accuracy (%) on five granularity ranges of experimental methods on the Im2GPS3k evaluation set

Im2GPS3k					
Method	1km	25km	200km	750km	2500km
Google S2 35K	13.3	39.2	52.1	68.7	82.9
Hierarchical classification	11.2	30.4	43.6	61.7	78.9
Top-k=1	<u>12.9</u>	<u>37.3</u>	<u>50.0</u>	<u>67.4</u>	<u>81.4</u>
GADM 26K	11.9	33.2	46.7	64.8	80.6
Cross-entropy loss	10.7	27.5	42.3	62.4	79.3

Table AIII- 6 Accuracy (%) on five granularity ranges of experimental methods on the Im2GPS3k evaluation set

YFCC4k					
Method	1km	25km	200km	750km	2500km
Google S2 35K	16.2	27.2	39.0	59.2	75.3
Hierarchical classification	12.0	20.4	31.8	53.3	70.8
Top-k=1	<u>13.7</u>	<u>25.3</u>	<u>38.5</u>	<u>58.1</u>	<u>74.7</u>
GADM 26K	11.6	21.3	33.6	53.3	71.4
Cross-entropy loss	9.6	19.3	31.0	52.6	69.8



vera.ai



vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.

The content of this document is © of the author(s) and respective referenced sources. For further information, visit veraai.eu.