vera.ai: VERification Assisted by Artificial Intelligence

# **D2.2** – Evaluation report

| | |
|---|---|
| **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.

| | |
|---|---|
| **Deliverable title** | Evaluation report |
| **Deliverable number** | D2.2 |
| **Deliverable version** | V1.0 |
| **Previous version(s)** | - |
| **Contractual Date of delivery** | 14.09.2025 |
| **Actual Date of delivery** | 11.10.2025 |
| **Nature of deliverable** | Report |
| **Dissemination level** | Public |
| **Partner Responsible** | AFP |
| **Author(s)** | Denis Teyssou, Valentin Porcellini, Bertrand Goupil (AFP), Anna Schild (DW), Lalya Gaye (EBU) |
| **Reviewer(s)** | Nedelina Mitankina (ONTO), Luisa Verdoliva (UNINA) |
| **EC Project Officer** | Peter Friess |

| | |
|---|---|
| **Abstract** | This report describes the methodology and the results obtained during the two years and six cycles of evaluation undertaken with end-users testing vera.ai technologies during the project lifetime. The evaluation activities had been carried out starting from the second year of the project until its completion. |
| **Keywords** | Evaluation, end-users |

## Copyright

# Revision History

| Version | Date | Modified by | Comments |
|---|---|---|---|
| V0.1 | 12/02/2025 | Denis Teyssou (AFP) | First draft |
| V0.2 | 09/04/2025 | Denis Teyssou (AFP) | Second draft |
| V0.3 | 15/07/2025 | Anna Schild (DW), Lalya Gaye (EBU) | Sections and content added |
| V0.4 | 23/07/2025 | Anna Schild (DW), Lalya Gaye (EBU) | EBU x DW sections finalized |
| V0.5 | 05/08/2025 | Denis Teyssou, Valentin Porcellini (AFP) | More sections added |
| V0.6 | 11/09/2025 | Giada Marino, Anwescha Chakraborty (UNIURB) | Refinement on Coordinated Behaviour Detection and Vera AI Alerts |
| V0.7 | 29/09/2025 | Denis Teyssou, Valentin Porcellini, Bertrand Goupil (AFP) | Adding evaluation metrics |
| V0.7.1 | 30/09/2025 | Denis Teyssou (AFP) | Adding final sections |
| V0.8 | 07/10/2025 | Nedelina Mitankina (ONTO), Luisa Verdoliva (UNINA) | Internal review |
| V0.9 | 08/10/2025 | Anna Schild (DW), Lalya Gaye (EBU), Denis Teyssou (AFP) | Implemented internal reviewers feedback |
| V0.9.1 | 09/10/2025 | Olga Papadopoulou, Symeon Papadopoulos (CERTH) | Ultimate formatting and consistency checks. Preparing for submission to EC |
| V1.0 | 11/10/2025 | Olga Papadopoulou, Symeon Papadopoulos (CERTH) | Deliverable sent to EC |

# Glossary

| Abbreviation | Meaning |
|---|---|
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| DoA | Description of Action |
| FIMI | Foreign Information Manipulation and Interference |
| GAN | Generative Adversarial Networks |
| GenAI | Generative AI |
| HCI | Human-Computer Interaction |
| IFCN | International Fact-Checking Network |
| IPTC | International Press Telecommunications Council |
| ITW | Abbreviation of In The Wild |
| LLR | Log-Likelihood Ratio |
| UI/UX | User Interface / User Experience |
| WP | Work Package |

# Table of Contents

# Index of Tables

# Index of Figures

# Executive Summary

Deliverable D2.2 details the main evaluation activities (T2.3) undertaken during the final two years of the vera.ai project. It outlines, among other tools and prototypes, how we tackled the challenge of detecting synthetic images in real-world scenarios, enhanced video verification with an innovative Keyframes Enhancement Service, and built innovative tools for social network analysis despite platforms' reluctance to share data.

Most of our evaluation activities were grounded in real-life fact-checking use cases where our project regularly had a strong impact, allowing fact-checkers to debunk fake and AI-generated content more effectively. After months of refinement, several vera.ai tools are now used by thousands of fact-checkers, journalists, and researchers around the world.

The project also developed many other prototypes and proofs of concept that were evaluated to upgrade and prepare their features for a production environment. In this respect, our project has been a lever for advancing verification during a very difficult period for the fact-checking industry, which faces economic downsizing, harassment from populist movements, and a disruptive wave of AI-fuelled disinformation.

Furthermore, we were able to validate and update a design framework for AI-based fact-checking tools that can function as a checklist for further developments in this domain.

We collaborated with France's Viginum agency to perform a quantitative data analysis of all models developed by the project to detect synthetic images. This partnership was particularly valued as it helped us better assess the recall and precision of our developments. This was crucial for investigating and mitigating false positives, which had been a top ethical priority throughout the six evaluations. Our goal was to avoid misleading our users while protecting both their reputation and that of the tools themselves.

Another critical issue we focussed on was the growing weaponization of detection results in the many violent and political conflicts shaping our world. By providing better explanations, offering guidelines for using the tools, and constantly supporting fact-checking work, we managed to avoid the controversies that now arise regularly in disputed elections and war zones across the globe.

This evaluation work tackled both innovative tools built by the project and the vera.ai enhancement of features developed in previous projects. This deliverable is the result of testing vera.ai tools and prototypes with two complementary methodologies: design thinking and participatory evaluation.

# 1 Introduction

This deliverable reports on the evaluation activities of the technologies and tools developed within the vera.ai project, conducted throughout six iterative cycles from mid-October 2023 to July 2025, in the last two years of the project.

As stated in the Description of the Action (DoA), the first phase of evaluation cycles aimed to ensure that the vera.ai technology is at the desired level of quality and usability. The second phase was conducted on a larger scale to evaluate the effectiveness and real-world performance of vera.ai technologies on social networks, by collecting usage statistics and qualitative feedback to refine and improve the project's verification features. Overall, the evaluation phase was set out to assess iteratively and together with target end-users, how well the tool prototypes were meeting the requirements established in year 1 (cf. deliverable D2.1).

Two complementary methodologies were applied throughout the evaluation: the design thinking testing phase (section 3.1) and participatory evaluation (Ehn, 1993) (section 3.2). These approaches ensured both a structured, real-world testing process and the active involvement of end-users in shaping and refining the tools.

The tools and services evaluated during the project are described in the following Table 1:

*Table 1 Matrix of the vera.ai tools under evaluation*

| vera.ai tools and services | GenAI detection tools | Analytical support tools and services |
|---|---|---|
| Innovative (new) tools | Synthetic Image Detector (Section 4.1)<br>Synthetic Audio Detector (Section 4.3) | Credibility Signals (with Machine-Generated Text Detector, section 5.5)<br>Coordinated Sharing Detection Service (Section 4.7)<br>Temporal Analysis (Section 4.5)<br>Audio Provenance (Section 4.2)<br>vera.ai Alerts (Section 4.8)<br>Claim Extractor (Section 4.9) |
| Improved tools (faster, more reliable) and enhanced (adding more functionalities) | Image Forensic AI filters (Section 5.4)<br>Deepfake Video Detector (Section 5.2) | Keyframes Selection and Enhancement Service (KSE, Section 5.1)<br>Database of Known Fakes (DBKF, Section 5.3)<br>CheckGIF (Section 5.5)<br>Location Extraction/Geolocalizer (Section 5.6) |

This deliverable is divided between innovative tools and services built *de novo* by the project and tools and services launched in previous projects, which have been improved (faster, more reliable) and enhanced (adding new functionalities).

GenAI detectors are new tools developed within the project to address the growing challenge posed by generative artificial content. They are detectors aiming to give at least an indication, at best a clue or even a proof that a particular content was made with AI.

Journalists typically do not solely rely on AI detectors due to the risk of inaccuracies (and potential reputation damage) that may arise from their use. Editors are even more cautious. Consequently, media editorial guidelines generally advise against using detection tools as sole proof. Instead, these tools should serve as additional evidence, providing scientific corroboration to inconsistencies identified through prior means.

However, GenAI introduces an epistemic disruption as the produced content often lacks clear context and verifiable origin, making the core work of fact-checkers and journalists significantly more complex and time-consuming.

Regarding GenAI detection tools, particular emphasis has been placed on preventing false positives that could mislead end-users and sow more confusion in the digital information landscape. It sometimes conducted the consortium to discard some algorithms and to actively retrain models to improve the output.

In the evaluation of the diverse vera.ai analytical support tools and services we put more focus on Keyframes Selection and Enhancement Service (KSE), the most widely used feature in the Verification Plugin, Database of Known Fakes, also widely used, and the new Credibility Signals and Coordinated Sharing Detection Service.

In this deliverable, we report on the evaluation methodologies applied on the study of all the above available tools, the process we followed, and the insights we obtained from end-users that helped to improve our tools.

# 2 Deliverable Overview

In this section we provide insights into the main outcomes of the evaluation; how we worked with users; how the evaluation fit within the project, responded to initial year 1 requirements, and interacted with other work packages; and a discussion of the impact and contribution to wider research (including peer-reviewed publications).

## 2.1 Main outcomes

Our project's main success has been its timely response to the challenge posed by realistic GenAI content, particularly synthetic images. This disruption is a paradigm shift affecting not just information verification specialists but society at large—an understanding we gained over a three-year process of developing and testing solutions in close collaboration with these professionals. Generative AI fundamentally upends the nature of image verification for end-users. The traditional process of comparing a manipulated file to an original, risks becoming obsolete, as GenAI often lacks a traceable origin. Furthermore, AI can generate high volumes of fictional content that exceed human verification capabilities and are used to fuel disinformation.

While our approach to tackling generative AI disinformation has shown promise, the challenge remains complex and requires continued attention. The challenge is constantly evolving, with hundreds of generative models for images, audio, and video being launched at a rapid pace. Compounding the issue, digital files shared on social media are often repeatedly copied, degraded, and compressed, which complicates the detection of GenAI.

File format standards are also evolving quickly. For approximately three decades, digital image forensics has been predominantly shaped by the characteristics of the JPEG format. However, the emergence of high-efficiency formats such as HEIF (promoted by Apple), AVIF, and WebP (promoted by Google) is forcing the discipline to urgently overhaul its methodologies and tools. Detecting the subtle traces of GenAI on these structurally complex and highly compressed formats is significantly more difficult.

The probabilistic nature of AI predictions introduces another major hurdle: the risk of generating false positives. This presents a massive challenge for the design, implementation and testing of the AI models developed during the project, as we shall see in this Evaluation deliverable.

The false positive dilemma also creates a serious ethical problem when deciding whether to make these tools public, given the risk of errors from misuse or over-interpretation. In competing tools observed during the project, we have seen how the potential for misuse or over-interpretation has, in political contexts, led to the weaponization of detection results.

Although the mitigation of risks to end-users was also a concern in previous projects, it has grown in importance due to the increasing complexity of the disinformation landscape and the rise of GenAI, which is disrupting traditional verification methods. In this regard, our initial intervention (with a joint keynote[1]) and participation in the UvA Digital Methods Winter School 2023[2] dedicated to "the use and misuse of Open-Source Intelligence" were instrumental in drawing more attention to this critical issue.

Beyond GenAI, we have managed to extend social network analysis to detect coordination behaviour, even if we were deprived of data access due to changes made by the main platforms to their terms and conditions. We also made progress in content analysis (Assistant Credibility Signals and Chatbot). We have also improved several important tools for video verification (KSE), image verification (Forensic analysis and CheckGIF), knowledge retrieval (DBKF) and location extraction.

There was additionally strong focus on how to meet the needs of journalistic end users when using AI-based tools to understand not only the results or outputs, but also how these tools came to such results. This was intended to allow users to make informed decisions about whether to rely on a result or not. Participatory evaluation sessions provided technical partners with actionable feedback for making tools fit into user workflows in a trustworthy and useful way. Furthermore, they allowed the validation and updating of a design framework for AI-based fact-checking tools.

Another significant result is that our project's work on synthetic image detection and evaluation drew the interest of Viginum, France's public service for vigilance and protection against foreign digital interference, leading to a sustained collaboration with AFP Medialab.

### 2.1.1 User Feedback Collected Through the vera.ai Survey

To better understand the impact of vera.ai tools on user satisfaction and feature adoption, the AFP Medialab team launched at the IFCN Global Fact 12[3] conference at the end of June a final survey[4] to gather feedback from the end-users of the Verification Plugin[5] (hereinafter also referred to as the plugin). This summary highlights the most significant findings and actionable insights we have gained.

The survey consisted of 16 questions about the participants, their main occupation, their frequent use of the plugin, their personal satisfaction with the toolbox, their evaluation of the new tools, the features' usefulness in their workflow, and more qualitative questions about the user interface, their trust in the results or new features they would like to see in updates. At the time of this deliverable's writing, 72 participants (fact-checkers, journalists, OSINT investigators, and researchers) have responded. Respondents (mainly from AFP, Al Jazeera, dpa, LeadStories, MythDetector, Deutsche Welle, France24, etc.) could declare several occupations. Their frequency of use of the plugin, measured on a Likert scale, is 3.75 with 19.44% of users using it a lot and 41.67% often. The overall satisfaction measured on the same scale reached 3.94 with 12.5% of very satisfied users and 70.83% of satisfied.

[1] by Kalina Bontcheva and Denis Teyssou: https://u.afp.com/DMI23-Keynote

[2] https://www.digitalmethods.net/Dmi/WinterSchool2023

[3] Global Fact is the main yearly worldwide gathering of fact-checkers.

[4] Online survey: u.afp.com/GF12pluginsurvey

[5] Available on the Chrome store: u.afp.com/plugin

More user feedback was collected about the plugin's usage in Table 2:

*Table 2 User feedback from 72 participants, measured using a Likert scale*

| User feedback | Score |
|---|---|
| Is the plugin easy to use? | 4.18 |
| Overall user satisfaction with the plugin | 3.94 |
| Are AI detectors easy to understand? | 3.93 |
| Do you trust the plugin results? | 3.83 |
| Is the design appealing? | 3.76 |
| Frequence of use | 3.75 |

The Keyframe Selection and Enhancement (KSE) service, formerly known as Keyframe Fragmentation, has been a cornerstone of our video verification plugin since its 2017 launch. It continues to be a critical asset, now augmented with valuable AI-driven features that offer end-users additional verification capabilities. Underscoring its utility, 47 of 72 surveyed professionals identified keyframes as the plugin's most useful feature for their daily work.

This tool got the highest score of 4.44 out of 5 with 38 survey takers rating it as "excellent", far ahead of the Plugin's Assistant which got a score of 3.56, CheckGIF (3.51), and Synthetic Image Detection (3.41). All the other tools got an above average score but were either less used, with fewer responses, or were not immediately perceived as essential in the participants' verification work.

Table 3 provides the final output of the survey's participants by tools (as named in the Plugin):

*Table 3 Final survey results by tool*

| Tools | Score |
|---|---|
| Keyframes | 4.44 |
| Assistant | 3.56 |
| CheckGIF | 3.51 |
| Synthetic Image Detection | 3.44 |
| Deepfake Video Detector | 3.22 |
| Geolocation | 3.02 |

We consider the above as a positive indicator for the project, as user appreciation remains inherently subjective, often determined by their ability to successfully complete their intended verification tasks, including complicated use cases. The user's experience is therefore directly linked to their operational success, making their sentiment about the tools a metric of perceived utility rather than an objective measure of their quality. Several respondents mentioned in the survey an issue with the tool's inability to load or accept certain URLs, especially from Facebook.

The following chart (Figure 1) summarises the overall appreciation on all the vera.ai features used in production through the Verification Plugin.

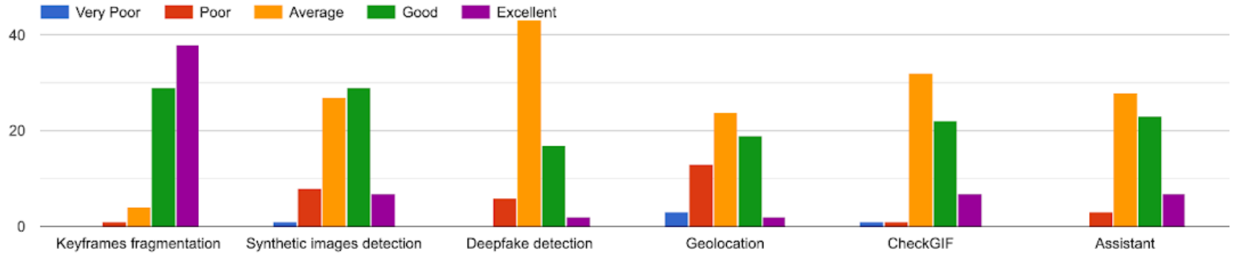


*Figure 1 Chart showing the vera.ai features appreciations of end-users in the final survey*

### 2.1.2 A Design Framework for Trustworthy-by-Design AI-Based Fact-Checking Tools

As the notion of trust was central to all discussions with end-users in participatory design sessions (see Deliverable D2.1), a concept that quickly emerged from these interactions as a highly relevant one for this project was that of 'trustworthiness by design', which in software engineering is defined as "ensuring that trustworthiness is at the core of design, implementation and maintenance" (Gol Mohammadi, 2019).

While model performance could be seen as the main aspect of a trustworthy system, end-users tend to put strong additional value on softer aspects pertaining to their use of the system: aspects that support the process of e.g. confirming results.

Trustworthiness in qualitative research has been identified as being based on credibility, transferability, dependability, and confirmability (Ahmed, 2024). This means, in the case of AI-based verification tools in journalistic research, that these demonstrate minimal bias, are consistently applicable in different situations, and need to provide as accurate results to its users as possible, as well as display impartiality and objectivity. Unlike disinformation agents, our end-users also need to understand not only the results but also how the AI-based tool they use led to it, so that they can make informed decisions about whether to rely on them. Based on this, on initial results from the participatory design sessions in year 1 (see Deliverable D2.1), and on literature about the notion of 'trustworthiness by design' in AI systems (Gol Mohamadi 2019, Poretschkin et al. 2023, Díaz-Rodríguez et al. 2023, Maia et al. 2024, Berman et al. 2024, Sykora 2023, Floridi 2019), we were able to outline an initial design framework for the design of AI-based verification tools: one that could potentially meet our target user group's core needs of trust and workflow. As a result of the participatory evaluation activities described below, we validated and refined this design framework (see Table 4), which now serves as a check-list of actionable points for future developers of AI-based fact-checking tools to consider in similar efforts (Gaye et al., 2025).

*Table 4 Design framework for AI-based fact-checking tools, based on participatory research results*

| Principle | Definition |
|---|---|
| Usability and effortlessness | Tools should be easy to use and immediate to learn, as users do not have time for a steep learning curve. |
| Speed | Tools should allow for quick use. |
| Autonomy and control | Tools should be usable autonomously without expert support and should put users in control of the process. |
| Literacy | Tools should encourage the user to learn how they work, so they can be confident in their results. |
| Layers of explainability | Tools should provide different layers of complexity in AI models and in explaining how they came to a result. |
| Familiarity in interpretability | Results should be easily interpretable by the end-user and be provided by the tools in an accessible and familiar language to them. |
| Built-in transparency about uncertainties | Tools should be transparent about error rate, false-positive rates, building them as e.g. likelihood ratios into how results are represented. |
| Verifiability | Tools should not only allow for humans to oversee their functioning but provide opportunities for cross-checks. |
| Augmented performance | Tools should rely on and amplify existing user expertise and thereby assist users rather than replace them. |
| Inclusiveness | Tools should be accessible to various abilities and cater to different geographical and work cultures. |
| Flexibility | Tools should allow for ad-hoc paths of use vs. guided ones when used in combination with one another. |
| Meaningfulness | Tools should align what results they display with what users need to get. |
| Redundancy | Tools should use redundant ways of displaying information. |
| Seamlessness | Going from using one tool to another in the same verification process should be seamless. |
| Levels of use | Tools should provide different levels of use from basic to advanced. |
| Empowerment in stakeholder engagement | All relevant stakeholders should not only be engaged in the design process, but even gain a level of empowerment over it. |

We recommend taking a user-centered and participatory approach based on the specific needs and values of one's target group if making use of this design framework. As important as keeping tools up to date by making them continuously evolve with the emergence of new forms of disinformation, developers need to continuously touch base with their end-users and make them evaluate changes, in order to keep a good fit with their professional use – a process that does not end.

## 2.2 Working with Users: Profiles, Process and Recruitment

The evaluation phase of vera.ai involved the target groups defined in the DoA, namely verification professionals (mainly fact-checkers, journalists and OSINT investigators working actively in information verification, disinformation debunking and digital investigations), journalists and media researchers. These included web, TV, radio and press journalists from a variety of public service media organizations (such as the BBC, France Televisions, RTVE, ZDF, BR, SRG-SSR, CBC/Radio-Canada, Radio France, DW), fact-checkers (AFP, France 24 observers, LeadStories, Ghana Fact, Pesacheck, Al Jazeera, Fact Crescendo, Boomlive, The Quint, etc.) and academic researchers tackling disinformation (University of Siegen, Tecnológico de Monterrey, Ca' Foscari University of Venice).

To recruit participants beyond large fact-checking newsrooms like AFP, we used a variety of methods. We sourced evaluators from the Verification Plugin's advanced user database, contacted respondents of the project survey on user needs[6], presented at dissemination events such as EDMO training sessions and Global Fact workshops, and leveraged our professional networks. This strategy enabled us to involve both users that are already familiar with previous versions of the tools and new ones with less experience of them.

With this approach, the users we work with are fully representative of the user communities that the project is set to cater for. Thanks to this diversity of users – in terms of professions, geographical distribution and, also, equal gender balance – we were able to study and address user needs across the board, and to best tackle the challenge of modalities, forms and propagation of disinformation, which is without borders and impacts all aspects of the information landscape.

Moreover, thanks to the regular and in-depth involvement of our user groups, we were able to stay on top of the evolving disinformation landscape and address verification challenges to the best of our abilities, for example in terms of those related to GenAI.

[6] For the users that have explicitly accepted to be contacted by the vera.ai project.

## 2.3 Evaluation Process with Two Complementary Methodologies

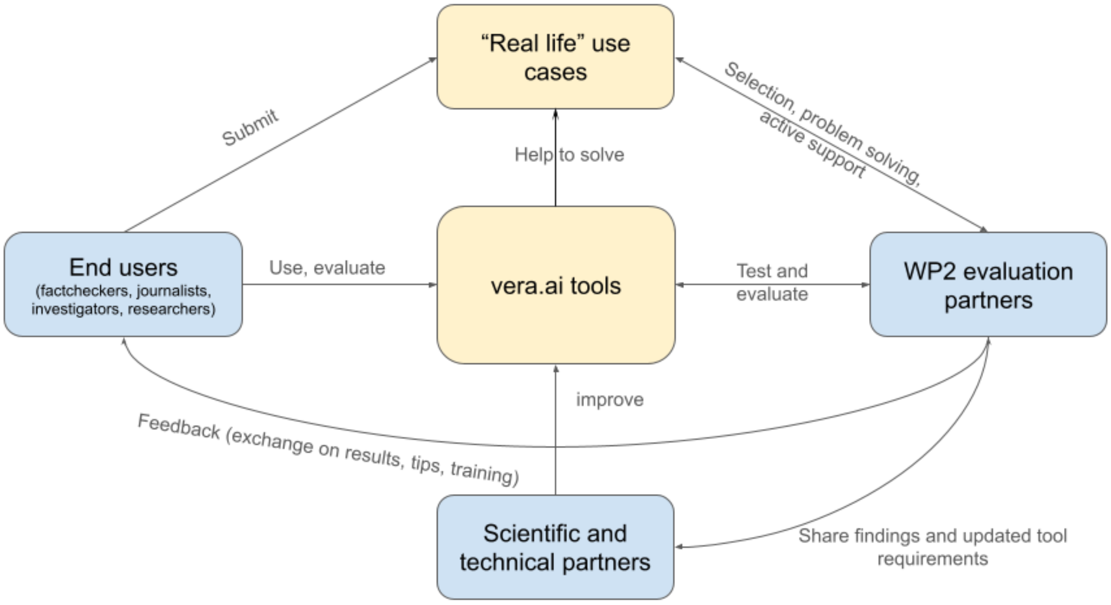

"Real life" use cases
Submit
Selection, problem solving, active support
Help to solve
End users (factcheckers, journalists, investigators, researchers)
Use, evaluate
vera.ai tools
Test and evaluate
WP2 evaluation partners
improve
Feedback (exchange on results, tips, training)
Scientific and technical partners
Share findings and updated tool requirements

*Figure 2 Evaluation within an iterative development process*

In this evaluation phase of the project, we focused on testing tool performance and limitations, as well as how they fit into user workflows (Figure 2).

The technical and design requirements defined in year 1 through large-scale surveys, targeted ethnographic studies and participatory design sessions, were provided to technical partners in periodic WP2, WP3 and WP4 online meetings. This helped inform iterations of technical development at the start of this phase, and insights from co-creation activities lead to new ideas in terms of features and model development. Improved prototypes were then evaluated, again with users, and insights fed again to technical development in an iterative manner.

In each iteration cycle, evaluation activities worked closely with the users described above (fact-checkers, journalists, researchers) and with usage data to assess how well further developments responded to previous requirements. Results were then compiled into thorough reports and lists of action points for technical partners and discussed among WP2 partners. By running these evaluations and co-creation activities on various specific aspects of prototypes and at different levels of maturity, we ensured that users took part in every aspect of the process and that the final results meet their needs as much as possible.

This strategy helped to refine the tools, as well as improve the results and the user experience. It led to the improvement and sometimes redesign of tool features such as user interfaces, information visualisation, transparent communication about limitations such as error rates and risks of false positives. It also provided insights into improving models and workflows, about how to design tools that users consider as trustworthy, and about how to integrate and combine tools within the same platform in a coherent way.

In order for this process to tackle all these aspects, we used two different and complementary approaches to assess the tools, each method focusing on various aspects of the development: an iterative beta testing in the wild through the Verification Plugin for the evaluation of tools performance and limitations, and a participatory approach for the user workflow angle, both with the goal to improve their usability and usefulness. The two evaluation approaches are described in more detail in Section 3 of this document.

# 3 Evaluation Methodologies

This section outlines the two evaluation methodologies we employed as part of our iterative project process, beginning with the Design Thinking approach, followed by the Participatory Evaluation.

## 3.1 Design Thinking Evaluation

In the design thinking methodology—used to build the Verification Plugin (formerly The InVID-WeVerify browser extension) since its inception in 2017 (Teyssou, 2019) —evaluation is conducted iteratively during testing phases. This process validates two key aspects:

- User focus: Whether the developed prototypes and features are user-centred and solve specific user problems effectively (considering efficiency, performance and performativity) in fact-checking context.
- Usability: whether the tools are easy and intuitive to use, thereby minimizing the cognitive load and interpretive burden on the end-user.

We ensure efficiency (reliable operation, fast execution, effective error handling), performance (scalability, reliability, and precision, particularly regarding false positives), and performativity (the tool's ability to optimize end-user verification tasks). This is achieved through a multi-faceted approach, including thorough testing with real image to detect false positives, continuous engagement with fact-checkers (through one-to-one conversations on their “real life” use cases, in-person and online meetings, and surveys), and detailed system log analysis (e.g. reviewing API error responses, timeouts, and malformed requests).

This approach involves constant interaction with verification professionals. The interaction involves both participant observations, enabling us to gain insights into user behaviours in their daily work, and an active, supportive role (intervening as needed) through which we assist users in accomplishing their verification goals, mostly by leveraging features being developed within the vera.ai project. Our collaborative engagement with end-users across more than 100 interventions throughout the project’s lifetime enabled us to build strong ties with a cohort of “lead users”—a concept defined by innovation theoretician Eric von Hippel as “users whose present strong needs will become general in a marketplace months or years in the future” (von Hippel, 1986).

The process schema presented in Figure 3 details the three main phases (Inspiration, Ideation, and Implementation). Each phase is divided into tasks (circles) while the ellipses symbolise the iterative process between those tasks. Understand (user needs), observe (end-users), express (point of view), ideate (solutions), prototype (co-creation) and test (evaluation) are all part of WP2. The prototyping is led technically by WP5 (Integration) on top of the provided scientific results (from WP3 and WP4).

The vera.ai features and services integrated into the Verification Plugin have been tested by hundreds of these “lead users” during the planned six evaluation phases. Insights from their feedback, observed use cases, difficulties, misunderstanding and errors have helped to continuously refine and improve the prototypes, with the goal of going beyond proofs of concept and providing an effective solution genuinely fitting complex challenges, such as detecting GenAI content in the wild on social networks.

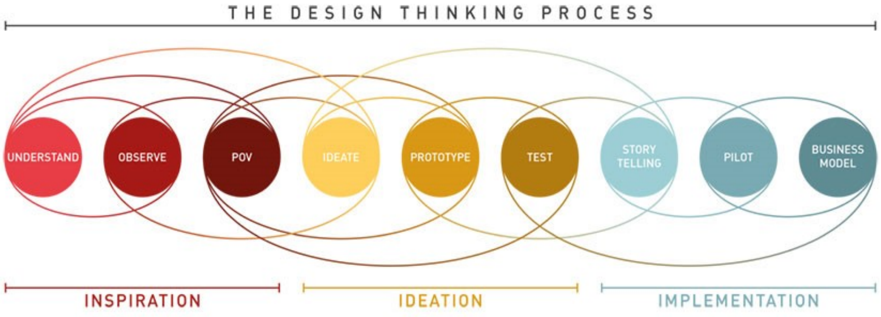


*Figure 3 View of the Design Thinking Process. (courtesy of © Ecole des Ponts Design school Paris)*

Table 5 lists the consecutive six evaluation cycles organised under WP2 with the evaluated tools, the organised side-events and the progression of advanced users involved during each cycle.

*Table 5 Schedule of six evaluation cycles*

| Cycles | Dates | Evaluated tools | Comments | Users' progression* |
|---|---|---|---|---|
| 1 | Oct 16 - Dec 15, 2023 | Synthetic Image Detection<br>Credibility Signals | Includes ESJ Lille live demo and UvA data sprint in January 2024 | 77 new registers users |
| 2 | Feb 15 - Mar 30, 2024 | Synthetic Image Detection<br>Deepfake Detection<br>KSE<br>vera.ai Alerts<br>Image Forensic Analysis | End-user survey on synthetic images | 38 new registered users |
| 3 | May 13 - Jun 21, 2024 | Synthetic Image Detection<br>Image Forensic Analysis<br>Deepfake Detection<br>Claim Extractor<br>CheckGIF<br>Credibility Signals | Verificathon UvA EU Elections | 79 new registered users |
| 4 | Oct 14 - Dec 6, 2024 | Synthetic Image Detection<br>Deepfake Detection<br>KSE<br>DBKF<br>Synthetic Audio Detection | | 87 new registered users |
| 5 | Feb 17 - Apr 11, 2025 | Synthetic Image Detection<br>Synthetic Audio Detection<br>Geolocation<br>Credibility Signals | Five new algorithms added | 81 new registered users |

| | | | | |
|---|---|---|---|---|
| 6 | May 15 - Jul 5, 2025 | Synthetic Image Detection<br>Geolocation<br>Credibility Signals | Quantitative analysis on detection and false positives<br>End-user final survey | 81 new registered users |

The user's progression is, for each cycle, a measure of the number of new applicants, who registered as advanced users in the Verification Plugin during the corresponding period. It only consists of people outside the vera.ai consortium. The plugin advanced tools, to which belong most tools developed within the project, are only accessible upon registration (to prevent counter-forensics activities and keep the computing load manageable and sustainable) and are reserved to journalists, fact-checkers, OSINT investigators and researchers working on disinformation. We also granted access to requests by human rights defenders, law enforcement and election integrity defenders.

Before starting the evaluation phase, we had 80 registered beta testers. At the time of writing this deliverable, we have a total of 1,388 beta testers (including the 443 registered during the six cycles) using the Plugin vera.ai advanced tools.

During the six individual evaluation cycles, we had an average of 82 unique users and 111 returning users per cycle, which represents in total 490 unique testers from Europe (59%), Asia (18%) and North America (17%), according to the Plugin's Matomo analytic software. The above figures do not consider KSE and the Plugin Assistant Credibility Signals which are fully open at the end of the project (without any registration) and are registering respectively 5500 and 1300 requests per month.

## 3.2 Participatory Evaluation

In parallel to the approach described above and as part of the iterative process, the participatory side of the evaluation phase (led by DW and EBU) consisted of qualitative studies focused specifically on user interaction with, and experience of the prototypes. This followed the participatory design phase in year 1 and aimed to determine the success of implementations for these services in relation to the design requirements defined in T2.1. This was done from the perspective of all involved stakeholders – in particular with regard to users' professional workflows and to incorporating the tools within these workflows rather than requiring users to change them. This process led to the development of a design framework for the development of AI-based tools for fact-checking that are 'trustworthy by design' (see section 2.1.2).

A participatory methodology requires bringing in end-users as active participants and vital stakeholders in this evaluation process: rather than developing the tools for them, these are developed and evaluated with them. Discussions among stakeholders should be in-depth, and happen on an equal footing. Participants take part in not only testing prototypes (at various levels of readiness, from mock-ups to advanced functional prototypes), but also in establishing evaluation criteria, and assessing whether these criteria were met in each iteration of the prototypes, from their own perspective. Some key techniques of participation in this context, which were used in this project, included: a) ensuring a common language among the participants devoid of jargon, and b) engaging them in the evaluation process and decision-making as actively as possible, on an equal-foot basis, and c) better understanding the context of use and people's workflow and processes within it.

The initial participatory design workshop series in year 1 had led to a number of user-centred design requirements, such as the need to provide access points for cross-examination; to adapt terminologies, information visualisation, and symbolic representations to ones that users are familiar with; for transparency regarding uncertainties; to provide redundant paths of interaction; and to make sure that users understand what they can expect from the AI model that the tool is built on. The participatory evaluation process was therefore set to assess how well the implementation of the design requirements matched these needs. Activities were thus centred on questions of workflow, literacy and trust (e.g. how tools would fit into user workflows and professional needs), as well as usability, transparency, explainability, interpretability, integration, modes of operations – all essential aspects when it comes to improving the usefulness and relevance of the tools developed in the project, and guarantee adoption from the target end-user group.

Online participatory evaluations were set from the second year of the project, to determine how prototypes fit user expectations and requirements. They also provided space for further exchanges between end-users and developers to foster a mutual understanding. While most of these end-users came from the original pool of participants of the project's first year, new end-users were brought in on a regular basis. Herewith, feedback loops and progress assessment were possible, as well as fresh perspectives. Overall, 52 individual testers contributed with written and oral qualitative feedback to the participatory evaluation sessions. Many of them did so in several sessions, resulting in 90 data points.

Since research focus, prototyping dimensions, and state of maturity varied among the services being evaluated, each of these evaluations was set-up slightly differently but along the lines of an initial internal testing to iron out minor usability issues that might get in the way of the evaluation; long-term asynchronous testing by participants and feedback; and/or follow-up sessions based on these responses, aiming at the collective in-depth exploration of emerged themes. Participants were provided with basic instructions and the prototype in question to test asynchronously in their own times, applying user scenarios in exercises and using them in their daily work, generally over a period of two weeks. During this time, users filled in qualitative questionnaires that served as a basis for group discussions amongst them and developers. After each evaluation, insights were translated into a report and series of action points for the developer's team to implement, such as modifications to the user interface, simplifying steps of use, adding functionalities, and more.

# 4 vera.ai Innovative Features

This section outlines the evaluation activities of the innovative features developed by the vera.ai project.

We start with Synthetic Image Detection, tackling a disruptive trend in (dis)-information production that broke soon after the project's start and in which our scientific partners already had valuable expertise.

We then report on evaluation work performed on audio tools (Audio Provenance and Synthetic Audio Detection), on information spread, from the perspective of coordinated sharing on social media and of the temporal spread of news topics across media organisations, as well as the vera AI Alerts aiming to detect automatically coordinated behaviour and the chatbot integrated into the Verification Plugin.

## 4.1 Synthetic Image Detection

The "boom" of AI image generation by the public took off after the launch of the vera.ai project, in late 2022 and early 2023, namely with the rise of Midjourney and Dall-E 2. Their accessibility, ease of use and social media impact contributed to this sudden rise, which involved the widespread distribution of fake images used for propaganda, destabilization or smearing.

As AI-generated images are created by models without (most of the time) any traceable origin and no reality, their verification remains very challenging: it requires other strategies than the traditional reverse image query on search engines to determine if the image already exists in another context.

Given this situation, the synthetic image detectors were heavily tested during all the six cycles, all through the Verification Plugin integration and the design thinking methodology.

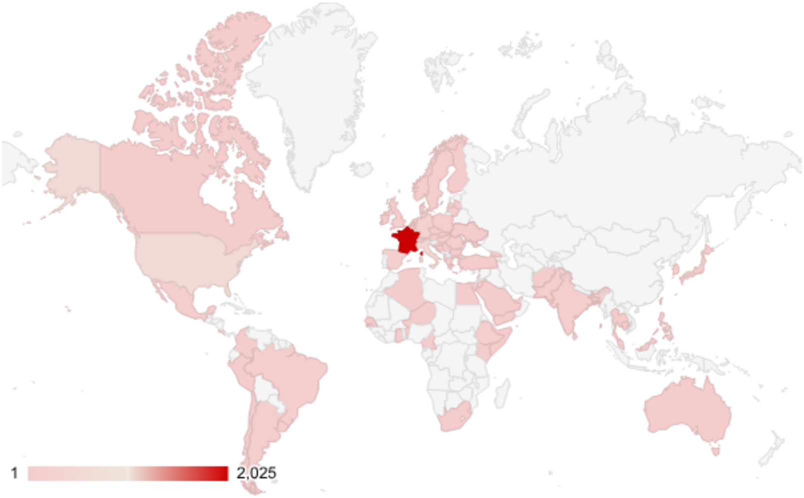


*Figure 4 View of the geographic origin of the Synthetic Image Detection queries made during the six cycles (coming from 6000 unique IP addresses)*

The above Figure 4 shows the geographic origin of the queries made by more than 6,000 unique IP addresses on the Synthetic Images Detection Service through the Verification Plugin, after eliminating the automated quantitative tests of cycle 6.

The first evaluation cycle was conducted with plugin releases v0.77 (October 11, 2023) and v0.77.2 (October 20, 2023). A total of 511 queries on the synthetic image services were made during this period.

The key findings of those early tests were:

- Difficulties were reported by participants to detect synthetic images created through Adobe Firefly or Playground AI, as well as images circulating on the Telegram platform.
- Images taken directly from the AI-generator (MidJourney, Dall-E) are better detected than later copies posted to social networks. This was particularly evident in an event at ESJ Lille School of Journalism, on November 16, 2023 where, in front of 160 students gathered in an amphitheatre, French photographer Gerald Holubowicz (Synth Media) showcased how to create synthetic image with Midjourney and Dall-E, while Denis Teyssou and a group of students were trying to detect them live with the Verification Plugin.

Figure 5 below shows how an image created and shared on the Midjourney discord channel was fully detected (100% result) by the UNINA Latent Diffusion Model algorithm (LDM).

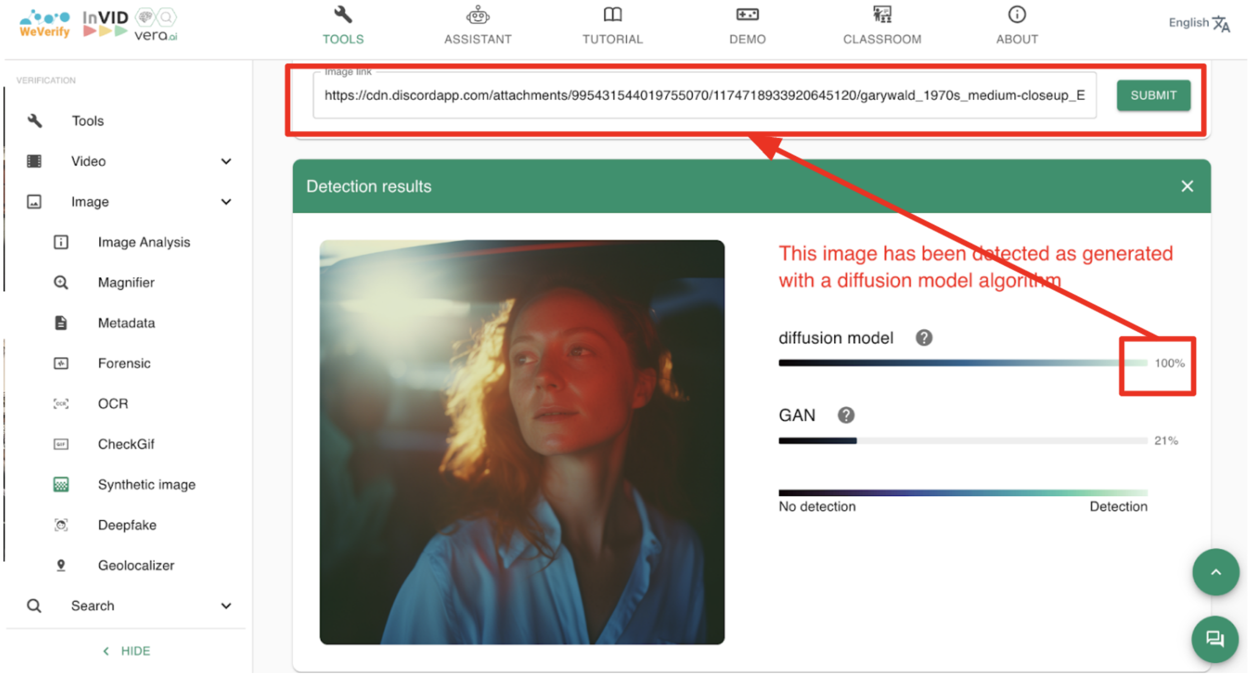


Image link: https://cdn.discordapp.com/attachments/995431544019755070/1174718933920645120/garywald_1970s_medium-closeup_E

This image has been detected as generated with a diffusion model algorithm

diffusion model 100%

GAN 21%

*Figure 5 Screenshot of the Verification Plugin synthetic image detector*

- Remarks on the limited information provided in the interface (the first prototype was re-using the forensic display of results with a progressive mako[7] colormap).
- A wide variety of photo formats (WebP, AVIF, HEIF) were found in usage data and were tested with lower results, during a Bellingcat hackathon at UvA in November 2023.
- Results variability between copies of the same image.

[7] https://cmap-docs.readthedocs.io/en/latest/catalog/sequential/seaborn:mako/

As the main algorithms were trained with synthetic images and forensic traces left by image generators, the latter point was particularly interesting to investigate as the widespread duplication of images on social media platforms posed a substantial usability issue for end-users.

We decided to tackle this immediately after cycle 1, by conducting a data sprint analysis workshop at University of Amsterdam Digital Methods Initiative (DMI) Winter School in January 2024. Analysing several images and their hundreds of copies circulating on social media, especially linked to Israel's war on Gaza, helped us to better understand the current limitations of our detection capabilities and to confirm that the detection rate decreases in time and most frequently with image geometric changes and degraded quality. The results, available on DMI website[8], led to the publication of a research paper (Karageogiou et al., 2024) presented at Trust What You learN (TWYN) ECCV 2024 Workshop.

In our research, we found that the first images indexed on Google, using the Google Fact Check Explorer Image Contexts (a tool in beta access), gave substantially better detection results. We therefore regularly shared the tip of trying to find an image the closest to the source/first indexation in all our training sessions with fact-checkers (including EDMO training and IFCN Global Fact workshop).

In the same data sprint, our multidisciplinary team (with vera.ai partners AFP, UvA, CERTH and ENS) ideated a technical solution to overcome this variability of results by archiving all the positive detections, above or equal to a 0.70 score using the CERTH Near-Duplicate Detection (NDD) service. It allows image similarity search to store and retrieve any detected copy of an image submitted through the Verification Plugin, and to populate positive detection samples retrieved by other users trying to verify a very similar image. The service returns the detection score, the algorithm used as well as a link to the analysed content. This improvement (Figure 6) was finally released, after co-creation and evaluation with lead users, in the Verification Plugin v0.82 in October 2024.

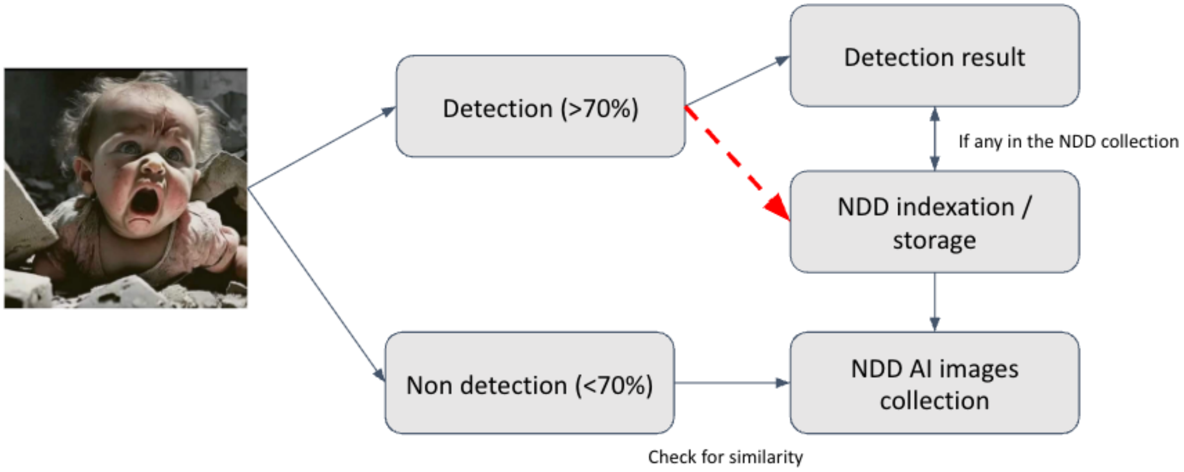
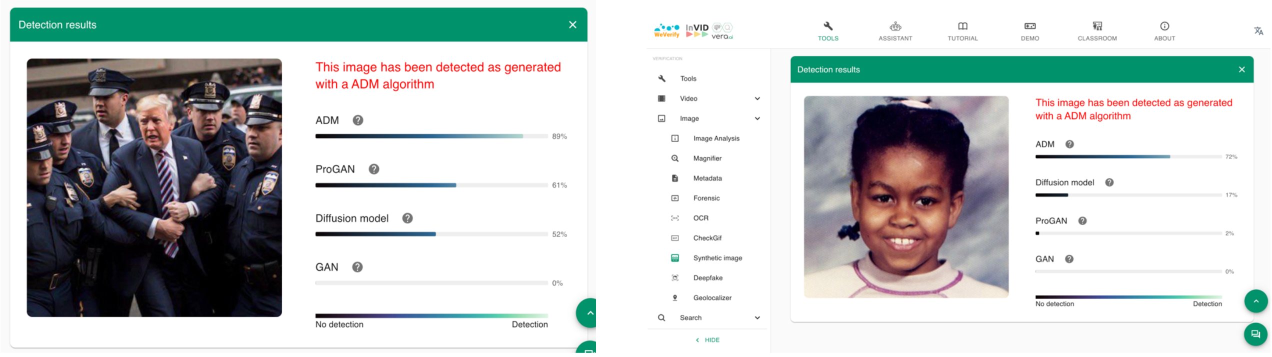


*Figure 6 Functional overview of the use of the NDD service to populate true positive results*

[8] https://www.digitalmethods.net/Dmi/WinterSchool2024SyntheticImages

During cycle 2, two new algorithms ProGAN and ADM (Adaptive Diffusion Model) were added to the toolbox (v0.78 in February 2024) and immediately tested. Both were able to detect GAN images[9], which nevertheless are rather scarce on social media since the rise of Diffusion Models (MidJourney, Dall-E and others).



Detection results

This image has been detected as generated with a ADM algorithm

ADM 89%

ProGAN 61%

Diffusion model 52%

GAN 0%

Detection results

This image has been detected as generated with a ADM algorithm

ADM 72%

Diffusion model 17%

ProGAN 2%

GAN 0%

*Figure 7 Juxtaposition of an ADM-detected fake image of Donald Trump and an ADM false positive on Michelle Obama as a child*

In our tests, ADM showed a different performance profile increasing its recall of synthetic images (compared to the previous LDM) but decreasing precision by generating false positives on real images (see Figure 7), while no such false positives were previously found with LDM. For this reason, ADM testing was kept within the consortium to avoid end-users making mistakes and flagging as AI-generated legitimate content.

Meanwhile, the LDM algorithm started to be adopted by fact-checkers to debunk disinformation like in a fake news campaign against a farmers protest in India[10]. Verification Plugin developer Valentin Porcellini (AFP) was interviewed on the results of the vera.ai feature, initiating a trend among fact-checkers consulting external experts to deepen their understanding of the AI-detection toolbox.

We completed this cycle with a short anonymous usability survey, essentially aimed at fact-checkers, to ask them if they found the feature helpful, useful (for them), explainable and unambiguous. We then encouraged them to share any issue they may have encountered with the tool and any suggestion for improvement.

Using the UEQ+ analysis method[11], we grouped our questions under two categories: Response quality and Comprehensibility. On 30 responses (see Figure 8), the consistency of the scale was positive with a Cronbach Alpha coefficient respectively of 0.96 and 0.95[12]. Results showed a mean at 0.97 on the Response Quality (measured on helpfulness and usefulness of the feature) versus -0.05 on Comprehensibility (explainable and unambiguous) for an overall KPI of 0.45 (with a standard deviation of 1.33).

[9] Generative Adversarial Network (GAN) consisting of two adversarial networks (a generator and a discriminator) competing against each other. Reference: https://en.wikipedia.org/wiki/Generative_adversarial_network

[10] https://observers.france24.com/en/fake-news-campaign-targets-protesting-indian-farmers

[11] https://ueqplus.ueq-research.org/

[12] Above 0.70 is generally considered as consistent according to the UEQ+ methodology.

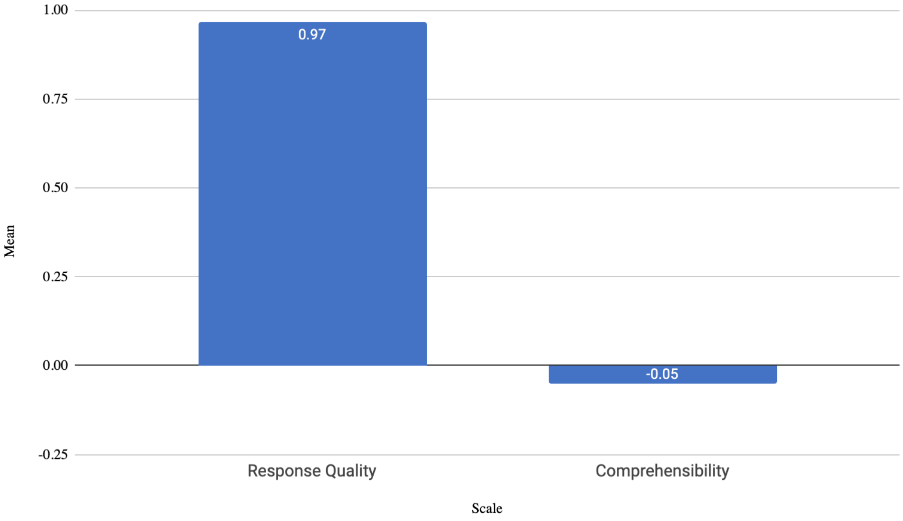
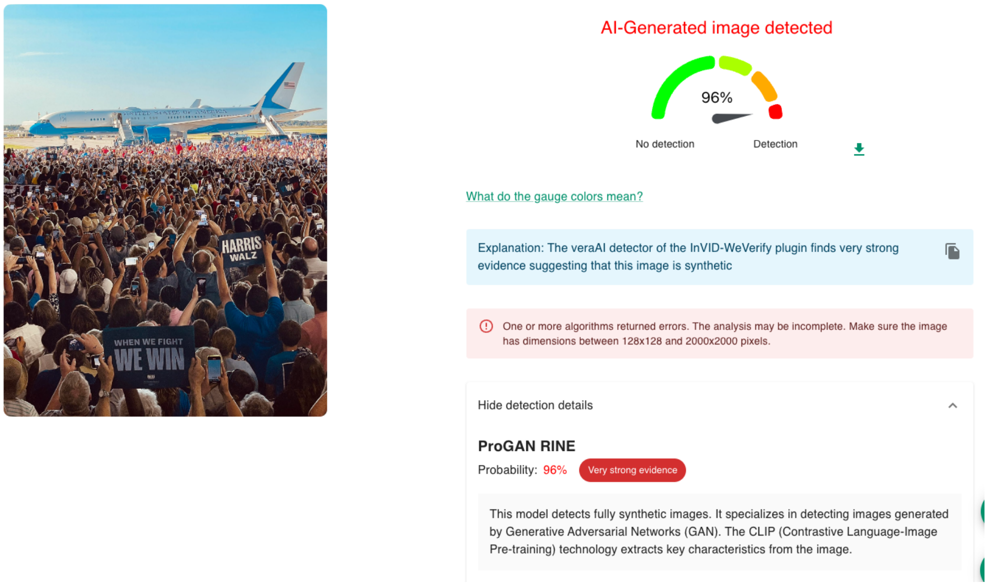


*Figure 8 Graph showing the outcome of the Synthetic Image Detection usability survey (UEQ+ methodology)*

The qualitative comments shed light on the above results. Fact-checkers had difficulties interpreting the coloured bar (inherited from the forensic tool), which indicates the result of the detection. Others reported that the toolbox was not detecting ChatGPT Pro (DALL-E) generated images. Very few advanced lead users, who had access to the new algorithms, found rapidly false positives with ADM and reported that this undermined their confidence in using the feature in a fact-checking production context.

The results of the survey were discussed during the February 2024 consortium plenary meeting in Thessaloniki, and we held an online co-creation session between partners to provide more explanations and affordances (gauge, scale and key findings). Those improvements on mock-ups were integrated and validated again by lead users in the following cycle 3.

Two new algorithms (ProGAN Rine and LDM Rine) were integrated into the API and the Verification Plugin to be tested during the third evaluation cycle. We performed false positives testing by proactively using news photos from reputed media organisations to find vulnerabilities. Unfortunately, testing several news photos often converted on the fly to the WebP format, on NYTimes or Reuters websites, resulted in several false positives.



AI-Generated image detected

96%

No detection Detection

What do the gauge colors mean?

Explanation: The veraAI detector of the InVID-WeVerify plugin finds very strong evidence suggesting that this image is synthetic

One or more algorithms returned errors. The analysis may be incomplete. Make sure the image has dimensions between 128x128 and 2000x2000 pixels.

Hide detection details

**ProGAN RINE**

Probability: 96% Very strong evidence

This model detects fully synthetic images. It specializes in detecting images generated by Generative Adversarial Networks (GAN). The CLIP (Contrastive Language-Image Pre-training) technology extracts key characteristics from the image.

*Figure 9 View of a ProGAN Rine false positive detection through the Verification Plugin*

The above example (Figure 9) coming from a picture shared on X is particularly problematic, as it represents Kamala Harris's first rally during the 2024 US Elections campaign in Detroit. Donald Trump accused his opponent of manipulating this picture with AI to increase the crowd size. This statement was baseless and other footage confirmed that this picture was not manipulated[13].

Nevertheless, a U.S. academic initiative launched to detect deepfakes during the election campaign falsely flagged the image as AI-generated. This incorrect assessment caused confusion for 24 hours before it was retracted following a human verifier's review.

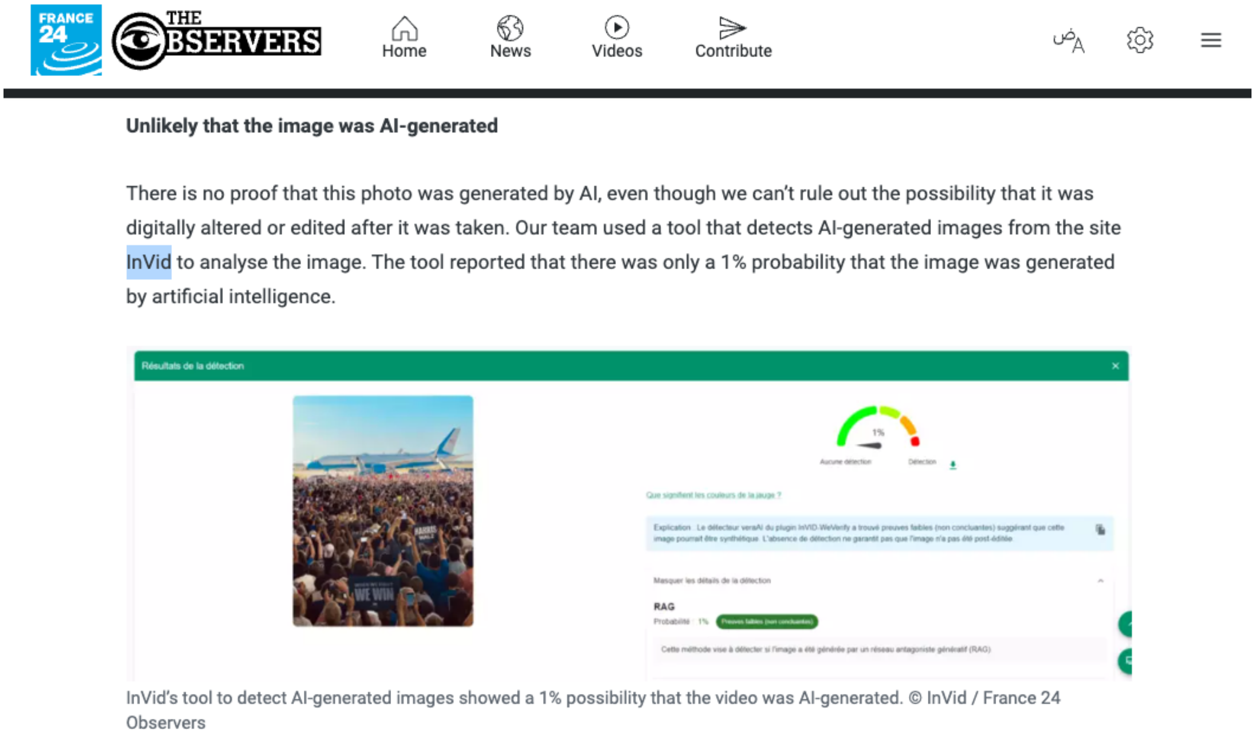
While a media organisation risks its reputation by using such a false positive in its fact-checking, the result can also be weaponised by a political side to sow confusion rather than establish facts. In the realm of international politics, this example underscores the critical importance of avoiding false positives. It is a profound ethical dilemma, as it inherently risks misinforming users and undermining the very purpose of the tool.

The production version of the Verification Plugin opened to fact-checkers did not include the above experimental algorithm. The Observers of France 24 external TV broadcaster used it (Figure 10) unexpectedly[14] as a clue that the above image was "unlikely" AI-generated[15].

[13] https://factcheck.afp.com/doc.afp.com.36E79DM

[14] Following the principle of falsifiability, the plugin is designed to help end-users debunk fake content, rather than confirming its authenticity.

[15] https://observers.france24.com/en/sorry-trump-the-crowd-that-gathered-to-see-kamala-harris-wasn-t-ai-generated


FRANCE 24
THE OBSERVERS
Home
News
Videos
Contribute
Résultats de la détection
1%
Aucune détection
Détection
Que signifient les couleurs de la jauge ?
Masquer les détails de la détection
Preuves faibles (non concluantes)

**Unlikely that the image was AI-generated**

There is no proof that this photo was generated by AI, even though we can't rule out the possibility that it was digitally altered or edited after it was taken. Our team used a tool that detects AI-generated images from the site InVid to analyse the image. The tool reported that there was only a 1% probability that the image was generated by artificial intelligence.

Explication : Le détecteur veraAI du plugin InVID-WeVerify a trouvé preuves faibles (non concluantes) suggérant que cette image pourrait être synthétique. L'absence de détection ne garantit pas que l'image n'a pas été post-éditée

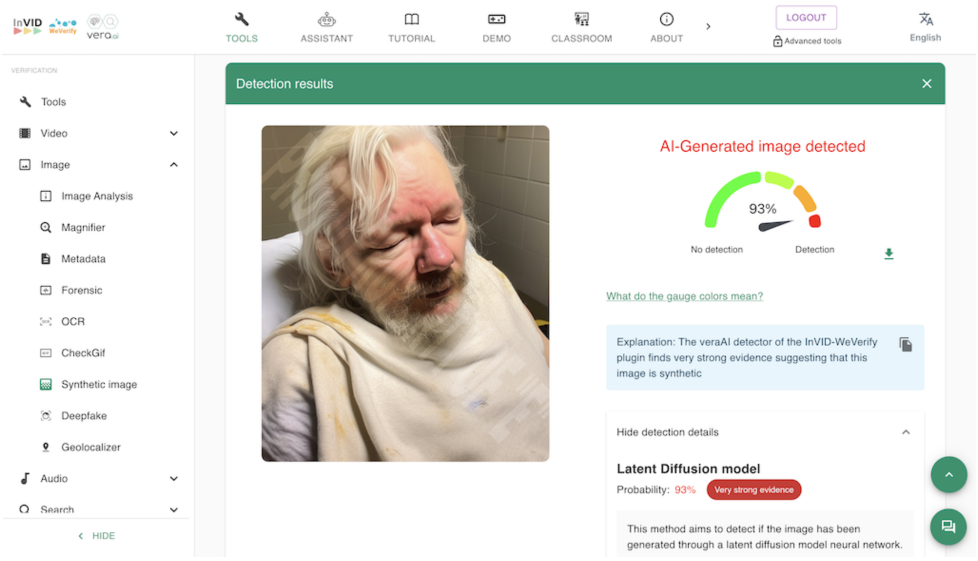
RAG
Probabilité : 1%

Cette méthode vise à détecter si l'image a été générée par un réseau antagoniste génératif (RAG)

InVid's tool to detect AI-generated images showed a 1% possibility that the video was AI-generated. © InVid / France 24 Observers

*Figure 10 View of the plugin interface on the same picture with a correct True Negative*

During the same cycle, we validated with lead users in a co-creation workshop at AFP a second version of the synthetic image detection graphical user interface (within the Verification Plugin) providing a clearer and more documented result as well as more affordances, to support fact-checkers. Figure 11 shows the result achieved for the synthetic image detection interface v2.



AI-Generated image detected

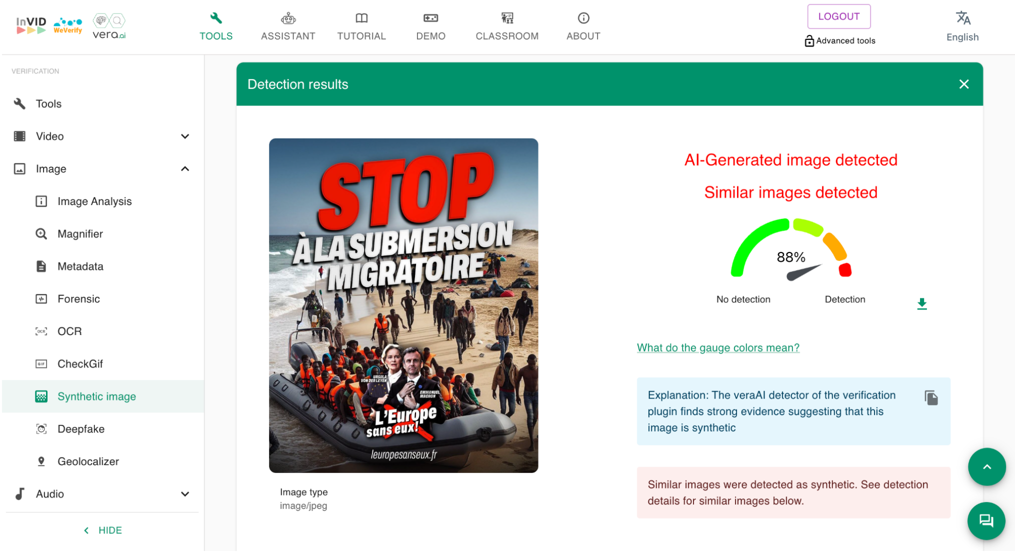
Explanation: The veraAI detector of the InVID-WeVerify plugin finds very strong evidence suggesting that this image is synthetic

Latent Diffusion model
Probability: 93%

This method aims to detect if the image has been generated through a latent diffusion model neural network.

*Figure 11 View of the Synthetic Image Detection service on an AI-generated image of Julian Assange*

On 6-7 of June 2024, partner UvA organised a "verificathon" on GenAI and Elections before the EU election. Polish fact-checker Demagog and EU non-profit organisation AI Forensics, specialized in platform algorithms accountability, joined forces with vera.ai partners UvA and AFP and 17 journalists from media organisations and researchers from various institutions. Analysing the political communication of alt-right on social media, we discovered a rather extensive usage of GenAI to create illustrations for the political campaign (Figure 12 below).



Detection results

AI-Generated image detected

Similar images detected

88%

No detection Detection

What do the gauge colors mean?

Explanation: The veraAI detector of the verification plugin finds strong evidence suggesting that this image is synthetic

Similar images were detected as synthetic. See detection details for similar images below.

Image type
image/jpeg

*Figure 12 GenAI image used by French Rassemblement National during the EU Election campaign*

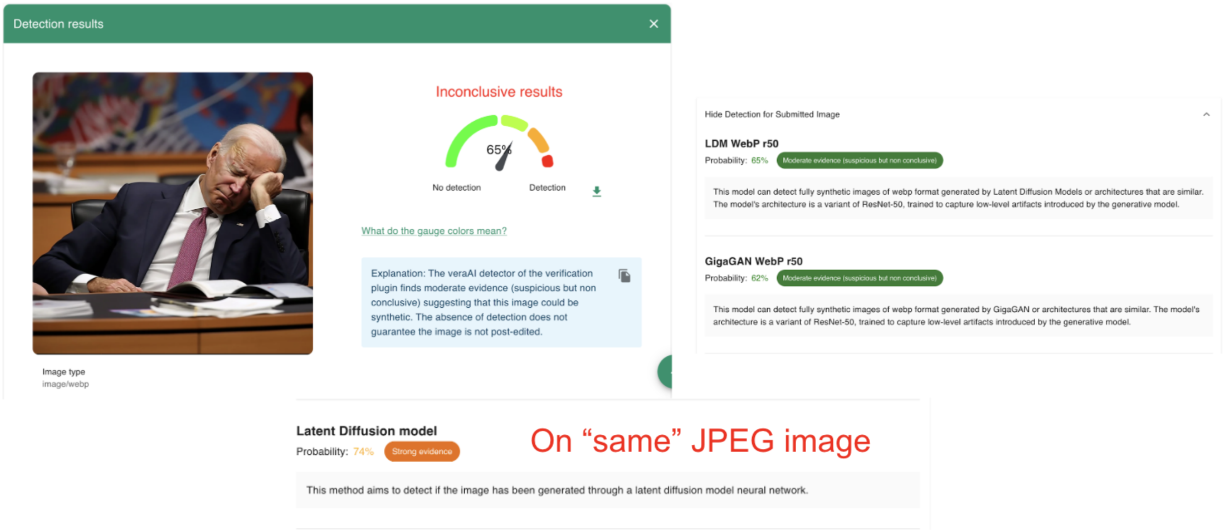
We found that AI-generated images were unlabelled, and their synthetic nature went unquestioned in comments. The findings of this data sprint reached several prominent media organisations with articles in Le Monde, Politico, Euronews and Demagog.pl.

In cycle 4, we focused on testing new algorithms specifically trained to detect WebP AI-generated images and the existing toolbox to detect Grok generated images, given that, unlike its competitors ChatGPT or Gemini, xAI[16] did not implement any guardrails, making it a good candidate to spread disinformation without any limitation.

To be able to compare our detection capabilities between JPEG and WebP images, we took images detected in previous cycles in the former format and converted them to WebP using Image Magick[17] before running the WebP algorithms on the resulting pictures.

Using this method, we found systematically a lower score in WebP format than in JPEG format, often below the empirical detection threshold of 0.70, as Figure 13 illustrates.

[16] xAI is the artificial intelligence company founded by Elon Musk that developed the conversational AI chatbot, Grok.
[17] https://github.com/ImageMagick/ImageMagick



Detection results
×
Inconclusive results
65%
No detection
Detection
What do the gauge colors mean?
Image type
image/webp
Hide Detection for Submitted Image
Moderate evidence (suspicious but non conclusive)
Moderate evidence (suspicious but non conclusive)
Strong evidence
On “same” JPEG image

Explanation: The veraAI detector of the verification plugin finds moderate evidence (suspicious but non conclusive) suggesting that this image could be synthetic. The absence of detection does not guarantee the image is not post-edited.

LDM WebP r50

Probability: 65%

This model can detect fully synthetic images of webp format generated by Latent Diffusion Models or architectures that are similar. The model's architecture is a variant of ResNet-50, trained to capture low-level artifacts introduced by the generative model.

GigaGAN WebP r50

Probability: 62%

This model can detect fully synthetic images of webp format generated by GigaGAN or architectures that are similar. The model's architecture is a variant of ResNet-50, trained to capture low-level artifacts introduced by the generative model.

Latent Diffusion model

Probability: 74%

This method aims to detect if the image has been generated through a latent diffusion model neural network.

*Figure 13 Annotated screenshot shared with WP3 on the WebP detection*

Then we compared the available algorithms by running a benchmark on a small dataset of 67 photorealistic images, split between 34 real and 33 synthetically generated. The synthetic images were retrieved from fact-checks for the images generated by Google’s AI, while others were generated by Grok using prompts written by Mistral AI that described real news agency pictures. LDM performed again best on the real and fake datasets, while ADM, LDM Rine and ProGAN Rine had respectively 6 (17,64%), 4 (11,76%) and 9 (26,47%) false positives making them unusable in a fact-checking context.

During the same period, several other publications featured the GenAI images detection such as AFP Factcheck[18] (also published on MSN and Yahoo), Factuel[19] and GhanaFact[20].

In cycle 5, we tested the newly released algorithms (ITW RINE and ITW SPAI from CERTH, ITW meaning In the Wild) using the same double dataset (fake and real) from the previous cycle. ITW Rine was integrated in the plugin version v0.84 in mid-February 2025 and ITW SPAI in v0.84.1 in early April. ITW Rine, after retraining during the cycle, and LDM performed best on detecting fake images (at around 50% detection), and ITW SPAI and LDM performed best by having zero false positives. ITW Rine had one unexplained false positive on a copy spreading on X of the viral AFP picture of Elon Musk performing a “Roman salute” after Donald Trump’s victory in the US Elections. Another key finding is that LDM showed better generalisation on newly created Grok3 pictures.

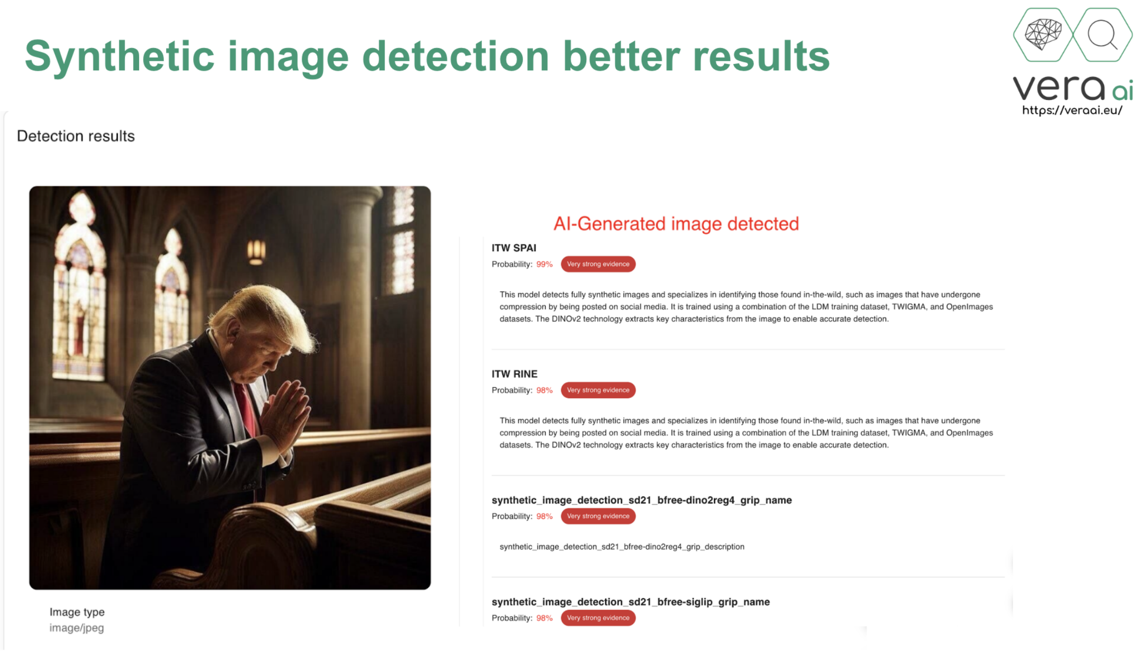
We also started to evaluate the automated AI-based label introduced by CERTH to replace the initial probability scoring and we used a development version of the plugin to test three more algorithms presented by GRIP-UNINA during the Naples February plenary meeting.

[18] https://factcheck.afp.com/doc.afp.com.36VU6WD

[19] https://factuel.afp.com/doc.afp.com.36UC3XN

[20] https://ghanafact.com/fact-check-photo-of-popular-nigerian-pastor-enoch-adeboye-in-an-islamic-veil-is-ai-generated/

Testing AFP real pictures, we unfortunately found several false positives but also much better results on previously undetected synthetic images (Figure 14).



Synthetic image detection better results

Detection results

AI-Generated image detected

ITW SPAI
Probability: 99% Very strong evidence

This model detects fully synthetic images and specializes in identifying those found in-the-wild, such as images that have undergone compression by being posted on social media. It is trained using a combination of the LDM training dataset, TWIGMA, and OpenImages datasets. The DINOv2 technology extracts key characteristics from the image to enable accurate detection.

ITW RINE
Probability: 98% Very strong evidence

This model detects fully synthetic images and specializes in identifying those found in-the-wild, such as images that have undergone compression by being posted on social media. It is trained using a combination of the LDM training dataset, TWIGMA, and OpenImages datasets. The DINOv2 technology extracts key characteristics from the image to enable accurate detection.

synthetic_image_detection_sd21_bfree-dino2reg4_grip_name
Probability: 98% Very strong evidence

synthetic_image_detection_sd21_bfree-dino2reg4_grip_description

synthetic_image_detection_sd21_bfree-siglip_grip_name
Probability: 98% Very strong evidence

Image type
image/jpeg

*Figure 14 View of a restitution slide of evaluation cycle 5*

Confronted with the dilemma of finding the best detection methods – some offering improved recall on synthetic images but concurrently increasing the false positive rate (thereby diminishing precision and potentially misguiding end-users) – we decided in our last evaluation cycle to automate our evaluations on a larger scale, across a more substantial volume of several hundred images, using our plugin middleware and a Python script. This effort incorporated not only the various datasets collected during the project but also a larger dataset made accessible by Truemedia.org[21].

As one of our main goals in this last cycle was to better evaluate the false positive rate, we used all the images real or fake with comments on their origin to investigate further the quality of the data. We found and corrected a few errors to finally keep a real image subset composed of 336 images and a fake image subset consisting of 535 images (both in JPEG format). We encountered several issues with the API as it returned errors up to 22% for some models due to the too small or too big file sizes (above 2Mpix for the GRIP family of models).

The following Table 6 shows the results obtained with the detection of our real images dataset. The main findings are:

- LDM has the lowest false positives (FPR) rate, half the score of the follower ITW RINE.
- The use of computed labels instead of probabilistic scores increases the FPR for ProGAN, LDM, ITW RINE, ITW SPAI and decreases it for the Bfree family.
- The new algorithms have a false positive rate ranging from 5.65% up to 16.07%.

[21] https://huggingface.co/datasets/nuriachandra/Deepfake-Eval-2024

As LDM did not show any false positives in our previous testing (and is the main algorithm accessible by end-users through the Verification Plugin), we checked the authenticity of the three photos detected as AI-generated. All of them are quite pixelated and are detected as "very likely resampled" by the Image Resampling Detection via Spectral Correlation and False Alarm Control of partner ENS. Early indexed copies in better resolution of one of those pictures, depicting a Turkish shooter at the Paris Olympic Games, were not detected as AI-generated. Another picture, released by AP news agency from Moscow Meshchansky District Court Press Service, featured a US citizen facing charges in a Russian tribunal. Equally pixelated, this picture is signed in the middle with a Russian tribunal QR code. The last one, two women holding a handmade banner (from 2020) in support of Donald Trump, shows various alterations in image forensics analysis.

Therefore, the authenticity and the provenance of those three "false positives" are not clearly demonstrated, which seems to acknowledge the precision of the LDM algorithm. At the time of this deliverable's writing, the investigation of the remaining algorithms' false positives is still ongoing to try to determine if we can safely open them to end-users. ProGAN Grip results are explained by the fact that this model only detects GAN-generated content, a technology much less common than the more modern diffusion models.

*Table 6 Results of the TrueMedia real images dataset (336 JPEG)*

| Algorithms / results | false positive | FPR (%) |
|---|---|---|
| Label ProGAN Grip | 33 | 9.82 |
| LDM GRIP | 3.00 | 0.89 |
| Label LDM Grip | 17 | 8.04 |
| ITW RINE | 6.00 | 2.61 |
| Label ITW RINE | 15 | 4.46 |
| ITW_SPAI | 19.00 | 5.65 |
| Label ITW_SPAI | 25 | 7.44 |
| B-Free | 47.00 | 13.99 |
| Label B-Free | 19 | 5.65 |
| DINO-multiGen | 19.00 | 5.65 |
| Label DINO-multiGen | 13 | 3.87 |
| B-Free-SigLIP | 54.00 | 16.07 |
| Label B-Free-SigLIP | 12 | 3.57 |

Regarding recall, the new algorithms performed much better. Table 7 below shows the results obtained on the TrueMedia fake images (with comments) JPEG dataset. Similarly, we present both the probabilistic scores and the AI-generated labels.

The main findings are:

- LDM has a low recall in the wild, which is not a surprise, as it is trained with fake images and forensic traces left by generators and we regularly recommend to fact-checkers to use early indexed previous copies of images to use this algorithm.
- The best recall rates are achieved by Bfree-dino and DINO-multiGen with scores above 60%.

*Table 7 Results of the TrueMedia fake images dataset (535 JPEG)*

| **Algorithms** | **Detections** | **Recall rate** |
|---|---:|---:|
| LDM GRIP | 91.00 | 17.01 |
| Label LDM Grip | 285.00 | 53.27102804 |
| ITW RINE | 287.00 | 52.28 |
| Label ITW RINE | 338.00 | 61.56648452 |
| ITW_SPAI | 280.00 | 51.00 |
| Label ITW_SPAI | 313.00 | 57.01275046 |
| B-Free | 358.00 | 66.92 |
| Label B-Free | 292.00 | 54.57943925 |
| DINO-multiGen | 347.00 | 64.86 |
| Label DINO-multiGen | 323.00 | 60.37383178 |
| B-Free (SigLIP) | 313.00 | 58.50 |
| Label B-Free (SigLIP) | 197.00 | 36.82242991 |

After participating in several conferences – EU DisinfoLab Conference in Riga (October 2024), Paris AI Summit (February 2025), and Infox sur Seine (March 2025) – with data scientists of the French public FIMI agency Viginum, we were contacted by them to test the vera.ai algorithms through the Verification Plugin middleware. This evaluation took place in July and August 2025.

Viginum tested 2,691 pictures, including the whole Deepfake-Eval-2024 dataset and the AMMeBa dataset[22] with in total 1,346 fake images and 1,345 reals. They also encountered errors returned by the API (12.6% for Grip's models and 4.3% for Mever's models).

Their results confirm on a larger scale the previous findings. LDM, beyond the specialized WebP model, achieves the lowest false positives but is being largely distanced by DINO-multiGen, ITW Rine and ITW SPAI in recall.

Figure 15 below shows the Receiver Operating Characteristic, or ROC curve of all available vera.ai models, aiming to assess their overall performance on the test dataset. The Area Under the Curve (AUC) gives a single score summarizing the overall prediction capacity of the models. The Mever models ITW SPAI and

[22] https://www.kaggle.com/datasets/googleai/in-the-wild-misinformation-media

ITW Rine achieved the best AUC (0.89) on the dataset, ahead of DINO-multiGen that have a lower false positive rate.

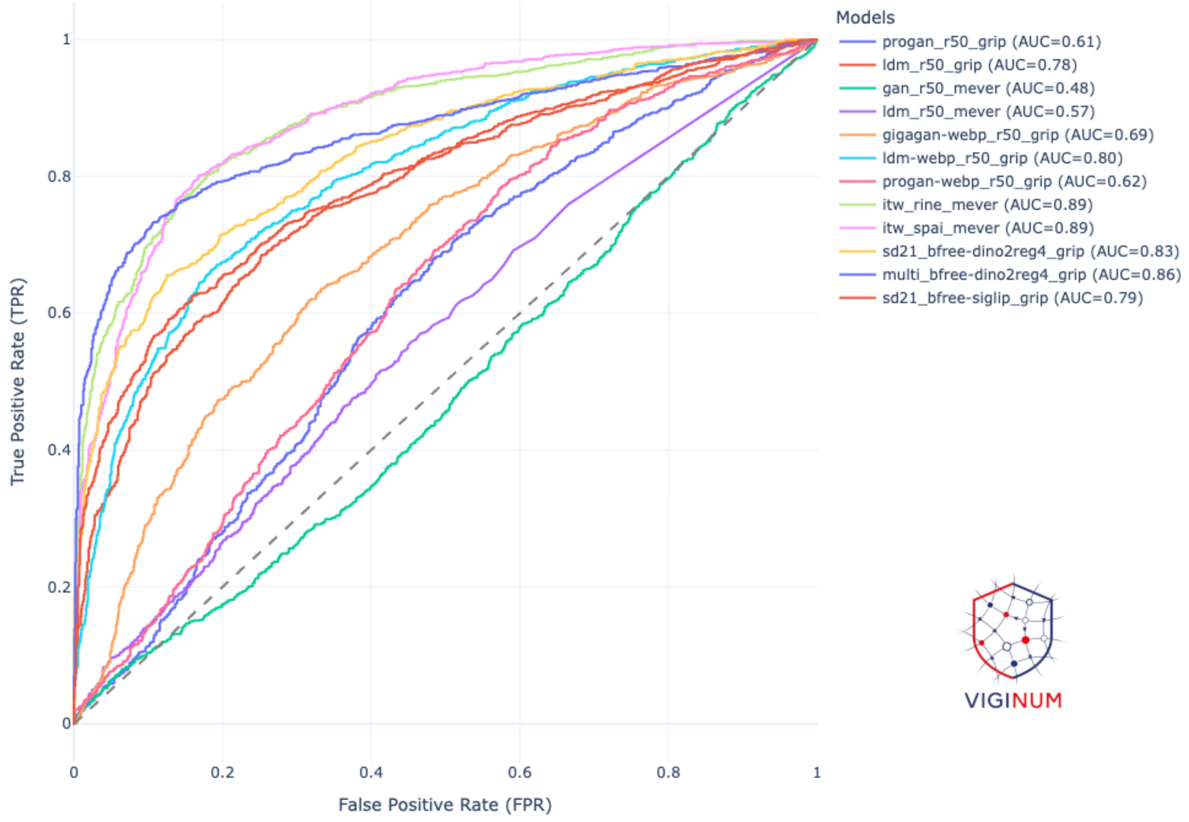


*Figure 15 ROC curves obtained by Viginum on the vera.ai synthetic images detection API*

The Precision-Recall Curves (Figure 16) give another interesting indication regarding the models' precision and therefore their capacity to avoid false positives. DINO-multiGen and ITW Rine appear on top with an Average Precision (AP) of 0.90 followed by ITW SPAI with 0.89.

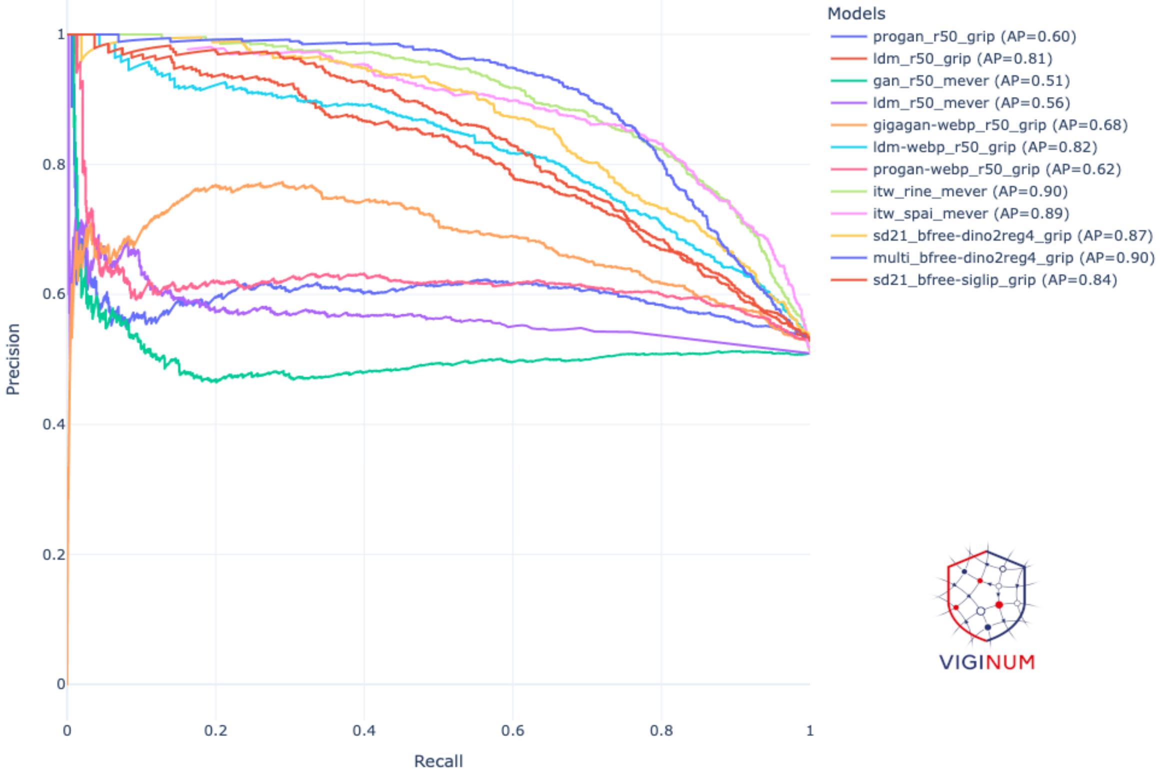


*Figure 16 Precision-Recall curves obtained by Viginum on the vera.ai synthetic image detection API*

Viginum provided us also with a result table (Table 8) of the same dataset for the probabilistic score of 0.70 that we took as an empirical threshold at the beginning of the evaluation cycles.

*Table 8 Results of Viginum's evaluation for a 0.70 detection threshold.*

| Models | accuracy | precision | recall (TPR) | FPR |
|---|---|---|---|---|
| progan_r50_grip | 0.48 | 0.78 | 0.01 | 0.00 |
| ldm_r50_grip | 0.57 | 0.94 | 0.20 | 0.02 |
| gan_r50_mever | 0.50 | 0.88 | 0.01 | 0.00 |
| ldm_r50_mever | 0.50 | 0.67 | 0.05 | 0.03 |
| gigagan-webp_r50_grip | 0.50 | 0.70 | 0.11 | 0.05 |
| ldm-webp_r50_grip | 0.59 | 0.91 | 0.24 | 0.03 |
| progan-webp_r50_grip | 0.48 | 0.95 | 0.01 | 0.00 |
| itw_rine_mever | 0.75 | 0.95 | 0.53 | 0.03 |
| itw_spai_mever | 0.74 | 0.91 | 0.55 | 0.06 |
| sd21_bfree-dino2reg4_grip | 0.75 | 0.84 | 0.66 | 0.14 |
| multi_bfree-dino2reg4_grip | 0.79 | 0.93 | 0.65 | 0.05 |
| sd21_bfree-siglip_grip | 0.72 | 0.84 | 0.59 | 0.13 |
| gigagan-webp_r50_grip_only_webp | 0.56 | 0.86 | 0.10 | 0.01 |
| ldm-webp_r50_grip_only_webp | 0.59 | 1.00 | 0.13 | 0.00 |
| progan-webp_r50_grip_only_webp | 0.53 | 1.00 | 0.02 | 0.00 |

Thanks to the above findings, we now have a better understanding of the overall performance of the models that will help us to further improve our user-facing tools in the concluding phase of the project.

In the final stages of preparing this deliverable, we achieved notable success[23] in detecting AI-modified content using vera.ai methods. This progress paves the way for wider adoption of these detectors among our advanced plugin users.

Although the project is officially complete, we have identified several key areas for further investigation.

Despite all the difficulties and limitations we faced, end-users valued the Synthetic Image Detection Service with a 3.44 score out of 5 on a Likert scale.

As we move towards project completion, our priorities are to focus on:

- Investigate false positives: Continue to investigate the source and cause of any false positives and explore the possibility to establish a detection threshold for selected models to verify the pertinence of the proposed labels (very strong evidence, strong evidence, moderate evidence and weak evidence).

[23] https://faktencheck.afp.com/doc.afp.com.76D269E , https://factuel.afp.com/doc.afp.com.74YM463 , https://napravoumiru.afp.com/doc.afp.com.77EM89E

- Deploy reliable methods: Release the most reliable and accurate detection methods to the broader fact-checking community.
- Improve UI explanations: Enhance the user interface with clear explanations to prevent any reproducible false positive cases.
- Evaluate re-sampling detection: Investigate incorporating a re-sampling detector into the pipeline. This could warn users about poor input image quality and help bridge the gap between a low-level technical anomaly and a meaningful, semantic alteration of the content.

## 4.2 Audio Provenance

Tracing the origins of audio content is often challenging, especially when working with datasets that combine files from many different internet or local sources, which may lack reliable metadata or clear descriptions. In such situations, it can be difficult to determine where a particular audio file comes from, or which version is closest to the original.

Audio provenance analysis addresses this challenge by offering a structured way to map relationships between audio recordings. The approach focuses on identifying segments that have been reused across different files and showing how these segments connect, ultimately creating a representation known as a provenance graph.

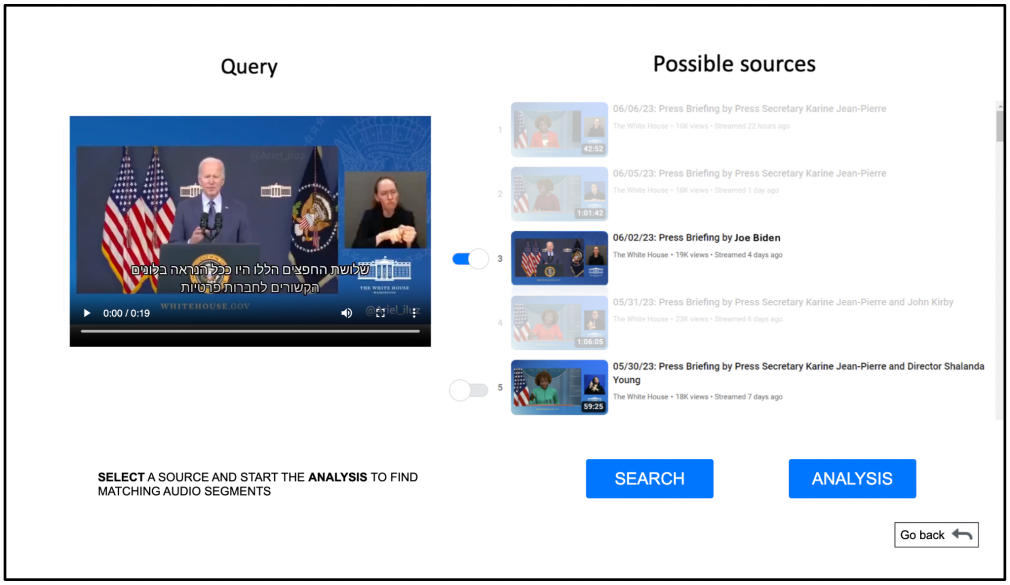
This process typically involves two key steps. First, provenance clustering, where reused material is detected and related audio files are grouped by detecting shared segments. Second, provenance graph building, where these relationships are organized into a graph that can illustrate how content has spread or been modified. Through these steps, audio provenance analysis helps to trace the reuse and transformation of audio material, which is particularly valuable for understanding how audio is manipulated or repurposed in cases of misinformation, disinformation, or decontextualization.

### 4.2.1 Participatory Evaluation

The participatory evaluation was conducted in May 2024, at a time when audio provenance analysis was still in its development phase. Consequently, only the first component – known as Provenance Clustering or Audio Reuse Detection – was presented to users and evaluated. Focus was on the question of how users interact with the tool, whether it is a useful addition to professionals, and where it needs refinement in order to become useful. Eight participants from France Télévisions, DW, Radio France, BR, and University of Siegen (3 journalists/fact-checkers, 1 news director, 3 innovation managers, 1 media research academic) were provided with user instructions, access to the prototype and respective material to test, which happened at hand of a user scenario. This was followed by a virtual exchange, in which feedback from testers was discussed.

Participants had several expectations, ranging from rather generic ones like the ease of use or the service to be trustworthy and transparent. Others were more concrete such as the wish for explained conclusions accompanied with respective highlights, the connection to an archive or ideally an attached reverse audio search.

Regarding the data upload, for the query input as well as possible source material, which functions as comparison audio, testers did not find this intuitive, nor was the used language easy to understand for them (see Figure 17).



1 06/06/23: Press Briefing by Press Secretary Karine Jean-Pierre
The White House • 15K views • Streamed 22 hours ago
42:52

2 06/05/23: Press Briefing by Press Secretary Karine Jean-Pierre
The White House • 18K views • Streamed 1 day ago
1:01:42

3 06/02/23: Press Briefing by Joe Biden
The White House • 19K views • Streamed 4 days ago

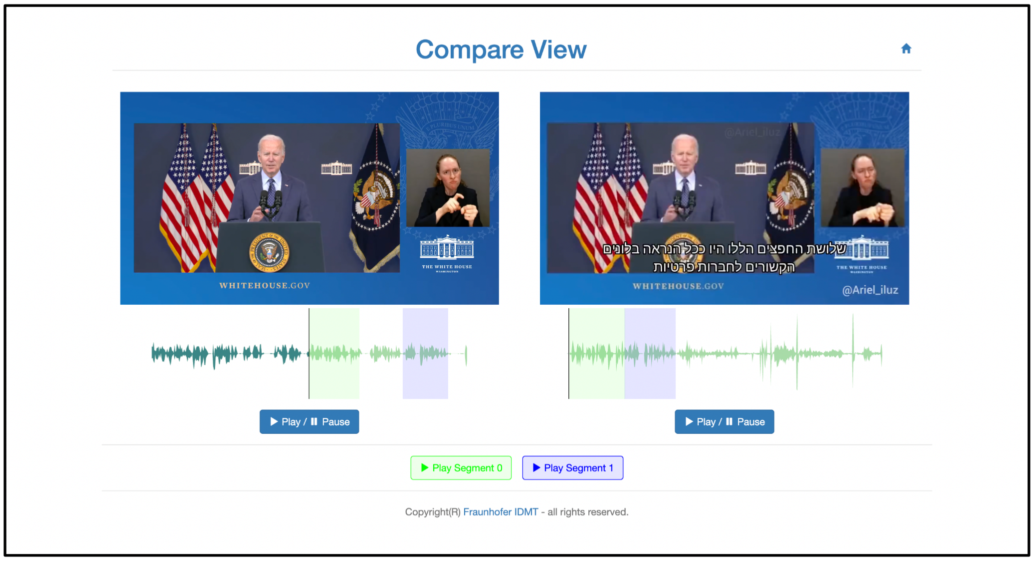
4 05/31/23: Press Briefing by Press Secretary Karine Jean-Pierre and John Kirby
The White House • 23K views • Streamed 6 days ago
1:06:05

5 05/30/23: Press Briefing by Press Secretary Karine Jean-Pierre and Director Shalanda Young
The White House • 18K views • Streamed 7 days ago
59:25

**SELECT** A SOURCE AND START THE **ANALYSIS** TO FIND MATCHING AUDIO SEGMENTS

*Figure 17 'Query' and 'Possible sources' page in the Audio Reuse Detection prototype*

They highlighted that the provision of possible sources seems a big burden and the process of narrowing down the list of potential sources was not well explained. For the analysis page the feedback was rather positive, as it was perceived as intuitive, simple to understand and to navigate through the audio waveforms. The colouring of matching segments is especially helpful in assessing the results (see Figure 18).



Copyright(R) Fraunhofer IDMT - all rights reserved.

*Figure 18 Analysis page and colouring of matching segments in the Audio Reuse Detection prototype*

.

Several ideas were provided on how to improve and enrich the service with new functionalities to make it a useful addition to user workflows:

- Automated search (such as a reverse audio search) or at least provision of hints for searching for possible sources.
- Allow for manual cutting of audio track or selection of a specific time frame of interest.
- Speech-to-text (transcript, subtitles) and translation functionality.
- Switch between the display of matching and non-matching segments.
- Add deepfake detection.

The overall result of the participatory evaluation of the prototype was that the tool could add value to the testers workflows, as it is easy to use and fast to learn. It additionally saves a lot of time by speeding up the process when analysing long audio files and it provides a technical proof next to the human assessment. However, the collection of suitable source material was perceived as too complicated in most cases and the fact that this material had to be archived locally and then uploaded compensated the saved time for analysis.

### 4.2.2 Design Thinking Evaluation

The Audio Provenance service aiming at detecting the provenance of an audio track within a dataset was released in May 2025 shortly before starting cycle 6. Its performance and usability were tested with the design thinking methodology.

The audio provenance service takes into input several audio files and gives back a JSON graph data file with nodes (for each audio fragment), links (or vertices linking those files) and clustered links providing the matching segments between the various files.

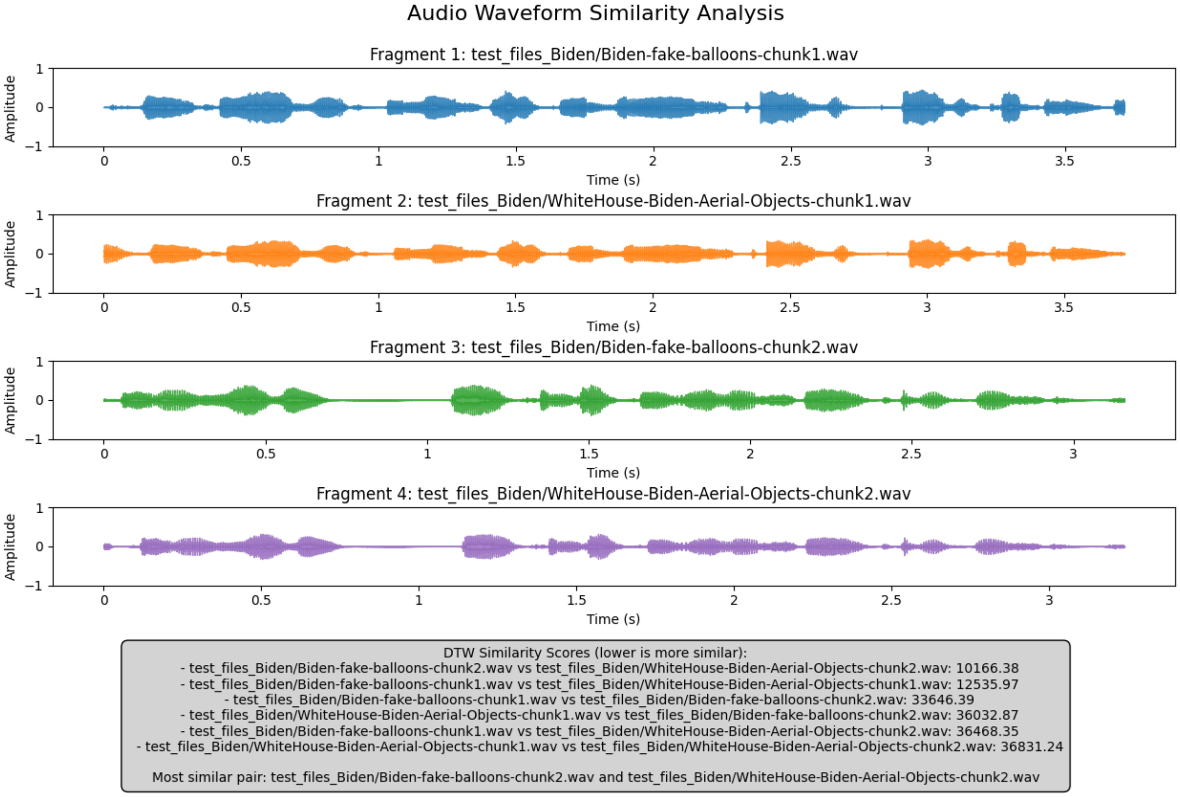
We first tested it with the “real life” use case coming from an AFP fact-check produced in March 2023[24], the same one which gave us the idea to propose building such a tool to automate the process of matching audio files segments. A video of the then US president Joe Biden, published by the Whitehouse official YouTube channel, had been remixed by a X account from Israel and tampered to trim his speech and making it inconsistent, as a kind of propaganda “proof” of a supposed senility.

As it happens frequently, the original video used to make the forgery was retrieved by a keyframes extraction and similarity search to identify the Whitehouse channel original video. Then, it took a couple of hours of manual analysis to review both videos and to identify, also by image similarity of the sign language interpreter near Mr Biden, the matching segments and to make a representation explaining the forgery by comparing both videos audio waveforms in an editing tool (Adobe Audition in our use case).

[24] https://sprawdzam.afp.com/doc.afp.com.33AC9GU

Using the results of our manual analysis of the above-mentioned AFP fact-check as ground truth, we found that the audio provenance service correctly matched both audio fragments by identifying the corresponding time sequences in both files, within just a few seconds of processing time.

Figure 19, made by the evaluator with an audio waveform similarity visualisation script written in Python, shows how the corresponding fragments match from one file to the other.



```
DTW Similarity Scores (lower is more similar):
- test_files_Biden/Biden-fake-balloons-chunk2.wav vs test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk2.wav: 10166.38
- test_files_Biden/Biden-fake-balloons-chunk1.wav vs test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk1.wav: 12535.97
- test_files_Biden/Biden-fake-balloons-chunk1.wav vs test_files_Biden/Biden-fake-balloons-chunk2.wav: 33646.39
- test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk1.wav vs test_files_Biden/Biden-fake-balloons-chunk2.wav: 36032.87
- test_files_Biden/Biden-fake-balloons-chunk1.wav vs test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk2.wav: 36468.35
- test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk1.wav vs test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk2.wav: 36831.24

Most similar pair: test_files_Biden/Biden-fake-balloons-chunk2.wav and test_files_Biden/WhiteHouse-Biden-Aerial-Objects-chunk2.wav
```

*Figure 19 Audio Waveform Similarity analysis performed with the Audio Provenance tool results*

The audio provenance service saves a lot of time for the end-user in this first use case as matching audio fragments between a short video posted on social networks and a longer previous video presser is very much time consuming.

Nevertheless, for building a proof of forgery, the end user still needs to use an editing software to compare the two audio files and reveal the manipulation by identifying, in this use case, the missing part circled in red in Figure 20 below which comes from the original fact-check article.

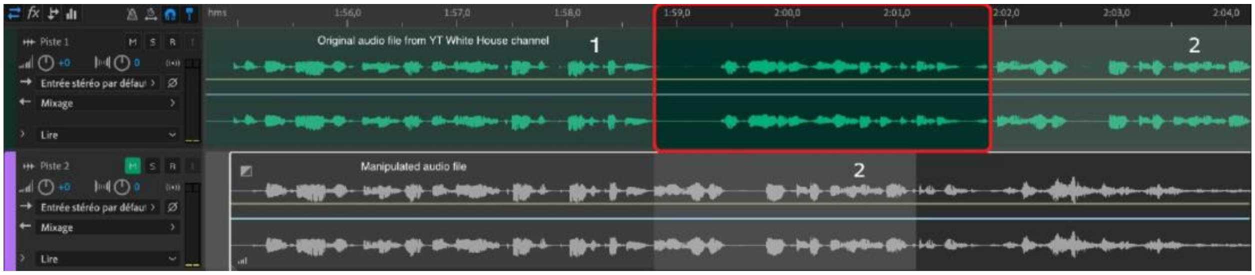


*Figure 20 View of the proof of the Biden's video forgery made by editing and removing an audio fragment*

A dedicated web application visualization is needed to enable detailed comparison of audio reuse and suppression between two or more audio tracks. To ease comparison, it is needed to present all the audio waveforms vertically stacked in the same timeline.

We then tested more complicated files like four videos of French president Emmanuel Macron being booed by the crowd, allegedly on May 8th, 2025.

While the audio provenance service captures well the copies between the various files, the phylogeny of those audio four fragments (from 11.4 till 15.72 seconds length) seems more complicated to establish through the following graph (Figure 21: also made by the evaluator with a Python script). This highlights an additional future direction: integrating an appropriate, interactive graph visualization of audio provenance results to facilitate easier tracking of the provenance of reused files and segments.

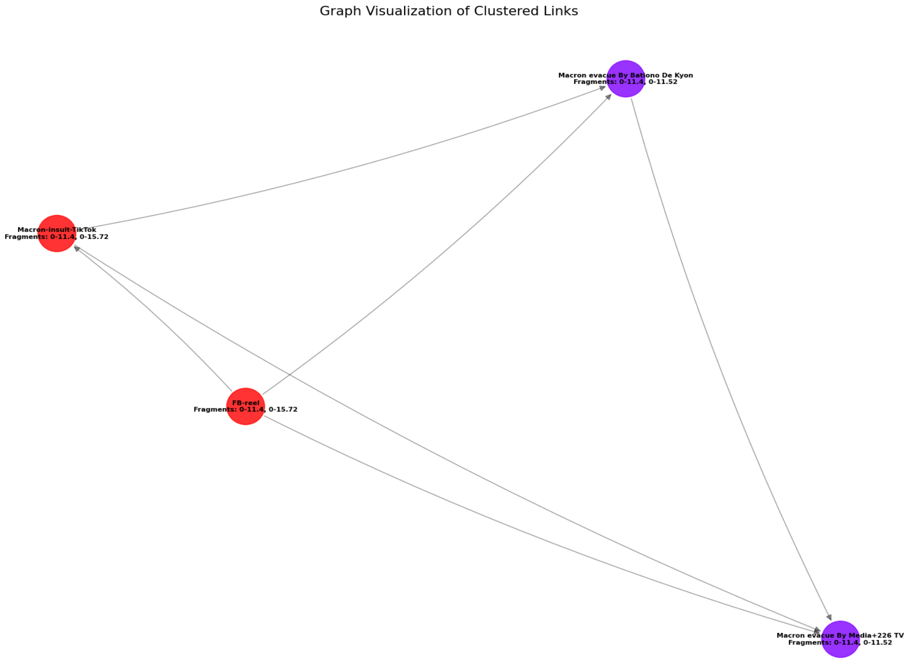

Graph Visualization of Clustered Links
Macron evacue By Bationo De Kyon
Fragments: 0-11.4, 0-11.52
Macron-insult-TikTok
Fragments: 0-11.4, 0-15.72
FB-reel
Fragments: 0-11.4, 0-15.72
Macron evacue By Media+226 TV
Fragments: 0-11.4, 0-11.52

*Figure 21 Graph visualization of the clustered links of the Audio provenance service results*

However, in a different factcheck published by AFP in Serbia[25], the tool failed to capture a two-second sentence spoken by a TV presenter who was commenting during the return to air immediately after the broadcast of the manipulated report. After iterative analysis with the developers, it was determined that the current version of the service was configured to detect segments with a minimum length of three seconds. By adjusting this minimum segment length parameter, the tool was able to detect the relevant segment in a subsequent iteration. Therefore, we emphasize as additional future work the importance of allowing users to adapt key algorithm parameters, such as segment length, based on specific use cases.

Finally, we ran a wider test, more for a research use case in media studies, by ingesting into the tool the seven hours of the Becoming Brigitte series, an absurd conspiracy story aired by US podcaster Candace Owens, targeting Brigitte Macron, the spouse of French president Emmanuel Macron. The goal was to find out whether the audio provenance tool could detect any reuse of audio footage during that video series.

We verified manually the tool's JSON file results and found among the 80 audio files that a total of 46 fragments were reused (42 podcast theme music, 2 ads pronounced by the podcaster and 2 mismatched fragments). The following graph visualization (Figure 22) illustrates the identified copies among all the audio fragments. This analysis of a larger dataset emphasizes the previously mentioned need for an interactive graph visualization solution to present the results of audio provenance analysis.

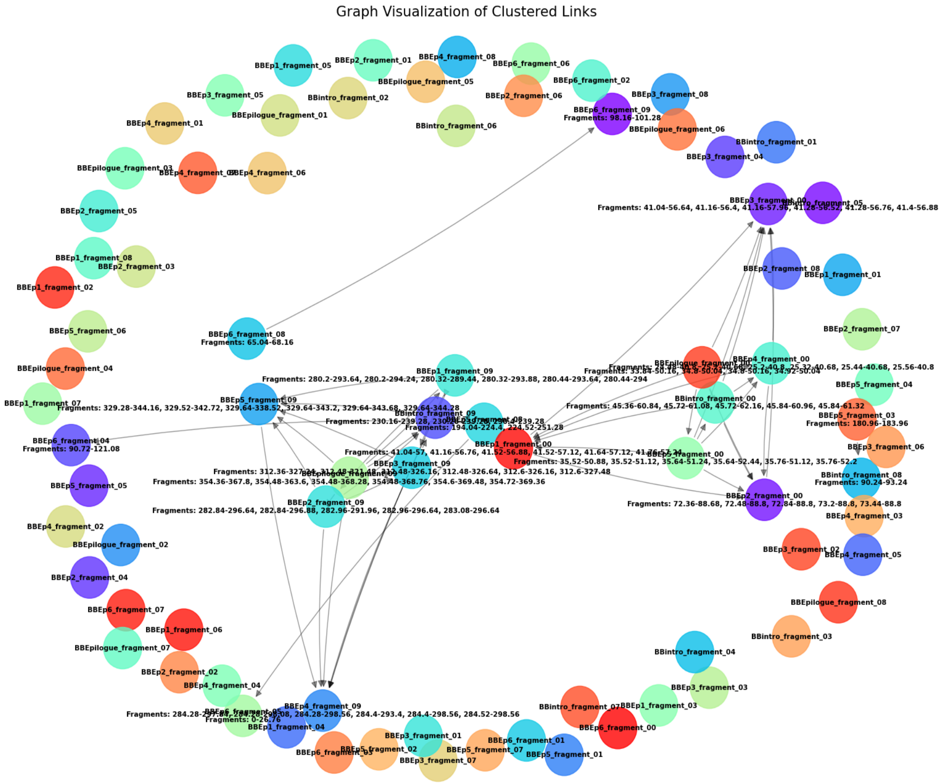


*Figure 22 Graph visualization of the audio fragments copies and provenance from a conspiracy video series*

[25] https://cinjenice.afp.com/doc.afp.com.68PZ8ZR

## 4.3 Synthetic Audio Detection

The Synthetic Audio Detection tool determines how likely parts of an audio file are to contain synthetic speech. It was evaluated by both the participatory evaluation process and the design thinking one, with the first focusing first on an early UX-focused interactive prototype with a short audio file and then on the version of the tool integrated into Truly Media, and the latter working with a larger number of previously debunked synthetic audio files of various lengths, using the API when made available. Here this section presents the results from both methodologies and their specific results, where initial insights about user interface and workflow from the participatory evaluation are refined and complemented with results from the design thinking evaluation, which also provides insights into the reliability of the detection models tested in this evaluation.

### 4.3.1 4.3.1 Participatory Evaluation

The participatory evaluation of the Synthetic Audio Detection tool stretched over two iterations. The first one – which took place in June 2024 and was based on an interactive UI/UX prototype coded in html and using a short audio file – focused on information visualisation and on the explanations of models and results. Seven testers working as journalists, fact-checkers and innovation managers (coming from France Télévisions, RTVE, ZDF, DW), familiarized themselves with the prototype based on user scenarios and tasks to interpret detection output. Testers provided written feedback that was then analysed.

Users had clear expectations of which aspects the service must fulfil for them to be useful regarding the evaluations' focus on visualisation and explanation. Most important to testers was transparency in how the service reaches its results, expected to come in the form of explanations accompanying the results and support in how to interpret these. Moreover, they wished for more general explanations on how the model works and what its limitations are. It was important to testers that the service does not only provide a single output, they expect a more fine-grained analysis on smaller chunks of the audio. Additionally, the service should not give the impression of it being absolutely sure, it was more important to understand the confidence of the service.

The feedback on this testing phase revealed that several of the initially collected expectations for this prototype were met. Key takeaways were:

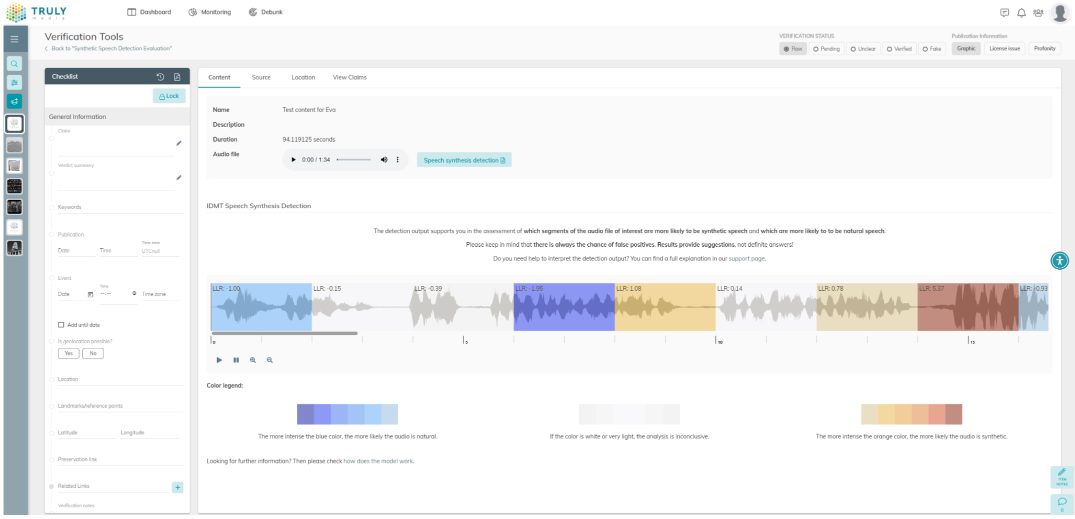
- Short explanation on **how to interpret values** (log-likelihood ratios[26] and uncertainty) together with a granular colour coding that supports easy interpretation of results and hence trust.
- A **combination of a minimalistic interface** that integrates aspects that foster users' trust into its design, with the option to find **more extensive explanations** on how the service works and what its limitations are works very well.
- Moving away from green and red for **colour coding** is promising as: 1) A synthetic audio detector is never a 100% sure of speech being natural or synthetic, this should therefore not be transported by the colours; 2) Blue is not loaded in any kind, while green implies "you are good to go".

[26] Log-likelihood ratios (LLR) in this service describe if the detected traces are more likely to appear in synthetic or in natural speech and how certain the service is of this assessment.

However, if something is not synthetic it can still be manipulated or dis-/misinformation of some kind; 3) People with red/green deficiency can easily use/interpret the detection output.

- Explaining **how the model works** and **how its results must be interpreted** – even in a rather brief way – helps users to **trust** the detection output.
- AI-based services should provide **different levels of explanations/transparency**: 1) very brief (as done here) for the first and fast glance; 2) extended version that still targets end-users with no special technical skills; 3) highly technical/scientific information (e.g. as provided in model cards).
- Being **transparent about potential uncertainty** of a detection output helps users to interpret the actual detection output and assess for themselves how much they want to rely on it.
- Providing **redundant ways of displaying information**, i.e. numbers and colour-coding.

In June and July 2025, the second iteration of the participatory evaluation of this service happened after its integration into the Truly Media platform (Figure 23), as a prototype with a functional analysis model. Insights from the design thinking evaluation – namely about horizontal scrolling – were taken into account for the interface design (see section 4.3.2). It aimed at exploring questions of usability and trustworthiness in a real-life setting and its integration into a user-facing platform. To make sure testers were able to give an assessment on how the tool fit into platform-specific workflows, this evaluation targeted organizations who use Truly Media in their daily work. In total eight participants from DW and ZDF took part. Backgrounds of testers spanned from journalists, information and archive experts to innovation managers, and researchers. Testers were granted access to the service in Truly Media and equipped with user instructions to guide the evaluation. They were assigned one out-of-the-wild example each which ranged in lengths – acknowledging the issue of audio file lengths raised in the design thinking evaluation (see section 4.3.2) – languages and quality to test (these were either AI, real, partially AI and one unknown). No further information on the respective audio was given. Additionally, testers were invited to try out additional self-chosen content items. Testing happened at a tester's own speed. Their feedback was provided in a written form.



Name Test content for Eva

Description

Duration 94.119125 seconds

Audio file 0:00 / 1:34

IDMT Speech Synthesis Detection

The detection output supports you in the assessment of **which segments of the audio file of interest are more likely to be synthetic speech** and **which are more likely to to be natural speech**.

Please keep in mind that **there is always the chance of false positives. Results provide suggestions**, not definite answers!

Do you need help to interpret the detection output? You can find a full explanation in our support page.

**Color legend:**

The more intense the blue color, the more likely the audio is natural.

If the color is white or very light, the analysis is inconclusive.

The more intense the orange color, the more likely the audio is synthetic.

Looking for further information? Then please check how does the model work.

*Figure 23 Synthetic Audio Detection results in Truly Media*

Overall, the expectations remained similar to the previous phase of the participatory evaluation, with a high focus on ‘trustworthiness by design’ aspects. However, testing in the integrated version at hand of real life examples, which are not always as unambiguous to interpret as for the prototype, revealed a series of additional aspects – positive ones as well as ones to further look into for upcoming versions of the service. In the following we highlight the most relevant aspects for participants:

- **Integration into Truly Media**: It is perceived as valuable to have this service within a platform already used by the testers, instead of having it standalone. The similar integration as other services fosters the ease of use and integration in users’ existing workflows. Additionally, Truly Media allows for easy collaboration with colleagues. This is highly appreciated by testers especially in this case of synthesis detection, which remains challenging and often requests expert knowledge.

- **Trust into detection outputs**: Participants had mixed opinions on whether they would trust the detection output. However, they highlighted that the provision of likelihood ratios with respective colour coding on short segments is a good way of displaying results. It gives users an indication where to have a closer look without giving the impression of a definite answer on the overall file. Ideas for improvement were about looking into additional ways to present the detection output, e.g. by displaying in a distribution graph how many segments are likely synthetic/natural.

- **Explanations**: The provided explanations on different levels are mostly perceived as sufficient but still rather technical. While effort should be made to facilitate explanations in non-scientific language, testers would also like to see them at different places in the interface. For example, it should be explained ‘how to interpret detection output’ and also ‘how the model works’ before actual results are displayed. This mostly accounts for first-time-users, thus the idea came up to provide a starter and an expert mode, where starters are guided more through the process and more extensive explanations are available. What several testers were missing is a more detailed explanation for individual segments and why they are assigned a certain LLR.

- **UI/UX and missing functionalities**: File upload should be possible via drag and drop as well as inserting a URL (currently, users select a file from folder). Established audio player functionalities like slowing down audio pace or easier movement in the waveform are requested. Transcription and translation functionalities are mandatory. The examples in this session came without any further contextual information on e.g. speakers or language. Assessing whether an audio is synthetic or not is already challenging when one understands what is said, it is however nearly impossible for foreign languages to the human listener. In the latter case especially, but in principle for any audio verification case, it is necessary to cross-check the services detection output with contextual information, such as the content of the audio.

The key takeaways from the first iteration of the participatory evaluation find support in the results of the second evaluation, and highlight the importance of **multi-level explanations** for AI models, of **redundancy** of results presentation, of **careful colour-coding**, and of **built-in transparency** to foster trust in AI-based services.

### 4.3.2 Design Thinking Evaluation

The initial technical implementation version of the Synthetic Audio Detection service was made available as a functional API in cycle 4. It was immediately tested through the design thinking methodology, as fact-checkers and researchers experienced a rise of "audio deepfakes" (voice impersonation through AI, also called voice cloning) in the previous months. We therefore tested "real life" examples from previous investigations such as voice cloning of the then US President Joe Biden, of the Ukrainian general Valerii Zaluzhny, of Oleksiy Danilov, the former secretary of the National Security and Defence Council of Ukraine, French President Emmanuel Macron, ex and future US President Donald Trump, and other audio impersonations of Elon Musk and Mark Zuckerberg in audio fakes crypto-currencies scams.

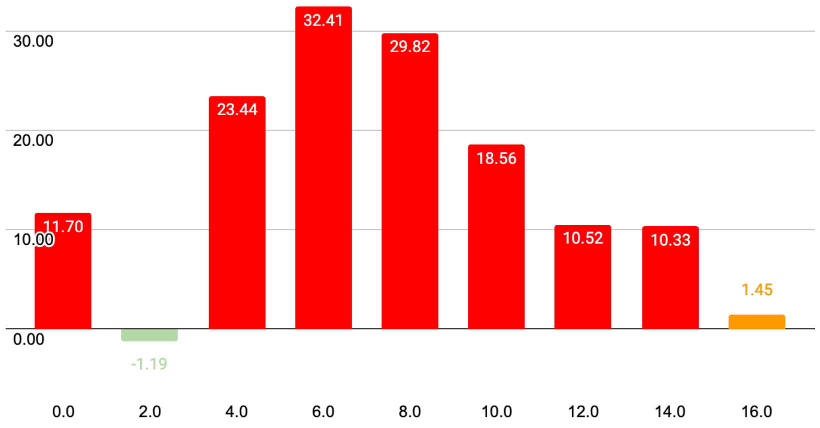


*Figure 24 Chart showing the detection result obtained on Macron's voice cloning. The x axis shows the fragment duration in seconds, and the y axis shows the detection score*

Figure 24 above shows the result of the detection on a short audio file falsely announcing on Facebook the resignation of French President Macron. The x axis shows the fragment duration in seconds, while the y axis shows the detection score by chunks of two seconds measured through the LLR commonly used in audio forensics to present a probabilistic statement. The higher the result, the more likely is the synthetic generation. Here with a median LLR result of 11.70, the file is clearly detected as synthetically generated.

AFP internally tested 10 audios proven synthetic, establishing ground truth from previous fact-checks debunked through third-party tools (deepfake-total and Loccus/Hiya). The feature achieved on that sample (see Figure 25), 40% detection, 20% partial detection (only a part of the audio track was detected as synthetic) and 40% non-detection.

We then did accuracy testing with a sample of 11 real voice fragments (mainly Donald Trump, Joe Biden, Kamala Harris, Emmanuel Macron, Nancy Pelosi). We obtained a high false positives rate (half of them partial) as outlined by Figure 26 below. This result raises concerns about the feature's usability in a fact-checking production environment due to the risk of errors and reputation damage. It also reveals a disconnection between the model theoretical false positive confidence rate and the results on our tests on real data.

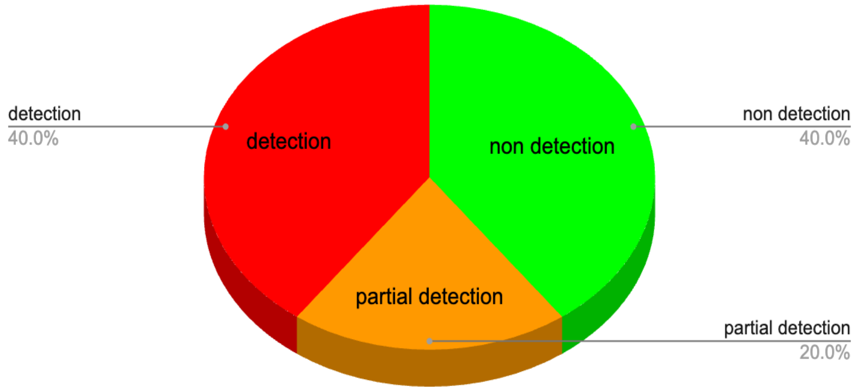


*Figure 25 Results chart on 10 synthetic audio files*

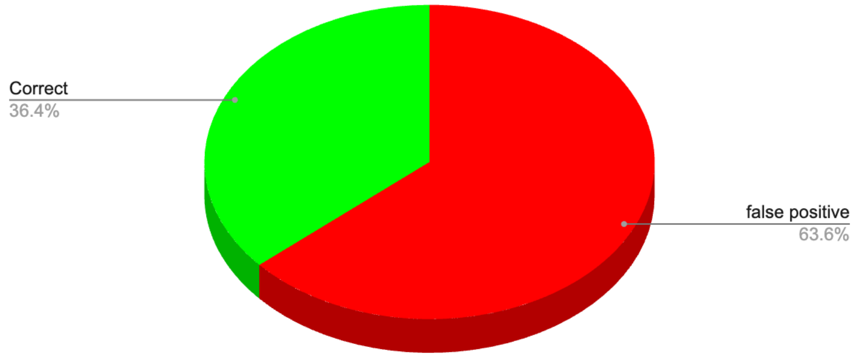
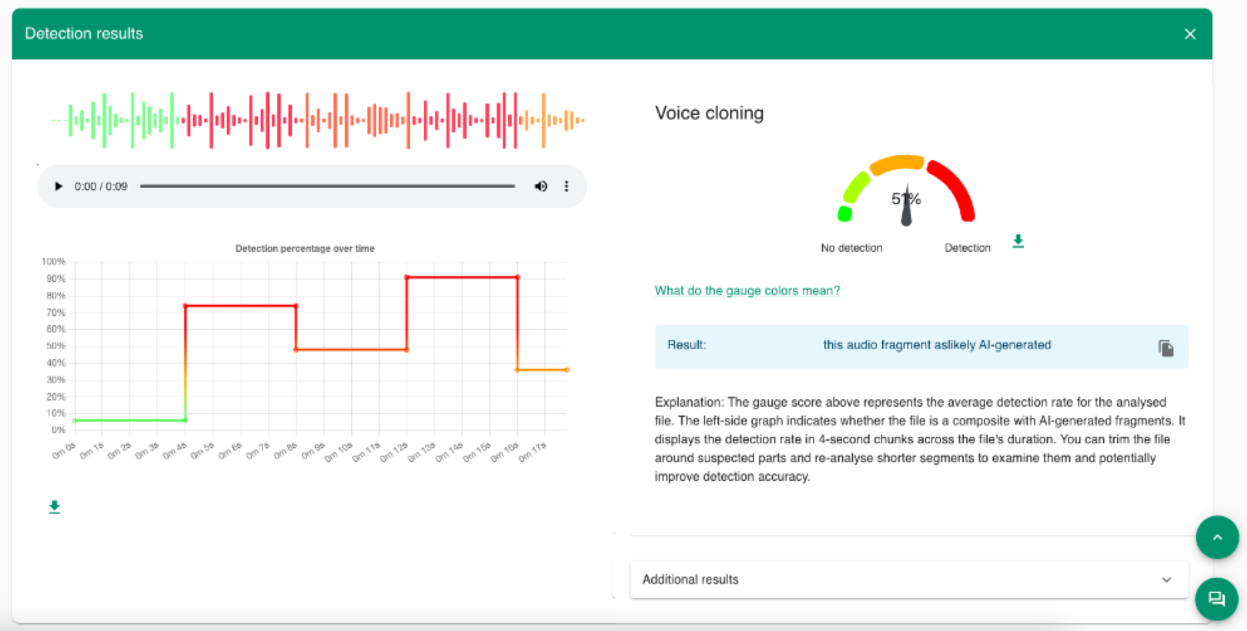


*Figure 26 Accuracy results chart on real voices audio files*

Regarding the usability of the initially proposed mock-up interface, we recommended to IDMT to avoid disclosing a theoretical false positive rate. Such communication, we learned from user feedback, fosters user reluctance to utilize the tool in their production setting. Our understanding is that false positive rate should be computed through thorough evaluation to inform the decision to make the tool available (or not) to end-users, thereby preventing misuse.

We equally recommended providing an average or median (as suggested by IDMT) global score to avoid using a multiple-chunk assessment as the latter complicates the interpretation of the overall outcome for the end-user. As an end-user said in our evaluation: a global score "is the first thing you look at". Additionally, the inclusion of multiple colours and horizontal scrolling, particularly unsuited for longer duration files, adds unnecessary complexity to the task.

To summarise our first evaluation findings in cycle 4, we suggested another interface mock-up for the Verification Plugin which incorporates a global score, a gauge and a detection scale, as well as a double view of the overall detection along the file duration, through a waveform and a chart using a more intuitive colour light traffic signal (Dobber et al., 2025) for universal understanding and reducing cognitive load (Figure 27).



Result: this audio fragment aslikely AI-generated

Explanation: The gauge score above represents the average detection rate for the analysed file. The left-side graph indicates whether the file is a composite with AI-generated fragments. It displays the detection rate in 4-second chunks across the file's duration. You can trim the file around suspected parts and re-analyse shorter segments to examine them and potentially improve detection accuracy.

*Figure 27 An alternative interface for Synthetic Audio Detection for the Verification Plugin*

In cycle 5, a second version of the service was provided by IDMT. Further internal testing showed a slight reduction in the false positive rate, but at the cost of a significant drop in recall. Out of 10 real voice samples, we found 6 correct assessments and 4 false positives (Figure 29). On a sample of 17 synthetic audio files, we found 23.5% of detected AI generated files (mostly partially like the Figure 28 example below) and 76.5% of non-detection (Figure 29).

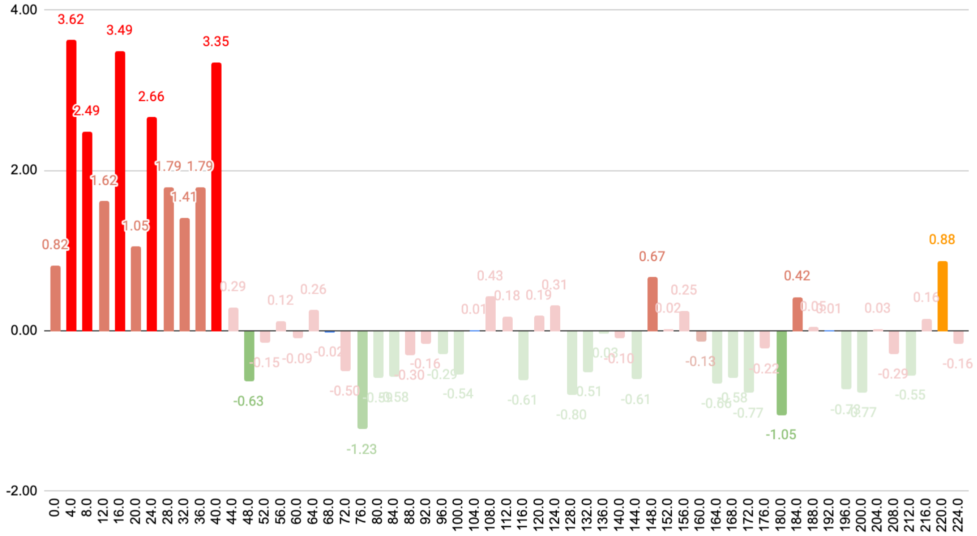


*Figure 28 A partial detection on Giorgia Meloni's audio deepfake falsely promoting a crypto-currency scam*

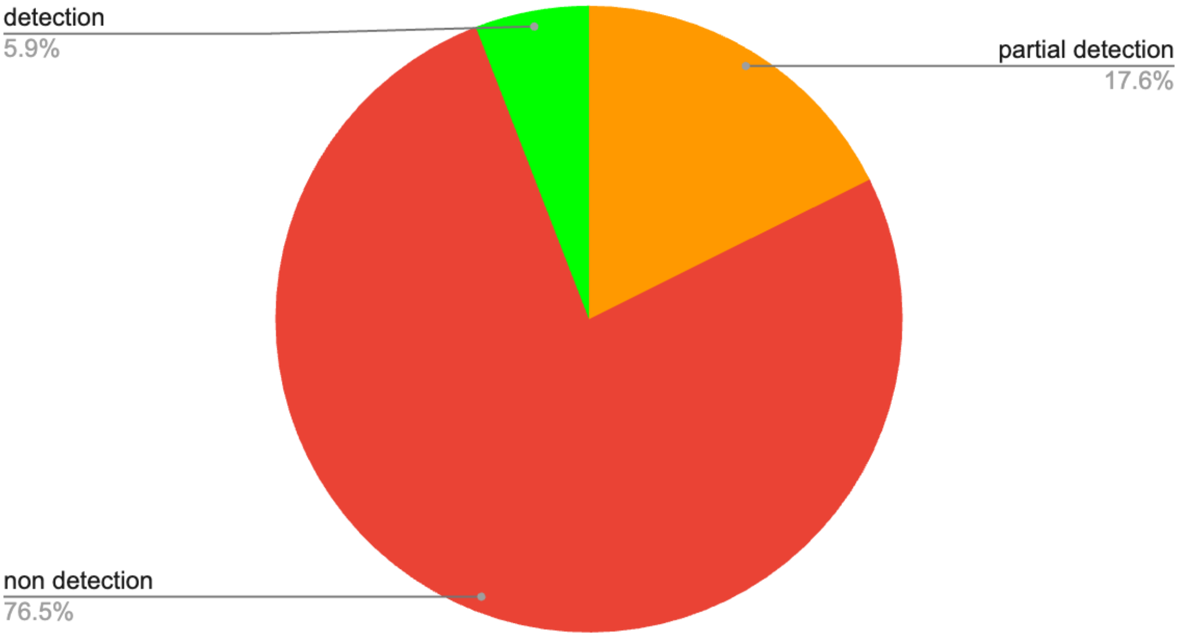


*Figure 29 Detection results of audio deepfakes samples with the Synthetic Audio Detector second version*

.

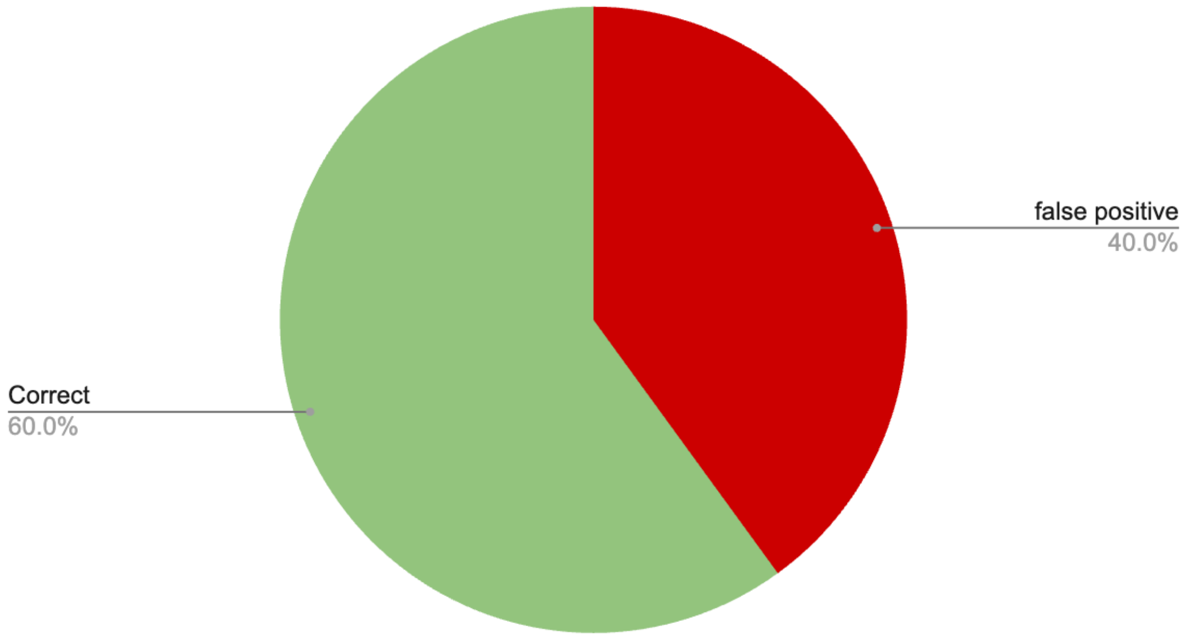


*Figure 30 Results of accuracy testing of the second version of the Synthetic Audio Detection Service*

The results (Figure 30) obtained at the time of this deliverable's writing suggest that the tool is not yet ready for release to end-users in a production environment.

To conclude this section, Fraunhofer IDMT's Speech Synthesis Detection (SSD) was designed to combine performance with transparency by guiding the model with hypotheses about speech characteristics and presenting results as LLRs rather than a single aggregated score.

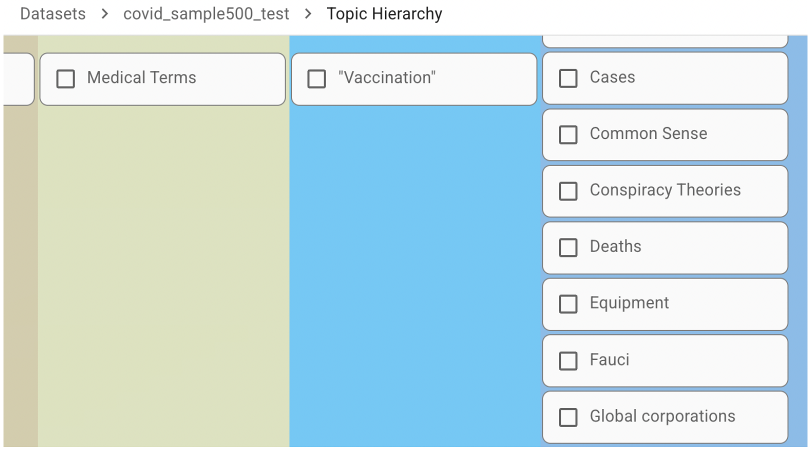
This approach resonated well with the first evaluation participants, but the second evaluation also revealed a substantial drop in performance on real-world examples compared to benchmark datasets such as ASVspoof, reducing accuracy to a level far below operational use.

Some issues were linked to training material and have since been fixed. However, broader challenges affect the entire field: real-world examples are rarely annotated in enough detail to reconstruct how they were created, making it hard to analyse failure cases, and attacker scenarios are not sufficiently reflected in public benchmarking datasets, as shown by the strong deviation observed for SSD. A deeper understanding and modelling of these factors will be key for improving real-world performance of detectors.

Further work will therefore focus on a deeper understanding and modeling of attack scenarios and ensuring they are better represented in datasets and testing procedures. In parallel, ongoing improvements at IDMT include enhanced robustness in audio processing, advancing a "net speech" approach, and the development of person-of-interest (POI) methods. These next steps will be pursued in collaboration and alignment with AFP, Deutsche Welle, and other vera.ai partners, and will continue beyond the current project. Speech synthesis detection will remain an evolving research field, with technological advances presenting persistent challenges. Sustained research is needed to achieve solutions that are both effective and trustworthy.

## 4.4 Topic Modelling and Temporal Analysis

The Topic Modelling and Temporal Analysis tools (developed by USFD) helps track the temporal evolution of a topic across published articles. The former builds topic trees through which users can navigate to explore the relationships between topics covered by news articles gathered in a CSV file, and the latter helps visualise these relationships through time to analyse how stories have evolved (see Figure 31Figure 32 andFigure 33).



Datasets › covid_sample500_test › Topic Hierarchy

□ Medical Terms

□ "Vaccination"

□ Cases

□ Common Sense

□ Conspiracy Theories

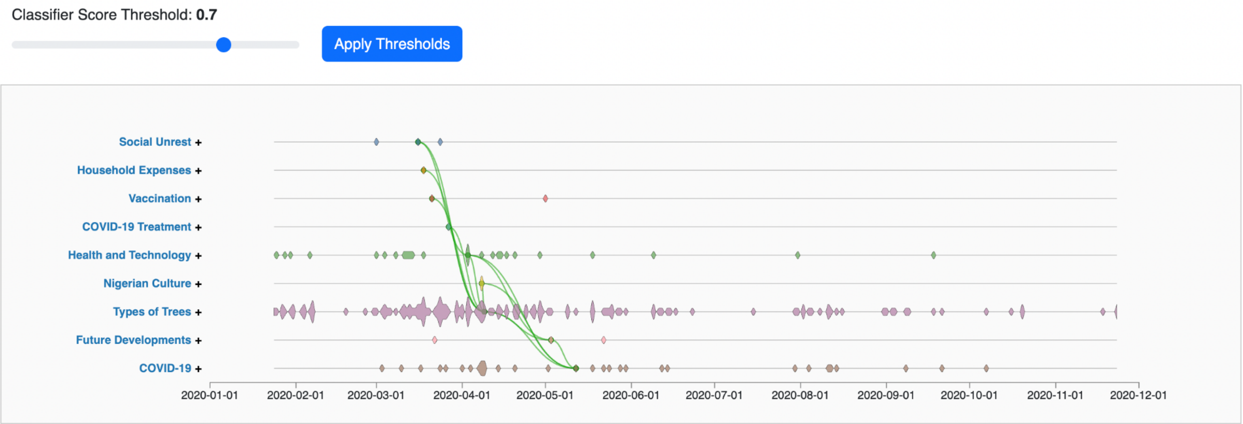
□ Deaths

□ Equipment

□ Fauci

□ Global corporations

*Figure 31 Topic Modelling tool tree structure*



Classifier Score Threshold: **0.7**

*Figure 32 Temporal Analysis tool: Visualisation of topic relationships through time*

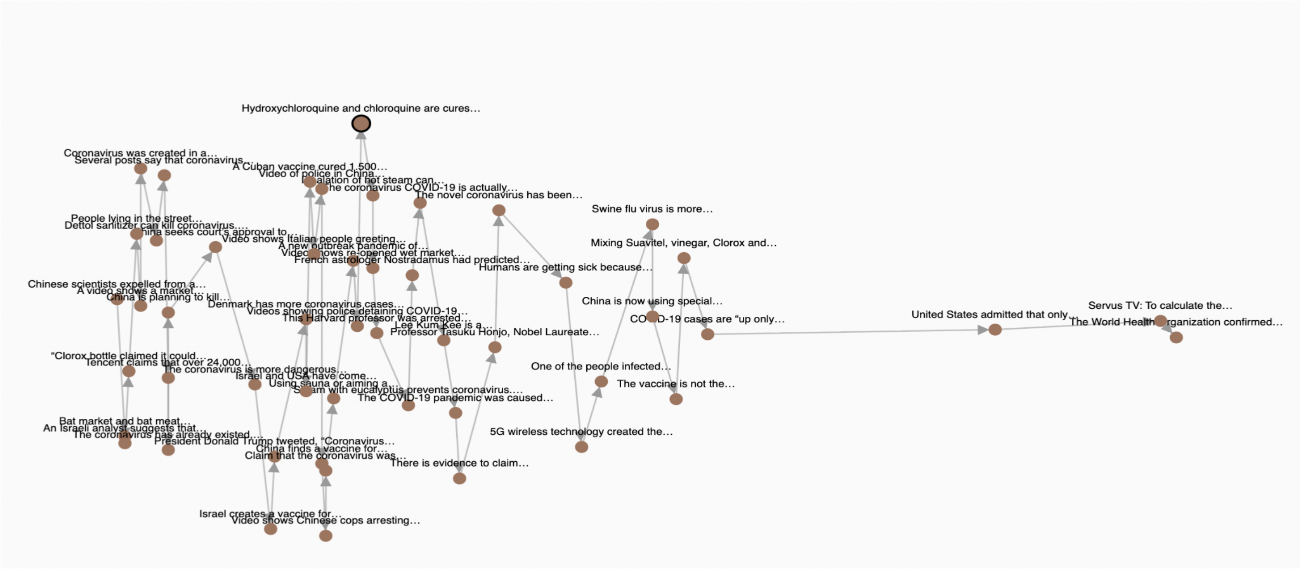

Hydroxychloroquine and chloroquine are cures…
Coronavirus was created in a…
Several posts say that coronavirus…
A Cuban vaccine cured 1,500…
Video of police in China…
Inhalation of hot steam can…
The coronavirus COVID-19 is actually…
The novel coronavirus has been…
Swine flu virus is more…
People lying in the street…
Dettol sanitizer can kill coronavirus.…
China seeks court's approval to…
Video shows Italian people greeting…
A new outbreak pandemic of…
Video shows re-opened wet market…
French astrologer Nostradamus had predicted…
Mixing Suavitel, vinegar, Clorox and…
Humans are getting sick because…
Chinese scientists expelled from a…
A video shows a market…
China is planning to kill…
Denmark has more coronavirus cases…
Videos showing police detaining COVID-19…
This Harvard professor was arrested…
Lee Kum Kee is a…
Professor Tasuku Honjo, Nobel Laureate…
China is now using special…
COVID-19 cases are "up only…
United States admitted that only…
Servus TV: To calculate the…
The World Health Organization confirmed…
"Clorox bottle claimed it could…
Tencent claims that over 24,000…
The coronavirus is more dangerous…
Israel and USA have come…
Using sauna or aiming a…
Steam with eucalyptus prevents coronavirus.…
The COVID-19 pandemic was caused…
One of the people infected…
The vaccine is not the…
Bat market and bat meat…
An Israeli analyst suggests that…
The coronavirus has already existed…
President Donald Trump tweeted, "Coronavirus…
China finds a vaccine for…
Claim that the coronavirus was…
There is evidence to claim…
5G wireless technology created the…
Israel creates a vaccine for…
Video shows Chinese cops arresting…

*Figure 33 Temporal Analysis tool: Evolution of topics through time across news articles*

### 4.4.1 Participatory Evaluation

The participatory evaluation of these two tools took place in two phases, in June and August 2025. The first phase focused on the second view of the Temporal Analysis tool (see Figure 33), i.e. the visualisation of pre-processed "corpuses" of news articles and their relationship through time. Since it was not yet in an advanced-enough stage to support user experimentation, it was demoed to a group of end-users from the participatory evaluation pool of users (this time from Radio Romania, Radio France, Atresmedia, Deutsche Welle, RTVE, University Tecnológico de Monterrey and MIT) for initial feedback. In the second phase (which followed an internal usability evaluation), a more advanced prototype of the Temporal Analysis tool, and additionally one of the Topic Modelling tool, were made available for testing: together, these allowed to explore tree hierarchies of extracted topics from a CSV file of news articles, and to visualise their relationships through time as well.

Due to the early stage nature of the prototypes, the participatory evaluation of these two tools was mainly focused on two aspects: the usefulness of the tools and their usability. While the testers found the tools potentially useful and promising, they had difficulties imagining concrete use cases for them, and they found that several usability issues made the tools complicated to use and interpret. In their further development, these tools could therefore follow the design framework – outlined in subsection 2.1.2 of the deliverable overview – in order to simplify functionalities and focus on aspects that help the tools fit into workflows and provide trustworthy and meaningful results.

### 4.4.2 Design Thinking Evaluation

The Topic Modelling and Temporal Analysis tools were evaluated internally by the AFP team after cycle 6 in several iterations with USFD during the summer 2025.

For this evaluation, we tested a subset of the same CrowdTangle dataset on "Ursula von der Leyen" used also in the Coordinated Behaviour Detection section (see section 4.7). After this latter tool identified signs of coordinated behaviour in this dataset, the main goal was to identify any potential spread of disinformation and to assess how much this coordination could be inauthentic, by examining how those narratives evolved in time.

We only kept from that initial dataset the date column and the text column (mapped on the Facebook posts) for the topic analysis. We removed with a regular expression formula all blank cells (as indicated in the guidelines) and protectively all the emojis to avoid any issue with the language processing pipeline.

A first subset of +6,000 posts was tested but the service went into error. We then tried a second subset of 5,632 posts and although it did not end either, we found that we could export a CSV file with 5,478 annotated messages after several hours of analysis.

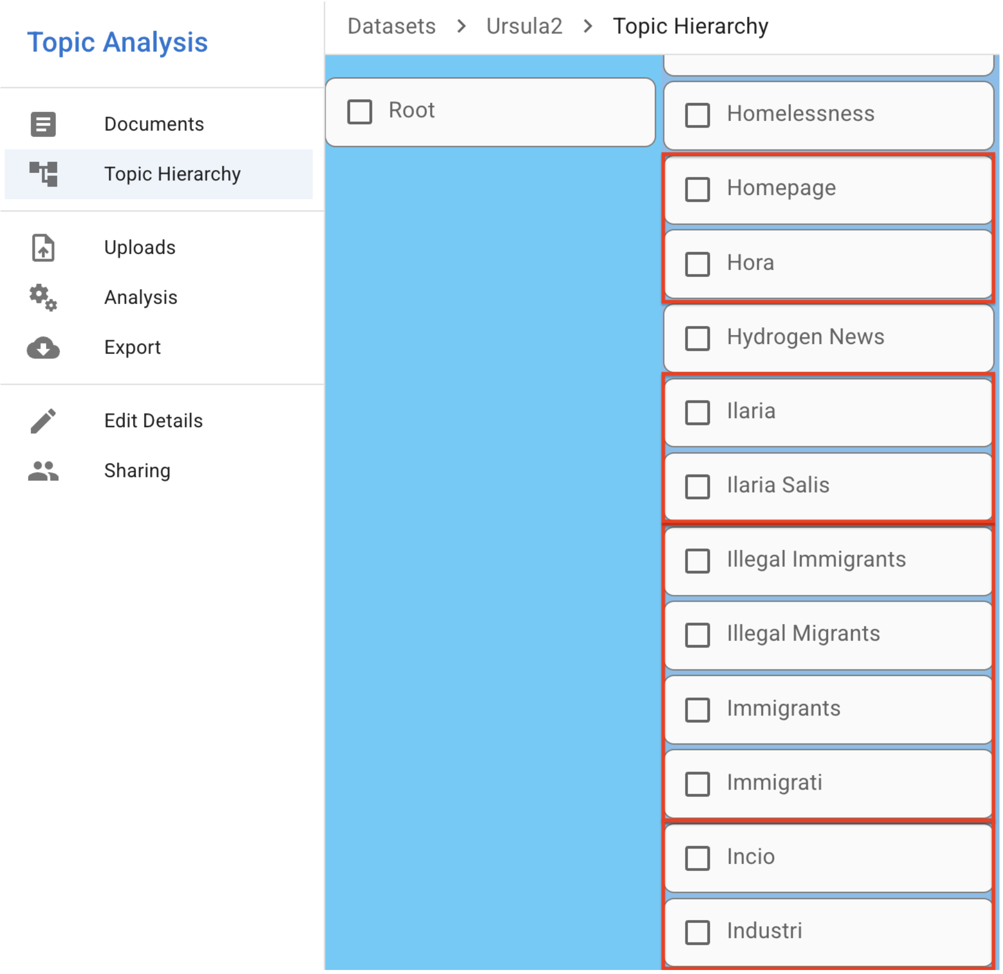
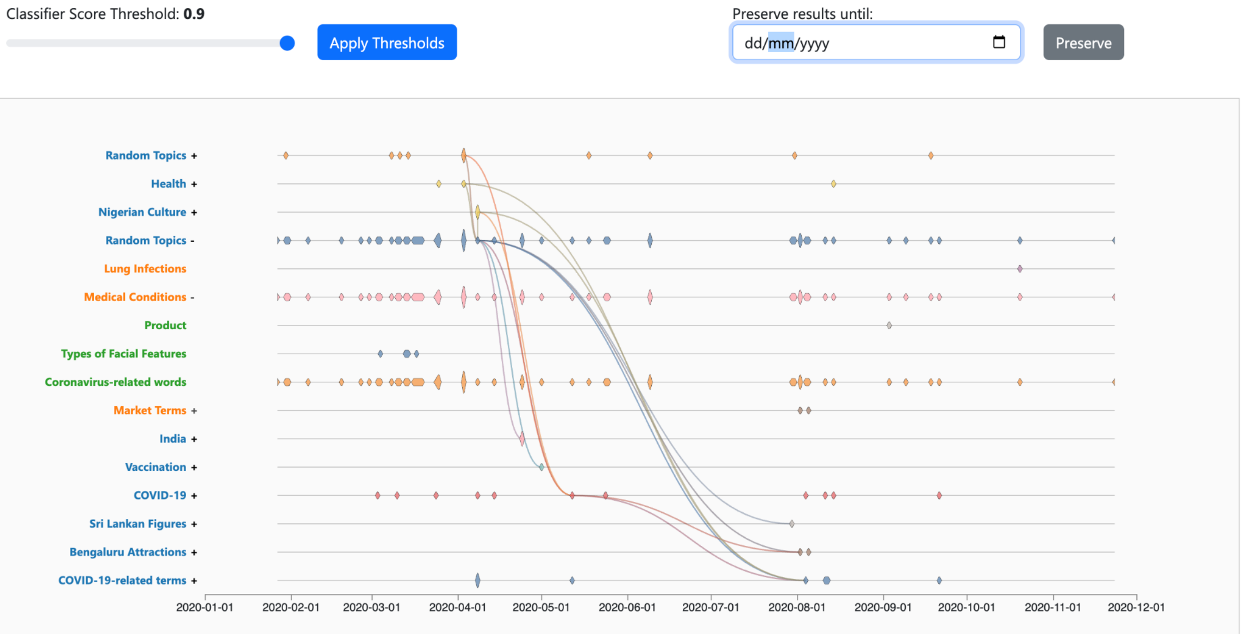


*Figure 34 Screenshot of the detected topics with a red box around the problematic ones*

The examination of the topics shows various issues (as exemplified in Figure 34 above):

- Some common keywords like homepage or hora (hour in Spanish) are identified as topics
- Hydrogen News seems to refer to a couple of web publications beyond hydrogen as a subtopic of energy.
- Ilaria Salis is not a topic but a person who, as an Italian member of the European parliament, should be detected (and not her first name) as a named entity.
- Several extracted topics are redundant (migrants, immigrants, immigranti, etc) and should be clustered
- The expression "illegal (im)migrants" has a loaded and pejorative connotation associated with the alt-right because it frames a human being as inherently illegal, rather than simply having a legal status that is non-compliant with local legislation. Therefore, this kind of expression should not be used as a topic but rather as a linguistic marker of a certain type of discourse.
- All accentuated letters are removed during the processing pipeline as exemplified by the words Incio (Início carries an accent in Portuguese) or Industri which seems to come from various verbal forms of "to industrialise" in Catalan, Spanish or Italian. We found many other examples of a systematic removal of accentuated letters in the topic list and in the processed documents. This may have had a negative impact on the dataset topic modelling.
- The interface offers to browse documents on each topic, but this functionality did not work in our tests (Chrome v.140.0.7339.214 (Official Build) (arm64) on Mac M1),
- There is a mix of language specific words, categories, named entities and common words that make the topic hierarchy unnecessarily complex.

The Temporal analysis could not be tested with the same dataset because the service falls into error. We examined an example provided by USFD on Covid-19 pandemic. To try to remove redundant topics and maximize accuracy, we used a maximal classifier score threshold of 0.9 as shown in Figure 35.

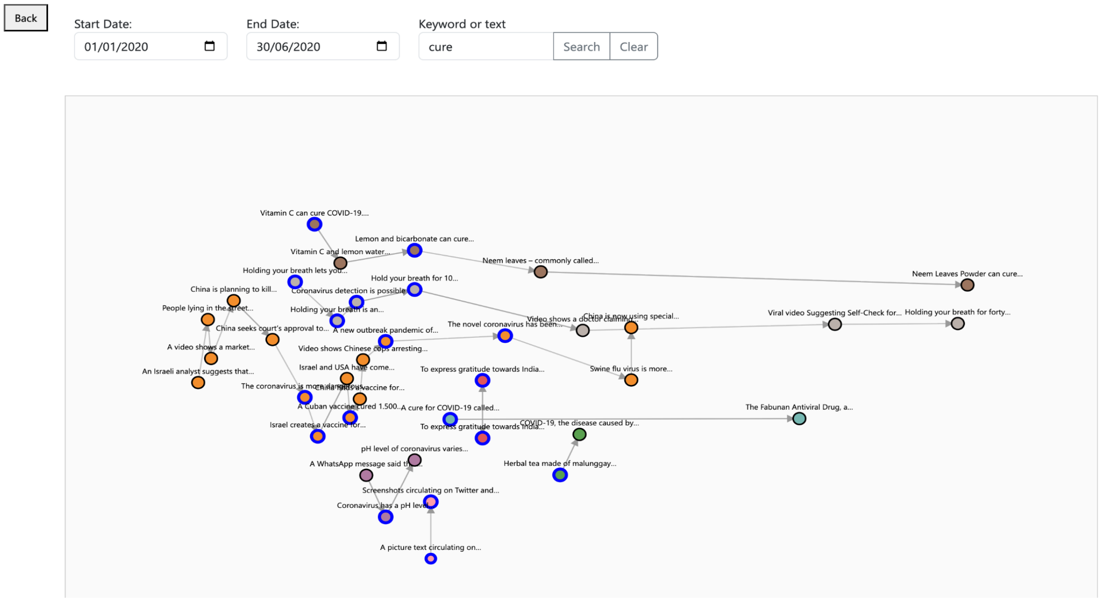


Classifier Score Threshold: **0.9**

Preserve results until: dd/mm/yyyy

*Figure 35 Screenshot of the Temporal Analysis service with a 0.9 classifier score threshold*

The result shows that at this score we still have two categories of random topics. Most documents were classified as random topics > medical conditions > coronavirus related words. By double-clicking on the timeline nodes, we entered a graph displaying the evolution of the underlying stories as shown in Figure 36.



Start Date: 01/01/2020

End Date: 30/06/2020

Keyword or text: cure

Back
Search
Clear
Vitamin C can cure COVID-19....
Lemon and bicarbonate can cure...
Vitamin C and lemon water...
Neem leaves – commonly called...
Holding your breath lets you...
Hold your breath for 10...
Neem Leaves Powder can cure...
China is planning to kill...
Coronavirus detection is possible
People lying in the street...
Holding your breath is an...
Video shows a doctor claiming...
China is now using special...
Viral video Suggesting Self-Check for...
Holding your breath for forty...
China seeks court's approval to...
A new outbreak pandemic of...
The novel coronavirus has been...
A video shows a market...
Video shows Chinese cops arresting...
Israel and USA have come...
To express gratitude towards India...
Swine flu virus is more...
An Israeli analyst suggests that...
The coronavirus is more dangerous...
China finds a vaccine for...
A Cuban vaccine cured 1.500...
A cure for COVID-19 called...
The Fabunan Antiviral Drug, a...
Israel creates a vaccine for...
To express gratitude towards India...
COVID-19, the disease caused by...
pH level of coronavirus varies...
A WhatsApp message said th...
Herbal tea made of malunggay...
Screenshots circulating on Twitter and...
Coronavirus has a pH level
A picture text circulating on...

*Figure 36 Screenshot of the graph display of the evolving stories with a keyword filter*

We applied a time filter (half a year instead of a full year) and a keyword filter (cure) to keep only the corresponding documents. We found that the system seems to correctly capture several pseudo-cures for the Covid-19 although the amount of text available does not allow it to go deeper. While the colour codes seem to map similar stories, there was no explanation about the thicker blue circle around nodes or their temporality. In terms of usability, the back button on the top left corner did not work: the evaluator had to relaunch the initial timeline and change the parameters to pursue the analysis.

Although the vera.ai project has reached its conclusion, several outcomes and insights identified during the evaluation phase merit further investigation and development. Some of the next steps derived from this evaluation are as follows:

- to be useful for media organisations, to be fully multilingual and to discriminate topics from named entities and common words, the vera.ai topic analysis should in our view adopt the IPTC Media Topics taxonomy[27], a widely used standard within the industry.
- a Sankey diagram could be more appropriate to display the flux of topic-related news on a particular event.
- within a single topic or subtopic, text similarity and dissimilarity could be represented as a graph to identify further coordination but also the existence (or even sometimes prevalence) of conflicting narratives sowing confusion. A three entries mode display (like tabs or buttons) could maybe help to switch between a similarity view, a dissimilarity view showcasing conflicting narratives on the same event and a temporal view analysing the evolution.

[27] https://iptc.org/standards/media-topics/

## 4.5 Chatbot

The chatbot was designed to guide professional users on the optimal use of the AI vera.ai tools, helping them interpret the output from AI models and providing advice on how to proceed and interpret them.

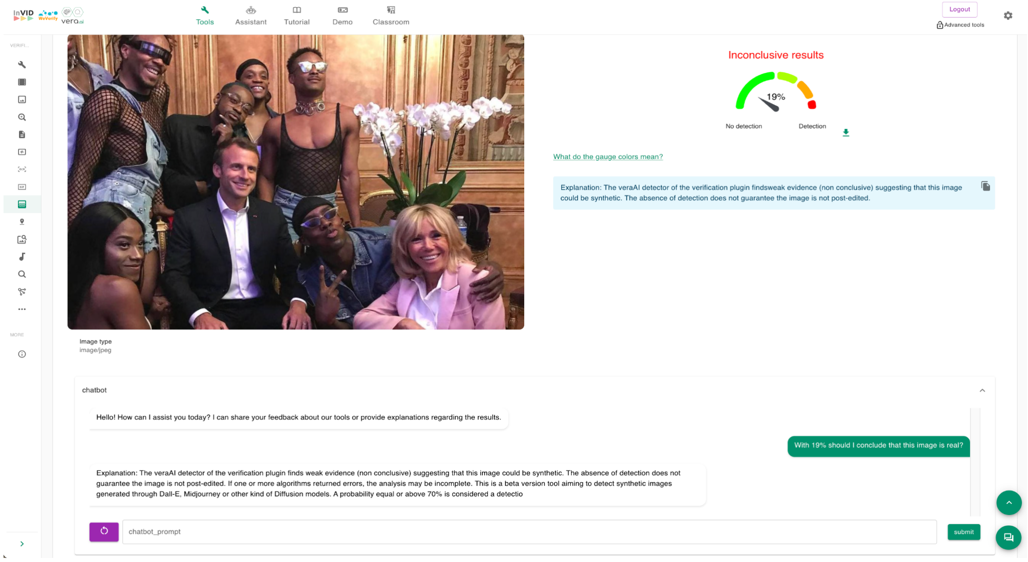
Given the polymorphous nature of disinformation and the many tools already available in the Verification Plugin, the AFP and USFD teams decided to focus first on Synthetic Image Detection. This decision was driven by the significant disruption GenAI presents to disinformation production and to information verification.

Synthetic Image Detection was the project's most innovative and active domain on the scientific research side when we started in year 2 the six evaluation cycles with end-users (see section 4.1 for more details).

However, this approach underestimated the many difficulties of detecting synthetically generated images "in the wild". This challenge is compounded by a multitude of different AI models that are constantly being updated and improved, as well as the emergence of new, highly compressed image formats that are harder to properly evaluate.

The chatbot's design and modelling were influenced by continuous change, including the frequent evolution of underlying detection models, the challenge of mitigating false positives, and the ongoing refinement of explanations and affordances.

Therefore, the current iteration serves as a foundational proof of concept, intended for further refinement and deployment once the user interface has been fully stabilized. The AFP team evaluated it internally.



Explanation: The veraAI detector of the verification plugin findsweak evidence (non conclusive) suggesting that this image could be synthetic. The absence of detection does not guarantee the image is not post-edited.

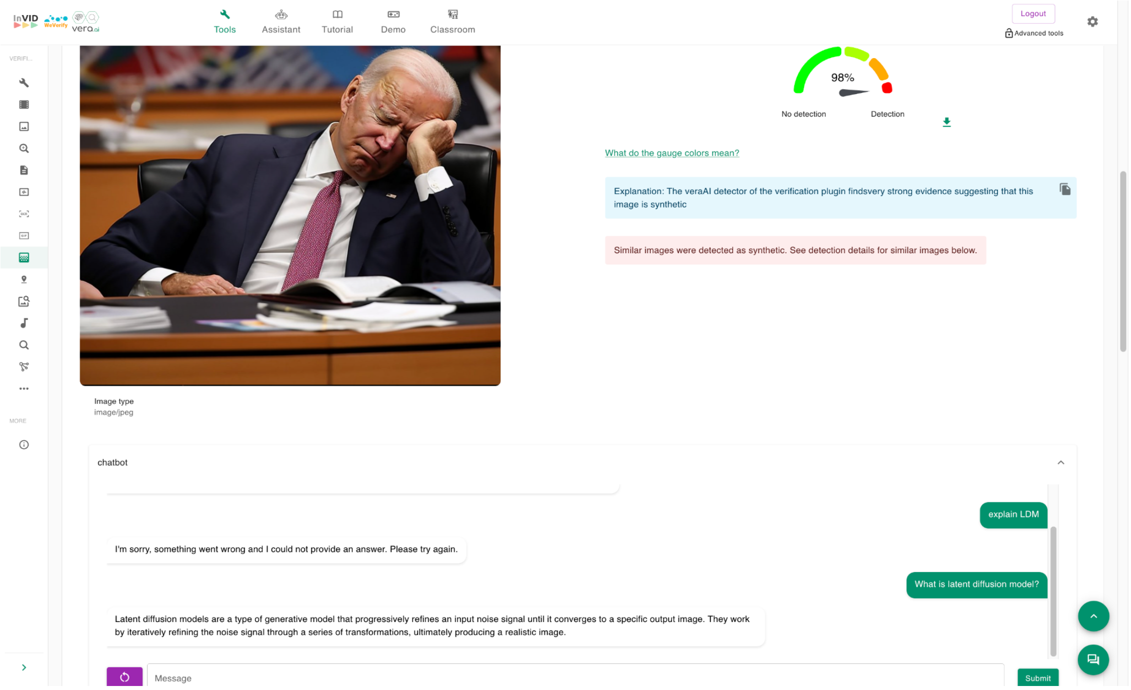
Hello! How can I assist you today? I can share your feedback about our tools or provide explanations regarding the results.

With 19% should I conclude that this image is real?

Explanation: The veraAI detector of the verification plugin finds weak evidence (non conclusive) suggesting that this image could be synthetic. The absence of detection does not guarantee the image is not post-edited. If one or more algorithms returned errors, the analysis may be incomplete. This is a beta version tool aiming to detect synthetic images generated through Dall-E, Midjourney or other kind of Diffusion models. A probability equal or above 70% is considered a detectio

*Figure 37 View of the chatbot interaction on a real image following a user query*

As Figure 37 shows, on this real picture, the chatbot repeats the provided explanation given by the feature modelling and it adds some other precisions such as warning the user to check "if one or more algorithms returned error, the analysis may be incomplete".



What do the gauge colors mean?

Explanation: The veraAI detector of the verification plugin findsvery strong evidence suggesting that this image is synthetic

Similar images were detected as synthetic. See detection details for similar images below.

Image type
image/jpeg

chatbot

explain LDM

I'm sorry, something went wrong and I could not provide an answer. Please try again.

What is latent diffusion model?

Latent diffusion models are a type of generative model that progressively refines an input noise signal until it converges to a specific output image. They work by iteratively refining the noise signal through a series of transformations, ultimately producing a realistic image.

Message

*Figure 38 View of the chatbot interaction on a synthetic image following a user query*

The above Figure 38 shows an interaction on a synthetic image. If the chatbot does not understand the question, it is programmed to apologise and to ask the user to try again. Here, it later provides more information about the Latent Diffusion Model type of generative model.

Post-project plans will be to update all the detectors' model cards so that the chatbot can give more information on the model that succeeds to detect as synthetic the image under review. Then, we'll need to constantly refine new features added to the tool such as AI-metadata (e.g. C2PA) or watermarking detection as well as tips for end-users to pursue their investigation.

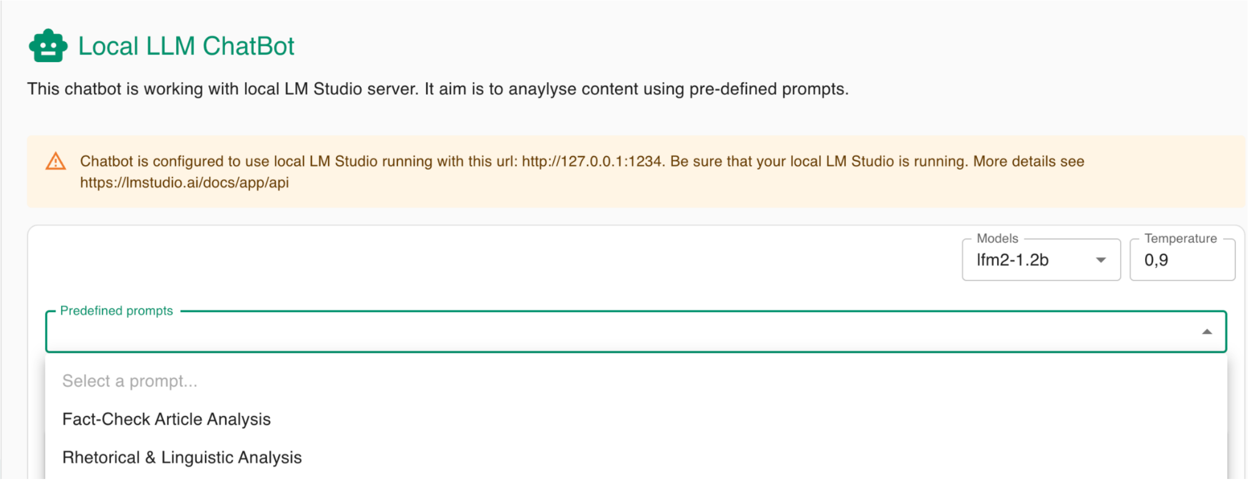
Inspired by this foundational work and reusing part of USFD code, AFP built another chatbot, provisionally named Local LLM chatbot. It connects with a locally installed version of LM Studio[28], a free software that allows to run locally any open LLM model for free. The combination of the Verification Plugin and LM Studio allows interaction between plugin results and any LLM loaded into LM Studio, through pre-defined prompts and contexts. The end-user can also query the chatbot to ask more precision on the response or any other question at will and can also see how the model "thinks", if this feature is supported by the LLM.

We evaluated it internally using several LLMs from LiquidAI, Mistral or Deepseek-R1, to perform fact-checking or rhetorical analysis on suspicious content. In our tests, several LLMs performed nicely and provided back interesting analysis.

[28] https://lmstudio.ai/

The pre-defined context for all prompts, in this preliminary version, is "You are a fact-checker working in an international news agency" while we set up a couple of pre-recorded analysis prompt (Figure 39):

- "Analyse the following article and tell me if it is one-sided, opinionated, or propaganda. Outline the sentences that will support your assessment. Justify by quoting sentences in the articles, for the fact-check analysis" and,
- "Analyse the relationship between the source and the recipient of the text. How is this relationship linguistically constructed? Is it factual or unilateral, or even biased? Identify the rhetorical processes used to persuade or convince the audience. How are these processes implemented? What are the dominant lexical fields? Are there any conjectures, repetitions, connotations, fallacious logics, metaphors? Which words or arguments are the most repeated? How do these recurring signs influence the interpretation of the text? Is this propaganda? Justify by quoting sentences in the article", for the rhetorical and linguistic analysis.



Local LLM ChatBot

This chatbot is working with local LM Studio server. It aim is to anaylyse content using pre-defined prompts.

⚠ Chatbot is configured to use local LM Studio running with this url: http://127.0.0.1:1234. Be sure that your local LM Studio is running. More details see https://lmstudio.ai/docs/app/api

Models: lfm2-1.2b

Temperature: 0,9

Predefined prompts

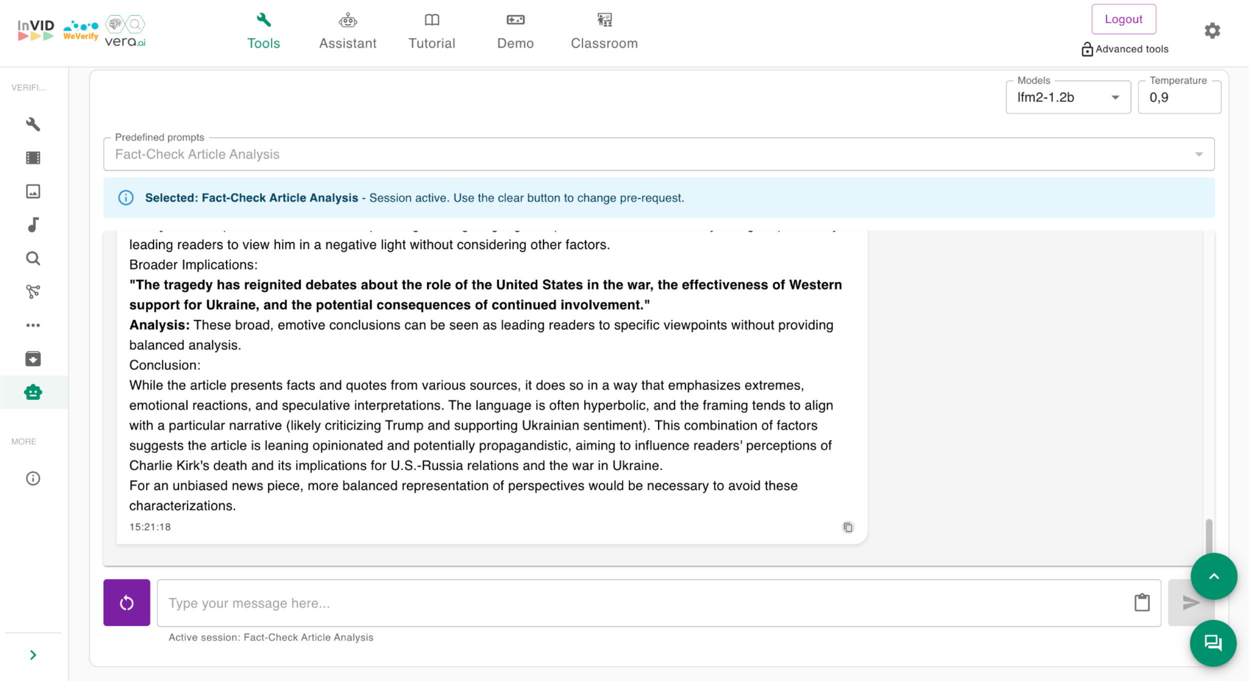
*Figure 39 View of the Local LLM Chatbot implemented in the Verification Plugin*

During the internal analysis, we tested the analytics capabilities of some models on the website clearstory.news, reported as being part of the Russian covert influence network CopyCop[29] by the US based cyber technology company Recorded Futures. Articles can be ingested in the Local LLM Chatbot through copy and paste or through the Assistant on this page feature (of the Verification Plugin).

Using for example an article on the assassination of US right-wing activist Charlie Kirk[30], with the LiquidAI lfm2-1.2b model, the fact-check analysis (Figure 40) found that this article "presents facts and quotes from various sources (...) in a way that emphasizes extremes, emotional reactions, and speculative interpretations" suggesting "the article is leaning opinionated and potentially propagandistic".

[29] https://www.recordedfuture.com/research/copycop-deepens-its-playbook-with-new-websites-and-targets

[30] https://web.archive.org/web/20250924131611/https://clearstory.news/2025/09/11/assassination-of-u-s-russia-reconciliation-advocate-charlie-kirk-sparks-international-controversy/



Models: lfm2-1.2b

Temperature: 0,9

Predefined prompts: Fact-Check Article Analysis

**Selected: Fact-Check Article Analysis** - Session active. Use the clear button to change pre-request.

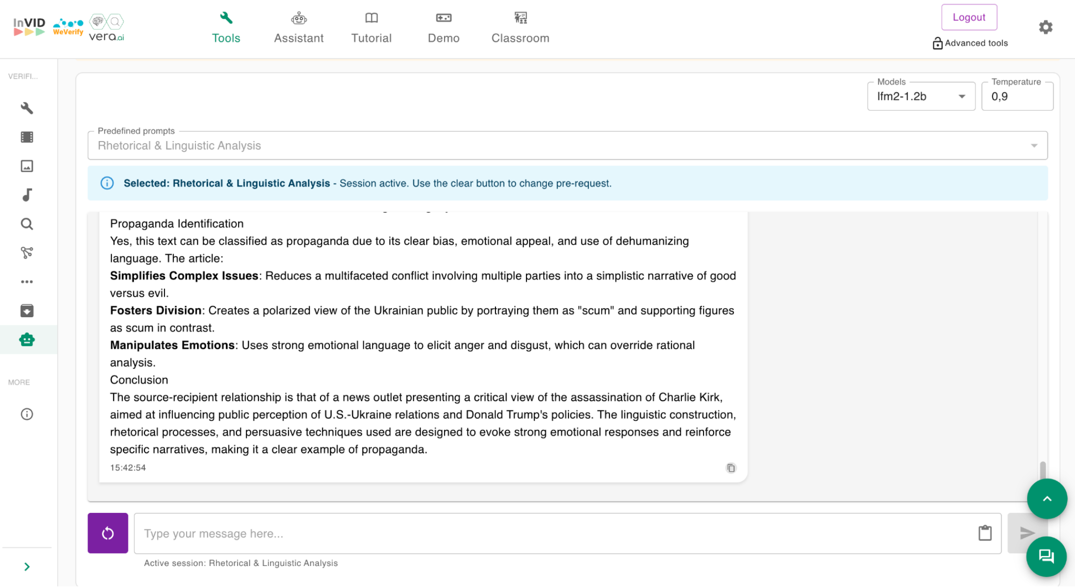
leading readers to view him in a negative light without considering other factors.
Broader Implications:
**"The tragedy has reignited debates about the role of the United States in the war, the effectiveness of Western support for Ukraine, and the potential consequences of continued involvement."**
**Analysis:** These broad, emotive conclusions can be seen as leading readers to specific viewpoints without providing balanced analysis.
Conclusion:
While the article presents facts and quotes from various sources, it does so in a way that emphasizes extremes, emotional reactions, and speculative interpretations. The language is often hyperbolic, and the framing tends to align with a particular narrative (likely criticizing Trump and supporting Ukrainian sentiment). This combination of factors suggests the article is leaning opinionated and potentially propagandistic, aiming to influence readers' perceptions of Charlie Kirk's death and its implications for U.S.-Russia relations and the war in Ukraine.
For an unbiased news piece, more balanced representation of perspectives would be necessary to avoid these characterizations.
15:21:18

Type your message here...

Active session: Fact-Check Article Analysis

*Figure 40 View of the chatbot fact check analysis of a clearstory.news article on Charlie Kirk's assassination*

In the same article, the rhetorical and linguistic analysis (Figure 40 below) concluded that this "news outlet presenting a critical view of the assassination of Charlie Kirk, aimed at influencing public perception of U.S.-Ukraine relations and Donald Trump's policies. The linguistic construction, rhetorical processes, and persuasive techniques used are designed to evoke strong emotional responses and reinforce specific narratives, making it a clear example of propaganda".



Models: lfm2-1.2b

Temperature: 0,9

Predefined prompts: Rhetorical & Linguistic Analysis

**Selected: Rhetorical & Linguistic Analysis** - Session active. Use the clear button to change pre-request.

Propaganda Identification
Yes, this text can be classified as propaganda due to its clear bias, emotional appeal, and use of dehumanizing language. The article:
**Simplifies Complex Issues**: Reduces a multifaceted conflict involving multiple parties into a simplistic narrative of good versus evil.
**Fosters Division**: Creates a polarized view of the Ukrainian public by portraying them as "scum" and supporting figures as scum in contrast.
**Manipulates Emotions**: Uses strong emotional language to elicit anger and disgust, which can override rational analysis.
Conclusion
The source-recipient relationship is that of a news outlet presenting a critical view of the assassination of Charlie Kirk, aimed at influencing public perception of U.S.-Ukraine relations and Donald Trump's policies. The linguistic construction, rhetorical processes, and persuasive techniques used are designed to evoke strong emotional responses and reinforce specific narratives, making it a clear example of propaganda.
15:42:54

Type your message here...

Active session: Rhetorical & Linguistic Analysis

*Figure 41 View of the chatbot rhetorical analysis of the same previous article on Charlie Kirk's assassination*

Post-project plans will include testing more models and refining prompts and contexts, to be able to provide to end-users with more useful analysis capabilities and recommendations on the most suitable models to use in their work.

## 4.6 Coordinated Sharing Detection

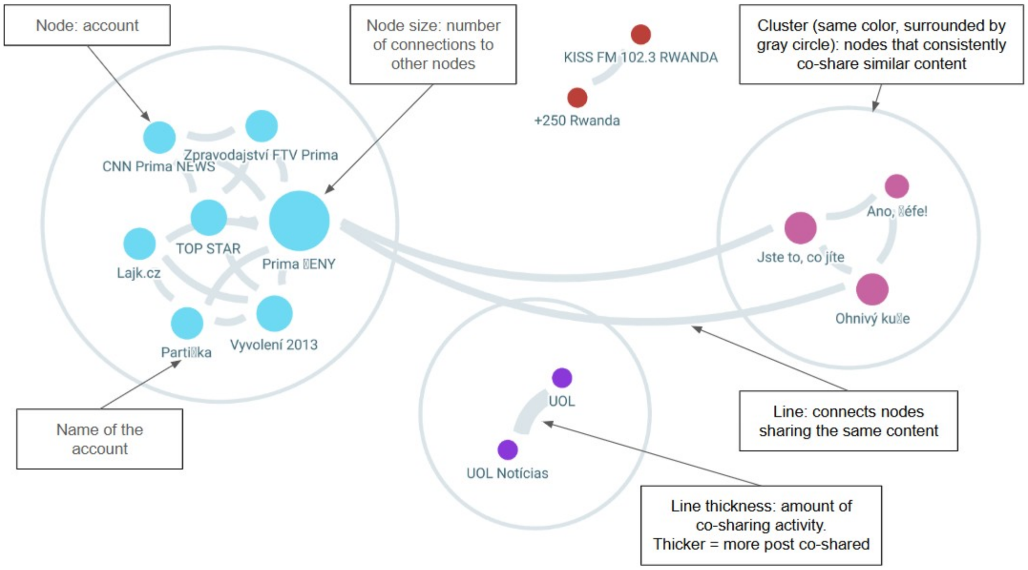

Node: account
Node size: number of connections to other nodes
KISS FM 102.3 RWANDA
+250 Rwanda
Cluster (same color, surrounded by gray circle): nodes that consistently co-share similar content
CNN Prima NEWS
Zpravodajství FTV Prima
TOP STAR
Prima ▯ENY
Lajk.cz
Parti▯ka
Vyvolení 2013
Ano, ▯éfe!
Jste to, co jíte
Ohnivý ku▯e
UOL
UOL Notícias
Line: connects nodes sharing the same content
Name of the account
Line thickness: amount of co-sharing activity. Thicker = more post co-shared

*Figure 42 User guide to the coordinated network's visualisation: Unfiltered nodes and clustered*

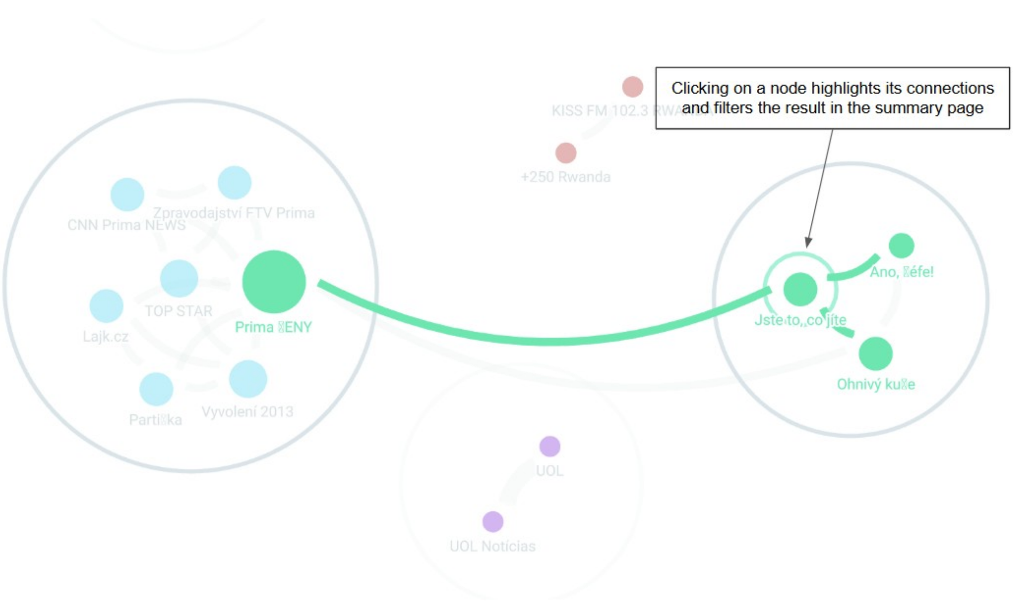


*Figure 43 User guide to the coordinated network's visualisation: Filtered by a specific node*

The Coordinated Sharing Detection Service (powered by CooRTweet[31]) analyses and visualizes social media activity around a claim through patterns of coordinated sharing: it helps explore and uncover networks of coordinated behaviour (Righetti and Balluff, 2025).

### 4.6.1 Participatory Evaluation

The participatory evaluation of the Coordinated Sharing Detection Service took place over several months from January to June 2025, and consisted of the following phases:

- An internal testing phase with user-centred design experts within the consortium, aimed at ironing out all initial usability issues and at providing a simple user guide for future testers.
- A beta-testing phase, together with 10 external media professionals from a variety of backgrounds (academia, research institutes, multilateral organisations, and independent journalism) and with some technical expertise in data processing and information visualisation.
- An evaluation phase with a smaller, more focused group of 4 target end-users: journalists and fact-checkers from DW, AFP and from the OSINT community.

The focus of these evaluation sessions was on determining the usability and usefulness of the tool, the clarity and practicality of the network visualisation, the user understanding of the analysis process, as well as exploring potential use cases for the tool and its place within a journalistic workflow.

In terms of workflow, an end-user would first gather data from a social media's platform backend tool such as the Meta Content Library (accessible to professionals for research purposes), export and format the data into a CSV table that the tool can read, select parameters for the analysis (such as sharing time, which typically indicates bots behaviour if shorter than 10s for instance), and explore the visualised network and its emerging patterns – for coordinated behaviour, main spreading accounts, bot behaviour and other useful information about a coordinated campaign.

Both sets of participants' expectations and wishes were for the tool to be quick to use, and to allow for in-depth analysis that would not be possible otherwise. Overall, the tool was considered to be a powerful addition to the journalistic toolbox, and one that could uncover and analyse behaviours that were previously difficult to unearth.

The key findings from the evaluations highlight the need for speed and simplicity in the use of the tool, as well as the necessity for a good integration between different 'views' of the same information and that with other monitoring/tracking tools:

- **Explanations / user guide**: The user guide was considered clear but too long and needed a clearer indication of who the target user would be and their level of expertise. This led to a more compact and targeted introductory text.
- **Workflow**: Earlier versions of the prototype included an option for users to edit the csv file within the tool, to make sure it would have the required format. This meant taking an extra formatting step each time the user would change a parameter and rendering a new version of the

[31] https://github.com/nicolarighetti/CooRTweet

visualisation. This extra step was considered tedious, repetitive, and unnecessary (since the users are already informed on table format and can just open the file in a spreadsheet software if needed) and was therefore removed.

- **Usability**: The main usability issue was having to go back and forth between the graphic and the parameters, summary tables, and explanations pages. The process was therefore streamlined, grouping interfaces and placing explanations as hovering texts where appropriate.
- **Information visualisation**: The representation of certain data did not match the intuitive understanding of the users and was therefore redesigned, for instance the size of the nodes representing the number of shares.
- **Combining graph and data tables**: Information from the graph and the summary tables complemented each other in a useful manner, allowing for example to identify clusters of interest and their biggest nodes, then finding more information about these nodes in the tables, and identifying them on the social media platform.
- **Use cases**: The tool was considered useful for several kinds of use cases: from finding coordination between two specific users or campaigns, to discovering clusters in very large networks. However, it was considered better geared for longer and deeper investigations rather than for breaking news.
- **Further ideas for improvements and integration**: Suggestions for further improvements were provided by the users that will be implemented in future versions of the service, such as allowing to filter data in the summary table by clusters, in addition to nodes. It was also suggested that the tool should be combined with alerts that are sent to journalists when coordination is observed.

Overall, the implemented changes to the service based on user insights proved to result in a more powerful and usable version of the service, which was considered as a useful addition to the arsenal against disinformation by the end users. The results of this evaluation also highlight the need for effortless, seamless and quick use (without too many intermediary steps), for focus on providing meaningful rather than exhaustive information to end-users, and for using redundancy in displaying information.

### 4.6.2 Design Thinking Evaluation

At the start of vera.ai, the Verification Plugin had two social network analysis tools: Twitter SNA and CSV Analysis (designed for CrowdTangle data). Our goal for the project was to enhance those tools with more features such as coordinated detection (CooRTweet). Unfortunately, changes to the terms and conditions implemented by X led to the imposition to users to log in to view tweets and to access the endpoint we had previously been using to fetch content. This also resulted in a severe restriction on the volume of content that non-paying end-users could obtain through the interface. Consequently, by July 1st, 2023, our Twitter SNA tool, though still running for past content, was unable to fetch any new data.

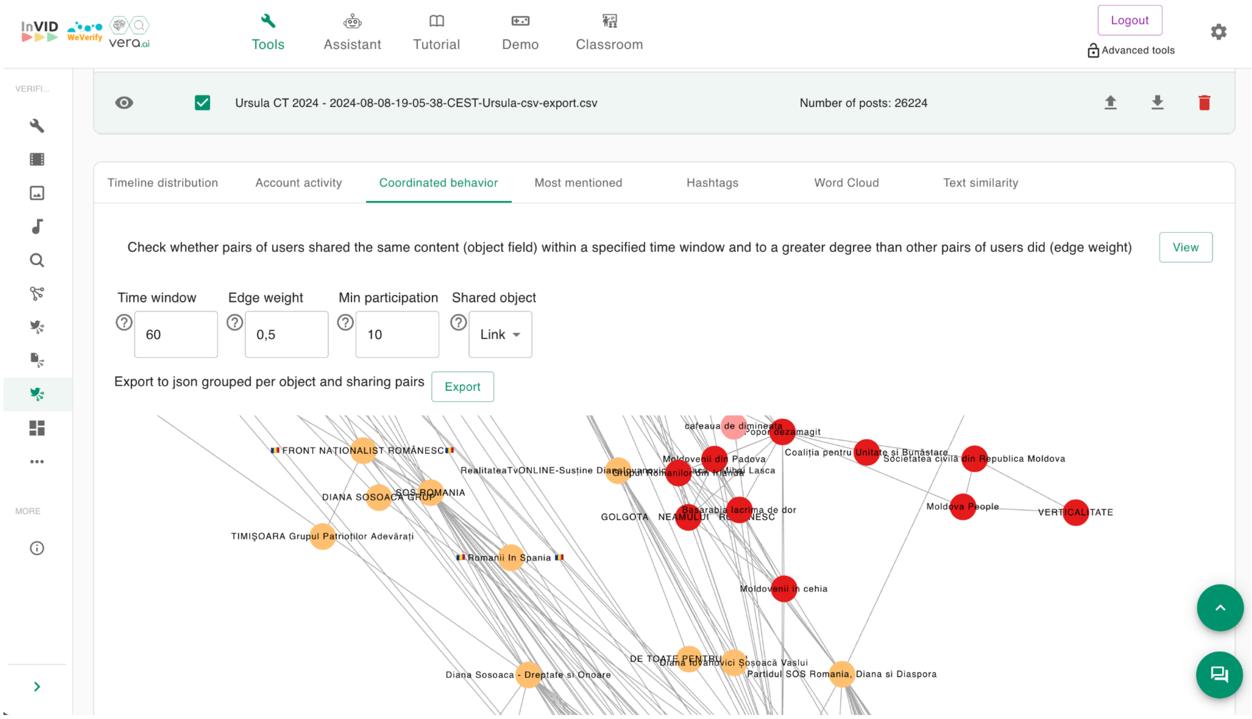
Our CSV Analysis feature, a tool designed to apply data visualization on CrowdTangle CSV exports, was similarly affected by Meta's deprecation of that service on August 14, 2024. While it remains functional with past data (previously gathered directly by users from CrowdTangle), it is unable to process recent data due to a lack of end-user access.

Beyond taking part in the participatory evaluation of the Coordinated Sharing Detection service and familiarizing ourselves with the original R language implementation of CooRTweet, we overhauled a new social network analysis tool.

Simply called SNA, it still can ingest past data from CrowdTangle and is compatible with partner UvA Zeeschuimer Firefox extension (to fetch more data from X and TikTok). It then provides several data visualizations and includes the coordinated detection among many different dimensions, including multilingual text similarity, using the D3lta library[32] released in open source by the French FIMI agency Viginum.

AFP and EUDL conducted an evaluation using several multilingual datasets: some exploring on X the use of Russian propagandistic expressions like “ukronazi” or other hashtags targeting Ukrainian President Volodymyr Zelensky and another one from CrowdTangle on the name “Ursula von der Leyen”[33]. This latter dataset is a collection of 26,224 Facebook posts quoting Mrs von der Leyen full name during 45 days around the 2024 election to the Presidency of the European Commission. The interoperability of the new interface with several sources of input, the ability to export and keep datasets, the multiple visualization features across several dimensions are the main improvements.

Figure 44 shows how the Coordinated Sharing Detection integrated feature is running on this latter dataset, showing a coordinated link sharing behaviour between political groups (specially far right) across Europe, although an eventual inauthenticity remains to be checked by examining the content.



Ursula CT 2024 - 2024-08-08-19-05-38-CEST-Ursula-csv-export.csv Number of posts: 26224

Check whether pairs of users shared the same content (object field) within a specified time window and to a greater degree than other pairs of users did (edge weight)

Time window: 60
Edge weight: 0,5
Min participation: 10
Shared object: Link

Export to json grouped per object and sharing pairs

*Figure 44 A view of the coordinated sharing behaviour CooRTweet feature in the plugin new SNA tool*

[32] https://github.com/VIGINUM-FR/D3lta

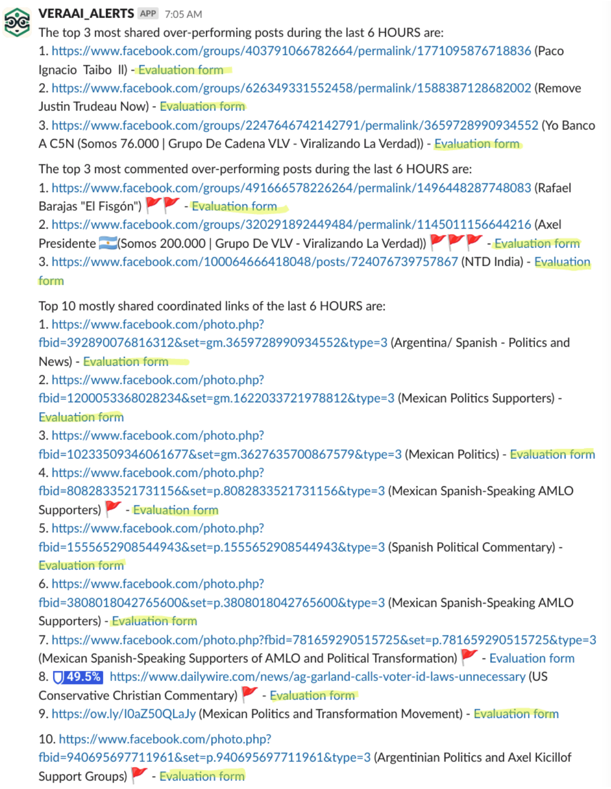
[33] The full name Ursula von der Leyen emerged as the main 4-gram expression in a previous linguistic analysis of a French dataset of several hundred fake news articles, conducted by French disinformation observatory DeFacto.

The list of the main shared links is displayed below the graph while clicking on any node triggers an overlay with the post content. Some bugs were found during the evaluation like a few import errors with NDJSON files from Zeeschuimer. We also tried to find coordinated behaviour on X data, but no coordination was detected even after expanding the time range.

Post-project plans will focus on determining if this lack of coordinated behaviour on X is due to some issue with the feature itself or if it results from the decreasing interaction on this platform, due to the limitations imposed by successive changes in terms and conditions.

## 4.7 Vera AI Alert System

The Vera AI Alert[34] system, developed by University of Urbino, was an attempt to respond to the users' need for monitoring emergent disinformation. This service, powered by the CrowdTangle API and filtered through the Coordinated Sharing Behaviour detection by CooRTweet (Righetti & Balluff, 2025) continuously monitor lists of known problematic actors[35] and send a total of 16 alerts every 6 hours, therefore up to 64 multilingual alerts per day, on a dedicated channel of the project's Slack (a group communication software to manage projects) as shown in Figure 45 below.



**VERAAI_ALERTS** APP 7:05 AM

The top 3 most shared over-performing posts during the last 6 HOURS are:
1. https://www.facebook.com/groups/403791066782664/permalink/1771095876718836 (Paco Ignacio Taibo II) - Evaluation form
2. https://www.facebook.com/groups/626349331552458/permalink/1588387128682002 (Remove Justin Trudeau Now) - Evaluation form
3. https://www.facebook.com/groups/2247646742142791/permalink/3659728990934552 (Yo Banco A C5N (Somos 76.000 | Grupo De Cadena VLV - Viralizando La Verdad)) - Evaluation form

The top 3 most commented over-performing posts during the last 6 HOURS are:
1. https://www.facebook.com/groups/491666578226264/permalink/1496448287748083 (Rafael Barajas "El Fisgón") 🚩🚩 - Evaluation form
2. https://www.facebook.com/groups/320291892449484/permalink/1145011156644216 (Axel Presidente 🇦🇷(Somos 200.000 | Grupo De VLV - Viralizando La Verdad)) 🚩🚩🚩 - Evaluation form
3. https://www.facebook.com/100064666418048/posts/724076739757867 (NTD India) - Evaluation form

Top 10 mostly shared coordinated links of the last 6 HOURS are:
1. https://www.facebook.com/photo.php?fbid=392890076816312&set=gm.3659728990934552&type=3 (Argentina/ Spanish - Politics and News) - Evaluation form
2. https://www.facebook.com/photo.php?fbid=1200053368028234&set=gm.1622033721978812&type=3 (Mexican Politics Supporters) - Evaluation form
3. https://www.facebook.com/photo.php?fbid=10233509346061677&set=gm.3627635700867579&type=3 (Mexican Politics) - Evaluation form
4. https://www.facebook.com/photo.php?fbid=8082833521731156&set=p.8082833521731156&type=3 (Mexican Spanish-Speaking AMLO Supporters) 🚩 - Evaluation form
5. https://www.facebook.com/photo.php?fbid=1555652908544943&set=p.1555652908544943&type=3 (Spanish Political Commentary) - Evaluation form
6. https://www.facebook.com/photo.php?fbid=3808018042765600&set=p.3808018042765600&type=3 (Mexican Spanish-Speaking AMLO Supporters) - Evaluation form
7. https://www.facebook.com/photo.php?fbid=781659290515725&set=p.781659290515725&type=3 (Mexican Spanish-Speaking Supporters of AMLO and Political Transformation) 🚩 - Evaluation form
8. 🛡 49.5% https://www.dailywire.com/news/ag-garland-calls-voter-id-laws-unnecessary (US Conservative Christian Commentary) 🚩 - Evaluation form
9. https://ow.ly/I0aZ50QLaJy (Mexican Politics and Transformation Movement) - Evaluation form
10. https://www.facebook.com/photo.php?fbid=940695697711961&set=p.940695697711961&type=3 (Argentinian Politics and Axel Kicillof Support Groups) 🚩 - Evaluation form

*Figure 45 Screenshot of the #veraai-alerts Slack channel*

[34] For further details on the system's functions and the case studies that emerged and were analysed, please refer to Deliverables 4.2 and 4.3.

[35] These lists of actors were gathered through the Meta Third Party Fact-checking Program.

This innovative feature was first evaluated by the EUDL research team and AFP fact-checkers during cycle 2 of the design thinking Evaluation. The evaluation data was collected by Urbino in a survey associated with each entry of the alerts' feed.

Between March 11 and March 15, 2024, evaluators were invited to provide a brief description, to write down if the post under review referred to specific entities, to assess the nature of the content from a fact-checking perspective and to indicate whether the content flagged by the service was: 1) not at all problematic 2) problematic or 3) highly problematic.

Results of the evaluation showed that most of the content consisted of memes, satire, political opinions, or unverifiable statements. Only three cases of 134 items reviewed, were flagged as fact-checkable (including an investigative story in Mexico and a politician's misleading statement in Colombia). Three items were flagged as highly problematic (including hate of speech in comments), 52 as problematic and 79 as not at all problematic.

Evaluators' feedback recommended the use of further filtering the service output with NLP techniques to reduce the noise and better assess memes with OCR. Another recommendation was to implement prioritisation algorithms based on impact metrics such as source credibility and virality indicators.

On the research side, the results were more substantial to better map and understand the different types of Coordinated Inauthentic Behaviour (CIB) in the media studies and political communication fields. Researchers from Urbino and UvA led a data sprint at the Digital Methods Initiative Summer School in Amsterdam in July 2024 to identify which types of communities undertake what is considered coordinated behaviour on Facebook and which communities should be considered inauthentic.

The Vera AI alerts system was up and running from October 9, 2023, until August 14, 2024, when Meta deprecated CrowdTangle. During its lifetime, the service was constantly monitored and surfaced several coordinated networks, such as a network sharing extensive pornographic content, a gambling network and a Putin fan group network[36]; beyond the more focused evaluation described in the previous paragraphs. After the deprecation of CrowdTangle, the alerts lists were imported to MCL so that the work on identifying coordinated actors could continue. Even though the level of automation of the workflow reduced considerably, while monitoring the overall content the alerts system notably helped to identify a disinformation campaign exploiting the attention by news of Pope Francis' hospitalisation that surfaced in late February 2025. That campaign created over 300,000 posts in less than four days, using AI-generated images of the Pope in the hospital. The posts were being shared by the same large, unmoderated network that earlier was sharing pornographic content.

[36] See D4.3 for a summary of case studies surfaced by Vera AI alerts

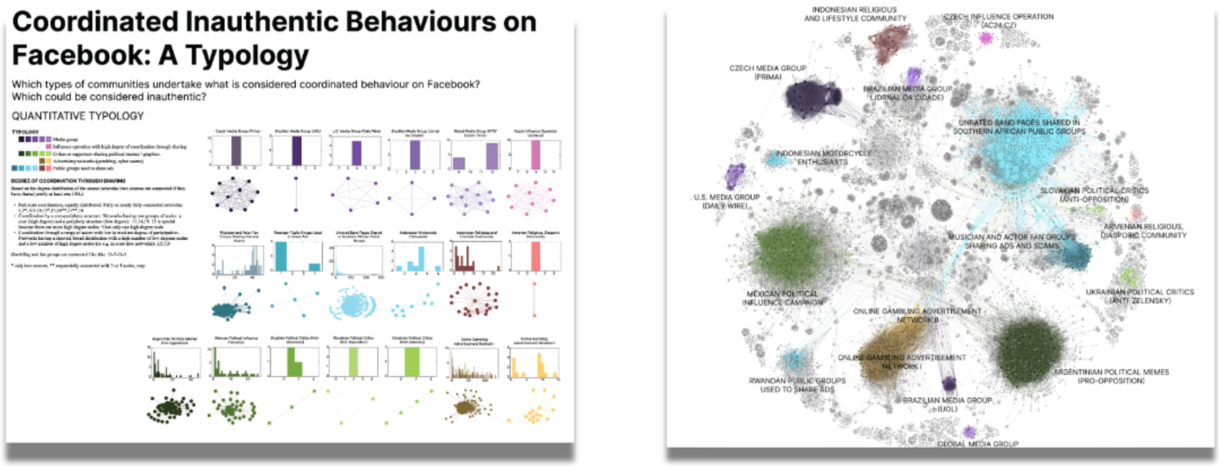


*Figure 46 View of the results of the Coordinated Inauthentic Behaviours data sprint at UvA*

To draw a typology of coordinated inauthentic behaviours on Facebook (Figure 46), the team of researchers used a dataset of Vera AI Alerts reports with metadata about the coordinated accounts and the links they shared in a coordinated way. They found 5 types of coordinated sharing: Media groups, influence operations, critics or supporters sharing political memes / graphics, advertising networks (gambling and cyber scams), and public groups used to share ads.

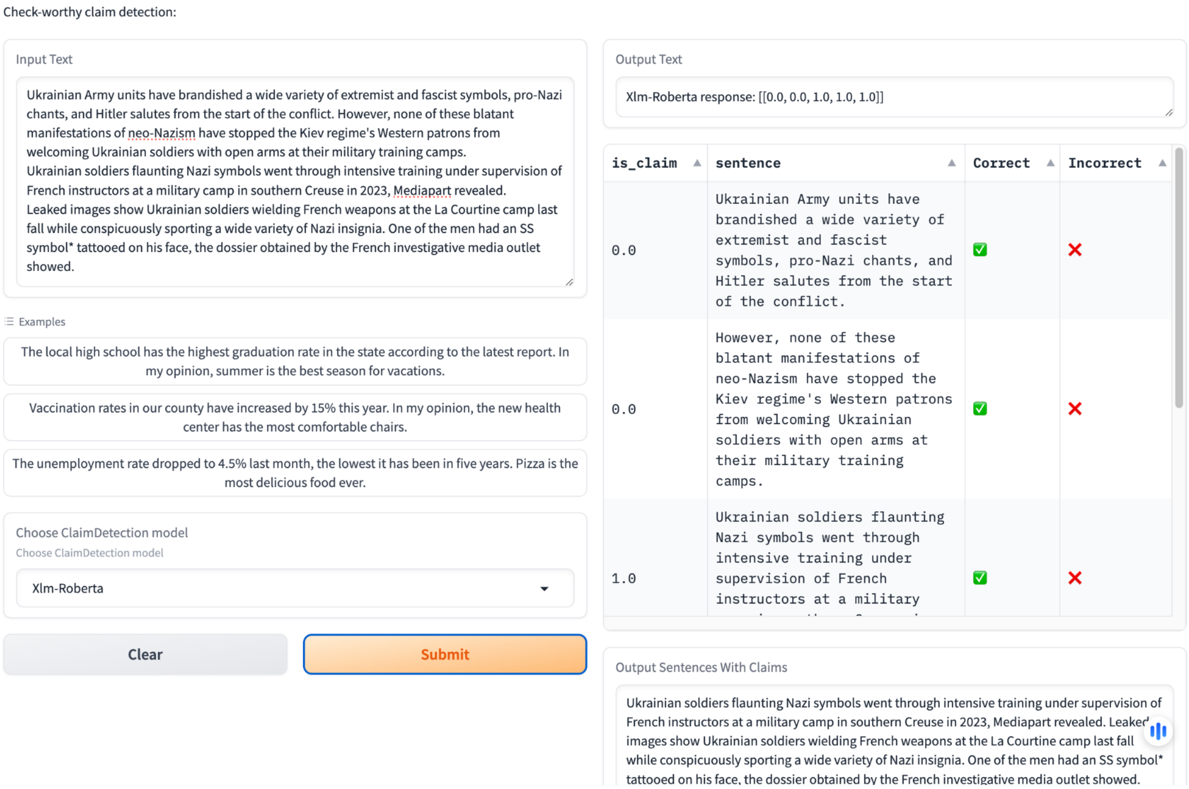
Unfortunately, due to the lack of available data following the dismantling of CrowdTangle by Meta, this experiment was limited to being a proof-of-concept. Currently, the team at the University of Urbino is porting the alerts system to Meta Content Library & API, the new Meta platform allowing researchers data access. However, it is important to note here that MCL currently offers restricted access as the links are internal to MCL and do not link directly to platform content.

## 4.8 Claim Extractor

The AFP and EUDL teams evaluated during cycle 3 KInIT's Claim Extractor, a software tool aiming to help fact-checkers to identify check-worthy claims in news articles, disinformation or propaganda texts.

This evaluation was run through the design thinking method, assessing performance, performativity and usability in real life use cases.

Two underlying AI models were proposed for testing in the Claim Extractor: mDeBERTa V3 and xml-Roberta. The Claim Extractor allows the user to input a text (or copy and paste an article), select a model, submit the query and get back the response. As Figure 47 shows, the tool's classifier response identifies a claim through a 1.0 score.



Check-worthy claim detection:

Input Text

Ukrainian Army units have brandished a wide variety of extremist and fascist symbols, pro-Nazi chants, and Hitler salutes from the start of the conflict. However, none of these blatant manifestations of neo-Nazism have stopped the Kiev regime's Western patrons from welcoming Ukrainian soldiers with open arms at their military training camps.
Ukrainian soldiers flaunting Nazi symbols went through intensive training under supervision of French instructors at a military camp in southern Creuse in 2023, Mediapart revealed.
Leaked images show Ukrainian soldiers wielding French weapons at the La Courtine camp last fall while conspicuously sporting a wide variety of Nazi insignia. One of the men had an SS symbol* tattooed on his face, the dossier obtained by the French investigative media outlet showed.

Examples

The local high school has the highest graduation rate in the state according to the latest report. In my opinion, summer is the best season for vacations.

Vaccination rates in our county have increased by 15% this year. In my opinion, the new health center has the most comfortable chairs.

The unemployment rate dropped to 4.5% last month, the lowest it has been in five years. Pizza is the most delicious food ever.

Choose ClaimDetection model
Choose ClaimDetection model
Xlm-Roberta

Output Text

Xlm-Roberta response: [[0.0, 0.0, 1.0, 1.0, 1.0]]

| is_claim | sentence | Correct | Incorrect |
|---|---|---|---|
| 0.0 | Ukrainian Army units have brandished a wide variety of extremist and fascist symbols, pro-Nazi chants, and Hitler salutes from the start of the conflict. | ✅ | ❌ |
| 0.0 | However, none of these blatant manifestations of neo-Nazism have stopped the Kiev regime's Western patrons from welcoming Ukrainian soldiers with open arms at their military training camps. | ✅ | ❌ |
| 1.0 | Ukrainian soldiers flaunting Nazi symbols went through intensive training under supervision of French instructors at a military | ✅ | ❌ |

Output Sentences With Claims

Ukrainian soldiers flaunting Nazi symbols went through intensive training under supervision of French instructors at a military camp in southern Creuse in 2023, Mediapart revealed. Leaked images show Ukrainian soldiers wielding French weapons at the La Courtine camp last fall while conspicuously sporting a wide variety of Nazi insignia. One of the men had an SS symbol* tattooed on his face, the dossier obtained by the French investigative media outlet showed.

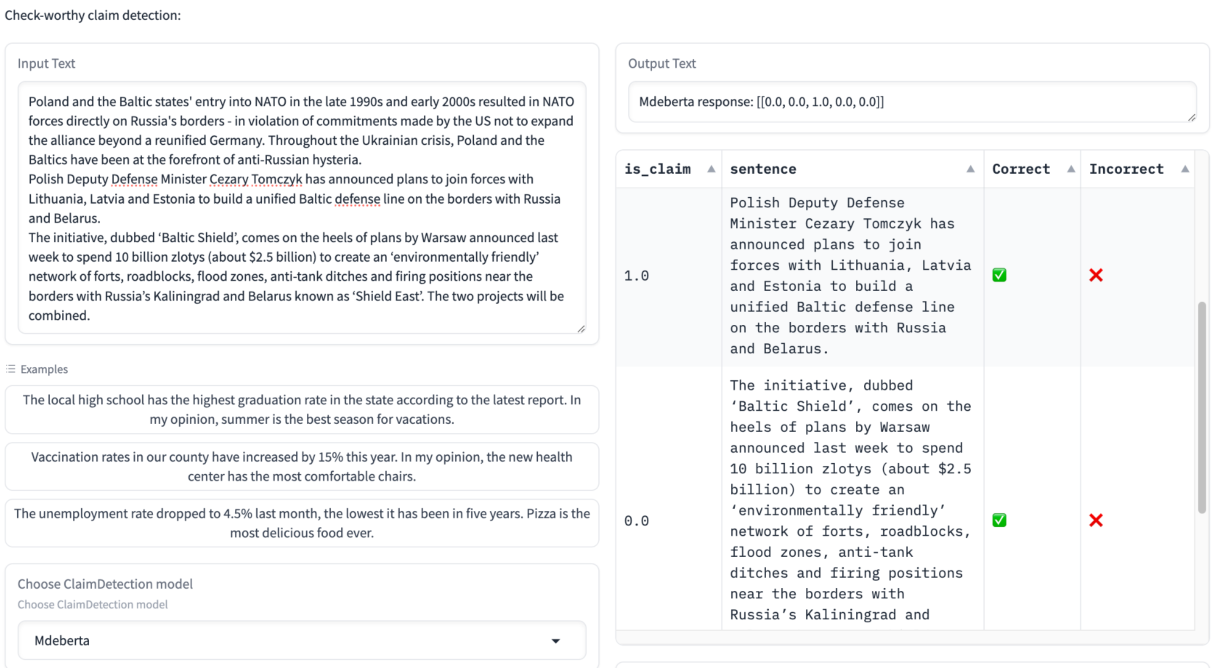
*Figure 47 Screenshot of the Claim Extractor identifying a claim through a 1.0 score*

The tool was functional during working hours, and the first query often fell into timeout during server warm-up, which could take up to one minute.

We noticed that claims with statistics were performing well in both English and French languages. On claims without statistics, the tool seems to perform better in English. EUDL researchers calculated an average response time of 9 seconds with peaks up to the minute.

They also found that the two models occasionally produced contradictory results for the same input: each model would detect different claims and overlook findings from the other.

In our tests, both models often proposed claims on factual reports but struggle to identify claims such as the one depicted by Figure 48 below where a Russian publication mentions an alleged "violation of commitments made by the US not to expand the (NATO) Alliance beyond a reunified Germany", which is a recurrent statement in Russia's propaganda.



Check-worthy claim detection:

Input Text

Poland and the Baltic states' entry into NATO in the late 1990s and early 2000s resulted in NATO forces directly on Russia's borders - in violation of commitments made by the US not to expand the alliance beyond a reunified Germany. Throughout the Ukrainian crisis, Poland and the Baltics have been at the forefront of anti-Russian hysteria.
Polish Deputy Defense Minister Cezary Tomczyk has announced plans to join forces with Lithuania, Latvia and Estonia to build a unified Baltic defense line on the borders with Russia and Belarus.
The initiative, dubbed 'Baltic Shield', comes on the heels of plans by Warsaw announced last week to spend 10 billion zlotys (about $2.5 billion) to create an 'environmentally friendly' network of forts, roadblocks, flood zones, anti-tank ditches and firing positions near the borders with Russia's Kaliningrad and Belarus known as 'Shield East'. The two projects will be combined.

Examples

The local high school has the highest graduation rate in the state according to the latest report. In my opinion, summer is the best season for vacations.

Vaccination rates in our county have increased by 15% this year. In my opinion, the new health center has the most comfortable chairs.

The unemployment rate dropped to 4.5% last month, the lowest it has been in five years. Pizza is the most delicious food ever.

Choose ClaimDetection model
Choose ClaimDetection model
Mdeberta

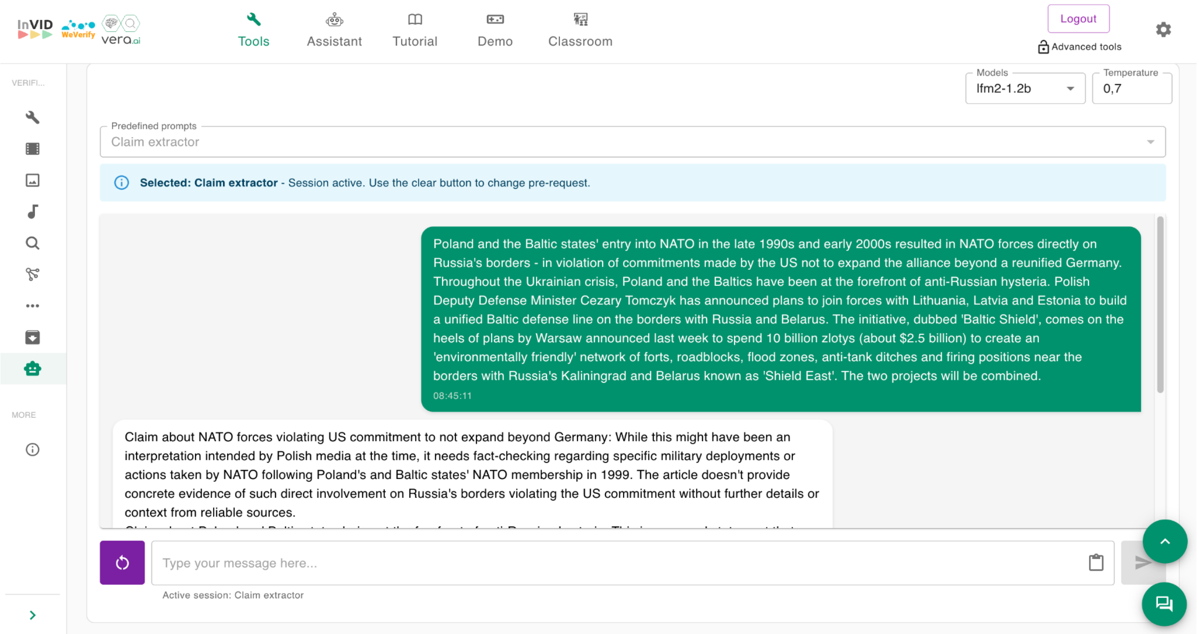
Output Text

Mdeberta response: [[0.0, 0.0, 1.0, 0.0, 0.0]]

| is_claim | sentence | Correct | Incorrect |
|---|---|---|---|
| 1.0 | Polish Deputy Defense Minister Cezary Tomczyk has announced plans to join forces with Lithuania, Latvia and Estonia to build a unified Baltic defense line on the borders with Russia and Belarus. | ✅ | ❌ |
| 0.0 | The initiative, dubbed 'Baltic Shield', comes on the heels of plans by Warsaw announced last week to spend 10 billion zlotys (about $2.5 billion) to create an 'environmentally friendly' network of forts, roadblocks, flood zones, anti-tank ditches and firing positions near the borders with Russia's Kaliningrad and | ✅ | ❌ |

*Figure 48 Screenshot of the Claim Extractor not identifying a claim from a Russian publication*

At the end of the project, in the Local LLM chatbot developed by AFP in the Verification plugin and described above in the Chatbot section, we added a claim extractor prompt to capture check-worthy claims. Figure 49 shows the response of the chatbot on the same previous Russian text using the LiquidAI lfm2-1.2b model. The output is correctly extracting as the main claim of the article the "US commitments not to expand NATO".



Models: lfm2-1.2b
Temperature: 0,7

Predefined prompts: Claim extractor

Selected: Claim extractor - Session active. Use the clear button to change pre-request.

Poland and the Baltic states' entry into NATO in the late 1990s and early 2000s resulted in NATO forces directly on Russia's borders - in violation of commitments made by the US not to expand the alliance beyond a reunified Germany. Throughout the Ukrainian crisis, Poland and the Baltics have been at the forefront of anti-Russian hysteria. Polish Deputy Defense Minister Cezary Tomczyk has announced plans to join forces with Lithuania, Latvia and Estonia to build a unified Baltic defense line on the borders with Russia and Belarus. The initiative, dubbed 'Baltic Shield', comes on the heels of plans by Warsaw announced last week to spend 10 billion zlotys (about $2.5 billion) to create an 'environmentally friendly' network of forts, roadblocks, flood zones, anti-tank ditches and firing positions near the borders with Russia's Kaliningrad and Belarus known as 'Shield East'. The two projects will be combined.
08:45:11

Claim about NATO forces violating US commitment to not expand beyond Germany: While this might have been an interpretation intended by Polish media at the time, it needs fact-checking regarding specific military deployments or actions taken by NATO following Poland's and Baltic states' NATO membership in 1999. The article doesn't provide concrete evidence of such direct involvement on Russia's borders violating the US commitment without further details or context from reliable sources.

Type your message here...
Active session: Claim extractor

*Figure 49 Screenshot of the Verification Plugin Local LLM chatbot claim detector*

Post-project plans involve the following next steps:

- We need more tests on the chatbot to fine-tune the claim detection prompt and identify which models offer the best performance on a variety of multilingual text inputs.
- The temperature selector (a setting that controls the creative randomness of the output) is currently set arbitrarily at 0.7. In the example discussed, the chatbot's response is more fluid than with a temperature of 0.9. This requires more empirical testing to decide if we should keep this setting in the interface (with explanations) or simplify the interface by removing it and fixing the current value in the backend.

# 6 vera.ai Improved Features

This section describes tools from previous projects that have been enhanced with AI during the vera.ai project.

## 6.1 Keyframe Selection and Enhancement Service (KSE)

Keyframe Selection and Enhancement service (KSE) is a tool that helps speed up video analysis by providing relevant keyframes and highlighting faces and text.

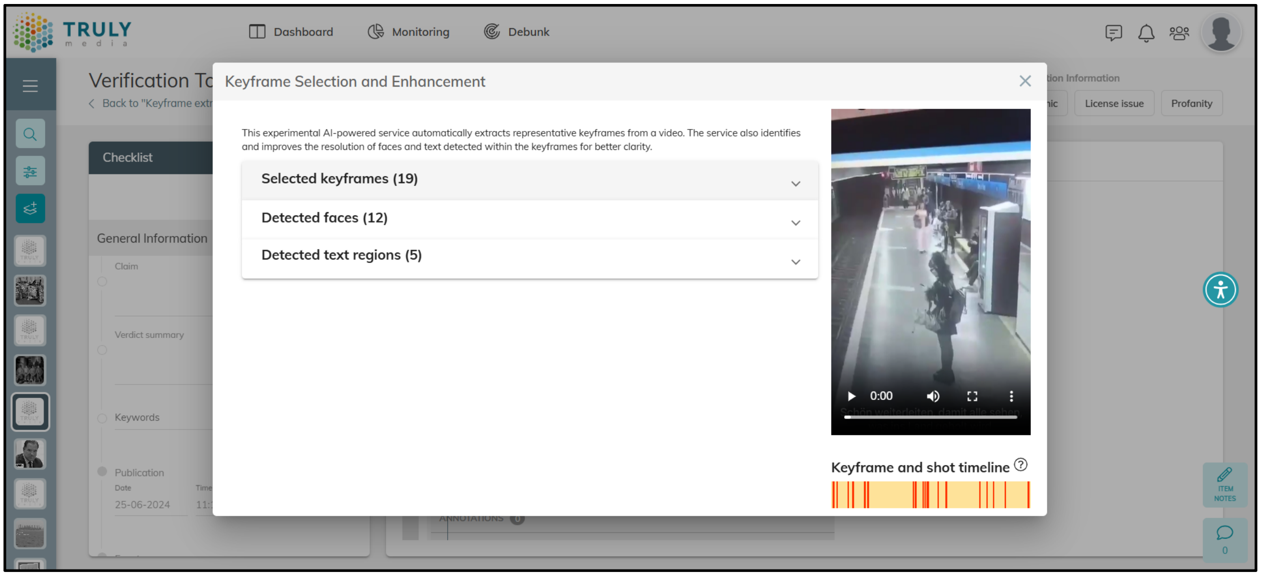
The goal of the keyframes fragmentation by scene of a video is to provide to end users dealing with video verification (mostly fact-checkers, OSINT investigators and human rights defenders) with an automated and quick solution to extract frames representing each scene and to perform directly (through the contextual menu) similarity search through any available search engine. Performing image similarity search on keyframes allows to find if a video was already known and indexed and in which context.

Most of the falsified videos circulating on social media are decontextualised videos, which have been downloaded and reposted, often with minimal alterations but presented in a misleading context. Other fake videos may be the result of a montage of various footage, while others are sometimes altered with video editing software.

In vera.ai, the goal was to go further by enhancing the selection of keyframes and extracting meaningful visual clues within the video such as faces and textual expressions (shops names, plane and boat registration numbers, car plates, tickers, etc.), which may permit to find more indices about the origin, the location, or the main actors involved in the video or in the making of it.

### 6.1.1 Participatory Evaluation

The KSE service was evaluated in two iterations through participatory evaluation. The first iteration in October 2024, a small-scale internal evaluation with four testers, was on the integrated version in Truly Media (Figure 50) to gather first insights into relevant aspects of the service, the specific integration in the platform and existing workflows. Participants had either a background in journalism or design.



Keyframe Selection and Enhancement

This experimental AI-powered service automatically extracts representative keyframes from a video. The service also identifies and improves the resolution of faces and text detected within the keyframes for better clarity.

Selected keyframes (19)

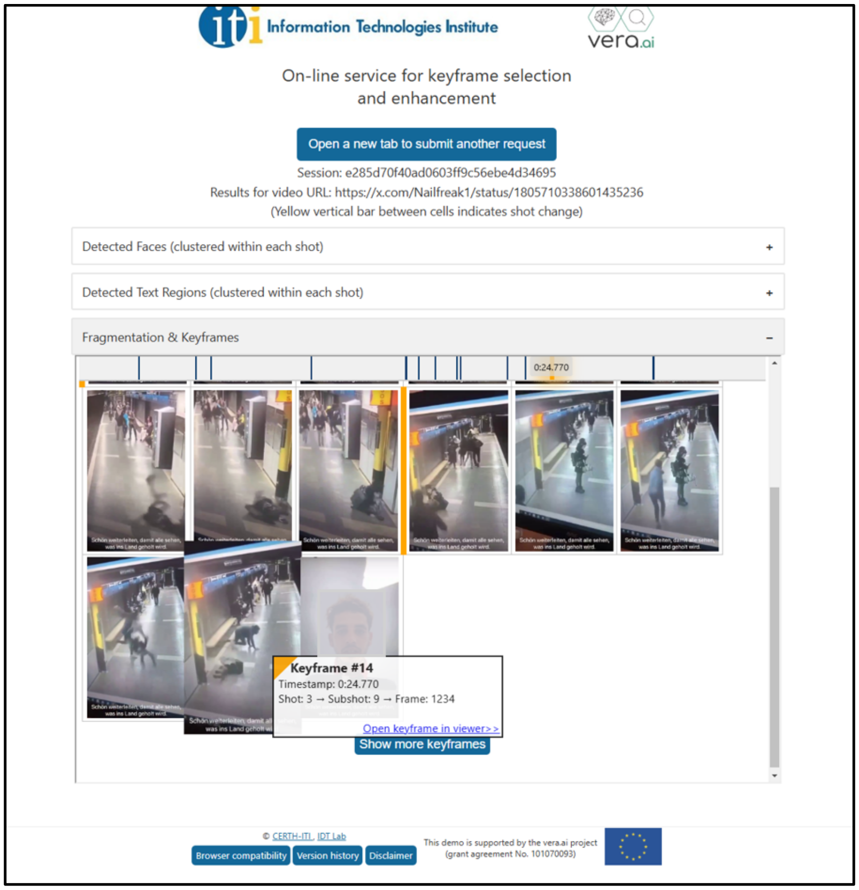
Detected faces (12)

Detected text regions (5)

Keyframe and shot timeline

Figure 50 KSE interface in Truly Media

The service was in its second iteration evaluated in its standalone version with 21 participants from December 2024 to February 2025. These came from University of Siegen, Ca' Foscari University of Venice, RTV, ZDF, CBC-Radio Canada, DW, Radio Romania, BR, Tagesschau, Springer Verlag, Artesmedia, and have backgrounds in fact-checking, media research, journalism, and news production. The evaluation was set-up in a twofold process: Starting with guided freestyle testing, for which testers were equipped with user instructions to test the service on their own for a while and provide written feedback (picking up from posed questions from first evaluation iteration). This was then followed by two remote exchange sessions to enable all participants to join. These sessions followed guided discussions, which were based on written feedback. Focus of this evaluation was the exploration of whether KSE with its extended functionalities of face and text detection, can efficiently support users in their video-verification process and whether it adds additional value to this process taking into consideration the existence of other keyframe tools.



Information Technologies Institute
vera.ai

On-line service for keyframe selection and enhancement

Open a new tab to submit another request

Session: e285d70f40ad0603ff9c56ebe4d34695
Results for video URL: https://x.com/Nailfreak1/status/1805710338601435236
(Yellow vertical bar between cells indicates shot change)

Detected Faces (clustered within each shot) +

Detected Text Regions (clustered within each shot) +

Fragmentation & Keyframes –

0:24.770
Schön weiterleiten, damit alle sehen, was ins Land geholt wird.

Keyframe #14
Timestamp: 0:24.770
Shot: 3 → Subshot: 9 → Frame: 1234
Open keyframe in viewer> >

© CERTH-ITI, IDT Lab

This demo is supported by the vera.ai project (grant agreement No. 101070093)

*Figure 51 KSE interface in standalone version*

Participants confirmed keyframe extraction services to be the most important tools in general and their use being a mandatory initial step in the verification of video content. KSE service (Figure 51) was described to be easy to use as it resembles functionalities from other comparable tools. It was especially the extraction of text regions and the indication of shot changes that can imply tempering, which add an additional value. On the other side, face detection in its current implementation and performance was not convincing to the testers.

Testers collected a number of ideas for additional functionalities and on UI/UX to further improve KSE and make it a service that adds meaningful value to the verification process:

- The importance of **connections to existing services** that are a logical follow-up action in the verification process, such as reverse image search, synthesis detection or audio analyses, was confirmed. Although already existing in the Verification Plugin's contextual menu, such features were considered especially important for the integration in Truly Media in an easy access from e.g. detected text via KSE combined with OCR services for further analysis.

- **Implementing analysis layers fast** (analysis that provides results quickly but maybe less accurate) **and deep** (analysis takes longer but higher accuracy): if users have to decide, they prefer accuracy over speed.

- **Language/text processing** such as adding transcription of video, language detection and OCR from extracted text.
- **More flexibility in keyframe selection** by allowing manual selection (on top of automatically extracted ones), filtering of which keyframes are extracted (overall vs. topical summarisation).
- **Batch analysis** to compare the same video from different sources.
- **Adding percentage rate to which extent one face resembles another one** in the video to allow for assessment of congruency throughout video and foster trustworthiness of the service.
- **UI/UX suggestions**: Upload from media gallery on phones; covering a wide range of formats; better indication of processing while uploading content; adding error messages; more visible representation of where in a video viewer is located.

The evaluation foremost proved the importance of advanced keyframe services that go beyond the pure selection of representative keyframes, which can seamlessly be integrated into users' workflows. However, workflows and users' expertise are multifold, which must be reflected in the tools' design, e.g. through **provision of different levels of use**.

### 6.1.2 Design Thinking Evaluation

Keyframes fragmentation is one of the foundational tools of the Verification Plugin, since its launch in July 2017 as part of the InVID Innovation Action project. It is the most used feature of the plugin, beyond image similarity search. Only during the first six months of 2025, more than 47,000 requests were made through the plugin, according to both our middleware logs and Matomo. Figure 52 shows the origin of the KSE queries worldwide through the plugin.

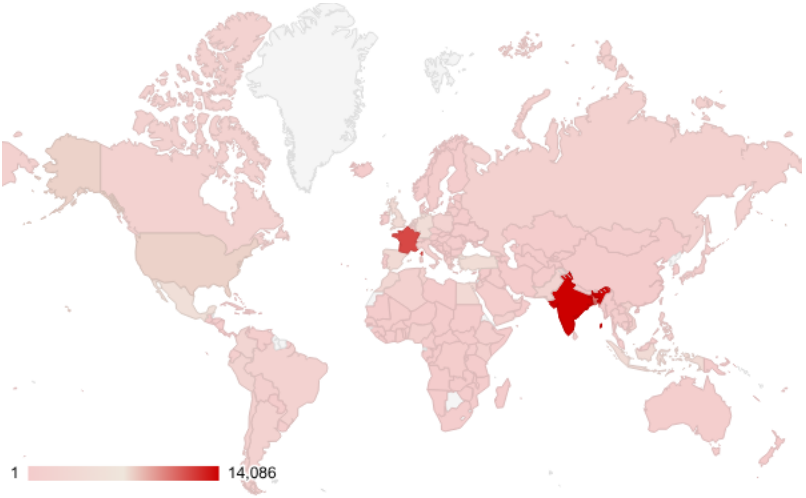


*Figure 52 Geographic distribution of KSE queries*

The new Keyframe Selection and Enhancement Service was tested through the design thinking evaluation in cycle 2, 4 and 6. In cycle 2, we paid more attention to ensuring that the service was performing as well as the previous one, already integrated into the Verification Plugin to avoid any regression. We found that a smaller number of keyframes were retrieved than the version in production and alerted the CERTH team about that.

Regarding new features, we also noticed that the extracted text from images was often partly cut. The method to capture those text boundaries needed to be improved. Regarding faces extracted, some were cut as well and some others look weird, giving the impression of a hallucination of the model (displaying a face not seen in the video). This behaviour was not reproduced after improvement of the model.

From the usage data, we identified a potential risk of misuse. This was demonstrated when an AFP fact-checker utilised an extracted keyframe of the Princess of Wales (Figure 53) to test our synthetic image detection. The use of super-resolution to reconstruct faces from a video fragment could be mistakenly identified as an AI-generated image, leading to a detection based on the AI-enhanced image rather than on the original footage. In that specific case, the submitted image was not classified as synthetic and remained below our empirical 70% detection threshold.

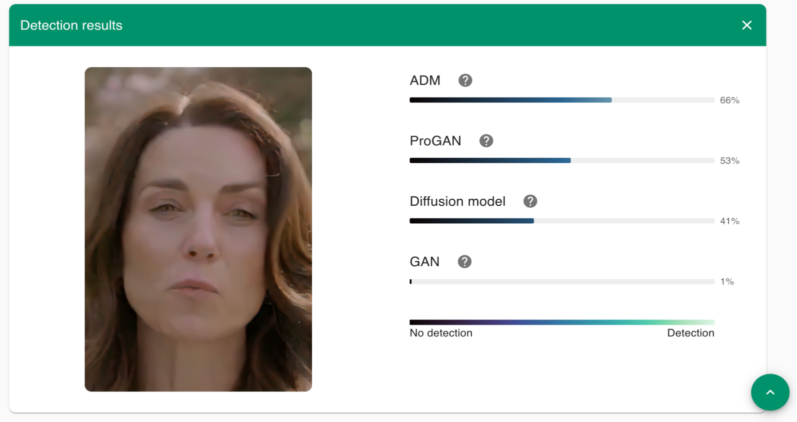
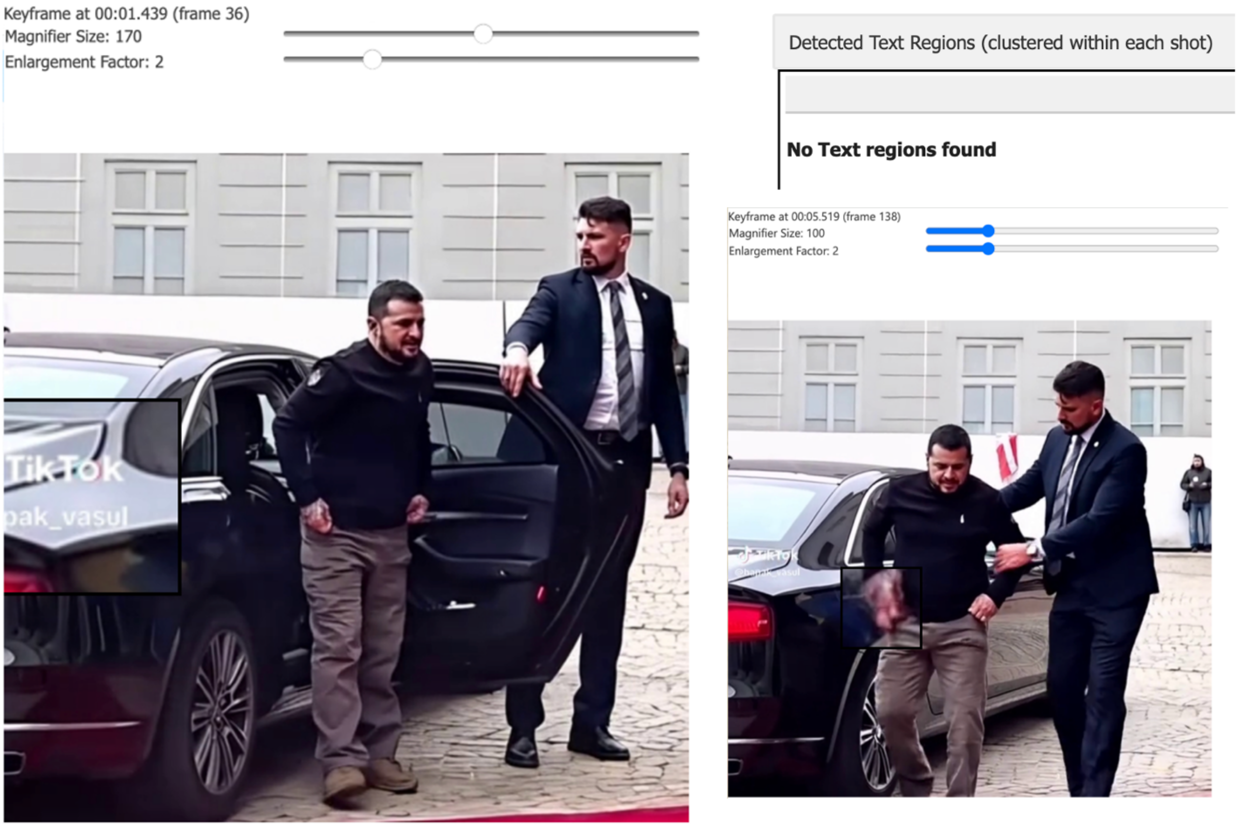


*Figure 53 View of a keyframe being tested for synthetic image detection*

In cycle 4, we suggested some improvements to the user interface such as extracting textual logos with account names to alert end-users about the likely provenance of a video as shown in the TikTok example below (Figure 54) where logo and account were not previously identified.



Keyframe at 00:01.439 (frame 36)
Magnifier Size: 170
Enlargement Factor: 2

Detected Text Regions (clustered within each shot)

**No Text regions found**

Keyframe at 00:05.519 (frame 138)
Magnifier Size: 100
Enlargement Factor: 2

*Figure 54 View of the KSE interface with a fake video denigrating Ukrainian President Volodymyr Zelensky*

We then proposed to CERTH another idea for consideration: to reveal automatically within KSE the flaws left by manipulative video editing. The above Zelensky's video example (Figure 54), decontextualised from an official visit to Poland, was altered to denigrate the Ukrainian President by showing him stumbling as he exited his car. While this video fragment was clearly tampered with (where a forearm and hand are missing for a brief moment in the right keyframe), systematically outlining flaws and anomalies in the video stream (e.g. with a visual affordance coloured frame) would be a great asset, as these are often the primary visual clues fact-checkers use to determine if a video was manipulated.

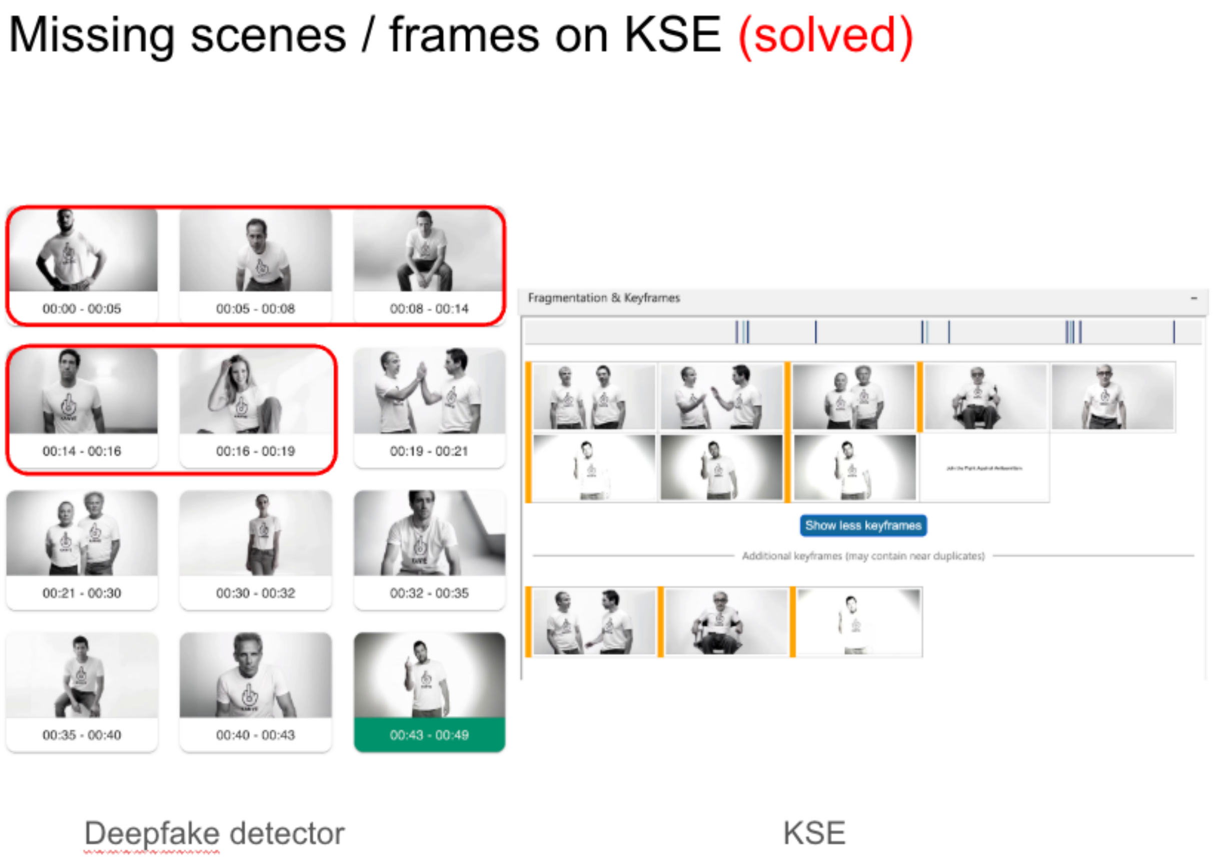
However, distinguishing between AI-generated anomalies and compression artifacts remains quite difficult. Several polemical examples with unverified claims of deepfakes have arisen previously during the project, like an encounter of Russian President Vladimir Putin with air hostesses[37] and the Princess of Wales's announcement of her illness, when her ring seemed to disappear in low resolution videos[38].

In January 2025, given the progress accomplished, KSE was integrated first in the file mode of the Verification Plugin, and later in the link mode (in production and open to all users), in replacement of the previous version.

[37] https://x.com/Shayan86/status/1500214745718312975

[38] https://fullfact.org/news/kate-cancer-video-ring-vanishes-false/

Shortly after, while helping fact-checkers with a viral video generated by artificial intelligence[39], depicting a virtual demonstration of disapproval by Jewish public figures against the rapper Kanye West, we discovered missing keyframes in the sequence between KSE and the video deepfake detector, as shown by Figure 55. This issue was very rapidly solved by CERTH.



Missing scenes / frames on KSE (solved)

Deepfake detector

KSE

*Figure 55 Evaluation restitution slide with a comparison of a keyframe extraction between the Deepfake Detector and KSE*

In this use case, the AI origin of the content, both acknowledged by the account owner of the creator on Instagram and by Meta, who flagged it as generated by their own AI, was no mystery. Nevertheless, KSE was useful to extract meaningful signs of manipulation through the multiple deformed and distorted faces present in the video, as exemplified in Figure 56 below.

[39] https://x.com/martinvars/status/1889390165321486632

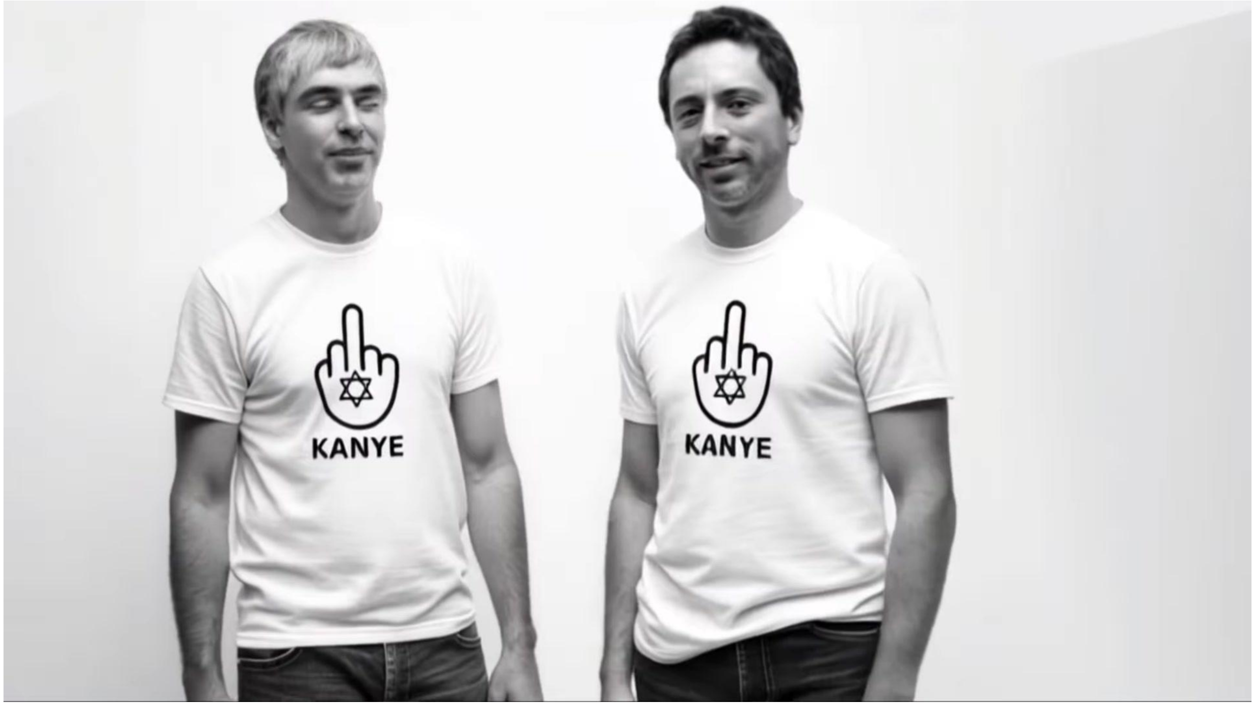


*Figure 56 View of a keyframe with distorted faces extracted by KSE from a GenAI video*

With the rise of fully AI-generated videos in the project last months, we had more similar use cases in cycle 6, where keyframe extraction was used several times, both for identifying anomalies in videos and to query successfully watermarking detection tools such as Google Deepmind SynthID Detector. Examples include a fake Rafale plane wreck[40] following an incident between India and Pakistan, a fake Ukrainian demo asking forgiveness to Russia[41], a fictional pro-Bolsonaro demo near the White House[42] and a viral fake video of Israelis allegedly demonstrating to apologise for bombing Iran[43].

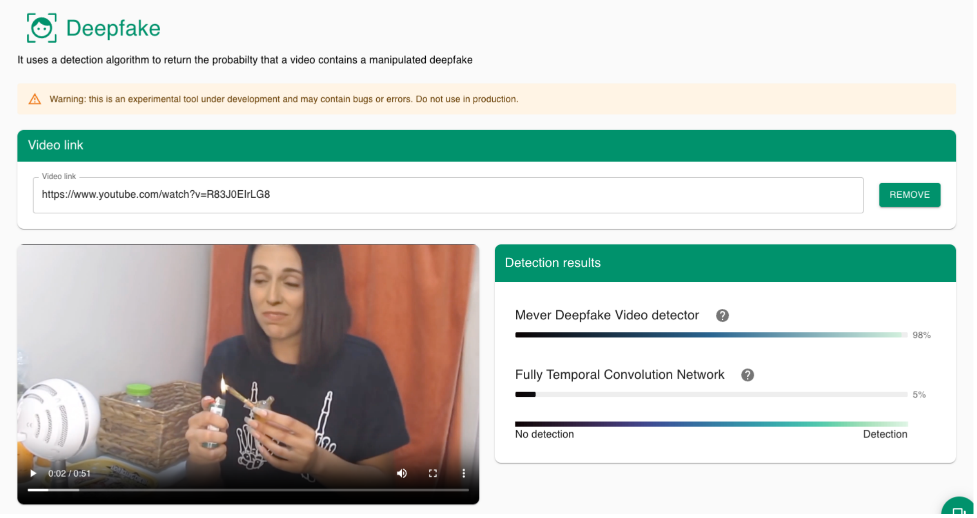
A final survey with end-users during the summer of 2025 (see section 2.1.1) showed that KSE remains the most valued feature of the Verification Plugin with a score above 4.44 out of 5 on a Likert scale. It is perceived by end-users as "quick, easy, convenient and very helpful", the "N1 feature for daily work" despite the known limitations imposed by video platforms to access and fetch their content.

## 6.2 Deepfake Video Detection

The Deepfake Video Detection service was tested through the design thinking methodology in cycles 2, 3 and 4. This detector was initially created during the previous WeVerify project and was only open to a small group of Verification Plugin beta testers, the initial 80 beta users we had before starting vera.ai evaluation sessions.

[40] https://factcheck.afp.com/doc.afp.com.47964WD
[41] https://fakty.afp.com/doc.afp.com.676V2NQ
[42] https://checamos.afp.com/doc.afp.com.68CR2ZT
[43] https://fakty.afp.com/doc.afp.com.64A7262, https://proveri.afp.com/doc.afp.com.64ZG863 and https://sprawdzam.afp.com/doc.afp.com.64JV3CQ

The underlying AI model has undergone retraining to improve its performance. Previously, we recommended end-users to avoid using it in production because of false positives. The new version showed better results on one particular deepfake modality: face swapping, which was the main goal of the retraining as shown in Figure 57.



Deepfake

It uses a detection algorithm to return the probabilty that a video contains a manipulated deepfake

⚠ Warning: this is an experimental tool under development and may contain bugs or errors. Do not use in production.

Video link

Video link: https://www.youtube.com/watch?v=R83J0EIrLG8

Detection results

Mever Deepfake Video detector ? 98%

Fully Temporal Convolution Network ? 5%

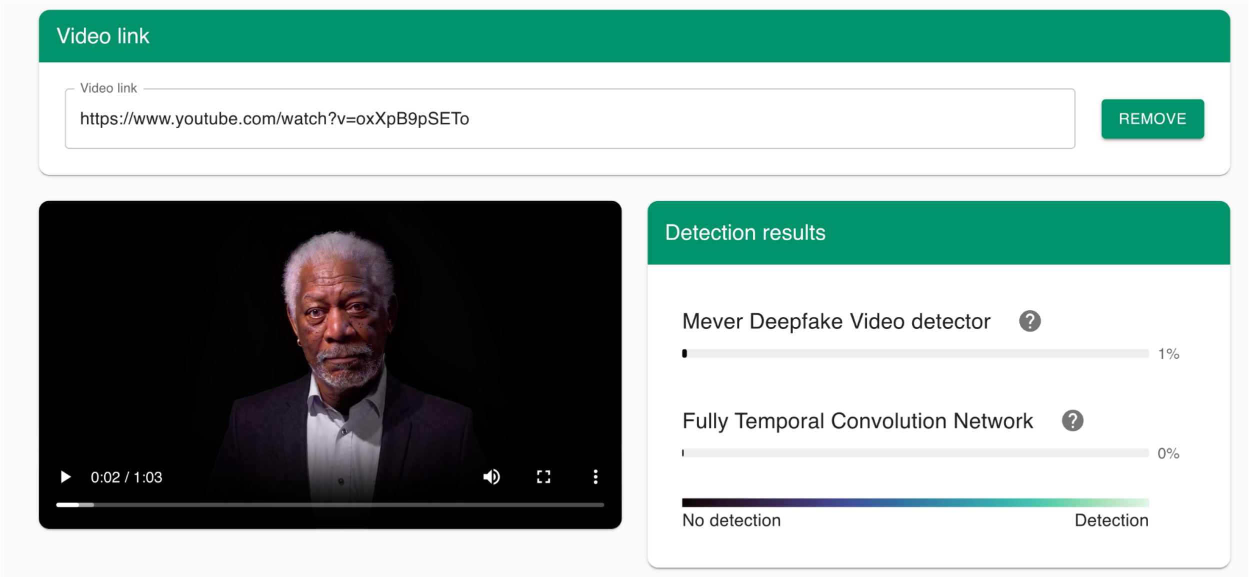
*Figure 57 Screenshot of a previous deepfake detection on a fake video of New Zealand PM Jacinda Ardern*

Early feedback received from fact-checkers emerged during a training session at the Global Fact 10 conference in Seoul in June 2023. There, a prominent speaker publicly demonstrated the plugin's limitations in deepfake detection, using a well-known deepfake video featuring American actor Morgan Freeman (see Figure 58 below).

This incident underscored the need for greater precision concerning the specific types of deepfake manipulation the model was designed to detect. The word "deepfake" functions as a portmanteau, serving as an umbrella word for numerous manipulation methods that are fundamentally dissimilar.

This is the case of the deepfake video featuring Morgan Freeman. Up to our knowledge, it is only detected as a deepfake by a lip-syncing method from the State University of New York at Buffalo's DeepFake-o-meter[44]. The inability of our vera.ai tool to detect lip-syncing stems from this being a relatively rare deepfake manipulation technique that was not included in its training data.

[44] https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/



Video link

Video link

https://www.youtube.com/watch?v=oxXpB9pSETo

Detection results

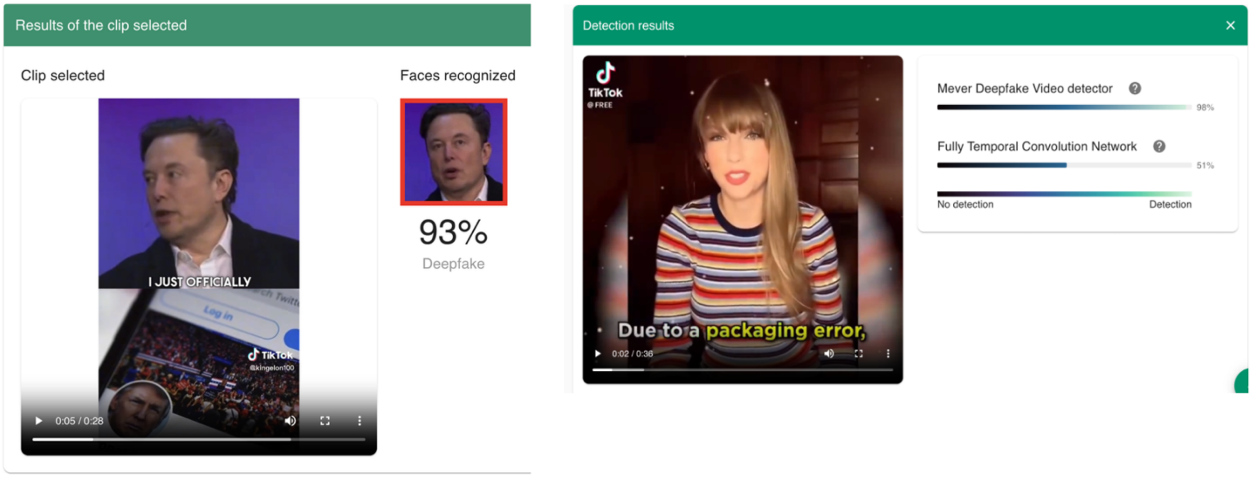
Mever Deepfake Video detector 1%

Fully Temporal Convolution Network 0%

No detection Detection

*Figure 58 Screenshot of a deepfake of Morgan Freeman*

Overall, the newly trained deepfake detector showed good performance to detect new samples circulating on social networks, including crypto-currencies scams and fake ads (see Figure 59 below).



Results of the clip selected

Clip selected

Faces recognized

93%
Deepfake

Detection results

Mever Deepfake Video detector 98%

Fully Temporal Convolution Network 51%

No detection Detection

*Figure 59 Combined view of a deepfake of Elon Mush on crypto-currencies and a deepfake of Taylor Swift promoting a fake ad*

Furthermore, the tool demonstrated improved performance in reducing the risk of previous false positives, for instance, by avoiding the misleading detection of "making-of" content as a deepfake, as exemplified by British TV Channel 4's Queen Elizabeth deepfake creation process (Figure 60 below).

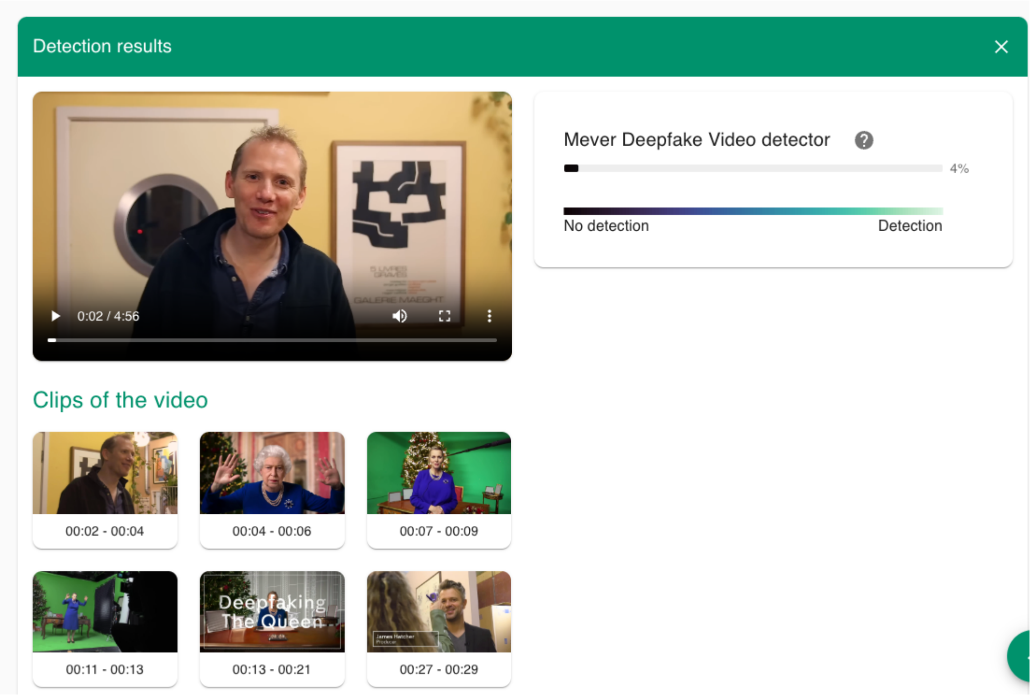
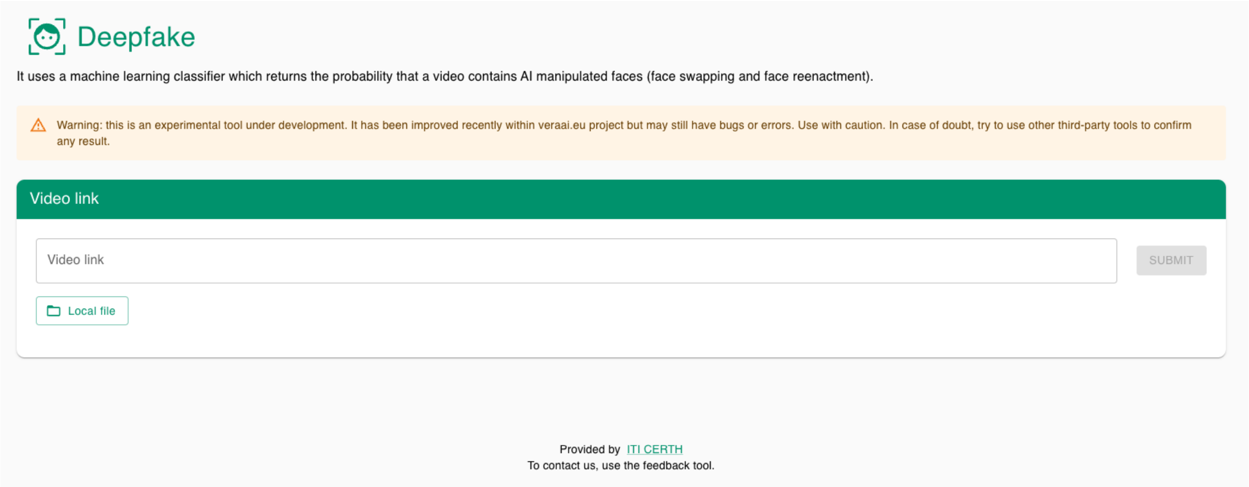


*Figure 60 Screenshot of a correct non-detection of a deepfake "making-of", broadcasted by British TV Channel 4*

Following the previous findings, we improved the interface by adding more details about the tool capabilities and by updating the warning as showcased in Figure 61 below.



It uses a machine learning classifier which returns the probability that a video contains AI manipulated faces (face swapping and face reenactment).

⚠ Warning: this is an experimental tool under development. It has been improved recently within veraai.eu project but may still have bugs or errors. Use with caution. In case of doubt, try to use other third-party tools to confirm any result.

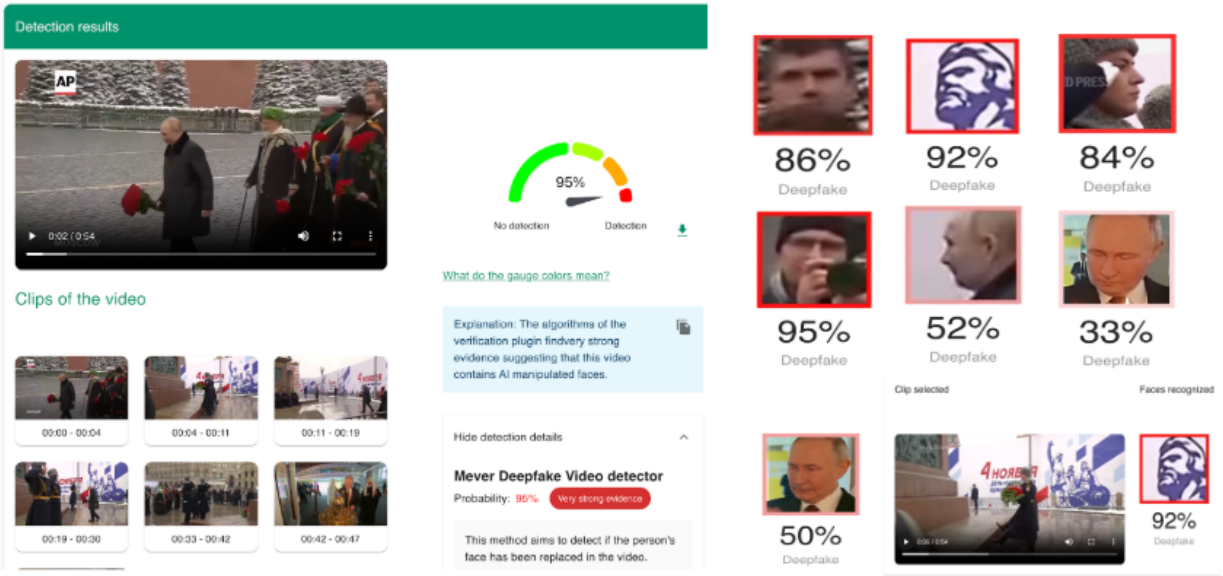
Provided by ITI CERTH
To contact us, use the feedback tool.

*Figure 61 View of the improved Deepfake Video Detection service interface in the plugin*

The tool now clearly indicates its scope of capabilities (face swapping and face re-enactment) and informs users of both recent improvements and remaining limitations, including giving tips to overcome them (Figure 61). We also reused the scale, gauge and colour schema implemented in the synthetic image detection to harmonise the Verification Plugin layout. "The UI evolved making the results understandable without ambiguity, the introduction of the gauge instead of a colour-graded bar demonstrated to be a key upgrade in the UI," according to EUDL's investigators. Another significant enhancement introduced was the capability to upload local files to the detector, given that the tool's backend frequently encountered increasingly stringent scraping limitations imposed by platforms.

In cycle 4, both EUDL and AFP reviewed more thoroughly the risk of false positives and found several flaws, especially with background faces in lower resolution videos.

https://www.youtube.com/watch?v=uW1T4INtm4A


Detection results

Explanation: The algorithms of the verification plugin findvery strong evidence suggesting that this video contains AI manipulated faces.

Hide detection details

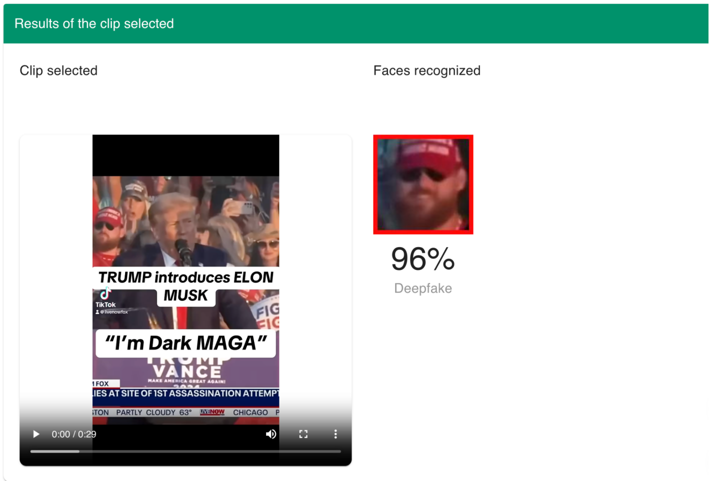
Mever Deepfake Video detector

Probability: 95% Very strong evidence

This method aims to detect if the person's face has been replaced in the video.

*Figure 62 Evaluation screenshot of a deepfake false positive likely due to excessive compression of background faces*

The above example (Figure 62), coming from an AP video on a Kremlin ceremony, was identified by EUDL investigators. It exemplifies how background faces, often blurred in lower resolution online videos, are mistakenly identified as face swapping or re-enactment.


Results of the clip selected

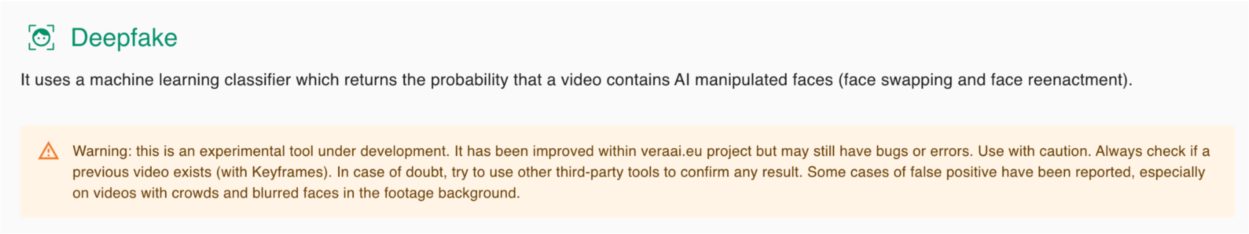
Clip selected

Faces recognized

*Figure 63 Screenshot of an incorrect detection on a background face*

The same kind of issues arose with political meetings in rallies of the 2024 US elections like in Figure 63 above, showing a Trump meeting where background faces are mistakenly identified. Those kinds of errors should not mislead an experimented fact-checker but cast doubt about the use of the tool in a production environment beyond giving just an indication. The tool should be primarily utilized when face-swapping manipulation is suspected (as it can also be cross-detected through similarity search).

We therefore updated the warning on top of the tool in the Verification Plugin to reflect our findings (Figure 64).



Deepfake

It uses a machine learning classifier which returns the probability that a video contains AI manipulated faces (face swapping and face reenactment).

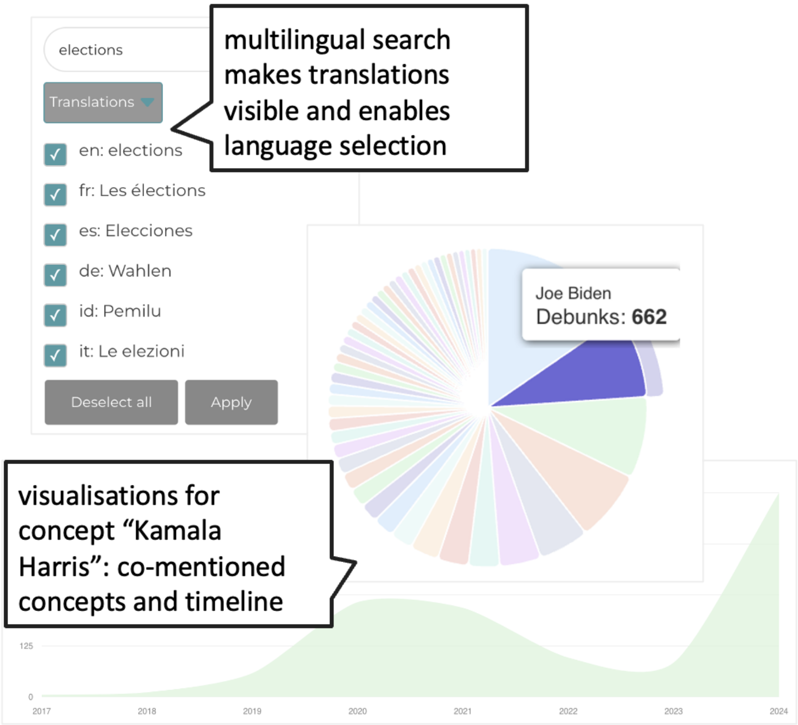
⚠ Warning: this is an experimental tool under development. It has been improved within veraai.eu project but may still have bugs or errors. Use with caution. Always check if a previous video exists (with Keyframes). In case of doubt, try to use other third-party tools to confirm any result. Some cases of false positive have been reported, especially on videos with crowds and blurred faces in the footage background.

*Figure 64 Update of the warning message following evaluation findings*

In cycle 6, we were confronted with another risk: that of the tool being used and weaponised in a use case from Israel's war on Gaza. A picture and two videos of a Palestinian journalist, covered with blood after an attack against a café in Gaza, in early July 2025, circulated online and went viral amid accusations of being staged. Both videos, one of them showing her removing blood from her face in a mirror, were detected as deepfakes, although there was no suspicion of face-swapping in both footages. We strongly recommended our fact-checking editor in charge to avoid using those results as a proof in such an uncertain situation.

In our final survey, the Deepfake Video Detection tools achieved a score of 3.22 out of 5 on a Likert scale which reflects the difficulties encountered. This, in turn, reveals the complexity of deepfakes and their multiple underlying manipulation techniques, which require specifically trained ad-hoc detectors and make any attempt at universal detection very difficult.

## 6.3 Database of Known Fakes (DBKF)



elections

Translations

☑ en: elections

☑ fr: Les élections

☑ es: Elecciones

☑ de: Wahlen

☑ id: Pemilu

☑ it: Le elezioni

multilingual search makes translations visible and enables language selection

visualisations for concept “Kamala Harris”: co-mentioned concepts and timeline

*Figure 65 Language filtering, search query-related concepts pie-chart and timeline in DBKF*

The Database of Known Fakes (DBKF, Figure 65) is a searchable archive initially developed in the WeVerify Innovation action, where users can check whether a claim, image or video has already been debunked by trusted fact-checking sources (International Fact-Checking Network (IFCN) signatories).

### 6.3.1 Participatory Evaluation

The participatory evaluation of the Database of Known Fakes took place in February 2024 and focused on the functionality of multilingual search, and on results presentation in this service. The process included:

- Internal testing to iron out usability issues that might get in the way of the evaluation.
- Asynchronous testing in the users’ own time, with the help of user instructions and including a questionnaire with a qualitative focus that revealed themes of interest to testers.
- Two live online sessions based on the questionnaire responses, aiming at exploration of the emerged themes.

The evaluation included 14 users (6 journalists, 5 innovation experts, 1 broadcasting news director, and 2 academic media researchers) from DW, SRG-SSR, France Télévisions, ERT, University of Siegen, and Ca' Foscari University of Venice – split into two evaluation groups.

User expectations were to be able to use this to search for debunked articles without having to rely on regular search engines and their many irrelevant search results, as well as do it fast with quick overviews of results and efficient language search support.

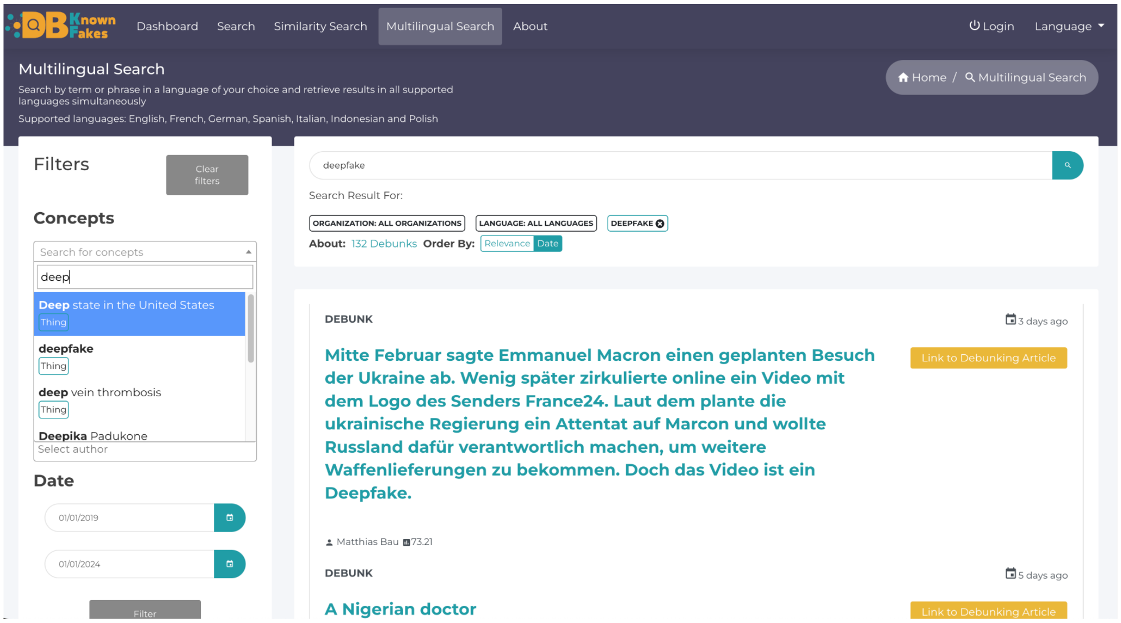
These expectations were met in that the users found the tool easy to use, reliable since based on content from IFCN-members, and focused on relevant information without commercial interference. They considered that it could be incorporated into their workflow, provided a few changes were made in terms of filtering and a clearer overview in the information visualisation, as well as ideas for new functionalities (such as alerts for new items in a search and an automatic translation functionality).

Based on evaluation feedback, the DBKF team made visible in the UI the AI-powered machine translations of search queries, which previously happened behind the scenes. This enables users to choose language(s) and to expect trustworthy search results. In terms of additional information visualisations, a separate tab in the search results page now features a timeline (to get a sense of how popular the search query was over time) and a pie chart of co-mentioned concepts, which are extracted by an AI algorithm (see Figure 65 above).

Besides considerations about language search, usability, and presentation of results, the key emerging themes central to this evaluation turned out to be trust and speed of use, which were considered as the most essential aspects in the use of this service.

## 6.3.2 5.3.2 Design Thinking Evaluation

The AFP team conducted two evaluations of the database during cycles 2 and 5. The first evaluation focused on the concept search (Figure 66), which had a recurring issue with mixing terms from different languages. This problem was resolved in the subsequent new release.



Multilingual Search

Search by term or phrase in a language of your choice and retrieve results in all supported languages simultaneously

Supported languages: English, French, German, Spanish, Italian, Indonesian and Polish

Concepts

Search for concepts

deep

**Deep** state in the United States
Thing

**deepfake**
Thing

**deep** vein thrombosis
Thing

**Deepika** Padukone

Select author

Date

01/01/2019

01/01/2024

deepfake

Search Result For:

ORGANIZATION: ALL ORGANIZATIONS | LANGUAGE: ALL LANGUAGES | DEEPFAKE

**About:** 132 Debunks **Order By:** Relevance Date

DEBUNK

3 days ago

**Mitte Februar sagte Emmanuel Macron einen geplanten Besuch der Ukraine ab. Wenig später zirkulierte online ein Video mit dem Logo des Senders France24. Laut dem plante die ukrainische Regierung ein Attentat auf Marcon und wollte Russland dafür verantwortlich machen, um weitere Waffenlieferungen zu bekommen. Doch das Video ist ein Deepfake.**

Matthias Bau 73.21

DEBUNK

5 days ago

**A Nigerian doctor**

*Figure 66 Screenshot of the DBKF search interface with a query on "deep" concept*

The concept search has been greatly enhanced by an auto-completion feature that streamlines the process of exploring the database. We found it particularly useful for retrieving deepfake examples to support a related study (Teyssou, 2025). Ultimately, the feature was very helpful, providing interesting information that complemented results from the Google Fact Check Explorer.[45]

Regarding usability, we suggested that ONTO create a clearer distinction between concepts/events and the other filters. We made this recommendation because concepts and events retrieve different types of pre-processed documents, while faceted search filters are designed to refine an existing user query.

This suggestion was later implemented effectively in the interface by introducing a "Semantic Search" section (containing Concepts and Event Types) separate from the standard refinement filters.

Separately, we noted that concepts were not listed in alphabetical order. ONTO explained that this is intentional; they use a PageRank-style algorithm to display the most relevant concepts first, rather than sorting them alphabetically.

When we tested the image similarity search using pictures from the database, we found it did not always retrieve every debunk associated with that image. The performance gap became clear when we compared the results to those from Google Fact Check Explorer. The ONTO team suggests a potential problem with image indexation during content ingestion could be the cause.

This remains a significant concern, as the Verification Plugin uses this functionality to power its core automatic verification features for text, images, and videos.

We improved the relevance of search results in the Verification Plugin Assistant by adjusting the textual similarity threshold to filter out unrelated content.

Further testing during cycle 5 by EUDL researchers identified several other usability issues. They noted the date search is not straightforward, and retrieved claims are often sorted improperly. They also found that the chatbot's responses to queries were limited compared to those from a keyword search.

## 6.4 Image Forensic Detection (AI methods)

Partners CERTH and UNINA introduced in vera.ai new AI-based digital forensic methods to detect image forgery. The AFP team first tested those new methods through the partners' demonstrator[46] during cycle 2, in March 2024.

We tested among fake pictures a recent viral one showing the late Russian opponent Alexei Navalny[47] allegedly making a nazi salute in a demo, with the body covered with tattoos, including one of Adolf Hitler.

This fake picture, although already debunked by Spanish fact-checker Maldita[48] three years before, surfaced again aiming to tarnish Navalny's memory at a time where his death was causing international

[45] https://toolbox.google.com/factcheck/explorer/

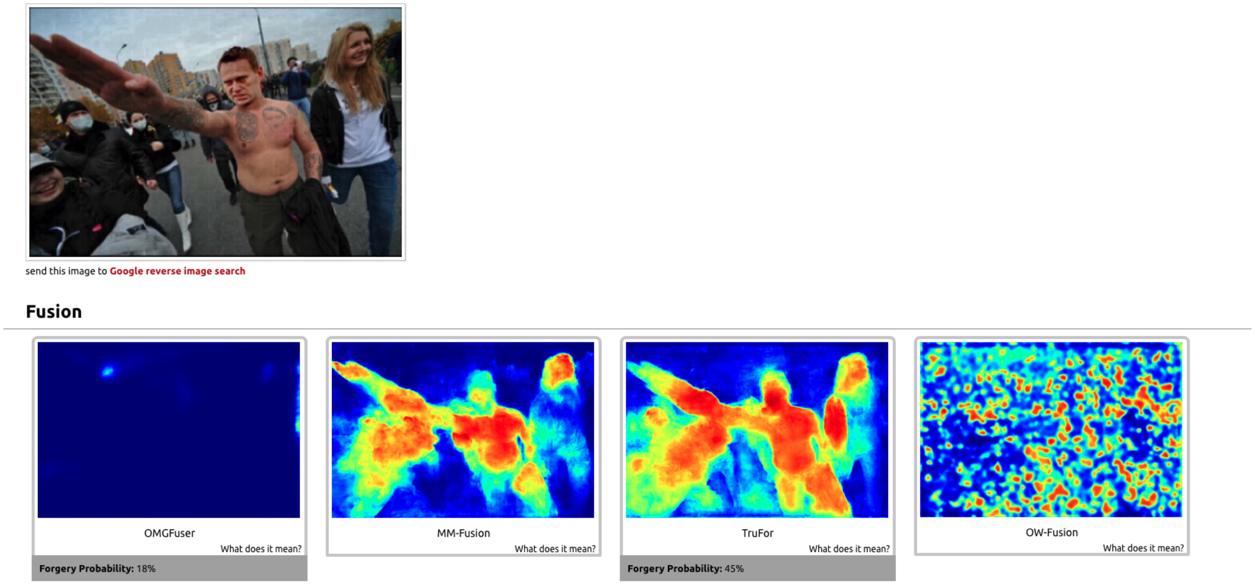
[46] https://mever.iti.gr/forensics/

[47] Navalny died as a political prisoner in a Western Siberia Russian prison on February 16, 2024.

[48] https://maldita.es/malditobulo/20210204/alexei-navalny-saludo-fascista/

outrage. It was debunked again[49] via a similarity search that found a very similar picture depicting a Moscow rally against immigration. In that photo, the face of the main character was swapped and replaced by Navalny's visage.

Previous forensic methods were inconclusive and inoperant. Two AI-methods, MM-Fusion (CERTH) and TruFor (UNINA) were able to detect the forgery as shown by Figure 67 below.



send this image to Google reverse image search

Fusion

OMGFuser
Forgery Probability: 18%

MM-Fusion

TruFor
Forgery Probability: 45%

OW-Fusion

*Figure 67 Forensic methods MM-Fusion and TruFor reveal the forgery of an Alexei Navalny's fake picture*

Other experimental methods like OMGFuser or Noisesniffer (ENS) were not conclusive on the same picture (as shown in Figure 67 and Figure 68 respectively).

[49] https://faktencheck.afp.com/doc.afp.com.34L7742 , https://dpa-factchecking.com/netherlands/240222-99-80195/ https://correctiv.org/faktencheck/2024/02/28/alexej-nawalny-zeigt-hitlergruss-nein-dieses-foto-ist-manipuliert/
https://observers.france24.com/en/fake-news-russian-activist-navalny-discredit-him-after-his-death

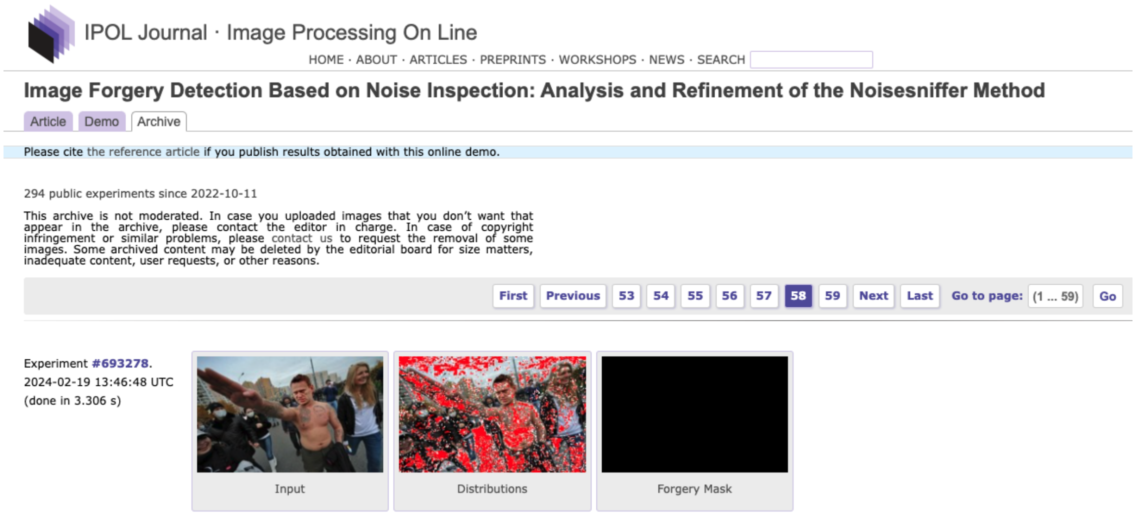


Image Forgery Detection Based on Noise Inspection: Analysis and Refinement of the Noisesniffer Method

Please cite the reference article if you publish results obtained with this online demo.

294 public experiments since 2022-10-11

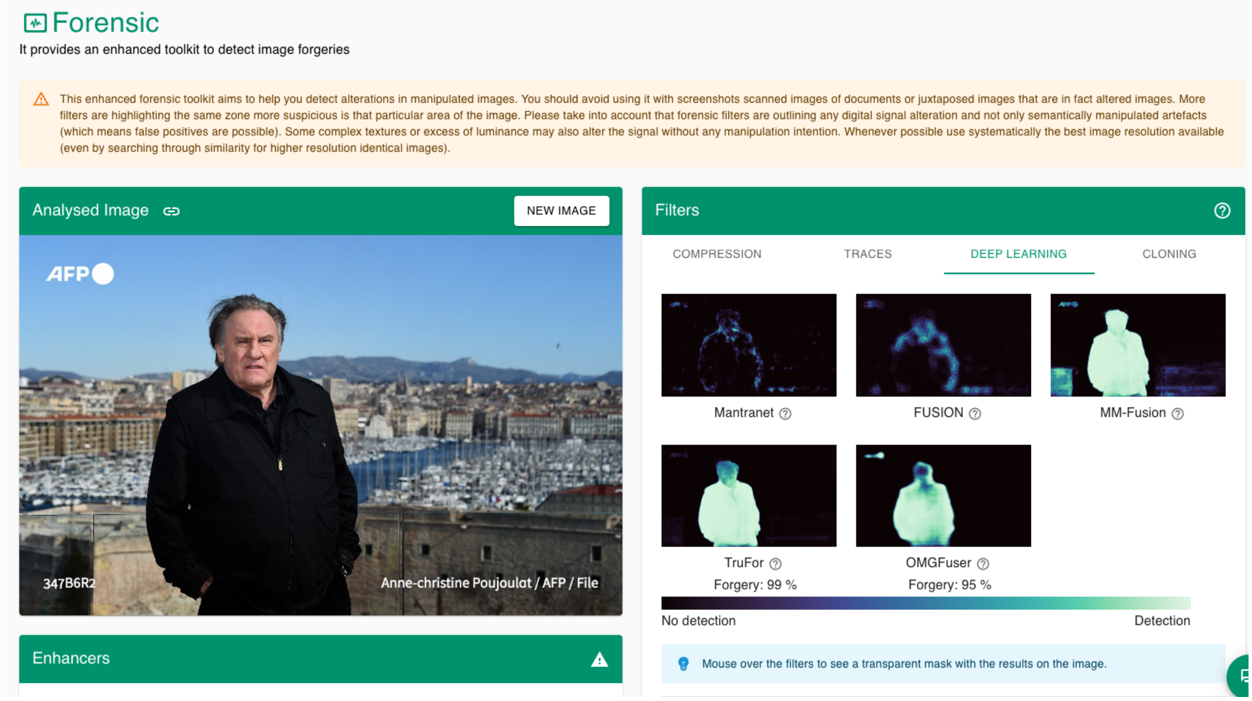
This archive is not moderated. In case you uploaded images that you don't want that appear in the archive, please contact the editor in charge. In case of copyright infringement or similar problems, please contact us to request the removal of some images. Some archived content may be deleted by the editorial board for size matters, inadequate content, user requests, or other reasons.

Experiment #693278.
2024-02-19 13:46:48 UTC
(done in 3.306 s)

*Figure 68 A view of the Noisesniffer method published in IPOL Journal with the fake Navalny's picture*

Following those preliminary tests, we integrated the AI-based methods into the Verification Plugin in version v0.80 and ran with partner EUDL new tests during cycle 4 and beyond.

We started to include in our tests some real images to check the new methods against false positives. The picture below (Figure 69) shows the French actor Gérard Depardieu in front of the old harbour in Marseille. It was taken by the AFP news agency and posted on its social media accounts. Despite being a real photo, it was flagged as a forgery by all three new methods.



It provides an enhanced toolkit to detect image forgeries

This enhanced forensic toolkit aims to help you detect alterations in manipulated images. You should avoid using it with screenshots scanned images of documents or juxtaposed images that are in fact altered images. More filters are highlighting the same zone more suspicious is that particular area of the image. Please take into account that forensic filters are outlining any digital signal alteration and not only semantically manipulated artefacts (which means false positives are possible). Some complex textures or excess of luminance may also alter the signal without any manipulation intention. Whenever possible use systematically the best image resolution available (even by searching through similarity for higher resolution identical images).

TruFor
Forgery: 99 %

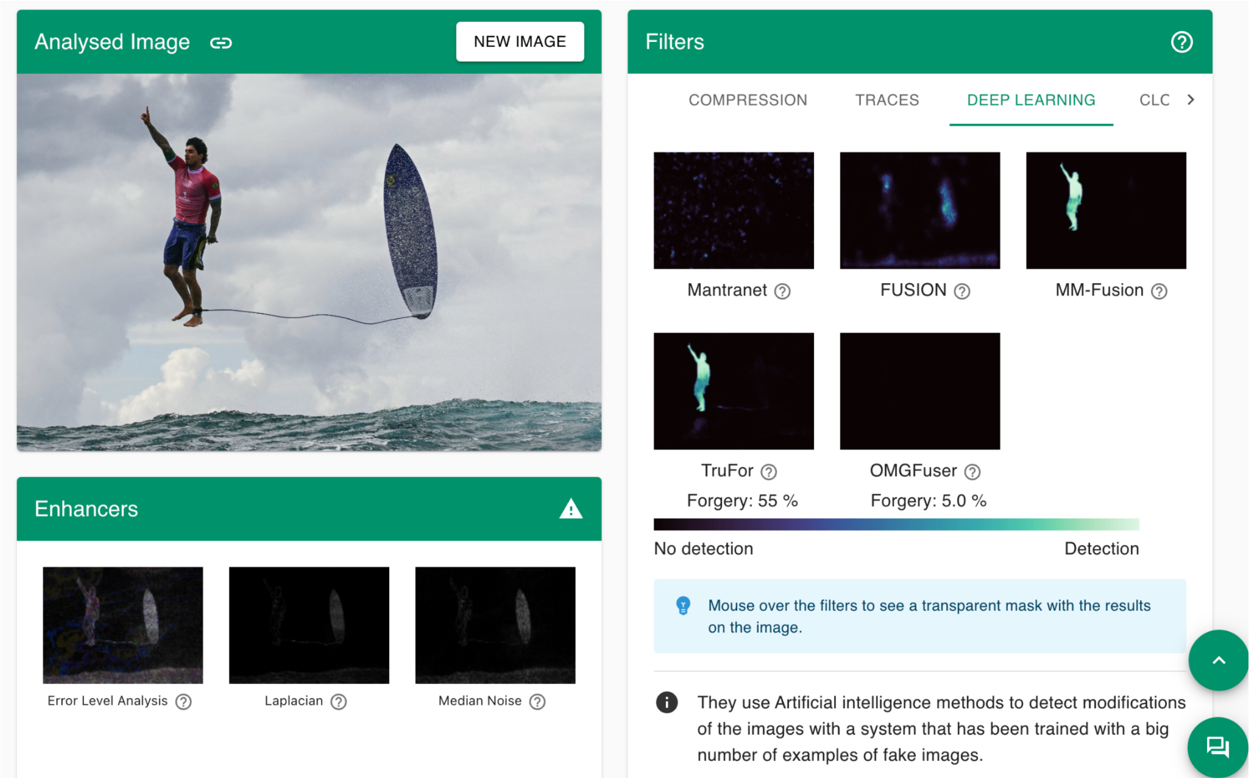
OMGFuser
Forgery: 95 %

Mouse over the filters to see a transparent mask with the results on the image.

*Figure 69 A false positive on an AFP picture of French actor Gérard Depardieu in Marseille*

Previous similar pictures not carrying the AFP logo, the photo number and the photographer's name are not detected as such. The process of adding the logo, the number and the name was made with a simple JavaScript application.

Two methods also detected as a forgery a copy (Figure 70) circulating on X social network of an iconic and offbeat AFP picture from the Olympic Games, taken during the surf competition in Tahiti's Island (French Polynesia). A Brazilian surfer was captured levitating above the sea's level with his surfboard positioned vertically a few feet to his right.



TruFor Forgery: 55 %

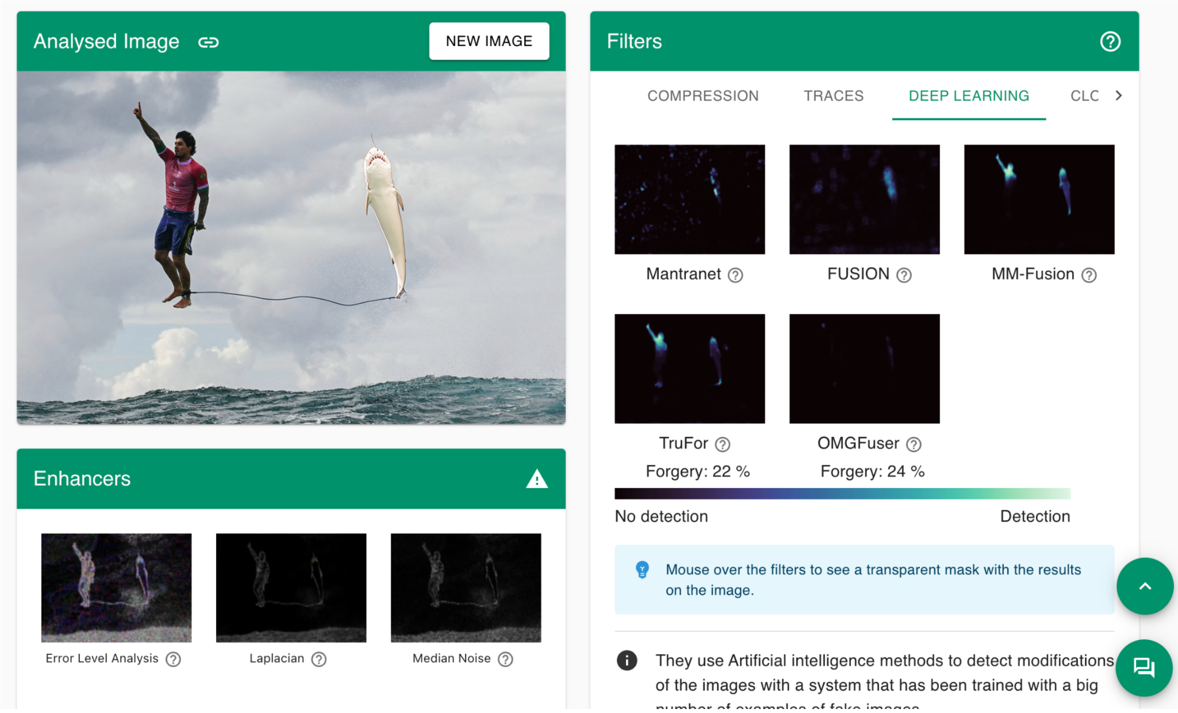
OMGFuser Forgery: 5.0 %

Mouse over the filters to see a transparent mask with the results on the image.

They use Artificial intelligence methods to detect modifications of the images with a system that has been trained with a big number of examples of fake images.

*Figure 70 A false positive on an AFP picture collected in HEIF during the Olympic Games*

The original copy from camera, retrieved from AFP, gives the same kind of false positive. The fact that the original image was collected in the more compressed HEIF format, then converted in JPEG format, therefore undergoing a double resampling (on colours and on re-compression) may have caused this detection. A humoristic forgery of the previous picture, shared on X, was paradoxically less detected as tampered when the surfboard was replaced by a shark (Figure 71).



TruFor Forgery: 22 %

OMGFuser Forgery: 24 %

Mouse over the filters to see a transparent mask with the results on the image.

They use Artificial intelligence methods to detect modifications of the images with a system that has been trained with a big number of examples of fake images

*Figure 71 Forensic analysis of a tampered image from an AFP picture of the Olympic Games*

The following example (Figure 72), which spread on social media in August 2024, was an effort by Trump's detractors to symbolically link him to Hitler by falsifying an old, mirrored Roger Viollet picture and mocking his iconic rescue image after the assassination attempt he underwent.

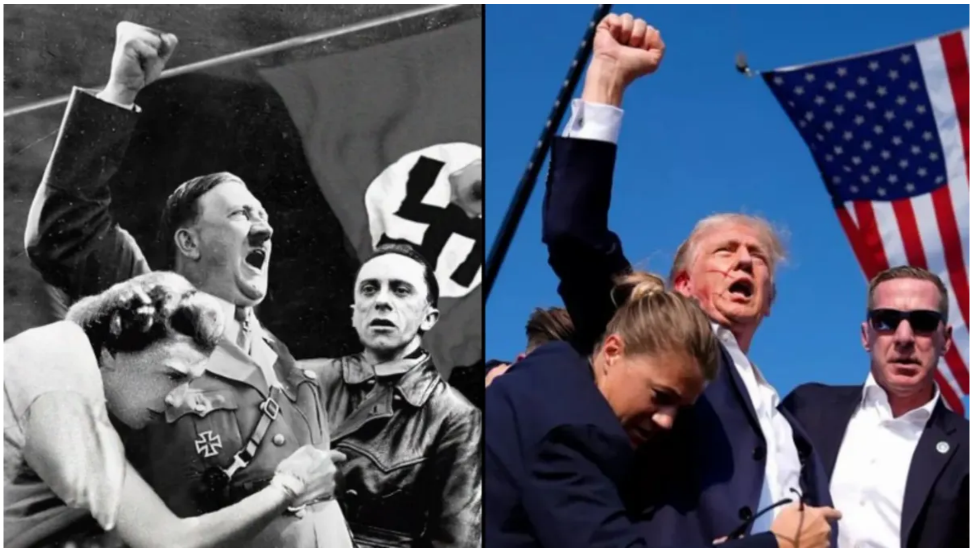
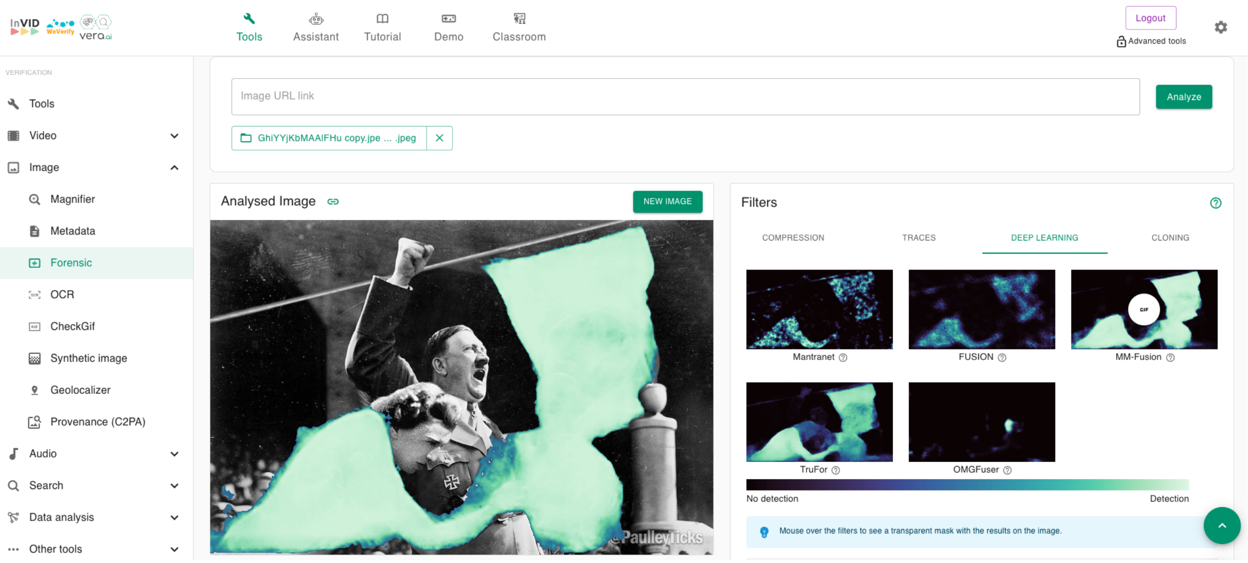


*Figure 72 A juxtaposition of pictures to create a symbolic link between present and a fabricated past*

In that case, the forgery of the older picture was fully detected by both AI methods Trufor and MM-Fusion as shown in Figure 73 below.

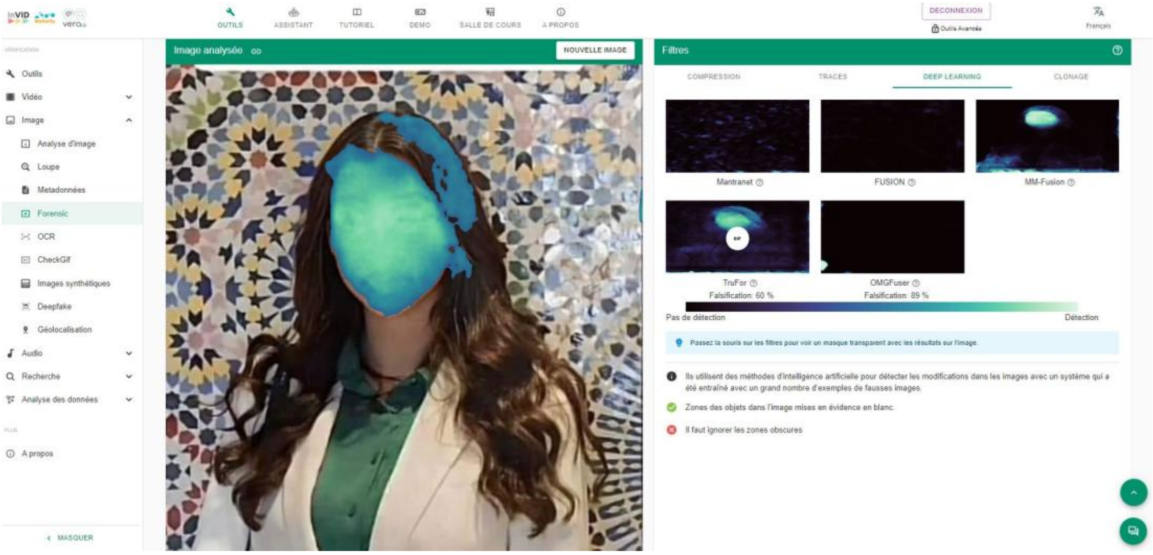


Mouse over the filters to see a transparent mask with the results on the image.

*Figure 73 View of a forgery detection by vera.ai forensic methods*

As with synthetic imagery, we again confront a similar dilemma: models capable of identifying previously undetected manipulations sometimes produce false positives that may mislead end-users. Investigating why certain authentic images are erroneously flagged as manipulated remains a critical area of research to improve the usability of these tools and facilitate their deployment within a fact-checking environment.

We still managed to use those new filters in fact-checking production as complementary evidence to debunk a manipulation, as in the case (Figure 74) of a viral fake picture of the daughter of the King of Morocco[50].



Passez la souris sur les filtres pour voir un masque transparent avec les résultats sur l'image.

Ils utilisent des méthodes d'intelligence artificielle pour détecter les modifications dans les images avec un système qui a été entraîné avec un grand nombre d'exemples de fausses images.

Zones des objets dans l'image mises en évidence en blanc.

Il faut ignorer les zones obscures

*Figure 74 Evidence of forgery detected by vera AI methods published in AFP Factuel*

[50] https://factuel.afp.com/doc.afp.com.34ZK2PL

Post-project plans will focus on investigating further false positives and especially pictures collected in HEIF format.

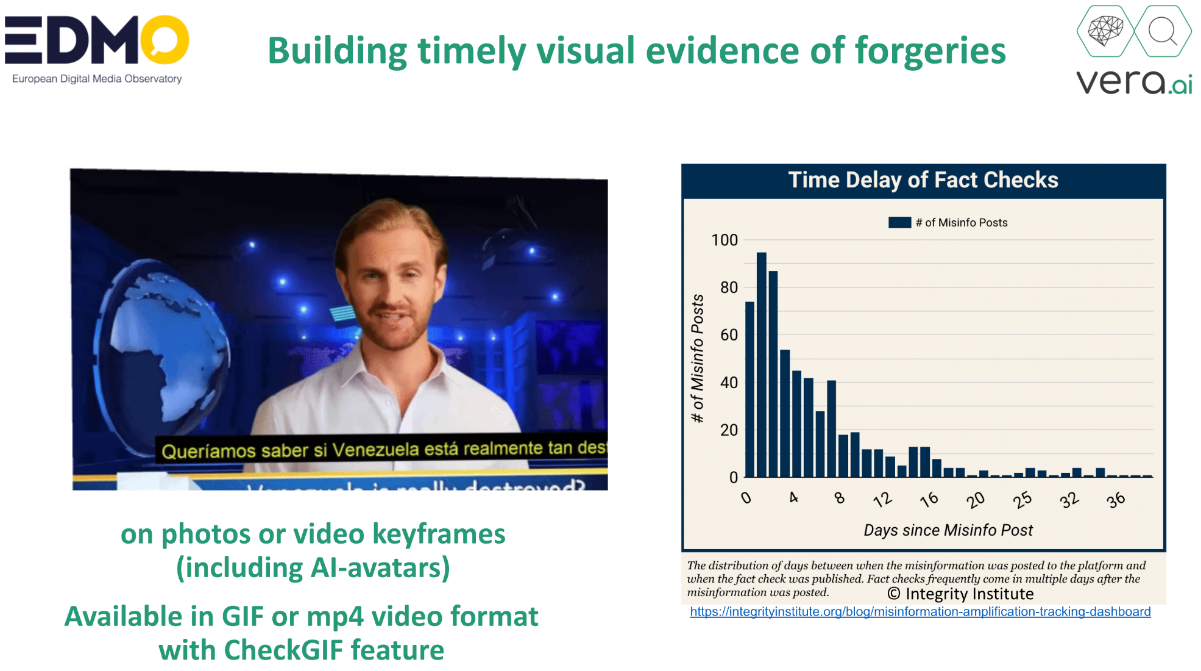
## 6.5 CheckGIF

The CheckGIF tool, a feature within the Verification Plugin, creates visual proof of image forgeries using a homography-based algorithm. This tool was developed in a previous project through a collaboration between AFP and ENS. The enhancements described here were evaluated with fact-checkers in cycle 3 using the design thinking methodology.

Our preceding deliverable, D2.1, highlighted a critical challenge: the time gap between the appearance of misinformation and its debunking by fact-checkers. It cited a study on the 2020 U.S. elections by the Integrity Institute, which stated that "fact-checkers need a suite of tools... to help increase their efficiency if they are able to keep up with the speed of misinformation on platforms."

Following discussions with fact-checkers in the AFP newsroom, we added the option to export CheckGIF results as MP4 videos. This enhancement was designed to overcome the structural limitations of legacy content management systems (CMS) that do not support GIFs and to facilitate the tool's adoption by television networks and other mainstream media.

We first showcased the concept of "building timely visual evidence of forgeries" with CheckGIF at the EDMO conference on May 25, 2023, demonstrating its application on both images and video keyframes (as shown in Figure 75).



EDMO
European Digital Media Observatory
vera.ai
Queríamos saber si Venezuela está realmente tan des
Venezuela is really destroyed?
Time Delay of Fact Checks
# of Misinfo Posts
100
80
60
40
20
0
# of Misinfo Posts
0
4
8
12
16
20
25
32
36
Days since Misinfo Post

Building timely visual evidence of forgeries

on photos or video keyframes (including AI-avatars)

Available in GIF or mp4 video format with CheckGIF feature

*The distribution of days between when the misinformation was posted to the platform and when the fact check was published. Fact checks frequently come in multiple days after the misinformation was posted.* © Integrity Institute

https://integrityinstitute.org/blog/misinformation-amplification-tracking-dashboard

*Figure 75 Copy of the CheckGIF slides shown at the EDMO conference 2023*

Subsequently, at Global Fact 10 in Seoul, renowned fact-checker Craig Silverman featured CheckGIF in his masterclass on digital investigation tools, specifically using it on video keyframes[51]. In a breakout session with 50 fact-checkers that same day, we presented several prototypes for an annotation layer, proposing the resulting animation as an immediate, shareable proof of forgery to curb disinformation on social media (Figure 76). The feedback was overwhelmingly positive, with attendees recommending a simple indicator to label the fake image and, if possible, coloured outlines to highlight the manipulated areas.

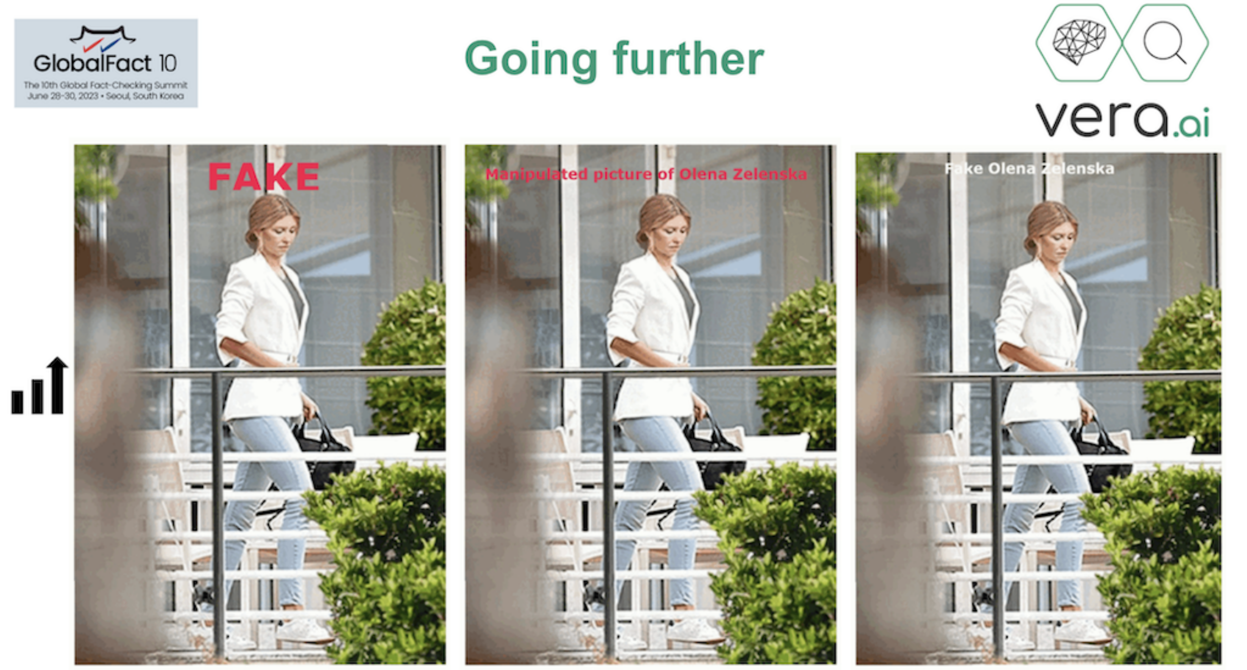
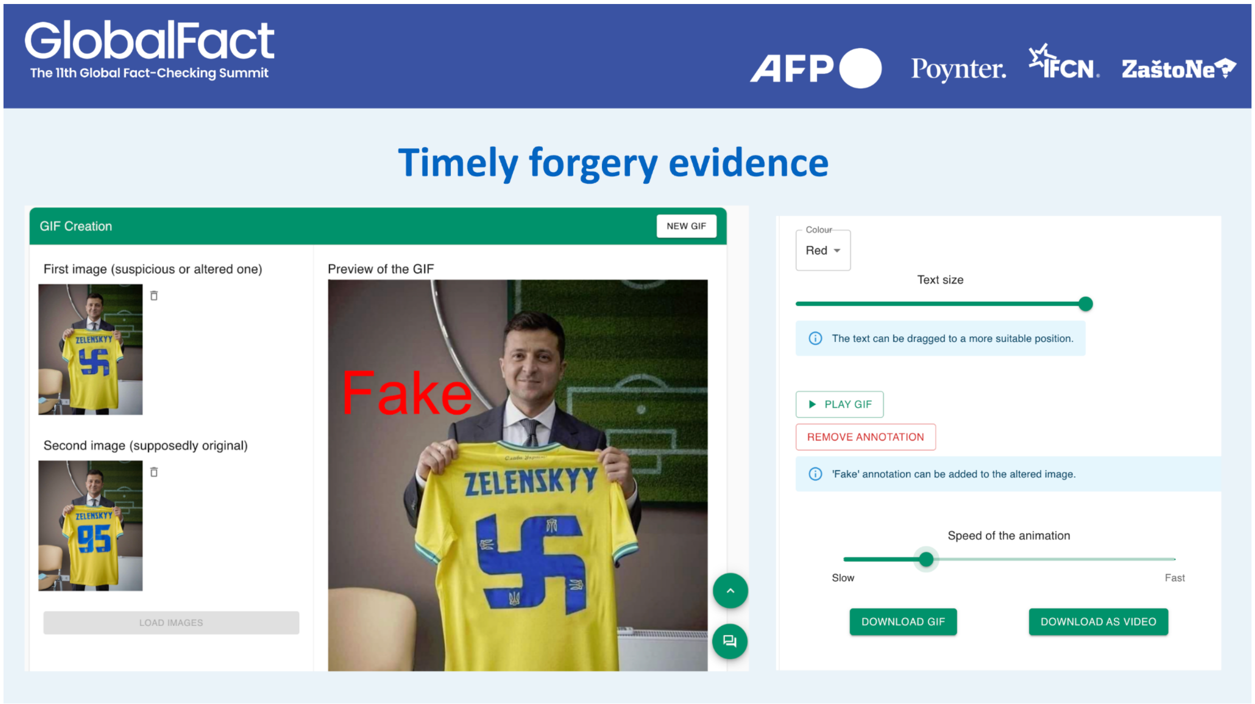

GlobalFact 10
The 10th Global Fact-Checking Summit
June 28-30, 2023 • Seoul, South Korea
Going further
vera.ai
FAKE
Manipulated picture of Olena Zelenska
Fake Olena Zelenska

*Figure 76 Slide of the vera.ai project presentation at Global Fact 10 showing CheckGIF annotations*

Six months later, in early December, we presented these prototypes again at the Google Trusted Media APAC Summit 2023 in Singapore, gathering feedback from the 60 Asian fact-checkers in attendance.

The enhanced feature, which includes the annotation layer, was finally released in the Verification Plugin v0.80 in early June 2024 after a final internal evaluation at AFP. This launch was strategically timed three weeks before our workshop at Global Fact 11 in Sarajevo (Figure 77), where it was also well received by the 70 fact-checkers present.

[51] https://x.com/dteyssou/status/1674574372034711553



GlobalFact
The 11th Global Fact-Checking Summit

## Timely forgery evidence

GIF Creation

First image (suspicious or altered one)

Second image (supposedly original)

Preview of the GIF

Colour: Red

The text can be dragged to a more suitable position.

'Fake' annotation can be added to the altered image.

*Figure 77 Copy of a slide presenting the CheckGIF feature with an annotation layer*

In mid-July 2024, AFP published a visual proof of forgery using a CheckGIF video on a picture mocking the assassination attempt of the then candidate to the U.S. presidency Donald Trump, in French, Brazilian and Greek fact-checking articles[52] and on X[53] network.

In our final survey, CheckGIF scored 3.51 out of 5 on a Likert scale and one respondent said that it "enables more intuitive image comparison".

As the vera.ai project concludes, we have two primary objectives for CheckGIF:

- Resolve geometric inconsistencies: We need to fix an issue reported by AFP fact-checkers in Latin America where white strips can appear around the GIF when the source images have significantly different dimensions.
- Create a video tutorial: In collaboration with digitalcourses.afp.com, we plan to produce a tutorial to further promote the tool's adoption. We've observed that while the MP4 format is more compatible than GIF, fact-checkers often lack direct control over their website settings to enable video looping, which is essential for simulating an animated GIF. The tutorial will aim to address such practical challenges.

[52] https://factuel.afp.com/doc.afp.com.36479ZG, https://checamos.afp.com/doc.afp.com.364B9JK , https://factcheckgreek.afp.com/doc.afp.com.364B4EJ

[53] https://x.com/AfpFactuel/status/1812868402207154665

## 6.6 Verification Plugin Assistant Credibility Signals Detection

The Verification Plugin Assistant is a suite of tools designed to help users determine which analysis tools are appropriate for their submitted content. By connecting to external services like the DBKF, it also provides alerts and warnings regarding the website or content under review.

The Assistant was first integrated into the Verification Plugin during the WeVerify project. The primary goal within vera.ai was to enhance this feature by adding a Credibility Signals Detection Service. This service analyses various attributes of the content, including its:

- Topic,
- Genre (factual or opinionated),
- Persuasion techniques (though linguistic categories),
- Subjectivity (whether the text is objective or subjective tone,
- A machine generated text detector (aiming to detect if the text is AI-generated).

This comprehensive analysis can be initiated in two ways: directly from the plugin’s launcher menu with the option “Assistant for current page” option, or through the Assistant interface by providing a link or a local file (such as a photo or a video).

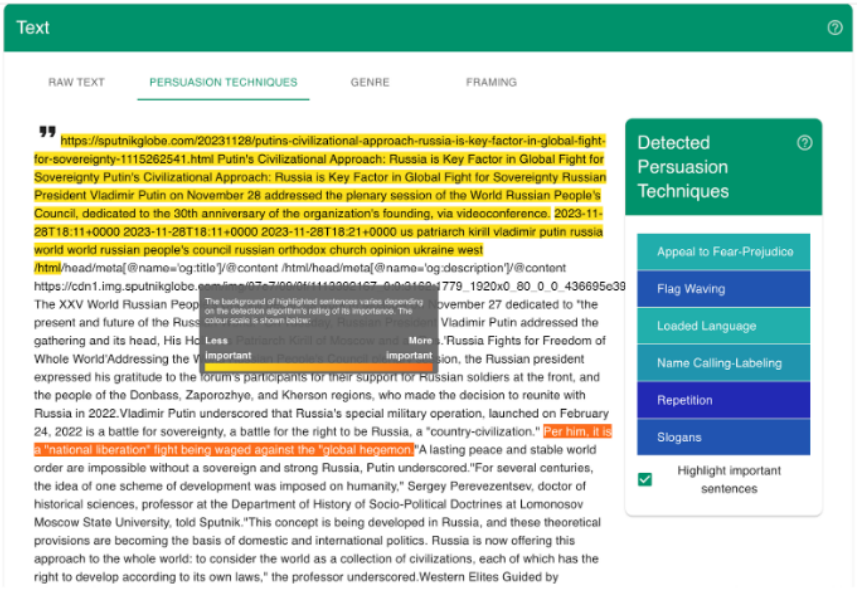
### 6.6.1 Design Thinking Evaluation

The initial evaluation of the Assistant's new features took place during cycle 1. The AFP team identified several key issues to be addressed:

- Issues with text extraction: The extracted text was sometimes messy, including raw code snippets (e.g., HTML, JSON) alongside the actual content.
- Issues with named entities: Named entities, labelled as "Test Topics," were not clickable and often had concatenation errors (missing spaces between words). We suggested linking these entities to a knowledge base like DBpedia to add functionality.
- Inconsistent technique detection: The detected persuasion techniques were not linked to the specific text segments where they appeared. The analysis frequently flagged very short phrases while overlooking other relevant sentences (Figure 78 below).
- Difficulty in genre classification: The system struggled to accurately determine the genre of a news article, especially when the content mixed reporting on current events with background information, such as quotes and statements as shown in Figure 79 below.

## Evaluation of assistant new features

Only two sentences highlighted. No way to find out which persuasion technique is detected in one particular sentence (no link between the right menu and the text). Other problematic sentences are not detected.



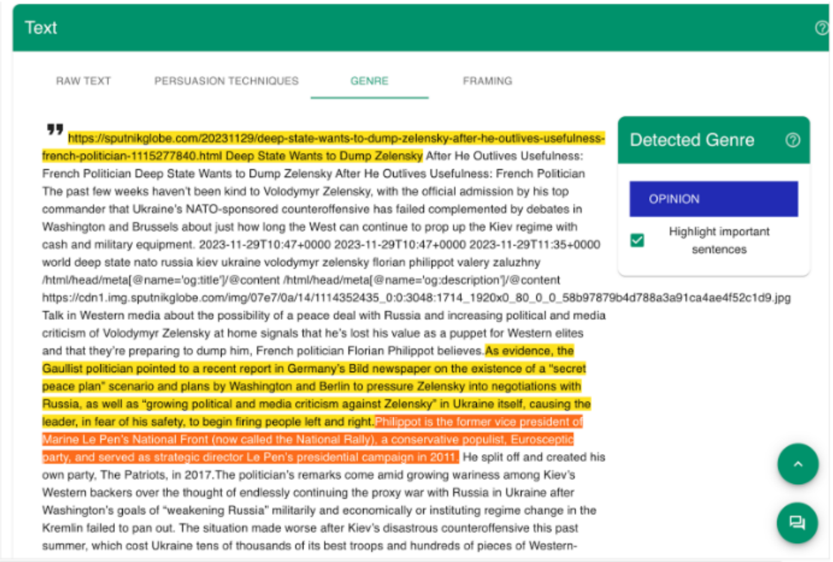
”https://sputnikglobe.com/20231128/putins-civilizational-approach-russia-is-key-factor-in-global-fight-for-sovereignty-1115262541.html Putin's Civilizational Approach: Russia is Key Factor in Global Fight for Sovereignty Putin's Civilizational Approach: Russia is Key Factor in Global Fight for Sovereignty Russian President Vladimir Putin on November 28 addressed the plenary session of the World Russian People's Council, dedicated to the 30th anniversary of the organization's founding, via videoconference. 2023-11-28T18:11+0000 2023-11-28T18:11+0000 2023-11-28T18:21+0000 us patriarch kirill vladimir putin russia world world russian people's council russian orthodox church opinion ukraine west /html/head/meta[@name='og:title']/@content /html/head/meta[@name='og:description']/@content https://cdn1.img.sputnikglobe.[illegible]1779_1920x0_80_0_0_436695e39 The XXV World Russian Peop[illegible] November 27 dedicated to "the present and future of the Russ[illegible] Tuesday, Russian President Vladimir Putin addressed the gathering and its head, His Ho[illegible] Patriarch Kirill of Moscow and a[illegible].'Russia Fights for Freedom of Whole World'Addressing the W[illegible] People's Council [illegible]sion, the Russian president expressed his gratitude to the forum's participants for their support for Russian soldiers at the front, and the people of the Donbass, Zaporozhye, and Kherson regions, who made the decision to reunite with Russia in 2022.Vladimir Putin underscored that Russia's special military operation, launched on February 24, 2022 is a battle for sovereignty, a battle for the right to be Russia, a "country-civilization." Per him, it is a "national liberation" fight being waged against the "global hegemon."A lasting peace and stable world order are impossible without a sovereign and strong Russia, Putin underscored."For several centuries, the idea of one scheme of development was imposed on humanity," Sergey Perevezentsev, doctor of historical sciences, professor at the Department of History of Socio-Political Doctrines at Lomonosov Moscow State University, told Sputnik."This concept is being developed in Russia, and these theoretical provisions are becoming the basis of domestic and international politics. Russia is now offering this approach to the whole world: to consider the world as a collection of civilizations, each of which has the right to develop according to its own laws," the professor underscored.Western Elites Guided by

The background of highlighted sentences varies depending on the detection algorithm's rating of its importance. The colour scale is shown below:

Less important — More important

*Figure 78 Evaluator-annotated screenshot outlining issues in persuasion techniques detection*

## Evaluation of assistant new features



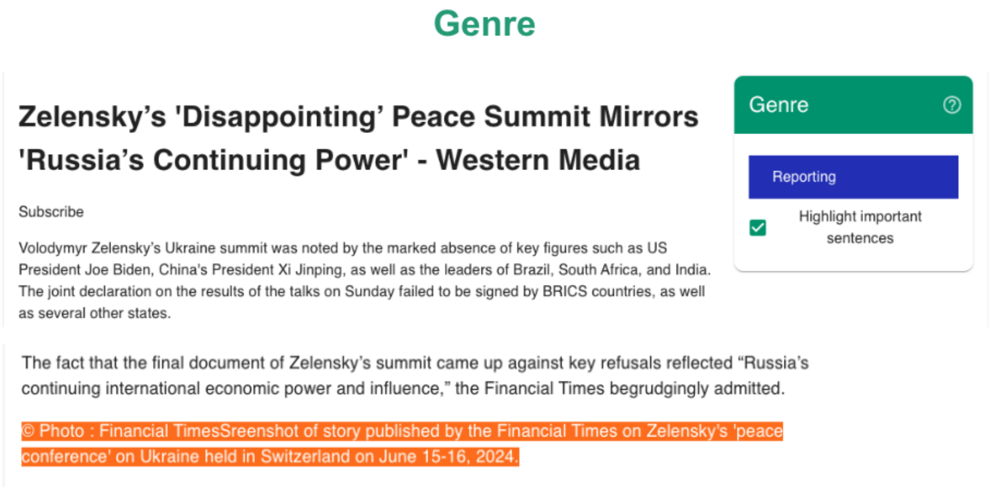
”https://sputnikglobe.com/20231129/deep-state-wants-to-dump-zelensky-after-he-outlives-usefulness-french-politician-1115277840.html Deep State Wants to Dump Zelensky After He Outlives Usefulness: French Politician Deep State Wants to Dump Zelensky After He Outlives Usefulness: French Politician The past few weeks haven't been kind to Volodymyr Zelensky, with the official admission by his top commander that Ukraine's NATO-sponsored counteroffensive has failed complemented by debates in Washington and Brussels about just how long the West can continue to prop up the Kiev regime with cash and military equipment. 2023-11-29T10:47+0000 2023-11-29T10:47+0000 2023-11-29T11:35+0000 world deep state nato russia kiev ukraine volodymyr zelensky florian philippot valery zaluzhny /html/head/meta[@name='og:title']/@content /html/head/meta[@name='og:description']/@content https://cdn1.img.sputnikglobe.com/img/07e7/0a/14/1114352435_0:0:3048:1714_1920x0_80_0_0_58b97879b4d788a3a91ca4ae4f52c1d9.jpg Talk in Western media about the possibility of a peace deal with Russia and increasing political and media criticism of Volodymyr Zelensky at home signals that he's lost his value as a puppet for Western elites and that they're preparing to dump him, French politician Florian Philippot believes. As evidence, the Gaullist politician pointed to a recent report in Germany's Bild newspaper on the existence of a "secret peace plan" scenario and plans by Washington and Berlin to pressure Zelensky into negotiations with Russia, as well as "growing political and media criticism against Zelensky" in Ukraine itself, causing the leader, in fear of his safety, to begin firing people left and right. Philippot is the former vice president of Marine Le Pen's National Front (now called the National Rally), a conservative populist, Eurosceptic party, and served as strategic director Le Pen's presidential campaign in 2011. He split off and created his own party, The Patriots, in 2017.The politician's remarks come amid growing wariness among Kiev's Western backers over the thought of endlessly continuing the proxy war with Russia in Ukraine after Washington's goals of "weakening Russia" militarily and economically or instituting regime change in the Kremlin failed to pan out. The situation made worse after Kiev's disastrous counteroffensive this past summer, which cost Ukraine tens of thousands of its best troops and hundreds of pieces of Western-

It seems that this article is rated as opinion based mostly on a background factual paragraph on an alt-right politician while the above sentence is more problematic. Would be interesting to have parts of factual content in one color and opinion in another.

*Figure 79 Evaluator-annotated screenshot revealing the difficulty to determine the genre of an article*

More usability issues were reported in cycle 3:

- Topic detection and genre can be inaccurate as it is sometimes based on very short text fragments like photo captions, as illustrated by an evaluator screenshot below (Figure 80)
- Categorizing truncated quotes from third party publications remains challenging (Figure 81).
- The system often misidentifies the common journalistic practice of repeating a headline as the lead text, flagging this simple repetition as a persuasion technique.



Zelensky's 'Disappointing' Peace Summit Mirrors 'Russia's Continuing Power' - Western Media

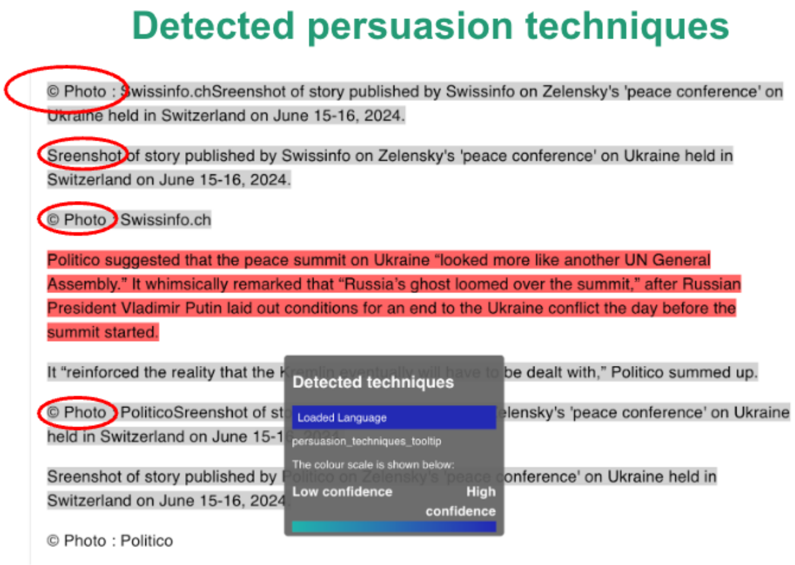
Subscribe

Volodymyr Zelensky's Ukraine summit was noted by the marked absence of key figures such as US President Joe Biden, China's President Xi Jinping, as well as the leaders of Brazil, South Africa, and India. The joint declaration on the results of the talks on Sunday failed to be signed by BRICS countries, as well as several other states.

The fact that the final document of Zelensky's summit came up against key refusals reflected “Russia's continuing international economic power and influence,” the Financial Times begrudgingly admitted.

© Photo : Financial TimesSreenshot of story published by the Financial Times on Zelensky's 'peace conference' on Ukraine held in Switzerland on June 15-16, 2024.

*Figure 80 Screenshot of the genre detection of a misleading article of Russian publication SputnikGlobe*


Detected persuasion techniques
Detected techniques
Loaded Language
persuasion_techniques_tooltip
The colour scale is shown below:
Low confidence
High confidence

© Photo : Swissinfo.chSreenshot of story published by Swissinfo on Zelensky's 'peace conference' on Ukraine held in Switzerland on June 15-16, 2024.

Sreenshot of story published by Swissinfo on Zelensky's 'peace conference' on Ukraine held in Switzerland on June 15-16, 2024.

© Photo : Swissinfo.ch

Politico suggested that the peace summit on Ukraine “looked more like another UN General Assembly.” It whimsically remarked that “Russia's ghost loomed over the summit,” after Russian President Vladimir Putin laid out conditions for an end to the Ukraine conflict the day before the summit started.

It “reinforced the reality that the Kremlin eventually will have to be dealt with,” Politico summed up.

© Photo : PoliticoSreenshot of st elensky's 'peace conference' on Ukraine held in Switzerland on June 15-1

Sreenshot of story published by Politico on Zelensky's 'peace conference' on Ukraine held in Switzerland on June 15-16, 2024

© Photo : Politico

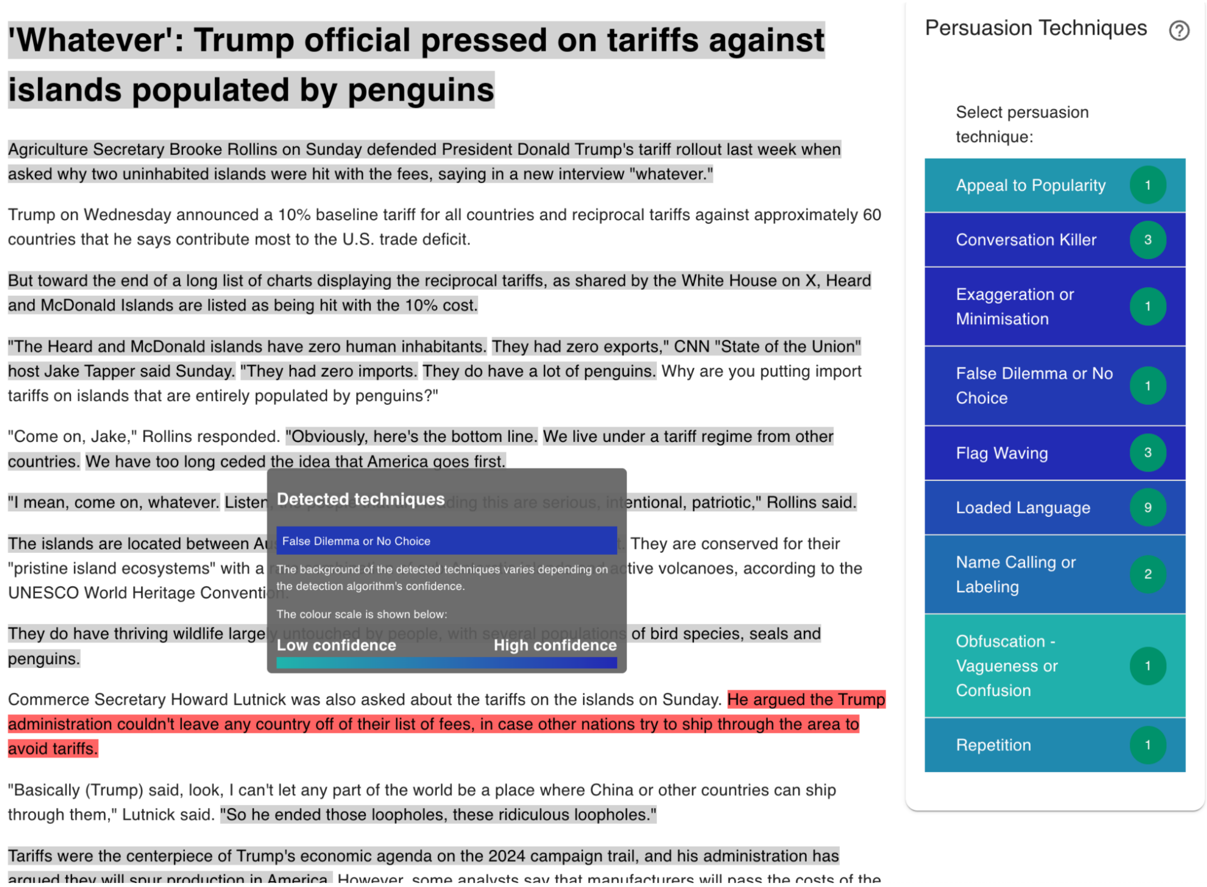
*Figure 81 Annotated screenshot with a truncated quote from SputnikGlobe*

As an evaluation experiment, we then analysed the same SputnikGlobe article[54] through several LLMs.

We prompted Gemini Advanced (Google), ChatGPT-4o (OpenAI) and Claude (Anthropic) asking on the same above article "Is the following text a factual report or a propaganda news article?".

While Gemini responded "I'm still learning how to answer this question, try Google Search", the two others clearly stated that this piece of text appears to be a propaganda news article rather than a factual report, identifying globally biased language, selective and cherry-picked quotations, one-sided perspective, lack of original repetition of negative points, Russian perspective, Framing and emotive screenshots.

In cycle 4, we made a key usability recommendation to address the confusing way the system displays confidence scores for detected persuasion techniques. The existing design used a multiple scale, where an initial indicator showed the detection of a technique, while a separate colour scale showed the confidence level as shown in the following screenshot (Figure 82).



'Whatever': Trump official pressed on tariffs against islands populated by penguins

Agriculture Secretary Brooke Rollins on Sunday defended President Donald Trump's tariff rollout last week when asked why two uninhabited islands were hit with the fees, saying in a new interview "whatever."

Trump on Wednesday announced a 10% baseline tariff for all countries and reciprocal tariffs against approximately 60 countries that he says contribute most to the U.S. trade deficit.

But toward the end of a long list of charts displaying the reciprocal tariffs, as shared by the White House on X, Heard and McDonald Islands are listed as being hit with the 10% cost.

"The Heard and McDonald islands have zero human inhabitants. They had zero exports," CNN "State of the Union" host Jake Tapper said Sunday. "They had zero imports. They do have a lot of penguins. Why are you putting import tariffs on islands that are entirely populated by penguins?"

"Come on, Jake," Rollins responded. "Obviously, here's the bottom line. We live under a tariff regime from other countries. We have too long ceded the idea that America goes first.

"I mean, come on, whatever. Listen ding this are serious, intentional, patriotic," Rollins said.

The islands are located between Au They are conserved for their "pristine island ecosystems" with a tive volcanoes, according to the UNESCO World Heritage Convention.

They do have thriving wildlife largely untouched by people, with several populations of bird species, seals and penguins.

Detected techniques

False Dilemma or No Choice

The background of the detected techniques varies depending on the detection algorithm's confidence.

The colour scale is shown below:

Low confidence High confidence

Commerce Secretary Howard Lutnick was also asked about the tariffs on the islands on Sunday. He argued the Trump administration couldn't leave any country off of their list of fees, in case other nations try to ship through the area to avoid tariffs.

"Basically (Trump) said, look, I can't let any part of the world be a place where China or other countries can ship through them," Lutnick said. "So he ended those loopholes, these ridiculous loopholes."

Tariffs were the centerpiece of Trump's economic agenda on the 2024 campaign trail, and his administration has argued they will spur production in America. However, some analysts say that manufacturers will pass the costs of the

Persuasion Techniques ?

Select persuasion technique:

| | |
|---|---|
| Appeal to Popularity | 1 |
| Conversation Killer | 3 |
| Exaggeration or Minimisation | 1 |
| False Dilemma or No Choice | 1 |
| Flag Waving | 3 |
| Loaded Language | 9 |
| Name Calling or Labeling | 2 |
| Obfuscation - Vagueness or Confusion | 1 |
| Repetition | 1 |

*Figure 82 Screenshot of the multiple scale to present a persuasion technique detection*

[54] https://web.archive.org/web/20250103215335/https://sputnikglobe.com/20240617/zelenskys-disappointing-peace-summit-mirrors-russias-continuing-power---western-media-1118994797.html

This approach increases the user's cognitive load and complicates interpretation, as the initial detection is finally qualified by the secondary colour indicator. To simplify this, we recommended merging the two indicators into a single, unified scale, such as a common slider that could clearly display the confidence score for each detection with a certain level of confidence.

Regarding colour schemes, we recommended adopting the traffic light veracity labels and gauges already used in synthetic image detection and deepfake detection within the Verification Plugin. This approach would ensure a consistent user experience across the different tools.

Furthermore, a test of the Machine Generated Text Service with a senior AFP fact-checking trainer in Hong-Kong proved successful. On a fact-check from the Philippines[55], we were asked to check if the content of this archive post initially published on Facebook[56] and already debunked because of the GenAI image, could be machine generated. The vera.ai detector gave a score of 76% AI-generated for this text, in line with the results of a dozen other online detectors.

To work around the stability issue we faced in the integrated Machine Generated Text Detection service in cycle 5, we added a standalone instance to the plugin. This separate tool allowed us to proceed with a more flexible evaluation using any text input, independent of the Assistant.

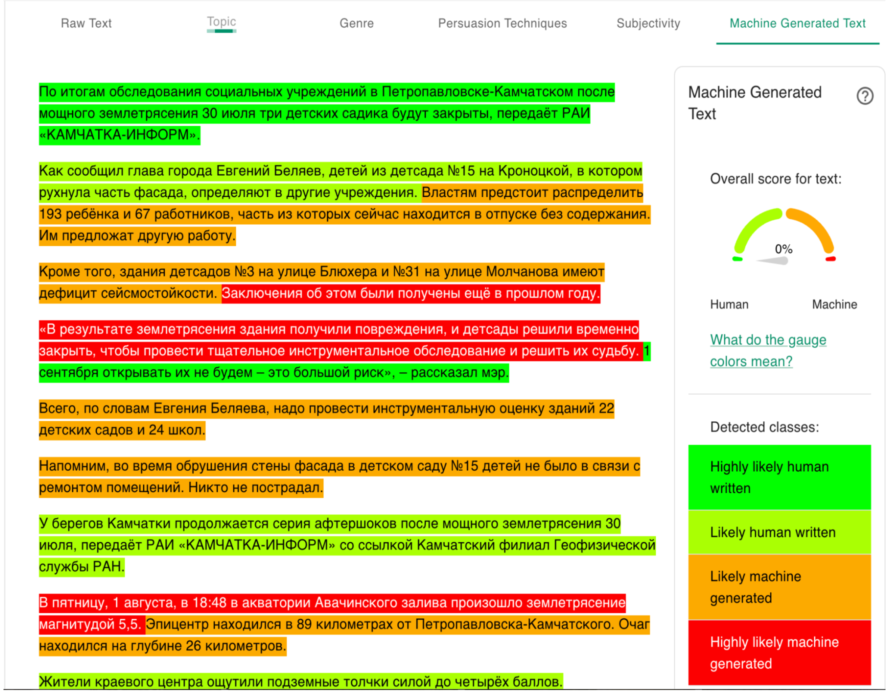
In cycle 6, we discovered a significant usability issue with the service's output. The tool provides two different analyses: a visible, more accurate overall score for the entire text and a hidden, sentence-by-sentence score used to highlight potentially generated phrases. This created a contradiction in a use case submitted by an AFP fact-checker who was trying to determine if a Russian website was real media or not. The service returned a global score of 0% (human-written), yet several sentences were coloured as problematic based on an undisplayed sentence-level score of 55.46%. As exemplified in Figure 82 below, this large gap between the two detection scores is quite confusing for the end-user.

In our final survey, the Assistant was praised for “making specific actions on a website easier” and received a global note of 3.56 on a Likert scale.

Post-project plans on credibility signals will focus on testing further LLMs reasoning and efficiency to analyse persuasion technique, subjectivity and genre. More tests are needed on machine-generated text to propose explanations without the confusion of a double scale between chunks and sentences (Figure 83).

[55] https://factcheck.afp.com/doc.afp.com.38MD3WD

[56] https://mvau.lt/media/883eebe8-bae2-4913-97fa-5b45bca4fca2



По итогам обследования социальных учреждений в Петропавловске-Камчатском после мощного землетрясения 30 июля три детских садика будут закрыты, передаёт РАИ «КАМЧАТКА-ИНФОРМ».

Как сообщил глава города Евгений Беляев, детей из детсада №15 на Кроноцкой, в котором рухнула часть фасада, определяют в другие учреждения. Властям предстоит распределить 193 ребёнка и 67 работников, часть из которых сейчас находится в отпуске без содержания. Им предложат другую работу.

Кроме того, здания детсадов №3 на улице Блюхера и №31 на улице Молчанова имеют дефицит сейсмостойкости. Заключения об этом были получены ещё в прошлом году.

«В результате землетрясения здания получили повреждения, и детсады решили временно закрыть, чтобы провести тщательное инструментальное обследование и решить их судьбу. 1 сентября открывать их не будем – это большой риск», – рассказал мэр.

Всего, по словам Евгения Беляева, надо провести инструментальную оценку зданий 22 детских садов и 24 школ.

Напомним, во время обрушения стены фасада в детском саду №15 детей не было в связи с ремонтом помещений. Никто не пострадал.

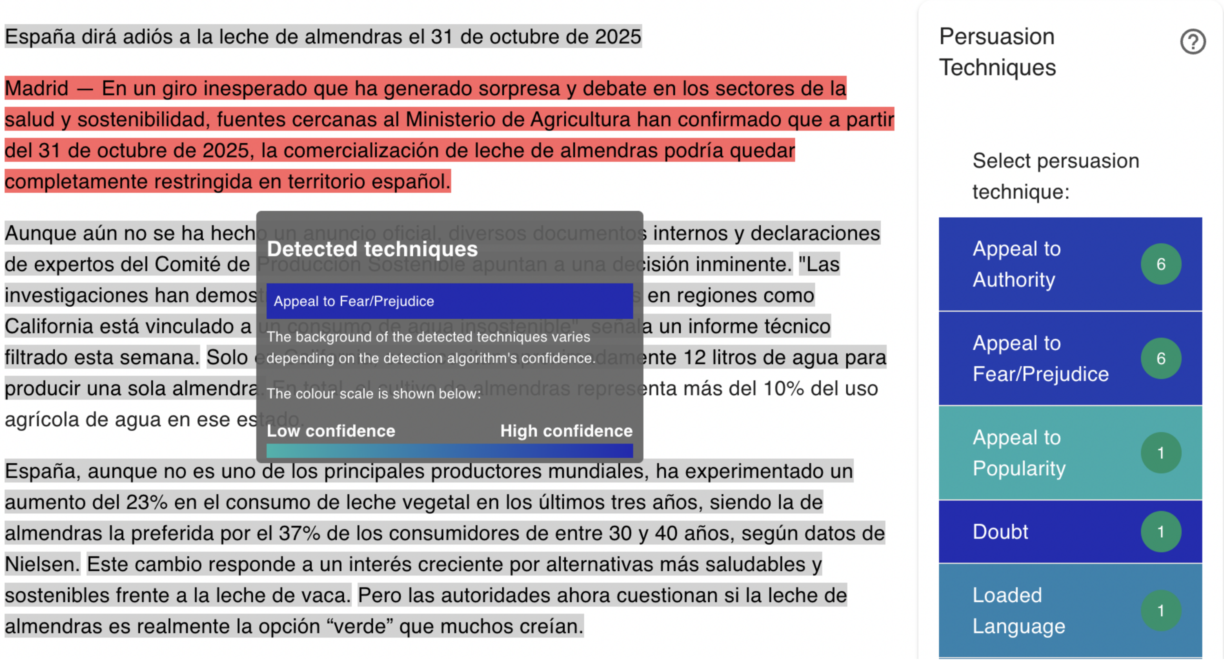
У берегов Камчатки продолжается серия афтершоков после мощного землетрясения 30 июля, передаёт РАИ «КАМЧАТКА-ИНФОРМ» со ссылкой Камчатский филиал Геофизической службы РАН.

В пятницу, 1 августа, в 18:48 в акватории Авачинского залива произошло землетрясение магнитудой 5,5. Эпицентр находился в 89 километрах от Петропавловска-Камчатского. Очаг находился на глубине 26 километров.

Жители краевого центра ощутили подземные толчки силой до четырёх баллов.

*Figure 83 Screenshot of a Machine Generated Text Detection showing a gap between the score and explainability*

## 6.6.2 Participatory Evaluation



España dirá adiós a la leche de almendras el 31 de octubre de 2025

Madrid — En un giro inesperado que ha generado sorpresa y debate en los sectores de la salud y sostenibilidad, fuentes cercanas al Ministerio de Agricultura han confirmado que a partir del 31 de octubre de 2025, la comercialización de leche de almendras podría quedar completamente restringida en territorio español.

Aunque aún no se ha hecho un anuncio oficial, diversos documentos internos y declaraciones de expertos del Comité de Producción Sostenible apuntan a una decisión inminente. "Las investigaciones han demos [illegible] en regiones como California está vinculado a un consumo de agua insostenible", señala un informe técnico filtrado esta semana. Solo [illegible]damente 12 litros de agua para producir una sola almendra. En total, el cultivo de almendras representa más del 10% del uso agrícola de agua en ese estado.

Detected techniques

Appeal to Fear/Prejudice

The background of the detected techniques varies depending on the detection algorithm's confidence.

The colour scale is shown below:

Low confidence High confidence

España, aunque no es uno de los principales productores mundiales, ha experimentado un aumento del 23% en el consumo de leche vegetal en los últimos tres años, siendo la de almendras la preferida por el 37% de los consumidores de entre 30 y 40 años, según datos de Nielsen. Este cambio responde a un interés creciente por alternativas más saludables y sostenibles frente a la leche de vaca. Pero las autoridades ahora cuestionan si la leche de almendras es realmente la opción “verde” que muchos creían.

*Figure 84 Persuasion techniques analysis of a fake news article in the Credibility Signals Assistant*

The participatory evaluation of the Assistant took place in June 2025 and focused on 3 different aspects of the analysis of a text's credibility: the persuasion techniques used in its writing, the likelihood of the text being machine-generated, and on authorship detection – these aspects being among the main indicators that the text might have been written with the intention to deceive (based on studies of how current disinformation texts are produced) (Srba et al., 2025).

The group of end-users we worked with consisted of 10 news media professionals and researchers from OSINT, France Télévisions, University of Siegen, RTVE, Atresmedia, Radio France, and Radio Romania. The users were provided with guidelines and with an exercise: the analysis (by themselves and with the service) of a blinded text in Spanish about a so-called "ban on almond milk" supposedly taking place in Spain in a few months. This text was AI-generated using ChatGPT and a prompt requesting the use of persuasion techniques and of credible sources. Users were also encouraged to try the service on their own texts and during the session they were presented with a demo of the authorship detection service.

The evaluation focused on user expectations, usability, the usefulness of the signals used in the analysis, the level of granularity in the presented results, whether and to what extent the tool changed or confirmed their initial analysis, whether users preferred to be provided with aggregated results vs. raw results as clues to help them decide themselves, and on discussions on possible use cases (in particular the authorship detection).

In terms of analysing the 'almond milk' text, the tool recognised a number of persuasion techniques that had indeed been used in the text (such as appeal to authority, loaded language, and appeal to fear/prejudice), and correctly determined that the text was machine-generated. Users on the other end were undecided about the origin of the text, while analysing the text themselves. They relied on their own knowledge of recent news and on other analogue analysis techniques: since they had never heard of a ban on almond milk, they were indeed suspicious. The fact that some of the sources were cited excessively (which the tool also flagged under the signal 'repetition') and that some seemed irrelevant also tipped them off, but without being able to know more about the source of the article and similarly relevant context, they were undecided about its authenticity. Once presented with the tool's results, they were able to refine their conclusions based on the highlighted persuasion techniques (Figure 84).

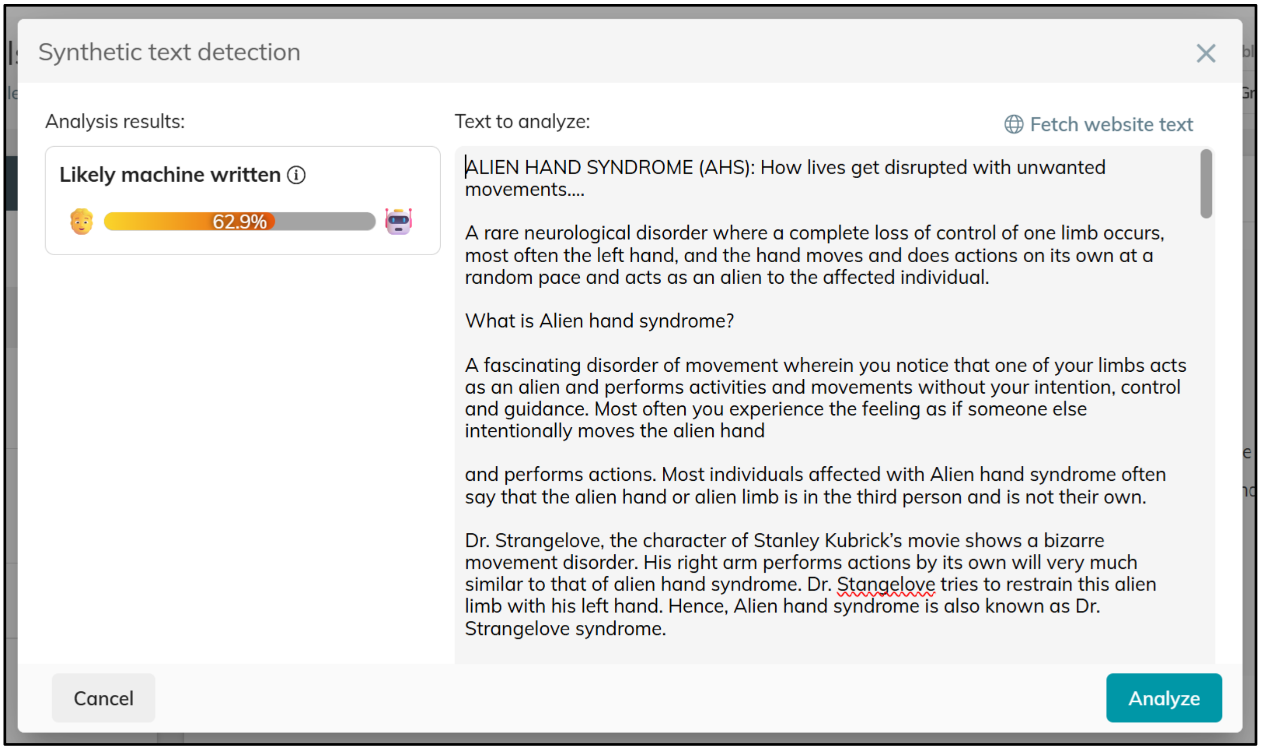
Key takeaways for the evaluation were the following:

- **Information overload**: Users were overwhelmed by the amount of information provided about the text and requested more balance: for instance an option to go from a broad overview (foregrounding core information) to a deeper analysis, i.e. from highlights of the most relevant persuasion techniques to more details about them and a greater number of signals.
- **Usability and suggested functionalities**: A more careful and hierarchical use of colour-coding should be used to highlight the most important pieces of information, but used sparingly, as highlighting too much is equivalent to highlighting nothing at all. The level of granularity also affected how fast they could use the tool. Multilingual translations need to be integrated into the tool itself rather than bringing the user outside of it to a different service. Finally, a suggestion was made to integrate the tool with a video transcript one to be able to also analyse videos besides text.

- **Aggregation vs. support through clues**: This need to reduce information overload should be balanced with a preference from the users to use the tool as a support to their own investigation through clues, rather than be provided with an aggregated summarised result. The tool should also provide guidance in how to interpret the clues.
- **Authenticity**: After using the service on their own organisation's news articles, some users discussed whether the use of persuasion techniques in a text can always be linked to an intention to deceive. Texts relaying important information for the public good or that report on facts that might be perceived as improbable, might for instance use a larger number of persuasion techniques than average due to the need to convince the reader of the facts' veracity. This aspect, which should be highlighted in the tool, is linked to the previous point about needing guidance in interpretation.
- **Importance of context**: The evaluation highlighted the users' reliance on contextual information, and thereby the need for contextual signal analysis as part of the service. The hierarchy of which signals to focus on vs. deprecate was however difficult to determine and a final decision was not reached.
- **Usefulness of MGT detection**: Although the detection of machine-generated text proved useful as a clue to determine the validity (or rather lack thereof) the example text, users pointed out that it was not a key signal in determining deceit, as sometimes MGT is used as a way to simplify a writing workflow and not necessarily as part of a disinformation attempt. The need for results granularity in the MGT detection was expressed, as the users were testing a version of the tool that was only providing a percentage of likelihood.
- **Authorship use cases**: The main use cases for this signal was to help determine the identity of an anonymous source and to cross-check the identity of an alleged author.

The results of this evaluation confirmed several principles of the design framework (see section 2.1.2), such as the need for such tools to support rather than replace the end-user's analysis processes, to focus on meaningful rather than exhaustive information, as well as the importance of speed of use, explainability, and intuitive information visualisation, and finally, the need for differentiated levels of use (from general to detailed).

MGT is also integrated in Truly Media (see Figure 85) and was subject to an expert testing with journalists and designers in August 2024. This led to insights about usability, transparency and trustworthiness. While the service was considered easy to use, participants lack use cases for MGT individually and consider the combination with other credibility signals as promising. Overall the provision of the service by trusted partners is perceived as trustworthy, still users wished for more information about the decision-making of the service.



ALIEN HAND SYNDROME (AHS): How lives get disrupted with unwanted movements....

A rare neurological disorder where a complete loss of control of one limb occurs, most often the left hand, and the hand moves and does actions on its own at a random pace and acts as an alien to the affected individual.

What is Alien hand syndrome?

A fascinating disorder of movement wherein you notice that one of your limbs acts as an alien and performs activities and movements without your intention, control and guidance. Most often you experience the feeling as if someone else intentionally moves the alien hand

and performs actions. Most individuals affected with Alien hand syndrome often say that the alien hand or alien limb is in the third person and is not their own.

Dr. Strangelove, the character of Stanley Kubrick's movie shows a bizarre movement disorder. His right arm performs actions by its own will very much similar to that of alien hand syndrome. Dr. Stangelove tries to restrain this alien limb with his left hand. Hence, Alien hand syndrome is also known as Dr. Strangelove syndrome.

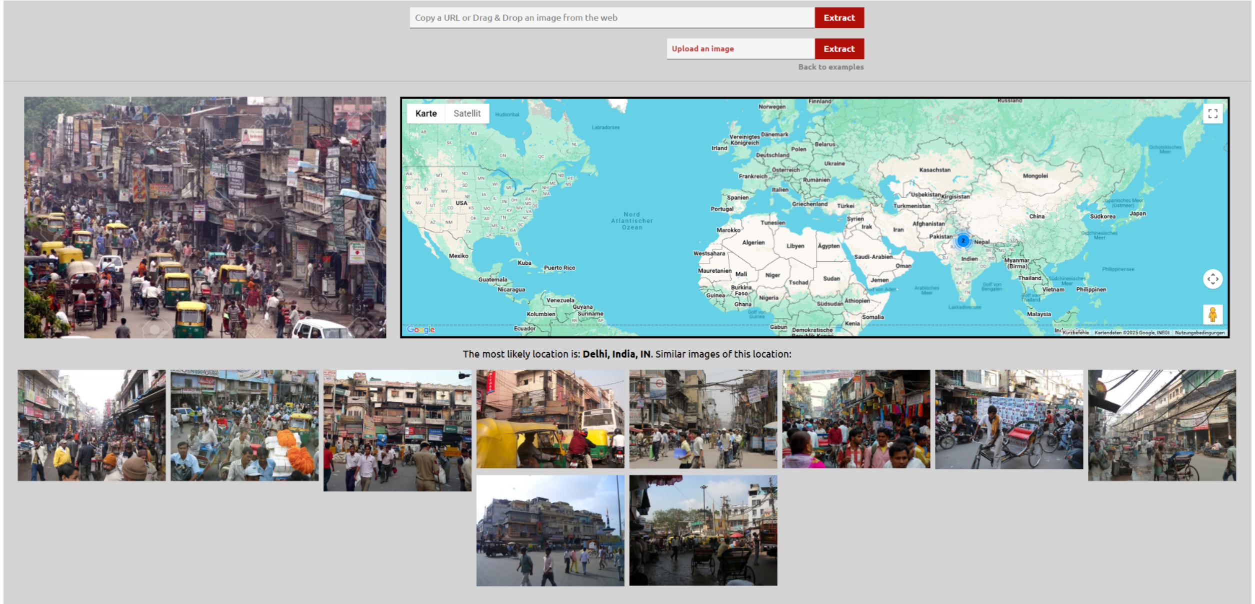
*Figure 85 MGT service integrated in Truly Media*

## 6.7 Location Extraction / Geolocalizer

The Location Extraction service (called Geolocalizer in the Verification Plugin) aims to infer the geographical location of a query image, using the visual content of the image.

### 6.7.1 Participatory Evaluation

In May 2025, the CERTH demonstrator in its standalone version (Figure 86) was subject of a participatory evaluation with seven expert testers from DW with extensive experience in tools (some even in their development) that support the geolocation tasks of journalists. They have backgrounds in innovation management, fact-checking, academic research and AI development. The goal of this evaluation was to explore the tools usefulness in its current state for journalism and focus on potential improvements.



The most likely location is: **Delhi, India, IN**. Similar images of this location:

*Figure 86 Location Extraction interface in standalone version*

Like for other evaluations, testers were provided with instructions and access to the service about a week prior to a remote exchange. This gave them enough time to familiarise themselves with the service and collect individual feedback. Within the remote exchange testers provided their feedback on a whiteboard, followed by a discussion. The synthesized results are described in the following paragraphs.

The general aim of the service was perceived as useful. This is realized due to aspects like the speed of analysis, easy understanding and use of the service. Displaying related images to a submitted image adds to the trustworthiness of estimated locations. This positive feedback was however provided under the condition that it is only then a useful tool, if the accuracy is right.

It is thus not surprising that several aspects that currently get in the way of the service's usefulness target trustworthiness. For example, the service is perceived as a black box in how it reaches a certain conclusion and that it is unclear what the percentage rate of an estimated location means. It additionally seems like the service is supporting the detection of an area instead of its intention to provide an exact location.

Testers had several ideas on how to improve the location estimation's usefulness across the areas of trustworthiness and UI/UX:

- Add **explanations** and further clues on how the service works, why a location was chosen and how to interpret the percentages and locations found.
- **Provide documentation and information** on training, e.g. allow for easier access to publications.
- **Search/insert data**: allow for more file formats, allow for users to fine grain the search in order to get better results, e.g. if a user knows the region of the image already.
- **Related images**: show sources and actual locations.

- **Display of results/map**: facilitate navigation on map as known from other maps, show map in satellite view by default, make streetview access easier and provide coordinates to estimated locations.
- **Result processing**: allow for users to correct the process afterwards to get more accurate results, enable further analyses, and provide export options.
- **Connection to other services**: follow up action needed to allow the user to continue its search, e.g. team up with services like SPOT[57].

As this service is also integrated in the Verification Plugin, the results were made available to the developing partner to improve the standalone version but also integration partner to improve UI/UX in the plugin.

This evaluation supports several takeaways from other sessions regarding the design of AI-based services. For example, the need for users to be able to easily cross-check results and the familiarity in result provision (e.g. in this case through satellite view and navigation of the map). Furthermore, it shows how important it is for users to seamlessly move from one tool to another to efficiently verify content.

### 6.7.2 Design Thinking Evaluation

The Geolocalizer tool, initially created in the WeVerify project[58] has been integrated into the Verification Plugin since the Plugin v0.77 in mid-October 2023 for evaluation. Due to its inconsistent previous performance, the tool was restricted to beta testers for benchmarking geolocation services. For this reason, the feature's description warned against using it in production until further notice (Figure 87).

The tool's underlying model was retrained by CERTH during vera.ai and the new updated service became available for evaluation in cycle 5 and cycle 6. To evaluate the Geolocaliser's performance, we built a dataset of smartphone pictures taken with geotagging. This dataset served as our ground truth, allowing us to estimate the deviation (through a Haversine distance[59]) between the model's predictions and the image-coded GPS coordinates.

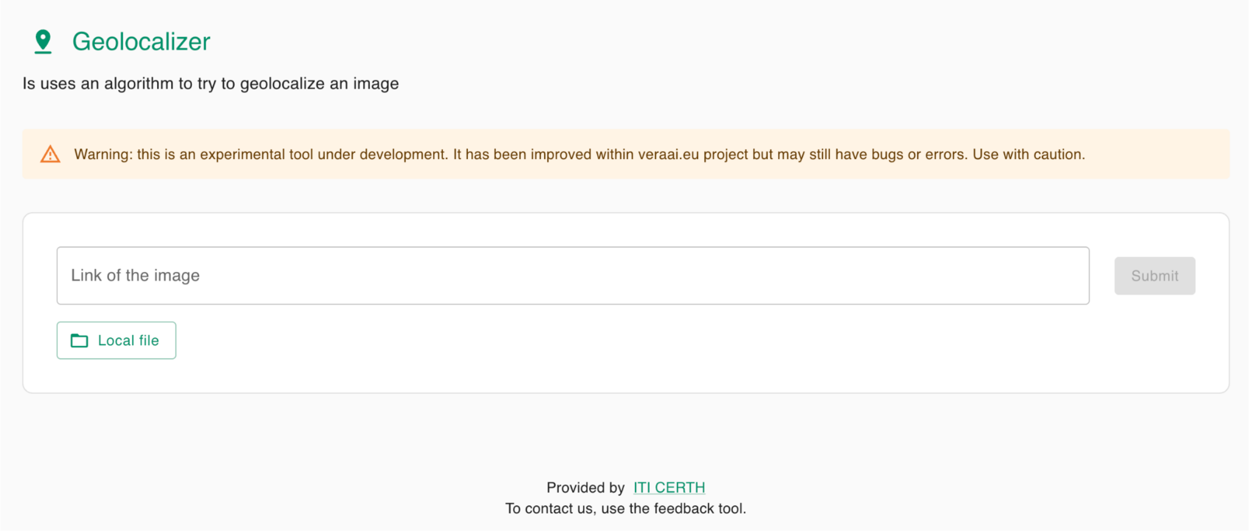
Our initial results from cycle 5's evaluation encouraged us to continue our work by building a geographically diverse bigger dataset of over 100 photos. We also enhanced the Geolocalizer interface by adding an upload option and updating the tool presentation (as shown in figure 85 below since plugin v0.85) as well as the predictions display.

The upload functionality was extended to many features after the plugin middleware's logs revealed that a substantial number of links were resulting in processing errors when the remote API was unable to fetch the content. Therefore, providing an upload feature allows end-users to bypass limitations imposed by websites and platforms – restrictions that have been significantly tightened since the rise of generative AI and strong concerns over the unauthorized use of online images.

[57] https://www.findthatspot.io/

[58] https://weverify.eu/#

[59] https://en.wikipedia.org/wiki/Haversine_formula



Geolocalizer

Is uses an algorithm to try to geolocalize an image

Warning: this is an experimental tool under development. It has been improved within veraai.eu project but may still have bugs or errors. Use with caution.

Provided by ITI CERTH
To contact us, use the feedback tool.

*Figure 87 Screenshot of the Geolocalizer feature in the Verification Plugin with an updated warning*

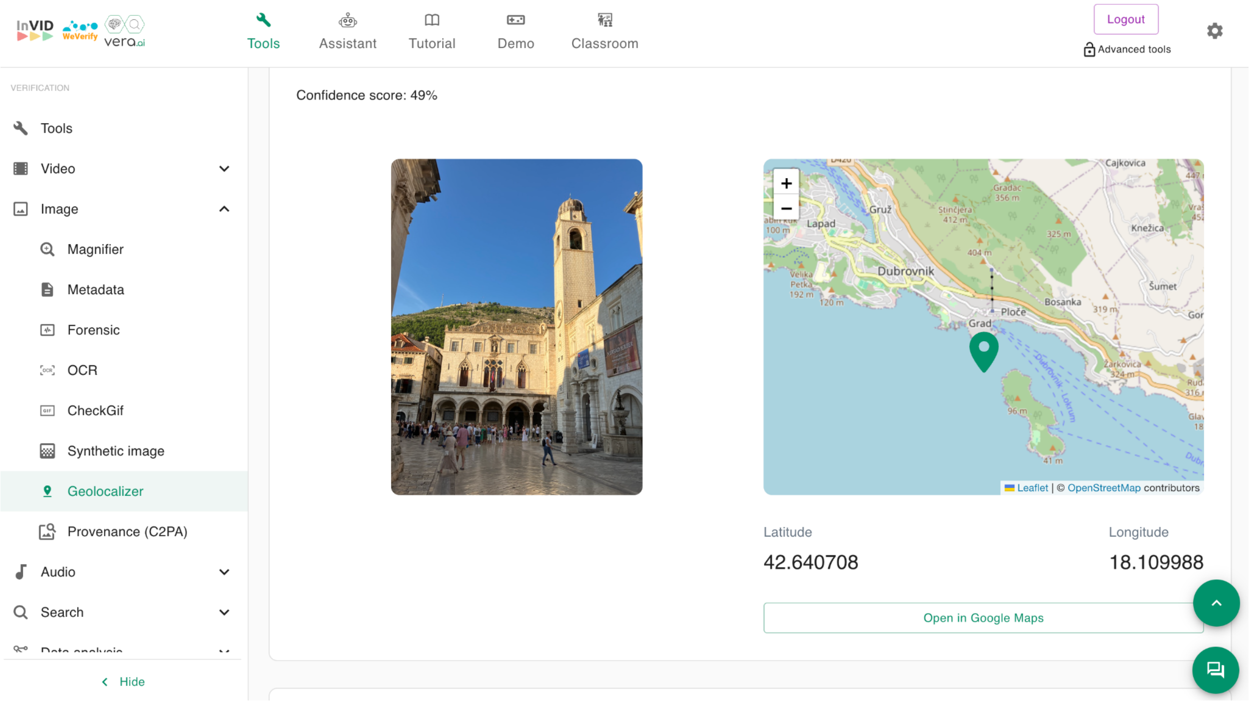
Our evaluation dataset of 105 images gave the following results on 202 predictions[60] in Table 9 below:

*Table 9 Overall predictions achieved for the ground truth dataset*

| **Geolocation performance** | **Metrics** |
|---|---|
| Average confidence score | 37% |
| Average Haversine distance | 565.69 km |
| Median Haversine distance | 12.01 km |
| Percentage of predictions < 5 km | 34.65% |
| Bigger Haversine distance (worst result) | 17691.64 km |
| Smallest Haversine distance (best result) | 0.045 km |
| Correlation coefficient (between confidence and Haversine distance) | -0.23 |

The best result (Figure 88) was obtained on a photo of the old city of Dubrovnik (Croatia) with a precision of 45 meters despite a confidence score of 49.57%, while the worst result was the fifth estimation (with a low confidence score of 10.64%) of a photo initially taken in Botafogo (Rio de Janeiro). Three out of five predictions put the same photo in Rio de Janeiro (one at only 1.38 km of distance from the GPS coordinates, with a confidence score of 18.01%), while two others put it in Hong Kong or near Macao.

[60] The model may propose several predictions for an image.



Confidence score: 49%

Latitude
42.640708

Longitude
18.109988

Open in Google Maps

*Figure 88 View of the best geolocalisation result obtained during evaluation*

To eliminate the bias of the model's multiple predictions, we duplicated our dataset and kept only the best prediction, which was not necessarily the highest confidence score as shown by the weak negative correlation coefficient between the confidence score and the Haversine distance, suggesting that the relationship between the two is not very strong or consistent.

Using the same dataset but keeping only the best prediction of each image, we got better results showcased in Table 10 below:

*Table 10 Best predictions achieved for each image of the ground truth dataset*

| **Geolocation Performance on first prediction** | **Metrics** |
|---|---|
| Average confidence score | 51% |
| Average Haversine distance | 266.41 km |
| Median Haversine distance | 3.88 km |
| Percentage of predictions < 5 km | 55.24% |
| Bigger Haversine distance (worst result) | 9567.38 km |
| Smallest Haversine distance (best result) | 0.045 km |
| Correlation coefficient (between confidence and Haversine distance) | -0.25 |

The above table 10 gives a better overview of the tool performance as well as the chart below (Figure 89).

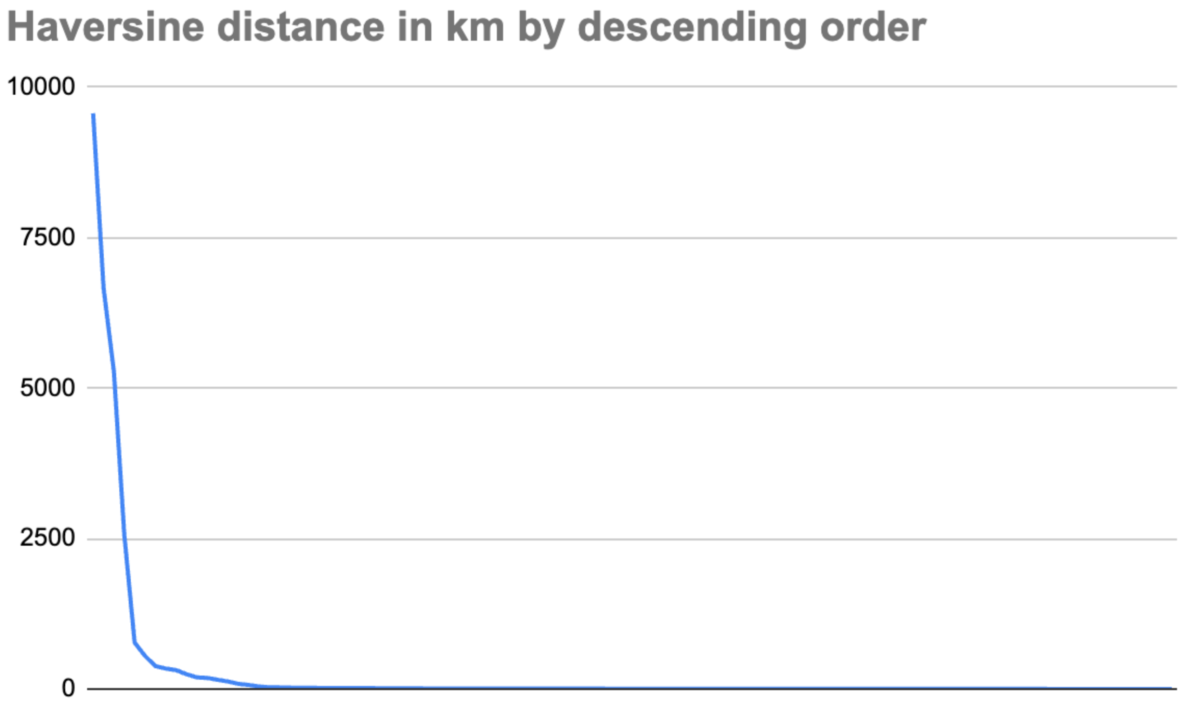


*Figure 89 Graph showing the rapid decrease of the erroneous predictions*

The above curb shows a rapid decrease of the erroneous predictions. In fact, 79% of the geolocation predictions of our dataset are within a precision of 16 km.

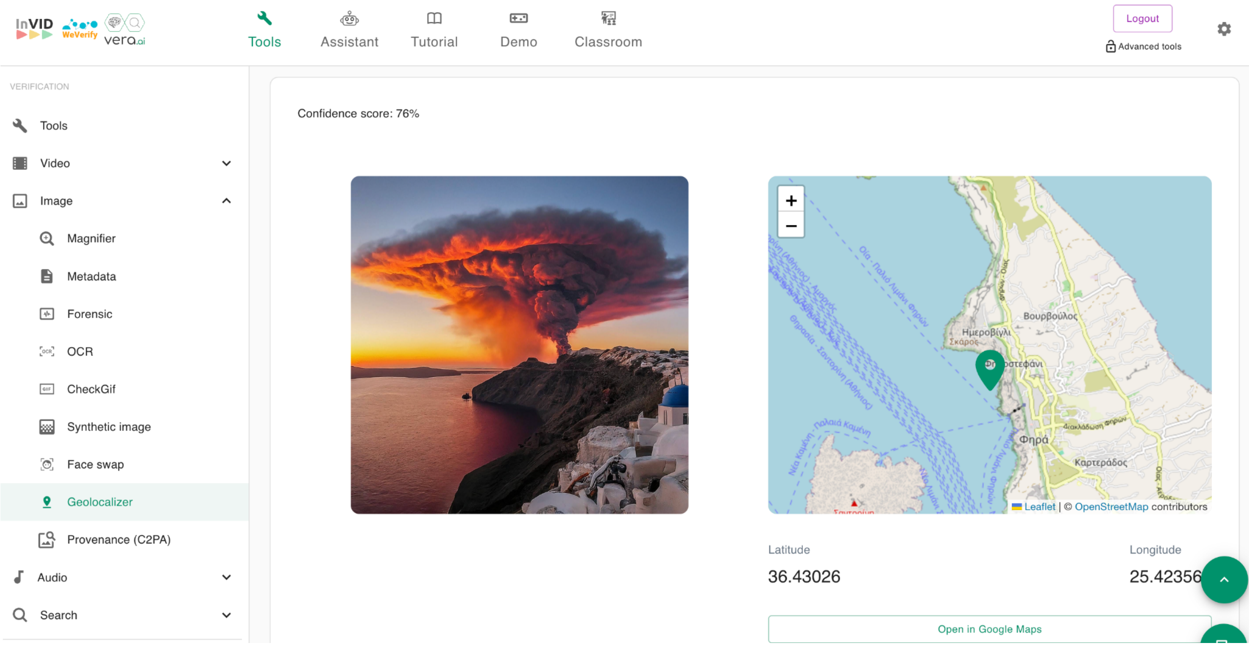
The 3 main errors or hallucinations of the model (all 3 predictions above 5,000 km with a confidence score below 20%) come from geographically diverse pictures (a view of Park Güell in Barcelona predicted in San Francisco, a view of the Salses Fortress (France) located in Jaipur, India, and a seaside landscape from Tamarindo Bay (Costa Rica) located in the sea off the Oregon (USA) coast). While the third one, a generic sea landscape, was difficult to spot, the two others are well-known sites with pages and photographs available on Wikipedia.

For post-project plans, relying more systematically on Wikimedia Commons to upgrade the training dataset could be an interesting approach to improve the current feature. Improving the correlation coefficient between the prediction and the confidence score would be particularly interesting, as it would allow us to filter results through a threshold in the user interface.

To conclude this section, we examine a compelling case from our cycle 5 testing phase that underscores the complexities surrounding generative AI's photorealism and its associated detection and geolocation challenges. In early 2025, an AI-generated image depicting a potential explosion on the Greek island of Santorini[61] went viral on social media, coinciding with authentic seismic disturbances on the island.

[61] https://www.volcanocafe.org/santorini-beauty-and-the-beast/comment-page-1/

Although three of our consortium's models correctly flagged the image as AI-generated, its visual content was still “accurately” geolocated to Santorini (Figure 90). This example demonstrates the impressive realism that GenAI can achieve. Furthermore, it exposes a significant paradox and challenge for detection models, especially when dealing with the output of nested AI systems.



Confidence score: 76%

Latitude
36.43026

Longitude
25.42356

*Figure 90 A view of an GenAI image logically but paradoxically geolocated in Santorini*

# 7 Conclusion: Insight, Challenges and Lessons Learned

We consider that through the insights that were gained thanks to these evaluation activities and the significant improvements they led to, the evaluation process achieved its goals, and was therefore successful.

Our evaluation work was largely conducted through an iterative process using real-world data and involved supporting high-impact fact-checking operations. This approach allowed us to address the ever-evolving verification challenges posed by multimodal Generative AI.

Particular emphasis was placed on mitigating and avoiding false positives, a critical ethical consideration given the reputational risk to fact-checking organisations and the potential for results to be weaponised in polemical online debates.

This focus reflects our commitment to developing responsible and constructive verification tools. Or to say it more trivially, to be part of the solution, not of the problem.

Following months of refinement, and as a direct result of this evaluation, several vera.ai tools integrated into the Verification Plugin and Truly Media are now used by thousands of fact-checkers, journalists, and researchers worldwide.

As a result of the Verification Plugin designed, “by fact-checkers for fact-checkers”, screenshots of tools’ results are frequently published by fact-checking publications as proof of forgery. While limitations remain, we have identified them; overcoming these will be part of our post-project plans and future research.

Furthermore, this project has yielded significant insights into the disruptions caused by Generative AI, including new file formats and novel manipulation techniques, such as the AI-driven modification of real images or video keyframes.

In terms of practice, we found that GenAI content is transforming the work of fact-checking. Beyond the classical "fake or real" claim, fact-checkers are often trying to respond to another nested claim: "is the content AI-generated or not". Such AI-related claims are spreading on social networks, sometimes as a tactic to denigrate legitimate content.

Then, regarding our knowledge of AI tools, our latest experiments are paving the way for a new generation of specialised AI agents to support further fact-checking and information verification at large.

In addition to this, the insights gained with regards to user workflow and professional-level needs helped us develop a design framework for trustworthy-by-design AI-based fact-checking tools (see Section 2.1.2), which is of benefit to both the AI HCI (human-computer interaction) and fact-checking communities.

# 8 References

Ahmed, S.K. (2024) The pillars of trustworthiness in qualitative research. Journal of Medicine, Surgery, and Public Health. 2, 100051.

Berman, A., De Fine Licht, K., Carlsson, V.: Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system (2024). Technology in Society. 76, 102471.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López De Prado, M., Herrera-Viedma, E., Herrera, F. (2023) Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion. 99, 101896.

Dobber, T., Kruikemeier, S., Votta, F., Helberger, N., & Goodman, E. P. (2025). The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media. Journal of Information Technology & Politics, 22(1), 82-97. https://doi.org/10.1080/19331681.2023.2224316

Ehn, P. (1993). Scandinavian design: On participation and skill, in Schuler, D. and Namioka, A. (eds.) Participatory Design: Principles and Practices. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 41-77.

Floridi, L. (2019) Establishing the rules for building trustworthy AI. Nature Machine Intelligence 1, 261–262.

Gaye, L., Schild, A., Lopez, E. (Accepted Manuscript) Designing for trustworthiness in AI-based fact-checking services. In: Degen, H., Ntoa, S. (eds) HCI International 2025 - Late Breaking Papers: Part XV, vol 16345. Springer, Cham.

Gol Mohammadi, N. (2019) Trustworthiness-by-design. Trustworthy Cyber-Physical Systems. pp. 79–118. Springer Fachmedien, Wiesbaden.

Karageogiou, D., Bammey, Q., Porcellini, V., Goupil, B., Teyssou, D., & Papadopoulos, S. (2024, September). Evolution of detection performance throughout the online lifespan of synthetic images. In European Conference on Computer Vision (pp. 400-417). Cham: Springer Nature Switzerland.

Maia, C.H., Ariel, P., Nunes, S. (2025) Adding human values on the deepfake: co-designing fact-checking solutions to combat misinformation. AI Ethics. 5, 3035–3050.

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A.B., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voß, A., Wrobel, S. (2023): Guideline for Designing Trustworthy Artificial Intelligence. Fraunhofer-Gesellschaft.

Righetti, N., & Balluff, P. (2025). CooRTweet: A Generalized R Software for Coordinated Network Detection. Computational Communication Research, 7(1), 1. https://doi.org/10.5117/CCR2025.1.7.RIGH

Srba, I., Razuvayevskaya, O., Leite, J. A., Moro, R., Baris Schlicht, I., Tonelli, S., Moreno García, F., Barrio Lottmann, S., Teyssou, D., Porcellini, V., Scarton, C., Bontcheva, K., Bielikova, M. (2025) A Survey on Automatic Credibility Assessment Using Textual Credibility Signals in the Era of Large Language Models. ACM Trans. Intell. Syst. Technol. https://doi.org/10.1145/3770077

Sykora, T. (2023). “AI UX Patterns”: Trustworthy AI with “fact-checking UI”. https://medium.com/@tomsyk/ai-ux-patterns-trustworthy-ai-with-fact-checking-ui-5e34aef66b10, last accessed 23/07/2025.

Teyssou, D. (2019). Applying the Design Thinking Methodology: The InVID Verification Plugin, in Mezaris, V et al. Video Verification in the Fake News Era, Springer Nature Switzerland (2019).

Teyssou, D. Document multimédia : Du deepfake au trucage hyperréel dans l'information d'actualité in Broudoux, E., Chartron, G., & Epron, B. (2025). Information et intelligence artificielle: Opportunités et risques. De Boeck Supérieur.

von Hippel, E. (1986) Lead Users: A Source of Novel Product Concepts, Management Science, Vol. 32, No. 7. https://doi.org/10.1287/mnsc.32.7.791