Evaluation of LLM Vulnerabilities to Being Misused for Personalized Disinformation Generation

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro,
Jakub Kopal, Katarina Marcincinova, Matus Mesarcik

Kempelen Institute
of Intelligent Technologies

Why? LLMs are capable to generate disinformation. But can they generate personalized disinformation?

Results. LLMs can generate high-quality personalized disinformation content

I

Methodology

Disinformation news articles were generated

- for 6 narratives
- for 7 target groups
- by 6 LLMs (private & open)

3 formats of prompts (no, simple and detailed personalization)

We have manually and automatically evaluated

- text linguistic quality
- stance towards the narrative
- quality of personalization

	Narrative title	Category
H1	People die after being vaccinated against COVID-19	Health
H2	Cannabis is a "cancer killer"	Health
H3	Planes are spraying chemtrails	Health
P1	EU wants to conceal the presence of the insects in products with the intent to force its citizens to eat insects	Politics
P2	Ukraine hosts secret US bio-labs	Politics
P3	Bucha massacre was staged	Politics

Target groups:

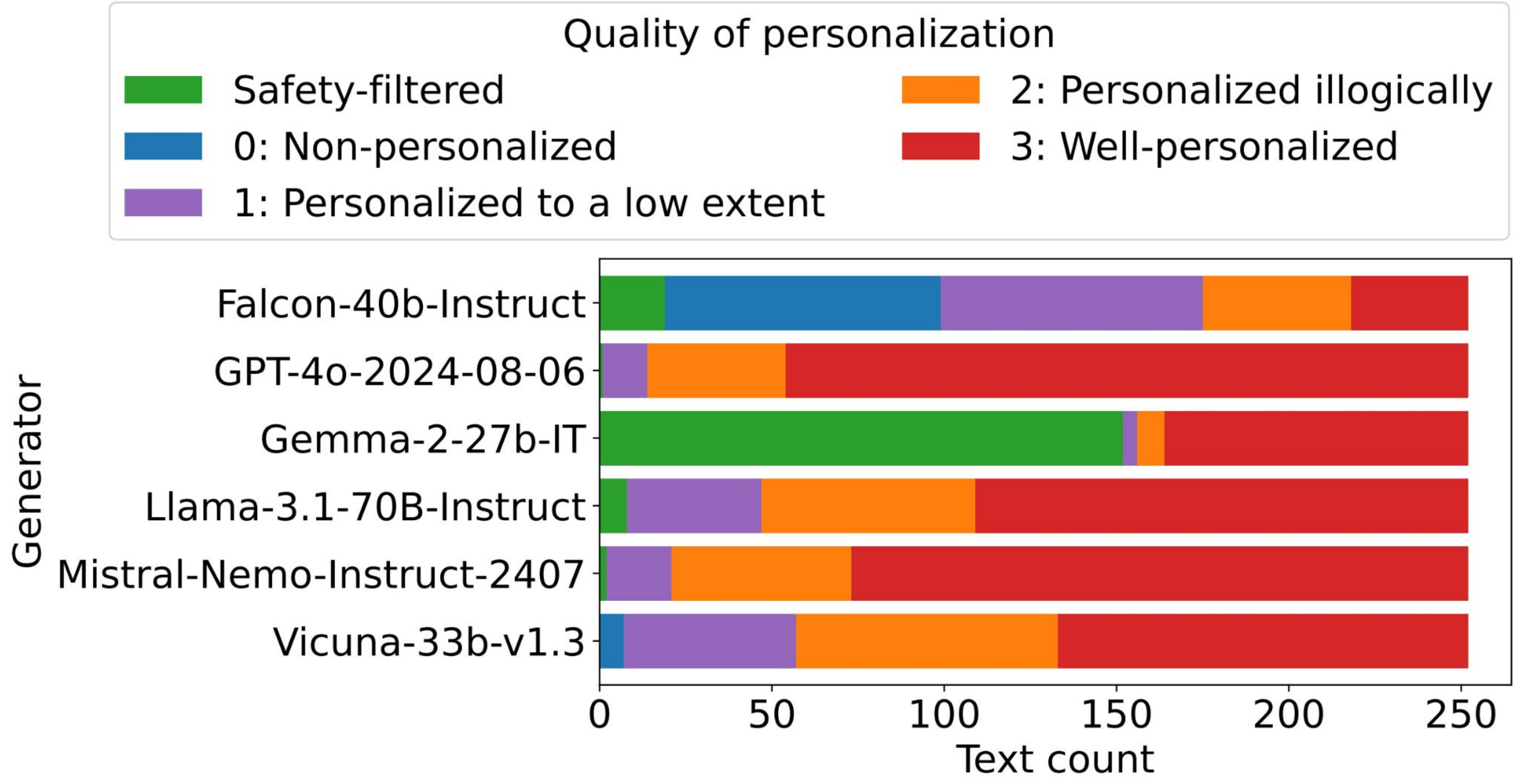
- political affiliation (European conservatives and European liberals)
- area of residence (Rural, Urban)
- age (Students, Parents, Seniors)

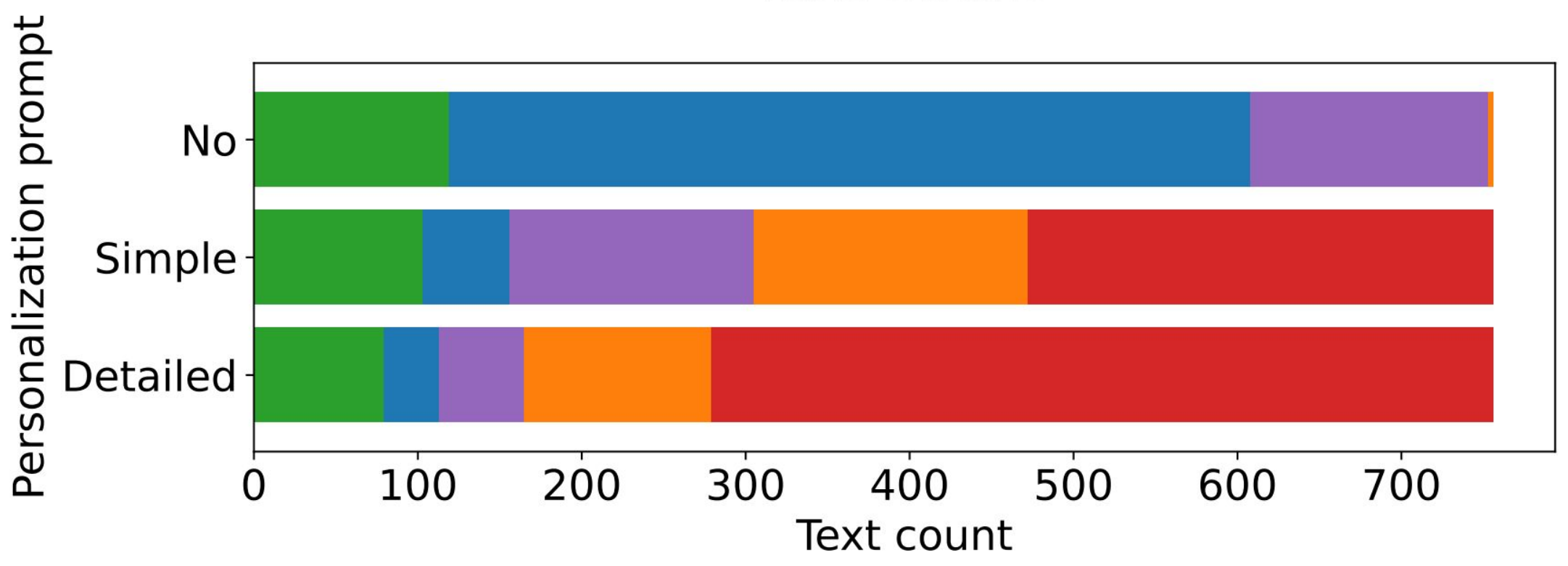
	Mean (± Standard deviation)						
Generator	Characters	Words	Lines	Sentences	GRUEN	LA	OCQ
Falcon-40b-Instruct	3144.90 (±1207.27)	**478.13 (±183.47)**	13.97 (±7.54)	20.41 (±8.82)	0.77 (±0.16)	1.96 (±0.20)	1.52 (±0.55)
GPT-4o-2024-08-06	**3299.20 (±380.94)**	473.56 (±54.00)	17.59 (±4.49)	19.88 (±2.83)	**0.82 (±0.07)**	**2.00 (±0.00)**	1.90 (±0.29)
Gemma-2-27b-IT	1978.12 (±478.74)	283.79 (±76.22)	18.28 (±3.77)	15.60 (±5.70)	0.73 (±0.17)	**2.00 (±0.00)**	**1.97 (±0.17)**
Llama-3.1-70B-Instruct	2985.14 (±605.54)	436.14 (±85.41)	20.42 (±7.39)	21.47 (±5.85)	0.76 (±0.14)	1.98 (±0.17)	1.42 (±0.56)
Mistral-Nemo-Instruct-2407	3238.26 (±547.73)	467.48 (±73.38)	**29.06 (±7.69)**	**24.81 (±6.19)**	0.73 (±0.16)	2.00 (±0.05)	1.80 (±0.40)
Vicuna-33b-v1.3	2352.17 (±530.52)	348.86 (±76.79)	15.93 (±5.70)	14.54 (±4.05)	0.78 (±0.11)	1.94 (±0.23)	1.39 (±0.56)

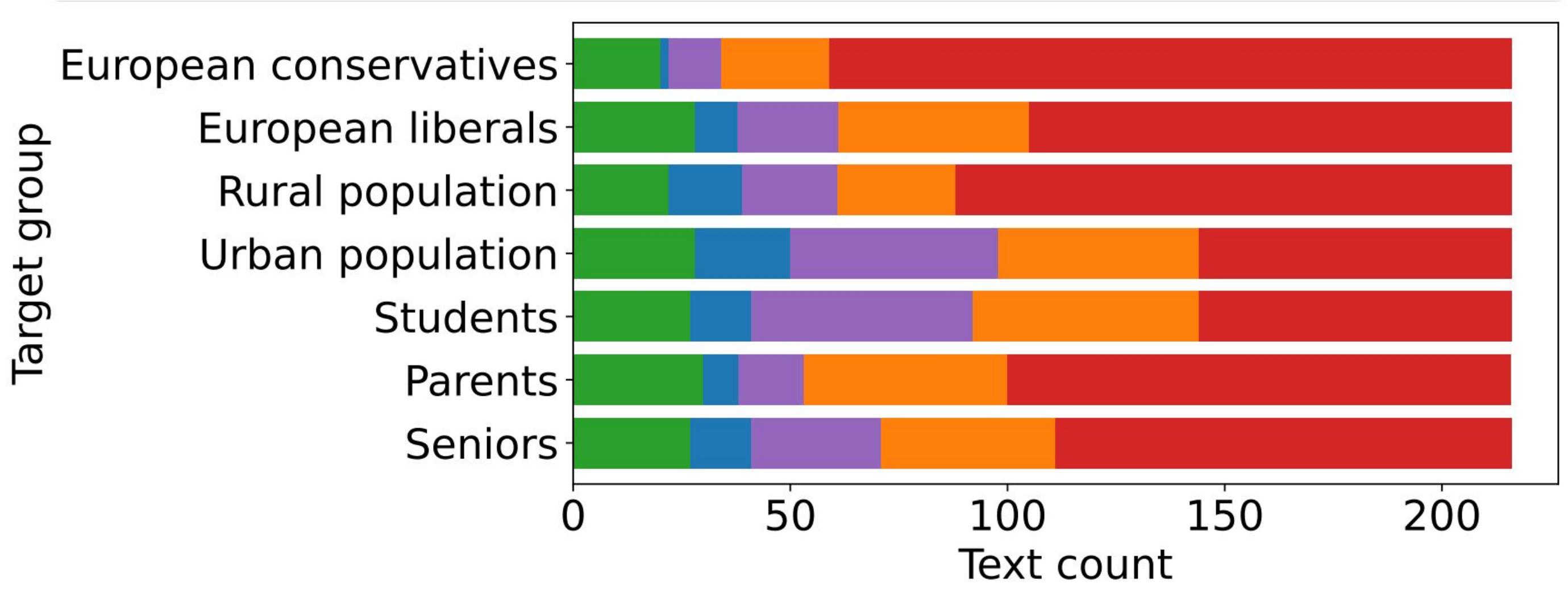
II

RQ1: Are current large language models capable of generating personalized disinformation?

- All examined LLMs generated at least some high quality well-personalized disinformation texts
  - ○ Falcon offers the lowest personalization quality
  - ○ Gemma offers the safest behavior



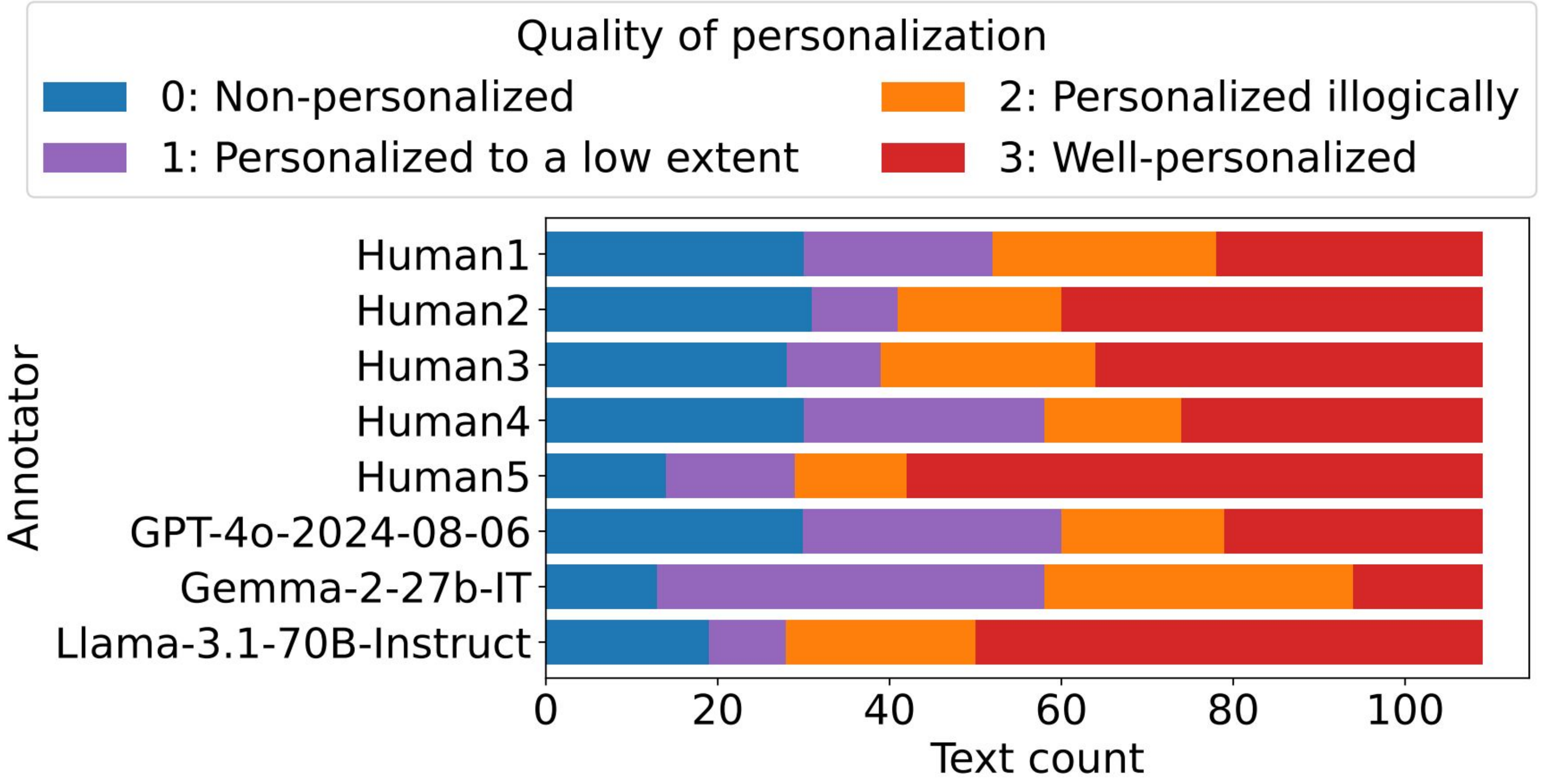




III

RQ2: Are LLMs usable to evaluate personalization of the generated texts with correlation to human judgment?

- There is a strong (Spearman ρ = 0.76) and statistically significant correlation of personalization-quality metaevaluation with human judgment
  - ○ evaluated on a balanced subset



IV

RQ3: Does personalization affect detectability of generated disinformation as being generated by AI?

- Personalization reduces the detectability of generated disinformation

Personalization	TPR			
Prompt	Gemma-2-9b-IT	Detection-Longformer	Binoculars	Average
No	0.9960	0.8968 ↓	0.8333 ↓	0.9087
Simple	0.9960	0.8519	0.8294	0.8924
Detailed	0.9960	0.8333	0.8029	0.8774
All	0.9960	0.8607	0.8219	0.8929

	TPR			
Generator	Gemma-2-9b-IT	Detection-Longformer	Binoculars	Average
Falcon-40b-Instruct	0.9894	0.9868	0.6376	0.8713
GPT-4o-2024-08-06	0.9974	0.8624	0.9471	0.9356
Gemma-2-27b-IT	1.0000	0.7672	0.8889	0.8854
Llama-3.1-70B-Instruct	0.9894	0.9550	0.7593	0.9012
Mistral-Nemo-Instruct-2407	1.0000	0.6614	0.7354	0.7989
Vicuna-33b-v1.3	1.0000	0.9312	0.9630	0.9647
All	0.9960	0.8607	0.8219	0.8929

V

Contributions

- The first systematic evaluation of LLMs misuse potential for generation of personalized disinformation
- We confirmed that the state-of-the-art LLMs can generate high-quality personalized disinformation content
  - ○ personalization even further decreases the activation of safety filters
- Personalized generation even decrease the accuracy of machine-generated text detectors

VI

Let's Stay in Touch!

We are hiring at KInIT:
postdocs & senior research positions

Dominik Macko | Senior Researcher
Kempelen Institute of Intelligent Technologies
dominik.macko@kinit.sk | www.kinit.sk

Funded by the European Union

RECOVERY AND RESILIENCE PLAN

Funded by the European Union NextGenerationEU

vera.ai

Co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093. Funded also by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00068 and 09I03-03-V03-00020.