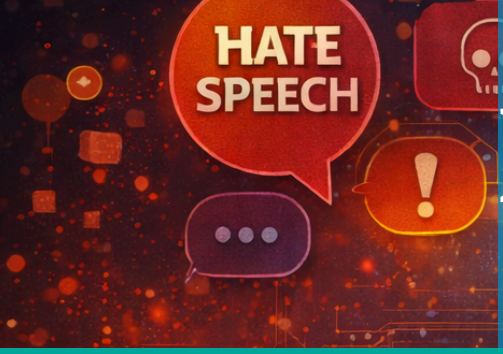

Decoding Hate

Funded by the European Union

UK Research and Innovation

Usage/License

MThe models can be used via the Hugging Face transformers library. The models are licensed under the Apache 2.0 License. Users are advised to be aware of potential biases present in the training data and understand that the models may misclassify some instances. Performance may vary across different domains. Code available under the Apache 2.0 License via the GitHub repository.

Download

The models can be accessed and downloaded from its HuggingFace collection page: https://huggingface.co/collections/irlab-udc/decoding-hate .

The code can be accessed via its GitHub repository https://github.com/HYBRIDS-MSCA/decoding-hate.

Researchers

- Paloma Piot
- Javier Parapar

These researchers are affiliated with the Information Retrieval Lab at the University of A Coruña.

Summary

This study analyses how large language models respond to hateful inputs in open text generation. Comparing seven state-of-the-art models, the authors assess whether they produce harmful content, neutral responses, or counter-speech. The work also examines mitigation strategies, such as fine-tuning and guardrails, and evaluates reactions to more subtle forms of hate speech. The findings reveal inconsistent behaviours across models and underline both risks and approaches for safer deployment.

Key Features

- Compares multiple state-of-the-art LLMs to analyse their responses to hate speech prompts.
- Highlights variability in outputs, from harmful content to neutral or counter-speech.
- Explores mitigation strategies, including fine-tuned “Stop-Hate” models and guardrails.
- Examines reactions to subtly or politically framed hate speech.

What We Offer

- Open-Source Codebase
- Pretrained Model Variants
- Evaluation Framework

Collaboration Objectives

- Safe AI Research: Collaboration on LLM safety and harmful content mitigation.
- Responsible Deployment: Improving guardrails for real-world systems.
- Model Improvement: Extending mitigation methods across languages and domains.
- Behavioural Benchmarking: Developing evaluation frameworks for hate-speech responses.

hybrids

UNIVERSIDADE DA CORUÑA

HYBRIDS Website
www.hybridsproject.eu

Get in Touch
paloma.piot@udc.es