DL4NLP - Project Proposal Group 15
Zero-Shot Human vs AI Generated text Detection using LLMs

Jose Carrilo
University of Amsterdam
jose.garcia.carrillo@student.uva.nl

Krystof Bobek
University of Amsterdam
krystof.bobek@student.uva.nl

Mara Dragomir
University of Amsterdam
mara.dragomir@student.uva.nl

Mahdi Rahimi
University of Amsterdam
mahdi.rahimi@student.uva.nl

Salvador Torpes
University of Amsterdam
salvador.baptista.torpes@student.uva.nl

Project Goal

In recent years AI development has accelerated rapidly, leading to the creation of increasingly sophisticated AI agents that incorporate powerful language models. This progress has raised concerns about the potential misuse of these resources, in contexts such as misinformation, content generation and education. From this, the need for AI generated text detection arises as a tool to help identify and mitigate the impact of such misuse. Among multiple attempts to address this issue, Mitchell et al.[6] proposed a zero-shot approach called *DetectGPT* that leverages probability curvature of perturbed human and AI-generated text. This approach is based on the observation that the loss curvature is consistently negative for AI-generated text and not clearly negative or positive for human-written text. Mitchell et al.'s [6] algorithm is based solely on the model's internal metrics, and it outperformed other methods including some that required fine-tuning and training, showing this approach's effectiveness. Further research lead Bao et al.[1] to propose *Fast-DetectGPT*, an improvement to the original algorithm that uses sampling from a LLM in order to produce the perturbed versions of both AI and human generated text.

Our goal is to implement *Fast-DetectGPT* and evaluate its performance in detecting AI-generated text using new models and a new dataset. Furthermore, some extensions may be explored in order to understand what are the limitations of this zero-shot approach.

Research Questions

1. Can we use a zero-shot approach to detect AI-generated text effectively?
2. How does the detection performance vary across different models?
3. How does the detection performance of a model compare when detecting text generated by itself versus text generated by other models?

Possible extensions We believe that the extensions will be discussed during the development of the project given that we haven yet been able to meet with our official TA. Currently, these are some of our ideas:

- Change the model used to create the perturbations on the *Fast-DetectGPT* algorithm;
- Use a dataset with a different language and use AI-generated text in that language.

Models and Datasets

We will use three different models for both synthetic dataset generation and detection, specifically: DeepSeekR1 [2], gpt-oss-20B [8], and Qwen3-4B 2507 [10].

When it comes to datasets, we will use HC3 [4], a large dataset containing questions with answers by both human and by GPT 3.5. The questions are separated by topics (finance, medicine, open question answering, Reddit and Wiki). We plan to choose a topic and sample a certain number of human sentences, and then use that to generate the synthetic datasets.

Apart from HC3, we also plan to use the four datasets originally tested by Mitchell et al.[6]: WritingPrompts [3] is a dataset with prompted stories and academic essays, SQuAD [9] contains Wikipedia paragraphs, XSum [7] contains news articles, and PubMedQA [5] is a group of long-form answers about biomedical research questions.

The metric used to measure performance on this task will be AUROC.

Workplan

Our workplan will focus on: (1) Building synthetic datasets with the three different models; The generation of AI text can be done either by prompting a model to rephrase the human sentence or to continue generating words based on the first words of the sentence, (2) Implementation of the *Fast-DetectGPT* algorithm and (3) Evaluation of the model's performance on the synthetic datasets with cross-evaluation, i.e., using each of the three models to detect AI-generated text from the other two models and also from itself.

References

[1] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature, 2024.

[2] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[3] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

[5] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023.

[7] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, 2018.

[8] OpenAI. gpt-oss-120b and gpt-oss-20b model card, 2025.

[9] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[10] Qwen Team. Qwen3 technical report, 2025.