Presenting Bust - A Benchmark for the Evaluation of System Detectors of LLM-generated Text

Joseph Cornelius and **Oscar William Lithgow Serrano** and **Sandra Mitrović**
Ljiljana Dolamic and **Fabio Rinaldi**
joseph.cornelius@idsia.ch

Abstract

The rapid advancement of Large Language Models (LLMs) like GPT-4 presents a growing challenge in distinguishing between human and machine-generated texts, heightening the risks of fraud and misinformation. Studies indicate that most adults struggle to differentiate between human and machine-generated content, underscoring the urgent need for reliable detection systems. To address this, there is a pressing requirement for a flexible benchmarking dataset that can effectively evaluate detection systems across various tasks.

In this work, a comprehensive approach was adopted, involving: 1) the creation of a dataset comprising instructions and responses obtained by different generator models; 2) the evaluation of different detectors' performance across different tasks; 3) the accomplishment of meta-analysis and development of surrogate models to provide in-depth insights into dataset characteristics and the potential to simulate detectors behaviour. The resulting benchmark dataset comprises over 25,000 texts derived from 3,180 instructions and synthetic responses generated by seven different instruction-tuned generators across ten tasks sourced from three different datasets. Five detectors (both close- and open-source), with varying training corpora and employing different detection strategies, were evaluated to assess their performance across these ten tasks.

The study revealed notable performance differences among detectors across different tasks. Furthermore, the detectors exhibited varying capabilities in detecting text produced by different generators. Surrogate models highlighted the difficulty in explaining the most performant detector, frequently relying on unexpected textual features.

Our (publicly available) automated pipeline integrating all analyses, facilitates detector selection based on text style and specific use-cases. This benchmarking effort marks a significant step forward in addressing the escalating challenges posed by the proliferation of machine-generated text, providing valuable insights for the development of more robust detection systems in combating fraud and misinformation.