A TwoStage Framework for LLM Generated Text Detection

Harshit Jaiswal
Chemical Engineering
IIT Kanpur
harshitj23@iitk.ac.in

Abstract

Large language models (LLMs) such as GPT-3.5, Claude Instant, and LLaMA-2 now generate human-quality text across diverse domains, but existing detectors—whether statistical, watermarking-based, or fine-tuned classifiers—are readily evaded by paraphrasing, polishing, or spelling-error attacks and often fail on unseen architectures.

In this work, we present a novel two-stage framework for detecting LLM-generated text that combines traditional fine-tuning with adversarial representation learning. First, we fine-tune a BERT encoder on RAID—a diverse, open-source benchmark of 2000+ texts spanning multiple domains, models, decoding strategies, and adversarial attacks—to establish a 92% baseline accuracy. To improve robustness against paraphrasing, polishing, and spelling-error attacks, we next recast detection as a "two-player" game by translating token-level attention maps into synthetic images and training a conditional GAN to learn latent features that distinguish human- and machine-authored passages. A lightweight regression head then classifies these GAN-extracted embeddings. On held-out RAID splits and the cross-domain DetectRL benchmark, our approach boosts accuracy from 68% (standard GAN) to 82% while outperforming zero-shot detectors such as DetectGPT and SIR. Ablations demonstrate that (1) attention-map visuals capture semantic invariants, and (2) the GAN's adversarial objective yields representations that generalize to unseen LLMs (GPT-3.5, Claude-instant, LLaMA-2). Our results suggest that multimodal feature learning can significantly enhance the reliability of LLM-text detection under real-world attack scenarios.