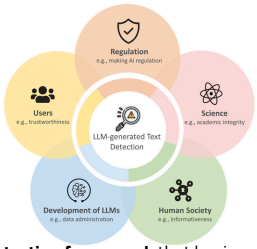
A TWOSTAGE FRAMEWORK FOR LLM GENERATED TEXT DETECTION

Harshit Jaiswal
Department of Chemical Engineering, IIT Kanpur

Prof. Tushar Sandhan
Department of Electrical Engineering, IIT Kanpur

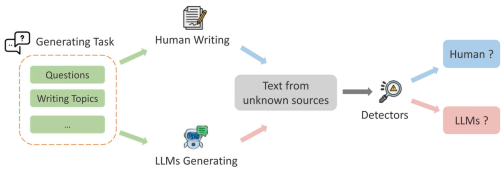
Abstract

The rapid advancement of large language models (LLMs) has heightened the need to reliably **distinguish human-written** from **machine-generated text variant** coming from a range of ever growing cohort of LLMs .



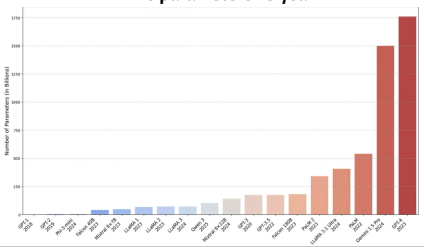
We propose a **two-stage detection framework** that begins with **fine-tuning BERT** for binary classification, then augments robustness using **GAN**-based adversarial training and a label-supervised LLaMA to bring **interpretability**.

Introduction



- **Proliferation of LLMs** : The past two years have seen an explosion in large language model usage (GPT-3.5, GPT-4, LLaMA 2), powering chatbots and more.
- **Our Goal** : Design a lightweight yet resilient detector that (1) delivers high raw accuracy, (2) resists paraphrase-and-attack tactics, and (3) scales to new model families.

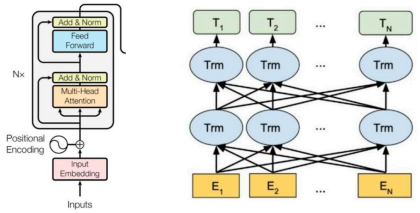
LLMs parameters vs year



Methodology

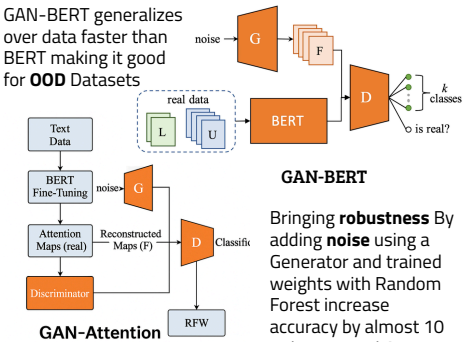
BERT- Finetuning

Pre-trained **BERT-Base** (110 M parameters)
Fine-tuned on CHEAT's human vs LLM corpus
Hyperparams: lr = 2e-5, batch = 16, epochs = 3



GAN-BERT

GAN-BERT generalizes over data faster than BERT making it good for **OOD** Datasets



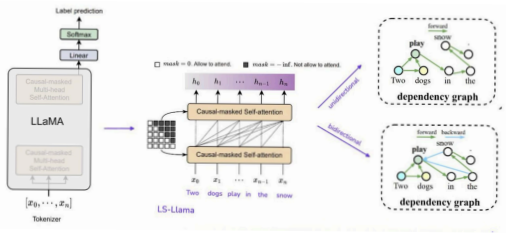
GAN-Attention

GAN-BERT

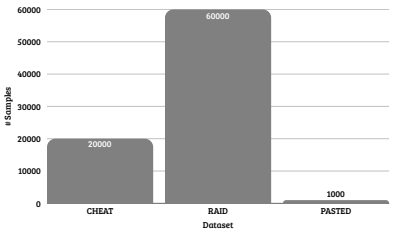
Bringing **robustness** By adding **noise** using a Generator and trained weights with Random Forest increase accuracy by almost 10 % than normal GANs

LLaMA as a Judge

Our attempt towards **interpretability** by using generation along with classification.



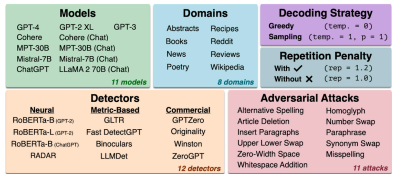
Dataset



CHEAT

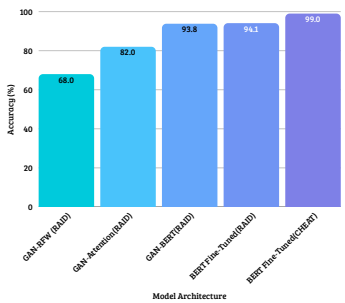
LLM and Human generated abstracts mostly from academic papers.

RAID

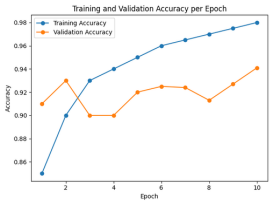


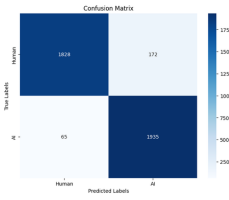
- Hardware & Software:
  - NVIDIA A100 GPUs (**80GB×4**)
  - PyTorch , Transformers
- Preprocessing:
  - Lowercasing, punctuation normalization
  - Tokenization with BERT tokenizer

Results and Analysis



GAN-BERT Plots





Cross-Model and Dataset Comparisons

Table 2: Performance across all models and datasets

Model	Dataset Description	Test Accuracy
GAN-RFW (concurrency matrix)	RAID(20k Traing , 4k Test)	68
GAN-Attention	RAID(20k Traing , 4k Test)	82
GAN-BERT	RAID(20k Traing , 4k Test)	93.8
Bert finetuned	RAID(20k Traing , 4k Test)	94.1
BERT finetuned	CHEAT (4000 Train,800 Test)	99

Important Analysis and Highlights

Highlight	Details	Performance
Top Performer on RAID	GAN-BERT vs. BERT	93.8 % vs. 94.1 %
Massive Adversarial Gain	GAN-Attention vs. GAN-RFW	82 % vs. 68 % (+14 pts)
Clean-Data Ceiling	BERT fine-tuned on CHEAT	99 %
Robustness Edge	GAN-BERT vs. GAN-RFW	+25.8 pts
Attack Impact	BERT CHEAT → RAID	99 % → 94.1 % (-4.9 pts)

Acknowledgement

- I extend my profound gratitude to **Prof. Tushar Sandhan** for his guidance and support throughout this work.
- I am equally thankful to **Dr. Jivnesh Sandhan** from **Kyoto University** for his insightful feedback and constant encouragement, which greatly enriched this project.

References

- J. Wu et al., "A Survey on LLM-Generated Text Detection," arXiv:2310.14724, 2023.
- L. Dugan et al., "RAID: Benchmark for Machine-Generated Text Detectors," arXiv:2405.07940, ACL 2024.
- Z. Li et al., "Label-Supervised LLaMA Finetuning," arXiv:2310.01208, 2023.
- J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," arXiv:1810.04805, 2018.
- D. Croce et al., "GAN-BERT: Generative Adversarial Learning for Robust Text Classification," ACL 2020.