Can We Trust LLM Detectors?

Jivnesh Sandhan[1]   Harshit Jaiswal[2]   Fei Cheng[1]   Yugo Murawaki[1]

[1]Kyoto University, Japan   [2]IIT Kanpur, India.

{jivnesh,feicheng,murawaki}@i.kyoto-u.ac.jp   harshitj23@iitk.ac.in

Abstract

The rapid adoption of LLMs has increased the need for reliable AI text detection, yet existing detectors often fail outside controlled benchmarks. We systematically evaluate 2 dominant paradigms (training-free and supervised) and show that both are brittle under distribution shift, unseen generators, and simple stylistic perturbations. To address these limitations, we propose a supervised contrastive learning (SCL) framework that learns discriminative style embeddings. Experiments show that while supervised detectors excel in-domain, they degrade sharply out-of-domain, and training-free methods remain highly sensitive to proxy choice. Overall, our results expose fundamental challenges in building domain-agnostic detectors.[1]

1   Introduction

The rapid adoption of large language models (LLMs) has intensified the need for reliable LLM detection, particularly in academic settings where pedagogical integrity and research authenticity are critical [1]. While many existing detectors report near-perfect in-domain accuracy, their reliability in real-world, out-of-distribution (OOD) settings remains questionable [2]. This mismatch between benchmark performance and practical robustness undermines trust in automated detection systems. In this work, we systematically evaluate the reliability of 2 dominant detection paradigms (training-free and supervised fine-tuning) and propose a new framework to mitigate their OOD failures.

Current LLM detectors mainly follow 2 approaches. Supervised classifiers, often based on fine-tuned models such as BERT [3], achieve high in-domain accuracy but tend to overfit to model-specific artifacts, leading to severe degradation on text from unseen generators. In contrast, training-free detectors, including DetectGPT [4], Fast-DetectGPT

[5], and Binoculars [6], rely on intrinsic statistical properties of generated text. Although appealing, their performance is highly sensitive to the choice of proxy model, introducing a distinct but equally problematic brittleness.

To address these limitations, we propose a supervised contrastive learning (SCL) framework that learns discriminative style embeddings separating human and machine-generated text. Using a DeBERTa-v3 backbone optimized with an InfoNCE loss, our method enables efficient few-shot adaptation to new LLMs with as few as 25 examples. Extensive experiments show that supervised detectors degrade substantially under distribution shift, while training-free methods remain fragile to proxy selection. We find that limited supervision improves generalization but does not fully resolve these issues. Our robustness analysis further reveals vulnerabilities to adversarial attacks and simple stylistic perturbations, highlighting severity of reliability these detectors. Our contributions are as follow:

- We demonstrate that both training-free and supervised detectors are unreliable under domain shift.
- We introduce a contrastive learning framework that enables adaptation to new LLMs.
- We provide comprehensive OOD and adversarial analyses, providing evidence that it is not possible to built universal detector for all domains.

2   The proposed framework

We use a pre-trained BERT model as our base model fine-tuned with binary classification objective. The model is trained by minimizing the binary cross-entropy loss function:

$$\mathcal{L}_{\mathrm{BCE}} = -\left[y \log(p) + (1-y) \log(1-p)\right] \qquad (1)$$

where y is the true label and p is the model's predicted probability. We propose a Supervised Contrastive Learning (SCL) module illustrated in Figure 1, inspired by recent successes in learning discriminative representations
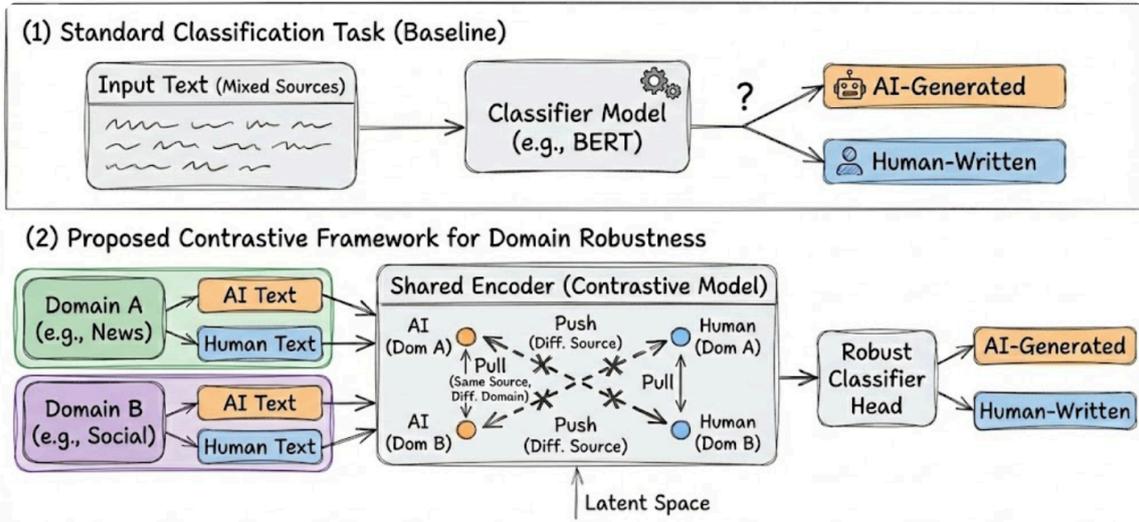

Figure 1: Overview of the AI text detection task and proposed framework. Our framework integrates a supervised classifier with a supervised contrastive learning module that learns style embeddings, enabling robust detection.

[7, 8, 9]. This module helps learning a meaningful style embeddings using a DeBERTa-v3 backbone and a projection head. It is trained with the InfoNCE loss function:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\left(\text{sim}(z_i, z_j^+)/\tau\right)}{\sum_{k=1}^{N} \exp(\text{sim}(z_i, z_k)/\tau)} \quad (2)$$

where z_i is an anchor embedding, z_j^+ is a positive sample, $\text{sim}(\cdot, \cdot)$ is cosine similarity, and τ is a temperature parameter. This objective structures the embedding space such that classes form distinct, coherent clusters.

3 Experiments

Datasets: We use 3 datasets: CHEAT, with 35,304 academic abstracts for high-performance supervised evaluation [10]; RAID, a large-scale benchmark with over 6 million texts from 11 LLMs across 8 domains for cross-domain comparison [11]; and M4, a multi-generator, multi-domain, and multi-lingual corpus designed for black-box detection that encompasses diverse linguistic settings and generative models to test detector generalization [12].

Baselines: We evaluate our method against two training-free and two supervised baselines. We use NVIDIA A100 and TITAN RTX GPUs with a learning rate of 2e-5 for models training.

- **Binocular:** A perplexity-based, training-free detector compares token-level likelihoods under a base language model and its perturbed variant to find statistical regularities characteristic of AI-generated text [6].
- **FastDetectGPT:** A discrepancy-based, training-free

approach that measures the divergence between the likelihood of a text under an original model and samples generated from it [5].

- **BERT:** A supervised BERT-base classifier fine-tuned for binary classification.
- **GAN-BERT:** We adopt GAN-BERT [13], which integrates adversarial learning with supervised text classification. The model comprises a generator and a discriminator: the generator maps Gaussian noise $z \sim \mathcal{N}(0,1)$ to synthetic embeddings that mimic BERT representations, while the discriminator receives either real BERT embeddings or generated ones and predicts $k + 1$ classes (the k task labels plus a "fake" class).
- **Ours:** The proposed method describe in §2.

Main Results: Table 1 summarizes results on the in-domain RAID dataset and the out-of-domain (OOD) CHEAT and M4 benchmarks. In-domain, supervised methods clearly outperform training-free approaches, with FastDetectGPT exhibiting extremely low recall, while Binoculars provides moderate gains. Among supervised models, GAN-BERT and BERT achieve strong performance, but our method attains the best overall results on RAID, achieving the highest accuracy and F1-score, perfect precision, and the lowest false positive rate. Under OOD evaluation, all methods show performance degradation. On CHEAT, our method generalizes well and achieves the best accuracy and F1-score with a low false positive

Table 1: Results of all baselines on the in-domain RAID and out-of-domain (CHEAT, M4) benchmarks. Supervised methods outperform training-free approaches in-domain, while all models degrade under OOD shifts. Our method achieves the best performance on RAID and generalizes to CHEAT, but fails on M4 due to substantial domain mismatch, highlighting the difficulty of building a universal LLM detector across domains and models.

Dataset	Method	Acc.	Prec.	Recall	F1	FPR
RAID	FastDetectGPT	34.63%	92.00%	2.00%	4.00%	65.37%
	Binoculars	82.81%	93.61%	79.65%	86.07%	17.19%
	GAN-BERT	95.72%	95.85%	97.81%	96.82%	8.50%
	BERT	95.31%	94.37%	**98.86%**	96.56%	9.00%
	Ours	**95.98%**	**100.00%**	94.00%	**97.00%**	**7.00%**
CHEAT	FastDetectGPT	54.33%	97.00%	9.00%	16.00%	45.67%
	Binoculars	97.41%	95.49%	**99.52%**	97.46%	2.59%
	GAN-BERT	83.56%	77.81%	93.90%	85.10%	26.80%
	BERT	76.75%	68.46%	99.22%	81.02%	34.00%
	Ours	**97.83%**	**98.00%**	98.00%	**98.00%**	**2.00%**
M4	FastDetectGPT	49.99%	0.00%	0.00%	0.00%	50.01%
	Binoculars	71.35%	**96.15%**	44.48%	60.82%	28.65%
	GAN-BERT	**76.12%**	77.44%	**73.72%**	**75.53%**	**21.50%**
	BERT	66.42%	69.44%	58.66%	63.59%	33.00%
	Ours	50.83%	66.00%	4.00%	7.00%	49.00%

rate, outperforming other supervised baselines. In contrast, on M4, performance drops sharply for all models; while GAN-BERT performs best among baselines, our method fails to generalize, with very low recall and F1. Overall, these results show that supervised detectors dominate in-domain settings, OOD shifts remain challenging, and achieving a universal LLM detector that generalizes across diverse domains and models remains difficult.

Why does our approach succeed on OOD CHEAT but fail on OOD M4? We analyze OOD generalization by evaluating models trained on RAID against the CHEAT and M4 benchmarks. Our framework transfers effectively to CHEAT, achieving 97.83% accuracy, due to strong stylistic alignment between the datasets: the inclusion of ArXiv abstracts in RAID enables the style encoder to learn representations that generalize well to academic text. In contrast, performance drops sharply on M4 because of severe distributional shift. M4 exhibits substantially higher character diversity, increased digit density, and longer average text length compared to RAID, reflecting a move from formal sources such as Wikipedia and news to informal, noisy domains like Reddit [12]. This mismatch introduces linguistic variability that the learned style representations fail to capture.

4 Analysis

4.1 Can our framework adapt to unseen LLMs with a few examples?

We study whether our method can adapt to unseen LLMs (GPT-4o and claude-3.5) using only a few examples. Experiments are conducted on the LMSYS Arena dataset, which contains conversations from 55 chatbots and supports both zero-shot and few-shot generalization. We first compute class centroids by averaging style embeddings of human and AI samples. For adaptation to a new, unseen LLM, we generate a small number of model-specific samples (e.g., 25) and update only the AI centroid, without retraining the base model. Figure 2 compares zero-shot and 25-shot performance, demonstrating that this lightweight adaptation yields substantial gains on both GPT-4o and claude-3.5. Overall, the results highlight the efficiency of the SCL framework for adapting to new LLMs.

4.2 How robust is our proposed approach to adversarial perturbations?

We evaluate the robustness of our framework under four adversarial settings. For white-box, gradient-based attacks, we assume full access to the detector and apply the Greedy Coordinate Gradient (GCG) attack [14]. We also test univer-
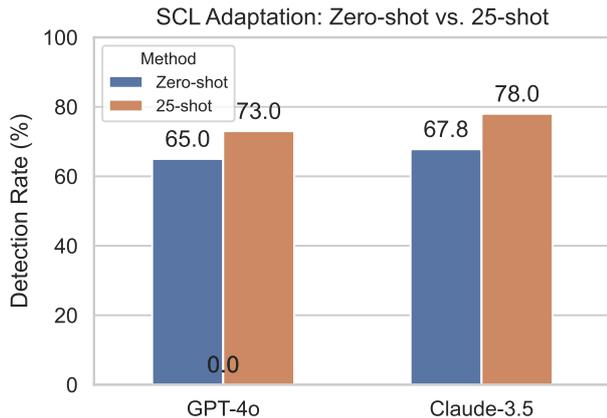
Figure 2: Zero-shot vs. 25-shot detection performance of our method on unseen LLMs, showing that lightweight adaptation without retraining yields substantial gains on both `GPT-4o` and `claude-3.5` LLMs.

sal trigger attacks using a fixed adversarial suffix [15], and consider simple black-box heuristic perturbations such as quotation marks, attribution cues, and typographical noise. White-box GCG attacks can flip predictions for each sample 99.3%, but this success comes at the cost of semantic naturalness. It means the optimized texts exhibit broken syntax, repetition, and nonsensical tokens, making them easily identifiable by humans. In contrast, GCG universal attempts to find universal trigger for all samples but it fail to do so. Notably, the detector is most sensitive to simple black-box stylistic perturbations: adding quotation marks and attribution reduces accuracy from 3.3% to 8.5%, while introducing minor typos unexpectedly improves accuracy. These results indicate that current detectors rely on brittle surface cues, underscoring the need for more robust, style-aware detection methods.

Table 2: Results showing how different perturbations induce misclassification of AI-generated text as human. Per-sample GCG is highly effective, while universal GCG fails to generalize; simple noise such as quotation marks, attribution cues, and typographical perturbations also significantly impact performance.

Attack / Perturbation	Success
GCG (per-sample)	99.3%
GCG (universal)	0%
Quotation/Attribution	8.5%
Typographical Noise	3.3%

5 Related Work

Although recent LLM detectors report near-perfect accuracy, they remain unreliable in practice, motivating analysis of their generalization limits and the feasibility of trustworthy detection. Existing approaches fall into supervised neural detectors and training-free methods [16]. Early detectors exploited surface statistics such as n-grams, entropy, and perplexity, achieving strong in-domain performance (e.g., 99% on CHEAT [10]) but exhibiting severe model brittleness under generator, domain, and decoding shifts. Training-free methods aim to remove labeled data dependence by leveraging intrinsic text statistics. DetectGPT [4], along with Fast-DetectGPT [5] and Binoculars [6], improves efficiency and scalability, but detection remains highly sensitive to the proxy language model, limiting robustness under distribution shift. Our experiments confirm that supervised detectors outperform training-free methods in-distribution but degrade substantially under OOD settings. Prior work addresses robustness via adversarial training (GAN-BERT [13]) and contrastive learning for OOD generalization [9, 8, 7]. Building on this, we introduce a contrastive objective into strong detectors and show that while robustness improves modestly, no universal detector generalizes reliably across models, domains, and adversarial strategies, revealing fundamental limits of current detection paradigms.

6 Conclusion

This work systematically examined whether current LLM detectors can be trusted under realistic deployment conditions. Through extensive in-domain, out-of-domain, and adversarial evaluations, we showed that both training-free and supervised detectors exhibit significant brittleness when confronted with unseen generators, domain shifts, or simple stylistic perturbations. While supervised methods dominate in-domain settings, their performance degrades sharply under distributional mismatch, and training-free approaches remain highly sensitive to proxy model choices. We introduced a supervised contrastive learning framework that improves robustness and enables efficient few-shot adaptation to new LLMs, demonstrating clear gains in certain OOD scenarios. Overall, our findings suggest that universal, domain-agnostic LLM detection remains infeasible with current paradigms.

Acknowledgments

References

[1] Michael Sheinman Orenstrakh, Oscar Karnalim, Carlos Aníbal Suárez, and Michael Liut. Detecting llm-generated text in computing education: Comparative study for chatgpt cases. In **2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)**, pp. 121–126, 2024.

[2] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2025.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: zero-shot machine-generated text detection using probability curvature. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.

[5] 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. In **The Twelfth International Conference on Learning Representations**, 2024.

[6] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text, 2024.

[7] Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning, 2024.

[8] Junfan Chen, Richong Zhang, Xiaohan Jiang, and Chunming Hu. Spcontrastnet: A self-paced contrastive learning model for few-shot text classification. **ACM Trans. Inf. Syst.**, Vol. 42, No. 5, April 2024.

[9] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R. Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. In **2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 15297–15307, 2021.

[10] Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. **IEEE Transactions on Big Data**, Vol. 11, No. 3, pp. 898–906, 2025.

[11] Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1369–1407, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[13] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2114–2119, Online, July 2020. Association for Computational Linguistics.

[14] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[15] Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. Universal adversarial attacks with natural triggers for text classification. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3724–3733, Online, June 2021. Association for Computational Linguistics.

[16] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. **Computational Linguistics**, Vol. 51, No. 1, pp. 275–338, March 2025.