Almost AI, Almost Human: The Challenge of Detecting AI-Polished Writing

Shoumik Saha
University of Maryland
College Park, USA
smksaha@umd.edu

Soheil Feizi
University of Maryland
College Park, USA
sfeizi@umd.edu

Abstract

The growing use of large language models (LLMs) for text generation has led to widespread concerns about AI-generated content detection. However, an overlooked challenge is AI-polished text, where human-written content undergoes subtle refinements using AI tools. This raises a critical question: should minimally polished text be classified as AI-generated? Such classification can lead to false plagiarism accusations and misleading claims about AI prevalence in online content. In this study, we systematically evaluate *twelve* state-of-the-art AI-text detectors using our AI-Polished-Text Evaluation **(APT-Eval) dataset**, which contains $15K$ samples refined at varying AI-involvement levels. Our findings reveal that detectors frequently flag even minimally polished text as AI-generated, struggle to differentiate between degrees of AI involvement, and exhibit biases against older and smaller models. These limitations highlight the urgent need for more nuanced detection methodologies.

github.com/ShoumikSaha/ai-polished-text

huggingface.co/datasets/smksaha/apt-eval

1 Introduction

The rapid advancement of LLMs has enabled AI to generate highly fluent, human-like text, raising concerns about detectability and prompting the development of various AI-text detectors (Gehrmann et al., 2019; Mitchell et al., 2023; Hu et al., 2023). However, the distinction between AI-generated and human-written text remains a gray area, particularly when human-authored content is refined using AI tools. *If a human-written text is slightly polished by AI, should it still be classified as human-written, or does it become AI-generated?* Classifying such text as AI-generated can lead to false plagiarism accusations and unfair penalties, especially when detectors flag minimally polished content as AI-generated.[1]

Additionally, reports suggesting that a significant share of online content is AI-generated – such as analyses indicating that over 40% of Medium posts are likely AI-written – often overlook the nuance of AI-polished text, which detection tools may misclassify.[2,3] These sweeping claims risk misrepresenting the actual extent of AI involvement, leading to misleading statistics and misplaced skepticism about human authorship. Motivated by these issues, our study systematically examines how AI-text detectors respond to AI-polished text and whether their classifications are both accurate and fair.

To investigate this issue, we introduce the **AI-Polished-Text Evaluation (APT-Eval)** dataset of size **15K**, which systematically examines how AI-text detectors respond to varying degrees of AI involvement in human writing. Our dataset is built from pre-existing human-written samples that are refined using different LLMs, such as GPT-4o (OpenAI, 2023), Llama3-70B (Dubey et al., 2024), DeepSeek-V3 (Liu et al., 2024), etc., applying degree and percentage based modifications. This allows us to assess how detectors respond to minor and major AI polishing. We analyze the classification accuracy, false positive rates, and domain-specific sensitivities of **12 state-of-the-art detectors**, spanning model-based, metric-based, and commercial systems.

Our findings reveal critical weaknesses in existing AI-text detection systems. AI-text detectors exhibit alarmingly high false positive rates, often flagging very minimally polished text as AI-generated. For instance, minimal polishing with GPT-4o can lead to detection rates ranging from 10% to 75%, depending on the detector. Furthermore, detectors struggle to differentiate between minor and major AI refinements, suggesting that they may not be as reliable as previously assumed. We also uncover biases against smaller or older

[1] USAToday News on false AI allegations

[2] Wired news on Medium [3] NewsBytes report on Medium

Findings of the Association for Computational Linguistics: ACL 2025, pages 25414–25431
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

LLMs, where polishing done by less advanced models is more likely to be flagged than text refined by state-of-the-art LLMs. On average, 46% of samples polished by LLaMA2-7B are classified as AI-generated, whereas this drops to 23% for DeepSeek-V3-polished samples. Furthermore, our study uncovers domain-specific inconsistencies in detection accuracy. Detectors show the greatest vulnerability for the 'speech' domain, while exhibiting comparatively higher robustness to 'paper-abstract' texts.

These findings raise concerns about the fairness and generalizability of current detection methods. By shedding light on these issues, our research provides valuable insights into the evolving challenges of AI-assisted writing and the limitations of current AI-text detection methodologies. Our code and dataset are publicly available.

2 APT(AI-Polished-Text) Eval Dataset

2.1 Initial Dataset

In this study, we begin with purely human-written texts (HWT) and refine them using various large language models (LLMs). Building on the work of Zhang et al. (2024), we utilize HWT samples from their 'MixSet' dataset. These samples are carefully selected based on two key criteria: (1) they were created prior to the widespread adoption of LLMs, and (2) they span six distinct domains. For clarity, we refer to this baseline HWT dataset as the 'No-Polish-HWT' set. This set comprises 300 samples, with 50 samples per domain (details in Table 3).

2.2 Dataset Preparation

As we generate the AI-polished versions of our No-Polish-HWT samples, we adjust the level of AI/LLM involvement. We employ two distinct polishing strategies: –

1. **Degree-Based Polishing:** The LLM is prompted to refine the text in four varying degrees of modification: – (1) extremely-minor, (2) minor, (3) slightly-major, and (4) major.

2. **Percentage-Based Polishing:** The LLM is instructed to modify a fixed percentage ($p\%$) of words in a given text. The percentage is systematically varied across the following values: $p\% = \{1, 5, 10, 20, 35, 50, 75\}\%$.

As a result, each HWT sample is transformed into 11 distinct AI-polished variants. For the LLM-polishing, we employ five different models: GPT-4o, Llama3.1-70B, Llama3-8B, Llama2-7B, and DeepSeek-V3. Each model is carefully prompted to generate the highest-quality output, preserving the original semantics of the text (details provided in Appendix A.2). Figure 1 illustrates a randomly selected sample from our dataset, which has been polished by GPT-4o with an extremely minor modification.

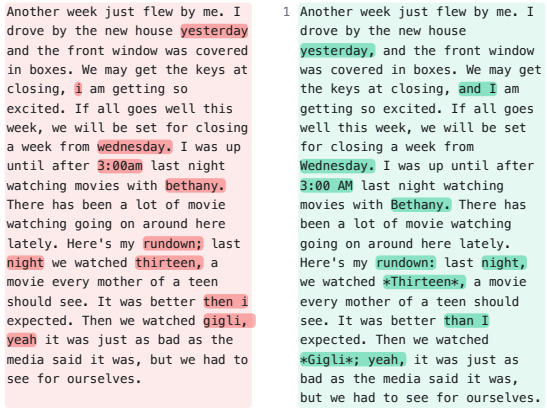


Figure 1: Random sample from our APT Eval dataset. Original HWT on left; Polished version on right.

2.3 Dataset Analysis

To assess the differences and deviations between pure HWT and AI-polished text, we employ three key metrics: Cosine semantic similarity, Jaccard distance, and Levenshtein distance. For semantic similarity, we compute the cosine similarity between the embeddings of the original and AI-polished texts (APT) using the BERT-base model. To ensure that the polished samples retain a strong resemblance to the original text, we filter out any samples with a semantic similarity below 0.85. Figure 2 shows the distribution for APTs (degree-based) with their mean value (see all metrics, and plots in Appendix A.3).

Polish Type	Polisher LLM					Total
	GPT-4o	Llama3.1 70B	Llama3 8B	Llama2 7B	DeepSeek V3	
no-polish / pure HWT	-	-	-	-	-	300
Degree based	1152	1085	1125	744	1141	5247
Percentage based	2072	2048	1977	1282	2078	9457
Total	3224	3133	3102	2026	3219	**15004**

Table 1: Our APT Eval Dataset

After filtering, the final APT-Eval dataset con-

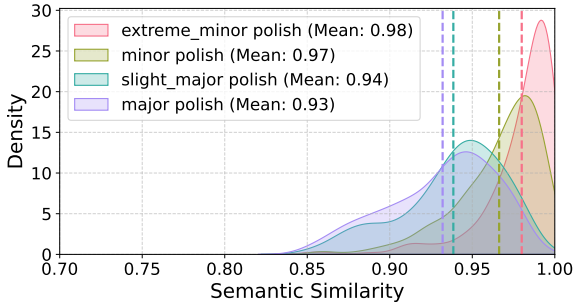


Figure 2: Distribution of Semantic Similarity for Degree-based AI-Polished Texts by GPT-4o.

sists of **15K** samples, providing a robust benchmark for evaluating AI-text detection systems. Table 1 shows the total number of samples for each strategy and polisher (more details in table 2).

3 AI-text Detectors

In this work, we evaluate a total of twelve detectors from three different categories:

1. **Model-based:** RADAR (Hu et al., 2023), RoBERTa-Base (ChatGPT) (Guo et al., 2023), RoBERTa-Base (GPT2), and RoBERTa-Large (GPT2) (OpenAI, 2019).

2. **Metric-based:** GLTR (Gehrmann et al., 2019), DetectGPT (Mitchell et al., 2023), Fast-DetectGPT (Bao et al., 2023), LLMDet (Wu et al., 2023), Binoculars (Hans et al., 2024).

3. **Commercial:** ZeroGPT [4], GPTZero [5], Pangram[6].

3.1 Detectors' Threshold

AI-text detectors generate a scalar score or prediction based on a given sequence of input tokens. To transform this score into a binary classification, an appropriate threshold must be determined. Dugan et al. (2024) highlight that a naive threshold selection can lead to poor accuracy or a high false positive rate (FPR). Therefore, we optimize the threshold for each detector to achieve maximum accuracy in detecting HWT and AI-text.

We evaluate these detectors on 300 samples of our 'no-polish-HWT' (pure human-written) set and 300 samples of pure AI-generated texts from the dataset of Zhang et al. (2024). Most detectors achieve $70\% - 88\%$ accuracy, with a false positive rate of $1\% - 8\%$. Table 6 shows the detector-specific threshold with their accuracy and FPR.

[4] https://www.zerogpt.com/ [5] https://gptzero.me/
[6] https://www.pangram.com/

4 Key Findings

We evaluate the detectors on our APT-Eval dataset from multiple perspectives to analyze their response to AI-polished text. Our key findings are as follows –

4.1 Alarming false positive rate by AI-text detectors for minor polishing.

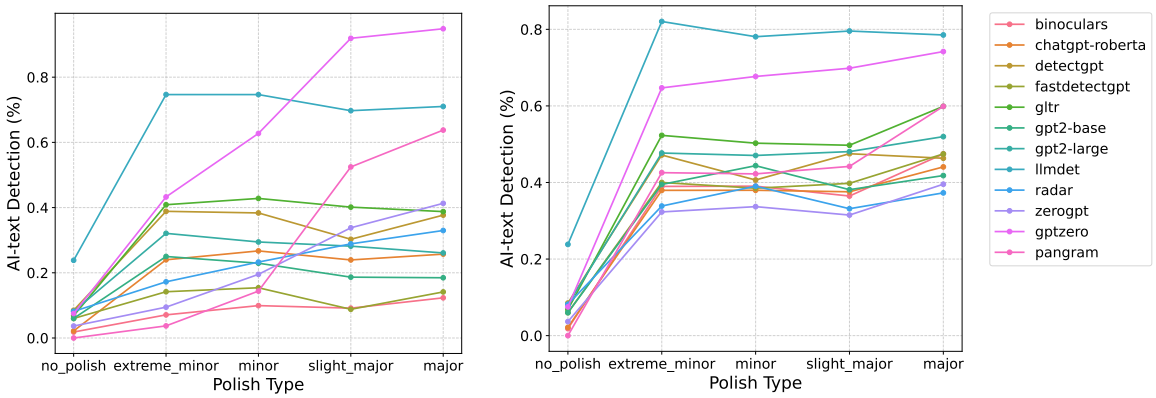
Though most detectors can achieve a low FPR on pure HWT (Table 6), most of them give a high FPR for any polishing, especially for extremely minor and minor polishing. For example, GLTR, with a 6.83% FPR on pure HWT, classifies 40.87% of extremely minor and 42.81% of minor-polished GPT-4o texts as AI-text. This trend extends to percentage-based polishing – GLTR flagging 26.85% of texts with only 1% AI edits. The issue persists across LLM polishers, with classification rates of: 34.11% (DeepSeek-V3), 39.19% (Llama3.1-70B), 44.86% (Llama3-8B), and 52.31% (Llama2-7B) for extremely minor AI-polishing. Figure 3 visualizes these AI-detection rates, with further details in Appendix C.1.

For high-stakes tasks like AI-text detection – where even a 5% FPR is very high – every detector we evaluated flagged a higher number of minimally polished, human-written texts than their unedited counterparts (Figure 3). Commercial systems are no exception: after applying only extremely minor edits with LLaMA-2-7B, the share of samples detected as AI-text jumped to 32.31% for ZeroGPT, 42.56% for Pangram, and 64.71% for GPTZero.

4.2 Most AI-text detectors fail to distinguish between minor and major polishing.

Most detectors not only flag a large portion of minor-polished texts but also struggle to differentiate between the degrees of AI involvement. For example, RoBERTa-large classifies 47.69% of minor-polished texts as AI-generated, yet its rate for major-polished texts is only slightly higher at 51.98%.

Surprisingly, some detectors mark out fewer major-polished texts than extremely minor ones, revealing a lack of sensitivity to modification extent. As shown in Figure 3, detectors like DetectGPT, FastDetectGPT, GLTR, RoBERTa-base, RoBERTa-large, and LLMDet follow this trend. FastDetectGPT, for instance, detects 10.07% of texts with 1% AI edits as AI, but only 9.59% for 75% polishing (Figure 14).



(a) GPT-4o (b) Llama2-7b

Figure 3: AI-text detection rate for degree-based AI-polished-texts (APT) by all detectors.

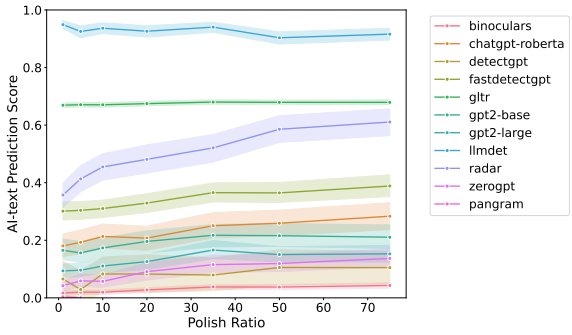


Figure 4: AI-text prediction score with 95% confidence interval for percentage-based AI-polished-texts by Llama3-8B.

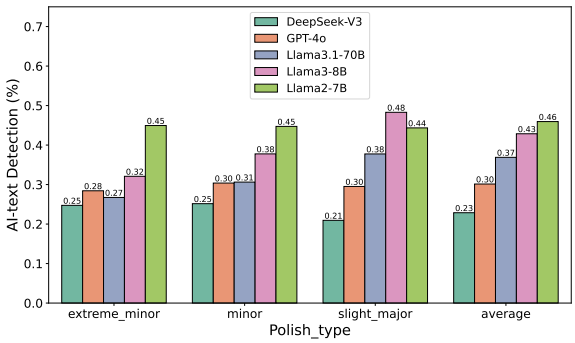


Figure 5: AI-text detection rate for degree-based AI-polished-texts from different polisher LLMs.

To provide a richer understanding of the detectors' behavior beyond binary classification, we also analyze the raw probability score. Figure 4 shows the mean prediction logits across different percents of AI-polishing by Llama3-8B, along with 95% confidence intervals. Notably, most detectors exhibit minimal variation in logits across different polishing degrees – revealing a key limitation: their inability to reliably distinguish between subtle and substantial levels of LLM-driven refinement.

4.3 Most detectors penalize more if the polisher LLM is older or smaller.

We analyze whether AI-text detectors exhibit biases across different LLMs and find a higher AI-detection rate for smaller and older models (Figure 5). For extremely minor polishing, Llama-2 has a higher AI-detection rate of 45%, while DeepSeek-V3, GPT-4o, and Llama-3 models range from 25% to 32%. The same trend was also found for our percentage-based polishing (Figure 15). Among the polishers, the latest released LLM DeepSeek-V3 (Dec 2024) gets the lowest AI-detection rate on average as 23%. The probable reason can be – with time, newer LLMs have become increasingly adept at generating human-like text, making detection more challenging over time. However, such an imbalance can create unfair scenarios, where a student using Llama-2 is flagged for minor polishing while another using Llama-3.1 is found innocent.

4.4 Some domains are more sensitive than others.

Since our HWT dataset spans six domains, we analyze detectors' AI-detection rate across them. Some domains are flagged more than others – 'speech' has the highest rate (33% – 56% for extreme-minor polishing), while 'paper_abstract' has the lowest (16% – 31%). This trend is noticed at any level of polishing. Figure 6 demonstrates the average AI-detection rate across different domains for GPT-4o polished texts. More detailed results are in Figure 17 (Appendix C.3).

Interestingly, detection rates do not always correlate with polishing levels. As shown in figure

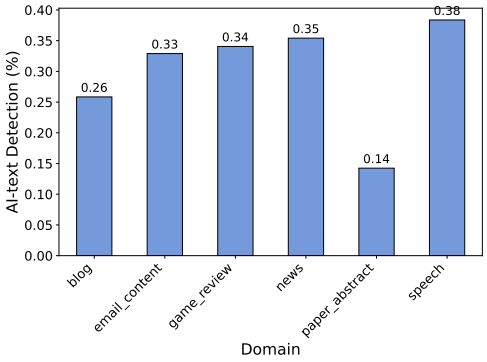


Figure 6: Average AI-detection rate for different domains (GPT-4o polishing)

18, for larger models like GPT-4o, DeepSeek-V3, and Llama3.1-70B, the AI-detection rate for 'paper_abstract' decreases as polishing increases – likely because with more freedom in polishing, they tend to generate more-human-like texts.

5 Related Work

Detecting AI-generated text is crucial as models become more human-like. Traditional methods use statistical metrics like perplexity and n-gram frequency (Gehrmann et al., 2019; Wu et al., 2023; Hans et al., 2024), while others rely on machine learning classifiers like BERT and RoBERTa (Hu et al., 2023; Guo et al., 2023; Solaiman et al., 2019). However, these approaches mainly differentiate pure AI and human-written text.

Some prior studies have explored paraphrasing as a means to evade AI detectors (Sadasivan et al., 2023; Krishna et al., 2023), but they do not specifically evaluate the unreliability of detection models in AI-polished text scenarios, which we address. Other works (Dugan et al., 2023; Zeng et al., 2024) focus on detecting the boundary between HWT and AI-generated text, treating the sentences as distinct entities. Verma et al. (2023) explores how perturbations to AI-text can affect detection outcomes. More recent research (Gao et al., 2024; Yang et al., 2024) has investigated LLM-assisted text polishing, but without considering varied degrees of AI involvement. We extend this by systematically analyzing AI-polished text across multiple levels, assessing detection limitations.

6 Discussion

The findings of this study reveal several critical limitations in current AI-text detectors, particularly in distinguishing between human-authored content that has been subtly refined by AI and fully machine-generated text. A key concern is the high false positive rate associated with minimally polished text. Many detectors classify such lightly edited content as AI-generated, which poses serious risks of unjust accusations of plagiarism or academic dishonesty.

To address this, we recommend moving beyond binary classification frameworks and adopting tiered or probabilistic labeling schemes that reflect varying degrees of AI involvement. A promising direction for future work is to train detectors not only on purely human-written and fully AI-generated texts but also on AI-polished samples. We hope our released APT-Eval dataset will serve as a valuable resource for developing and evaluating such models. Furthermore, rather than producing a definitive label, detectors should output prediction probabilities, enabling users to better interpret and trust the system's verdict.

We also observe detector biases against older or smaller LLMs, and recommend training on a more diverse range of LLM outputs—including from newer, more human-like models—to improve fairness. Domain-specific biases can also be mitigated by fine-tuning detectors for particular genres or text types. Also, in the early stage of the pipeline, there can be another model to find the text-domain that will trigger a specific detector accordingly.

Lastly, we emphasize the importance of interpretability and human oversight in detection tools. Developing interpretable detectors that can highlight suspicious segments or stylistic anomalies will empower users to make informed decisions. In high-stakes scenarios, integrating human-in-the-loop review mechanisms can further enhance the reliability and fairness of the process. Ultimately, addressing these challenges requires a multi-faceted approach that balances technical sophistication with transparency, fairness, and adaptability.

7 Conclusion

Our study exposes key flaws in AI-text detectors when handling AI-polished text, showing high false positive rates and difficulty distinguishing minor from major AI refinements. Detectors also exhibit biases against older or smaller models, raising fairness concerns. We highlight the need for more nuanced detection methods and release our APT-Eval dataset to support further research.

Limitations

While our study provides valuable insights into the challenges of AI-text detection for AI-polished texts, several limitations should be acknowledged. First, our dataset, APT-Eval, is built using a specific set of LLMs (GPT-4o, Llama3-70B, etc.), which may not fully represent the diversity of AI models available. Future research should explore a broader range of models to assess generalizability. Additionally, while our dataset spans six distinct domains of human-written text (HWT), incorporating more domains could provide a more comprehensive evaluation of AI-text detection across different writing contexts.

Second, our findings highlight biases in detection models, particularly against smaller or older LLMs, but further investigation is needed to understand the root causes of these biases. Moreover, while this study focuses on identifying limitations in current AI-text detection systems, the development of more nuanced, fine-grained detection frameworks remains an open challenge. Future work should explore adaptive AI-text detectors capable of distinguishing varying levels of AI involvement, ensuring both accuracy and fairness in AI-assisted writing evaluation.

Acknowledgments

This project was supported in part by a grant from an NSF CAREER AWARD 1942230, ONR YIP award N00014-22-1-2271, ARO's Early Career Program Award 310902-00001, Army Grant No. W911NF2120076, the NSF award CCF2212458, NSF Award No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS), a MURI grant 14262683, an award from meta 314593-00001 and an award from Capital One.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

The Data Beast. 2021. Ted talk transcripts (2006-2021) dataset. urlhttps://www.kaggle.com/datasets/thedatabeast/ted-talk-transcripts-2006-2021.

Enron Corporation. 2004. Enron email dataset. urlhttps://www.cs.cmu.edu/ enron/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771.

Ashkan Farhangi, Ning Sui, Nan Hua, Haiyan Bai, Arthur Huang, and Zhishan Guo. 2022. Protoformer: Embedding prototypes for transformers. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 447–458. Springer.

Chujie Gao, Dongping Chen, Qihui Zhang, Yue Huang, Yao Wan, and Lichao Sun. 2024. Llm-as-a-coauthor: The challenges of detecting llm-human mixcase. *arXiv preprint arXiv:2401.05952*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Najzeko. 2021. Steam reviews 2021 dataset. urlhttps://www.kaggle.com/datasets/najzeko/steam-reviews-2021.

OpenAI. 2019. Gpt-2 output dataset (detector). urlhttps://github.com/openai/gpt-2-output-dataset/tree/master/detector.

OpenAI. 2023. Gpt-4 research. urlhttps://openai.com/index/gpt-4-research/.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A third party large language models generated text detection tool. *arXiv preprint arXiv:2305.15004*.

Lingyi Yang, Feng Jiang, Haizhou Li, et al. 2024. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text. *APSIPA Transactions on Signal and Information Processing*, 13(2).

Zijie Zeng, Lele Sha, Yuheng Li, Kaixun Yang, Dragan Gašević, and Guangliang Chen. 2024. Towards automatic boundary detection for human-ai collaborative hybrid essay in education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22502–22510.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.

A Dataset

A.1 Dataset Details

Polish Type	Polish Operation	Polisher LLM					Total
		GPT-4o	Llama3.1 70B	Llama3 8B	Llama2 7B	DeepSeek V3	
	no-polish pure HWT	-	-	-	-	-	300
Degree-based	**extreme-minor**	297	297	293	196	300	1383
	minor	293	294	286	188	297	1358
	slight-major	285	282	280	182	293	1322
	major	277	212	266	178	251	1184
Percentage-based	**1%**	299	299	294	193	265	1350
	5%	298	299	283	173	244	1297
	10%	297	298	285	176	225	1281
	20%	297	296	277	177	257	1304
	35%	295	294	287	199	247	1322
	50%	293	285	279	176	273	1306
	75%	293	277	272	188	276	1306
Total		3224	3133	3102	2026	2928	**14713**

Table 2: Details of our APT-Eval dataset

Domain	Year	Source
Blog Content	2006	Blog (Schler et al., 2006)
Email Content	2015	Enron email dataset (Corporation, 2004)
News Content	2006	BBC news (Greene and Cunningham, 2006)
Game Reviews	2021	Steam reviews (Najzeko, 2021)
Paper Abstract	2022	ArXiv-10 (Farhangi et al., 2022)
Speech Content	2021	Ted Talk (Beast, 2021)

Table 3: Details on our 'no-polish-HWT' set of our Dataset.

A.2 Prompts to Polisher LLM

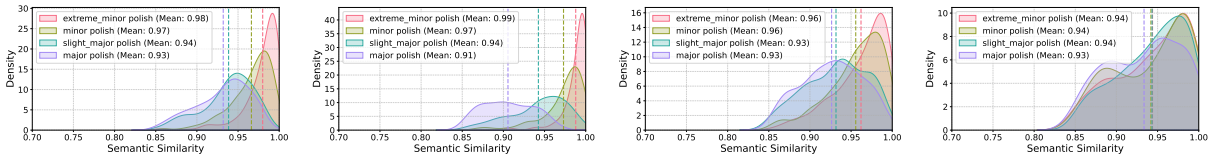
System Prompt: You are a helpful chatbot who always responds with helpful information. You are asked to provide a polished version of the following text. Only generate the polished text.
User Prompt: Polish the given original text below with `{polish_type}` polishing. The difference between original and polished text must be `{polish_type}`. The semantic meaning of polished text must be the same as original text. The given original text: `{original_text}`

Figure 7: Prompt for the Degree-based AI-polishing

System Prompt: You are a helpful chatbot who always responds with helpful information. You are asked to provide a polished version of the following text. Only generate the polished text.
User Prompt: Polish the given text below. The text has a total of `{text_length}` words. Make sure that you edit exactly `{polish_word_limit}` words. Do not change or polish more than `{polish_word_limit}` words. Also, make sure that the semantic meaning does not change with polishing. Only output the polished text, nothing else. The given text: `{original_text}`

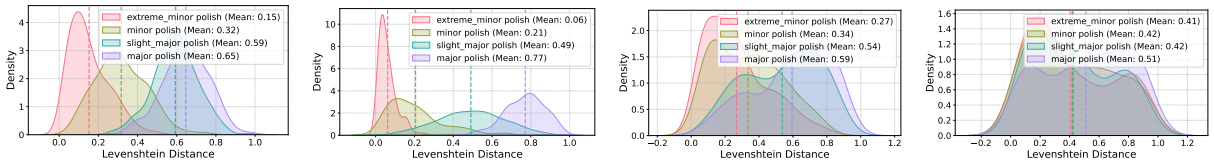
Figure 8: Prompt for the Percentage-based AI-polishing

A.3 Dataset Analysis



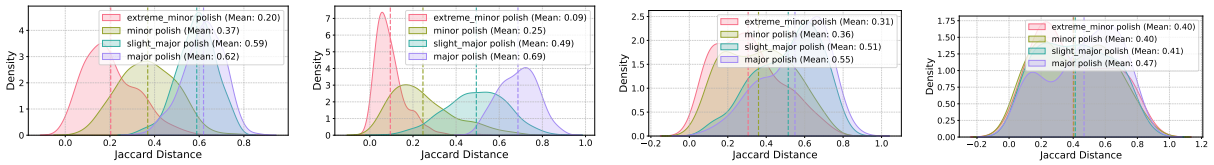
(a) GPT-4o (b) Llama3.1-70b (c) Llama3-8b (d) Llama2-7b

Distribution of Semantic Similarity



(e) GPT-4o (f) Llama3.1-70b (g) Llama3-8b (h) Llama2-7b

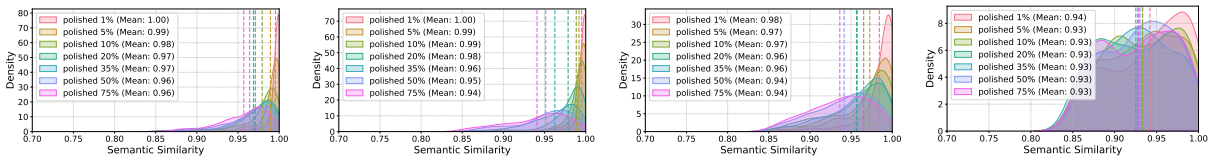
Distribution of Levenshtein Distance



(i) GPT-4o (j) Llama3.1-70b (k) Llama3-8b (l) Llama2-7b

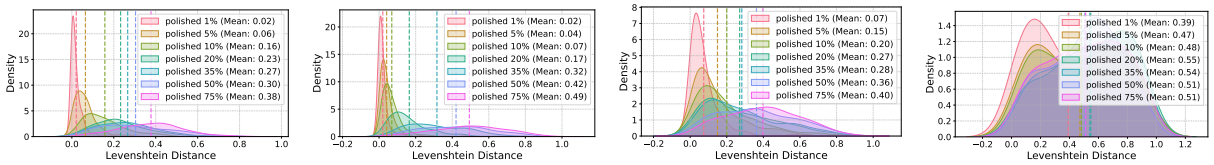
Distribution of Jaccard Distance

Figure 9: Distribution of cosine semantic similarity, levenshtein distance, and jaccard distance for **degree-based** AI-polished texts by different polisher.



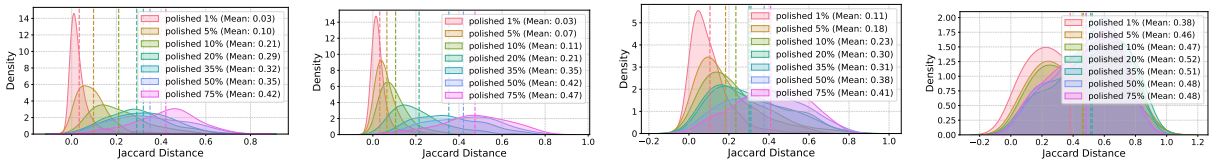
(a) GPT-4o (b) Llama3.1-70b (c) Llama3-8b (d) Llama2-7b

Distribution of Semantic Similarity



(e) GPT-4o (f) Llama3.1-70b (g) Llama3-8b (h) Llama2-7b

Distribution of Levenshtein Distance



(i) GPT-4o (j) Llama3.1-70b (k) Llama3-8b (l) Llama2-7b

Distribution of Jaccard Distance

Figure 10: Distribution of cosine semantic similarity, levenshtein distance, and jaccard distance for **percentage-based** AI-polished texts by different polisher.

Polish Type	GPT-4o			Llama3.1-70B		
	Mean Semantic Similarity	Mean Levensthein Distance	Mean Jaccard Distance	Mean Semantic Similarity	Mean Levensthein Distance	Mean Jaccard Distance
extreme-minor	0.98	0.15	0.20	0.99	0.06	0.09
minor	0.97	0.35	0.37	0.97	0.21	0.25
slight-major	0.94	0.59	0.59	0.94	0.49	0.49
major	0.93	0.65	0.62	0.91	0.77	0.69

Table 4: Similarity and Distance between original HWT and AI-polished texts (degree-based)

Polish %	GPT-4o			Llama3.1-70B		
	Mean Semantic Similarity	Mean Levensthein Distance	Mean Jaccard Distance	Mean Semantic Similarity	Mean Levensthein Distance	Mean Jaccard Distance
1	1.00	0.02	0.03	1.00	0.02	0.03
5	0.99	0.06	0.10	0.99	0.04	0.07
10	0.98	0.16	0.21	0.99	0.07	0.11
20	0.97	0.23	0.29	0.98	0.17	0.21
35	0.97	0.27	0.32	0.96	0.32	0.35
50	0.96	0.30	0.35	0.95	0.42	0.42
75	0.96	0.38	0.42	0.94	0.49	0.47

Table 5: Similarity and Distance between original HWT and AI-polished texts (percentage-based)

A.4 Difference Between APT-Eval and MixSet Dataset

While our APT-Eval dataset builds upon the human-written portion of the MixSet dataset, it is fundamentally different in its objectives and design. MixSet includes AI-polished text using token- and sentence-level paraphrasing, but it does not control or quantify the degree of AI involvement in the polishing process. In contrast, our APT-Eval dataset systematically explores varying levels of AI assistance – both in terms of degree-based (e.g., extremely minor to major edits) and percentage-based modifications. Our goal is to introduce a fine-grained continuum of AI involvement, allowing us to rigorously evaluate how current detectors respond to subtle versus substantial AI polish, rather than just simple polish.

A.5 Samples of our APT-Eval Dataset

```
1 Another week just flew by me. I drove by the new house yesterday and the front window was
  covered in boxes. We may get the keys at closing, i am getting so excited. If all goes well this
  week, we will be set for closing a week from wednesday. I was up until after 3:00am last night
  watching movies with bethany. There has been a lot of movie watching going on around here
  lately. Here's my rundown; last night we watched thirteen, a movie every mother of a teen should
  see. It was better then i expected. Then we watched gigli, yeah it was just as bad as the media
  said it was, but we had to see for ourselves.
```

(a) Original HWT

```
1 Another week just flew by me. I drove by the new house yesterday, and the front window was
  covered in boxes. We may get the keys at closing, and I am getting so excited. If all goes well
  this week, we will be set for closing a week from Wednesday. I was up until after 3:00 AM last
  night watching movies with Bethany. There has been a lot of movie watching going on around here
  lately. Here's my rundown: last night, we watched *Thirteen*, a movie every mother of a teen
  should see. It was better than I expected. Then we watched *Gigli*; yeah, it was just as bad as
  the media said it was, but we had to see for ourselves.
```

(b) GPT-4o Polished

```
1 Another week just flew by me. I drove by the new house yesterday and the front window was
  covered in boxes. We may get the keys at closing; I am getting so excited. If all goes well this
  week, we will be set for closing a week from Wednesday. I was up until after 3:00 am last night
  watching movies with Bethany. There has been a lot of movie watching going on around here
  lately. Here's my rundown: last night we watched Thirteen, a movie every mother of a teen should
  see. It was better than I expected. Then we watched Gigli; yeah, it was just as bad as the media
  said it was, but we had to see for ourselves.
```

(c) Llama3.1-70b Polished

```
1 Another week flew by me. I drove by the new house yesterday and the front window was covered in
  boxes. We may get the keys at closing, I am getting so excited. If everything goes well this
  week, we will be set for closing a week from Wednesday. I was up until after 3 00 am last night
  watching movies with Bethany. There has been a lot of movie watching going on around here
  lately. Here's my rundown last night we watched thirteen, a movie every mother of a teenager
  should see. It was better than I expected. Then we watched Gigli, yes, it was just as bad as the
  media said it was, but we had to see for ourselves.
```

(d) Llama3-8b Polished

```
1 Another week just flew by. I drove by the new house yesterday, and the front window was covered
  in boxes. We may get the keys at closing\u2014I'm getting so excited. If all goes well this
  week, we'll be set for closing a week from Wednesday. I was up until after 3:00 AM last night
  watching movies with Bethany. There's been a lot of movie-watching going on around here lately.
  Here's my rundown: last night we watched *Thirteen*, a movie every mother of a teen should see.
  It was better than I expected. Then we watched *Gigli*\u2014yeah, it was just as bad as the
  media said it was, but we had to see for ourselves.
```

(e) DeepSeek-V3 Polished

Figure 11: Example sample from our APT-Eval dataset, where the original HWT is polished (extreme-minor) by different polisher LLMs.

1 I experienced no performance problems even at release, and after all these patches it plays really well. Imo, this is a really great and fun game in truth, although the format may be somewhat annoying (where the free-roam city area repopulates with new enemies every time you progress the story). I particularly enjoyed the 'dual play' segments (where you can play batman + 1 more), and the ability to use the controller for most segments of the game but the mouse and keyboard for the tank combat segments was greatly appreciated. Driving your batmobile around the free-roam city, including when chasing the apc commanders, is actually quite enjoyable once you figure out how to use the breaks to skid around corners (and when you master that, driving around the city is really quite enjoyable!)

(a) Original HWT

1 I experienced no performance problems even at release, and after all these patches, it plays really well. In my opinion, this is a really great and fun game, although the format may be somewhat annoying (where the free-roam city area repopulates with new enemies every time you progress the story). I particularly enjoyed the 'dual play' segments (where you can play as Batman plus one more), and the ability to use the controller for most segments of the game, but the mouse and keyboard for the tank combat segments, was greatly appreciated. Driving your Batmobile around the free-roam city, including when chasing the APC commanders, is actually quite enjoyable once you figure out how to use the brakes to skid around corners (and when you master that, driving around the city is really quite enjoyable!)

(b) Extreme-minor Polishing

1 I experienced no performance issues even at release, and after all these patches, it plays really well. In my opinion, this is a really great and fun game, although the format may be somewhat annoying, where the free-roam city area repopulates with new enemies every time you progress the story. I particularly enjoyed the 'dual play' segments, where you can play as Batman and another character, and the ability to use the controller for most segments of the game, but the mouse and keyboard for the tank combat segments, was greatly appreciated. Driving the Batmobile around the free-roam city, including when chasing the APC commanders, is actually quite enjoyable once you figure out how to use the brakes to skid around corners, and when you master that, driving around the city is really quite enjoyable.

(c) Minor Polishing

1 I encountered no performance issues at launch, and with the numerous patches, the game runs smoothly. In my opinion, this is a truly great and enjoyable game, despite the somewhat frustrating format where the free-roam city area repopulates with new enemies upon progressing through the story. I particularly enjoyed the 'dual play' segments, which allow you to play as Batman alongside another character. The option to use a controller for most of the game, while switching to mouse and keyboard for tank combat segments, was a welcome feature. Driving the Batmobile around the free-roam city, including during APC commander chases, is quite enjoyable once you master the use of brakes to skid around corners, making navigation through the city a truly enjoyable experience.

(d) Slight-major Polishing

1 From my experience, the game has consistently delivered a seamless performance, even from its initial release, and has only continued to improve with each subsequent patch. In my opinion, this is an exceptional and engaging title that truly lives up to its potential, despite some minor drawbacks in its design. Specifically, the game's tendency to repopulate the open-world city area with new enemies after each story progression can be somewhat frustrating. However, this is more than offset by the game's numerous strengths, including the innovative 'dual play' feature, which allows players to control two characters simultaneously, and the flexibility to switch between controller and mouse-and-keyboard controls for different segments of the game, such as the tank combat sequences. One of the most enjoyable aspects of the game is the opportunity to explore the open-world city in the Batmobile, particularly when engaging in high-speed pursuits of APC commanders. Mastering the art of using the brakes to execute precision skids around corners adds a thrilling layer of complexity to the driving experience, making it a truly exhilarating aspect of the game.

(e) Major Polishing

Figure 12: Example sample from our APT-Eval dataset, where the original HWT is polished with different degrees by Llama3.1-70B model.

B AI-text Detectors

We developed our code-base on the framework of RAID (Dugan et al., 2024), and kept the hyperparameters for all detectors the same as RAID for a fair evaluation.

Additionally, we identify the threshold corresponding to a 5% false positive rate (FPR) – or the lowest possible FPR if it exceeds 5%. We notice that – for most detectors, the thresholds for 'best accuracy' and '5% FPR' do not vary much. Moreover, since our primary focus is on misclassification rates for both minor-polished and major-polished texts, optimizing the threshold for overall accuracy is more appropriate than minimizing FPR alone.

Table 6 shows the threshold that we found by optimizing the accuracy, and used for the evaluation of our APT-Eval dataset.

	Detector	Threshold	Accuracy	FPR
Model-Based	**RADAR**	0.8989	0.8017	0.082
	RoBERTa (ChatGPT)	0.333	0.8617	0.0217
	RoBERTa-Base (GPT2)	0.091	0.7917	0.06
	RoBERTa-Large (GPT2)	0.0408	0.8	0.0817
Metric-Based	**GLTR**	0.7038	0.845	0.0683
	DetectGPT	0.355	0.725	0.085
	Fast-DetectGPT	0.778	0.8317	0.06
	LLMDet	0.9798	0.605	0.2383
	Binoculars	0.1075	**0.88**	*0.018*
Commercial	**ZeroGPT**	0.2525	0.8067	0.0367
	GPTZero	0.03	0.862	0.075
	Pangram	0.01	*0.875*	**0.00**

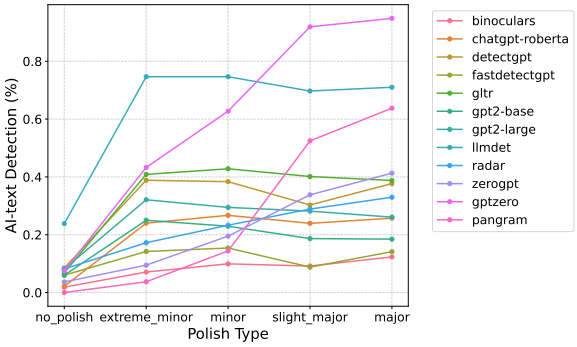
Table 6: Detector-based Threshold, Accuracy, and False Positive Rate. The best performance is in bold, and the second best is in italics.

For computational resources, we employed:

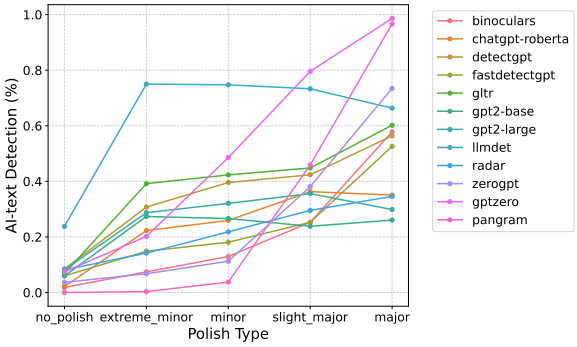
- One NVIDIA RTX A5000 GPU for running model-based and metric-based detectors.
- One NVIDIA RTX A6000 GPU for the Binoculars detector.
- API subscriptions for ZeroGPT, GPTZero and Pangram

C Results and Findings

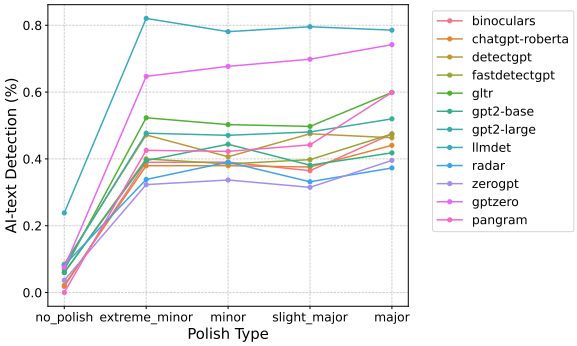
C.1 Results for All Detectors



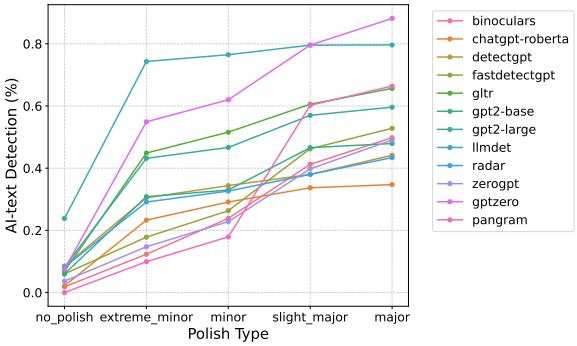
(a) GPT-4o



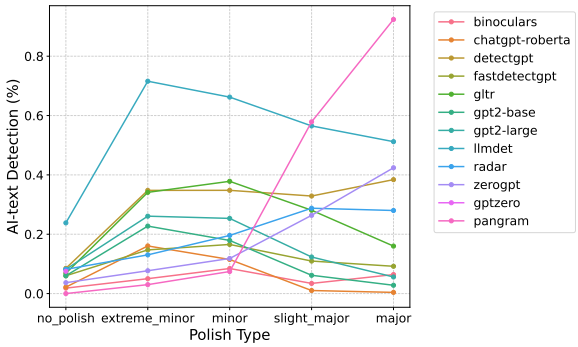
(b) Llama3.1-70b



(c) Llama2-7B

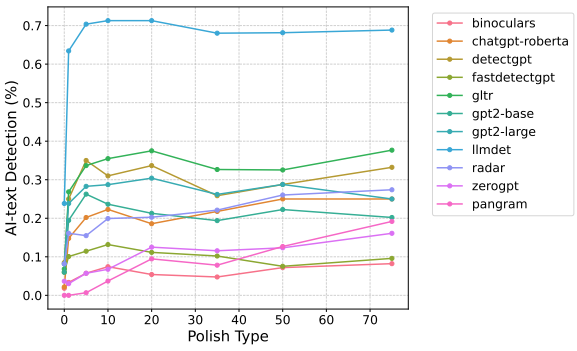


(d) Llama3-8B

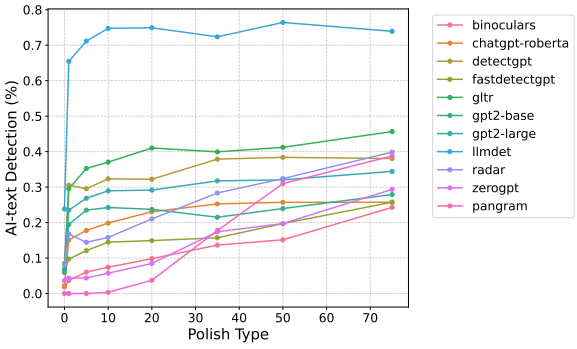


(e) DeepSeek-V3

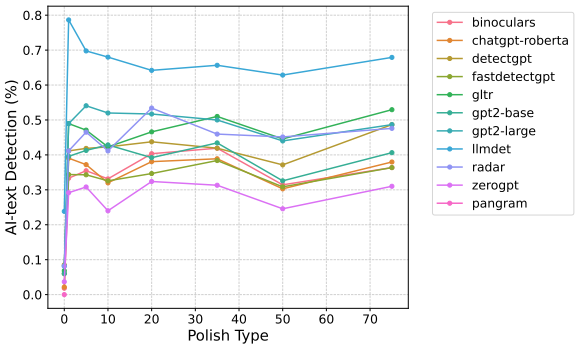
Figure 13: AI-text detection rate for **degree-based** AI-polished-texts (APT) by all detectors.



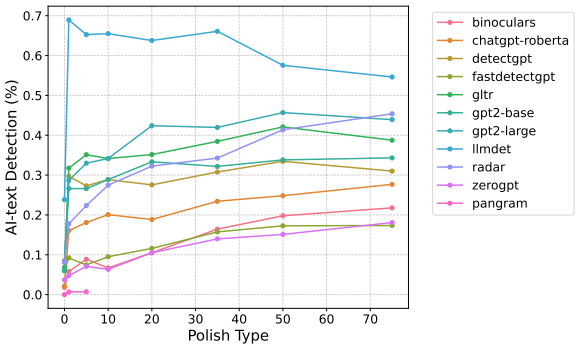
(a) GPT-4o



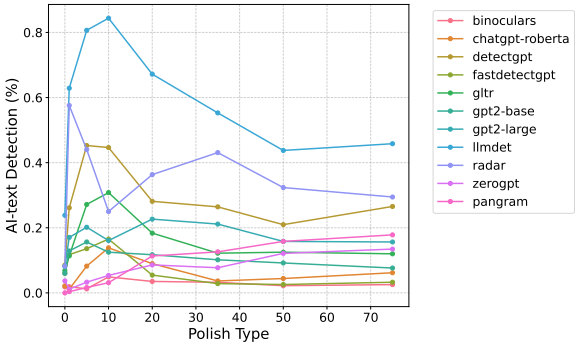
(b) Llama3.1-70b



(c) Llama2-7B



(d) Llama3-8B



(e) DeepSeek-V3

Figure 14: AI-text detection rate for **percentage-based** AI-polished-texts (APT) by all detectors.

C.2 Polisher LLM Specific Results

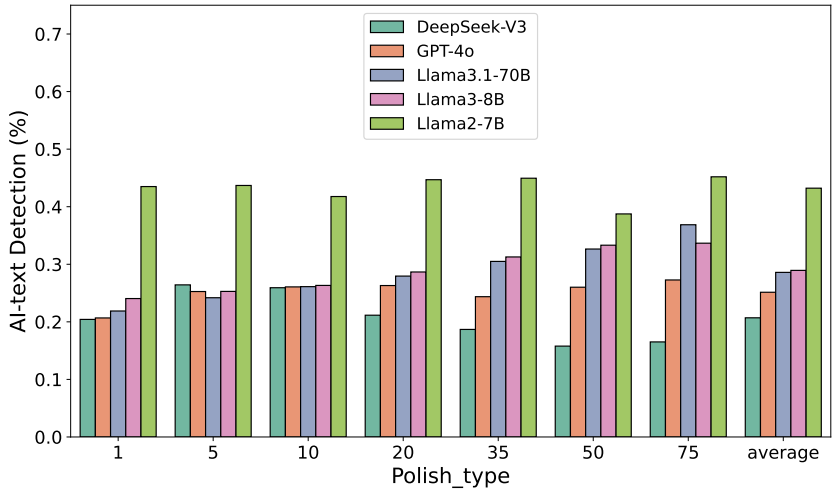


Figure 15: AI-text detection rate for **percentage-based** AI-polished-texts from different polisher LLMs.

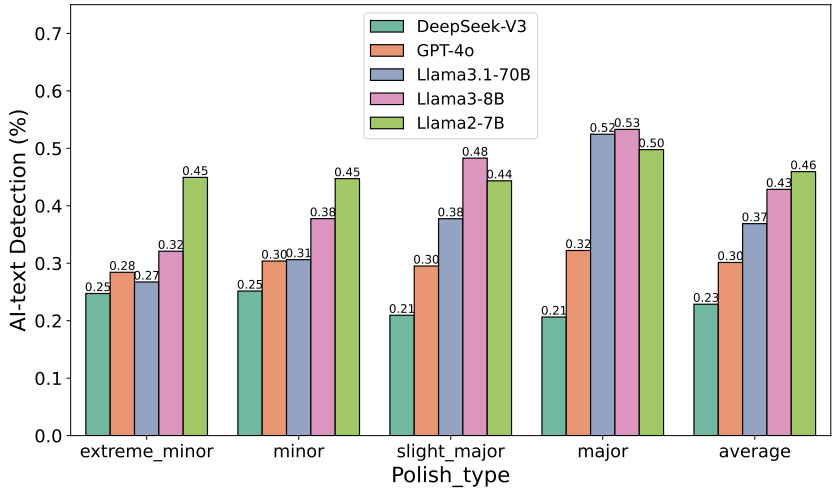
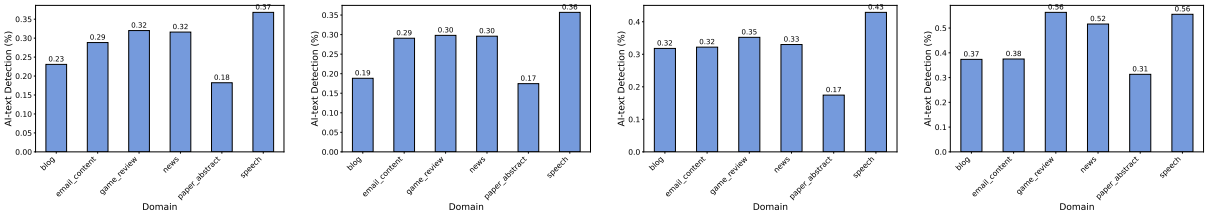


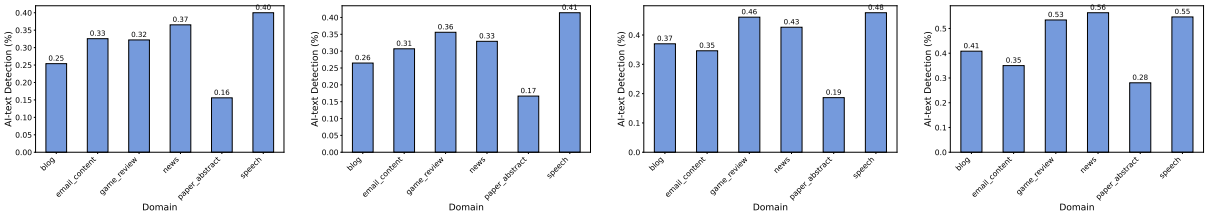
Figure 16: AI-text detection rate for **degree-based** AI-polished-texts from different polisher LLMs.

C.3 Domain Specific Results



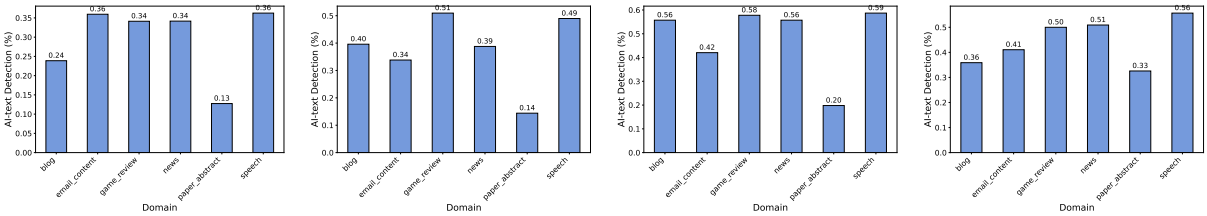
(a) GPT-4o (b) Llama3.1-70b (c) Llama3-8b (d) Llama2-7b

Extreme-minor Polishing



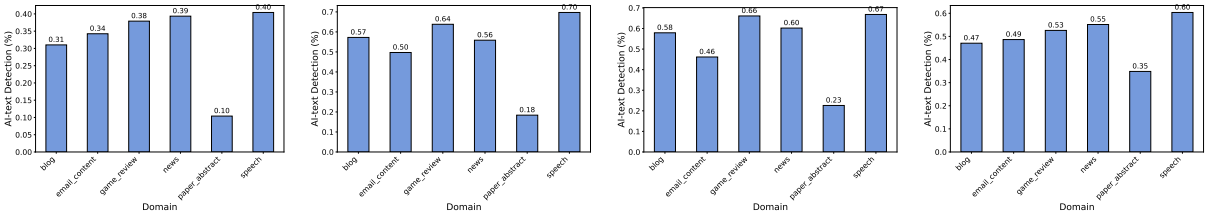
(e) GPT-4o (f) Llama3.1-70b (g) Llama3-8b (h) Llama2-7b

Minor Polishing



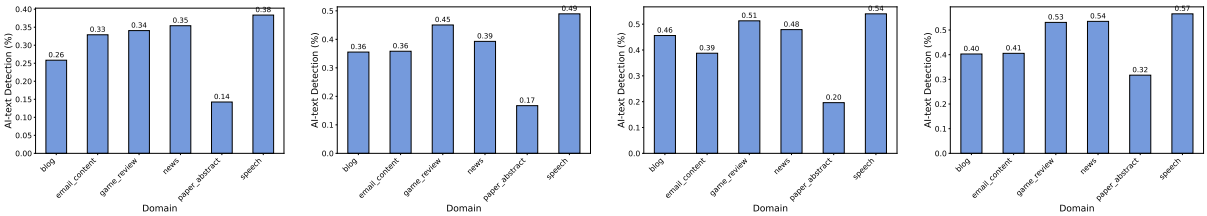
(i) GPT-4o (j) Llama3.1-70b (k) Llama3-8b (l) Llama2-7b

Slight-major Polishing



(m) GPT-4o (n) Llama3.1-70b (o) Llama3-8b (p) Llama2-7b

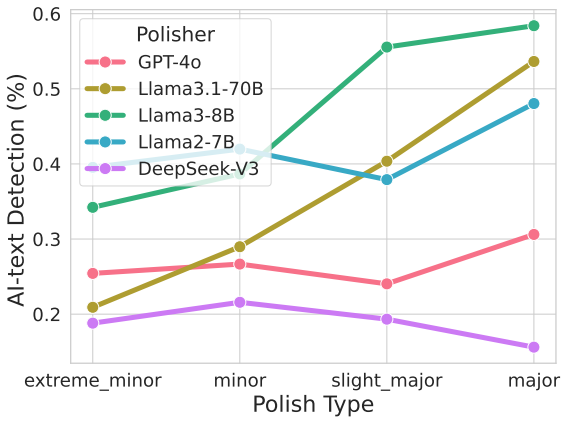
Major Polishing



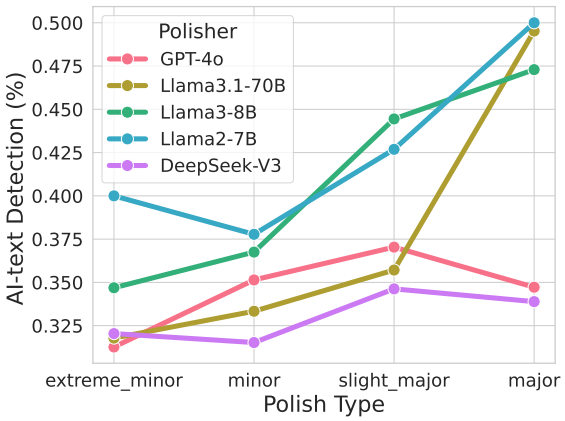
(q) GPT-4o (r) Llama3.1-70b (s) Llama3-8b (t) Llama2-7b

Average

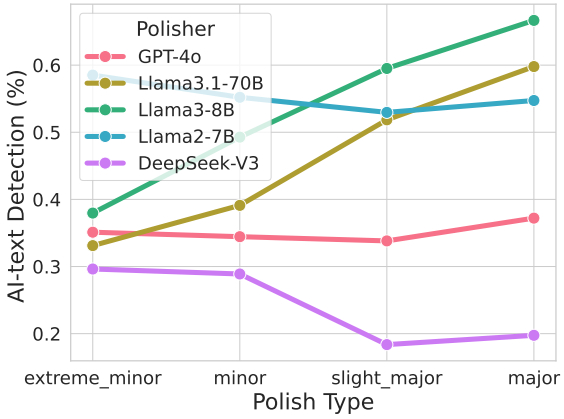
Figure 17: AI-text detection rate across all domains for different degree-based polishing.



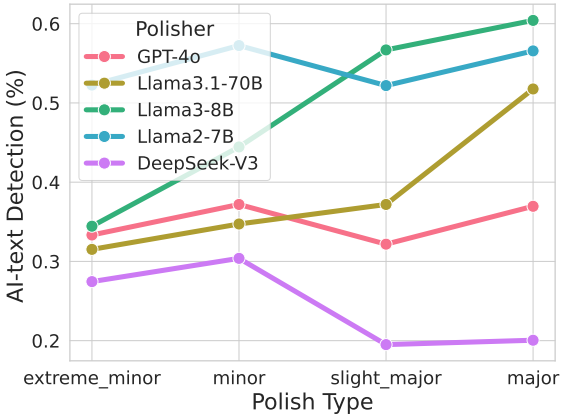
(a) Blog



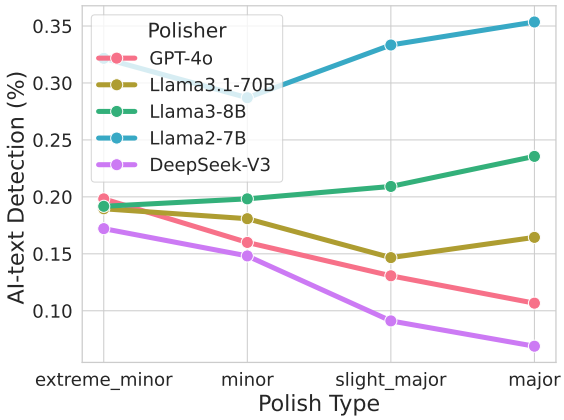
(b) Email content



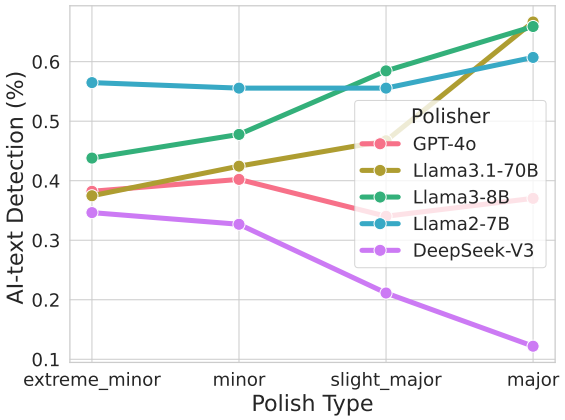
(c) Game review



(d) News



(e) Paper Abstract



(f) Speech

Figure 18: Average AI-text detection for different domains across all polisher LLMs.