MAGA-Bench: Machine-Augment-Generated Text via Alignment Detection Benchmark

Anonymous ACL submission

Abstract

Large Language Models (LLMs) alignment is constantly evolving. Machine-Generated Text (MGT) is becoming increasingly difficult to distinguish from Human-Written Text (HWT). This has exacerbated abuse issues such as fake news and online fraud. Fine-tuned detectors' generalization ability is highly dependent on dataset quality, and simply expanding the sources of MGT is insufficient. Further augment of generation process is required. According to HC-Var's theory, enhancing the alignment of generated text can not only facilitate attacks on existing detectors to test their robustness, but also help improve the generalization ability of detectors fine-tuned on it. Therefore, we propose **M**achine-**A**ugment-**G**enerated Text via **A**lignment (MAGA). MAGA's pipeline achieves comprehensive alignment from prompt construction to reasoning process, among which **R**einforced **L**earning from **D**etectors **F**eedback (RLDF), systematically proposed by us, serves as a key component. In our experiments, the RoBERTa detector fine-tuned on MAGA training set achieved an average improvement of 2.06% in generalization detection accuracy. MAGA test set caused an average decrease of 5.63% in the accuracy of the selected detectors, expecting to provide indicative significance for future research on the generalization detection ability of detectors.

1 Introduction

In recent years, LLMs have advanced rapidly, with remarkable improvements in their language comprehension capabilities, precision in instruction execution, and ability to generate complex text for fulfilling diverse tasks (Brown et al., 2020; Ouyang et al., 2022). Human can hardly distinguish between MGT and HWT (Chang et al., 2024). This powerful capability poses substantial abuse risks. (Hanley and Durumeric, 2024) notes that in 2022, the proportion of AI-generated news on websites notorious for disseminating false rumors surged by approximately 457%. Such misinformation is not only used in fake news (Zellers et al., 2019; Dugan et al., 2023a) but also utilized by numerous bad actors for various unethical or even illegal activities, including online fraud (Weidinger et al., 2021; Ayoobi et al., 2023) and academic misconduct (Stokel-Walker, 2022; Kasneci et al., 2023). Therefore, to address and mitigate its potential risks in information dissemination, the development of efficient and robust MGT detectors is of paramount importance.

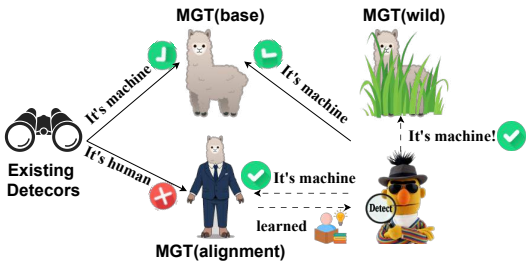

Figure 1: More aligned MGT not only evades detection by existing detectors but also facilitates fine-tuning of neural-based detectors, which enhancing their generalization capability for wild detection.

Toward this need, many exciting detection methods and benchmarks have emerged, each with pros and cons (Crothers et al., 2023). Previous datasets boosted detection difficulty via overall distribution or data source diversity (first peaking at M4 (Wang et al., 2023b)), but their single text generation was simple—relying on few fixed prompt templates. RAID (Dugan et al., 2024) is the first benchmark highly prioritizing detector robustness. Its attacks (e.g., random adding '\n', deleting articles like "a"/"an"/"the") raise MGT's Perplexity (PPL) and successfully attack metric-based detectors. Binoculars (Hans et al., 2024) noted their near-zero accuracy for extremely

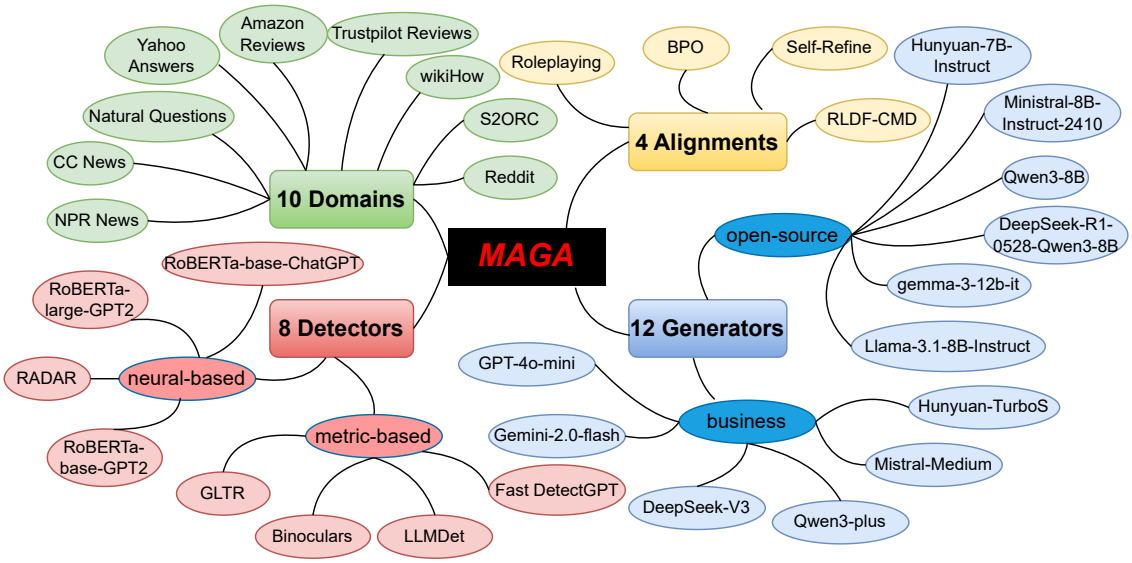

Figure 2: Our construction process is highly balanced. For 10 domains and 12 generators, each domain is generated by 12 generators, and each generator produces data for 10 domains (Detailed statistical descriptions are provided in Appendix A). We evaluate selected detectors on our dataset.

random MGT. However, this text manipulation damages semantics, hindering fine-tuning of neural-based detectors represented by RoBERTa (Liu et al., 2019). Its value lies more in testing robustness to specific attacks than advancing generalized detection.

HC-Var (Xu et al., 2023) notes that more aligned MGT is closer to HWT in relevant features, helping neural-based detectors learn better decision boundaries and boost generalized detection. Thus, while ensuring broad data sources (10 domains, 12 generators, various decoding strategies), we augment the generation process. The augment is based on MAGA's pipeline (combining 4 alignment methods centered on RLDF) to make MGT more aligned. For comparison, we also built an unenhanced dataset, Machine-Generated-Text-base (MGB). We tested generalization by using RoBERTa fine-tuned on our training set to detect on multiple influential datasets proposed by previous works. We find our method achieves promising attack performance against existing detectors, whether neural-based or metric-based ones, and that RoBERTa fine-tuned on MAGA also shows notably improved generalization ability.

2 Related Work

Table ?? presents a comparison of publicly available generated text sources, among which Li et al. (2024)—which built a dataset of 447k generations from 7 language model families across 10 domains to study detector robustness—is the most relevant work. Other resources usually focus on specific sub-areas, such as multilingual text (Macko et al., 2023; Wang et al., 2023b), code (Wang et al., 2023a), question-answering (Guo et al., 2023; Xu et al., 2023; Su et al., 2024), and scientific papers (Mosca et al., 2023); shared tasks like AuTextTification (Sarvazyan et al., 2023b) and RuATD (Shamardina et al., 2022) have also provided datasets and driven centralized evaluation and competition. However, most of these shared resources, despite covering multiple generative models and domains, rarely include adversarial attacks and none account for variations in decoding strategy (often even failing to specify the strategy used), resulting in insufficiently challenging datasets that inflate reports of detector accuracy. Another approach to evaluating robustness is small-scale comparative studies: by varying one aspect of generated text and comparing detector accuracy across these variations, such studies have revealed detectors' lack of robustness to unseen generative models (Stiff and Johansson, 2022; Pu et al., 2023b; Chakraborty et al., 2023), domains (Pagnoni et al., 2022; Pu et al., 2023a; Rodriguez et al., 2022), decoding strategies (Ippolito et al., 2020; Solaiman et al., 2019), prompts (Koike et al., 2023; Kumarage et al., 2023; Lu et al., 2023), repetition penalties (Fishchuk

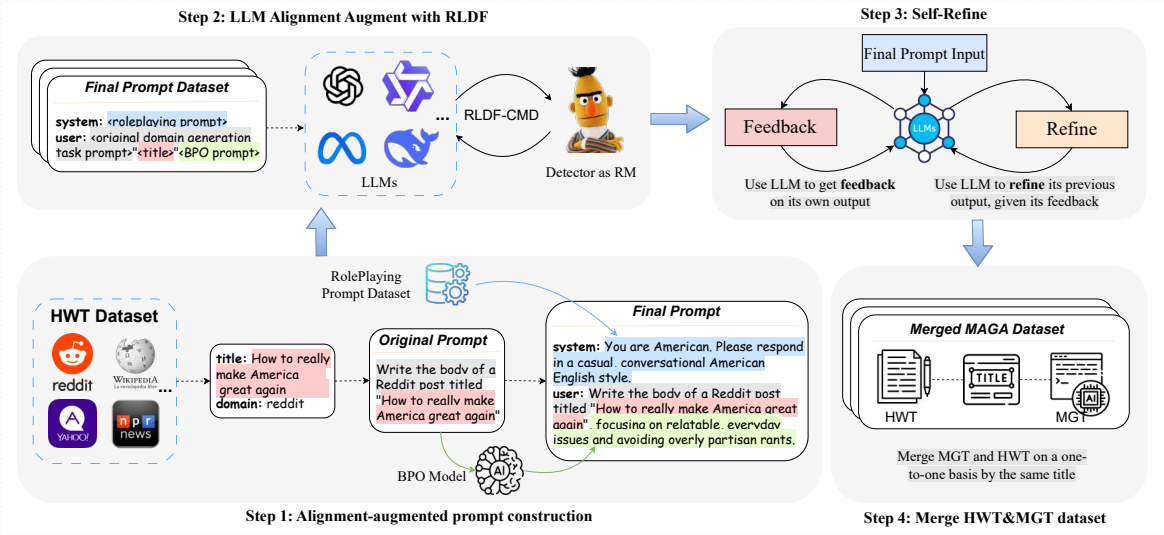

Figure 3: MAGA Pipeline

and Braun, 2023), and human edits (Gao et al., 2024). Meanwhile, work focusing on adversarial robustness has shown detectors are vulnerable to homoglyph attacks (Gagiano et al., 2021; Wolff, 2020; Macko et al., 2024), whitespace insertion (Cai and Cui, 2023), sentiment/factual alterations (Bhat and Parthasarathy, 2020), paraphrase attacks (Krishna et al., 2023; Sadasivan et al., 2023), and synonym replacement (Kulkarni et al., 2023; Pu et al., 2023a). The present work builds on this foundation, synthesizing multiple elements of robustness literature into a single systematic benchmark study.

3 The MAGA Dataset

3.1 Overview

Figure 2 presents the components of the MAGA dataset. To create MAGA, we first sampled 72k human texts (with titles) in total uniformly from 10 target domains (§3.2), splitting them into a 60k training set and a 12k test set. For each human text, we generated a matching machine text using its title. We used 12 generators((§3.3). We first created an unenhanced MGB dataset (train + test) for comparison, generated via the original prompt template ((§3.4); full original prompt in Appendix A). MAGA' s augment involves four methods (§3.4): roleplaying, BPO, self-refine, and RLDF-CMD (§3.5). The MAGA pipeline, integrating these four methods, was applied to the MAGA dataset (train + test). For roleplaying, BPO, and self-refine, we separately provided MAGA-extra for additional indicative testing (Appendix A). Full statistics are in Appendix A.

3.2 Domains

Various domains raise detection difficulty (Li et al., 2024; Dugan et al., 2023b) and improve the fine-tuned RoBERTa' s performance (Wang et al., 2024). Our sources cover social conversations (Reddit), summarization (S2ORC), general knowledge (Wikipedia, wikiHow), reviews (Trustpilot, Amazon Reviews), direct Q&A (Yahoo Answers, Natural Questions), and news creation (CC News, NPR News), which challenge various capabilities of LLM. To boost generalization, we prioritized domain diversity in selection. For each capability, we chose two sources where possible: one more targeted and conversational (e.g., Yahoo Answers), and the other more general and formal (e.g., Natural Questions). Details about the HWT sources of these domains is provided in Appendix A.

3.3 Generators

Compared with previous work, we further expanded the diversity of generator LLM selections. Instead of choosing multiple models of different parameter sizes from the same series, we prioritized enriching model series variety, including GPT, Llama, Gemini/gemma, DeepSeek, Qwen, Mistral/Ministral, and Hunyuan. Following (Sarvazyan et al., 2023a), we first selected the largest (commercial) model in each series as possible, then chose models with smaller parameter sizes. We also ensured selected models are as

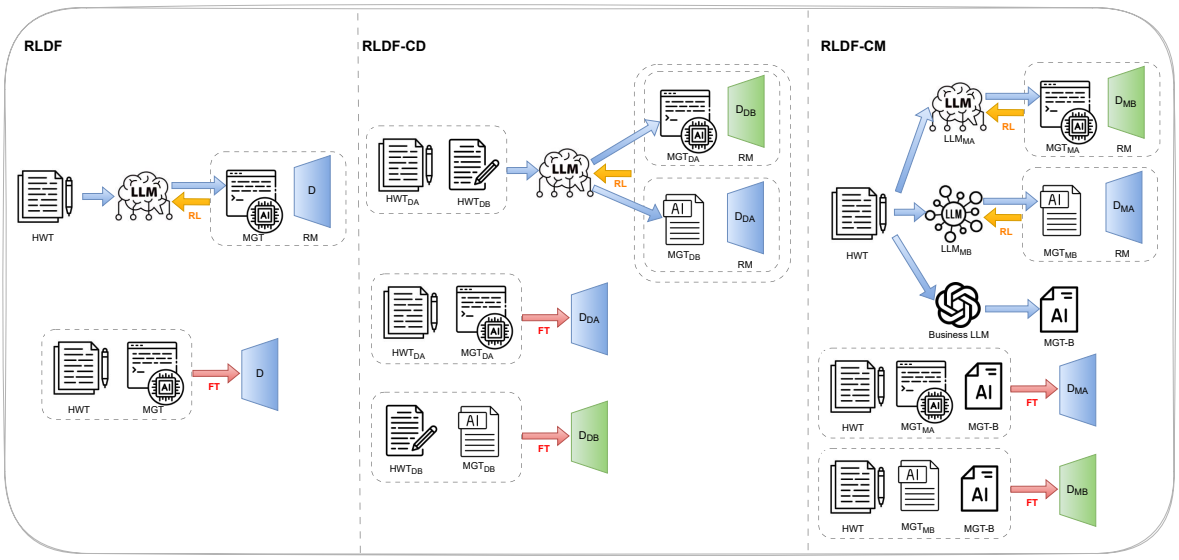

Figure 4: RLDF

up-to-date as possible to include the latest insights. Details of selections are shown in Figure 2.

3.4 Alignment Methods & MAGA Pipeline

We used four non-overlapping, combinable alignment methods: three classic alignment enhancement approaches and our proposed RLDF-CMD.

Roleplaying: Previous works have not fully leveraged the powerful role-designation function of the system role. According to (Chen et al., 2024)'s survey in roleplaying area , roleplaying prompt datasets proposed in existing previous works are often fine-grained, accompanied by extremely long identity and background information. LLMs struggle to perform these roles well and may even reveal obvious inconsistencies, making this an area still under research and specific optimization. Concise coarse-grained roleplaying prompts can yield excellent results and are needed in our work. By referring to (Shivagunde et al., 2023) and incorporating LLMs' summaries of fine-grained role prompts, we manually crafted 80 coarse-grained roleplaying prompts (details are provided in Appendix A).

BPO (Cheng et al., 2023): BPO is a prompt optimi zation model. Its optimization aims to ensure that the text generated by BPO-optimized prompts is better aligned with HWT. BPO appends additional information to the original prompt.

Self-Refine (Madaan et al., 2023): Self-Refine is a reasoning process optimization technique. It enables LLMs to provide feedback on the generated results and then optimize the results based on this feedback. This process can be executed in multiple rounds, with LLMs determining whether the optimization is sufficient. Single-round criticism and optimization can always bring slight improvements compared to the initial output. Since we found that LM-as-a-judge seems less reliable in MGT Detection, we ultimately adopted the single-round approach to ensure stable improvement. The detailed prompts we used are available in Appendix A.

RLDF-CMD: Details on RLDF will be talked about in §3.5. We adopted fully enhanced RLDF-CMD, which integrates RLDF-CM and RLDF-CD with cross-reward between the two groups both cross-domain and cross-model. As analyzed in Section 6, it's better to minimize within-group differences while maximizing between-group differences. Members of each group are listed in Appendix A.

MAGA Pipeline: Figure 3 shows the final pipeline we used to construct MAGA, which integrates four alignment methods into four steps. All these alignment methods complement each other and can be combined. Roleplaying adds a prompt prefix, while BPO adds a prompt suffix; one at the beginning and the other at the end, they expand the diversity of prompts and help improve the performance of RLDF. In turn, RLDF raises the starting point of self-refine. Notably, we ultimately merged HWT and MGT in a strict 1:1 ratio based on their titles. We believe this may help neural-based detectors better learn subtle differences, and

	ACC		TPR		TNR	AUC		ACC(@FPR=5%)	
Alignment?	✗	✓	✗	✓	-	✗	✓	✗	✓
R-B GPT2	58.39	52.00(6.39↓)	33.88	21.11(12.77↓)	82.90	56.64	45.18(11.46↓)	59.84	54.49(5.35↓)
R-L GPT2	56.53	51.50(5.03↓)	20.02	9.96(10.06↓)	93.03	53.09	40.51(12.58↓)	56.53	51.62(4.91↓)
R-B CGPT	57.50	47.26(10.24↓)	33.39	12.90(20.49↓)	81.62	63.79	50.60(13.19↓)	54.45	49.22(5.23↓)
RADAR	59.62	54.96(4.66↓)	29.16	19.84(9.32↓)	90.08	63.13	53.79(9.34↓)	60.03	55.98(4.05↓)
neutral.avg	58.01	51.43(6.58↓)	29.11	15.95(13.16↓)	86.91	59.16	47.52(11.64↓)	57.71	52.83(4.88↓)
GLTR	-	-	-	-	-	71.09	62.25(8.84↓)	70.04	60.39(9.65↓)
F-DetectGPT	79.86	78.63(1.23↓)	75.92	73.42(2.50↓)	83.83	84.01	83.38(0.63↓)	79.75	78.83(0.92↓)
LLMDet	50.50	44.04(11.46↓)	28.52	15.60(12.92↓)	72.48	47.89	35.58(12.31↓)	21.28	13.64(7.64↓)
Binoculars	84.33	79.29(5.04↓)	83.83	73.75(10.08↓)	84.83	86.76	79.68(7.08↓)	82.96	75.25(7.71↓)
metric.avg	71.56	67.32(4.24↓)	62.76	54.26(8.50↓)	80.38	72.44	65.22(7.22↓)	63.51	57.03(6.48↓)
avg	63.82	58.24(5.58↓)	43.53	32.37(11.16↓)	84.11	65.80	56.37(9.43↓)	60.61	54.93(5.68↓)

Table 1: MAGA Bench. An "✗" in the "Alignment" column indicates the MGB dataset, while the opposite indicates the MAGA dataset; all data are generated in a single run. Among these metrics, ACC, TPR, and TNR are evaluated under the default threshold, and all five metrics are described in detail in §4.

it also facilitates our data analysis 6.

3.5 RLDF

When alignment is mentioned, the first thing that comes to mind is Reinforcement Learning (RL). Inspired by RADAR, we systematically propose RLDF and explore the reasons for the need for RLDF-CD (cross-domain) and RLDF-CM (cross-model).

RLDF uses detector D as the Reward Model (RM) to fine-tune LLM via RL. This fine-tuning aims to make the text generated by the LLM harder to detect and better aligned with human-like characteristics.

RLDF constructs a dataset using HWT and more aligned corresponding MGT. Utilize this dataset to fine-tune neural-based detector D (e.g. RoBERTa), resulting in an enhanced detector D with stronger generalized detection capabilities. Through this process, multi-round adversarial fine-tuning is achieved.

RLDF-CD: MGT in the domain DA/DB ($MGT_{DA/DB}$) is used to fine-tune detectors respectively, obtaining detectors D_{DA} and D_{DB} that are applicable to domains DA and DB. These detectors are then cross-applied to score and reward MGT_{DB} and MGT_{DA} , which in turn are used for RL-based fine-tuning of the LLM.

RLDF-CM: MGT from LLMs MA/MB ($MGT_{MA/MB}$) is used to fine-tune detectors respectively, yielding detectors D_{MA} and D_{MB} that are specific to models MA and MB . These detectors are cross-applied to score and reward MGT_{MB} and MGT_{MA} . MGT used to fine-tune detector D, MGT-B (MGT from commercial LLMs) can be incorporated to enhance the fine-tuning process.

For adversarial fine-tuning of RADAR for paraphrasers, a key question is whether paraphrased texts count as machine-generated for training, causing uncertain detection performance. Recent studies propose 3/4-class tasks (beemo) in MGT detection. RADAR integrates RoBERTa fine-tuning into PPO, leading to hard convergence and limited experience for detectors and LLMs. When decoupled into RLDF, pre-trained RoBERTa reaches 100% accuracy on training sets and usally about 98% in cross-title scenarios (due to in-distribution overfitting), making it unable to be a reward model. Moreover, the small portion of text that RoBERTa classifies as human-written mainly relies on irrelevant features to achieve classification through shortcuts, and thus is not reliable. Our RLDF-CD/CM uses RoBERTa' s OOD detection trait (relying on relevant features) in HC-Var's theory to fine-tune LLMs effectively. The RL algorithm we adopted is GRPO; hyperparameters are in Appendix A.

4 Detectors

Our detector selection is basically the same as RAID. Based on the distinct characteristics of detection methods, detectors are categorized into two types: neural-based detectors (RoBERTa-Base (GPT2), RoBERTa-Large (GPT2), RoBERTa-Base (ChatGPT), RADAR) and metric-based detectors (GLTR, Binoculars, Fast DetectGPT, LLMDet).

These detectors use different thresholds

	SE24-M4-t	CL25-v-M4GT	CL25-v-HC3	CL25-v-MAGE	avg
R-B GPT2	64.32	61.04	94.51	64.18	71.01
R-L GPT2	54.30	50.60	89.38	57.22	62.88
R-B CGPT	67.19	63.91	99.27	62.92	73.32
RADAR	70.28	68.24	90.79	65.01	73.58
F-DetectGPT	83.43	80.96	93.57	72.16	82.53
LLMDet	52.25	50.58	72.74	34.56	52.53
Binoculars	89.76	82.36	87.95	73.27	83.34
R-B MGB(ours)	81.15	70.42	93.85	64.22	77.41
R-B MAGA(ours)	85.12(↑3.97)	72.68(↑2.26)	95.79(↑1.94)	64.28(↑0.06)	79.47(↑2.06)

Table 2: Four popular datasets from past work were selected, including the test set of SemEval2024-M4, and the three subsets of the validation set of COLING2025: M4GT, HC3, and MAGE.

for final classificatison. Thus, we compared multiple indicators: ACC, AUC, and ACC (@FPR=5%). For ACC, we used detectors' default thresholds—0.5 for neural-based detectors, Fast DetectGPT, and LLMDet (which provide probabilities); 0.9015310749276843 for Binoculars which Binoculars provided. GLTR has no default threshold, so its original ACC was not evaluated. AUC reliably assesses binary classification performance across varying thresholds. To address unfairness from different thresholds, inappropriate default thresholds, and emphasize MGT detection accuracy, we introduced ACC (@FPR=5%)—ACC when HWT detection error rate is fixed at 5%. For default ACC, we additionally showed TPR (MGT accuracy) and TNR (HWT accuracy).

5 Experiments and Results

5.1 attack

Table 1 compares mainstream AI text detectors on unaligned (MGB) and aligned (MAGA) datasets, with a core finding: all detectors suffer significant performance degradation on MAGA. Overall, alignment most severely weakens detectors' ability to identify AI text (reflected by TPR), followed by their overall discrimination capability (AUC). Notably, their ability to recognize human text (TNR) remains stable, proving alignment only interferes with AI-text judgment. Grouped analysis shows clear differences: the neutral group (e.g., R-B GPT2) is highly sensitive to alignment, with sharper performance drops; the metric group (e.g., F-DetectGPT) has stronger anti-interference, among which F-DetectGPT maintains the most stable performance, while LLMDet declines sharply. These results confirm text alignment is a core challenge for current detectors, and capturing deep-seated AI-generated features (not surface traits) is key to enhancing anti-alignment capability—providing a direction for future detector design.

5.2 generalization

Table 2 shows the performance changes of the models on four datasets (the higher the value, the stronger the ability). Overall, Binoculars and F-DetectGPT have become top-performing models with their consistent robustness. Although the upper limit of detection capability of such metric-based methods is not as high as that of neural-based methods, their generalization ability is indeed strong, and our model cannot be compared with them. Our enhanced model, R-B MAGA, significantly outperforms the unenhanced baseline (R-B MGB) on all datasets, which proves the effectiveness of MAGA. However, the improvement is negligible on the MAGE dataset, which seems to be indeed highly challenging. Notably, our method has comprehensively outperformed other neural-based models represented by RADAR, which again proves the significance of our work.

6 Ablation Study & Analysis

We took 10% of the original prompts and used only RLDF, roleplaying, BPO, and self-refine to generate the corresponding training and test sets. We then observed two metrics: first, the average performance decline of the selected detectors compared to the unenhanced state; second, the

	avg.ACC@attack	avg.AUC@attack	avg.ACC@generalization	avg.AUC@generalization
MGB	63.82	65.80	77.41	83.15
/w RLDF	61.88 (1.94 ↓)	62.89 (2.91 ↓)	79.28 (1.87 ↑)	85.13 (1.98 ↑)
/w roleplaying	61.73 (2.09 ↓)	62.13 (3.67 ↓)	79.33 (1.92 ↑)	85.27 (2.12 ↑)
/w BPO	63.49 (0.33 ↓)	65.62 (0.18 ↓)	77.45 (0.04 ↑)	83.23 (0.08 ↑)
/w self-refine	61.68 (2.14 ↓)	62.78 (3.02 ↓)	79.47 (2.06 ↑)	85.45 (2.30 ↑)

Table 3: Ablation study.

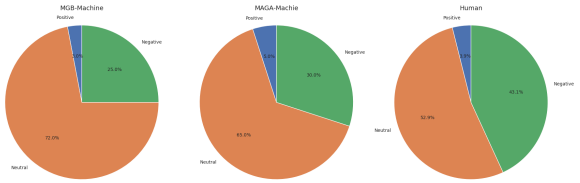

Figure 5: Sentiment Analysis

average performance improvement of RoBERTa, which was fine-tuned on these training sets, when tested on other selected test sets. Ablation experiments show that all four alignment methods are effective. RLDF, roleplaying, and self-refine yield similar effects, while BPO is significantly weaker, as shown in table 3. However, BPO is of great significance and value for expanding the diversity of prompts, and can also enhance the effectiveness of RLDF.

Figure 5 provides a visual analysis, comparing the MGB machine, MAGA machine, and humans. It can be seen that the enhanced machine text is emotionally closer to human text.

7 Conclusion

We constructs the MAGA dataset, which covers 10 domains and 12 LLMs, and generates machine-generated text (MGT) closer to human-written text (HWT) using four alignment methods: roleplaying, BPO, self-refine, and RLDF-CMD. Experiments validate MAGA' s dual value: it significantly impairs the performance of all mainstream detectors (with an average 5.58% drop in ACC and 11.16% drop in TPR), while RoBERTa fine-tuned on MAGA achieves a 2.06% average ACC improvement on external datasets like SE24-M4 and CL25-HC3—outperforming the baseline fine-tuned on unenhanced MGB and classic detectors such as RADAR. Ablation tests further confirm RLDF, roleplaying, and self-refine as core effective alignment methods, while BPO, though less impactful, aids in prompt diversity and RLDF enhancement. Overall, MAGA provides a new benchmark for detector robustness testing and a practical path to boost detector generalization.

Limitations

We still have some limitations. First, BPO is far less effective than other alignment methods, failing to narrow the feature gap between MGT and HWT, requiring more efficient prompt optimization strategies. Second, generalization testing only covers 4 external datasets, lacking scenarios like low-resource languages or professional fields (e.g., medicine, law), limiting validation of MAGA' s effectiveness in extreme or niche cases. Third, experiments only use 8 detectors from the RAID benchmark, excluding latest zero-shot detectors based on large models (e.g., GPT-4), and do not test detector performance under "adversarial modification + alignment" dual interference, reducing conclusion generalizability.

References

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention. In *Proceedings of the 34th ACM conference on hypertext and social media*, pages 1–10.

Meghana Moorthy Bhat and Srinivasan Parthasarathy. 2020. How effectively can machines defend against machine-generated fake news? an empirical study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 48–53, Online. Association for Computational Linguistics.

T Brown and 1 others. 2020. Language models are few-shot learners. advances in neural information. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. Retrived from https://proceedings. neurips. cc/pa*.

Shuyang Cai and Wanyun Cui. 2023. Evade chatgpt detectors via a single space. *Preprint*, arXiv:2307.02599.

Megha Chakraborty, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar Barman, Chandan Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Counter Turing test (CT2): AI-generated text detection is not as easy as you may think - introducing

AI detectability index (ADI). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2206–2239, Singapore. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

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. 2023a. 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.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023b. Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Vitalii Fishchuk and Daniel Braun. 2023. Efficient black-box adversarial attacks on neural text detectors. *Preprint*, arXiv:2311.01873.

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Zhang, and Jennifer Biggs. 2021. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 119–127, Online. Australasian Language Technology Association.

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. *Preprint*, arXiv:2401.05952.

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. *Preprint*, arXiv:2301.07597.

Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 542–556.

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. *Preprint*, arXiv:2401.12070.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. How you prompt matters! even task-oriented constraints in instructions affect llm-generated text detection. *Preprint*, arXiv:2311.08369.

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. *Preprint*, arXiv:2303.13408.

Pranav Kulkarni, Ziqing Ji, Yan Xu, Marko Neskovic, and Kevin Nolan. 2023. Exploring semantic perturbations on grover. *Preprint*, arXiv:2302.00509.

Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. How reliable are ai-generated-text detectors? an assessment framework using evasive soft prompts. *Preprint*, arXiv:2310.05095.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. Mage: Machine-generated text detection in the wild. *Preprint*, arXiv:2305.13242.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. 2023. Large language models can be guided to evade ai-generated text detection. *Preprint*, arXiv:2305.10847.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. *Preprint*, arXiv:2401.07867.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

J. Pu, Z. Sarwar, S. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath. 2023a. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1613–1630, Los Alamitos, CA, USA. IEEE Computer Society.

Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. 2023b. On the zero-shot generalization of machine-generated text detectors. *Preprint*, arXiv:2310.05165.

Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, Seattle, United States. Association for Computational Linguistics.

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

Areg Mikael Sarvazyan, José Ángel González, Paolo Rosso, and Marc Franco-Salvador. 2023a. Supervised machine-generated text detectors: Family and scale matters. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 1821, 2023, Proceedings*, page 121132, Berlin, Heidelberg. Springer-Verlag.

Areg Mikael Sarvazyan, Jos é ngel Gonz á lez, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023b. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Preprint*, arXiv:2309.11285.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. In *Computational Linguistics and Intellectual Technologies*. RSUH.

Namrata Shivagunde, Vladislav Lialin, and Anna Rumshisky. 2023. Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. *arXiv preprint arXiv:2303.16445*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv preprint*.

Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13:363–383.

Chris Stokel-Walker. 2022. Ai bot chatgpt writes smart essays-should professors worry? *Nature*.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2024. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. *Preprint*, arXiv:2309.02731.

Jian Wang, Shangqing Liu, Xiaofei Xie, and Yi Li. 2023a. Evaluating aigc detectors on code content. *Preprint*, arXiv:2304.05193.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, and 1 others. 2024. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *Preprint*, arXiv:2305.14902.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Max Wolff. 2020. Attacking neural text detectors. *Preprint*, arXiv:2002.11768.

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2023. On the generalization of training-based chatgpt detection methods. *Preprint*, arXiv:2310.01307.

R Zellers, A Holtzman, H Rashkin, Y Bisk, A Farhadi, F Roesner, and Y Choi. 2019. Defending against neural fake news. advances in neural information processing systems, 32.

A Hyperparameters & training time

We trained using GRPO + LoRA on 8 L20 GPUs. The smallest model (7B) was trained for 50 hours, and the largest model (100B) for 100 hours. Our micro_batch_size is 1, batch_size is 1024, and learning rate is 5e6. We trained for 1 epoch, while RLDF was trained for 2 rounds.