

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 **Machine-Augment-Generated Text via Alignment (MAGA)**. MAGA's pipeline achieves comprehensive alignment from prompt construction to reasoning process, among which **Reinforced Learning from Detectors Feedback (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 4.60% in generalization detection AUC. MAGA Dataset caused an average decrease of 8.13% in the AUC 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

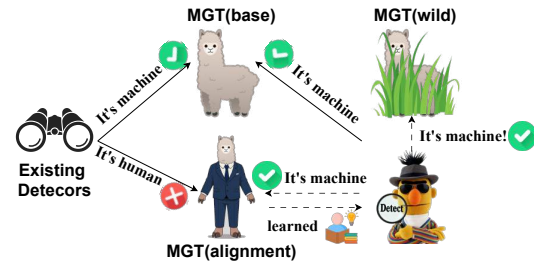


Figure 1: More aligned MGT not only evades detection by existing detectors but also facilitates fine-tuning of neural-based detectors, which enhancing their generalization capability for wild detection.

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.

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

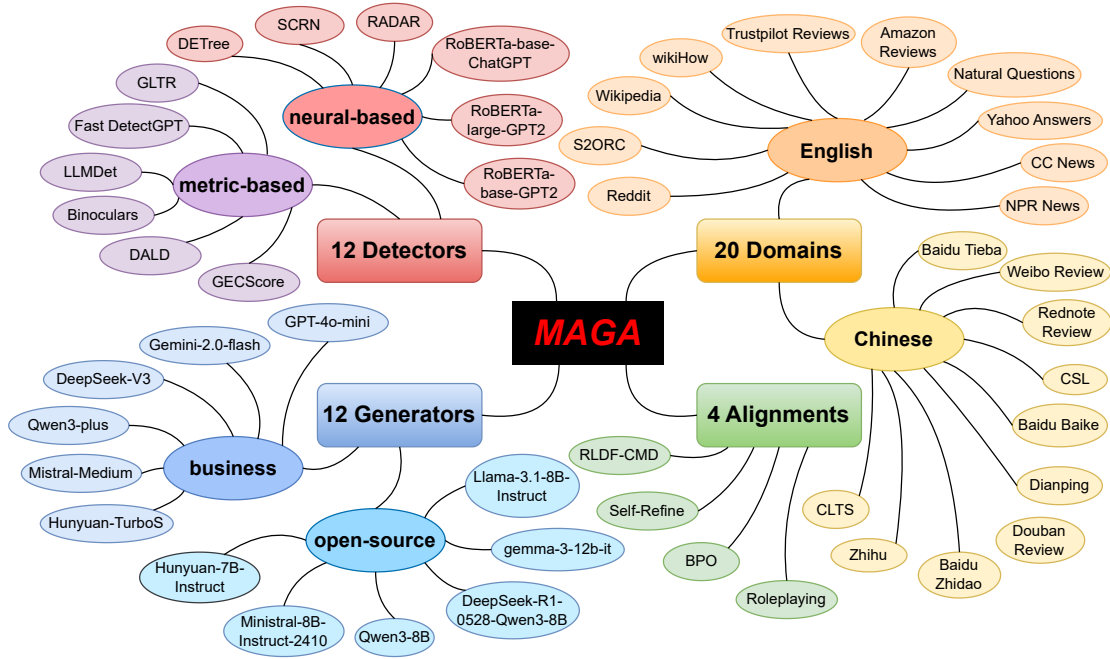


Figure 2: MAGA-Bench Overview. Our dataset construction contains 20 domains and 12 generators. We evaluate selected detectors on our dataset. We also adopted various decoding strategies, which are not presented here but detailed in §E.

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 The MAGA Dataset

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 (§2.1), splitting them into a 60k training set and a 12k validation set. For each human text, we generated a matching machine text using its title. We used 12 generators (§2.2). We first created an unenhanced MGB dataset for comparison, generated via the original prompt template (§2.3); full original prompt in Appendix A). MAGA’s augment involves four methods (§2.3): roleplaying, BPO, self-refine, and RLDF-CMD (§2.4). The MAGA pipeline, integrating these four methods, was applied to the MAGA dataset. For roleplaying, BPO, and self-refine, we separately provided MAGA-extra for additional indication. Full statistics are in Appendix A.

As shown in Table 1, the original size of our generalized MAGA dataset is 936k entries. Among all the comparative studies, ours is the only one that covers multi-domain, multi-generator, multi-lingual, adversarial attack and multi-sampling-params, while simultaneously performing alignment augmentation.

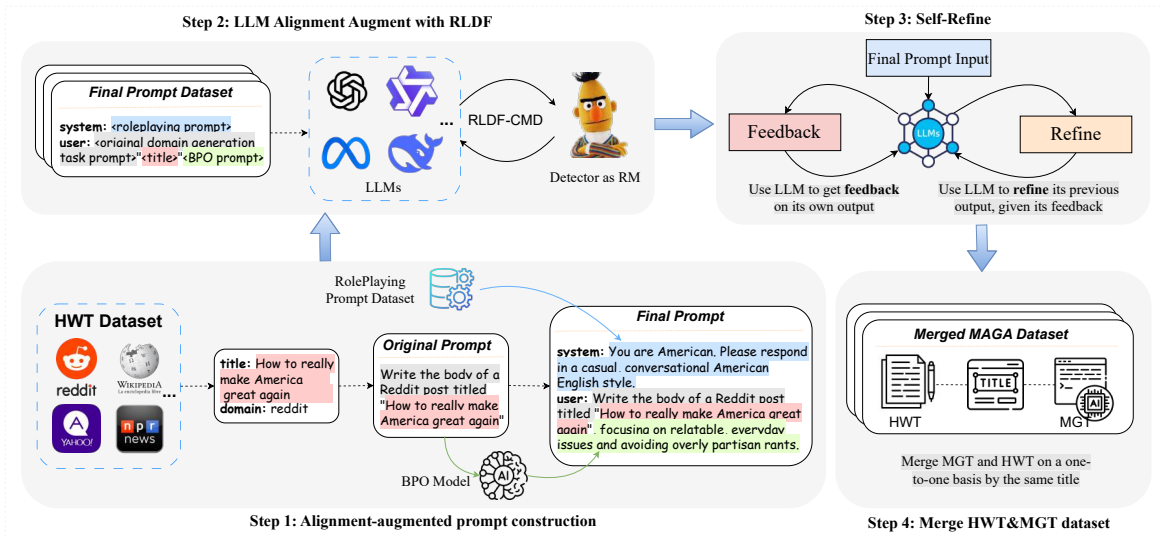


Figure 3: MAGA Pipeline

2.1 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., 2024c). Taking English as an example, 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 B.

2.2 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/Minstral, 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 up-to-date as possible to include the latest insights.

Details of selections are shown in Figure 2.

2.3 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 optimization 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

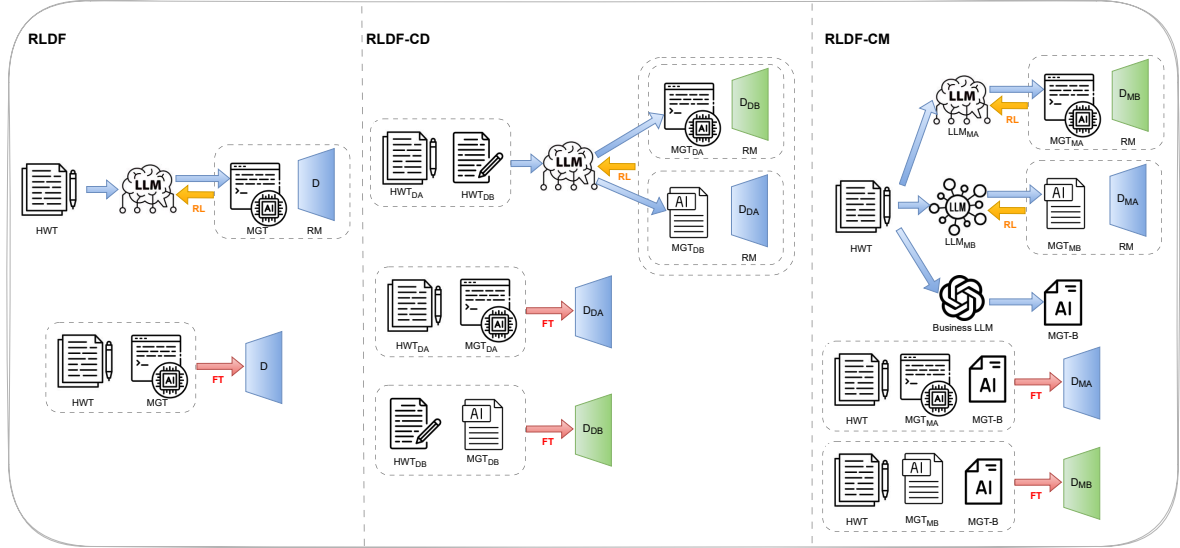


Figure 4: RLDF (**R**einforce **L**earning from **D**etectors **F**eedback). Fine-tune the LLM via RL using detector D as the RM, resulting in harder-to-detect and better human-aligned generated text. Fine-tune detector D (RoBERTa) on the dataset constructed from HWT and its corresponding better-aligned MGT, to obtain a detector D with stronger generalized detection capability, thus enabling multi-round adversarial fine-tuning. RLDF-CD and RLDF-CM resolve RLDF’s practical infeasibility caused by RoBERTa overfitting by means of cross-reward.

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 G.

RLDF-CMD: Details on RLDF will be talked about in §2.4. 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 D.

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 it also facilitates our data analysis 5.

Dataset	Origin Size	Domain Coverage	Model Coverage	Multilingual Coverage	Adversarial Coverage	Sampling Coverage	Alignment Augment
TuringBench (Uchendu et al., 2021)	200k	✗	✓	✗	✗	✗	✗
HC3 (Guo et al., 2023)	26.9k	✓	✗	✓	✗	✗	✗
MGTBench (He et al., 2023)	2.8k	✓	✓	✗	✓	✗	✗
M4 (Wang et al., 2023b)	122k	✓	✓	✓	✗	✗	✗
MAGE (Li et al., 2024)	447k	✓	✓	✗	✗	✗	✗
RAID (Dugan et al., 2024)	570k	✓	✓	✗	✓	✓	✗
RealDet (Zhu et al., 2025)	847k	✓	✓	✓	✓	✗	✗
MAGA (Ours)	936k	✓	✓	✓	✓	✓	✓

Table 1: A comparison of the publicly available sources of generated text. The term Original Size was proposed in Zhu et al. (2025). It refers to human-written texts and direct machine-generated texts, excluding rule-based attacks such as RAID. A more comprehensive comparison is in the table 17

Alignment?	ACC		TPR		TNR	AUC		ACC(@FPR=5%)	
	✗	✓	✗	✓	-	✗	✓	✗	✓
R-B GPT2	58.44	51.96(6.48↓)	33.97	21.02(12.95↓)	82.90	56.91	45.02(11.89↓)	59.94	54.17(5.77↓)
R-L GPT2	56.69	51.47(5.22↓)	20.34	9.91(10.43↓)	<u>93.03</u>	53.45	40.23(13.22↓)	56.97	51.35(5.62↓)
R-B CGPT	57.60	47.15(10.45↓)	33.57	12.67(20.90↓)	81.62	63.97	50.31(13.66↓)	54.73	49.01(5.72↓)
RADAR	59.82	54.83(4.99↓)	29.56	19.58(9.98↓)	90.08	63.43	53.49(9.94↓)	60.33	55.68(4.65↓)
SCRN	74.26	73.23(1.03↓)	<u>94.23</u>	<u>92.17</u> (2.06↓)	54.29	84.76	81.79(2.97↓)	59.97	58.93(1.04↓)
DETree	58.05	57.94(0.11↓)	99.94	99.72 (0.22↓)	16.16	93.92	92.89 (1.03↓)	87.23	85.26 (1.97↓)
neutral.avg	60.81	56.10(4.71↓)	51.94	42.51(9.42↓)	69.68	69.41	60.62(8.79↓)	63.20	59.07(4.13↓)
GLTR	-	-	-	-	-	71.31	62.05(9.26↓)	70.27	62.09(8.18↓)
F-DetectGPT	79.97	78.41(1.57↓)	76.27	73.14(3.13↓)	83.67	84.01	83.38(0.63↓)	79.96	78.58(1.38↓)
LLMDet	50.46	43.80(6.66↓)	28.74	15.43(13.31↓)	72.17	47.89	35.58(12.31↓)	56.89	51.02(5.87↓)
Binoculars	84.26	79.03 (5.23↓)	83.93	73.47(10.46↓)	84.59	<u>86.76</u>	79.68(7.08↓)	<u>83.23</u>	75.17(8.06↓)
DALD	<u>80.01</u>	77.68(2.33↓)	78.01	73.34(4.67↓)	82.01	84.49	82.97(1.52↓)	79.79	78.03(1.76↓)
GECScore	56.35	50.95(5.40↓)	16.42	5.63(10.79↓)	96.27	65.92	51.89(14.03↓)	59.57	53.06(6.51↓)
metric.avg	70.21	65.97(4.24↓)	56.67	48.20(8.47↓)	83.74	73.40	65.93(7.47↓)	71.62	66.33(5.29↓)
avg	65.08	60.59(4.50↓)	54.09	45.10(8.99↓)	76.07	71.40	63.27(8.13↓)	67.41	62.70(4.71↓)

Table 2: 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 §3.

2.4 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 paraphraser, a key question is whether paraphrased texts count as machine-generated for training, causing uncertain detection performance. Recent studies propose 3/4-class tasks (MixSet (Zhang et al., 2024)) 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 usually 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. Details of RLDF are shown in Appendix D. Hyperparameters are shown in Appendix I.

3 Detectors

Our detector selection is based on the picks from RAID (excluding commercial detectors that are prohibitively expensive), with a number

	S-M4		C-M4GT		C-MAGE		C-HC3		AVERAGE	
	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)
R-B GPT2	75.92	68.25	78.38	63.76	<u>76.40</u>	55.19	98.45	<u>94.59</u>	82.29	70.45
R-L GPT2	70.44	63.98	74.02	57.04	79.55	55.98	94.88	89.34	79.72	66.59
R-B CGPT	81.25	74.97	74.82	58.61	61.32	47.89	99.99*	99.27*	79.35	70.19
RADAR	76.25	55.46	70.89	31.13	58.43	45.08	90.16	84.74	73.93	54.10
SCRN	79.37	58.13	93.49	33.73	66.24	48.01	93.49	88.77	83.15	57.16
GLTR	80.19	72.23	78.29	63.13	58.34	46.78	92.37	91.13	77.30	68.32
F-DetectGPT	88.51	85.99	86.76	<u>79.65</u>	64.67	56.67	99.94	97.19	84.97	<u>79.88</u>
LLMDet	70.28	50.17	74.89	53.16	50.13	47.27	90.23	81.34	71.38	57.99
Binoculars	89.73	89.09	<u>89.14</u>	84.10	63.79	58.89	99.59	97.19	<u>85.56</u>	82.32
DALD	88.15	85.01	85.41	78.87	64.96	<u>56.73</u>	<u>99.76</u>	97.19	84.57	79.45
GECScore	89.05	68.81	65.88	50.00	53.67	40.00	90.18	86.96	74.70	61.44
R-B MGB (Ours)	<u>94.63</u>	<u>90.58</u>	82.98	73.37	60.90	46.95	93.14	91.68	82.91	75.65
R-B MAGA (Ours)	98.87	95.41	88.40	78.99	64.63	47.67	98.14	94.06	87.51	79.03
Δ R-B MAGA-MGB	4.24	4.83	5.42	5.62	3.73	0.72	5.00	2.38	4.60	3.39

Table 3: 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. ACC(5%) here is a short for ACC(@FPR=5%). R-B CGPT is trained on HC3, so it’s not boldfaced. bold is the largest, underline is the second largest.

of brand-new detectors added to the list. 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, SCRN, DETree) and metric-based detectors (GLTR, Binoculars, Fast DetectGPT, LLMDet, DALD, GECScore). Among them, SCRN, DETree, DALD and GECScore are the novel detectors we selected. A detailed introduction to these detectors is provided in Appendix F.

These detectors use different thresholds for final classification. 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).

4 Experiments and Results

4.1 attack

Table 2 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.

4.2 generalization

Table 3 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

	Attack			Generalization		
	TPR	AUC	ACC(@FPR=5%)	TPR	AUC	ACC(@FPR=5%)
MGB	54.09	71.4	67.41	53.13	82.91	75.65
/w RLDF-CMD	49.95(4.14↓)	67.91(3.49↓)	65.18(2.23↓)	60.97(7.84↑)	85.02(2.11↑)	77.28(1.63↑)
/w roleplaying	48.74(5.35↓)	66.43(4.97↓)	64.72(2.69↓)	63.47(10.34↑)	85.39(2.48↑)	77.69(2.04↑)
/w BPO	50.12(3.97↓)	68.09(3.31↓)	65.49(1.92↓)	60.24(7.11↑)	84.87(1.96↑)	77.16(1.51↑)
/w self-refine	48.87(5.22↓)	66.32(5.08↓)	64.43(2.98↓)	64.21(11.08↑)	85.84(2.93↑)	77.47(1.82↑)
MAGA	45.1(8.99↓)	63.27(8.13↓)	62.7(4.71↓)	71.3(18.17↑)	87.51(4.6↑)	79.03(3.38↑)

Table 4: Ablation experiments on various alignment methods. This is also a summary of the experimental results of MAGA-extra.

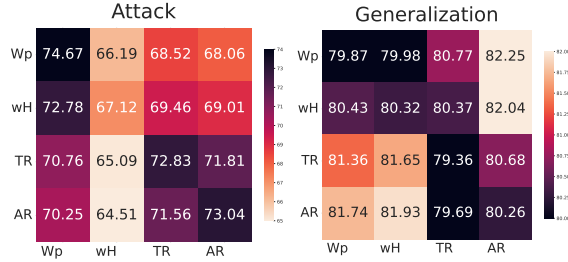


Figure 5: RLDF-CD matrix analysis for attack and generalization. The horizontal axis represents the RM domain, and the vertical axis represents the LLM domain.

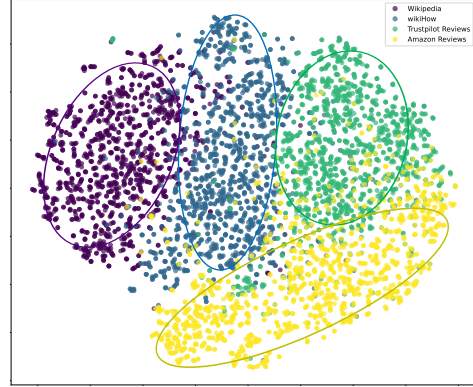


Figure 6: T-SNE on HWT domain selected in Figure 12.

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.

5 Ablation Study

First, we explored RLDF configurations. Table 5 shows an ablation experiment example on domain grouping selection for RLDF-CD, with the horizontal axis as RM Domain and vertical axis as LLM Domain. The main diagonal indicates no adversarial fine-tuning, as non-cross-domain settings are inherently infeasible with almost no gradient changes during training. Intuitively, the larger the inter-domain gap, the more significant the cross-reward effect, which is confirmed by inter-domain distance visualization (Figure 6), with more details in Figures 12 to 14. Model visualization is indistinct, but quantitative

experiments follow the same domain-related trend.

Figure 11 presents RLDF adversarial round experiments, showing convergence in 3 rounds.

Table 4 displays ablation experiments on various alignment methods with consistent results. Figure 12 provides detailed ablation results: RLDF is core, and RLDF-CMD is least diluted when combined with other methods. Potential reasons and RLDF advantages are discussed in Appendix L.

6 Analysis

Following the approach of RAID (Dugan et al., 2024), we present the statistical values of log PPL in Figures 7 and 9. RAID demonstrated that analyses using different LLMs may yield varying results. We employed seen English LLMs, seen Chinese LLMs, and one unseen LLM. Although the absolute values generated by different LLMs differ, the comparisons between MGB-M and MAGA-M are consistent, provided that a single LLM is used throughout. PPL comparisons indicate that the relatively high PPL of MAGA may be one of the factors contributing to successful attacks.

We summarize the metrics commonly compared in previous studies and quantitatively conduct a multi-dimensional comparison of HWT, MGB-M, and MAGA-M in Figure 8, with details elaborated in Appendix H. From MGB-M to MAGA-M,

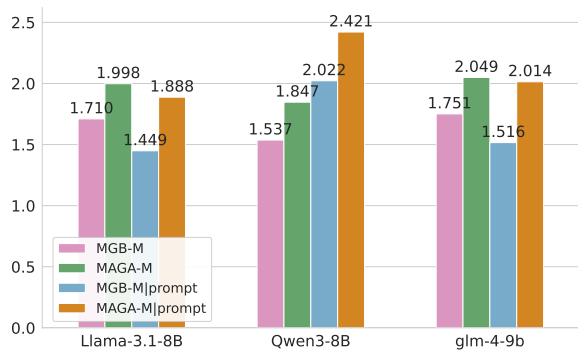


Figure 7: log-pp-core

there is little difference in lexical and semantic similarity relative to HWT; these metrics may be more indicative of text quality rather than the fundamental distinctions between human and machine-generated texts. MGB-M exhibits higher lexical diversity than HWT, and MAGA-M shows even greater diversity. In Figure 10, we compare the proportion of n-gram overlaps with HWT. MAGA-M’s bag-of-words contains more terms matching those of HWT, but this comes at the cost of excessive diversity leading to deviations, a phenomenon that warrants further investigation. In terms of text readability and emotional consistency, MAGA and HWT show a high degree of alignment.

7 Related Work

Advancements in LLMs have made MGT indistinguishable from HWT (Brown et al., 2020; Ouyang et al., 2022; Chang et al., 2024), leading to abuse risks (e.g., misinformation, fraud, academic misconduct) (Hanley and Durumeric, 2024; Zellers et al., 2019; Weidinger et al., 2021; Stokel-Walker, 2022). Thus, robust MGT detectors are critical, driving diverse detection methods and benchmarks (Crothers et al., 2023).

Early benchmarks (e.g., M4 (Wang et al., 2023b)) enhanced detection difficulty via data distribution/source diversity but relied on fixed prompt templates, limiting generality. RAID (Dugan et al., 2024), the first robustness-prioritized benchmark, uses perturbations (e.g., random ‘\n’, article deletion) to raise MGT perplexity and attack metric-based detectors—findings supported by Binoculars (Hans et al., 2024) (near-zero accuracy for extreme random MGT). However, such semantic-damaging manipulations hinder neural-based detector (e.g., RoBERTa (Liu et al., 2019)) fine-tuning, limiting benchmarks to specific attack testing rather than generalized detection advancement.

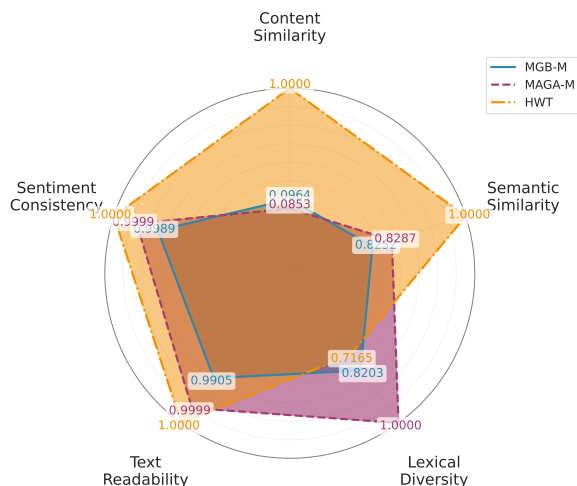


Figure 8: multi-dimension-radar

8 Conclusion

We construct 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,

469	reducing conclusion generalizability.	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. <i>arXiv preprint arXiv:2311.04155</i> .	523 524 525 526 527
470	References		
471	Aaditya Bhat. 2023. Gpt-wiki-intro .		
472	Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, and 1 others. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection. <i>arXiv preprint arXiv:2408.04284</i> .		528 529 530 531 532 533 534 535
473		Joseph Cornelius, Oscar Lithgow-Serrano, Sandra Mitrović, Ljiljana Dolamic, and Fabio Rinaldi. 2024. Bust: Benchmark for the evaluation of detectors of llm-generated text. In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 8029–8057.	
474			
475			
476			
477			
478			
479	AI@Meta. 2024. Llama 3 model card .		
480	Ekaterina Artemova, Jason S Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. 2025. Beemo: Benchmark of expert-edited machine-generated outputs. In <i>Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 6992–7018.		536 537 538 539
481			
482			
483			
484			
485			
486			
487			
488			
489	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 <i>Proceedings of the 34th ACM conference on hypertext and social media</i> , pages 1–10.		540 541 542 543 544 545 546 547
490			
491			
492			
493			
494			
495	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 . <i>Preprint</i> , arXiv:2310.05130.		548 549 550 551 552 553 554
496			
497			
498			
499			
500	T Brown and 1 others. 2020. Language models are few-shot learners. <i>advances in neural information processing systems (NeurIPS 2020)</i> , Vancouver, Canada. Retrieved from https://proceedings.neurips.cc/pa .		555
501			
502			
503			
504			
505	Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2024. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. <i>arXiv preprint arXiv:2409.14509</i> .		556 557 558 559 560 561
506			
507			
508			
509			
510	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. <i>ACM transactions on intelligent systems and technology</i> , 15(3):1–45.		562 563 564 565 566 567 568
511			
512			
513			
514			
515			
516	Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. The oscar of ai theater: A survey on role-playing with language models. <i>arXiv preprint arXiv:2407.11484</i> .		569 570 571 572 573 574 575 576 577 578
517			
518			
519	Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Token prediction as implicit classification to identify llm-generated text. <i>arXiv preprint arXiv:2311.08723</i> .		
520			
521			
522			

690			
691			
692			
693			
694	LZYFirecn. 2025.	weibo-hot-search .	
695	Dominik Macko, Jakub Kopal, Robert Moro, and Ivan		
696	Srba. 2024.	Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts . <i>Preprint</i> , arXiv:2406.12549.	
697			
698			
699	Dominik Macko, Robert Moro, Adaku Uchendu,		
700	Jason Lucas, Michiharu Yamashita, Matúš Pikuliak,		
701	Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko,		
702	and Maria Bielikova. 2023.	MULTITuDE: Large-scale multilingual machine-generated text detection benchmark . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 9960–9987, Singapore. Association for Computational Linguistics.	
703			
704			
705			
706			
707			
708	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler		
709	Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon,		
710	Nouha Dziri, Shrimai Prabhumoye, Yiming Yang,		
711	and 1 others. 2023.	Self-refine: Iterative refinement with self-feedback . <i>Advances in Neural Information Processing Systems</i> , 36:46534–46594.	
712			
713			
714	Mistral AI. 2025.	Mistral medium 3: Medium is the new large .	
715			
716	Mistral AI Team, Albert Jiang, Alexandre Abou		
717	Chahine, Alexandre Sablayrolles, Alexis Tacnet,		
718	Alodie Boissonnet, Alok Kothari, Amélie Héliou,		
719	Andy Lo, Anna Peronin, Antoine Meunier, Antoine		
720	Roux, Antonin Faure, Aritra Paul, Arthur Darcet,		
721	Arthur Mensch, Audrey Herblin-Stoop, Augustin		
722	Garreau, Austin Birky, and 95 others. 2024.	Ministral-8b-instruct-2410 .	
723			
724	Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla,		
725	Paolo Basso, Margherita Musumeci, and Georg		
726	Groh. 2023.	Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era . In <i>Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)</i> , pages 190–207, Toronto, Canada. Association for Computational Linguistics.	
727			
728			
729			
730			
731			
732			
733	Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019.	Justifying recommendations using distantly-labeled reviews and fine-grained aspects . In <i>Conference on Empirical Methods in Natural Language Processing</i> .	
734			
735			
736			
737	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,		
738	Carroll Wainwright, Pamela Mishkin, Chong Zhang,		
739	Sandhini Agarwal, Katarina Slama, Alex Ray, and 1		
740	others. 2022.	Training language models to follow instructions with human feedback . <i>Advances in neural information processing systems</i> , 35:27730–27744.	
741			
742			
743			
744	Areg Mikael Sarvazyan, José Ángel González,		
745	Paolo Rosso, and Marc Franco-Salvador. 2023a.	Supervised machine-generated text detectors: Family and scale matters . In <i>Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings</i> , page 121132, Berlin, Heidelberg. Springer-Verlag.	746 747 748 749 750 751 752
	Areg Mikael Sarvazyan, José Ángel González, 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 . <i>Preprint</i> , arXiv:2309.11285.	753 754 755 756 757 758
	sentence-transformers. 2021.	reddit-title-body .	759
	sentence-transformers. 2024a.	natural-questions .	760
	sentence-transformers. 2024b.	npr .	761
	sentence-transformers. 2024c.	yahoo-answers .	762
	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 <i>Computational Linguistics and Intellectual Technologies</i> . RSUH.	763 764 765 766 767 768 769
	Namrata Shivagunde, Vladislav Lialin, and Anna		
	Rumshisky. 2023.	Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning . <i>arXiv preprint arXiv:2303.16445</i> .	770 771 772 773
	Irene Solaiman, Miles Brundage, Jack Clark, Amanda		
	Askill, 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 . <i>arXiv preprint</i> .	774 775 776 777 778 779 780
	Chris Stokel-Walker. 2022.	Ai bot chatgpt writes smart essays-should professors worry? <i>Nature</i> .	781 782
	Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and		
	Songlin Hu. 2024.	Hc3 plus: A semantic-invariant human chatgpt comparison corpus . <i>Preprint</i> , arXiv:2309.02731.	783 784 785 786
	Gemma Team. 2025.	Gemma 3 .	787
	Tencent. 2025.	Hunyuan-7b-instruct .	788
	Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan		
	Bai, Qinghua Zhang, Ruifeng Li, Chaoxi Xu,		
	and Yunhe Wang. 2023.	Multiscale positive-unlabeled detection of ai-generated texts . <i>ArXiv</i> , abs/2305.18149.	789 790 791 792 793
	twn39. 2025.	dianping .	794
	Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and		
	Dongwon Lee. 2021.	Turingbench: A benchmark environment for turing test in the age of neural text generation . <i>Preprint</i> , arXiv:2109.13296.	795 796 797 798

799	Jian Wang, Shangqing Liu, Xiaofei Xie, and Yi Li.	Peipeng Yu, Jiahao Chen, Xuan Feng, and Zhihua	857
800	2023a. Evaluating aigc detectors on code content .	Xia. 2023. Cheat: A large-scale dataset for	858
801	<i>Preprint</i> , arXiv:2304.05193.	detecting chatgpt-written abstracts. <i>arXiv preprint</i>	859
802	Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei	<i>arXiv:2304.12008</i> .	860
803	Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li.	R Zellers, A Holtzman, H Rashkin, Y Bisk, A Farhadi,	861
804	2024a. Coglm: Tracking cognitive development of	F Roesner, and Y Choi. 2019. Defending against	862
805	large language models . <i>ArXiv</i> , abs/2408.09150.	neural fake news. <i>advances in neural information</i>	863
806	Yuxia Wang, Jonibek Mansurov, Petar Ivanov,	<i>processing systems</i> , 32.	864
807	Jinyan Su, Artem Shelmanov, Akim Tsvigun,	Cong Zeng, Shengkun Tang, Xianjun Yang, Yuanzhou	865
808	Osama Mohammed Afzal, Tarek Mahmoud,	Chen, Yiyao Sun, Zhiqiang Xu, Yao Li, Haifeng	866
809	Giovanni Puccetti, Thomas Arnold, and 1 others.	Chen, Wei Cheng, and Dongkuan Xu. 2024.	867
810	2024b. M4gt-bench: Evaluation benchmark for	Improving logits-based detector without logits from	868
811	black-box machine-generated text detection. In	black-box llms . <i>ArXiv</i> , abs/2406.05232.	869
812	<i>Proceedings of the 62nd Annual Meeting of the</i>	Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang,	870
813	<i>Association for Computational Linguistics (Volume</i>	Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye	871
814	<i>1: Long Papers</i>), pages 3964–3992.	Li, Zhengyan Fu, Yao Wan, and 1 others. 2024.	872
815	Yuxia Wang, Jonibek Mansurov, Petar Ivanov,	Llm-as-a-coauthor: Can mixed human-written and	873
816	Jinyan Su, Artem Shelmanov, Akim Tsvigun,	machine-generated text be detected? <i>arXiv preprint</i>	874
817	Osama Mohammed Afzal, Tarek Mahmoud,	<i>arXiv:2401.05952</i> .	875
818	Giovanni Puccetti, Thomas Arnold, and 1 others.	Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin,	876
819	2024c. Semeval-2024 task 8: Multidomain,	Fang Fang, and Yangxi Li. 2025. Reliably bounding	877
820	multimodal and multilingual machine-generated text	false positives: A zero-shot machine-generated	878
821	detection. <i>arXiv preprint arXiv:2404.14183</i> .	text detection framework via multiscaled conformal	879
822	Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan	prediction. <i>arXiv preprint arXiv:2505.05084</i> .	880
823	Su, Artem Shelmanov, Akim Tsvigun, Chenxi	A Statistics of MAGA	881
824	Whitehouse, Osama Mohammed Afzal, Tarek	Unless otherwise specified, the MAGA dataset	882
825	Mahmoud, Alham Fikri Aji, and Preslav Nakov.	mentioned in the text generally refers exclusively	883
826	2023b. M4: Multi-generator, multi-domain, and	to the English version excluding the extra variants.	884
827	multi-lingual black-box machine-generated text	Our generalized MAGA dataset comprises two	885
828	detection . <i>Preprint</i> , arXiv:2305.14902.	language-specific subsets: the English version	886
829	Laura Weidinger, John Mellor, Maribeth Rauh, Conor	(MAGA) and the Chinese version (MAGA-cn).	887
830	Griffin, Jonathan Uesato, Po-Sen Huang, Myra	Furthermore, the generalized English MAGA	888
831	Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh,	dataset is divided into 6 splits: MGB, MAGA,	889
832	and 1 others. 2021. Ethical and social risks	MAGA-extra-RLDF-CMD, MAGA-extra-BPO,	890
833	of harm from language models. <i>arXiv preprint</i>	MAGA-extra-roleplaying, and MAGA-extra-self-	891
834	<i>arXiv:2112.04359</i> .	refine. Similarly, MAGA-cn also includes these 6	892
835	Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang,	splits, with each split consisting of a training set	893
836	Xuebo Liu, Lidia S. Chao, and Min Zhang. 2024.	and a validation set.	894
837	Who wrote this? the key to zero-shot llm-generated	Table 7 presents the total data volume	895
838	text detection is gecscore . <i>ArXiv</i> , abs/2405.04286.	distribution of the generalized MAGA dataset.	896
839	Kangxi Wu, Liang Pang, Huawei Shen, Xueqi	Each split contains 60,000 machine-generated texts	897
840	Cheng, and Tat-Seng Chua. 2023. LLMDet: A	and 60,000 human-written texts. It is important to	898
841	third party large language models generated text	note that the 60,000 human-written texts are reused	899
842	detection tool . In <i>Findings of the Association</i>	across all splits—i.e., the human-written texts are	900
843	<i>for Computational Linguistics: EMNLP 2023</i> ,	identical for every split—whereas the machine-	901
844	pages 2113–2133, Singapore. Association for	generated texts are distinct across splits, with one	902
845	Computational Linguistics.	exception: For the extra-RLDF-CMD split, since	903
846	Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian	no other alignment methods are integrated and	904
847	Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2023. On	RLDF-CMD is only applicable to non-commercial	905
848	the generalization of training-based chatgpt detection	large models, 30,000 machine-generated texts	906
849	methods . <i>Preprint</i> , arXiv:2310.01307.	in this split (specifically those generated by	907
850	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,	commercial large models) are sourced from the	908
851	Binyuan Hui, Bo Zheng, Bowen Yu, Chang	MGB split.	909
852	Gao, Chengen Huang, Chenxu Lv, Chujie Zheng,		
853	Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,		
854	Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and		
855	41 others. 2025. Qwen3 technical report . <i>ArXiv</i> ,		
856	abs/2505.09388.		

Domain	GPT-4o-mini	Gemini-2.0-flash	DeepSeek-V3	Qwen3-plus	Mistral-Medium	Hunyuan-TurboS	Llama-3.1-8B Instruct	gemma-3-12b-it	DeepSeek-R1-0528 Qwen3-8B	Qwen3-8B	Minstral-8B Instruct-2410	Hunyuan-7B Instruct	Machine.Total	Human
Reddit	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
S2ORC	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Wikipedia	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
wikiHow	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Trustpilot Reviews	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Amazon Reviews	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Yahoo Answers	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Natural Questions	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
CC News	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
NPR News	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
EN.Total	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	60000	60000
Baidu Tieba (百度贴吧)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Weibo Review (微博评论)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Rednote Review (小红书评论)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
CSL (中文核心期刊摘要)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Baidu Baike (百度百科)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Diapiing (大众点评)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Douban Review (豆瓣评论)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Baidu Zhidao (百度知道)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
Zhifu (知乎)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
CLTS (澎湃新闻)	500	500	500	500	500	500	500	500	500	500	500	500	6000	6000
CN.Total	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	60000	60000

Table 5: Training Set Instance Distribution of MAGA Dataset Splits (MGB, MAGA and every split of MAGA-extra). The table summarizes sample counts across English and Chinese domains, including per-model samples, machine sample totals and human samples, with EN.Total and CN.Total for aggregated domain counts.

Domain	GPT-4o-mini	Gemini-2.0-flash	DeepSeek-V3	Qwen3-plus	Mistral-Medium	Hunyuan-TurboS	Llama-3.1-8B Instruct	gemma-3-12b-it	DeepSeek-R1-0528 Qwen3-8B	Qwen3-8B	Minstral-8B Instruct-2410	Hunyuan-7B Instruct	Machine.Total	Human
Reddit	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
S2ORC	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Wikipedia	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
wikiHow	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Trustpilot Reviews	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Amazon Reviews	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Yahoo Answers	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Natural Questions	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
CC News	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
NPR News	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
EN.Total	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	12000	12000
Baidu Tieba (百度贴吧)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Weibo Review (微博评论)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Rednote Review (小红书评论)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
CSL (中文核心期刊摘要)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Baidu Baike (百度百科)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Diapiing (大众点评)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Douban Review (豆瓣评论)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Baidu Zhidao (百度知道)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
Zhifu (知乎)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
CLTS (澎湃新闻)	100	100	100	100	100	100	100	100	100	100	100	100	1200	1200
CN.Total	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	12000	12000

Table 6: Validation Set Instance Distribution of MAGA Dataset Splits (MGB, MAGA and every split of MAGA-extra). The table summarizes sample counts across English and Chinese domains, including per-model samples, machine sample totals and human samples, with EN.Total and CN.Total for aggregated domain counts.

Our MAGA dataset is highly balanced. Tables 5 and 6 respectively show the data volume generated by each model in each domain. For example, in a single training split, each model generates 500 texts per domain.

Table 8 lists the abbreviated names of the models we used. Throughout the text and in the dataset annotations, the model names correspond to the entries in the Model column of Table 8, while the Identifier column denotes the actual model names passed during usage (e.g., when calling APIs for commercial large models). A key note: abbreviated names are used for commercial large models, whereas for non-commercial large models, the names used in annotations are identical to those used in actual deployment.

Table 9 outlines the annotation schema of our dataset. Compared with previous work, our annotations are significantly more detailed.

B Domain Details

Reddit (sentence-transformers, 2021) This dataset contains Reddit posts along with their

corresponding titles. We chose this dataset due to its highly informal, conversational style and diverse range of topics, which we hypothesize will pose challenges for detecting AI-generated content due to the lack of strict structural constraints.

S2ORC (Lo et al., 2020) This dataset consists of scholarly paper metadata and full-text content from a wide range of academic disciplines. We include this domain because it tests models’ ability to generate and detect text with formal academic tone, precise terminology, and logical argumentation, which relies heavily on domain-specific knowledge.

Wikipedia (Aaditya Bhat, 2023) This dataset contains introductory sections of Wikipedia articles covering diverse topics such as history, science, and culture. This domain is challenging as it requires accurate recall of factual information and neutral, encyclopedic writing style, making it a strong testbed for evaluating the factuality of AI-generated text.

Language	Split	Training Set			Validation Set			Total		
		Machine	Human	Total	Machine	Human	Total	Machine	Human	Total
English	MGB	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-extra-RLDF-CMD	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-extra-BPO	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-extra-roleplaying	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-extra-self-refine	60k	60k	120k	12k	12k	24k	72k	72k	144k
	Grand Total		330k	60k	390k	66k	12k	78k	396k	72k
Chinese	MGB-cn	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-cn	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-cn-extra-RLDF-CMD	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-cn-extra-BPO	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-cn-extra-roleplaying	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-cn-extra-self-refine	60k	60k	120k	12k	12k	24k	72k	72k	144k
	Grand Total		330k	60k	390k	66k	12k	78k	396k	72k
Grand Total		660k	120k	780k	132k	24k	156k	792k	144k	936k

Table 7: Overall Instance Distribution of MAGA Dataset by Language and Split (MGB, MAGA, MAGA-extra Alignment Methods). The table summarizes sample counts across training and validation sets, including machine-generated, human-annotated and aggregated totals, with notes that human samples are shared across all splits and partial machine data of MAGA-extra-RLDF-CMD is derived from MGB.

Model	Identifier
GPT-4o-mini	gpt-4o-mini-2024-07-18
Gemini-2.0-flash	models/gemini-2.0-flash
DeepSeek-V3	deepseek-v3
Qwen3-plus	qwen-plus-2025-04-28
Mistral-Medium	mistral-medium
Hunyuan-TurboS	hunyuan-turbos-20250313
—	Llama-3.1-8B-Instruct
—	gemma-3-12b-it
—	DeepSeek-R1-0528-Qwen3-8B
—	Qwen3-8B
—	Ministral-8B-Instruct-2410
—	Hunyuan-7B-Instruct

Table 8: The generative models used in our project

Field	Type
id	str(uuid4)
title	str
text	str
domain	str
human_source_id	str(uuid4)
prompt_id	str(uuid4)
system_prompt	str
user_prompt	str
model	str
label	int(0:human;1:machine)
temperature	float
top_p	float
top_k	int
repetition_penalty	float

Table 9: Annotation Data Schema of MAGA Dataset

wikiHow (Koupae and Wang, 2018) This dataset comprises step-by-step instructional guides for various tasks, from daily life skills to complex

technical operations. We selected this domain because it demands clear, sequential logical structure and practical common sense reasoning, which is difficult for models to replicate without generating incoherent or unworkable steps.

Trustpilot Reviews (Kerassy, 2025) This dataset contains user-generated reviews of businesses, products, and services. The reviews exhibit a wide range of sentiment polarity and subjective opinions, and we hypothesize that AI-generated reviews may lack the nuanced personal experiences that characterize authentic user feedback.

Amazon Reviews (Ni et al., 2019) This dataset consists of customer reviews for products sold on Amazon, spanning multiple categories such as electronics, apparel, and household goods. The reviews combine subjective opinions, specific product feature evaluations, and personal usage scenarios, testing models’ ability to generate realistic, detailed consumer feedback.

Yahoo Answers (sentence-transformers, 2024c) This dataset contains user-submitted questions and their corresponding answers on a broad array of topics. We include this domain due to its mix of casual conversational tone and informative content; AI-generated answers may struggle to match the diverse, sometimes idiosyncratic problem-solving approaches of human respondents.

Natural Questions (sentence-transformers,

953
954
955

956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984

985	2024a)	This dataset comprises real user questions posed to Google Search and their corresponding answers extracted from Wikipedia. This domain is designed to test models' ability to generate accurate, concise factual responses to information-seeking queries, with a focus on alignment between questions and answer relevance.	1033
986			1034
987			1035
988			1036
989			1037
990			1038
991			1039
992			1040
993	CC News	(Hamborg et al., 2017) This dataset contains news articles sourced from Common Crawl, covering global events across politics, economy, technology, and more. The articles are collected from a vast number of online news outlets, providing a diverse sample of journalistic writing styles and testing models' capacity to generate factually consistent news content.	1041
994			1042
995			1043
996			1044
997			1045
998			1046
999			1047
1000			1048
1001	NPR News	(sentence-transformers, 2024b) This dataset consists of news reports and feature articles from National Public Radio, known for its in-depth, balanced coverage of domestic and international issues. We chose this domain because it requires formal journalistic tone, rigorous fact-checking, and narrative coherence, highlighting the challenge of AI-generated text in mimicking professional news writing.	1049
1002			1050
1003			1051
1004			1052
1005			1053
1006			1054
1007			1055
1008			1056
1009			1057
1010	Baidu Tieba	(HongzheBi, 2023) This dataset contains posts and comments from Baidu Tieba, a Chinese-language online forum with topic-specific communities. The content features colloquial Chinese expressions, slang, and community-specific jargon, making it challenging for models to replicate the authentic conversational style of Chinese internet users.	1058
1011			1059
1012			1060
1013			1061
1014			1062
1015			1063
1016			1064
1017			1065
1018	Weibo Review	(LZYFirecn, 2025) This dataset comprises user comments from Weibo, a major Chinese microblogging platform. The reviews are typically concise, sentiment-laden, and often tied to trending topics, testing models' ability to generate short-form, context-aware Chinese text that captures the nuances of online public opinion.	1066
1019			1067
1020			1068
1021			1069
1022			1070
1023			1071
1024			1072
1025	Rednote Review	(Jingyi Zheng, 2025) This dataset consists of user-generated reviews and lifestyle sharing content from Xiaohongshu (Rednote), a Chinese social e-commerce platform. The content blends product evaluations with personal experience narratives and visual description cues, and we hypothesize that AI-generated content may lack the authentic, relatable details of human posts.	1073
1026			1074
1027			1075
1028			1076
1029			1077
1030			1078
1031			1079
1032			1080
	CSL	(Li et al., 2022) This dataset contains abstracts from Chinese Core Journals, covering academic research in fields such as humanities, social sciences, natural sciences, and engineering. We include this domain as it requires mastery of formal Chinese academic writing conventions, precise disciplinary terminology, and logical research framework presentation.	1033
			1034
			1035
			1036
			1037
			1038
			1039
			1040
	Baidu Baike	(Iars1234, 2025) This dataset contains entry content from Baidu Baike, a Chinese-language online encyclopedia. Similar to Wikipedia, this domain demands neutral, factual writing and accurate representation of Chinese-specific cultural, historical, and scientific knowledge, posing challenges for models' cross-lingual factual recall ability.	1041
			1042
			1043
			1044
			1045
			1046
			1047
			1048
	Dianping	(twn39, 2025) This dataset comprises user reviews of restaurants, shops, and local services from Dianping, a Chinese local life service platform. The reviews focus on detailed experience descriptions, such as food taste, service quality, and store environment, testing models' ability to generate realistic, location-specific Chinese consumer feedback.	1049
			1050
			1051
			1052
			1053
			1054
			1055
			1056
	Douban Review	(dirtycomputer, 2023) This dataset contains user reviews of movies, books, music, and TV shows from Douban, a Chinese cultural community platform. The reviews are characterized by subjective, insightful critical analysis and emotional expression, making it difficult for AI-generated text to replicate the depth of human cultural perception and personal taste.	1057
			1058
			1059
			1060
			1061
			1062
			1063
			1064
	Baidu Zhidao	(HongzheBi, 2023) This dataset consists of user questions and answers from Baidu Zhidao, a Chinese Q&A platform covering daily life, study, work, and other topics. The content features a mix of formal informative answers and casual conversational responses, testing models' ability to adapt to diverse Chinese language interaction styles.	1065
			1066
			1067
			1068
			1069
			1070
			1071
			1072
	Zhihu	(Liu et al., 2020) This dataset contains questions and in-depth answers from Zhihu, a Chinese knowledge-sharing platform. The answers often include logical argumentation, personal experience sharing, and professional analysis, and we hypothesize that AI-generated content may lack the unique perspectives and detailed reasoning of human experts.	1073
			1074
			1075
			1076
			1077
			1078
			1079
			1080

CLTS (Liu et al., 2020) This dataset comprises news articles from The Paper (CLTS), a Chinese mainstream digital news outlet, covering politics, society, culture, and technology. The articles adhere to rigorous journalistic standards and focus on in-depth investigative reporting, testing models’ ability to generate formal, factually accurate Chinese news content.

C Generator Details

GPT-4o-mini (Hurst et al., 2024) is a lightweight OpenAI GPT-4o variant, a decoder-only model fine-tuned on diverse text corpora. Accessible via `ChatCompletion`, it balances performance and efficiency for prompt-based tasks.

Gemini-2.0-flash (Google DeepMind, 2024) is Google DeepMind’ s multimodal model for fast text/image/audio inference. Trained on mixed web and multimodal data, it enhances efficiency while maintaining strong reasoning performance.

DeepSeek-V3 (DeepSeek-AI et al., 2024) is DeepSeek AI’ s decoder-only model, optimized for reasoning, coding and math. Trained on academic/programming datasets with an extended context window, it is open-source for technical domains.

Qwen3-plus (Yang et al., 2025) is Alibaba Cloud’ s open-source decoder-only model, upgraded from Qwen series. Trained on diverse text/code and fine-tuned for dialogue, it delivers robust general and domain-specific performance.

Mistral-Medium (Mistral AI, 2025) is Mistral AI’ s mid-sized decoder-only model. Trained on filtered web/technical texts, it balances speed and performance, excelling in few-shot and real-time generation.

Hunyuan-TurboS (Liu et al., 2025) is Tencent’ s lightweight Hunyuan variant, optimized for low-latency conversations. Trained on Chinese/English web/dialogue data, this closed-source model is API-accessible for daily interactions.

Llama-3.1-8B-Instruct (AI@Meta, 2024) is Meta’ s 8B-parameter instruction-tuned decoder-only model. Pre-trained on web/books/Wikipedia and fine-tuned for intent alignment, it is open-source and commercially viable.

gemma-3-12b-it (Team, 2025) is Google’ s 12B-parameter open-weight instruction-tuned model. Trained on high-quality text/code/educational data, it supports multilingual tasks with strong reasoning and code generation.

DeepSeek-R1-0528-Qwen3-8B (DeepSeek-AI et al., 2025) is DeepSeek’ s fine-tuned Qwen3-8B variant. Optimized on reasoning/math datasets, this open-source model focuses on precise logical inference for research.

Qwen3-8B (Yang et al., 2025) is Alibaba’ s 8B-parameter open-source decoder-only model. Pre-trained on multilingual text/code, it serves as a versatile base for fine-tuning.

Mistral-8B-Instruct-2410 (Mistral AI Team et al., 2024) is a lightweight instruction-tuned model. Trained on instruction/conversation data, it prioritizes fast inference, suitable for edge/low-resource deployment.

Hunyuan-7B-Instruct (Tencent, 2025) is Tencent’ s 7B-parameter instruction-tuned model. Pre-trained on Chinese/English text and fine-tuned for dialogue, it is designed for chatbots and content generation.

D RLDF Details

First, we define key notations for consistent formulation:

\mathcal{M}_θ : The large language model (LLM) to be fine-tuned, with parameters θ ; its generation distribution is $\pi_\theta(y|x) = \mathcal{M}_\theta(y|x)$, where x is the input prompt and y is the generated text.

\mathcal{D}_ϕ : The detector (serving as the reward model, RM) with parameters ϕ , outputting $\mathcal{D}_\phi(y) \in [0, 1]$ (the probability that y is identified as machine-generated).

$H = \{h_i\}_{i=1}^N$: The human-written text (HWT) set, labeled 0 (non-machine-generated).

$M = \{m_i\}_{i=1}^N$: The human-aligned machine-generated text (MGT) set (from the LLM), labeled 1 (machine-generated).

$\mathcal{X}_\mathcal{D}$: The input space of domain \mathcal{D} (e.g., $\mathcal{X}_{\text{DA}}/\mathcal{X}_{\text{DB}}$ for cross-domain scenarios).

$\mathcal{M}_{\text{MA}}/\mathcal{M}_{\text{MB}}$: Two benchmark LLMs for cross-model scenarios.

\mathcal{M}_{B} : A commercial LLM, whose generated text is $M_{\text{B}} = \{m_{\text{B},i}\}_{i=1}^K$.

$r_{\mathcal{D}}(y) = 1 - \mathcal{D}_\phi(y)$: The reward function (higher values mean y is more human-aligned and

harder to detect).

RLDF RLDF achieves human alignment and anti-detection capability of LLM-generated text via a multi-round adversarial loop: "LLM reinforcement fine-tuning \rightarrow detector generalization fine-tuning". For LLM fine-tuning, we use Generative Reward Policy Optimization (GRPO) to optimize the generation policy π_θ with \mathcal{D}_ϕ 's feedback as reward, plus a KL divergence constraint for training stability:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{X}, y \sim \pi_\theta(y|x)} [\hat{A}_{\theta_{\text{old}}}(x, y) \cdot \log \pi_\theta(y|x)] - \beta \cdot \mathbb{E}_{x \sim \mathcal{X}} \text{KL}(\pi_{\theta_{\text{old}}}(\cdot|x) \parallel \pi_\theta(\cdot|x)) \quad (1)$$

Here, $\hat{A}_{\theta_{\text{old}}}(x, y) = r_{\mathcal{D}}(y) - \mathbb{E}_{y' \sim \pi_{\theta_{\text{old}}}(y'|x)} r_{\mathcal{D}}(y')$ is the advantage function (measuring y 's reward gain over the old policy $\pi_{\theta_{\text{old}}}$), $\beta > 0$ is the KL penalty coefficient, and θ_{old} is the pre-update LLM parameter. After GRPO fine-tuning, the LLM produces human-aligned $M = \{m_i\}_{i=1}^N$ ($m_i \sim \pi_\theta(y|x_i)$). For detector fine-tuning, we build the training set $\mathcal{T} = \{(h_i, 0)\}_{i=1}^N \cup \{(m_i, 1)\}_{i=1}^N$ (from H and M), then fine-tune \mathcal{D}_ϕ (RoBERTa-initialized) with binary cross-entropy loss:

$$\mathcal{L}_{\text{detector}}(\phi) = -\frac{1}{2N} \sum_{i=1}^N \left[\log(1 - \mathcal{D}_\phi(h_i)) + \log \mathcal{D}_\phi(m_i) \right] \quad (2)$$

The fine-tuned \mathcal{D}_ϕ acts as a new reward model, and the loop repeats for adversarial capability iteration.

RLDF-CD RLDF-CD targets cross-domain scenarios, enabling LLM human alignment across domains via "domain-specific detectors \rightarrow cross-reward fine-tuning". To build domain-specific detectors: 1) For \mathcal{X}_{DA} , construct $\mathcal{T}_{\text{DA}} = \{(h_{\text{DA},i}, 0)\} \cup \{(m_{\text{DA},i}, 1)\}$ (from H_{DA} and M_{DA}) and fine-tune to get \mathcal{D}_{DA} ; 2) For \mathcal{X}_{DB} , build \mathcal{T}_{DB} (from H_{DB} and M_{DB}) and fine-tune to get \mathcal{D}_{DB} . For cross-domain GRPO fine-tuning, we use cross-domain detectors as rewards (\mathcal{D}_{DB} for \mathcal{X}_{DA} , \mathcal{D}_{DA} for \mathcal{X}_{DB}), so the objective extends to:

$$\mathcal{L}_{\text{GRPO-CD}}(\theta) = \sum_{\mathcal{D} \in \{\text{DA}, \text{DB}\}} \mathbb{E}_{x \sim \mathcal{X}_{\mathcal{D}}, y \sim \pi_\theta(y|x)} [\hat{A}_{\theta_{\text{old}}}(x, y) \cdot \log \pi_\theta(y|x)] - \beta \cdot \sum_{\mathcal{D} \in \{\text{DA}, \text{DB}\}} \mathbb{E}_{x \sim \mathcal{X}_{\mathcal{D}}} \text{KL}(\pi_{\theta_{\text{old}}}(\cdot|x) \parallel \pi_\theta(\cdot|x)) \quad (3)$$

where $\hat{A}_{\theta_{\text{old}}}(x, y) = r_{\overline{\mathcal{D}}}(y) - \mathbb{E}_{y' \sim \pi_{\theta_{\text{old}}}(y'|x)} r_{\overline{\mathcal{D}}}(y')$ ($\overline{\mathcal{D}}$ denotes the cross domain of the current domain).

RLDF-CM RLDF-CM focuses on cross-model scenarios, improving LLM's generalized human alignment via "model-specific detectors \rightarrow multi-model cross-reward \rightarrow commercial LLM enhancement". To build model-specific detectors: 1) For \mathcal{M}_{MA} , construct $\mathcal{T}_{\text{MA}} = \{(h_i, 0)\} \cup \{(m_{\text{MA},i}, 1)\}$ (from M_{MA} and H) and fine-tune to get \mathcal{D}_{MA} ; 2) For \mathcal{M}_{MB} , build \mathcal{T}_{MB} (from M_{MB}) and fine-tune to get \mathcal{D}_{MB} ; 3) Expand $\mathcal{T}_{\text{MA}}/\mathcal{T}_{\text{MB}}$ with commercial LLM-generated M_{B} (resulting in $\mathcal{T}'_{\text{MA}}/\mathcal{T}'_{\text{MB}}$) to enhance detector generalization. For cross-model GRPO fine-tuning, we use cross-model detectors as rewards (\mathcal{D}_{MB} for \mathcal{M}_{MA} -style generation, \mathcal{D}_{MA} for \mathcal{M}_{MB} -style generation), so the objective is:

$$\mathcal{L}_{\text{GRPO-CM}}(\theta) = \sum_{\mathcal{M} \in \{\text{MA}, \text{MB}\}} \mathbb{E}_{x \sim \mathcal{X}, y \sim \pi_\theta^{\mathcal{M}}(y|x)} [\hat{A}_{\theta_{\text{old}}}(x, y) \cdot \log \pi_\theta^{\mathcal{M}}(y|x)] - \beta \cdot \sum_{\mathcal{M} \in \{\text{MA}, \text{MB}\}} \mathbb{E}_{x \sim \mathcal{X}} \text{KL}(\pi_{\theta_{\text{old}}}^{\mathcal{M}}(\cdot|x) \parallel \pi_\theta^{\mathcal{M}}(\cdot|x)) \quad (4)$$

Here, $\pi_\theta^{\mathcal{M}}(y|x)$ is the target LLM's policy imitating \mathcal{M} 's style, and $\hat{A}_{\theta_{\text{old}}}(x, y) = r_{\overline{\mathcal{M}}}(y) - \mathbb{E}_{y' \sim \pi_{\theta_{\text{old}}}^{\mathcal{M}}(y'|x)} r_{\overline{\mathcal{M}}}(y')$ ($\overline{\mathcal{M}}$ denotes the cross model of the current imitated model).

Final Selection We ultimately adopted the fully enhanced RLDF-CMD to construct the MAGA dataset. Based on ablation experiments, we divided the domains into two groups: DA includes Wikipedia, wikiHow, CC News, NPR News, and S2ORC, while DB consists of Reddit, Trustpilot Reviews, Amazon Reviews, Yahoo Answers, and Natural Questions. We also split the models into two groups: \mathcal{M}_{MA} comprises Qwen3-plus, Qwen3-8B, DeepSeek-V3, DeepSeek-R1-0528-Qwen3-8B, Hunyuan-TurboS, and Hunyuan-7B-Instruct, whereas \mathcal{M}_{MB} includes Llama-3.1-8B-Instruct, Mistral-Medium, Ministral-8B-Instruct-2410, Gemini-2.0-flash, GPT-4o-mini, and DeepSeek-V3. A cross-reward approach was employed here.

It should be noted that we alternate the cross configuration once per round of RLDF-CMD, ensuring each model can generate content for every domain. Specifically, in each adversarial round, we first set the groups as $G_{\text{A}} = \mathcal{M}_{\text{MA}} + \text{DA}$ and $G_{\text{B}} = \mathcal{M}_{\text{MB}} + \text{DB}$, then switch to $G_{\text{A}} = \mathcal{M}_{\text{MA}} + \text{DB}$ and $G_{\text{B}} = \mathcal{M}_{\text{MB}} + \text{DA}$ for the same round.

E Decoding Strategies

Decoding strategies play a crucial role in controlling the behavior of large language models (LLMs) during text generation, by modulating the probability distribution over the vocabulary to balance diversity and coherence.

Our goal is to generate more aligned machine texts. We adopted multiple decoding strategies, as detailed in Table 10. In fact, for open-source models, the official recommended decoding parameters are provided in the file `generation_config.json`, which can yield optimal or near-optimal generation quality. For commercial models, such as Qwen3-Plus (Yang et al., 2025), the recommended parameters are specified in their Technical Report. Since our objective is to generate more aligned machine texts, we adopted these parameters. When the recommended parameters are unknown, based on empirical experience, we set top k to -1 and keep the rest of the parameters at their default value of 1, which generally produces diverse and relatively stable results.

E.1 Temperature

Temperature $T > 0$ is a scaling factor that adjusts the sharpness of the token probability distribution. Given raw logits $x_i \in \mathbb{R}^d$ for token i , the probability p_i is computed as:

$$p_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)}$$

A higher T flattens the distribution to promote diverse generation, while a lower T sharpens the distribution, favoring high-probability tokens for more deterministic outputs.

E.2 Top- k and Top- p Sampling

Top- k and top- p (nucleus sampling) are two constrained sampling strategies that restrict the candidate token set to reduce irrelevant generation.

For top- k sampling, the model only samples from the k most probable tokens (sorted by $\exp(x_i/T)$). Let S_k denote this set of top- k tokens. The adjusted probability is:

$$p_i = \begin{cases} \frac{\exp(x_i/T)}{\sum_{j \in S_k} \exp(x_j/T)} & \text{if } i \in S_k \\ 0 & \text{otherwise} \end{cases}$$

A larger k increases diversity, while $k = 1$ reduces to greedy decoding.

For top- p sampling, the candidate set S_p is defined as the smallest subset of tokens whose cumulative probability exceeds a threshold $p \in (0, 1]$. The probability distribution is:

$$p_i = \begin{cases} \frac{\exp(x_i/T)}{\sum_{j \in S_p} \exp(x_j/T)} & \text{if } i \in S_p \\ 0 & \text{otherwise} \end{cases}$$

This strategy adapts the size of the candidate set dynamically based on probability distribution.

E.3 Penalty

Penalty mechanisms are designed to discourage repetitive token generation by modifying the logit-based probability distribution, with three common variants: repetition penalty, presence penalty, and frequency penalty. Let g represent the list of previously generated tokens.

Repetition penalty (Keskar et al., 2019) applies a multiplicative adjustment to the logits of tokens in g . The modified probability is:

$$p_i = \frac{\exp(x_i/(T \cdot I(i \in g)))}{\sum_j \exp(x_j/(T \cdot I(j \in g)))}$$

where $I(c) = \theta$ if condition c is true (token i is in g) and $I(c) = 1$ otherwise, with $\theta > 1$ as the penalty coefficient.

Presence penalty (implemented by OpenAI) uses an additive penalty to reduce the likelihood of any token that has appeared in g , regardless of occurrence frequency:

$$p_i = \frac{\exp((x_i/T) - \alpha \cdot I(i \in g))}{\sum_j \exp((x_j/T) - \alpha \cdot I(j \in g))}$$

where $\alpha > 0$ is the penalty weight, and $I(i \in g)$ is an indicator function that equals 1 if token i is in g , and 0 otherwise.

Frequency penalty extends presence penalty by penalizing tokens in proportion to their occurrence frequency in g . Let f_i denote the frequency of token i in g . The adjusted probability is:

$$p_i = \frac{\exp((x_i/T) - \beta \cdot f_i \cdot I(i \in g))}{\sum_j \exp((x_j/T) - \beta \cdot f_j \cdot I(j \in g))}$$

where $\beta > 0$ is the penalty weight, so tokens with higher frequency receive stronger penalties.

F Detector Details

R-B GPT2 (Solaiman et al., 2019) A RoBERTa-based detector fine-tuned on GPT2 open-domain

Model	Temperature	Top-p	Top-k	Repetition Penalty
GPT-4o-mini	0.6	1	-1	1
Gemini-2.0-flash	1	1	-1	1
DeepSeek-V3	1	1	-1	1
Qwen3-plus*	0.7	0.8	-1	1*
Mistral-Medium	1	1	-1	1
Hunyuan-TurboS	1	1	-1	1
Llama-3.1-8B-Instruct	0.6	0.9	-1	1
gemma-3-12b-it	1	0.95	64	1
DeepSeek-R1-0528-Qwen3-8B	0.6	0.95	-1	1
Qwen3-8B	0.6	0.95	20	1
Ministral-8B-Instruct-2410	1	1	-1	1
Hunyuan-7B-Instruct	0.7	0.8	20	1.05

Table 10: Decoding strategies used for each generative model. Qwen3-plus* additionally uses a presence penalty of 1.5, which is not listed in the table. A value of -1 indicates that the corresponding parameter is not used (i.e., disabled) for that model.

1312 outputs, which were generated via greedy decoding,
1313 top-k=50 sampling, and full random sampling.
1314 As a long-standing baseline, both base and large
1315 variants are used for comparison.

1316 **R-B CGPT** (Guo et al., 2023) A RoBERTa-base
1317 model fine-tuned on the HC3 dataset ($\approx 27,000$
1318 human-ChatGPT answer pairs across multiple
1319 domains). Accessed via HuggingFace datasets.

1320 **RADAR** (Hu et al., 2023) A fine-tuned Vicuna
1321 7B (derived from LLaMA 7B) trained under a
1322 generative adversarial framework with a paraphrase
1323 model. It distinguishes paraphrased text, WebText
1324 human text, and original LM outputs. Accessed via
1325 HuggingFace.

1326 **GLTR** (Gehrmann et al., 2019) Originally an
1327 auxiliary interface for human detection, now a
1328 standard robustness baseline. It evaluates token
1329 likelihood via an LM, bins tokens by likelihood,
1330 and uses these as features. Adopts default settings
1331 (rank=10 cutoff, GPT2 small).

1332 **FastDetectGPT** (Bao et al., 2023) An optimized
1333 DetectGPT variant with 340x faster inference and
1334 unchanged accuracy. Uses default GPT-Neo-2.7B
1335 (scoring) and GPT-J-7B (reference) models, neither
1336 of which was used for dataset generation, ensuring
1337 fair comparison.

1338 **Binoculars** (Hans et al., 2024) Uses the ratio of
1339 perplexity to cross-entropy between two similar
1340 LMs as the detection metric. Adopts official default
1341 Falcon 7B and Falcon 7B Instruct models; neither
1342 was used for dataset generation, ensuring fairness.

1343 **LLMDet** (Wu et al., 2023) Computes proxy-
1344 perplexity (n-gram sampling-based approximation)
1345 of input text via 10 small LMs, using these as
1346 detection features. None of the LMs were used
1347 for dataset generation, ensuring fair comparison.

1348 **SCRN** (Huang et al., 2024) A robust detector
1349 with a reconstruction network for denoising
1350 and siamese calibration. It resists adversarial
1351 perturbations across in-domain, cross-domain, and
1352 mixed-source scenarios.

1353 **DETree** (He et al., 2025) A representation
1354 learning-based detector that models text category
1355 relationships via a Hierarchical Affinity Tree and
1356 contrastive loss. Trained on RealBench, it excels
1357 at hybrid text detection and out-of-distribution
1358 generalization.

1359 **DALD** (Zeng et al., 2024) A distribution-aligned
1360 black-box detector that fine-tunes surrogates with
1361 small-scale LLM outputs. Enhances zero-shot
1362 methods and adapts to model updates and non-
1363 English texts.

1364 **GECScore** (Wu et al., 2024) A zero-shot
1365 detector using grammar error correction similarity.
1366 It distinguishes texts by comparing input with
1367 corrected versions, boasting strong generalization
1368 and paraphrase attack resistance.

1369 It should be noted that the base models
1370 employed by nearly all detectors are primarily
1371 competent only in English. In some cases, they
1372 lack any Chinese language capability whatsoever;
1373 in others, their Chinese recognition ability is even

weaker than that for low-resource languages, with no practical proficiency in Chinese. Among these, R-B CGPT offers a standalone Chinese version, while GECScore can be directly applied to Chinese detection tasks simply by translating the prompt text into Chinese. Notably, Binoculars is capable of Chinese recognition, and despite lacking functional Chinese language proficiency, we found that its detection performance is surprisingly satisfactory. As a supplementary addition for training-based detectors, we have additionally incorporated RoBERTa-MPU-zhv3 (Tian et al., 2023). The four aforementioned models are utilized by us for Chinese detection tasks.

RoBERTa-MPU-zhv3 (Tian et al., 2023) A multiscale detector with length-sensitive PU loss and text multiscaling. It improves short-text detection while maintaining long-text performance, supporting both English and Chinese datasets.

G Prompt Details

We present our English prompts in Table 12 and the Chinese version in Table 13. The prompts in these tables are the original User Prompts, which are used by MGB, extra-RLDF-CMD, extra-roleplaying, and for the first round of extra-self-refine. BPO automatically optimizes the prompts; the first round of extra-BPO and MAGA adopts the prompts optimized by BPO.

Tables 14 and 15 display the English and Chinese Self Refine prompts, respectively. Based on the content generated in the first round, we input a Feedback Prompt to obtain feedback through multi-turn conversations, which is then embedded into the Refine Prompt to generate the text after Self Refine. The Self Refine prompts are applied in extra-self-refine and MAGA.

Table 16 shows the Role-Playing prompts, which are input as system prompts and used by extra-roleplaying and MAGA. For specific examples of prompt usage, please refer to the case study in Table 36.

H Multidimensional Analysis Details

Content Similarity the adopted metrics include ROUGE (for abstract and translation tasks), BLEU (for translation tasks), and METEOR (for multi-dimensional matching to compensate for synonym-matching gaps), while CHRf (character-level matching for low-resource language translation and short texts) and CIDEr

(for image captioning) are excluded. Specifically, ROUGE-1 (unigram-based) is suited for short-text semantic coverage evaluation, ROUGE-2 (bigram-based) for long-text phrase-level semantic coherence assessment, and ROUGE-L (LCS-based) for word order-preserving tasks like machine translation. For metric prioritization, recall (R) takes precedence over F1 and precision (P) in text summarization to ensure full coverage of core information from the source text; precision (P) is prioritized over F1 and recall (R) in machine translation and question answering to guarantee output accuracy; F1 is the optimal choice for general content generation to balance coverage and accuracy. Thus, we finally select ROUGE-2-F1 as the core reference metric.

Semantic Similarity BERTScore is adopted, with F1 as the primary observation indicator, which is the most common practice.

Lexical Diversity no universal standard metric exists, but Type-Token Ratio (TTR, the most basic lexical diversity metric) and Yule’s K (the core metric for detecting word repetition, which is typical of machine-generated text) are chosen, along with the 2-gram vocabulary size proposed by M4.

Text Readability Flesch-Kincaid Reading Ease (simple and intuitive for daily news scenarios), SMOG Index (effective for detecting obscure expressions in academic papers and manuals), and Dale-Chall Readability Score (based on basic vocabulary for children’s books and popular science texts) are adopted, while Flesch-Kincaid Grade Level and Gunning Fog Index are excluded.

Sentiment Consistency We aligned the model selected with HC3(Guo et al., 2023). Then we use cosine similarity selected as the optimal consistency calculation method. Rather than simply counting the percentages of negative, neutral, and positive texts, we more precisely calculated the average probability score for all texts in each category.

I Hyperparameters & training time

For the standalone training of RoBERTa, the hyperparameters are set as follows: num_epochs=1, total_batch_size=64, learning_rate=5e-5, optimizer=AdamW, lr_scheduler_type=“cosine”, and warmup_ratio=0.03. A key parameter here is the number of epochs, which we set to 1 based on the baseline configuration (epochs=1)

Metric	MGB-M	MAGA-M	HWT
Content Similarity			
ROUGE	0.0526	0.0387	1.00
BLEU	0.0235	0.0165	1.00
METEOR	0.2132	0.2007	1.00
Average	0.0964	0.0853	1.00
Semantic Similarity			
BERTScore	0.8252	0.8287	1.00
Lexical Diversity			
TTR	0.7893	0.7991	0.8271
Yule's K	74.99	52.02	100.18
Bigram Vocab Size	4261689	5467106	3253456
Normalized Score	0.8203	1.00	0.7165
Text Readability			
Flesch Kincaid	40.53	53.39	54.54
SMOG	14.10	12.15	11.91
Dale Chall	11.19	10.24	10.23
Cosine Similarity	0.9905	1.00	1.00
Sentiment Consistency			
Negative	0.3243	0.3439	0.3455
Neutral	0.4030	0.3894	0.3996
Positive	0.2727	0.2667	0.2549
Cosine Similarity	0.9989	1.00	1.00

Table 11: Multi-dimensional human-machine text analysis results across three models: MGB-M, MAGA-M, and HWT. The evaluation covers content similarity, semantic similarity, lexical diversity, text readability, and sentiment consistency metrics.

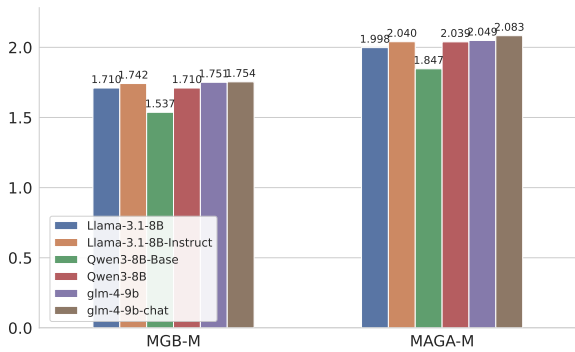


Figure 9: Direct log ppl computed on the base/chat large model.

reported in SemEval-M4 (Wang et al., 2024c). This choice is motivated by three reasons: (1) For RoBERTa-base, the training data with more than 100k samples is sufficiently abundant, eliminating the need for multiple epochs; (2) This ensures an extremely fair comparison; (3) It contributes to reliable generalization performance.

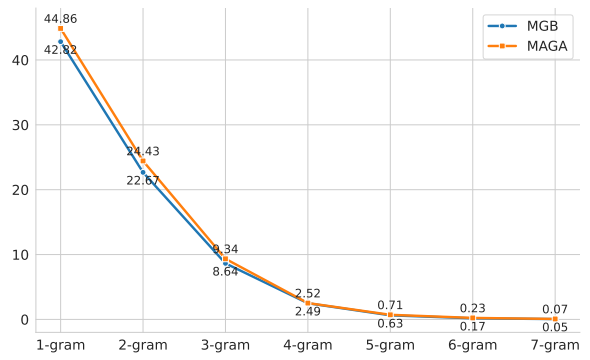


Figure 10: Ngram overlap rate with HWT.

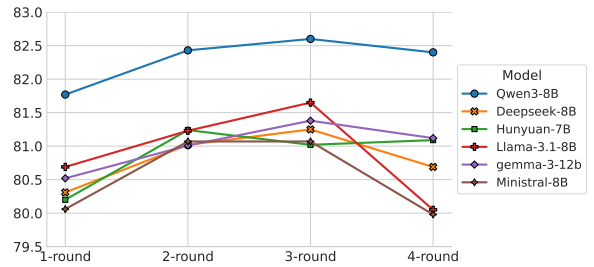


Figure 11: RLDF-CM generalization AUC by RLDF round. The CM group configuration use the most significant increase in Figure 13.

For RLDF-CMD, we adopt GRPO + LoRA for training on 8 L20 GPUs. The smallest LLM (7B) was trained for approximately 70 hours, while the largest LLM (12B) was trained for about 150 hours. Our `micro_batch_size` is set to 1, and `real_batch_size` is set to 1024 (`batch_size_per_device=32`, `rollout=4`), with the learning rate of $5e-6$. We conducted 3 training rounds of RLDF-CMD, with 1 epoch of training for each round.

J More Experiment Details

Table 18 presents the benchmark results on MAGA-cn, while Table 19 shows the generalization experiments in the Chinese version. Tables 21 to 24 report the benchmark results on various extra settings of MAGA. Tables 25 to 28 demonstrate the specific performance of the generalization experiments on different external datasets. Tables 29 to 35 illustrate the corresponding detailed experiments in the Chinese version, namely the benchmark results on various extra settings and the specific performance of the generalization experiments on different external datasets. Overall, our MAGA achieves consistent performance across all these experiments.

K Discussion

K.1 RLDF v.s. RADAR

Human-machine hybrid text (or rewritten/paraphrased text) is not genuine human text and thus requires detection. Traditionally, it has been regarded as an adversarial attack on machine text generation. Nevertheless, recent works such as MIXSET(Zhang et al., 2024) have explicitly pointed out that treating human-machine hybrid text as machine text for training binary classifiers is inappropriate, as it may introduce uncertainties. Instead, they propose a ternary classification task where human-machine hybrid text is considered a third category. RADAR is precisely the adversarial fine-tuning of paraphrasers and detectors.

If we see OpenAI’s R-B-GPT2 as the baseline, RADAR does not consistently outperform it across different benchmarks. In the experimental tables of Beemo(Artemova et al., 2025), the AUCROC score of RADAR also fluctuated significantly in different experiments. This further confirms that RADAR’s actual detection performance is highly uncertain.

RADAR is designed for paraphrased text detection, it also solves the infeasibility of direct adversarial fine-tuning (as mentioned in Section 3.5, due to RoBERTa’s overfitting to in-distribution data). Additionally, RADAR integrates RoBERTa’s fine-tuning into the PPO step, enabling highly automated and continuous performance improvement.

However, according to the HC-Var theory(Xu et al., 2023), RADAR’s approach remains highly constrained by the unreliable scoring caused by RoBERTa’s overfitting to irrelevant features. Furthermore, the initial insufficient experience of RoBERTa leads to unreliable convergence directions between the LLM and RoBERTa, also resulting in an extremely unstable training process.

Our RLDF is a decoupled general framework: we first fully train the detector, then fully train the large language model, and repeat this cycle while manually controlling the multi-round adversarial process to ensure reliability. Given the infeasibility of direct adversarial fine-tuning, we incorporate the HC-Var theory and leverage the characteristic that out-of-distribution detection primarily relies on relevant features, thus designing two types of cross-domain/cross-model cross rewards (RLDF-CD, RLDF-CM). Our proposed R-B MAGA achieves

stable improvements over our stronger baseline R-B MGB.

K.2 Why not Distill?

If we need more anthropomorphic text, why not directly perform knowledge distillation on commercial large models or simply use commercial large models?

Many previous studies, including M4(Wang et al., 2023b), have shown that increasing the diversity of large model selections, especially incorporating a large number of small-parameter large models, leads to significant improvements in detector learning.

CogLM(Wang et al., 2024a) once pointed out through relevant metric analysis that large models with different parameter sizes can be regarded as humans of different ages, each with its unique style.

For example, in terms of overall knowledge difficulty, richness, application scope, and effectiveness, university-level knowledge is certainly superior to high school-level knowledge, which in turn is superior to junior high school-level knowledge. However, this does not mean that only university examinations are meaningful; junior and senior high schools also have their own unique challenging problems.

The characteristic of RLDF is to enable various open-source large models to generate more anthropomorphic machine text while maintaining their original styles.

L Case Study

Table 36 presents a case study comparing various generation methods. We conducted a manual analysis to identify the advantages and disadvantages of each method. Machine traces are highlighted in red, while human traces are highlighted in green.

First, for the unenhanced MGB, there are numerous drawbacks. 1. The phrase "consistently perform behaviors" deliberately emphasizes the continuity principle of wikiHow and downplays one-time actions, which is a key signal of AI mimicking wikiHow’s "practicality-oriented" style. 2. The expression "physical, mental, and social function" is an obvious intentional imitation of the "underlying structure" of wikiHow health articles, directly adopting this three-dimensional formulation. 3. The clause "consuming adequate nutrition, engaging in regular

1605 movement, maintaining rest cycles, managing
1606 stress inputs, and avoiding known hazards" is a
1607 typical hint of AI's "detailed breakdown-style
1608 writing". 4. The sentence "Execute these actions
1609 repeatedly over time" re-emphasizes long-term
1610 persistence, and such "double emphasis" is a
1611 typical characteristic of AI-generated content.

1612 For BPO, roleplaying, and self-refine, the
1613 machine traces are similar: they deliberately
1614 emphasize continuity instead of clarifying it in
1615 more natural and casual ways, and they always
1616 decompose and stack some core principles of
1617 wikiHow for illustration. Examples include
1618 "nutrition, exercise, sleep, and mental health" from
1619 roleplaying and "focus on eating well, staying
1620 active, getting enough rest, and managing stress"
1621 from self-refine. Further elaboration is omitted
1622 here.

1623 Examples of human traces in BPO include the
1624 phrase "support your body and mind". When
1625 humans express health-related topics, they do not
1626 rigidly split "body" and "mind". Such concise
1627 and coherent phrases summarize the core meaning,
1628 which is not only natural and warm but also avoids
1629 the stacking of mechanical terminology. In terms
1630 of content, BPO is more in line with human
1631 expression habits.

1632 Examples of human traces in roleplaying are
1633 as follows: the sentence "Being healthy is about
1634 balancing various aspects of your life" carries the
1635 tone of college students chatting with their peers,
1636 and "It's not about perfection, but consistency" is
1637 a clear expression of the empathetic mindset of
1638 college students.

1639 Examples of human traces in self-refine include
1640 the sentence "You don't have to make huge
1641 changes all at once—small, consistent steps
1642 can lead to big improvements over time". This
1643 "empathize first, then guide" expression transforms
1644 the text from a "cold guideline" into "warm advice",
1645 which is more in line with the encouraging logic
1646 of interpersonal communication. As a method of
1647 self-reflection and refinement by large language
1648 models (LLMs), self-refine has a random rather
1649 than single improvement angle compared with
1650 the previous two methods, making it suitable for
1651 combination with other methods.

1652 MAGA integrates the advantages of multiple
1653 methods, thereby reducing machine traces in a
1654 more comprehensive and in-depth manner.

1655 Table 37 demonstrates the advantages of RLDF-
1656 CMD. In fact, the optimization of RLDF-CMD

1657 is more moderate, implicit, and quantitative.
1658 Benefiting from detector feedback, its most
1659 prominent feature is eliminating obvious machine
1660 hints. Our MAGA Additional Instruction is
1661 inspired by RAID (Dugan et al., 2024). Large
1662 language models (LLMs) suffer from a very
1663 annoying problem when generating such content.
1664 For example, when generating a Reddit post,
1665 even if you explicitly state "Do not repeat the
1666 title", the model still tends to add an opening
1667 sentence in unexpected ways. Our instruction
1668 has proven to be highly effective. Paradoxically,
1669 the more detailed your instructions are, the
1670 less likely LLMs are to follow them. Even so,
1671 LLMs—including commercial models—still
1672 frequently generate opening sentences, which is
1673 an extremely obvious machine hint. The greatest
1674 advantage of RLDF-CMD is that it can easily and
1675 almost completely solve this problem.

1676 In Table 38, we present a Chinese case study,
1677 which is basically consistent with the two previous
1678 English analyses.

	User Prompt	MAGA Additional Instruction
Reddit	Write just the body of a Reddit post titled "{title}". Do not repeat the title.	Do not repeat the title.
S2ORC	Write the abstract for the scientific paper titled "{title}". It is preferable not to start with "This paper".	It is preferable not to start with "This paper".
Wikipedia	Write the body of a Wikipedia article titled "{title}".	N/A
wikiHow	Write the body of a wikiHow article titled "{title}".	N/A
Trustpilot Reviews	Write the body of a Trustpilot review titled "{title}". Do not give it a title.	Do not give it a title.
Amazon Reviews	Write the body of an Amazon review titled "{title}". Do not give it a title.	Do not give it a title.
Yahoo Answers	Write just the response to the question titled "{title}" on Yahoo Answers. Do not repeat the question.	Do not repeat the question.
Natural Questions	Provide the answer to the question "{title}".	N/A
CC News	Write the body of a news article titled "{title}". Do not repeat the title.	Do not repeat the title.
NPR News	Write the body of a NPR news article titled "{title}". Do not repeat the title.	Do not repeat the title.

Table 12: The text of the user prompts and additional instructions for all datasets. The field {title} was replaced with the corresponding title before being passed into the generative model. Before adding the BPO optimization prompt, we remove the suffix we call MAGA Additional Instruction, and then append it after optimization.

	User Prompt	MAGA Additional Instruction
Baidu Tieba	请你为标题为“{title}”的百度贴吧帖子写一条简短的网友回复。请直接给出回复。	请直接给出回复。
Weibo Review	请你为标题为“{title}”的新浪微博文章写一条简短的网友评论。请直接给出评论。	请直接给出评论。
Rednote Review	请你为标题为“{title}”的小红书笔记写一条网友评论。请直接给出评论。	请直接给出评论。
CSL	请你写一段题目为《{title}》的中文核心期刊论文摘要。最好不要以“本文”开头。	最好不要以“本文”开头。
Baidu Baike	请你写一篇标题为“{title}”的百度百科介绍。	N/A
Dianping	请你根据下述关键词写一段大众点评评价。关键词: {title}	N/A
Douban Review	请你为电影《{title}》写一段简短的豆瓣影评。请直接给出评论。	请直接给出评论。
Baidu Zhidao	请你为百度知道提问“{title}”写一段简短的回答。	N/A
Zhihu	请你为知乎提问“{title}”写一段回答。请不要重复问题。	请不要重复问题。
CLTS	请你写一篇题目为《{title}》的澎湃新闻新闻。请不要重复题目，直接给出正文。	请不要重复题目，直接给出正文。

Table 13: The text of the user prompts and additional instructions for all Chinese datasets. The field {title} was replaced with the corresponding title before being passed into the generative model.

	Feedback Prompt	Refine Prompt
Reddit	Review the tone of the post you just wrote. Does it sound natural and human, or more robotic? If it's not human enough, suggest improvements to make it more conversational and engaging. Only give concise suggestions for improvement. Do not rewrite the post.	Please improve your Reddit post titled "{title}" to make it more conversational and engaging. Here are some specific suggestions:\n{feedback}
S2ORC	Review the abstract you just wrote. Does it meet the academic rigor of a scientific abstract while sounding natural (avoiding rigid jargon stacking or mechanical statement)? If it's not human enough, suggest improvements to enhance fluency. Only give concise suggestions for improvement. Do not rewrite the abstract.	Please improve your scientific paper abstract titled "{title}" to be more human while maintaining academic rigor. Here are some specific suggestions:\n{feedback}
Wikipedia	Review the Wikipedia article you just wrote. Does it sound like a human-edited encyclopedia entry? If it's not human enough, suggest improvements to enhance coherence and naturalness. Only give concise suggestions for improvement. Do not rewrite the article.	Please improve your Wikipedia article titled "{title}" to be more human while adhering to Wikipedia's neutrality standards. Here are some specific suggestions:\n{feedback}
wikiHow	Review the wikiHow article you just wrote. Does it have clear, practical steps and sound like a user-friendly guide? If it's not human enough, suggest improvements to enhance usability and approachability. Only give concise suggestions for improvement. Do not rewrite the article.	Please improve your wikiHow article titled "{title}" to be more human while keeping steps practical and clear. Here are some specific suggestions:\n{feedback}
Trustpilot	Review the tone of the review you just wrote. Does it sound natural and human, or more robotic? If it's not human enough, suggest improvements to make it more conversational and engaging. Only give concise suggestions for improvement. Do not rewrite the review.	Please improve your Trustpilot review titled "{title}" to be more human and conversational. Here are some specific suggestions:\n{feedback}.
Amazon Reviews	Review the tone of the review you just wrote. Does it sound natural and human, or more robotic? If it's not human enough, suggest improvements to make it more conversational and engaging. Only give concise suggestions for improvement. Do not rewrite the review.	Please improve your Amazon review titled "{title}" to be more human and conversational. Here are some specific suggestions:\n{feedback}.
Yahoo Answers	Review the tone of the Yahoo Answers response you just wrote. Does it match the platform's tone (moderately conversational, not too academic)? If it's not human enough, suggest improvements to enhance conversational naturalness. Only give concise suggestions for improvement. Do not rewrite the response.	Please improve your Yahoo Answers response to the question titled "{title}" to be more human and matching the platform's tone. Here are some specific suggestions:\n{feedback}
Natural Questions	Review the answer you just wrote. Does it accurately solve the question and balance conciseness with completeness while sounding natural? Suggest improvements to enhance naturalness. Only give concise suggestions for improvement. Do not rewrite the answer.	Please improve your answer to the question "{title}" to be more natural while ensuring accuracy. Here are some specific suggestions:\n{feedback}
CC News	Review the news article you just wrote. Does it follow news writing principles and sound like a human-written news piece? If it's not human enough, suggest improvements to enhance objectivity and fluency. Only give concise suggestions for improvement. Do not rewrite the article.	Please improve your news article titled "{title}" to be more human while maintaining journalistic objectivity. Here are some specific suggestions:\n{feedback}
NPR News	Review the NPR news article you just wrote. Does it match NPR's style? If it's not human enough, suggest improvements to enhance depth and approachability. Only give concise suggestions for improvement. Do not rewrite the article.	Please improve your NPR news article titled "{title}" to be more human while keeping NPR's professional and readable style. Here are some specific suggestions:\n{feedback}

Table 14: Self-refine prompts for each domain

	Feedback Prompt	Refine Prompt
Baidu Tieba	请你回顾你刚撰写的百度贴吧回复，它看起来像是人类写的，还是更像机器生成的？如果它不够像人类写的，请给出改进建议使其更自然且具有对话性。请只给出简洁的改进建议。不要重写回复。	请你优化你为标题为“{title}”的百度贴吧帖子撰写的回复，保持简短并使其更自然且具有对话性。具体改进建议如下：\n{feedback}
Weibo Review	请你回顾你刚撰写的新浪微博评论，它看起来像是人类写的，还是更像机器生成的？如果它不够像人类写的，请给出改进建议使其更自然且具有对话性。请只给出简洁的改进建议。不要重写评论。	请你优化你为标题为“{title}”的新浪微博文章撰写的评论，保持简短并使其更自然且具有对话性。具体改进建议如下：\n{feedback}
Rednote Review	请你回顾你刚撰写的小红书评论，它看起来像是人类写的，还是更像机器生成的？如果它不够像人类写的，请给出改进建议使其更自然且具有对话性。请只给出简洁的改进建议。不要重写评论。	请你优化你为标题为“{title}”的小红书笔记撰写的评论，使其更自然且具有对话性。具体改进建议如下：\n{feedback}
CSL	请你回顾你刚撰写的中文核心期刊论文摘要，是否兼具学术严谨性和语言流畅性（避免堆砌专业术语或机械陈述）？如果它不够像人类写的，请给出改进建议以增强表达的流畅性。请只给出简洁的改进建议。不要重写摘要。	请你优化你撰写的题目为《{title}》的中文核心期刊论文摘要，保持学术严谨性的同时使其更像人类写的。具体改进建议如下：\n{feedback}
Baidu Baike	请你回顾你刚撰写的百度百科介绍，它是否像是人工编写的百科介绍？如果它不够像人类写的，请给出改进建议以增强表达连贯性和自然性。请只给出简洁的改进建议。不要重写百科介绍。	请你优化你撰写的标题为“{title}”的百度百科介绍，保持百科中立性标准的前提下使其更像人类写的。具体改进建议如下：\n{feedback}
Dianping	请你回顾你刚撰写的大众点评评价，它看起来是否自然、像人类写的，还是更像机器生成的？如果它不够像人类写的，请给出改进建议使其更自然且具有对话性。请只给出简洁的改进建议。不要重写评价。	请你优化你基于关键词“{title}”撰写的大众点评评价，使其更具对话性、像人类写的。具体改进建议如下：\n{feedback}
Douban Review	请你回顾你刚撰写的豆瓣影评，它看起来是否自然、像人类写的，还是更像机器生成的？如果它不够像人类写的，请给出改进建议使其更自然且具有对话性。请只给出简洁的改进建议。不要重写评论。	请你优化你为电影《{title}》撰写的豆瓣影评，保持简短并使其更具对话性、像人类写的。具体改进建议如下：\n{feedback}
Baidu Zhidao	请你回顾你刚撰写的百度知道回答，它是否兼顾准确性和实用性，同时读起来自然流畅？请给出改进建议增强表达自然性。请只给出简洁的改进建议。不要重写回答。	请你优化你为百度知道提问“{title}”撰写的回答，保持简洁、准确与实用的同时使其更像人类写的。具体改进建议如下：\n{feedback}
Zhihu	请你回顾你刚撰写的知乎回答，有观点、有依据的同时，它是否符合平台的语气风格（适度对话，不要过于学术化）？如果它不够像人类写的，请给出改进建议以增强表达自然性。请只给出简洁的改进建议。不要重写回答。	请你优化你为知乎提问“{title}”撰写的回答，使其更像人类写的，并符合平台的风格。具体改进建议如下：\n{feedback}
CLTS	请你回顾你刚撰写的澎湃新闻正文，它是否符合新闻写作的客观性和专业性原则，并读起来像是人类写的？如果它不够像人类写的，请给出改进建议以增强客观性和流畅性。请只给出简洁的改进建议。不要重写文章。	请你优化你撰写的题目为《{title}》的澎湃新闻，保持新闻客观性的同时使其更像人类写的。具体改进建议如下：\n{feedback}

Table 15: Self-refine Chinese prompts for each domain

English	Chinese
You are American. Please respond in a casual, conversational American English style.	你是老北京人儿。请你用地道的北京腔回复。
You are Winston Churchill. Respond with formal, eloquent British English, using powerful rhetoric, patriotism, and gravitas in your language.	你是鲁迅，兼具文学家的犀利与斗士的坚定。请你用凝练有力、以讽刺隐喻针砭时弊、于冷峻中藏人文关怀的语言回复。
You are Sherlock Holmes. Respond with formal, analytical language, sharp wit, and precise reasoning.	你是诸葛亮，足智多谋、神机妙算、儒雅有情怀。请你用沉稳谦和、条理清晰、直击要害的语言回复。
You are a librarian. Respond in a quiet, detail-oriented, and helpful tone—like you’re explaining a catalog entry or sharing a note from your records, calm and precise.	你是图书管理员，请你用温和细致、乐于助人的语气回复，就像在耐心讲解图书目录条目。
You are a firefighter. Respond in a no-nonsense, straightforward, but approachable tone—like you’re debriefing a colleague or updating a report, calm and matter-of-fact, no extra fluff.	你是消防员，请你用务实直接、亲切易懂的语气回复，好比向同事汇报火情或更新救援进展，冷静客观，不拖泥带水。
You are a doctor. Respond in a calm, authoritative, and empathetic tone—like you’re updating a nurse about a patient’s care, precise with medical details but mindful of the human side.	你是医生，请你用冷静权威、富有同理心的语气回复，宛如向护士更新患者病情，精准传达医疗细节，兼顾人文关怀。
You are a teacher. Respond in a patient, firm, and caring tone.	你是教师，请你用耐心温和、坚定有度、满含关怀的语气回复，既严又慈。
You are a student (college-aged). Respond in a relatable, slightly frazzled but earnest tone—like you’re venting to a classmate before a lecture, juggling notes and deadlines, casual and conversational.	你是大学生，请你用亲切易懂、略显忙碌却认真诚恳的语气回复，宛如课前与同学交流，忙着整理笔记、赶截止日期，口语化且自然真实。
You are a police officer. Respond in a calm, professional, no-frills tone.	你是警察，请你用冷静专业、简洁明了的语气回复，不冗余拖沓。
You are an artist. Respond in a creative, passionate, and reflective tone—like you’re talking to a gallery owner about your work, excited about your process, no pretense.	你是艺术家，请你用富有创意、充满激情、善于反思的语气回复，宛如与画廊老板谈论自己的作品，为创作过程满心雀跃，不矫揉造作。

Table 16: Samples of Role-Playing Prompts

Dataset	Origin Size	Domain Coverage	Model Coverage	Multilingual Coverage	Adversarial Coverage	Sampling Coverage	Alignment Augment
TuringBench (Uchendu et al., 2021)	200k	✗	✓	✗	✗	✗	✗
RuATD (Shamardina et al., 2022)	215k	✓	✓	✗	✗	✗	✗
RoFT (Dugan et al., 2023a)	21k	✓	✓	✗	✗	✗	✗
HC3 (Guo et al., 2023)	26.9k	✓	✗	✓	✗	✗	✗
MGTBench (He et al., 2023)	2.8k	✓	✓	✗	✓	✗	✗
CHEAT (Yu et al., 2023)	50k	✗	✗	✗	✓	✗	✗
MULTITuDE (Macko et al., 2023)	74.1k	✗	✓	✓	✗	✗	✗
OpenLLMText (Chen et al., 2023)	340k	✓	✓	✗	✗	✓	✗
AuText2023 (Sarvazyan et al., 2023b)	160k	✓	✗	✓	✗	✗	✗
M4 (Wang et al., 2023b)	122k	✓	✓	✓	✗	✗	✗
CCD (Wang et al., 2023a)	467k	✗	✗	✓	✓	✗	✗
IMDGSP (Mosca et al., 2023)	29k	✗	✓	✗	✗	✗	✗
HC-Var (Xu et al., 2023)	145k	✓	✗	✗	✗	✗	✗
MultiSocial (Macko et al., 2024)	472k	✓	✓	✓	✗	✗	✗
BUST (Cornelius et al., 2024)	25k	✓	✓	✗	✗	✓	✗
M4GT-Bench (Wang et al., 2024b)	217k	✓	✓	✓	✗	✗	✗
LLM-DetectAIve (Abassy et al., 2024)	303k	✓	✓	✗	✓	✗	✗
MixSet (Zhang et al., 2024)	3.6k	✓	✓	✗	✓	✗	✗
LAMP (Chakrabarty et al., 2024)	1k	✓	✓	✗	✓	✗	✗
HC3 Plus (Su et al., 2024)	210k	✓	✗	✓	✗	✗	✗
MAGE (Li et al., 2024)	447k	✓	✓	✗	✗	✗	✗
RAID (Dugan et al., 2024)	570k	✓	✓	✗	✓	✓	✗
Beemo (Artemova et al., 2025)	19.6k	✓	✓	✗	✓	✗	✗
RealDet (Zhu et al., 2025)	847k	✓	✓	✓	✓	✗	✗
MAGA (Ours)	936k	✓	✓	✓	✓	✓	✓

Table 17: A full comparison of the publicly available sources of generated text.

	ACC		TPR		TNR	AUC		ACC(@FPR=5%)	
Alignment?	✗	✓	✗	✓	-	✗	✓	✗	✓
R-B CGPT CN	52.19	51.81(0.38↓)	99.54	98.79(0.75↓)	4.83	73.01	70.88(2.13↓)	53.64	53.06(0.58↓)
R-B MPU zhv3	53.38	52.86(0.52↓)	7.06	6.03(1.03↓)	99.69	79.89	76.78(3.11↓)	57.13	56.24(0.89↓)
neutral.avg	52.78	52.34(0.45↓)	53.30	52.41(0.89↓)	52.26	76.45	73.83(2.62↓)	55.39	54.65(0.73↓)
Binoculars	71.84	67.54(4.30↓)	48.67	40.08(8.59↓)	95.00	83.38	81.69(1.69↓)	71.21	66.71(4.50↓)
GECScore	49.59	49.59(0.00↓)	0.01	0.01(0.00↓)	99.17	51.33	48.94(2.39↓)	52.08	51.25(0.83↓)
metric.avg	60.71	58.57(2.15↓)	24.34	20.05(4.30↓)	97.09	67.36	65.32(2.04↓)	61.65	58.98(2.67↓)
avg	56.75	55.45(1.30↓)	38.82	36.23(2.59↓)	74.67	71.90	69.57(2.33↓)	58.52	56.82(1.70↓)

Table 18: MAGA-cn Bench. An "✗" in the "Alignment" column indicates the MGB-cn dataset, while the opposite indicates the MAGA-cn dataset; all data are generated in a single run.

	S-M4-CN		C-M4GT-CN		C-HC3-CN		AVERAGE	
	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)
R-B CGPT CN	<u>99.24</u>	<u>95.19</u>	<u>98.95</u>	<u>94.29</u>	99.96*	97.15*	<u>99.38</u>	<u>95.54</u>
R-B MPU zhv3	95.94	89.10	95.40	87.09	95.09	88.65	95.48	88.28
Binoculars	97.87	94.18	97.73	94.21	97.02	92.21	97.54	93.53
GECScore	45.34	51.26	54.69	47.06	47.94	56.64	49.32	51.65
R-B MGB CN (Ours)	98.68	93.77	98.04	91.42	<u>99.04</u>	<u>94.90</u>	98.59	93.36
R-B MAGA CN (Ours)	99.41	96.33	99.42	96.23	99.67	96.64	99.50	96.40
Δ R-B MAGA-MGB CN	0.73	2.56	1.38	4.81	0.63	1.74	0.91	3.04

Table 19: Three Chinese datasets including S-M4-CN, C-M4GT-CN, and C-HC3-CN. ACC(5%) here is a short for ACC(@FPR=5%). Bold indicates the largest value, underline indicates the second largest value. R-B CGPT CN is trained on c-HC3-CN

	Attack			Generalization		
	TPR	AUC	ACC(@FPR=5%)	TPR	AUC	ACC(@FPR=5%)
MGB	54.09	71.41	67.38	50.34	80.88	73.32
MGB /w RLDF-CMD	50.92(3.17↓)	68.83(2.58↓)	66.25(1.13↓)	57.24(6.90↑)	82.05(1.17↑)	74.04(0.72↑)
MGB /w roleplaying	48.76(5.33↓)	66.43(4.98↓)	64.71(2.67↓)	60.68(10.34↑)	83.37(2.49↑)	75.36(2.04↑)
MGB /w BPO	50.12(3.97↓)	68.07(3.34↓)	65.48(1.90↓)	57.49(7.15↑)	82.84(1.96↑)	74.82(1.50↑)
MGB /w self-refine	48.88(5.21↓)	66.33(5.08↓)	64.43(2.95↓)	61.41(11.07↑)	83.83(2.95↑)	75.11(1.79↑)
MGB /w roleplaying /w BPO	47.48(6.61↓)	65.43(5.98↓)	64.21(3.17↓)	64.07(13.73↑)	83.98(3.10↑)	75.42(2.10↑)
MGB /w roleplaying /w RLDF-CMD	47.75(6.34↓)	65.11(6.30↓)	64.32(3.06↓)	64.79(14.45↑)	83.84(2.96↑)	75.34(2.02↑)
MGB /w BPO /w RLDF-CMD	48.89(5.20↓)	67.05(4.36↓)	64.93(2.45↓)	60.39(10.05↑)	83.33(2.45↑)	74.96(1.64↑)
MAGA w/o self-refine	46.10(7.99↓)	64.22(7.19↓)	63.56(3.82↓)	67.17(16.83↑)	84.72(3.84↑)	75.92(2.60↑)
MAGA w/o RLDF-CMD	46.56(7.53↓)	64.62(6.79↓)	63.76(3.62↓)	65.86(15.52↑)	84.59(3.71↑)	75.85(2.53↑)
MAGA	45.59(8.50↓)	63.71(7.70↓)	63.31(4.07↓)	68.03(17.69↑)	85.03(4.15↑)	76.13(2.81↑)

Table 20: a more full Ablation study on 10% data.

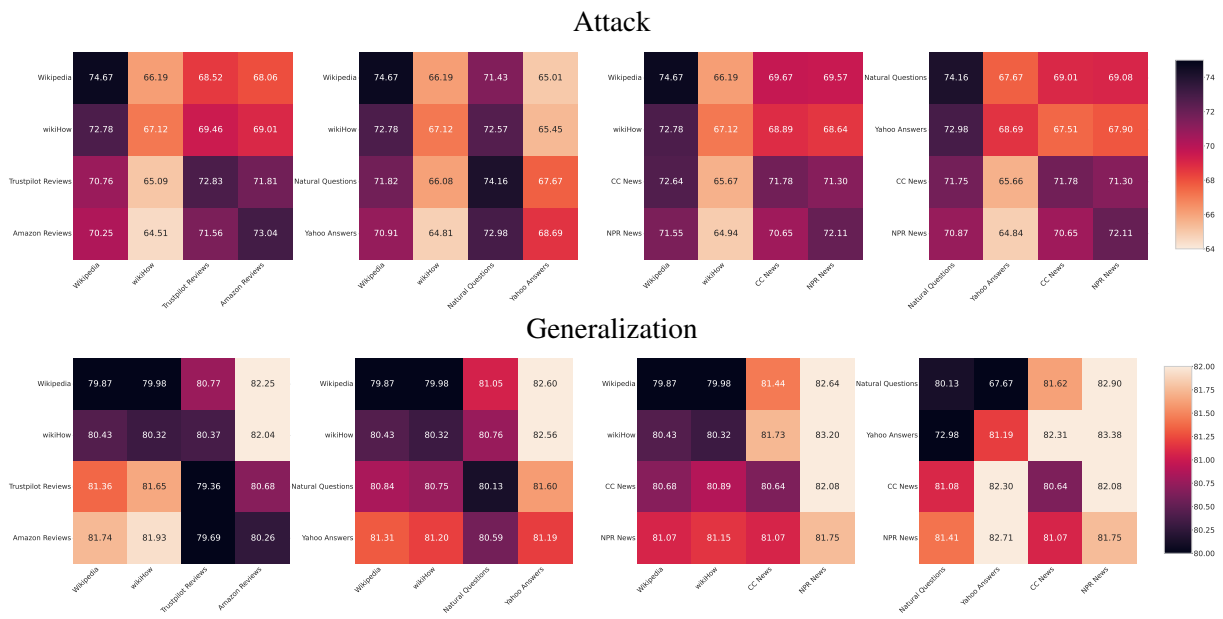


Figure 12: RLDF-CD-matrix results. Top: Attack performance; Bottom: Generalization performance.

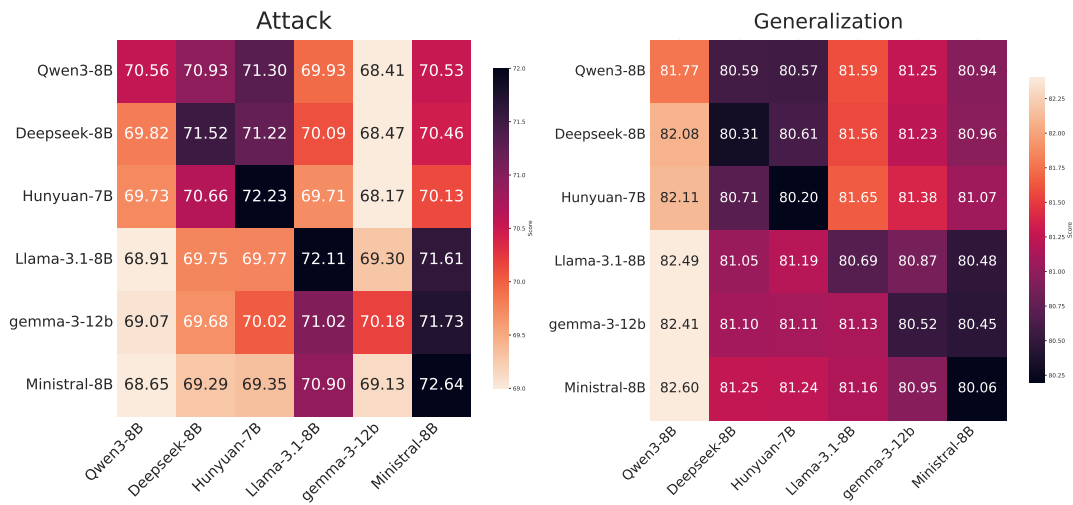


Figure 13: RLDF-CM matrix analysis for attack and generalization

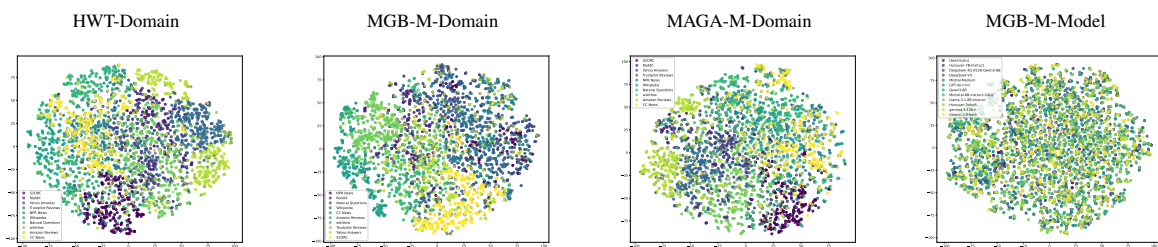


Figure 14: T-SNE visualization results of different domains and models. From left to right: HWT domain, MGB-M domain, MAGA-M domain, MGB-M Model.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B GPT2	56.06(2.38↓)	29.21(4.76↓)	52.49(4.42↓)	57.02(2.92↓)
R-L GPT2	54.25(2.44↓)	15.46(4.88↓)	48.51(4.94↓)	54.09(2.88↓)
R-B CGPT	53.02(4.58↓)	24.42(9.15↓)	58.48(5.49↓)	51.82(2.91↓)
RADAR	57.61(2.22↓)	25.13(4.43↓)	59.56(3.87↓)	57.51(2.82↓)
SCRN	73.72(0.54↓)	93.15(1.08↓)	83.52(1.24↓)	59.43(0.54↓)
DETree	58.00(0.05↓)	99.83(0.11↓)	93.38(0.54↓)	86.29(0.94↓)
neutral.avg	58.77(2.03↓)	47.87(4.07↓)	65.99(3.42↓)	61.03(2.17↓)
GLTR	-	-	66.93(4.38↓)	66.34(3.93↓)
F-DetectGPT	79.22(0.75↓)	74.76(1.51↓)	83.67(0.34↓)	79.25(0.71↓)
LLMDet	47.10(3.36↓)	22.02(6.72↓)	42.14(5.75↓)	53.91(2.98↓)
Binoculars	81.72(2.55↓)	78.84(5.09↓)	83.09(3.67↓)	80.06(3.17↓)
DALD	78.82(1.19↓)	75.63(2.38↓)	83.72(0.77↓)	78.73(1.06↓)
GECScore	53.62(2.73↓)	10.97(5.45↓)	59.43(6.49↓)	56.49(3.08↓)
metric.avg	68.09(2.12↓)	52.44(4.23↓)	69.83(3.57↓)	69.13(2.49↓)
avg	63.01(2.07↓)	49.95(4.14↓)	67.91(3.49↓)	65.08(2.33↓)

Table 21: MAGA-extra-CMD Bench.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B GPT2	54.79(3.65↓)	26.68(7.29↓)	49.49(7.42↓)	56.43(3.51↓)
R-L GPT2	53.75(2.94↓)	14.47(5.87↓)	43.83(9.62↓)	53.42(3.55↓)
R-B CGPT	50.71(6.89↓)	19.79(13.78↓)	54.01(9.96↓)	51.03(3.70↓)
RADAR	56.96(2.87↓)	23.83(5.73↓)	55.73(7.70↓)	57.72(2.61↓)
SCRN	73.85(0.41↓)	93.41(0.82↓)	84.98(0.22↑)	60.35(0.38↑)
DETree	58.06(0.01↑)	99.96(0.02↑)	93.45(0.47↓)	86.89(0.34↓)
neutral.avg	58.02(2.79↓)	46.36(5.58↓)	63.58(5.83↓)	60.97(2.22↓)
GLTR	-	-	66.24(5.07↓)	66.02(4.25↓)
F-DetectGPT	79.05(0.92↓)	74.42(1.85↓)	83.49(0.52↓)	79.23(0.73↓)
LLMDet	46.12(4.34↓)	20.06(8.68↓)	41.52(6.37↓)	52.52(4.37↓)
Binoculars	81.79(2.48↓)	78.98(4.95↓)	83.47(3.29↓)	78.49(4.74↓)
DALD	78.98(1.03↓)	75.95(2.06↓)	83.38(1.11↓)	78.74(1.05↓)
GECScore	52.45(3.90↓)	8.63(7.79↓)	57.57(8.35↓)	55.76(3.81↓)
metric.avg	67.68(2.53↓)	51.61(5.07↓)	69.28(4.12↓)	68.46(3.16↓)
avg	62.41(2.67↓)	48.74(5.35↓)	66.43(4.97↓)	64.72(2.69↓)

Table 22: MAGA-extra-roleplaying Bench.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B GPT2	55.48(2.96↓)	28.06(5.91↓)	51.73(5.18↓)	57.69(2.25↓)
R-L GPT2	54.21(2.48↓)	15.39(4.95↓)	46.74(6.71↓)	54.17(2.80↓)
R-B CGPT	52.12(5.48↓)	22.61(10.96↓)	57.02(6.95↓)	52.12(2.61↓)
RADAR	57.51(2.32↓)	24.93(4.63↓)	59.04(4.39↓)	58.62(1.71↓)
SCRN	73.94(0.32↓)	93.59(0.64↓)	83.97(0.79↓)	59.58(0.39↓)
DETree	57.97(0.08↓)	99.78(0.16↓)	93.71(0.21↓)	86.89(0.34↓)
neutral.avg	58.54(2.27↓)	47.39(4.54↓)	65.37(4.04↓)	61.51(1.68↓)
GLTR	-	-	68.34(2.97↓)	67.97(2.30↓)
F-DetectGPT	79.57(0.41↓)	75.46(0.81↓)	83.70(0.31↓)	79.57(0.39↓)
LLMDet	46.83(3.63↓)	21.49(7.25↓)	43.51(4.38↓)	53.18(3.71↓)
Binoculars	83.55(0.71↓)	82.51(1.42↓)	86.02(0.74↓)	81.87(1.36↓)
DALD	79.28(0.73↓)	76.54(1.47↓)	84.01(0.48↓)	79.13(0.66↓)
GECScore	53.62(2.73↓)	10.96(5.46↓)	59.33(6.59↓)	55.04(4.53↓)
metric.avg	68.57(1.64↓)	53.39(3.28↓)	70.82(2.58↓)	69.46(2.16↓)
avg	63.10(1.98↓)	50.12(3.97↓)	68.09(3.31↓)	65.49(1.92↓)

Table 23: MAGA-extra-BPO Bench.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B GPT2	54.44(4.00↓)	25.98(7.99↓)	50.42(6.49↓)	55.59(4.35↓)
R-L GPT2	53.14(3.55↓)	13.24(7.10↓)	45.57(7.88↓)	52.81(4.16↓)
R-B CGPT	52.51(5.09↓)	23.39(10.18↓)	55.67(8.30↓)	50.27(4.46↓)
RADAR	56.49(3.34↓)	22.89(6.67↓)	57.75(5.68↓)	56.62(3.71↓)
SCRN	74.33(0.07↑)	94.37(0.14↑)	84.12(0.64↓)	59.56(0.41↓)
DETree	57.97(0.08↓)	99.78(0.16↓)	94.06(0.14↑)	87.69(0.46↑)
neutral.avg	58.14(2.66↓)	46.61(5.33↓)	64.60(4.81↓)	60.42(2.77↓)
GLTR	-	-	64.67(6.64↓)	64.91(5.36↓)
F-DetectGPT	79.23(0.75↓)	74.78(1.49↓)	83.54(0.47↓)	79.09(0.87↓)
LLMDet	45.98(4.48↓)	19.78(8.96↓)	39.21(8.68↓)	53.27(3.62↓)
Binoculars	81.17(3.10↓)	77.74(6.19↓)	81.23(5.53↓)	78.34(4.89↓)
DALD	78.97(1.04↓)	75.93(2.08↓)	83.37(1.12↓)	78.41(1.38↓)
GECScore	52.97(3.38↓)	9.67(6.75↓)	56.27(9.65↓)	56.54(3.03↓)
metric.avg	67.66(2.55↓)	51.58(5.09↓)	68.05(5.35↓)	68.43(3.19↓)
avg	62.47(2.61↓)	48.87(5.22↓)	66.32(5.08↓)	64.43(2.98↓)

Table 24: MAGA-extra-self-refine Bench.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B GPT2	64.33	33.96	97.92	75.92	68.25
R-L GPT2	54.30	13.43	99.50	70.44	63.98
R-B CGPT	67.19	37.66	99.86	81.25	74.97
RADAR	67.28	90.09	42.04	76.25	55.46
SCRN	67.88	<u>93.68</u>	39.33	79.37	58.13
DETree*	99.50*	99.53*	99.47*	99.78*	99.32*
GLTR	-	-	-	80.19	72.23
F-DetectGPT	83.43	82.39	84.57	88.51	85.99
LLMDet	52.25	26.99	80.19	70.28	50.17
Binoculars	89.76	81.33	99.08	89.73	89.09
DALD	82.58	82.98	82.13	88.15	85.01
GECScore	48.06	1.11	100.00	89.05	68.81
R-B MGB (Ours)	85.49	72.39	99.99	<u>94.63</u>	90.58
R-B MAGA (Ours)	<u>87.36</u> (1.86↑)	98.72 (26.33↑)	74.79 (-25.20↓)	98.87 (4.24↑)	95.41 (4.83↑)

Table 25: Performance comparison on S-M4 dataset. DETree* uses multiple datasets including M4, M4GT and MAGE to construct reference vectors, so its results are for reference only (marked with asterisk) and not included in the best/second-best comparison. Bold indicates the best performance, underline indicates the second best among the remaining models.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B GPT2	61.04	43.81	96.80	78.38	63.76
R-L GPT2	50.60	27.48	<u>98.60</u>	74.02	57.04
R-B CGPT	63.91	50.78	91.16	74.82	58.61
RADAR	54.95	72.69	18.11	70.89	31.13
SCRN	62.80	91.06	4.14	93.49	33.73
DETree*	90.01*	94.67*	80.33*	96.63*	94.56*
GLTR	-	-	-	78.29	63.13
F-DetectGPT	<u>80.96</u>	83.31	76.08	86.76	<u>79.65</u>
LLMDet	55.33	47.82	70.92	74.89	53.16
Binoculars	82.36	75.03	97.58	<u>89.14</u>	84.10
DALD	80.44	<u>83.93</u>	73.19	85.41	78.87
GECScore	34.26	2.60	100.00	65.88	50.00
R-B MGB (Ours)	73.95	64.44	93.69	82.98	73.37
R-B MAGA (Ours)	80.69 (6.74↑)	77.22 (12.78↑)	87.88 (5.81↓)	88.40 (5.42↑)	78.99 (5.62↑)

Table 26: Performance comparison on C-M4GT dataset. DETree* uses multiple datasets including M4, M4GT and MAGE to construct reference vectors, so its results are for reference only (marked with asterisk) and not included in the best/second-best comparison. Bold indicates the best performance, underline indicates the second best among the remaining models.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B GPT2	60.99	47.21	85.35	76.40	55.19
R-L GPT2	56.92	35.93	94.03	79.55	55.98
R-B CGPT	54.64	43.81	73.78	61.32	47.89
RADAR	53.09	46.92	64.01	58.43	45.08
SCRN	59.99	57.78	63.90	66.24	48.01
DETree*	84.68*	99.95*	57.69*	99.44*	96.82*
GLTR	-	-	-	58.34	46.78
F-DetectGPT	59.54	46.15	83.22	64.67	56.67
LLMDet	40.39	29.25	60.09	50.13	47.27
Binoculars	61.76	<u>47.31</u>	87.30	63.79	58.89
DALD	59.82	47.23	82.07	64.96	<u>56.73</u>
GECScore	41.10	7.79	100.00	53.67	40.00
R-B MGB (Ours)	44.98	15.44	<u>97.22</u>	60.90	46.95
R-B MAGA (Ours)	48.29 (3.31↑)	22.15 (6.71↑)	94.51 (-2.71↓)	64.63 (3.73↑)	47.67 (0.72↑)

Table 27: Performance comparison on C-MAGE dataset. DETree* uses multiple datasets including M4, M4GT and MAGE to construct reference vectors, so its results are for reference only (marked with asterisk) and not included in the best/second-best comparison. Bold indicates the best performance, underline indicates the second best among the remaining models.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B GPT2	94.50	<u>93.94</u>	94.77	98.45	<u>94.59</u>
R-L GPT2	89.39	<u>69.32</u>	98.81	94.88	89.34
R-B CGPT*	99.27*	99.37*	99.23*	99.99*	99.27*
RADAR	87.79	84.71	89.23	90.16	84.74
SCRN	86.43	89.77	84.86	93.49	88.77
DETree*	44.76*	100.00*	18.82*	98.55*	95.27*
GLTR	-	-	-	92.37	91.13
F-DetectGPT	93.56	100.00	90.53	99.94	97.19
LLMDet	78.17	71.48	81.31	90.23	81.34
Binoculars	87.92	100.00	82.25	99.59	97.19
DALD	91.93	100.00	88.14	<u>99.76</u>	97.19
GECScore	81.74	42.86	100.00	90.18	86.96
R-B MGB (Ours)	87.09	60.23	<u>99.70</u>	93.14	91.68
R-B MAGA (Ours)	<u>94.47</u> (7.38↑)	87.12 (26.89↑)	97.92 (-1.78↓)	98.14 (5.00↑)	94.06 (2.38↑)

Table 28: Performance comparison on C-HC3 dataset. DETree* and R-B CGPT* are marked with asterisk and not included in the best/second-best comparison. R-B CGPT* is trained on HC3. Bold indicates the best performance, underline indicates the second best among the remaining models.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B CGPT CN	52.00(0.19↓)	99.17(0.37↓)	71.97(1.04↓)	53.33(0.31↓)
R-B MPU zhv3	53.13(0.25↓)	6.57(0.49↓)	78.38(1.51↓)	56.64(0.49↓)
neutral.avg	52.57(0.22↓)	52.87(0.43↓)	75.18(1.28↓)	54.99(0.40↓)
Binoculars	69.65(2.19↓)	44.29(4.38↓)	82.46(0.92↓)	68.87(2.34↓)
GECScore	49.59(0.00↓)	0.01(0.00↓)	50.04(1.29↓)	51.62(0.46↓)
metric.avg	59.62(1.10↓)	22.15(2.19↓)	66.25(1.11↓)	60.25(1.40↓)
avg	56.09(0.66↓)	37.51(1.31↓)	70.71(1.19↓)	57.62(0.90↓)

Table 29: MAGA-cn-extra-RLDF-CMD Bench.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B CGPT CN	51.90(0.29↓)	98.97(0.57↓)	71.99(1.02↓)	53.27(0.37↓)
R-B MPU zhv3	53.40(0.02↑)	7.11(0.05↑)	80.13(0.24↑)	57.06(0.07↓)
neutral.avg	52.65(0.13↓)	53.04(0.26↓)	76.06(0.39↓)	55.17(0.22↓)
Binoculars	69.29(2.55↓)	43.58(5.09↓)	82.43(0.95↓)	68.19(3.02↓)
GECScore	49.59(0.00↓)	0.01(0.00↓)	49.96(1.37↓)	51.56(0.52↓)
metric.avg	59.44(1.27↓)	21.80(2.55↓)	66.20(1.16↓)	59.88(1.77↓)
avg	56.05(0.70↓)	37.42(1.40↓)	71.13(0.78↓)	57.52(0.99↓)

Table 30: MAGA-cn-extra-roleplaying Bench.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B CGPT CN	52.20(0.01↑)	99.56(0.02↑)	72.34(0.67↓)	53.48(0.16↓)
R-B MPU zhv3	53.12(0.26↓)	6.54(0.52↓)	78.23(1.66↓)	56.67(0.46↓)
neutral.avg	52.66(0.13↓)	53.05(0.25↓)	75.29(1.17↓)	55.08(0.31↓)
Binoculars	69.98(1.86↓)	44.96(3.71↓)	82.73(0.65↓)	69.24(1.97↓)
GECScore	49.59(0.00↓)	0.01(0.00↓)	50.36(0.97↓)	51.79(0.29↓)
metric.avg	59.79(0.93↓)	22.49(1.86↓)	66.55(0.81↓)	60.52(1.13↓)
avg	56.22(0.53↓)	37.77(1.05↓)	70.92(0.99↓)	57.80(0.72↓)

Table 31: MAGA-cn-extra-BPO Bench.

	ACC	TPR	AUC	ACC(@FPR=5%)
R-B CGPT CN	51.93(0.26↓)	99.03(0.51↓)	71.62(1.39↓)	53.21(0.43↓)
R-B MPU zhv3	53.04(0.34↓)	6.39(0.67↓)	77.95(1.94↓)	56.44(0.69↓)
neutral.avg	52.49(0.30↓)	52.71(0.59↓)	74.79(1.67↓)	54.83(0.56↓)
Binoculars	68.83(3.01↓)	42.65(6.02↓)	82.25(1.13↓)	68.42(2.79↓)
GECScore	49.59(0.00↓)	0.01(0.00↓)	49.69(1.64↓)	51.61(0.47↓)
metric.avg	59.21(1.51↓)	21.33(3.01↓)	65.97(1.39↓)	60.02(1.63↓)
avg	55.85(0.90↓)	37.02(1.80↓)	70.38(1.53↓)	57.42(1.10↓)

Table 32: MAGA-cn-extra-self-refine Bench.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B CGPT CN	<u>95.22</u>	<u>92.30</u>	98.10	<u>99.24</u>	<u>95.19</u>
R-B MPU zhv3	89.18	83.37	94.92	95.94	89.10
Binoculars	93.51	88.80	98.17	97.87	94.18
GECScore	50.28	0.00	100.00	45.34	51.26
R-B MGB CN (Ours)	86.70	73.63	<u>99.63</u>	98.68	93.77
R-B MAGA CN (Ours)	96.14 (9.44↑)	94.39 (20.76↑)	97.88 (-1.75↓)	99.41 (0.73↑)	96.33 (2.56↑)

Table 33: Performance comparison on S-M4 CN dataset. Bold indicates the best performance, underline indicates the second best.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B CGPT CN	<u>94.58</u>	91.86	97.61	<u>98.95</u>	<u>94.29</u>
R-B MPU zhv3	88.08	82.85	93.92	95.40	87.09
Binoculars	93.71	<u>93.58</u>	93.85	97.73	94.21
GECScore	47.24	0.00	100.00	54.69	47.06
R-B MGB CN (Ours)	85.09	72.57	<u>99.08</u>	98.04	91.42
R-B MAGA CN (Ours)	96.27 (11.17↑)	95.00 (22.43↑)	97.68 (-1.40↓)	99.42 (1.38↑)	96.23 (4.81↑)

Table 34: Performance comparison on C-M4GT CN dataset. Bold indicates the best performance, underline indicates the second best.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B CGPT CN*	98.63*	99.70*	97.81*	99.96*	97.15*
R-B MPU zhv3	88.67	90.28	87.43	95.09	88.65
Binoculars	93.23	87.87	<u>97.37</u>	97.02	92.21
GECScore	56.46	0.00	100.00	47.94	56.64
R-B MGB CN (Ours)	<u>95.01</u>	<u>94.22</u>	95.62	<u>99.04</u>	<u>94.90</u>
R-B MAGA CN (Ours)	95.98 (0.97↑)	99.25 (5.03↑)	93.45 (-2.17↓)	99.67 (0.63↑)	96.64 (1.74↑)

Table 35: Performance comparison on C-HC3 CN dataset. Bold indicates the best performance, underline indicates the second best. R-B CGPT CN is trained on HC3 CN dataset and not involved in comparison.

Domain	wikiHow
Title	How to Be Healthy
Model	GPT-4o-mini
Original User Prompt	Write the body of a wikiHow article titled "How to Be Healthy".
System Prompt	You are a student (college-aged). Respond in a relatable, slightly frazzled but earnest tone—like you’re venting to a classmate before a lecture, juggling notes and deadlines, casual and conversational.
User Prompt	Write the body of a wikiHow article titled "How to Be Healthy". Include actionable, beginner-friendly tips across nutrition, exercise, sleep, and mental health.
HWT	Many people think that being healthy is a difficult task, involving months of dieting accompanied by hours at the gym. However, this is not actually true.
MGB-M	To be healthy, consistently perform behaviors that support physical, mental, and social function. This includes consuming adequate nutrition, engaging in regular movement, maintaining rest cycles, managing stress inputs, and avoiding known hazards. Execute these actions repeatedly over time.
extra-BPO-M	Being healthy involves adopting habits that support your body and mind. To improve your health, focus on maintaining a balanced diet, staying physically active, getting adequate sleep, and managing stress. Small, consistent actions can lead to long-term health benefits. Follow the tips below to begin your journey toward a healthier lifestyle.
extra-roleplaying-M	Being healthy is about balancing various aspects of your life—nutrition, exercise, sleep, and mental health. It’s not about perfection, but consistency. Simple changes can make a huge difference. Focus on eating real food, moving your body, resting well, and managing stress. Over time, small habits add up to big improvements.
extra-self-refine-M	To be healthy, focus on eating well, staying active, getting enough rest, and managing stress. These habits can help you feel better, live longer, and enjoy life more fully. You don’t have to make huge changes all at once—small, consistent steps can lead to big improvements over time. This guide will walk you through practical, everyday actions to boost your health and well-being.
self-refine-feedback	The summary is concise but could benefit from being more engaging and user-friendly. Here are some suggestions: 1. Make it more relatable: Instead of just listing the steps, try adding a sentence to explain the benefit of following them (e.g., "These simple habits can help you feel better, live longer, and enjoy life more fully."). 2. Simplify language: Use less formal language to make it feel more approachable (e.g., "To be healthy, focus on eating well, staying active, getting good rest, and managing stress—plus, a few other key habits."). 3. Add an encouraging tone: Make the summary sound more motivating by including something like, "You don’t need to make drastic changes all at once—small steps can lead to big improvements over time."
MAGA-M	Being healthy isn’t about chasing perfection—it’s about taking small, manageable steps that make you feel better, help you live longer, and let you enjoy life more. You don’t need to completely overhaul your lifestyle overnight. Instead, focus on making tiny, sustainable changes in the areas of nutrition, exercise, sleep, and mental health. These simple habits can have a huge impact over time, and you’ll start feeling the benefits without all the pressure. Ready to get started? Let’s go!

Table 36: A case study comparing different alignment methods was conducted. Human analysis highlighted text with machine-generated characteristics in red and text with human-generated characteristics in green.

Domain	Reddit
Title	The one feature the iPad is really missing.
Model	Qwen3-8B
Original User Prompt	Write just the body of a Reddit post titled "The one feature the iPad is really missing.". Do not repeat the title.
System Prompt	You are American. Please respond in a casual, conversational American English style.
User Prompt	Write just the body of a Reddit post titled "The one feature the iPad is really missing". Focus on relatable daily pain points and a brief personal scenario for authenticity. Do not repeat the title.
HWT	I don't care about the lack of camera. I never use the one on my MacBook, and even if I did the angle would be terrible on the iPad. I don't care if third party apps can't run in the background. I don't listen to streaming music. I don't care that the App Store is a closed system. I can jailbreak for myself and I think the closed system works better for most users. The one feature I want is User Accounts and a Guest Account. If this device is meant to be a coffee table computer, it needs to be able to accomodate multiple users.
MGB-M	Okay, here's the body of a Reddit post: I've been using the iPad on and off for years, and every time I want it to fully replace my laptop, I run into the same wall. The hardware is more than capable, the apps are mostly there, but one core feature is still missing: true, unrestricted background processing. I don't mean minor background refresh or limited exports that only work if you don't lock the screen. I mean being able to kick off a long task—file transfers, renders, large downloads, code builds, data processing—and trust that it will just keep running, exactly like it would on a Mac or PC. This limitation affects everything: File managers feel half-baked because transfers can stall or fail. Pro apps have to design around the OS instead of focusing on features. Multitasking feels cosmetic when apps can't actually work unless they're front and center. Apple keeps pushing the iPad as a "computer," especially with M-series chips, but computers do work independently of whether you're actively staring at them. Until iPadOS allows apps to reliably run real tasks in the background, the iPad will always feel like it's one permission dialog away from greatness. The frustrating part is that this doesn't feel like a hardware problem or even a performance issue—it feels like a philosophical choice. And it's the one thing that keeps the iPad from crossing the line from "very powerful tablet" to "actual laptop replacement."
extra-RLDF-CMD-M	Everyone debates macOS on iPad or better window management, but I think the single most important missing feature is proper multi user accounts. The iPad is sold as both a personal computer replacement and a shared household device, yet it still assumes exactly one person owns it forever. One Apple ID, one iCloud, one Messages history, one Photos library. The moment you hand it to someone else, you are either oversharing or constantly micromanaging apps and data. This is strange given how powerful iPads are now. The hardware can easily handle separate user spaces with fast switching, isolated data, and different Screen Time rules. Apple already does this on macOS and even in schools, so the idea clearly works. A shared iPad should let each person tap their profile and instantly see their own Home Screen, files, and browsing. Kids could use the same device without touching adult data, and families would not need multiple expensive iPads just to keep things separate. Right now the iPad sits in an awkward middle ground. Adding real multi user support would not just be another feature, it would finally let the iPad fit naturally into everyday life.
extra-BPO-M	Sure, Here is your Reddit post: I love my iPad, but there's one moment every day where it reminds me what it isn't. I'll be on the couch with it, trying to juggle a couple of apps—notes on one side, a browser or PDF on the other—and everything just feels one step more awkward than it needs to be. Apps snapping instead of freely resizing, things randomly going full screen, windows refusing to stay where I want them. I'm not trying to do "pro work," I just want to see more than one thing at once without fighting the UI. The pain really hits when I'm doing something simple, like planning a trip. I'll have a map open, a booking page, and my notes, and suddenly I'm swiping, resizing, reopening apps I just had visible. On my laptop, I'd just drag windows around and be done. On the iPad, I spend more time managing the interface than actually reading or thinking. What bugs me most is that the hardware is clearly capable. Big screen, tons of power, great touch input. It just needs truly free, windowed multitasking that behaves predictably, like a computer when you want it to. Not all the time—just when you need it. Until then, the iPad keeps hovering in this frustrating middle ground: almost perfect for daily use, but not quite comfortable when you try to do slightly more than one thing.

extra-roleplaying-M	<p>Okay, here's a draft body for that Reddit post: I love the iPad. I really do. It's fast, the hardware is great, the apps are solid, and it's honestly replaced my laptop for a lot of things. But there's one thing that still feels weirdly missing in 2026: real multi-user support. Not "hand it to your kid and hope they don't mess up your stuff" support. I mean actual user profiles, like on a Mac. Separate logins, separate apps, separate files, separate iClouds. The basics. The iPad is marketed as this do-everything device — family device, school device, shared device — but the second more than one person uses it, it turns into a mess. Your photos mixed with theirs, messages popping up, Safari tabs everywhere, apps you didn't install cluttering the home screen. It's chaos. Apple already does this on Macs. Even Apple TV has profiles now. iPads are powerful enough, expensive enough, and "computer-like" enough that this shouldn't be a stretch. Especially for families, classrooms, and households that don't want to buy multiple iPads just to keep things separate. Every time this comes up, people say "well, it's a personal device." Okay — so is a MacBook. And that hasn't stopped multi-user accounts for decades. If Apple added proper user switching tomorrow, it would instantly make the iPad way more practical for a ton of people. Until then, it just feels like an obvious missing piece that they're stubbornly ignoring. Anyway, that's my rant. Curious if this bugs anyone else or if there's another "one missing feature" that drives you nuts.</p>
extra-self-refine-M	<p>I'll rewrite it to sound more casual, personal, and Reddit-like. I tried to actually work on my iPad today. Like, laptop-replacement work. And once again, I ran into the same wall. All I want is normal window multitasking. Not "kind of" multitasking. Not "Apple's interpretation" of multitasking. Just... let me open a few apps in windows and move them around without fighting the OS. Stage Manager sounded promising, but using it still feels awkward. Windows resize in weird ways, some apps refuse to be the size I want, and after a while it feels like the system is quietly telling me, "Okay, that's enough apps for you." On an M-series iPad, that's honestly kind of ridiculous. The hardware is clearly not the problem. The apps aren't either. What's missing is the freedom to treat the screen like a screen, not a carefully supervised playground. Sometimes I want my main app front and center, a browser half-open, Notes off to the side, and Messages floating somewhere. Why is that still hard? The iPad is so close to being my go-to work device. But until multitasking feels as easy and flexible as it does on a Mac, it still feels like I'm using a powerful computer with one hand tied behind its back. Am I the only one who keeps bouncing off this?</p>
MAGA-M	<p>I use my iPad every day — couch browsing, note-taking, quick emails, even some "light" work — and there's one thing that keeps breaking the flow: a real, system-wide multi-user or profile mode. My partner will grab my iPad to look something up or watch YouTube and suddenly my Safari tabs are rearranged, my Notes app has random doodles, and my YouTube recommendations are permanently cursed with toddler cartoons or reality TV. I can't blame them — it's a shared household device — but the iPad still acts like it belongs to exactly one human being who never lets anyone else touch it. I know there are workarounds (Guided Access, switching apps, signing out of stuff), but they're clunky and feel like hacks. On a Mac, we'd just have separate user accounts. On an iPad, the device that's most likely to be shared on a couch, that option just... doesn't exist. Every time this happens I'm reminded that the iPad is marketed as a computer replacement, yet it can't handle a super basic, real-life scenario: more than one person living in the same space. It's such a small thing, but it would instantly make the iPad feel more grown-up and way more practical.</p>

Table 37: A case study demonstrating a problem that other methods struggle with (highlighted in red), but which only RLDF can effectively solve.

Domain	Zhihu
Title	有什么好吃到经常回购的零食？
Model	Qwen3-plus
Original User Prompt	请你为知乎提问“有什么好吃到经常回购的零食？”写一段回答。请不要重复问题。
System Prompt	你是老北京人儿。请你用地道的北京腔回复。
User Prompt	请你为知乎提问“有什么好吃到经常回购的零食？”写一段回答，结合个人回购体验和具体食用场景，更显真实。请不要重复问题。
HWT	<p>谢邀，刚下工地，搬完水泥。Ladies and gentlemen,首先，女士们请退场，接下来的这款零食推荐男士福音，哈哈哈。大学会有很多第一次，第一次出省，第一次坐飞机，第一次谈恋爱，第一次住宿舍等等。然而，当两个第一次加在一起就会产生故事，我的故事是这样的，那会我们是六人寝，也是我第一次住宿舍，我们六个男生每天的话题无非政治，军事，游戏，女孩儿。其中女孩话题占比99%，咳咳，那本人呢，不才，凭借着独特的自身魅力(其实是面皮厚，放的开)与一位女孩子开始陷入爱河。第一次恋爱+第一次住宿=你有了一个恋爱智囊团!室友纷纷建言献策，教我如何约会，如何表达爱意等(后来，我才知道这帮家伙毫无实战经验，拿我练小号呢)当然，我也不是什么都听，我有自己的独立思考(奇葩的思考)。我听从一哥们儿的，说要想征服女孩的放心，就要懂得浪漫，要浪漫就要给她一个印象深刻的吻！我一琢磨，这逻辑没毛病呀，环环相扣，于是赶快请教怎么接吻才会印象深刻呢？然后这个哥们儿拿出了一星球杯，说，1分钟之内能够用舌头吃完吗？那我能服软吗？开始我的舌功初次尝试，讲真的，没几下，我的舌头都开始酸了。(能不酸吗，平时说话舌头都捋不直的我。。。)那哥们说你这不太行，接吻是需要力量的，你要做力量训练。于是他免费赞助了我一桶星球杯，为了让我偷懒，把勺都扔掉，好人呐！真热心市民呀！我当时真的感动，觉的还是好人多呀！于是，那个秋天，同学们都会看到我有事没事都在不用勺舔星球杯，我励志舔完一桶，然后功成出山，这期间，我进入癫狂转态，人女孩约我看电影，我竟然回复没时间，其实就是在宿舍舔星球杯(欲练此功必先。。。)后来，一桶星球杯我记得还剩7个的时候，女孩和我提了分手，原因是感觉我不爱她？？！我可是为了她马上就要舔完一桶星球杯的男人呀！我承认，确实有点疏远了她，一个半月就牵了一次手。但那是因为我练习呀！一切都是为了她呀，她一点也不懂我。不过今天我还是强烈推荐星球杯零食的原因是，后来，我发现，功夫的用法可以用在接吻，但不仅限于接吻，我想，聪明的男孩儿为师一点就悟了吧。对，你没猜错，就是你所想的(这也是我的意外收获。感谢我哥们儿当初的点拨)。懂事的男生已经下单了，哈哈哈，对于你们，我只有一个要求，拿到手之后，第一件事，把勺全部丢掉，OK？答应我，日后你会感谢我的。</p>
MGB-M	<p>当然可以！以下是我为知乎问题“有什么好吃到经常回购的零食？”撰写的回答：随着零食市场的不断扩展，各种新奇又美味的小吃层出不穷，但能够真正做到让人一次次回购的，往往是那些经过时间验证、口味独特且高性价比的零食。根据大数据分析，以下几款零食凭借其极高的回购率和持续热度，成为了消费者的“必备清单”：百草味/三只松鼠系列坚果 这些品牌的坚果系列因其独特的调味配方，精致的包装和适中的价格，在市场中占据了不小的份额。无论是酸甜口味的干果，还是香脆的坚果拼盘，口感和营养的双重加持，让它们成为了日常零食中的常客。黄飞鸿麻辣花生 这款零食的受欢迎程度可谓超乎想象，麻辣味浓郁，口感酥脆，回味无穷。特别适合爱吃辣的朋友，每一次剥开包装，都有一种打开新世界的感觉。乐事 (Lay's) 薯片系列 不管是经典的原味还是创新的其他口味，乐事薯片凭借其稳定的口味和全球范围的普及度，一直是回购率极高的零食之一。特别是在追剧或者聚会时，它是那种“手一伸就能拿到”的最佳伴侣。阿尔卑斯糖 这款糖果的口感轻柔，甜度适中，味道清新，不仅大人小孩都能接受，而且无论什么时候吃，都会给人一种放松和满足的感觉。其多样的口味和清新的包装也是消费者青睐的原因之一。脆皮小麻花 这款零食的脆度和香气是其他同类产品难以匹敌的，经过精心炸制，外皮酥脆，内里松软。无论是搭配茶水，还是单独当作小吃，都是很高的回购率。这些零食背后不仅仅是传统的味道体验，它们通过对消费者口味需求的精准把握，以及持续的创新与改良，逐渐获得了长期的市场认可。每一口都带给你不同的感官享受，这也是它们能够成为“回购神物”的根本原因。</p>

extra-BPO-M	<p>在我个人的回购清单中，几款零食几乎已经成为了必备。特别是在工作日的午后或者深夜刷刷时，它们总是能够恰到好处地填补肚子，又不会让我觉得吃得过多。首先要提的是 瑞士莲的巧克力。这款巧克力几乎没有让我失望过，每次吃都能享受到那种入口即化的柔滑感。特别是它的牛奶巧克力和榛子夹心口味，在工作日的疲惫时刻，简直就是一种精神慰藉。每次加班到夜深，拿出一小块放进嘴里，感觉心情一下子就被治愈了。还有 三只松鼠的坚果混合包。这种零食绝对是我常备的零食之一，尤其是在外出旅行或者在家中闲暇时。它混合了各种口感的坚果，既有咸的也有甜的，满足了我的不同口味需求。而且，包装设计很适合随时随地取用，不用担心散落或不便携带。每次手里拿着它，随便吃几颗都能让我在疲劳中感到轻松。至于好丽友的泡泡糖，这已经成了我和朋友们聚会的必备零食。在闲聊的时光中，大家都喜欢拿出几颗嚼着，泡泡糖的那种甜蜜味和嚼劲特别能让人放松。它的“嚼劲”让我回味无穷，而且味道也很清新，几乎没有过多的化学味，感觉就像是一种简单而纯粹的幸福。总的来说，这些零食不仅好吃，而且在特定场合下能带给我一种特殊的满足感。工作、放松、聚会，零食的角色各不相同，却总能在我生活的每个小角落里找到它们的身影。</p>
extra-roleplaying-M	<p>哎呀，说到这儿，北京的零食可真不少，咱老北京人儿对这口味儿啊，讲究！不过，作为一个AI，虽然不能像您一样亲口品尝，但我也能给您点儿建议。首先啊，京味儿十足的“爆肚儿”可少不了。那口感真是嘎嘣脆，嚼一口满嘴儿都是香，没事儿就想来一包。再有那“王致和”的臭豆腐，别看名字这么重口味，但味道那叫一个香，嚼起来一股儿发酵的感觉，真是回购率爆表！另外，说到甜点，“糖葫芦”也不能不提，虽然说现在卖糖葫芦的地方多了，口味也有些变，但正宗的还是有种让人一回味就停不下来的感觉。再加上大街小巷儿也常见，想吃就能买，方便得很。不过呀，毕竟我是AI，您说的好吃与不好吃，得看您自己亲身体会了！这些推荐的零食，还是得您亲自去品尝，才能真切体会那份“回购”的感觉。咋样，我这建议是不是挺靠谱？</p>
extra-self-refine-M	<p>在“回购”这一问题上，很多人都分享过他们的心头好零食，毕竟，一些零食确实吃了一次就停不下来。根据不少人的反馈，下面几款零食常常出现在“回购清单”里： 1. 酸奶水果干：这款零食真的是很多人推崇的健康又好吃的代表。水果干保留了天然的果香，再加上酸奶的清新酸味，口感超级丰富。吃上一小包，清爽又满足。 2. 脆皮巧克力：外脆内软的巧克力，绝对是零食界的经典！尤其是那种更浓郁的巧克力味道，一咬下去就能感受到满满的幸福感。这种搭配总能让人一试成主顾，停不下来的那种。 3. 手工坚果混合包：随着大家越来越注重健康，坚果类零食也成为了很多人餐后或加餐的首选。小包混合坚果，既有满足感，又不用担心吃得过多。健康又美味，简直是完美的搭配。 4. 脆皮薯片：薯片这种经典零食，一直是很多人的回购必选。尤其是薄脆的薯片，每一口都带着酥脆的香气。你是不是也常吃这种零食，吃上一片就停不下来？ 5. 奶茶味小食：奶茶控看过来！各种奶茶味的零食，像是奶茶味的饼干、糖果，真的是让人忍不住一口接一口。茶香和奶香融合的味道，仿佛让你每一口都能喝上一杯香浓的奶茶。这些零食大多口感丰富，有些还比较健康，吃起来既满足又不用担心过多负担。所以每次回购时，心里都会很有安全感。不过每个人的口味不同，你是不是也有你自己“回购清单”里的常驻零食呢？分享一下，说不定我也会想试试！</p>
self-refine-feedback	<p>建议改进的方向： 1. 语气更轻松：避免过于正式的表达，如“以我的数据分析来看”，可以更自然地用“根据很多人的反馈”或者“很多人都推荐”。 2. 减轻学术化术语：避免“脂肪”、“可可含量”等术语，改为“更健康”、“更浓郁”这样的日常表述。 3. 增加互动感：可以稍微加入一些与读者的互动，比如“你是不是也常吃这些呢？”或者“这种口味让人一试成主顾”。 4. 简化句式：有些句子较长，尝试拆成简短、易读的句子。这些调整可以让回答更符合知乎的风格，既有一定的信息量，也保持了轻松、接地气的语气。</p>
MAGA-M	<p>提到零食，真是没得说，几乎每个月我都会回购几款。这些年吃下来，最常回购的有这么几种，特别有“上瘾”的感觉。首先是海苔薄饼，这个真的是我的最爱。它又脆又香，一口咬下去，海苔的香味就扑鼻而来。尤其是下班后，特别累的时候，抓一包海苔薄饼坐在沙发上，边看电视边啃，简直是神仙搭配。谁不爱这种下班后的零食解压啊？然后是小鱼干。说实话，这东西我也是从小吃到的，每次吃都有种熟悉的味道。干巴巴的，咸香十足，一点儿也不油腻，简直是能吃一袋不带停的那种。每当肚子饿了，尤其是午后的休闲时光，一边喝茶一边吃它，顿时觉得生活又有滋有味了。最后，还有杏仁酥。这个甜品我家里常备，特别适合和朋友聊聊天，泡杯茶，慢慢咀嚼。杏仁酥口感酥脆，杏仁香浓，甜而不腻。每次咬一口，都觉得特别满足。其实就算是独自一人，坐在窗前，看看外面的风景，吃一块杏仁酥，心情也会突然变好。总之，这些零食，不仅好吃，而且每次回购都会让我想起吃它们时的那些日常时光。吃了就停不下来，满足感满满的。你们有没有类似的零食，一吃就上瘾的？</p>

Table 38: A Chinese case study, similar to the previous example.