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

Anonymous ACL submission

Abstract

Machine-Generated Text (MGT) is becoming increasingly difficult to distinguish from Human-Written Text (HWT). This trend has exacerbated malicious activities such as fake news and online fraud. The generalization ability of fine-tuned detectors relies heavily on dataset quality, and simply expanding the sources of MGT may become increasingly insufficient. Further augmentation of the generation process is required. Based on HC-Var's theory, enhancing the human-like alignment of MGT not only facilitates robustness testing of existing detectors but also boosts the generalization ability of detectors fine-tuned on such aligned MGT datasets. Therefore, we propose the **M**achine-**A**ugment-**G**enerated Text via **A**lignment (MAGA) Detection Benchmark. MAGA integrates several alignment methods, ranging from prompt construction to **G**enerator-**D**etector **A**dversarial **R**einforcement **L**earning (GDARL) and the reasoning process. In our experiments, the RoBERTa detector fine-tuned on MAGA achieves an average improvement of 4.60% in generalization AUC. Conversely, the aligned MGTs in MAGA also lead to an average decrease of 8.13% in the AUC of selected detectors. We hope the MAGA Benchmark will provide valuable insights for future research on the generalization ability of MGT detectors.

1 Introduction

In recent years, Large Language Models (LLMs) have advanced rapidly, with remarkable improvements in their language comprehension capabilities, precision in instruction execution, and ability to generate complex text to fulfill diverse tasks (Brown et al., 2020; Ouyang et al., 2022). It has become increasingly difficult for humans to distinguish between Machine-Generated Text (MGT) and Human-Written Text (HWT) (Chang et al., 2024). This powerful capability, however, poses substantial risks of abuse. Hanley and Durumeric (2024) noted that in 2022, the proportion of AI-generated news on websites notorious for disseminating false rumors increased by approximately 457%. Such misinformation is not only used in fake news (Zellers et al., 2019; Dugan et al., 2023) but also utilized by numerous malicious 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 these potential risks in information dissemination, the development of efficient and robust MGT detectors is of paramount importance.

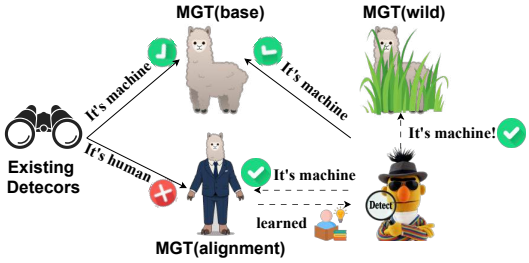

Figure 1: More human-like aligned MGT not only evades detection by existing detectors but also facilitates the fine-tuning of neural-based detectors, thereby enhancing their generalization capability for in-the-wild detection.

To address this need, numerous specialized detection methods and benchmarks have been developed. Prior datasets increased detection difficulty by incorporating diverse distributions, a direction that marked its first pinnacle in M4 (Wang et al., 2024d). However, the generation of individual texts in these datasets remained straightforward, relying on a limited set of fixed prompt templates. RAID (Dugan et al., 2024) is the first benchmark to significantly prioritize the robustness of detectors against adversarial attacks.

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	✓	✓	✗	✓	✗	✗
MULTITuDE (Macko et al., 2023)	74.1k	✗	✓	✓	✗	✗	✗
M4 (Wang et al., 2024d)	122k	✓	✓	✓	✗	✗	✗
HC-Var (Xu et al., 2024)	145k	✓	✗	✗	✗	✗	✗
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 publicly available data sources of MGT. The term "Original Size", introduced by RealDet (Zhu et al., 2025), refers to the collection of HWTs and direct MGT outputs, excluding rule-based transformations such as those employed in RAID (Dugan et al., 2024). A more comprehensive comparison is provided in Table 5.

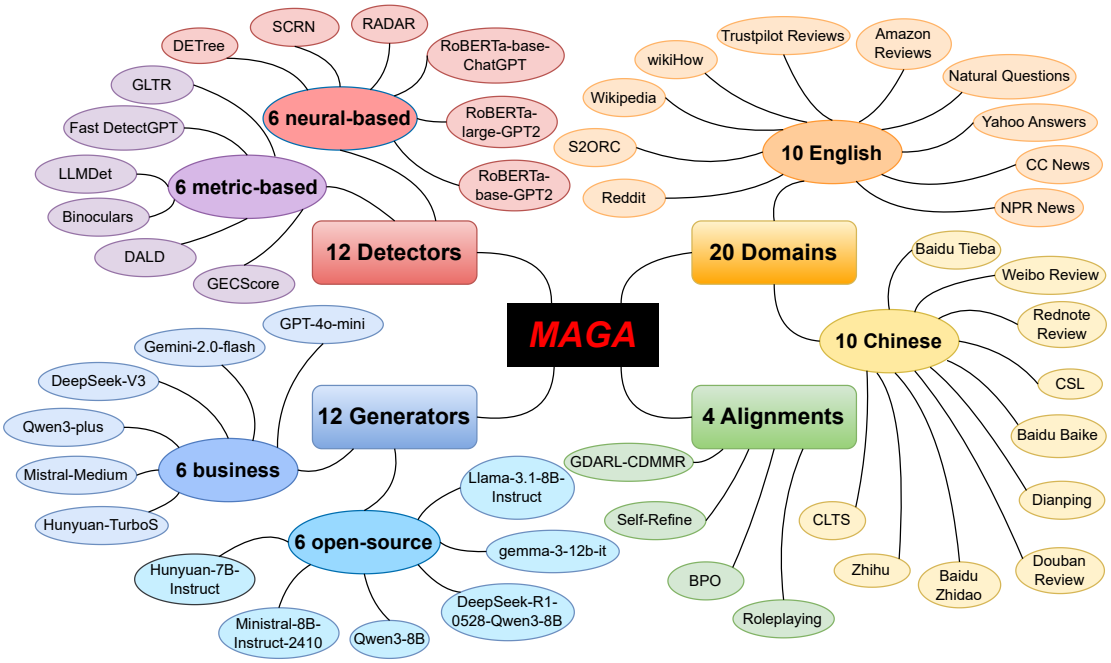

Figure 2: Overview of MAGA-Bench. The construction of the dataset encompasses 20 domains, 12 generators, 2 languages and 4 alignment methods. Selected detectors are evaluated on this dataset. Various decoding strategies are also employed, with details provided in Appendix E.

The implemented attacks, such as homoglyph transformations (e.g., randomly transforming 'a'-'z' into 'a'-'z'), successfully compromise detector performance, particularly for metric-based detectors. In line with this, Hans et al. (2024) observed that the performance of their proposed Binoculars detector drops to near zero when evaluating highly randomized MGT. Nevertheless, this form of text manipulation disrupts the surface structure of the text, thereby hindering neural-based detectors, exemplified by RoBERTa (Liu et al., 2019), from learning and understanding semantics during fine-tuning, even though the semantic meaning remains interpretable to human readers.

HC-Var (Xu et al., 2024) indicates that MGT with a higher degree of human-like alignment exhibits relevant features closer to those of HWT. This proximity assists neural detectors in learning more effective decision boundaries, thereby enhancing generalized detection. Therefore, we augment the generation process while ensuring the diversity of data sources across 20 domains, 12 generators, 2 languages, and various decoding strategies. Our main contributions are: (i) we propose the **M**achine-**A**ugment-**G**enerated Text via **A**lignment Detection Benchmark (MAGA-Bench), which

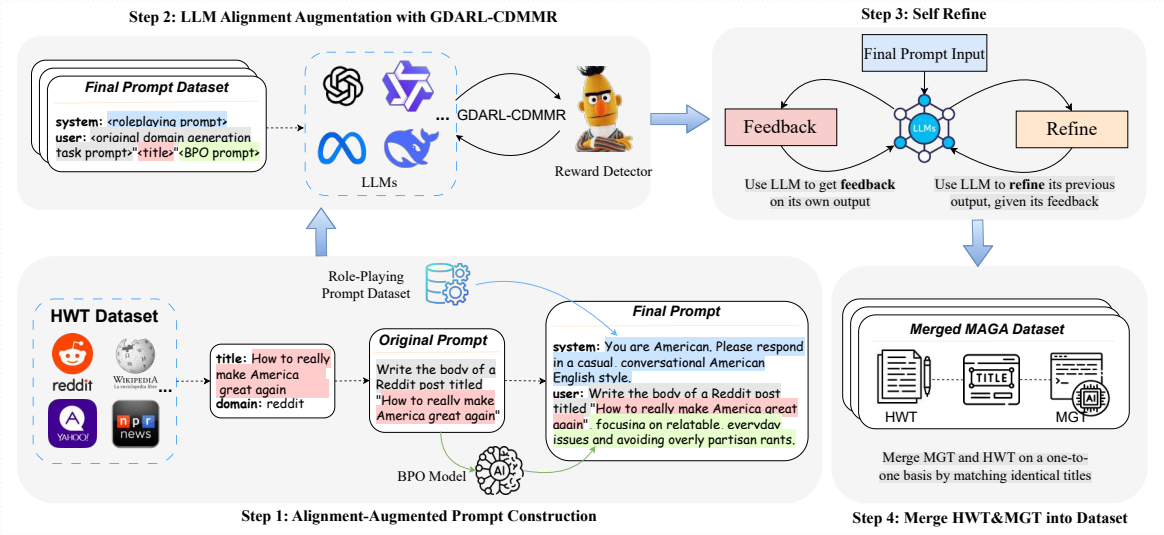

Figure 3: MAGA Pipeline, illustrating the construction process for the MAGA split of the MAGA dataset. MAGA split is the fully enhanced split that integrates all four alignment methods (see Table 6, Appendix A for detailed statistics of the splits).

incorporates 4 human-like alignment methods to enhance MGT generation; (ii) we propose a **C**ross-**M**utual **R**eward (CMR) component, which facilitates the alignment process of large language models in support of the **G**enerator-**D**etector **A**dversarial **R**einforcement **L**earning (GDARL) framework; (iii) comprehensive experiments demonstrate that MAGA reduces the average AUC of 12 selected evaluation detectors by 8.13%, while the RoBERTa detector fine-tuned on MAGA improves the average generalization AUC by 4.60% across 4 selected external datasets.

2 Related Work

Previous MGT detection benchmarks (Table 1 for key representatives and Table 5 in the Appendix for a comprehensive compilation) mainly focused on expanding the diversity of the overall dataset distribution. Since TuringBench (Uchendu et al., 2021) extended the multi-generator (or multi-model) dimension, the dimensions used for expansion have become increasingly diverse. RuATD (Shamardina et al., 2022) and RoFT (Dugan et al., 2023) were among the earliest works to introduce the multi-domain dimension. HC3 (Guo et al., 2023) pioneered the multi-lingual dimension. MGTBench (He et al., 2023) and CHEAT (Yu et al., 2023) pioneered the introduction of adversarial considerations, and OpenLLMText (Chen et al., 2023) was among the first to explicitly incorporate multiple decoding strategies. This evolutionary trajectory established a significant milestone with M4 (Wang et al., 2024d), which fully integrated three core dimensions: multi-domain, multi-generator, and multi-lingual. Subsequently, RAID (Dugan et al., 2024) advanced the field by comprehensively introducing a wide range of adversarial methods for the first time and covering sampling strategies alongside the multi-domain and multi-generator dimensions. RealDet (Zhu et al., 2025) addressed the absence of multi-lingual support in the original RAID release but did not include sampling coverage.

Other MGT detection benchmarks in Table 5 concentrate more on specific research scenarios. RuATD, AuText2023 (Sarvazyan et al., 2023a), CCD (Wang et al., 2023), IMDGSP (Mosca et al., 2023), and MultiSocial (Macko et al., 2025) are dedicated to MGT detection for Russian, Spanish, source code, academic papers, and social media texts, respectively. LLM-DetectAIve (Abassy et al., 2024), MixSet (Zhang et al., 2024), and LAMP (Chakrabarty et al., 2025) focus specifically on text paraphrasing attacks, which Beemo (Artemova et al., 2025) further enhances through expert editing. MULTITuDE (Macko et al., 2023) conducted the first study on cross-linguistic family generalization, and MAGE (Li et al., 2024) was the first to establish multiple out-of-distribution testbeds for generalization research. Despite these targeted efforts, the generalization problem

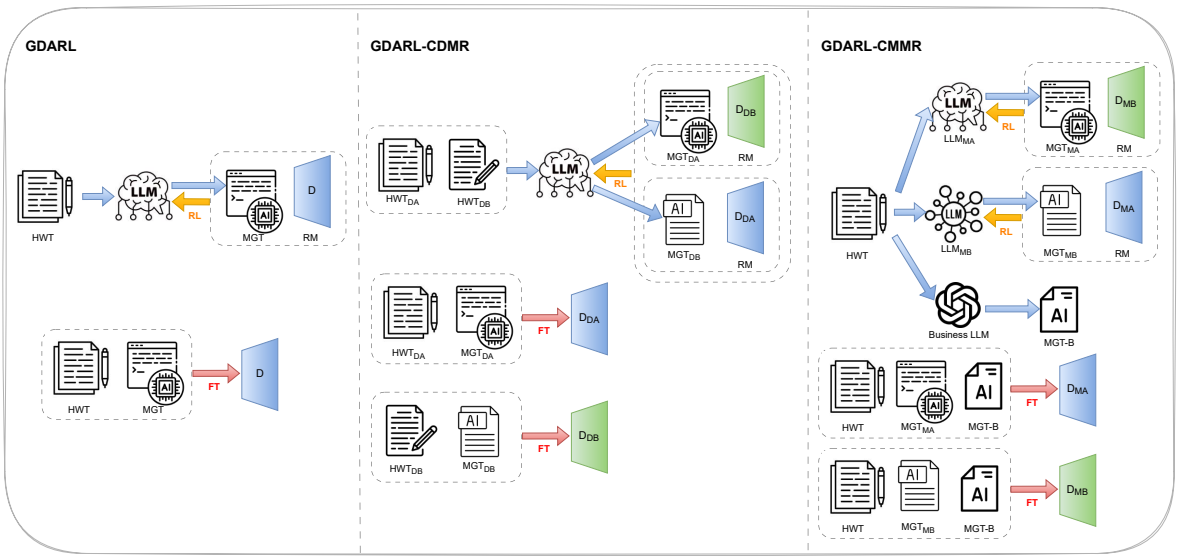

Figure 4: GDARL and its integration with the **C**ross-**M**utual **R**eward (CMR) component. The term GDARL serves as a collective designation for GDARL-like frameworks and also represents their most general form. The illustration depicts the two standard applications of CMR, namely GDARL-CDMR (Cross Domain Mutual Reward) and GDARL-CMMR (Cross Model Mutual Reward).

remains one of the most critical unresolved challenges in MGT detection. HC-Var (Xu et al., 2024) found that RoBERTa fine-tuned on more human-like aligned MGTs exhibits stronger generalization in detection tasks. Our work builds upon HC-Var's theory and aims to contribute to the research on the generalization problem in MGT detection.

3 The MAGA Dataset

Figure 2 presents the components of the MAGA dataset. The MAGA dataset consists of two subsets: the English MAGA and MAGA-cn. Each subset comprises six splits, where MGB represents the unenhanced split, MAGA denotes the split integrating four distinct alignment methods, and the remaining four splits individually apply each single alignment method. To construct the MAGA dataset, taking the English subset as an example, we first sample 72k HWTs annotated with titles, which are consistently reused across all splits. For each split, we generate one MGT corresponding to each HWT by embedding the title into a prompt template. Consequently, each split contains 144k texts. Detailed statistics and data examples are provided in Appendix A.

3.1 Domains

Various domains raise detection difficulty (Li et al., 2024; Dugan et al., 2023) and improve the fine-tuned RoBERTa's performance (Wang et al., 2024b). Taking English as an example, our sources cover social conversations (Reddit), summarization (S2ORC), general knowledge (Wikipedia, wikiHow), reviews (Trustpilot Reviews, Amazon Reviews), direct Q&A (Yahoo Answers, Natural Questions), and news creation (CC News, NPR News), which challenge various capabilities of LLM. Details about all the domains (including the Chinese domains) are provided in Appendix B.

3.2 Generators

Following Sarvazyan et al. (2023b), we first selected the largest (commercial) LLM 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. The models we selected are shown in Figure 2, with details provided in Appendix C.

3.3 Alignment Methods

A detailed theoretical analysis and introduction of all alignment methods we used, along with their specific execution, are discussed in Appendix D.

Roleplaying To facilitate roleplaying, we construct a custom prompt set named MAGA-ROLE-80, from which we randomly sample a roleplaying prompt as the system prompt for each original prompt. To the best of our

	ACC		TPR(M.)		TNR(H.)	AUC		ACC(@FPR=5%)	
Alignment?	✗	✓	✗	✓	-	✗	✓	✗	✓
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↓)	93.03	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↓)	94.23	92.17(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↓)
neural.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	86.76	79.68(7.08↓)	83.23	75.17(8.06↓)
DALD	80.01	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: Attack performance against evaluation detectors on the MAGA split. The "✗" symbol in the "Alignment" column denotes the unenhanced MGB split, whereas "✓" represents the MAGA split. ACC, TPR, and TNR are evaluated based on the default threshold of each detector. Labels M. and N. are introduced to facilitate understanding. TPR(M.) denotes the ACC exclusively on MGT, and TNR(H.) denotes the ACC exclusively on HWT. The highest values are bolded, and the second highest values are underlined.

knowledge, among all machine-generated text detection benchmarks listed in Table 5, our work is the only one that leverages system prompts. Our roleplaying prompts incorporate the core philosophy of "Talk Less, Call Right" from RRP (Ruangtanusak et al., 2025), where the first sentence defines the baseline identity and the second sentence specifies the linguistic style.

BPO (Cheng et al., 2024): BPO is a prompt optimization model. Its optimization aims to ensure that the texts generated by BPO-optimized prompts are better aligned with HWTs. The BPO model always appends additional information to the original prompt. In this work, we directly employ the BPO model released by Cheng et al. (2024)[1].

Self-Refine Self Refine[2] denotes an iterative refinement procedure wherein a large language model generates an initial text, reflects upon its own output to provide feedback, and subsequently rewrites the content based on this feedback. The specific prompts deployed for this self-refinement process are provided in Tables 27, 28, and are discussed in detail in Appendix G.

[1] https://huggingface.co/zai-org/BPO

[2] Although Self Refine supports multi-round iterations, we adopt a single-round execution in our practical implementation, the rationale for which is discussed in Appendix N.3.

GDARL-CDMMR GDARL is the general term for existing GDARL algorithms such as RADAR (Hu et al., 2023) and HUMPA (Wang et al., 2025a), and it also represents the vanilla formulation of such approaches. We observe that the vanilla GDARL exhibits minimal change throughout training due to reward overfitting under identical data distributions[3]. To address this limitation, we introduce a **Cross-Mutual Reward (CMR) component**. This component incorporates the theoretical insights from HC-Var, which state that cross-distribution generalization primarily relies on the latent essential relative features between HWT and MGT (Xu et al., 2024).To illustrate CMR, we consider the application of GDARL-CDMMR (Cross Domain&Model Mutual Reward). Specifically, the text generation models (LLMs) and the target domains of the generated text are divided into two distinct groups, denoted as GA and GB. In the adversarial round t, the LLMs in group GA first generate MGTs, denoted as M. HWTs represented as H. The detector $\mathcal{D}$ within group GA is trained with parameters ϕ according to the following objective function:

[3] See Appendix N.1 for further discussion about this as well as why we did not utilize a specific GDARL framework and introduce a CMR component.

	S-M4		C-M4GT		C-MAGE		C-HC3		AVERAGE	
Metric	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	76.40	55.19	98.45	94.59	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
DETree*	99.78*	99.32*	96.63*	94.56*	99.44*	96.82*	98.55*	95.27*	98.60*	96.49*
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	79.65	64.67	56.67	**99.94**	**97.19**	84.97	79.88
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	89.14	**84.10**	63.79	**58.89**	99.59	**97.19**	85.56	**82.32**
DALD	88.15	85.01	85.41	78.87	64.96	56.73	99.76	**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)	94.63	90.58	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	**4.24**	**4.83**	**5.42**	**5.62**	**3.73**	**0.72**	**5.00**	**2.38**	**4.60**	**3.39**

Table 3: Experiments evaluating generalization capability. Performance of evaluation detectors as baseline alongside our R-B MGB and R-B MAGA on external datasets. R-B CGPT is fine-tuned on HC3 and is marked with an asterisk. DETree utilizes reference vectors constructed from dataset texts including M4, M4GT, and MAGE for detection, and is also marked with an asterisk. The highest values excluding those with asterisks are bolded, and the second highest values are underlined. ACC(5%) is an abbreviation for ACC(@FPR=5%).

$$\mathcal{J}_{\mathcal{D}}(\phi^{GA,t}, t) = -\mathbb{E}_{x_m^{GA,t} \sim M^{GA,t}} \left[\log \mathcal{D}_{\phi}^{GA,t}(x_m^{GA,t}) \right] - \mathbb{E}_{x_h^{GA} \sim H^{GA}} \left[\log \left(1 - \mathcal{D}_{\phi}^{GA,t}(x_h^{GA}) \right) \right] \quad (1)$$

The detector in group GB is trained in an identical manner. Subsequently, the detector from group GB provides rewards for the text generated by the text generation models in group GA to facilitate reinforcement learning. The reward function is formulated as follows:

$$r^{GA,t}(x, \phi^{GB,t+1}) = 1 - \mathcal{D}_{\phi}^{GB,t+1}(x) \quad (2)$$

The training process for the text generation models in group GB follows the same procedure. Appendix L provides a detailed discussion regarding the hyperparameter configurations and the final grouping strategy for GDARL-CDMMR.

3.4 MAGA Pipeline

Figure 3 illustrates the MAGA pipeline for constructing the MAGA split, which integrates all four methods to elucidate our data construction process. In the first stage, we construct the prompt set utilizing roleplaying and BPO. This prompt set is subsequently applied to the GDARL-CDMMR process in the second stage, followed by self-refining in the third stage. In the final stage, the HWT and MGT are merged on a one-to-one basis to produce the definitive MAGA split.

4 Experiments

4.1 Experimental Setup

Evaluation Detectors Our selection of open-source Evaluation Detectors extends the choices by RAID (Dugan et al., 2024) by incorporating four more recent detectors. The Evaluation Detectors we selected are listed in Figure fig:MAGA Overview, where SCRN, DETree, DALD, and GECScore are the four recent detectors. A detailed introduction to these detectors is provided in Appendix F.

Evaluation Metrics Our evaluation spans accuracy (ACC), True Positive Rate (TPR), True Negative Rate (TNR), Area Under the ROC Curve (AUC), and ACC(@FPR=5%), with False Positive Rate denoted as FPR. Setting the MGT label to 1, TPR and TNR signify the standalone accuracies on MGT and HWT, respectively. ACC, TPR, and TNR are assessed under default probability thresholds (e.g., 0.5 for RoBERTa; see Table 12 for the full list). ACC(@FPR=5%) is the metric adopted by RAID (Dugan et al., 2024) and denotes the detection accuracy evaluated at a specific threshold, which is determined by fixing the ACC of HWT to 95%. Given that the default threshold is not always appropriate, we tend to focus more on AUC and ACC(@FPR=5%).

	Attack			Generalization		
Metric	TPR	AUC	ACC(@FPR=5%)	TPR	AUC	ACC(@FPR=5%)
MGB	54.09	71.4	67.41	53.13	82.91	75.65
MAGA-extra-GDARL-CDMMR	49.95(4.14↓)	67.91(3.49↓)	65.18(2.23↓)	60.97(7.84↑)	85.02(2.11↑)	77.28(1.63↑)
MAGA-extra-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↑)
MAGA-extra-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↑)
MAGA-extra-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	45.10(8.99↓)	63.27(8.13↓)	62.70(4.71↓)	71.30(18.17↑)	87.51(4.60↑)	79.03(3.38↑)

Table 4: Experimental results of attack performance and generalization capabilities on the MAGA-extra splits. "Attack" represents the average performance of evaluation detectors under the experimental setup specified in Table 2. "Generalization" represents the average generalization performance on external datasets of the RoBERTa fine-tuned on the corresponding split, following the identical experimental setup as Table 3. This analysis also serves as an ablation study, with more comprehensive details provided in Table 15, Appendix J.

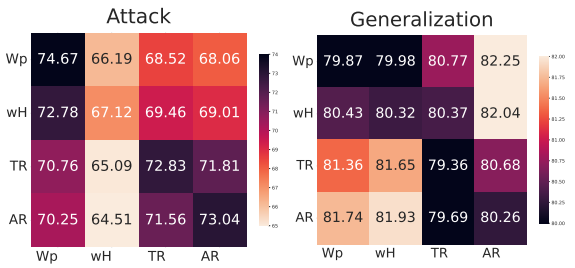

Figure 5: Matrix analysis of GDARL-CDMR for Attack and Generalization (meanings are consistent with Table 4) AUC performance. The horizontal axis denotes the domain of the text used to train the reward detector, while the vertical axis denotes the LLM generation domain. The abbreviations Wp, wH, TR, and AR correspond to Wikipedia, wikiHow, Trustpilot Reviews, and Amazon Reviews, respectively.

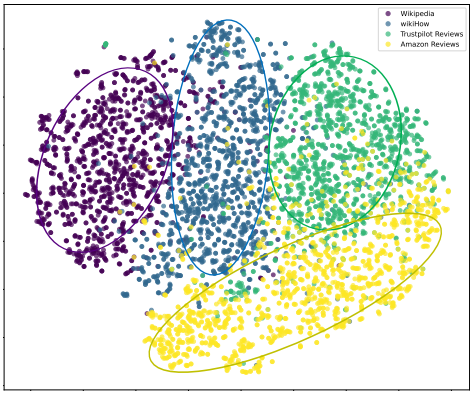

Figure 6: T-SNE visualization of HWTs on the domains specified in Figure 5.

External Datasets We incorporate workshop competition versions[4] of four datasets: M4 (Wang et al., 2024d), M4GT (Wang et al., 2024c), MAGE (Li et al., 2024), and HC3 (Guo et al., 2023), detailed in Appendix I. For brevity, these are denoted as S-M4 (Wang et al., 2024b), C-M4GT, C-MAGE, and C-HC3 (Wang et al., 2025c), respectively.

4.2 Main Experiment Results

Table 2 presents the empirical results of our Attack experiments. Compared to the MGB split, the Evaluation Detectors on the MAGA split exhibit an average decrease of 8.13% in AUC and 4.71% in ACC(@FPR=5%). The average performance degradation identified in neural-based detectors tightly aligns with that of metric-based detectors. This consistent reduction indicates that our MAGA relatively achieves a comprehensive and stable attack effect across different detection mechanisms rather than showing a biased preference toward a specific category, such as rule-based random text perturbations.

Table 3 outlines the results of our generalization assessment. On External Datasets, the R-B MAGA delivers an average improvement of 4.60% in AUC and 3.39% in ACC(@FPR=5%) compared to the baseline R-B MGB.

Furthermore, the results on the Chinese MAGA-cn subset are similar[5].

4.3 Ablation Study

Table 4 presents additional experimental results across various MAGA-extra splits, which also serves as a ablation study to validate the effectiveness of each alignment method. Detailed results are provided in Appendix J. Figure 5 illustrates the ablation study regarding the grouping configurations of GDARL-CDMR, where the main diagonal represents the performance prior

[4]See Appendix I for detailed introduction to External Datasets.

[5]See Appendix J for experiments on Chinese subset and other experiment details.

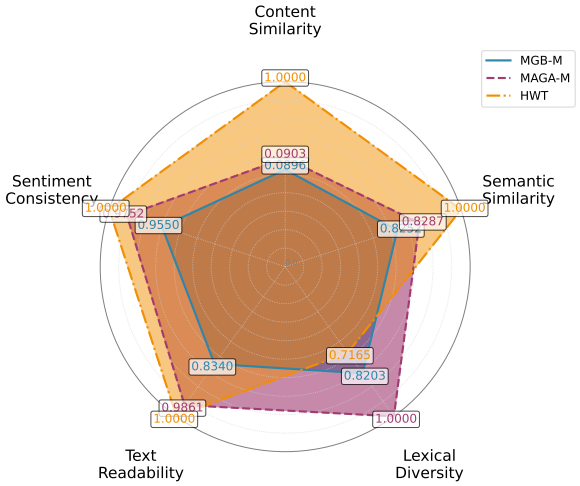

Figure 7: Multidimensional Text Analysis. MGB-M and MAGA-M represent MGTs on the MGB split and MAGA split, respectively. Detailed metrics and specific values for each dimension are presented in Table 16 and Appendix K.1.

to augmentation. As indicated in Table 22, vanilla GDARL yields nearly no change in performance due to reward overfitting. Figures 5 and 6 demonstrate that larger domain discrepancies lead to more substantial performance gains for GDARL-CDMR[6]. These observations demonstrate the significance and efficacy of the CMR component.

5 Analysis

We consolidate several quantitative analysis metrics from prior literature regarding the discrepancies between MGT and HWT into five dimensions, as illustrated in Figure 7. The value for each dimension represents the mean Normalize Score of all constituent metrics calculated across the entire corpus. Appendix K.1 elaborates on the detailed mathematical formulations.

Content Similarity and Semantic Similarity HC-Var (Xu et al., 2024) demonstrates that more human-like MGT exhibits higher MAUVE (Pillutla et al., 2021) similarity with HWT. Corresponding to MAUVE, we select BERTScore as the analysis metric for semantic similarity. Content similarity is evaluated concurrently via ROUGE, BLEU, and METEOR. The results indicate that MAGA-M achieves slightly higher content and semantic similarity scores with HWT than MGB-M.

[6]A more detailed analysis is in Figures 14, 15, and 16, which are discussed in Appendix J.

Lexical Diversity M4 (Wang et al., 2024d) reveals that utilizing multiple generators significantly increases the vocabulary size of unigrams and bigrams, thereby enriching the vocabulary of MGTs and facilitating the training and evaluation of MGT detectors. Furthermore, the unigram and bigram vocabulary size of MGT reported in M4 (Wang et al., 2024d) already exceeds that of HWT. Extending beyond bigram vocabulary size, we incorporate TTR and Yule's K metric. Our analysis shows that the lexical diversity of MGB-M surpasses that of HWT, and MAGA-M further amplifies this gap. It shows a divergence, although this aligns with the findings in M4 (Wang et al., 2024d) and may provide benefits.

Text Readability AA (Uchendu et al., 2020) notes that the Flesch-Kincaid score of MGT is higher than that of HWT. In addition to Flesch-Kincaid, we incorporate SMOG and Dale-Chall into our analysis. The empirical results confirm that the Flesch-Kincaid score of HWT is significantly higher than that of MGB-M. MAGA-M substantially narrows this gap to HWT, demonstrating that its textual expressions are neither overly simplistic nor fragmented. On SMOG and Dale-Chall, MGB-M scores significantly higher than HWT, whereas MAGA-M exhibits a marked decrease, approaching HWT closely. This trend reflects a substantial reduction in obscure and difficult vocabulary.

Sentiment Consistency HC3 (Guo et al., 2023) conducts a comparative sentiment consistency analysis between HWT and MGT across negative, neutral, and positive categories. HC3 concludes that MGT is generally conservative, exhibiting a pronounced deficit in negative sentiment. Our experiments yield consistent findings.

6 Conclusion

We introduce MAGA-Bench to address generalization bottlenecks in MGT detection. Evaluations show that aligned MGTs within MAGA degrade existing detectors' AUC by 8.13% on average. Conversely, fine-tuning on MAGA boosts detector robustness, improving generalization AUC across diverse external datasets by 4.60%. We hope alignment augmentation serves as a promising path toward robust and generalized MGT detection.

Limitations

There remain several limitations to our work: 1. while the combination of Self Refine with other methods proves effective, this integration is still insufficient, suggesting that the introduction of agentic reinforcement learning may be necessary. 2. MAGA-M exhibits a further and significant increase in lexical diversity for reasons that currently remain unclear, although this trend aligns to some extent with the findings reported in M4 (Wang et al., 2024d). 3. a notable limitation in our experiments is that R-B MAGA yields a marginal improvement in ACC(@FPR=5%) on C-MAGE. Clarifying the underlying cause of this limited gain could potentially advance research on the generalization of machine-generated text detection. 4. regarding multilingual evaluation, our current study only incorporates English and Chinese, leaving the efficacy of our method on other languages undetermined. 5. our focus primarily centers on vanilla MGT, without accounting for specialized text forms such as continued writing or paraphrased text.

References

Aaditya Bhat. 2023. Gpt-wiki-intro.

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, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, and 5 others. 2024. LLM-DetectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343, Miami, Florida, USA. Association for Computational Linguistics.

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 *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)*, pages 6992–7018, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–33.

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

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

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Token prediction as implicit classification to identify LLM-generated text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13112–13120, Singapore. Association for Computational Linguistics.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219, Bangkok, Thailand. Association for Computational Linguistics.

John W Chotlos. 1944. Iv. a statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75.

Wing-Yee Chow, Cybelle Smith, Ellen Lau, and Colin Phillips. 2016. A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5):577–596.

Joseph Cornelius, Oscar Lithgow-Serrano, Sandra Mitrovic, Ljiljana Dolamic, and Fabio Rinaldi. 2024. BUST: Benchmark for the evaluation of detectors of LLM-generated text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8029–8057, Mexico City, Mexico. Association for Computational Linguistics.

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

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Isaac David and Arthur Gervais. 2025. Authormist: Evading ai text detectors with reinforcement learning. *Preprint*, arXiv:2503.08716.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

dirtycomputer. 2023. douban_movie_review.

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

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

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Google DeepMind, Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2024. Introducing gemini 2.0: our new ai model for the agentic era.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633638.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Hans W. A. Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):542–556.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *Preprint*, arXiv:2303.14822.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. 2025. Detree: Detecting human-ai collaborative texts via tree-structured hierarchical representation learning. *ArXiv*, abs/2510.17489.

HongzheBi. 2023. Dureader2.0.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. Are AI-generated text detectors robust to adversarial perturbations? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6005–6024, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Albert Jiang, Alexandre Abou Chahine, Alexandre Sablayrolles, Alexis Tacnet, Alodie Boissonnet, Alok Kothari, Amélie Héliou, Andy Lo, Anna Peronnin, Antoine Meunier, Antoine Roux, Antonin Faure, Aritra Paul, Arthur Darcet, Arthur Mensch, Audrey Herblin-Stoop, Augustin Garreau, Austin Birky, Avinash Sooriyarachchi, and 94 others. 2024. Ministral-8b-instruct-2410.

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

Kerassy. 2025. trustpilot-reviews-123k.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Preprint*, arXiv:1909.05858.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):2125821266.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305.

lars1234. 2025. baidu-baike-dataset.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. CSL: A large-scale Chinese scientific literature dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3917–3923, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaojun Liu, Chuang Zhang, Xiaojun Chen, Yanan Cao, and Jinpeng Li. 2020. Clts: A new chinese long text summarization dataset. In *Natural Language Processing and Chinese Computing*.

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

LZYFirecn. 2025. weibo-hot-search.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Mistral AI. 2025. Mistral medium 3: Medium is the new large.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

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

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Charlotte Nicks, Eric Mitchell, Rafael Rafailov, Archit Sharma, Christopher Manning, Chelsea Finn, and Stefano Ermon. 2023. Language model detectors are easily optimized against. In *Socially Responsible Language Modelling Research*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Saksorn Ruangtanusak, Pittawat Taveekitworachai, and Kunat Pipatanakul. 2025. Talk less, call right: Enhancing role-play llm agents with automatic prompt optimization and role prompting. *Preprint*, arXiv:2509.00482.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023a. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Preprint*, arXiv:2309.11285.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

sentence-transformers. 2021. reddit-title-body.

sentence-transformers. 2024a. natural-questions.

sentence-transformers. 2024b. npr.

sentence-transformers. 2024c. yahoo-answers.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. In *Computational Linguistics and Intellectual Technologies*, page 497511. RSUH.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Namrata Shivagunde, Vladislav Lialin, and Anna Rumshisky. 2023. Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2094–2107, Singapore. Association for Computational Linguistics.

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

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

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

Tencent-Hunyuan. 2025. Hunyuan-7B.

Tencent Hunyuan Team, Ao Liu, Botong Zhou, Can Xu, Chayse Zhou, ChenChen Zhang, Chengcheng Xu, Chenhao Wang, Decheng Wu, Dengpeng Wu, Dian Jiao, Dong Du, Dong Wang, Feng Zhang, Fengzong Lian, Guanghui Xu, Guanwei Zhang, Hai Wang, Haipeng Luo, and 236 others. 2025. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. *Preprint*, arXiv:2505.15431.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, QINGHUA ZHANG, Ruifeng Li, Chao Xu, and Yunhe Wang. 2024. Multiscale positive-unlabeled detection of AI-generated texts. In *The Twelfth International Conference on Learning Representations*.

twn39. 2025. dianping.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2025a. Humanizing the machine: Proxy attacks to mislead LLM detectors. In *The Thirteenth International Conference on Learning Representations*.

Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2025b. CogLM: Tracking cognitive development of large language models. In *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)*, pages 73–87, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

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

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, and 7 others. 2025c. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xuebo Liu, Lidia S. Chao, and Min Zhang. 2025. Who wrote this? the key to zero-shot LLM-generated text detection is GECScore. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10275–10292, Abu Dhabi, UAE. Association for Computational Linguistics.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMDet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, Singapore. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2024. On the generalization of training-based ChatGPT detection methods. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7223–7243, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongzhou Yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2025. Finemedlm-o1: Enhancing medical knowledge reasoning ability of llm from supervised fine-tuning to test-time training. *Preprint*, arXiv:2501.09213.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *Preprint*, arXiv:2304.12008.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Cong Zeng, Shengkun Tang, Xianjun Yang, Yuanzhou Chen, Yiyou Sun, Zhiqiang Xu, Yao Li, Haifeng Chen, Wei Cheng, and Dongkuan Xu. 2024. Dald:

Improving logits-based detector without logits from black-box llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 54947–54973. Curran Associates, Inc.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jingyi Zheng, Tianyi Hu, Yule Liu, Zhen Sun, Zongmin Zhang, Wenhan Dong, Zifan Peng, and Xinlei He. 2025a. Chasm: Unveiling covert advertisements on chinese social media.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025b. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. 2025. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12298–12319, Vienna, Austria. Association for Computational Linguistics.

Contents of Appendix

A Details and Statistics of MAGA Dataset 16

B Domain Details 17

C Generator Details 21

D Alignment Method Details 22
- D.1 Roleplaying 22
- D.2 BPO 24
- D.3 Self Refine 25
- D.4 GDARL 25

E Decoding Strategies 28
- E.1 Temperature 28
- E.2 Top-k and Top-p Sampling 28
- E.3 Penalty 29

F Evaluation Detector Details 29

G Prompt Details 31

H Data Clean Details 32
- H.1 HWT Clean 33
- H.2 MGT Clean 33

I Introduction to External Datasets for Generalization Ability Evaluation 34

J More Experiment Details 34

K Quantitative Text Analysis Details 37
- K.1 Multidimensional Text Analysis . 37
- K.2 PPL Analysis 39

L Training Hyperparameters 41
- L.1 Common Hyperparameters 41
- L.2 GDARL-CDMMR Rounds and Group 42

M Computational Overhead Analysis 42

N Discussion 43
- N.1 Cross-Mutual Reward Component and Existing GDARL Frameworks 43
- N.2 Why not Distill? 45
- N.3 Why Single-Round Self Refine? . 45

O Human Analysis Case Study 45

A Details and Statistics of MAGA Dataset

In the main paper, we primarily focus on the MAGA split, which represents the fully enhanced MAGA split of the English subset. This split is constructed via the MAGA pipeline integrated with all four alignment methods we adopted.

Our generalized MAGA dataset comprises two language-specific subsets, namely the English subset MAGA and the Chinese subset MAGA-cn. Specifically, the English MAGA subset contains six distinct splits: MGB, MAGA, MAGA-extra-GDARL-CDMMR, MAGA-extra-BPO, MAGA-extra-roleplaying, and MAGA-extra-self-refine. The MAGA-cn subset follows an identical structure with these six splits. Each split consists of a training subsplit and a validation subsplit. MGB serves as the baseline split constructed without any alignment methods, while each extra split is constructed using a single corresponding alignment technique.

Table 6 presents the overall distribution of the generalized MAGA dataset. Each split contains 72,000 machine-generated texts and 72,000 human-written texts (60,000 training + 12,000 validation). It should be noted that the 72,000 human texts are reused across all splits and remain identical. The machine-generated texts vary across splits with one exception: in the MAGA-extra-GDARL-CDMMR split, since GDARL-CDMMR only applies to non-commercial large language models and does not integrate other alignment methods, the 36,000 machine texts generated by commercial models are sourced from MGB.

Our MAGA dataset maintains a high degree of balance. Table 7 and Table 8 detail the number of samples generated by each model within each domain. For instance, in a single training subsplit, each model generates 500 samples per domain.

Table 9 provides the detailed annotation schema for our MAGA dataset to ensure comprehensive documentation. Figure 27 presents a HWT sample from MGB_train.jsonl, which also appears across all data splits. Figures 28, 29, 30, 31, 32, 34 illustrate MGT examples from each respective split, all of which correspond directly to the same HWT shown in Figure 27. These examples facilitate a more intuitive understanding of the meaning of our annotations.

Table 10 lists the abbreviations for the models used in this study. The model names appearing in the text and dataset annotations correspond to the

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., 2023)	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., 2023a)	160k	✓	✗	✓	✗	✗	✗
CCD (Wang et al., 2023)	467k	✗	✗	✓	✓	✗	✗
IMDGSP (Mosca et al., 2023)	29k	✗	✓	✗	✗	✗	✗
M4 (Wang et al., 2024d)	122k	✓	✓	✓	✗	✗	✗
HC-Var (Xu et al., 2024)	145k	✓	✗	✗	✗	✗	✗
BUST (Cornelius et al., 2024)	25k	✓	✓	✗	✗	✓	✗
M4GT-Bench (Wang et al., 2024c)	217k	✓	✓	✓	✗	✗	✗
LLM-DetectAIve (Abassy et al., 2024)	303k	✓	✓	✗	✓	✗	✗
MixSet (Zhang et al., 2024)	3.6k	✓	✓	✗	✓	✗	✗
HC3 Plus (Su et al., 2024)	210k	✓	✗	✓	✗	✗	✗
MAGE (Li et al., 2024)	447k	✓	✓	✗	✗	✗	✗
RAID (Dugan et al., 2024)	570k	✓	✓	✗	✓	✓	✗
LAMP (Chakrabarty et al., 2025)	1k	✓	✓	✗	✓	✗	✗
Beemo (Artemova et al., 2025)	19.6k	✓	✓	✗	✓	✗	✗
MultiSocial (Macko et al., 2025)	472k	✓	✓	✓	✗	✗	✗
RealDet (Zhu et al., 2025)	847k	✓	✓	✓	✓	✗	✗
MAGA (Ours)	**936k**	✓	✓	✓	✓	✓	✓

Table 5: A more detailed comparison of the publicly available sources of machine-generated text.

Subset (Language)	Split	Training Subsplit			Validation Subsplit			**Total**		
		Machine	Human	Total	Machine	Human	Total	Machine	Human	Total
MAGA (English)	MGB	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA	60k	60k	120k	12k	12k	24k	72k	72k	144k
	MAGA-extra-GDARL-CDMMR	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	468k
MAGA-cn (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-GDARL-CDMMR	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	468k
Grand Total		660k	120k	780k	132k	24k	156k	792k	144k	**936k**

Table 6: 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-GDARL-CDMMR is derived from MGB.

Model column in Table 10, while the Identifier column specifies the actual model names used during implementation, such as API calls. We employ abbreviations for all commercial models, whereas for non-commercial models, the actual identifiers remain consistent with the annotations.

B Domain Details

We prioritize diversity as a major consideration for domain selection. As discussed in §3.1, we select domains for the English subset to challenge various model capabilities, and we apply a similar principle to the Chinese subset. Specifically, we include ten distinct Chinese domains: social conversations (Baidu Tieba, Weibo Review, Rednote Review), summarization (CSL), general knowledge (Baidu Baike), reviews (Dianping, Douban Review), direct Q&A (Baidu Zhidao, Zhihu), and news creation (CLTS).

To enhance coverage, we prioritize selecting two distinct data sources with certain differences for each capability where possible. For the

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	Ministral-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
Dianping (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
Zhihu (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 7: 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	Ministral-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
Dianping（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
Zhihu（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 8: 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.

Field	**Type**	**Description**
id	`str`	A UUID4 that uniquely identifies the content of the text.
title	`str`	Title from the annotation of the raw HWT data, used in our prompt.
text	`str`	The content of the text.
domain	`str`	The genre or platform of the text (e.g., Reddit).
human_source_id	`str`	A UUID4 referencing the human-written source text.
prompt_id	`str`	A UUID4 uniquely identifying the prompt (system+user).
system_prompt	`str`	The system prompt used for generation.
user_prompt	`str`	The user prompt used for generation.
model	`str`	The LLM that generated the text; labeled as "human" for HWT.
label	`int`	Binary indicator: 0 for HWT and 1 for MGT.
temperature	`float`	The decoding parameter for generation.
top_p	`float`	The decoding parameter for generation.
top_k	`int`	The decoding parameter for generation.
repetition_penalty	`float`	The decoding parameter for generation.

Table 9: Annotation Data Schema of the MAGA Dataset

English datasets, Wikipedia provides rigorous, fact-oriented content, whereas wikiHow offers more informal, instructional text. Regarding reviews, Amazon focuses on specific product transactions and feedback, while Trustpilot Reviews maintains a broader scope by including evaluations of services, brands, and physical stores. In the context of question answering, Natural Questions emphasizes scientific accuracy, while Yahoo Answers exhibits a more community-driven style. Similarly, CC News consists of general textual news, whereas NPR News adopts a

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 10: The generative models used in our project

specialized broadcast style. We note that the selection process for Chinese domains does not strictly adhere to this dual-source strategy due to the relative scarcity of open-source human-written Chinese text datasets.

All our source HWT datasets across selected domains possess a common attribute representing the ground-truth title (even if the specific field naming varies, such as the `Movie_Name` field in the Douban Review domain). This enables the integration of titles into prompt templates to generate the corresponding MGT.

The source HWTs we collected are generally clean, yet minimal further inspection and cleaning are required. We provide a detailed description of our data cleaning procedures in Appendix H.

Furthermore, as discussed in Appendix A, taking the MAGA English subset as an example, this subset is constructed by initially sampling 7,200 human-written texts from the original human text dataset of each domain, resulting in a total of 72,000 human texts across all domains. All subsequent experiments and the construction of every splits rely exclusively on these 72,000 human texts. This sampling procedure exerts no influence on the experimental results, and no further sampling for data filtering occurs in the following experimental stages.

The following provides a detailed introduction to all our selected domains:

Reddit (sentence-transformers, 2021) This dataset consists of posts sourced from Reddit. We chose this dataset due to its highly informal, conversational style and diverse range of topics, which we hypothesize will pose challenges for detecting machine-generated content due to the lack of strict structural constraints. For this domain, we use the post title as the title and the specific content of each post as the target text.

S2ORC (Lo et al., 2020) This Semantic Scholar Open Research Corpus (S2ORC) 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. For this domain, we use the paper title as the title and the abstract as the target text.

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 machine-generated text. For this domain, we use the title of a Wikipedia page as the title and the article content as the target text.

wikiHow (Koupaee and Wang, 2018) This dataset contains the introductory text of wikiHow articles, spanning from daily life skills to complex technical operations. We selected this domain because it demands practical common sense and an inherent procedural logic, as generating the content requires an understanding of the necessary steps and underlying structure of the task. For this domain, we use the title of a wikiHow page as the title and the article content as the target text.

Trustpilot Reviews (Kerassy, 2025) This dataset contains user-written reviews of businesses, products, and services. The reviews exhibit a wide range of sentiment polarity and subjective opinions, and we hypothesize that machine-generated reviews may lack the nuanced personal experiences that characterize authentic user feedback. For this domain, we use the title of a review as the title and the review content as the target text.

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. For this domain, we use the title of a review as the title and the review content as the target text.

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; machine-generated answers may struggle to match the diverse, sometimes idiosyncratic problem-solving approaches of human respondents. For this domain, we use the question as the title and the answer as the target text.

Natural Questions (sentence-transformers, 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. For this domain, we use the question as the title and the answer as the target text.

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. For this domain, we use the title of the news article as the title and the article content as the target text.

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 machine-generated text in mimicking professional news writing. For this domain, we use the title of the news article as the title and the article content as the target text.

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. For this domain, we use the title of a post as the title and a reply to that post as the target text.

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. For this domain, we use the title of a Weibo article as the title and a review of that article as the target text.

Rednote Review (Zheng et al., 2025a) This dataset consists of user-written reviews and lifestyle sharing content from Xiaohongshu (Rednote), a leading Chinese social media platform with a predominantly female user base. Centered on authentic personal experiences, the platform emphasizes highly subjective, first-person narratives and visual cues rooted in daily life. We hypothesize that machine-generated content may struggle to replicate the nuanced, relatable details and the emotional resonance characteristic of these organic community posts. For this domain, we use the title of a Rednote post as the title and a review of the post as the target text.

CSL (Li et al., 2022) This Chinese Scientific Literature (CSL) 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. For this domain, we use the paper title as the title and the abstracts as the target text.

Baidu Baike (lars1234, 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. For this domain, we use the title of a Baidu Baike page as the title and the article content as the target text.

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. For this domain, the title we use consists of the products and services user purchased, while the target text is the review provided by the user for this purchase.

Douban Review (dirtycomputer, 2023) This dataset contains user reviews of movies from Douban, a Chinese cultural community platform. The reviews are characterized by subjective, insightful critical analysis and emotional expression, making it difficult for machine-generated text to replicate the depth of human cultural perception and personal taste. For this domain, we use the movie name as the title and the user review as the target text.

Baidu Zhidao (HongzheBi, 2023) This dataset consists of Q&A pairs from Baidu Zhidao. Characterized by rigorous content moderation and professional verification, the answers provide a source of high-quality, standardized Chinese text. This helps evaluated models distinguish between machine-generated content and highly structured, verified human expertise. For this domain, we use the question as the title and the answer as the target text.

Zhihu (Liu et al., 2020) This dataset contains content from Zhihu, a platform known for its casual and spontaneous user interactions. Unlike professional databases, these answers are often subjective and conversational, reflecting the unconstrained writing styles of general users. Such data is crucial for testing whether detection models can identify the subtle gap between AI outputs and fluid, non-expert human expression. For this domain, we use the question as the title and the answer as the target text.

CLTS (Liu et al., 2020) This Chinese Long Text Summarization (CLTS) dataset comprises news articles from "Pengpai Xinwen" (thepaper.cn), 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. For this domain, we use the title of the news article as the title and the article content as the target text.

C Generator Details

In Table 10 we list the exact generative models used in our project along with their unique identifiers. All open-source models were run using the vLLM library [7] and all closed-source models were run using the proprietary APIs from OpenAI [8]. The following is a detailed list of the generative models used in the project.

GPT-4o-mini (Hurst et al., 2024) is a lightweight variant of GPT-4o developed by OpenAI. We utilize the `gpt-4o-mini-2024-07-18` checkpoint released on July 18, 2024. This model significantly outperforms GPT-3.5 Turbo in reasoning and coding tasks while maintaining high efficiency and low latency, making it the industry standard for cost-effective intelligence.

Gemini-2.0-flash (Google DeepMind et al., 2024) is Google DeepMind's multimodal model released on March 12, 2025. It is designed for high-speed inference across text, image, and audio. It offers state-of-the-art multimodal reasoning with near-instantaneous response times. We queried the model for the MAGA English subset from September 6 to 7, 2025, and for the MAGA-cn Chinese subset from October 13 to 14, 2025.

DeepSeek-V3 (DeepSeek-AI et al., 2025) is a strong decoder-only model developed by DeepSeek AI, released on December 16, 2024. It employs a Mixture-of-Experts (MoE) architecture, excelling in math, coding, and logical reasoning to a level comparable with top-tier proprietary models. We queried the model for the MAGA English subset from September 6 to 7, 2025, and for the MAGA-cn Chinese subset from October 13 to 14, 2025.

Qwen3-plus (Yang et al., 2025) is a large-scale model developed by Alibaba Cloud. We utilize the `qwen-plus-2025-04-28` checkpoint released on April 28, 2025. As a flagship model in the Qwen3 series, it demonstrates superior performance in complex instruction following and multi-turn dialogue compared to its predecessors.

[7] https://docs.vllm.ai/en/stable/

[8] https://platform.openai.com/docs/introduction

Mistral-Medium (Mistral AI, 2025) is developed by Mistral AI and was released on May 7, 2025. It strikes an optimal balance between the efficiency of smaller models and the reasoning power of the Large series, particularly excelling in multilingual understanding. We queried the model for the MAGA English subset from September 6 to 7, 2025, and for the MAGA-cn Chinese subset from October 13 to 14, 2025.

Hunyuan-TurboS (Tencent Hunyuan Team et al., 2025) is Tencent's lightweight variant optimized for low-latency streaming. We use the `hunyuan-turbos-20250313` checkpoint released on March 13, 2025. This model is specifically fine-tuned for high-concurrency scenarios and daily conversational interactions.

Llama-3.1-8B-Instruct (Grattafiori et al., 2024) is an 8B-parameter instruction-tuned model developed by Meta, released in July 2024. It features an expanded 128K context window and significantly improved tool-use capabilities, setting a benchmark for open-source models in its weight class.

gemma-3-12b-it (Gemma Team et al., 2025) is an open-weight instruction-tuned model developed by Google, released in March 2025. Building on the Gemma 3 architecture, it introduces native multimodal support and exhibits robust performance in logical deduction and mathematical problem-solving.

DeepSeek-R1-0528-Qwen3-8B (Guo et al., 2025) is developed by DeepSeek AI and released in May 2025. It integrates DeepSeek's R1 reinforcement learning methodology into the Qwen3-8B base, specifically designed to enhance "Chain-of-Thought" (CoT) reasoning for complex tasks.

Qwen3-8B (Yang et al., 2025) is Alibaba Cloud's 8B-parameter open-source model released in April 2025. It serves as a highly versatile base model with competitive performance across general knowledge and coding benchmarks for small-scale deployment.

Ministral-8B-Instruct-2410 (Jiang et al., 2024) is a specialized "edge" model developed by Mistral AI, released in October 2024. It is optimized for local inference on resource-constrained devices while maintaining high instruction-following accuracy.

Hunyuan-7B-Instruct (Tencent-Hunyuan, 2025) is a 7B-parameter model developed by Tencent, released in July 2025. It is highly optimized for Chinese linguistic nuances and cultural context, making it particularly effective for Chinese content generation tasks.

D Alignment Method Details

D.1 Roleplaying

To implement coarse-grained roleplaying within the MAGA-extra-roleplaying and MAGA splits, we first constructed the MAGA-ROLE-80 prompt set. As discussed in Section 3 and Appendix A, a one-to-one basis exists between MGT and HWT. Taking the MAGA English Subset as an example, we utilize the title fields from 72k HWT data to create 72k original prompts, which serve as the basis for the MGTs in the MGB split. Then each original prompt is randomly assigned one of the 80 roleplaying prompts. Once this assignment is complete, the prompt set remains fixed for all subsequent uses to eliminate sampling randomness and we believe it is helpful for GDARL's learning stability. Details about all prompts including rolepaying prompts are provided in Appendix G.

All previous MGT detection benchmarks listed in Table 5 do not explicitly report the use of system prompt. Based on our reading and analysis and to the best of our knowledge, some additional instructions found in these previous works appear to have been included within the user prompt, rather than being utilized as a system prompt. In our work, roleplaying prompts serve as system prompts to provide high-level instructions and persona settings.

In the design of LLM conversational API, the structured input method that distinguishes the three core "Messages" roles—System, User, and Assistant—is known as Chat Markup Language (ChatML)[9]. ChatML addresses the difficulty models face in distinguishing between instructions and conversational content, a framework originally introduced with GPT-3.5[10]. System prompts are now frequently used to define a model's behavioral guidelines, identity and background, output constraints, or specific knowledge scope,

[9] https://github.com/openai/openai-python/blob/120d225b91a8453e15240a49fb1c6794d8119326/chatml.md

[10] https://openai.com/index/introducing-chatgpt-and-whisper-apis/

effectively allowing the model to maintain a specific tone.

Regarding its mode of input operation, taking the example in Figure 3 as an illustration, when only a user prompt is present, the actual input fed to the LLM after being formatted into the ChatML template is: '<lim_start|>user\nWrite the body of a Reddit post titled "How to really make America great again".<lim_end|>\n'. When a system prompt is added, the actual input fed to the LLM after formatting into the ChatML template becomes: '<lim_start|>system\nYou are American. Please respond in a casual, conversational American English style.<lim_end|>\n<lim_start|>user\nWrite the body of a Reddit post titled "How to really make America great again".<lim_end|>\n'.

Why do we adopt coarse-grained role-playing? What are the characteristics of MAGA-ROLE-80, and why is it effective? The following is a detailed discussion and analysis, grounded in relevant prior work and presented according to a logical theoretical progression:

Coarse-grained and Fine-grained Role-playing Chen et al. (2024) provide a systematic survey of the role-playing landscape in LLMs, categorizing the granularity of role-playing into coarse-grained and fine-grained approaches. Coarse-grained role-playing is typically persona-based, focusing on basic demographic information such as age, gender, location, and brief personality traits, or in even coarser-grained instances, it may be limited solely to identity traits. In contrast, fine-grained role-playing is character-based and involves comprehensive attributes including lineage, talents, belongings, and experiences, as well as complex interpersonal relations, environmental scenes, and temporal personas that evolve over time. The evaluation of role-playing generally encompasses four dimensions: conversational ability, role-persona consistency, role-behavior consistency, and role-playing attractiveness. According to Chen et al. (2024), the primary objective of fine-grained role-playing is to achieve a deep restoration of specific characters while maintaining consistency across these four dimensions. However, relying solely on prompt engineering for fine-grained tasks often proves problematic. Extensive prompts tend to saturate the context window, which leads to a decline in generation quality and coherence. Furthermore, excessively detailed personas can interfere with the native conversational capabilities of the model, resulting in rigid, unnatural, or logically fractured responses. Fine-grained role-playing typically requires dimension expanded high-quality data collection, and fine-tuning to improve dialogue quality. Without fine-tuning, LLMs often exhibit a decrease in fluency and human-like performance when performing fine-grained tasks. For improvements in basic style and human-likeness, coarse-grained role-playing is generally sufficient and yields satisfactory results through optimized prompt engineering. Experiments in RoleLLM (Wang et al., 2024a) report similar findings, demonstrating that coarse-grained role-playing achieves excellent performance using only prompt engineering.

RRP (Rule-based Role Prompting, Ruangtanusak et al. (2025) RRP focuses on enhancing coarse-grained role-playing within the Commonsense Persona-grounded Dialogue Challenge (CPDC) 2025. The core philosophy of RRP is "Talk Less, Call Right". Beyond the specific tool-calling challenges addressed in their work, their experimental results and analysis indicate that optimization in prompt design alone significantly improves the coarse-grained role-playing performance of LLMs. Specifically, prompts that are brief, clearly structured, explicit in intention, and possess a certain rule-based nature facilitate better role-playing outcomes.

ROLE-88 (Ettinger, 2020; Chow et al., 2016) ROLE-88 is originally designed for verb prediction tasks involving argument role reversals in sentences. For example, in the sentences "the restaurant owner forgot which customer the waitress had ___" and "the restaurant owner forgot which waitress the customer had ___"—where the role nouns *customer* and *waitress* have been swapped—predict the verb. Setting aside the specific tasks associated with ROLE-88, let us focus solely on the roles themselves (excluding the very few non-human entities, such as birds or dragons). Generally speaking, ROLE-88 primarily encompasses a curated selection of the most typical, high-frequency professional roles comparatively speaking encountered in everyday social life roles that, to a significant extent, serve as a representative sample of daily social interactions. Moreover, these roles are sufficiently easy to comprehend and well-suited for role-playing.

MAGA-ROLE-80 Consequently, MAGA adopts a coarse-grained role-playing approach. As suggested by the findings in Chen et al. (2024)'s survey, RoleLLM (Wang et al., 2024a), and RRP (Ruangtanusak et al., 2025), coarse-grained role-playing is the appropriate choice for our objective of human-like alignment in our task, offering enough satisfactory performance via prompt engineering.

To ensure social coverage and diversity, we select the majority of roles from ROLE-88 (Ettinger, 2020; Chow et al., 2016). Since ROLE-88 is not directly applicable as a set of conversational prompts, we construct the MAGA-ROLE-80 prompt set based on these roles. Our prompt design incorporates the core ideas of RRP's "Talk Less, Call Right" principle (Ruangtanusak et al., 2025), ensuring that the prompts are concise, structured, and easy to inhabit. Table 29 provides ten examples of these prompts. These prompts possess a simple, uniform rule-based nature, such as "You are American. Please respond in a casual, conversational American English style.", or "You are a teacher. Respond in a patient, firm, and caring tone.". Each prompt basically consists of two sentences: the first defines the basic identity, and the second specifies the speaking style, which aligns with the requirements of our task.

The MAGA-ROLE-80 set contains 80 prompts each for English and Chinese. While primarily professional roles, we include a few well-known figures like Sherlock Holmes to enhance diversity. The English and Chinese versions are largely translations of each other, with minor cultural adaptations, such as mapping an American role in English to a Beijinger role in Chinese. All prompts are manually authored.

Similar to ROLE-88, a set of 80 roles is considered sufficient for broad representation. As detailed in Appendix A, each LLM generates 7200 samples per split (6000 for training and 1200 for validation). The square root of this total is approximately 80, which ensures a balance between the number of roles and the amount of text generated per role. We believe this balance is conducive to the learning process of GDARL.

Additionally, it should also be noted that, our MAGA-ROLE-80 prompt set actually contains 79 specific roles plus one null string, representing a default or role-less role. We believe this inclusion benefits the overall quality of the dataset and the stability of GDARL training.

D.2 BPO

The Black-box Prompt Optimization (BPO) (Cheng et al., 2024) method is an automated alignment model that optimizes user instructions to bridge the understanding gap between humans and large language models. The underlying principle involves constructing a preference dataset where an initial prompt is paired with both preferred and rejected responses. By utilizing a powerful teacher model to reverse engineer optimized instructions that elicit the preferred responses, the authors train a sequence to sequence model to map original user inputs to refined versions. A defining characteristic of the BPO model is its operational tendency to preserve the core intent of the user while appending a descriptive suffix that clarifies the task requirements. For example, a simple instruction such as "Tell me about Harry Potter" is transformed into an expanded version such as "Tell me about Harry Potter, providing a comprehensive summary of the book series including its main characters, central plot, and major themes."

In this work, we directly employ the BPO model released by Cheng et al. (2024)[11]. To ensure the reliability and consistency of the optimized instructions, we specifically utilize the stable mode of the model. The execution process involves wrapping the user prompt into a specialized template: "[INST] You are an expert prompt engineer. Please help me improve this prompt to get a more helpful and harmless response: user_prompt [/INST]". During the inference stage, the model generates a suffix to optimize the original prompt using a temperature of 0.6, a top_p value of 0.9, and a max_new_tokens of 1024. This configuration prioritizes stable and grounded refinements over the more stochastic outputs of the aggressive mode, thereby providing a robust foundation for the subsequent generation tasks of the target model.

To the best of our knowledge, all previous machine-generated text (MGT) detection benchmarks listed in Table 5 exclusively utilize static, human-designed prompts. The primary advantage of BPO lies in its capacity to ensure that each individual prompt is unique, substantially enhancing prompt diversity.

It is essential to note that when incorporating a title into the original prompt, human designers cannot anticipate the specific content of each title in

[11] https://huggingface.co/zai-org/BPO

advance, or more precisely, it is simply impossible for human to tailor the prompt individually for every conceivable title. BPO addresses this limitation by providing specific optimization suggestions for every unique title. Additionally, it should also be noted that, role-playing do not encounter such unpredictable elements as titles and possess inherent specificity, thereby reducing and therefore do not necessitate such highly automated prompt optimization.

Additionally, the BPO model Cheng et al. (2024) released only supports English inputs and outputs, lacking native support for Chinese. To address this limitation, we employ the Hunyuan-MT-7B (Zheng et al., 2025b) translation model. Our selection of Hunyuan-MT-7B is motivated by two primary factors. First, it demonstrates exceptional translation quality, as Zheng et al. (2025b) report that its performance is comparable to powerful commercial models such as GPT-4.1 and DeepSeek-V3-0324. Second, as a specialized translation model, it directly produces translation outputs without generating redundant content, and eliminates the need for complex prompt engineering or post-processing controls. Specifically, we first translate the original Chinese prompts into English using Hunyuan-MT-7B and then process them through the BPO model for optimization. The resulting English suffixes generated by BPO are translated back into Chinese via Hunyuan-MT-7B. Finally, these optimized suffixes are appended to the original prompts to construct the BPO-optimized Chinese prompts.

D.3 Self Refine

The Self Refine (Madaan et al., 2023) framework enables the autonomous enhancement of large language model outputs through a recursive process involving critique and revision. The procedure begins with the generation of an initial response y_0 for a given input x. At each iteration t, the model performs self-evaluation via in-context learning (ICL) to produce multi-aspect feedback that identifies necessary corrections. Based on this feedback and the preceding output y_t, the model constructs a refined response y_{t+1}. This sequence continues iteratively, where the model employs in-context learning to determine whether the current output satisfies the predefined objectives. The loop terminates when the self-judgment indicates that the goal is reached or when the process hits a maximum number of allowed iterations. Since the framework utilizes a single model to function as the generator, feedback provider, and refiner, it avoids the need for supervised fine-tuning or reinforcement learning.

In our implementation for machine-generated text detection, we adhere to the original algorithmic structure while introducing a specific modification. While large language models are reliable as general evaluators for self-critique, their performance in the explicit detection of machine-generated patterns remains a challenging task that can exhibit inconsistency across multiple rounds. To ensure the stability of the detection results, we restrict the process to a single refinement round instead of multiple iterations. This approach leverages the structural benefits of feedback-informed refinement while mitigating the potential for error accumulation or instability in multi-round detection. We provide a more comprehensive discussion and experimental analysis regarding the choice of iteration rounds in Appendix N.3. The Self-Refine prompts for English and Chinese are shown in Table 27 and Table 28, respectively, which are described in Appendix G.

D.4 GDARL

It is necessary to emphasize that Generator-Detector Adversarial Reinforcement Learning (GDARL) is the general term for existing GDARL algorithms such as RADAR (Hu et al., 2023) and HUMPA (Wang et al., 2025a) in the field of MGT detection, and it represents the vanilla formulation of such approaches rather than a newly proposed innovative method.

However, We observe that the vanilla GDARL suffers from reward overfitting, which leads to negligible updates of the trained LLM. Based on the HC-Var theory, we further propose the Cross-Mutual Reward (CMR) component and present its three basic application, namely GDARL-CDMR, GDARL-CMMR and GDARL-CDMMR.

The CMR component operates as an external process attached to the ARL paradigm and does not modify the internal algorithmic framework of GDARL. Therefore, it does not conflict with algorithms such as RADAR or HUMPA and is able to be used in conjunction with them.

This section formally defines the algorithmic procedures of GDARL, GDARL-CDMR, GDARL-CMMR and GDARL-CDMMR. Detailed discussion about the CMR component and existing ARL frameworks is provided in Appendix N.1.

First, we predefine the following notations and variables:

Q: The set of input prompts.

t: The current round of adversarial learning, where $t \geq 0$ is a non-negative integer and $t = 0$ denotes the initial model state with pre-trained weights.

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

$\mathcal{D}_\phi^t$: The RoBERTa reward detector at round t with parameters ϕ^t that serves as GRPO's reward model (RM) outputs the probability $\mathcal{D}_\phi^t(x) \in [0, 1]$ that text x is machine-generated.

H: The set of human-written text (HWT), labeled as 0.

M^t: The set of machine-generated text (MGT) at round t, labeled as 1. It is composed of one single corresponding text generated by $\mathcal{M}_\theta^t$ for each individual prompt in Q on a one-to-one basis.

η: The learning rate for updates of the LLM parameters θ.

α: The learning rate for updates of the detector parameters ϕ.

GDARL First, feed each prompt in Q into $\mathcal{M}_\theta^t$; the model generates one MGT for each prompt on a one-to-one basis to form the MGT set $\mathcal{M}^t$. The set $\mathcal{M}^t$ is then used to train the detector $\mathcal{D}^t$ and yield the updated detector $\mathcal{D}^{t+1}$:

$$\mathcal{J}_\mathcal{D}(\phi^t, t) = -\mathbb{E}_{x_m^t \sim M^t}\left[\log \mathcal{D}_\phi^t(x_m^t)\right] - \mathbb{E}_{x_h \sim H}\left[\log\left(1 - \mathcal{D}_\phi^t(x_h)\right)\right] \tag{3}$$

$$\phi^{t+1} = \phi^t - \alpha \nabla_\phi \mathcal{J}_\mathcal{D}(\phi^t, t) \tag{4}$$

The trained detector $\mathcal{D}_\phi^{t+1}$ subsequently functions as the reward model to assign human-like scores. We employ Group Relative Policy Optimization (GRPO) to train the LLM, thereby achieving adversarial RL (ARL). Specifically, for each question q, GRPO samples a group of outputs $O = \{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}^t$ and optimizes the policy model π_θ^t by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta^t, t) = \mathbb{E}\left[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}^t(O|q)\right] \frac{1}{G}\sum_{i=1}^{G}\Bigg(\min\left(\frac{\pi_\theta^t(o_i|q)}{\pi_{\theta_{old}}^t(o_i|q)} A_i^t, \text{clip}\left(\frac{\pi_\theta^t(o_i|q)}{\pi_{\theta_{old}}^t(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right) A_i^t\right) - \beta \mathbb{D}_{KL}(\pi_\theta^t || \pi_{ref}^t)\Bigg), \tag{5}$$

$$\mathbb{D}_{KL}(\pi_\theta^t || \pi_{ref}^t) = \frac{\pi_{ref}^t(o_i|q)}{\pi_\theta^t(o_i|q)} - \log\frac{\pi_{ref}^t(o_i|q)}{\pi_\theta^t(o_i|q)} - 1, \tag{6}$$

where ε, β, and G denote the PPO clipping threshold, the KL regularization weight, and the number of responses generated per question for group-wise averaging, respectively. A_i^t is the advantage, computed using a group of rewards $\{r_1^t, r_2^t, \ldots, r_G^t\}$ corresponding to the outputs within each group:

$$A_i^t = \frac{r_i^t - \text{mean}(\{r_1^t, r_2^t, \ldots, r_G^t\})}{\text{std}(\{r_1^t, r_2^t, \ldots, r_G^t\})}, \tag{7}$$

$$r^t(x, \phi^{t+1}) = 1 - \mathcal{D}_\phi^{t+1}(x). \tag{8}$$

Then we can obtain the LLM $\mathcal{M}_\theta^{t+1}$:

$$\theta^{t+1} = \theta^t + \eta \nabla_\theta \mathcal{J}_{GRPO}(\theta^t, t) \tag{9}$$

GDARL-CDMR GDARL-CDMR (Cross Domain Mutual Reward) divides all texts into two groups: Domains DA and Domains DB. Each group has its own separately trained reward detectors $\mathcal{D}_\phi^{DA,t}$ and $\mathcal{D}_\phi^{DB,t}$. During the training of the larger model, Cross-Mutual Rewards are applied: $\mathcal{D}_\phi^{DB,t}$ scores text in DA, and $\mathcal{D}_\phi^{DA,t}$ scores text in DB. Specifically:

We divide H into H^{DA} and H^{DB}, with corresponding Q into Q^{DA} and Q^{DB}. The MGT set generated using $\mathcal{M}_\theta^t$ is also divided into $M^{DA,t}$ and $M^{DB,t}$, respectively. First, the detectors are trained as follows:

$$\mathcal{J}_\mathcal{D}(\phi^{DA,t}, t) = -\mathbb{E}_{x_m^{DA,t} \sim M^{DA,t}}\left[\log \mathcal{D}_\phi^{DA,t}(x_m^{DA,t})\right] - \mathbb{E}_{x_h^{DA} \sim H^{DA}}\left[\log\left(1 - \mathcal{D}_\phi^{DA,t}(x_h^{DA})\right)\right] \tag{10}$$

$$\phi^{DA,t+1} = \phi^{DA,t} - \alpha \nabla_\phi \mathcal{J}_\mathcal{D}(\phi^{DA,t}, t) \tag{11}$$

$$\mathcal{J}_{\mathcal{D}}(\phi^{DB,t}, t) = -\mathbb{E}_{x_m^{DB,t} \sim M^{DB,t}} \left[\log \mathcal{D}_\phi^{DB,t}(x_m^{DB,t})\right] - \mathbb{E}_{x_h^{DB} \sim H^{DB}} \left[\log \left(1 - \mathcal{D}_\phi^{DB,t}(x_h^{DB})\right)\right] \quad (12)$$

$$\phi^{DB,t+1} = \phi^{DB,t} - \alpha \nabla_\phi \mathcal{J}_{\mathcal{D}}(\phi^{DB,t}, t) \quad (13)$$

Then the trained detector $\mathcal{D}_\phi^{DA,t+1}, \mathcal{D}_\phi^{DB,t+1}$ is used as the reward model cross-scoring for the text x^{DB}, x^{DA} on Domains DB and Domains DA. LLM is trained using GRPO, with the following reward:

$$r^t = \begin{cases} 1 - \mathcal{D}_\phi^{DB,t+1}(x^{DA}), & x^{DA} \sim \mathcal{M}_\theta^t(o|q^{DA}),\ q^{DA} \in Q^{DA} \\ 1 - \mathcal{D}_\phi^{DA,t+1}(x^{DB}), & x^{DB} \sim \mathcal{M}_\theta^t(o|q^{DB}),\ q^{DB} \in Q^{DB} \end{cases} \quad (14)$$

Then we can obtain the LLM $\mathcal{M}_\theta^{t+1}$:

$$\theta^{t+1} = \theta^t + \eta \nabla_\theta \mathcal{J}_{GRPO}(\theta^t, t) \quad (15)$$

GDARL-CMMR In our task, in practice, we adopt multiple LLMs and split the entire generation work, with each LLM responsible for a separate part. GDARL-CMMR (Cross Model Mutual Reward) directly leverages this pipeline.

We divide all texts into two groups corresponding to LLM Models MA and LLM Models MB. A detector $\mathcal{D}_\phi^{MA,t}$ and a detector $\mathcal{D}_\phi^{MB,t}$ are trained independently for each group respectively. During the training of LLMs, cross-model mutual reward is implemented: $\mathcal{D}_\phi^{MB,t}$ scores the texts generated by MA, and $\mathcal{D}_\phi^{MA,t}$ scores the texts generated by MB. It is worth noting that the pipeline also involves texts generated by commercial LLMs, which is simply denoted as MGT-B in Figure 4. This part is also grouped in our work, as described in Appendix L.2. For ease of understanding and to avoid overly cumbersome formula expressions, we omit the description of this component in the formulas herein. Specifically:

We partition H into H^{MA} and H^{MB}, with corresponding prompts Q^{MA} and Q^{MB}. These prompts are then assigned to $\mathcal{M}_\theta^{MA,t}$ and $\mathcal{M}_\theta^{MB,t}$, respectively, for generation; the resulting sets of generated MGTs are $M^{MA,t}$ and $M^{MB,t}$. First, we train the detectors:

$$\mathcal{J}_{\mathcal{D}}(\phi^{MA,t}, t) = -\mathbb{E}_{x_m^{MA,t} \sim M^{MA,t}} \left[\log \mathcal{D}_\phi^{MA,t}(x_m^{MA,t})\right] - \mathbb{E}_{x_h^{MA} \sim H^{MA}} \left[\log \left(1 - \mathcal{D}_\phi^{MA,t}(x_h^{MA})\right)\right] \quad (16)$$

$$\phi^{MA,t+1} = \phi^{MA,t} - \alpha \nabla_\phi \mathcal{J}_{\mathcal{D}}(\phi^{MA,t}, t) \quad (17)$$

$$\mathcal{J}_{\mathcal{D}}(\phi^{MB,t}, t) = -\mathbb{E}_{x_m^{MB,t} \sim M^{MB,t}} \left[\log \mathcal{D}_\phi^{MB,t}(x_m^{MB,t})\right] - \mathbb{E}_{x_h^{MB} \sim H^{MB}} \left[\log \left(1 - \mathcal{D}_\phi^{MB,t}(x_h^{MB})\right)\right] \quad (18)$$

$$\phi^{MB,t+1} = \phi^{MB,t} - \alpha \nabla_\phi \mathcal{J}_{\mathcal{D}}(\phi^{MB,t}, t) \quad (19)$$

Subsequently, the trained detectors $\mathcal{D}_\phi^{MB,t+1}$ and $\mathcal{D}_\phi^{MA,t+1}$ serve as reward models to cross-score Models $\mathcal{M}_\theta^{MA,t}$ with parameters $\theta^{MA,t}$ and Models $\mathcal{M}_\theta^{MB,t}$ with parameters $\theta^{MB,t}$, respectively. The large models are then trained via GRPO, with the rewards defined as follows:

$$r^{MA,t}(x, \phi^{MB,t+1}) = 1 - \mathcal{D}_\phi^{MB,t+1}(x) \quad (20)$$

$$r^{MB,t}(x, \phi^{MA,t+1}) = 1 - \mathcal{D}_\phi^{MA,t+1}(x) \quad (21)$$

Here, $r^{MA,t}$ is utilized for the GRPO training of Models $\mathcal{M}_\theta^{MA,t}$, while $r^{MB,t}$ is utilized for the GRPO training of Models $\mathcal{M}_\theta^{MB,t}$, yielding the following updates:

$$\theta^{MA,t+1} = \theta^{MA,t} + \eta \nabla_\theta \mathcal{J}_{GRPO}(\theta^{MA,t}, t) \quad (22)$$

$$\theta^{MB,t+1} = \theta^{MB,t} + \eta \nabla_\theta \mathcal{J}_{GRPO}(\theta^{MB,t}, t) \quad (23)$$

GDARL-CDMMR The final form of GDARL-CDMMR (Cross Domain & Model Mutual Reward) is constructed by integrating GDARL-CDMR and GDARL-CMMR. More generally, we can simply designate the two groups as GA and GB. In group GA, the LLMs responsible for generation are MA, and the domains of the generated texts are DA; in group GB, the LLMs responsible for generation are MB, and the domains of the generated texts are DB. The

specific grouping scheme utilized by MAGA is detailed in Appendix .

We partition H into H^{GA} and H^{GB}, with the corresponding prompts being Q^{GA} and Q^{GB}. These prompts are then assigned to $\mathcal{M}_\theta^{GA,t}$ and $\mathcal{M}_\theta^{GB,t}$, respectively, for text generation; the corresponding sets of MGTs produced are $M^{GA,t}$ and $M^{GB,t}$. We first train the detectors:

$$\mathcal{J}_\mathcal{D}(\phi^{GA,t}, t) = -\mathbb{E}_{x_m^{GA,t} \sim M^{GA,t}} \left[\log \mathcal{D}_\phi^{GA,t}(x_m^{GA,t})\right] - \mathbb{E}_{x_h^{GA} \sim H^{GA}} \left[\log \left(1 - \mathcal{D}_\phi^{GA,t}(x_h^{GA})\right)\right] \quad (24)$$

$$\phi^{GA,t+1} = \phi^{GA,t} - \alpha \nabla_\phi \mathcal{J}_\mathcal{D}(\phi^{GA,t}, t) \quad (25)$$

$$\mathcal{J}_\mathcal{D}(\phi^{GB,t}, t) = -\mathbb{E}_{x_m^{GB,t} \sim M^{GB,t}} \left[\log \mathcal{D}_\phi^{GB,t}(x_m^{GB,t})\right] - \mathbb{E}_{x_h^{GB} \sim H^{GB}} \left[\log \left(1 - \mathcal{D}_\phi^{GB,t}(x_h^{GB})\right)\right] \quad (26)$$

$$\phi^{GB,t+1} = \phi^{GB,t} - \alpha \nabla_\phi \mathcal{J}_\mathcal{D}(\phi^{GB,t}, t) \quad (27)$$

Subsequently, the trained detectors $\mathcal{D}_\phi^{GB,t+1}$ and $\mathcal{D}_\phi^{GA,t+1}$ serve as reward models to cross-score Models $\mathcal{M}_\theta^{GA,t}$ with parameters $\theta^{GA,t}$ and Models $\mathcal{M}_\theta^{GB,t}$ with parameters $\theta^{GB,t}$, respectively. The large models are then trained via GRPO, with the rewards defined as follows:

$$r^{GA,t}(x, \phi^{GB,t+1}) = 1 - \mathcal{D}_\phi^{GB,t+1}(x) \quad (28)$$

$$r^{GB,t}(x, \phi^{GA,t+1}) = 1 - \mathcal{D}_\phi^{GA,t+1}(x) \quad (29)$$

Here, $r^{GA,t}$ is utilized for the GRPO training of Models $\mathcal{M}_\theta^{GA,t}$, while $r^{GB,t}$ is utilized for the GRPO training of Models $\mathcal{M}_\theta^{GB,t}$, yielding the following updates:

$$\theta^{GA,t+1} = \theta^{GA,t} + \eta \nabla_\theta \mathcal{J}_{GRPO}(\theta^{GA,t}, t) \quad (30)$$

$$\theta^{GB,t+1} = \theta^{GB,t} + \eta \nabla_\theta \mathcal{J}_{GRPO}(\theta^{GB,t}, t) \quad (31)$$

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 employed various decoding strategies, as detailed in Table 11. For open-source models, the official recommended decoding parameters are typically provided in the `generation_config.json` file to ensure optimal or near-optimal generation quality. For commercial models, such recommendations are usually disclosed in their Technical Reports; for instance, the recommended parameters for Qwen3-Plus are specified in the Qwen3 Technical Report (Yang et al., 2025). In cases where specific recommended parameters are unavailable, we follow empirical sub-optimal or near-sub-optimal practices by adopting the default configurations of vLLM (e.g., setting top_k to -1 and keeping other parameters at 1), which generally yield diverse and 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.

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 11: 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.

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 Evaluation Detector Details

In this section we provide a detailed description of all evaluation detectors used in the evaluations of the MAGA dataset. Table 12 presents the default thresholds for these detectors, which are established by their respective original authors.

R-B GPT2 & R-L GPT2 (Solaiman et al., 2019) These two detectors[12] are RoBERTa-base and RoBERTa-large models respectively, fine-tuned on GPT2 open-domain outputs, which were generated via greedy decoding, top-k=50 sampling, and full random sampling. As a long-standing baseline, both base and large variants are used for comparison.

[12]We download the model hosted by OpenAI from the following links https://openaipublic.azureedge.net/gpt-2/detector-models/v1/detector-base.pt and https://openaipublic.azureedge.net/gpt-2/detector-models/v1/detector-large.pt

Detector	Default Threshold
R-B GPT2	0.5
R-L GPT2	0.5
R-B CGPT	0.5
RADAR	0.5
SCRN	0.5
DETree	0.03
Fast DetectGPT	0.5
LLMDet	0.5
Binoculars	0.901531074927684
DALD	0.5
GECSCORE	0.9243697428995128
R-B MPU	0.5

Table 12: Default thresholds of evaluation detectors. It should be noted that Binoculars Score is a human-like score, where texts exceeding the threshold are classified as HWT. In contrast, others are machine-like scores (or specifically, RoBERTa-based detector consists of two-dimensional probability scores, the dimension corresponding to the machine-like likelihood can be selected).

R-B CGPT (Guo et al., 2023) This detector is a RoBERTa-base model fine-tuned on the HC3 dataset (roughly 27,000 human-ChatGPT answer pairs across multiple domains).

RADAR (Hu et al., 2023) This detector is a RoBERTa-based model trained via adversarial reinforcement learning between a RoBERTa classifier and a Vicuna-7B paraphraser. This framework enables the detector to better distinguish between human text, machine-generated text, and paraphrased machine outputs.

GLTR (Gehrmann et al., 2019) This detector was originally an auxiliary interface for human detection, now a standard robustness baseline. It evaluates token likelihood via an LM, bins tokens by likelihood, and uses these as features. We adopt its default settings (rank=10 cutoff, GPT2 small) for detection.

FastDetectGPT (Bao et al., 2024) This detector is an optimized DetectGPT (Mitchell et al., 2023) variant that achieves 340x faster inference without any reduction in accuracy. We use their default settings: GPT-Neo-2.7B as the scoring model and GPT-J-7B as the reference model. We felt that this was a reasonable choice, as neither model was used for generating in our dataset.

Binoculars (Hans et al., 2024) This detector uses the ratio of perplexity to cross-entropy between two similar language models as the detection metric. Adopts official default Falcon-7B and Falcon-7B-Instruct models; neither was used for dataset generation, ensuring fairness.

LLMDet (Wu et al., 2023) This detector computes proxy-perplexity (n-gram sampling-based approximation) of input text via 10 small language models, using these as detection features. None of the language models were used for dataset generation, ensuring fair comparison.

SCRN (Huang et al., 2024) This detector is a robust model with a reconstruction network for denoising and siamese calibration. It resists adversarial perturbations across in-domain, cross-domain, and mixed-source scenarios.

DETree (He et al., 2025) This detector is a representation learning-based model that models text category relationships via a Hierarchical Affinity Tree and contrastive loss. Trained on RealBench, it excels at hybrid text detection and out-of-distribution generalization. RealBench aggregates several existing datasets, including MAGE (Li et al., 2024), M4 (Wang et al., 2024d), M4GT-Bench (Wang et al., 2024c), TuringBench (Uchendu et al., 2021), OUTFOX (Koike et al., 2024), and RAID (Dugan et al., 2024), and undergoes data augmentation by simulating various realistic collaborative writing paradigms such as rewriting, expansion, polishing, and translation. DETree represents the text in RealBench as vectors and stores them in a Faiss vector database. During the specific detection phase, DETree transforms the input text into a vector representation and retrieves k nearest neighbors via the k-nearest neighbors algorithm, which subsequently yields the probability of the text being human-written or AI-generated through a probability-weighted voting mechanism. This underlying intuition is conceptually similar to Retrieval-Augmented Generation (RAG, Lewis et al. (2020). We set the parameter k to their default value of 10, and the reference vectors constructed from RealBench are sourced from DETree's official Hugging Face repository [13].

DALD (Zeng et al., 2024) This distribution-aligned detector mitigates the distribution

[13] https://huggingface.co/datasets/heyongxin233/RealBench

mismatch in black-box scenarios by fine-tuning a surrogate model with limited target samples. It serves as a flexible enhancement to boost zero-shot detectors against model iterations. In this work, we adopt its default and preferred configuration, which operates on top of Fast-DetectGPT (Bao et al., 2024) for conditional probability curvature estimation.

GECSCORE (Wu et al., 2025) This detector is a zero-shot model using grammar error correction similarity. It distinguishes texts by comparing input with corrected versions, boasting strong generalization and paraphrase attack resistance.

RoBERTa-MPU-zhv3 (Tian et al., 2024) This detector is a multiscale model with length-sensitive PU loss and text multiscaling. It improves short-text detection while maintaining long-text performance. We select its Chinese version as an additional evaluation detector, for MAGA-cn evaluation.

It should be noted that the base language models employed by nearly all detectors are primarily competent only in English. In some cases, these base language models lack any Chinese language capability whatsoever; in others, their Chinese recognition ability is weak like low-resource languages, with no practical proficiency in Chinese. Among all our selected evaluation detectors, 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 its base language models Falcon-7B and Falcon-7B-Instruct lack functional Chinese language proficiency, we found that its detection performance is surprisingly satisfactory. However, we did not individually verify whether other metric-based detectors still retain a certain level of capability in detecting Chinese text. As a supplementary addition for training-based detectors, we have additionally incorporated RoBERTa-MPU-zhv3 (Tian et al., 2024). The four aforementioned models are utilized by us for Chinese detection tasks. Then the number of neural-based and metric-based detectors used in our MAGA-cn evaluation is consistent.

G Prompt Details

Original Prompt and BPO-Optimized Prompt Table 25 and Table 26 present the English and Chinese versions of the orginal prompts respectively. These original prompts serve as the user prompts to construct the MGB, MAGA-extra-GDARL-CDMMR, and MAGA-extra-roleplaying splits (roleplaying has additional system prompts). Also, these original prompts are utilized to generate the initial MGT for the MAGA-extra-self-refine split before the feedback and refinement stage. BPO automatically optimizes these original prompts and the MAGA-extra-BPO split adopts the BPO-optimized prompts as user prompts (Figures 28 and 29 serve as a concrete example.). Similarly, the MAGA split (integrates all our alignment methods) also utilizes BPO-optimized prompts as user prompts for generating the initial text prior to get feedback and refinement.

MAGA Additional Instruction The MAGA Additional Instruction, as detailed in Table 25 and Table 26, functions as a suffix to the original prompt. This design is inspired by RAID (Dugan et al., 2024); for instance, the additional instruction for the Reddit domain, "Do not repeat the title.", is also utilized within RAID's Reddit prompts. We observe that LLMs tend to include introductory remarks during text generation, often incorporating the title in unpredictable ways. Such patterns are absent in human-written texts and are nearly impossible to remove by rule-based cleaning. Despite repeated attempts to provide explicit instructions forbidding prologues or title repetitions, models often fail to comply. We find that as prompt complexity and length increase, model adherence tends to deteriorate. Conversely, concise instructions such as Do not repeat the title" suggested by RAID (Dugan et al., 2024) result in a lower probability of the model outputting either the title or unnecessary introductory phrases. Furthermore, as illustrated in Figures 28 and 29, the prompts optimized by BPO also incorporate the MAGA Additional Instruction. To maintain clarity regarding the core BPO mechanism, this detail is omitted in Figure 3. Specifically, the implementation involves a two-step process. The Original User Prompt in Table 25 and Table 26, which is used directly in the MGB split, already contains the MAGA Additional Instruction. For the BPO optimization, we first remove this instruction to facilitate the optimization process. Once BPO generates the refined prompt, the MAGA Additional Instruction is re-appended as a suffix.

Self-Refine Prompt Table 27 and Table 28 illustrate the Self-Refine prompts for English and Chinese contexts. For the initial raw text generated before refinement, a Feedback Prompt is first introduced to produce specific feedback. This feedback is subsequently embedded into the Refine Prompt to generate the final self-refined text. The self-refinement prompts are applied within both the MAGA-extra-self-refine and MAGA splits. To facilitate a clearer understanding of the operational process, specifically we illustrate the procedure using the ChatML template. The large language model first receives the input sequence '<|im_start|>user\n{user_prompt}<|im_end|>' to generate the initial text. Subsequently, we employ a multi-turn dialogue approach where the input '<|im_start|>user\n{user_prompt}<|im_end|>\n <|im_start|>assistant\n{initial_text}<|im_end|>\n <|im_start|>user\n{feedback_prompt}<|im_end|>' yields the feedback response. The process then reverts to the second dialogue turn, where the obtained feedback is embedded into the Refine Prompt to produce the polished output. Specifically, the model processes the sequence '<|im_start|>user\n{user_prompt}<|im_end|>\n <|im_start|>assistant\n{initial_text}<|im_end|>\n <|im_start|>user\n{refine_prompt}<|im_end|>' to generate the final refined text. The two-turn dialogues can be compressed into a single turn using template. Two-turn may be clearer to LLM. Actually, this choice is not important, merely a minor implementation detail worth mentioning. Figures 32, 33, 34, 35 provide specific examples of the outputs.

Roleplaying Prompt Table 29 displays samples of our role-playing prompts, which function as system prompts for the MAGA-extra-roleplaying and MAGA splits. We provide a detailed introduction to our role-playing prompts' characteristics and usage methods in Appendix D.1.

It should be noted that, as we have repeatedly mentioned across various sections, such as Appendices D.1 and D.4, that our prompt set is pre-constructed; in this section, we provide a systematic explanation of this aspect. Taking the MAGA English Subset as an example, all splits contain the same 72k Human-Written Text (HWT) samples (see Table 7), each associated with a title annotation. We consistently utilize these titles to construct the prompt sets. Specifically, embedding the 72k titles into the original prompt template yields the original prompt set for the MGB split. Each original prompt generates one MGT, and the combination of these 72k MGTs with their corresponding 72k HWTs constitutes the MGB split. Similarly, for the MAGA-extra-roleplaying split, each of the 72k original prompts is randomly assigned a roleplaying system prompt according to the procedure in Appendix D.1. In the MAGA-extra-BPO split, the 72k original prompts are automatically optimized by BPO model and the 72k optimized prompts are served as the generation prompts. For the MAGA-extra-GDARL-CDMMR split, we fix the set of 72k prompts to generate MGT for training the RoBERTa detector and to serve as the prompt set for reinforcement learning as detailed in Appendix D.4. The MAGA split represents a synthesis of these methodologies, where all 72k prompts undergo BPO optimization and include roleplaying system prompts before executing the GDARL-CDMMR algorithm to train the model, followed by self-refinement to enhance the inference process. Consequently, we emphasize that our prompt sets are pre-constructed and remain invariant, maintaining a strict one-to-one correspondence where each HWT sample relates to exactly one prompt and one generated MGT.

H Data Clean Details

It is essential to clarify that we perform identical data cleaning procedures across all texts in the entire MAGA dataset. Consequently, the cleaning process is orthogonal to our comparative experiments and does not influence the core conclusions. For instance, the primary experiments in the main paper compare the attack and generalization performance between the MGB and MAGA splits before and after human-like alignment. Since both splits undergo the same cleaning pipeline, the results remain consistent. Our cleaning efforts are solely intended to ensure the overall quality of the dataset.

The source HWTs we collected are generally clean, as they were pre-processed by their respective publishers, particularly for the English data. Thus, only minimal further inspection and cleaning are required. For machine-generated texts (MGTs), inspired by M4 (Wang et al., 2024d), we observe that simple artifacts such as multiple newlines and bullet points occasionally persist.

These patterns may allow detectors to exploit superficial shortcuts rather than learning robust features for generalization.

We checked the text quality by randomly sampling 10 to 20 cases for each domain and generator. Our key observations and subsequent processing steps are detailed below.

H.1 HWT Clean

Based on the sample inspection findings, the major domains under data cleaning and their corresponding cleaning details are presented below.

S2ORC We unify the format by removing leading words such as "Abstract" and "abstract" in the string prefixes, along with any redundant whitespace.

Wikipedia Although the original publishers removed references (e.g., [1], [2]), URLs, multiple newlines, and bullet points, we identify some instances where article titles appear in the string prefixes. We remove these titles to ensure consistency.

wikiHow It should be noted that the original publishers retained only the first paragraph, the introduction of a wikiHow article, and this affects how we clean MGTs on wikiHow. In addition to processing the HWTs, the original publisher removed the "How to" prefix present in every wikiHow title from the title field. We restore the "How to" prefix to the title field. We believe this prefix potentially assists large language models in contextual understanding.

Baidu Tieba and Baidu Zhidao These domains are part of the DuReader2.0 (HongzheBi, 2023) dataset. We separate the data for these two sources. For Baidu Tieba, where multiple user replies are provided for each post, we retain only the first reply because subsequent entries may often consist of nested replies.

Weibo Review We remove unnecessary usernames and Chinese colons from the prefixes of each review string.

CLTS The original data contains numerous web-based whitespace characters, such as '\t', '\v', '\f', '\xa0', '\u2000' to '\u200b', " ", " ", " ". We remove these characters to the greatest extent possible. Furthermore, we identify and remove specific journalist remarks at the end of some news articles, such as "（澎湃新闻记者 xxx 整理）", as these serve as explicit hints of HWT.

H.2 MGT Clean

Based on the sample inspection findings, the major domains under data cleaning and their corresponding cleaning details are presented below.

S2ORC We observe that machines frequently generate various title-like prefixes such as "**Abstract**". We remove "Abstract", "abstract", all special characters such as '*' and '#', and redundant newlines from the prefixes. Notably, machine-generated S2ORC abstracts frequently begin with the phrase "This paper". Based on sampling survey, we found that—while not numerous—a distinct subset of scientific journal abstracts in human written papers indeed begin with the phrase "This paper." Inspired by the additional instructions in RAID (Dugan et al., 2024) (e.g., they appended the additional instruction "Do not repeat the title." to the prompt for the Reddit domain), we append "It is preferable not to start with \"This paper\"" to our prompt for S2ORC (specifically detailed in Appendix G.).

Wikipedia To align with HWTs, we remove redundant titles from the string prefixes and eliminate markdown syntax, including bullet points (-), bold markers (**), and header symbols (#, ##, ###).

wikiHow We remove redundant titles and whitespace from the string prefixes and retain only the first paragraph to align with HWTs.

CSL We remove all occurrences of "《{title}》" and "摘要：" from the string prefixes, along with redundant whitespace. Similar to the S2ORC English abstracts, we add the additional instruction to the CSL prompt: "最好不要以“本文”开头。".

Baidu Baike We remove titles appearing in various formats from the string prefixes (including "标题：{title}", "**{title}**", "【{title}】", "# {title}") and markdown symbols, including bullet points (-), bold markers (**), and header symbols (#, ##, ###). We also eliminate leading numbers in paragraphs (e.g., 1., 2.) and consolidate multiple consecutive newlines into a single newline.

Zhihu To align with HWTs, we replace all newlines with spaces. Specifically, multiple

consecutive newlines are collapsed into a single space.

CLTS To align with HWTs, we remove all newlines from the MGTs.

I Introduction to External Datasets for Generalization Ability Evaluation

We select the optimized versions of four datasets, M4 (Wang et al., 2024d), M4GT (Wang et al., 2024c), MAGE (Li et al., 2024), and HC3 (Guo et al., 2023), as external datasets to evaluate the generalization ability of detectors. These datasets are four representatives of the most diverse and challenging datasets among all benchmarks listed in Table 5.

The optimized versions specifically refer to the adjusted and refined releases provided for academic competition workshops focusing on MGT detection. For M4 (Wang et al., 2024d), we adopt the test set they released for SemEval-2024 Task 8 (Wang et al., 2024b). We denoted it as S-M4. Detailed statistics regarding its size and domain distribution are available in the corresponding overview paper (Wang et al., 2024b) and the official repository[14]. Regarding M4GT (Wang et al., 2024c), MAGE (Li et al., 2024), and HC3 (Guo et al., 2023), we utilize the validation sets from the COLING 2025 Workshop on MGT Detection Task 1 (Wang et al., 2025c), which we denote as C-M4GT, C-MAGE, and C-HC3, respectively. Detailed statistics regarding its size and domain distribution of these datasets are available in the corresponding overview paper (Wang et al., 2025c) and the official repository[15].

For English, we use the original English versions of these datasets. For Chinese, we extract all Chinese data from their original multilingual releases based on provided annotations. Notably, C-MAGE contains no Chinese data, so it is omitted from the experiments on our Chinese version.

J More Experiment Details

All our MGTs are generated in a single run, and all evaluation experiments process the entire split continuously within a single execution. We supplement additional experiment details as follows:

[14] https://github.com/mbzuai-nlp/SemEval2024-task8

[15] https://github.com/mbzuai-nlp/COLING-2025-Workshop-on-MGT-Detection-Task1

Experiments on the MAGA-cn Subset Table 13 presents the benchmarking results on the MAGA-cn split, which corresponds to Table 2 for the English split. Table 14 displays the generalization experiments for the final trained RoBERTa detector evaluated on MAGA-cn, serving as the Chinese counterpart to Table 3. The evaluation metrics and methodologies remain strictly identical to those used in the English version. At the end of Appendix F, we justify the selection of the four specific evaluation detectors, namely R-B CGPT CN, R-B MPU zhv3, Binoculars, and GECScore, because the vast majority of our selected detectors lack support for Chinese text detection. For the generalization experiments, we note at the end of Appendix I that C-MAGE is omitted due to the absence of Chinese data.

As shown in Table 13 and Table 14, the experimental results on the MAGA-cn split remain consistent with those observed on the MAGA split. On the MAGA-cn split, the average AUC of evaluation detectors decreases by 2.33% compared to the MGB-cn split, where neural-based detectors and metric-based detectors drop by 2.62% and 2.04%, respectively. Similarly, the average ACC(@FPR=5%) of evaluation detectors decreases by 1.70%, with neural-based detectors dropping by 0.73% and metric-based detectors decreasing by 2.67%. Similar to the trends on the MAGA split, the performance decrease discrepancy between neural-based detectors and metric-based detectors on the MAGA-cn split is marginal. In generalization experiments, R-B MAGA CN improves the average AUC on external datasets by 0.91% and the average ACC(@FPR=5%) by 3.04% compared to R-B MGB CN.

The relatively modest performance degradation in evaluation detectors is potentially due to the imbalanced detection capabilities of existing Chinese detectors. For instance, R-B CGPT CN exhibits a TPR close to 100% and a TNR close to 0%, whereas R-B MPU zhv3 and GECScore show a TNR close to 100% and a TPR close to 0%. This scarcity of available Chinese detectors might lead to an insufficient evaluation of the MAGA-cn split, which constitutes one of our limitations. Regarding generalization experiments, R-B MGB CN already achieves an average AUC of 98.59%, which restricts the margin for potential improvement. Consequently, the AUC enhancement achieved by R-B MAGA CN may be more substantial than

	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↓)
neural.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 13: Attack Experiment on MAGA-cn. An "✗" in the "Alignment" column indicates the MGB-cn split, while the opposite indicates the MAGA-cn split.

	S-M4-CN		**C-M4GT-CN**		**C-HC3-CN**		**AVERAGE**	
Metric	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)	AUC	ACC(5%)
R-B CGPT CN	99.24	95.19	98.95	94.29	99.96*	97.15*	99.38	95.54
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	99.04	94.90	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	**0.73**	**2.56**	**1.38**	**4.81**	**0.63**	**1.74**	**0.91**	**3.04**

Table 14: 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**		
Metric	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 GDARL-CDMMR*	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 BPO /w GDARL-CDMMR	48.89(5.20↓)	67.05(4.36↓)	64.93(2.45↓)	60.39(10.05↑)	83.33(2.45↑)	74.96(1.64↑)
MGB /w BPO /w self-refine	47.45(6.64↓)	65.40(6.01↓)	64.02(3.36↓)	63.92(13.58↑)	84.25(3.37↑)	75.50(2.18↑)
MGB /w BPO /w roleplaying	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 GDARL-CDMMR	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 roleplaying /w self-refine	47.36(6.73↓)	65.51(5.90↓)	63.93(3.45↓)	64.10(13.76↑)	84.13(3.25↑)	75.46(2.14↑)
MGB /w GDARL-CDMMR /w self-refine	47.85(6.24↓)	65.38(6.03↓)	64.00(3.38↓)	63.96(13.62↑)	84.26(3.38↑)	75.42(2.10↑)
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 BPO	46.69(7.40↓)	64.56(6.85↓)	63.62(3.76↓)	66.34(16.00↑)	84.50(3.62↑)	75.88(2.56↑)
MAGA w/o roleplaying	46.92(7.17↓)	64.79(6.62↓)	63.79(3.59↓)	65.45(15.11↑)	84.33(3.45↑)	75.80(2.48↑)
MAGA w/o GDARL-CDMMR	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 15: A more detailed ablation study on 10% data.

the absolute numerical increase suggests. But anyway, a joint analysis of Table 13 and Table 14 may reveal that both MGB-cn and MAGA-cn pose significantly greater detection challenges compared to existing Chinese datasets.

Experimental Results on MAGA-extra Splits The ablation studies presented in Table 4 provide a consolidated summary of the average experimental results across all individual MAGA-extra splits. Tables 30, 31, 32, 33 report the performance of evaluation detectors across all four extra splits under the English MAGA subset, which include

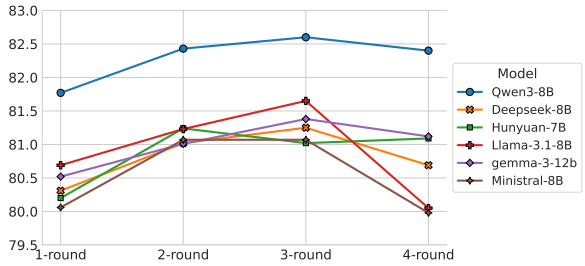

Figure 8: GDARL-CMMR generalization AUC by GDARL round. The CM group configuration use the most significant increase in Figure 15.

MAGA-extra-CDMMR, MAGA-extra-roleplaying, MAGA-extra-BPO, and MAGA-extra-self-refine. Analogously, Tables 38, 39, 40, 41 detail the detection results on the four corresponding Chinese extra splits. Minor performance fluctuations occur occasionally in these experiments, where specific evaluation detectors show either a marginal decrease or a slight increase in detection efficacy on certain extra splits.

Detailed Experimental Results on External Datasets Table 3 and Table 14 provide the comprehensive generalizations for the English and Chinese subsets, respectively, focusing primarily on the critical metrics of AUC and ACC(@FPR=5%). In the Tables described in the below in this paragraph, we provide additional documentation for the complete evaluation results, which encompass all four metrics of ACC, TPR, AUC, and ACC(@FPR=5%). Specifically, Tables 34, 35, 36, 37 compile the granular outcomes on S-M4, C-M4GT, C-MAGE, and C-HC3. Furthermore, Tables 42, 43, 44 display the detailed breakdown on S-M4-CN, C-M4GT-CN, and C-HC3-CN, respectively.

Ablation Study Details Table 15 provides a more detailed expansion of the experiments in Table 4. This analysis is based on 10% of the data, that is, based on 10% of the prompts and 10% of the HWTs, corresponding to the generation of 10% of the MGTs. The following are several key observations from Table 15.

First, removing any of the four alignment methods individually from MAGA leads to a performance decline compared to the full framework. This ablation study demonstrates that each of the four alignment methods contributes positively to MAGA.

Second, the performance improvement becomes increasingly marginal as the number of integrated methods increases. Overall, this trend indicates that substantial mutual dilution exists among these alignment methods.

Third, removing either self-refine or GDARL-CDMMR from MAGA leads to a performance decline, which demonstrates that integrating both self-refine and GDARL-CDMMR is beneficial and that they can mutually enhance performance beyond each other.

Fourth, the performance drop upon removing self-refine from MAGA is extremely marginal, despite being present. This suggests that the impact of self-refine is largely diluted, potentially because self-refine does not participate in the RL process, a limitation that agentic-RL may need to be introduced.

Fifth, taking the AUC under attack as an example, MGB /w roleplaying /w GDARL-CDMMR induces an additional 1.32% decrease in evaluation detector performance compared to MGB /w roleplaying, while MGB /w BPO /w GDARL-CDMMR leads to an extra 1.02% drop compared to MGB /w BPO. Meanwhile, MAGA w/o self-refine, which is equivalent to MGB /w roleplaying /w BPO /w GDARL-CDMMR, causes an additional 1.21% reduction compared to MGB /w roleplaying /w BPO. The marginal degradation brought by GDARL-CDMMR when introduced after combining BPO and roleplaying is largely comparable to when it is introduced after selecting only one of them. Despite the increased number of integrated methods, GDARL-CDMMR exhibits no further dilution, indicating to some extent that the combination of BPO and roleplaying is helpful to promote the efficacy of GDARL-CDMMR.

Figure 8 presents the ablation study on the number of adversarial training rounds in GDARL-CMMR. The overall performance generally converges after 3 rounds, whereas a 4-round configuration leads to apparent overfitting, resulting in a consistent performance decline. Figure 14 and Figure 15 explore the grouping mechanisms across domains and models for GDARL-CDMR and GDARL-CMMR, respectively. Specifically, Figure 15 provides a more comprehensive exploration that complements Figure 5. Figure 16 illustrates the t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations of texts across different domains and models, extending the details shown in

Figure 6. We utilize the bge-large-en-v1.5 model[16] (Xiao et al., 2024) for embeddings extraction in the t-SNE visualization, a choice validated by FineMedLM-o1 (Yu et al., 2025) for its suitability in t-SNE representations. Notably, as illustrated in Figure 16, while distinct domains exhibit clear clustering patterns, texts generated by different models may not be adequately characterized by two-dimensional features, resulting in the absence of discernible clusters within the t-SNE visualization. Furthermore, as shown in Figure 13, we analyze the log PPL computed mutually across texts generated by different generators. A given generator consistently assigns a low PPL to its own generated text but yields a substantially higher PPL when evaluating texts from other generators, indicating a pronounced discrepancies in viewpoints among different generators. This divergence in perspective is particularly prominent between predominantly Chinese-language generators and predominantly English-language generators. Consistent with the discussions in §4.3, the outcomes of these analyses and exploratory investigations guide our final hyperparameter configurations, which we document in Appendix L.

K Quantitative Text Analysis Details

K.1 Multidimensional Text Analysis

In this section, we detail the selection of dimension-specific metrics for multidimensional quantitative text analysis alongside the corresponding settings (e.g. parameter, model selection) utilized during calculation, and then present a more detailed report of the evaluation results.

As described in §5, the final value for each dimension is the average of the computed scores of all constituent metrics within that dimension. Metrics with values falling outside the $[0, 1]$ range undergo normalization. We implement two distinct strategies, namely the normalized score and the cosine similarity. Although cosine similarity is inherently suited for evaluating distributional alignment, we primarily rely on the normalized score (displayed in Figures 7, 9) as we found it demonstrates differences between text groups more prominently.

The normalized score is formulated to quantify the proximity between machine-generated texts (MGTs) and human-written texts (HWTs). Let S_M and S_H denote the raw metric values computed from all MGTs and HWTs on a specific split, respectively. Since our objective is to measure the degree of similarity, an optimal MGT performance implies that S_M closely approaches S_H. Consequently, when $S_M < S_H$, a higher S_M is desired, and we scale the metric by treating S_H as the baseline unit, yielding $S_M^{norm} = \frac{S_M}{S_H}$. Conversely, when $S_M > S_H$, a lower S_M is preferred, and we apply an inverse ratio defined as $S_M^{norm} = \frac{1/S_M}{1/S_H} = \frac{S_H}{S_M}$. The final normalized score for a given dimension is then derived by taking the arithmetic mean of the normalized scores across all underlying metrics.

To illustrate this process more clearly, using sentiment consistency as an example, Table 16 shows that MGB-M yields average negative, neutral, and positive sentiment scores of 0.3243, 0.4030, and 0.2727, respectively, while HWT obtains scores of 0.3455, 0.3996, and 0.2549. The resulting normalized score is computed as $(\frac{0.3243}{0.3455} + \frac{0.3996}{0.4030} + \frac{0.2549}{0.2727})/3 = 0.9550$.

For the cosine similarity, the calculation follows the standard algebraic definition, where the sentiment consistency between MGB-M and HWT is evaluated as the cosine similarity between the vectors $[0.3243, 0.4030, 0.2727]$ and $[0.3455, 0.3996, 0.2549]$.

It should be noted that lexical diversity is treated exceptionally. Instead of positioning HWT as the fixed baseline of 1.0, we designate the maximum observed value across groups as 1.0 to serve as the reference target, for reasons elaborated in §5. Such a setting may not be appropriate either.

We select only a single representative sub-metric for each metric (for example, for ROUGE, we selected only ROUGE-2-F1). This choice ensures a consistent number of scalar variables across different metrics, thereby facilitating unbiased averaging. The detailed specifications for each dimensional metric are provided below.

Content Similarity We select Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), and Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005). Among the various sub-metrics available for ROUGE, we choose ROUGE-2-F1 to represent the content similarity performance based on qualitative justifications.

[16] https://huggingface.co/BAAI/bge-large-en-v1.5

Specifically, ROUGE-1 focuses on unigrams and is suitable for evaluating semantic coverage in short texts, whereas ROUGE-2 targets bigrams and better captures phrase-level semantic coherence in longer documents. ROUGE-L utilizes the longest common subsequence and fits tasks that heavily rely on word order, such as machine translation. Regarding the optimization priorities of these sub-metrics, text summarization often prioritizes recall over precision and F1 score to ensure comprehensive coverage of core information from the source text. Conversely, machine translation and question answering favor precision over recall and F1 score to guarantee output accuracy. For general content generation tasks, the F1 score represents the most suitable compromise because it balances both coverage and accuracy. Consequently, we select ROUGE-2-F1 as the primary sub-metric.

Semantic Similarity We select BERTScore (Zhang* et al., 2020) as our metric and utilize the corresponding F1 score. To compute this metric, we extract text embeddings using the bert-base-uncased model (Devlin et al., 2019), which is the default configuration in the official bert_score package. Prior research on semantic similarity analysis, HC-Var (Xu et al., 2024), implements the MAUVE metric (Pillutla et al., 2021) to calculate the overall similarity between two distinct text datasets. Because our data generation follows a paired, one-to-one basis strategy, we can compute the BERTScore between individual MGT and HWT pairs directly, subsequently reporting the arithmetic mean across the entire dataset split.

Lexical Diversity We incorporate the Type-Token Ratio (TTR) (Chotlos, 1944), Yule's K (Yule, 2014), and the 2-gram vocabulary size metric introduced by M4 (Wang et al., 2024d). Although a universally accepted benchmark for lexical diversity remains absent in the literature, TTR and Yule's K are widely established as foundational indicators of vocabulary variation, making their selection a reliable baseline for our evaluation. Furthermore, empirical findings from M4 (Wang et al., 2024d) demonstrate that constructing datasets with multi-generators substantially increases the vocabulary size of unigrams and bigrams. This multi-generator paradigm enhances the lexical richness of machine-generated texts within the resulting detection datasets, which ultimately benefits the training and evaluation of robust MGT detectors. Notably, the vocabulary sizes of unigrams and bigrams for MGTs reported by M4 (Wang et al., 2024d) already exceed those observed in HWTs.

Text Readability We select Kincaid Reading Ease (Kincaid et al., 1975), which we denote as Flesch Kincaid for brevity, alongside Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969) and Dale Chall (Dale and Chall, 1948). These three metrics are widely recognized as foundational standards for text readability evaluation, which justifies our selection. Previously, AA (Uchendu et al., 2020) implemented Flesch Kincaid in their analytical study within the machine-generated text detection domain. The Flesch Kincaid framework actually contains multiple sub-metrics. We specifically select Kincaid Reading Ease to maintain a single sub-metric value, deliberately excluding Flesch-Kincaid Grade level which maps performance to school ages.

Sentiment Consistency We employ twitter-xlm-roberta-base-sentiment[17] (Barbieri et al., 2022) as our sentiment classification model, which aligns with the methodology established by HC3 (Guo et al., 2023). Rather than merely reporting the discrete classification distributions shown in Figure 26, we evaluate dimensional similarity by calculating the arithmetic mean of the continuous probability scores across three sentiment dimensions. To illustrate this approach more clearly, consider a toy dataset containing two sentences. If the first sentence receives negative, neutral, and positive probability scores of 0.3, 0.3, and 0.4, respectively, and the second sentence receives scores of 0.2, 0.3, and 0.5, then the final scores for the entire dataset are derived by averaging these individual probabilities, yielding $(0.3 + 0.2)/2 = 0.25$ for negative, $(0.3 + 0.3)/2 = 0.30$ for neutral, and $(0.4 + 0.5)/2 = 0.45$

More Result Details and Analysis Table 16 presents the specific values for the individual metrics that correspond to the radar chart illustrated in Figure 7. Table 17 expands upon Table 16 by incorporating detailed statistics from the machine-generated texts across all MAGA extra splits, which are visually aggregated in

[17] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

the comprehensive radar chart in Figure 9. To provide a more granular examination of these metrics, Figures 17, 18, 19, 20, 21, 22, 23, 24, and 25 display box plots representing the metric distributions computed from individual text instances across these splits, highlighting key descriptive statistics such as the median and the interquartile range. Additionally, Figure 26 depicts the proportional distribution of predicted sentiment classes across all splits, where the sentiment label for each text instance is determined by the highest prediction probability. As discussed in §5, our analytical findings remain broadly consistent with conclusions established in prior literature. In addition to the observations already detailed in §5, we provide a supplementary discussion here regarding the sentiment classification results, which also align with the conclusions reported in HC3 (Guo et al., 2023). Specifically, the categorical distributions in Figure 26 reveal that human-written texts exhibit a low frequency of positive sentiment alongside a higher prevalence of negative and neutral expressions, with neutral sentiment being the most dominant. Conversely, the baseline MGB-M, which lacks alignment enhancement, tends to be overly conservative, yielding a disproportionately higher rate of neutral sentiment and a lower rate of negative sentiment relative to HWTs, a discrepancy that leads to a sentiment mismatch between MGB-M and HWTs. Texts generated via alignment augment, such as MAGA-M, mitigate this disparity substantially. Furthermore, the granular averaging of continuous sentiment probability scores in Table 17 also supports similar conclusions while revealing a relatively more balanced distribution across the three probability categories. This trend indicates that the underlying sentiment within a single sentence may possess inherent complexity. A more granular analysis may be necessary to some extent.

K.2 PPL Analysis

Existing MGT Detection Benchmarks typically report the average perplexity (PPL) of MGTs within their datasets, often utilizing multiple evaluation models, as demonstrated in RAID (Dugan et al., 2024). Following this convention, we present our statistical report in Figure 11 and Figure 12. During the statistical process, several critical methodological questions arise, specifically whether to employ base models or chat (instruct) models, and whether to include the generation prompts during evaluation. These inquiries motivate the statistical experiments detailed in this section. It is worth noting that this section primarily analyzes and discusses these procedural

	MGB-M	MAGA-M	HWT
Content Similarity			
ROUGE	0.0479	0.0387	1.00
BLEU	0.0235	0.0165	1.00
METEOR	0.1974	0.2158	1.00
Average	0.0896	0.0903	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
Normalized Score	0.8340	0.9861	1.00
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
Normalized Score	0.9550	0.9752	1.00
Cosine Similarity	0.9989	1.00	1.00

Table 16: Multi-dimensional human-machine text analysis results across three texts: MGB-M, MAGA-M, and HWT. The evaluation covers content similarity, semantic similarity, lexical diversity, text readability, and sentiment consistency metrics.

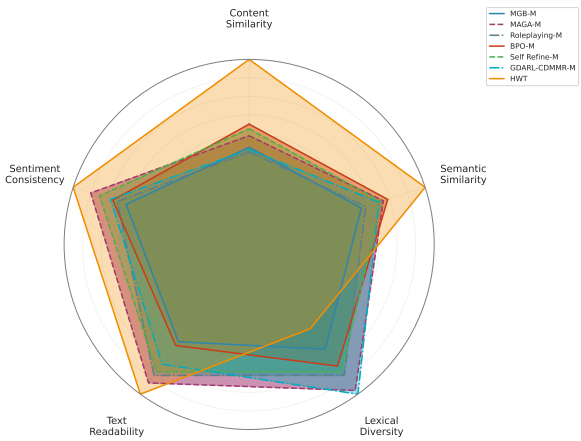

Figure 9: Multidimensional Text Analysis - Full.

	MGB-M	MAGA-M	Roleplaying-M	BPO-M	Self Refine-M	GDARL-CDMMR-M	HWT
Content Similarity							
ROUGE	0.0479	0.0387	0.0411	0.0505	0.0473	0.0434	1.00
BLEU	0.0235	0.0165	0.0189	0.0219	0.0205	0.0204	1.00
METEOR	0.1974	0.2158	0.1971	0.2119	0.2093	0.1984	1.00
Average	0.0896	0.0903	0.0857	0.0948	0.0924	0.0874	1.00
Semantic Similarity							
BERTScore	0.8252	0.8287	0.8256	0.8302	0.8285	0.8284	1.00
Lexical Diversity							
TTR	0.7893	0.7991	0.8306	0.7895	0.7916	0.8496	0.8271
Yule's K	74.99	52.02	55.94	74.44	62.41	50.76	100.18
Bigram Vocab Size	4261689	5467106	4050008	4909749	4665052	5511160	3253456
Normalized Score	0.7931	0.9694	0.8733	0.8340	0.8638	1.00	0.6902
Text Readability							
Flesch Kincaid	40.53	53.39	50.76	40.59	50.35	46.83	54.54
SMOG	14.10	12.15	12.66	13.97	12.65	13.98	11.91
Dale Chall	11.19	10.24	10.28	10.98	10.46	10.91	10.23
Normalized Score	0.8340	0.9861	0.9555	0.8428	0.9476	0.8827	1.00
Cosine Similarity	0.9905	1.00	0.9995	0.9912	0.9994	0.9966	1.00
Sentiment Consistency							
Negative	0.3243	0.3439	0.3553	0.3268	0.3461	0.3264	0.3455
Neutral	0.4030	0.3894	0.3819	0.4037	0.3841	0.4065	0.3996
Positive	0.2727	0.2667	0.2718	0.2705	0.2698	0.2672	0.2549
Normalized Score	0.9550	0.9752	0.9553	0.9594	0.9681	0.9606	1.0000
Cosine Similarity	0.9989	1.00	0.9990	0.9991	0.9993	0.9992	1.00

Table 17: Multi-dimensional human-machine text analysis results across seven texts: MGB-M, MAGA-M, RolePlaying-M, BPO-M, Self Refine-M, GDARL-CDMMR-M, and HWT. The evaluation covers content similarity, semantic similarity, lexical diversity, text readability, and sentiment consistency metrics.

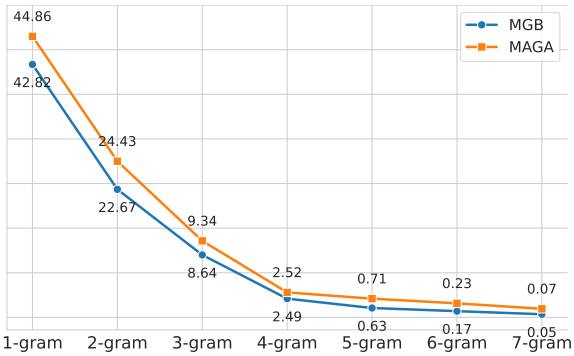

Figure 10: Ngram overlap rate with HWT. The calculation measures the percentage of the MGT's Ngram vocabulary that overlaps with the HWT's Ngram vocabulary, which is the intersection size divided by the HWT's Ngram vocabulary size.

questions as a supplementary investigation to our primary study, which maintains a relatively distant relation to the core tasks of MGT detection.

As discussed in §4.3 and Appendix J, and illustrated in Figure 13, our core finding reveals a pronounced discrepancy in viewpoints among different generators. The variations in these discrepancies closely correspond to the performance gains achieved by different generator combinations within the GDARL-CMMR framework. We next address the methodological questions raised at the beginning of this section.

To conduct this analysis, we select three categories of language models, evaluating both the base and instruct variants for each. The first category comprises English-centric models used during the dataset generation process, specifically Llama-3.1-8B and Llama-3.1-8B-Instruct. The second category consists of Chinese-centric models also utilized in dataset generation, where we select Qwen3-8B-Base and Qwen3-8B. The third category includes models completely withheld from the dataset generation process, for which we choose glm-4-9b and glm-4-9b-chat.

In Figure 12, we evaluate both the base and chat versions of these models by calculating the log PPL directly without prompts. The results

indicate that the choice between base and chat architectures yields negligible differences in direct perplexity estimation. This demonstrates that when computing PPL without prompts, the choice between base and instruct architectures introduces minimal variance, though base models exhibit marginally lower values.

In Figure 11, we compare the PPL calculated under conditional probability given the prompt against the PPL calculated without the prompt. In this holistic analysis, evaluation with prompts does not consistently yield lower PPL values and occasionally results in significantly higher values, a phenomenon potentially driven by the divergent viewpoints among different models. Intuitively, calculating PPL with prompts can better highlight these conflicting viewpoints across models. However, the differences observed on glm-4-9b remain extremely minor.

Both Figure 12 and Figure 11 demonstrate that the perplexity of MAGA-M is slightly higher than that of MGB-M. An increase in PPL does not fundamentally contradict alignment enhancement, as it likely stems from the viewpoint discrepancies among different models, especially since the calculation models are not the original authors of the text. Figure 13 clearly indicates that models assign a substantially lower PPL to the MAGA-M texts they generated themselves. Consequently, while PPL remains a valuable statistic, it is not always a reliable metric for judging alignment quality.

Regarding the questions posed initially, we draw the following conclusions. First, for a dataset containing texts generated by a sufficiently diverse set of generators (which can be viewed as generator-agnostic analysis), utilizing either an internal model involved in generation or an external unseen model yields similar PPL results. Second, in generator-agnostic PPL analysis, the distinction between base and instruct models is minimal, and evaluating with or without prompts shows negligible variation. Therefore, researchers can reliably implement any single configuration consistently throughout their analysis.

Intuitively, using an external model, utilizing prompts whenever available, and prioritizing base models when prompts are absent might indeed seem preferable. Our findings demonstrate that these methodological variations introduce negligible differences, indicating that researchers can confidently adopt and maintain any single

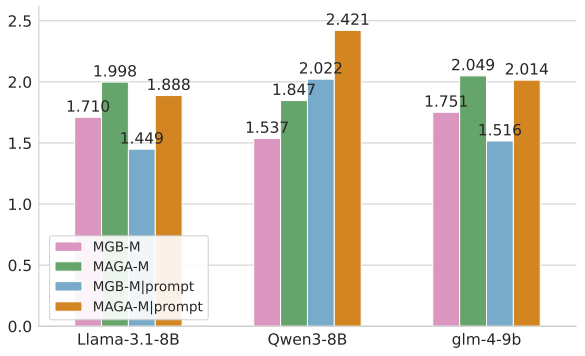

Figure 11: Compare the log PPL calculations with and without the prompt condition.

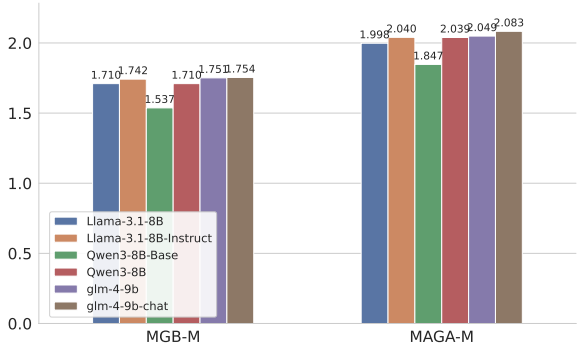

Figure 12: Direct log PPL computed on the base/chat LLM without the prompt condition.

calculation pipeline without concern. Or in other words, it is preferable to follow the intuitively superior approach, as it appears more reliable intuitively, but doing so is not strictly essential, and one need not harbor any additional apprehensions if unable to do so.

L Training Hyperparameters

L.1 Common Hyperparameters

The fine-tuning of RoBERTa consistently employs a learning rate of 5e-5, a total batch size of 64, the AdamW optimizer with a weight decay of 0.01, a warmup ratio of 0.03, and a cosine learning rate scheduler. The classification head is initialized using Kaiming uniform initialization, while pre-trained parameters remain intact before fine-tuning. Following the baseline configuration reported in SemEval-M4 (Wang et al., 2024b), the training duration is limited to one epoch. This single-epoch strategy is justified by three factors: the extensive training set exceeding 100k samples provides sufficient signal for convergence, the setting ensures a strictly fair comparison with established benchmarks, and it promotes robust

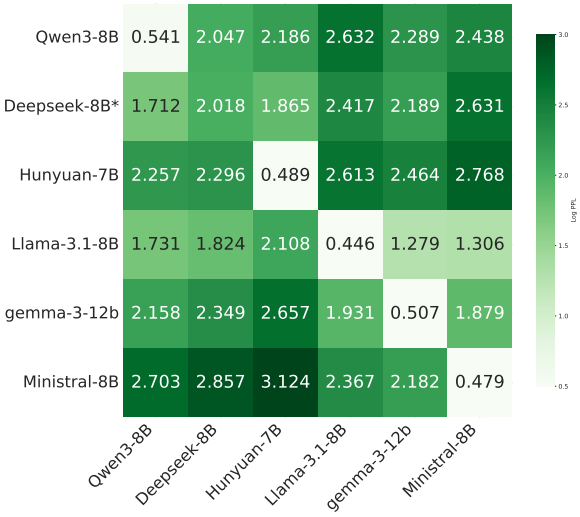

Figure 13: The log PPL calculated on MAGA-M, with the x-axis representing the LLM of the generated text and the y-axis representing the LLM used to calculate the log PPL. The calculation included prompts as conditions, but DeepSeek did not include chains-of-thoughts (COT) in its calculation.

Group	Domain	Model
GA	Wikipedia, wikiHow, CC News, NPR News, S2ORC	Qwen3-plus, Qwen3-8B, DeepSeek-V3, DeepSeek-R1-0528-Qwen3-8B, Hunyuan-TurboS, Hunyuan-7B-Instruct
GB	Reddit, Trustpilot Reviews, Amazon Reviews, Yahoo Answers, Natural Questions	Llama-3.1-8B-Instruct, Mistral-Medium, Ministral-8B-Instruct-2410, Gemini-2.0-flash, GPT-4o-mini, DeepSeek-V3

Table 18: Overview of domains and evaluation models across different groups.

generalization by preventing over-optimization.

For GDARL's RL process, we implement the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) integrated with Low-Rank Adaptation (LoRA) (Hu et al., 2021) to facilitate efficient training on our 8 × NVIDIA Tesla L20 48G GPUs. The LoRA configuration consists of a rank of 16 and an alpha value of 32. We maintain a micro batch size of 1 and a global batch size of 1024, achieved through a per-device batch size of 32 and a rollout factor of 4. The learning rate is 5e-6 with a cosine scheduler. The training duration is one epoch in each round of GDARL's RL process.

L.2 GDARL-CDMMR Rounds and Group

The GDARL-CDMMR process comprises three adversarial training rounds. The final grouping of GDARL-CDMMR is presented in Table 18. These configurations are analyzed and discussed in detail within the Ablation Study Details in Appendix J.

M Computational Overhead Analysis

Table 19 presents the computational time overhead for constructing each split within the English MAGA subset. The construction duration for the MAGA-cn subset exhibits similarly. The entire MAGA dataset requires approximately 2,800 hours to construct. Within each split, 72k MGTs are generated, with each individual model producing 6k MGTs. For open-source models, we employ vLLM for batch inference, which requires approximately 35 minutes per model to generate 6k MGTs. An exception is DeepSeek-R1-0528-Qwen3-8B, which requires approximately 1 hour due to the generation of chain-of-thought tokens. For commercial models, invoking APIs via a single-threaded process to generate 6k MGTs takes approximately 4 hours per model. The Self Refine process requires approximately four times the inference time of direct inference because it involves acquiring feedback and subsequent self-refinement. Regarding GDARL, among the models utilized, the smallest 7B model requires approximately 70 hours of training, whereas the largest 12B model requires approximately 150 hours. The training time for RoBERTa is negligible for GDARL. The BPO optimization of 72,000 prompts requires approximately 2 hours with a batch size of 64, implemented via the Hugging Face library. Distinct from other operations, the final trained RoBERTa detector on each split is fine-tuned on a training subsplit containing 120k instances, which consist of 60k MGTs and 60k HWTs, requiring approximately 35 minutes.

In general, regarding the numerical performance in our experiments, these methods exhibit comparable results, yet their computational time overheads vary significantly. GDARL-CDMMR demonstrates substantially lower practicality

Split	Training Time	Inference Time	Total Time	Hardware
MGB	-	28h	28h	8 × NVIDIA Tesla L20 48G GPUs
MAGA-extra-GDARL-CDMMR	530h	28h	558h	8 × NVIDIA Tesla L20 48G GPUs
MAGA-extra-BPO	-	31h	31h	8 × NVIDIA Tesla L20 48G GPUs
MAGA-extra-roleplaying	-	29h	29h	8 × NVIDIA Tesla L20 48G GPUs
MAGA-extra-self-refine	-	108h	108h	8 × NVIDIA Tesla L20 48G GPUs
MAGA	543h	112h	655h	8 × NVIDIA Tesla L20 48G GPUs
Total	1073h	336h	1409h	8 × NVIDIA Tesla L20 48G GPUs

Table 19: Training, inference, and total time consumption across different splits on the English MAGA Dataset along with the hardware configuration specifications.

	ACC	TPR	TNR	AUC
MGB-train	99.78	99.72	99.84	99.99
MGB-validation	99.38	99.15	99.61	99.97

Table 20: R-B MGB Performance on the MGB training and validation subsplits.

Training		Validation	
Score	Percentage	Score	Percentage
> 0.9999	46.97%	> 0.9999	47.22%
> 0.9995	85.18%	> 0.9995	84.35%
> 0.999	91.74%	> 0.999	90.80%
> 0.995	96.79%	> 0.995	95.97%
> 0.99	97.85%	> 0.99	96.93%
> 0.9886	98.00%	> 0.98	97.59%
> 0.98	98.55%	> 0.9581	98.00%
> 0.9564	99.00%	> 0.6304	99.00%
> 0.8192	99.50%	> 0.1547	99.50%
> 0.0670	99.90%	> 0.0056	99.90%

Table 21: Cumulative distribution of machine-like scores given by R-B MGB on MGB training and validation splits and their corresponding percentage.

	Wikipedia	wikiHow	Trustpilot Reviews	Amazon Reviews
MGB	74.67	67.12	72.83	73.04
/w GDARL	74.67	67.08	73.50	73.16

Table 22: AUC performance with or without GDARL.

than other approaches, and Self Refine also presents slightly lower practicality than BPO and roleplaying. Nevertheless, our human analysis in Appendix O indicates that the intuitive effects of these methods are different. Furthermore, the ablation study details in Appendix J clearly demonstrate that integrating these methods consistently enhances mutual performance, despite a severe dilution effect.

N Discussion

N.1 Cross-Mutual Reward Component and Existing GDARL Frameworks

Existing ARL Frameworks To the best of our knowledge, existing Generator-Detector Adversarial Reinforcement Learning (GDARL) frameworks primarily include RADAR (Hu et al., 2023), AuthorMist (David and Gervais, 2025), llm-detector-evasion (Nicks et al., 2023), and HUMPA (Wang et al., 2025a).

RADAR (Hu et al., 2023) addresses the vulnerability of existing machine-generated text detectors to paraphrased text by proposing a GDARL framework. This framework employs the proximal policy optimization (PPO) algorithm (Schulman et al., 2017) to implement a co-training scheme where the generator and the detector alternate step by step within each epoch. The loss function requires simultaneously distinguishing among original AI text, AI-paraphrased text, and human text. By adjusting the loss weight λ, the framework balances detection performance between original and paraphrased texts.

Similarly, AuthorMist (David and Gervais, 2025) focuses on GDARL for AI-paraphrased text. The authors select a suite of detectors and conduct adversarial learning against one detector at a time. They then evaluate the evasion effectiveness of the resulting AI-generated text across all selected detectors.

The llm-detector-evasion framework (Nicks et al., 2023) utilizes the direct preference optimization (DPO) algorithm (Rafailov et al., 2023) to achieve GDARL. The authors demonstrate that current mainstream AI text detectors are highly susceptible to evasion, indicating that existing detection solutions lack reliability.

HUMPA (Wang et al., 2025a) improves upon the method proposed in llm-detector-evasion (Nicks et al., 2023). The authors point out that directly fine-tuning a language model with a small parameter scale often compromises training

	Attack			Generalization		
Metric	TPR	AUC	ACC(@FPR=5%)	TPR	AUC	ACC(@FPR=5%)
MGB	54.09	71.40	67.41	53.13	82.91	75.65
/w RADAR's framework	48.73(5.36↓)	66.19(5.21↓)	64.73(2.68↓)	55.61(2.48↑)	76.87(6.04↓)	65.83(9.82↓)
/w RADAR's framework + CDMMR	47.82(6.27↓)	65.60(5.80↓)	64.35(3.06↓)	55.26(2.13↑)	79.39(3.52↓)	71.56(4.09↓)
/w GDARL-CDMMR	49.95(4.14↓)	67.91(3.49↓)	65.18(2.23↓)	60.97(7.84↑)	85.02(2.11↑)	77.28(1.63↑)

Table 23: Ablation experiments vs RADAR.

stability. To address this, their first step involves scoring text using a detector and fine-tuning a small language model via DPO. Subsequently, they introduce an attack intensity hyperparameter α to proxy the attack during the decoding stage. This approach avoids modifying the parameters of the large language model. Instead, during the generation of each token, it applies an offset to the output of the large model based on the logit difference between the fine-tuned small model and its original counterpart.

Why don't we use a specific existing GDARL framework? We do not adopt specific existing GDARL frameworks due to several distinct reasons. First, RADAR (Hu et al., 2023) explicitly reports their limitations that its method may show degraded detection performance against native LLM-generated texts (without paraphrasing). In the main experiments presented by RADAR, while its capacity to detect paraphrased text significantly outperforms the selected baselines, its performance on native text remains inferior to several baseline methods, demonstrated by that AUC is approximately 4% lower than that of R-B GPT2. AuthorMist is likely similar, as it also focuses on paraphrased text. Second, the attack strength α of HUMPA (Wang et al., 2025a) may be difficult to set based on the specific task. Although HUMPA (Wang et al., 2025a) reports that its BERTScore decreases by no more than 0.02 and its ROUGE-1 score decreases by no more than 0.03, which can be interpreted as marginal changes, this performance degradation contradicts our objectives to some extent.

Why do we need Cross-Mutual Reward? GDARL is the general term for existing GDARL algorithms. We also denote the vanilla formulation of these frameworks as GDARL, which we implement. To ensure optimal training stability, we eschew the co-training methodology of RADAR and instead execute alternation on an epoch-by-epoch basis, where the generator and the detector undergo alternate training for one full epoch per round. Under this setup, severe overfitting in RoBERTa prevents the LLM from receiving an effective reward signal. Table 20 presents the performance of R-B MGB on the MGB training and validation subsplits, where the detection accuracy reaches approximately 100% on the training subsplit and remains as high as 99.38% on the validation subsplit. Table 20 also outlines the machine-like scores assigned by R-B MGB to the machine-generated texts within these subsplits. Notably, on the training subsplit, 98.55% of the machine-generated texts receive a score exceeding 0.98, and 99% score above 0.9564. Consequently, the human-like score for almost all machine-generated text approaches zero, which eliminates viable reward gradients.

Table 22 provides the supplementary results when integrating GDARL into the main diagonal AUC values of Figure 5. Without any tailored modifications, applying standard GDARL yields virtually no performance gains.

We emphasize that our Cross-Mutual Reward (CMR) mechanism serves as a direct application of the theoretical foundations established in HC-Var (Xu et al., 2024). HC-Var distinguishes between irrelevant and relevant features (Xu et al., 2024), demonstrating that RoBERTa prone to overfitting on in-distribution data. When variables such as the prompts, domains, and generators used to produce MGT remain identical, RoBERTa relies heavily on irrelevant features (e.g. text length) for classification (Xu et al., 2024). Scores derived under these conditions may also be unreliable to some degree. Conversely, out-of-distribution generalization must depend primarily on relevant features (Xu et al., 2024), which represent the latent, intrinsic differences between human and machine text that remain fundamentally unidentified (Xu et al., 2024).

Additionally, we present our initial trial experiments in Table 23. As reported by RADAR (Hu et al., 2023), its detection capability against

vanilla MGT indeed sometimes tends to decline, rendering RADAR more suitable for identifying paraphrased text. It is critical to emphasize that our CMR component operates as an external procedural control mechanism, which remains inherently compatible with the inner workings of these algorithms rather than creating conflicts. As demonstrated in Table 23, integrating the CMR component with RADAR consistently enhances overall performance under our experiment.

N.2 Why not Distill?

If we need more human-like text, why not directly perform knowledge distillation on commercial LLMs or simply use commercial LLMs? After all, commercial LLMs generally possess superior overall capabilities.

Many previous studies, including M4(Wang et al., 2024d), have shown that increasing the diversity of LLM selections, especially incorporating a number of small-parameter LLMs, leads to significant improvements in detector learning.

CogLM(Wang et al., 2025b) once pointed out through relevant metric analysis that LLMs 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 GDARL is to enable various open-source LLMs to generate more human-like MGT while maintaining their original styles.

N.3 Why Single-Round Self Refine?

Table 24 presents the ablation study on the iteration rounds of Self-Refine. The experiment settings employed in this experiment is consistent with the analysis presented in Table 4. Instead of allowing the LLM to determine the termination condition dynamically, we simply strictly enforce a predefined number of iteration rounds. The empirical results indicate that multiple iterations yield marginal performance variance, characterized primarily by fluctuations, within the scope of our study. As discussed in Appendix M, the temporal overhead of a single-round Self Refine is approximately four times that of direct generation. Given this substantial increase in computational cost alongside the potential negligible performance gains, we systematically adopt a single-round configuration for Self Refine across all subsequent dataset construction and experimental procedures.

O Human Analysis Case Study

Table 45 presents our human analysis case study comparing various alignment methods. We conduct a manual analysis to identify the advantages and disadvantages of each method. Machine-generated traces are highlighted in red, whereas human-like traces are highlighted in green.

First, several drawbacks are observed in the MGT within the unenhanced MGB split. (1) The phrase "consistently perform behaviors" deliberately emphasizes the continuity principle of wikiHow and downplays one-time actions, which serves as a key signal of AI mimicking the practicality-oriented style of wikiHow. (2) The expression "physical, mental, and social function" appears to be an 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 movement, maintaining rest cycles, managing stress inputs, and avoiding known hazards" is a typical indicator of the detailed breakdown-style writing of AI. (4) The sentence "Execute these actions repeatedly over time" re-emphasizes long-term persistence; such double emphasis is a characteristic feature of AI-generated content.

For BPO, roleplaying, and self-refine, the machine traces exhibit similarities: these methods deliberately emphasize continuity instead of employing more natural and casual expressions, and they consistently decompose and stack core principles of wikiHow for illustration. Examples include "nutrition, exercise, sleep, and mental health" from roleplaying and "focus on eating well, staying active, getting enough rest, and managing stress" from self-refine.

For BPO, examples of human-like traces in BPO include the phrase "support your body and mind". When humans discuss health-related topics, they typically do not rigidly bifurcate "body"

	Attack			Generalization		
Metric	TPR	AUC	ACC(@FPR=5%)	TPR	AUC	ACC(@FPR=5%)
MGB	54.09	71.4	67.41	53.13	82.91	75.65
/w self-refine 1 round	48.87(5.22↓)	66.32(5.08↓)	64.43(2.98↓)	64.21(11.08↑)	85.84(2.93↑)	77.47(1.82↑)
/w self-refine 2 round	47.98(6.11↓)	65.87(5.53↓)	64.27(3.14↓)	63.02(9.89↑)	85.59(2.68↑)	77.53(1.88↑)
/w self-refine 3 round	49.03(5.06↓)	66.34(5.06↓)	64.58(2.83↓)	63.14(10.01↑)	85.07(2.16↑)	77.31(1.66↑)
/w self-refine 4 round	48.69(5.40↓)	66.74(4.66↓)	64.81(2.60↓)	62.85(9.72↑)	85.23(2.32↑)	77.18(1.53↑)

Table 24: Self-refine rounds ablation experiments results.

and "mind". Such concise and coherent phrases summarize the core meaning in a manner that is natural and empathetic while avoiding the accumulation of mechanical terminology. In terms of content, BPO aligns more closely with human expression habits.

For roleplaying, examples of human-like traces are as follows: the sentence "Being healthy is about balancing various aspects of your life" adopts the tone of college students conversing with peers, and "It is not about perfection, but consistency" reflects the empathetic mindset typical of such a demographic.

For self refine, examples of human-like traces include the sentence "You do not have to make huge changes all at once—small, consistent steps can lead to big improvements over time". This "empathy-first" guidance transforms the text from a cold guideline into warm advice, which better conforms to the encouraging logic of interpersonal communication. As a method involving self-reflection and refinement by LLMs, self-refine actually exhibits a stochastic rather than a single improvement dimension compared to the previous two methods, making it suitable for integration with other approaches to some degree.

Table 46 presents the advantages of GDARL-CDMMR, which provides a more moderate, implicit, and quantifiable optimization process. Enabled by the feedback mechanism of the detector, the most prominent characteristic of this method lies in its capability to eliminate explicit machine-generated artifacts. Even when the MAGA Additional Instruction is applied to explicitly prohibit the generation of introductory remarks, an implementation inspired by RAID (Dugan et al., 2024) and detailed in Appendix G, various LLMs, including commercial variants, still tend to insert opening sentences in unexpected manners. GDARL-CDMMR can mitigate such distinct machine artifacts that prompt engineering struggles to control, though its primary limitation is that it applies exclusively to open-source models rather than commercial ones.

The MGT in the MAGA split mitigates machine-specific artifacts across a wider perspective and at a deeper level by integrating the advantages of these four methodologies.

Table 47 presents a case study in Chinese. The human analysis results across Tables 46, 47, and Table 45 exhibit similar patterns. Obvious machine-generated hints are highlighted in orange, while standard machine-generated and human-written traces are marked in red and green, respectively. Further redundant elaborations about Tables 46, 47 are omitted due to the similar traces.

	Original 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 25: 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.

	Original 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。\n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。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 26: 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 27: 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 28: 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 29: Samples of Role-Playing Prompts

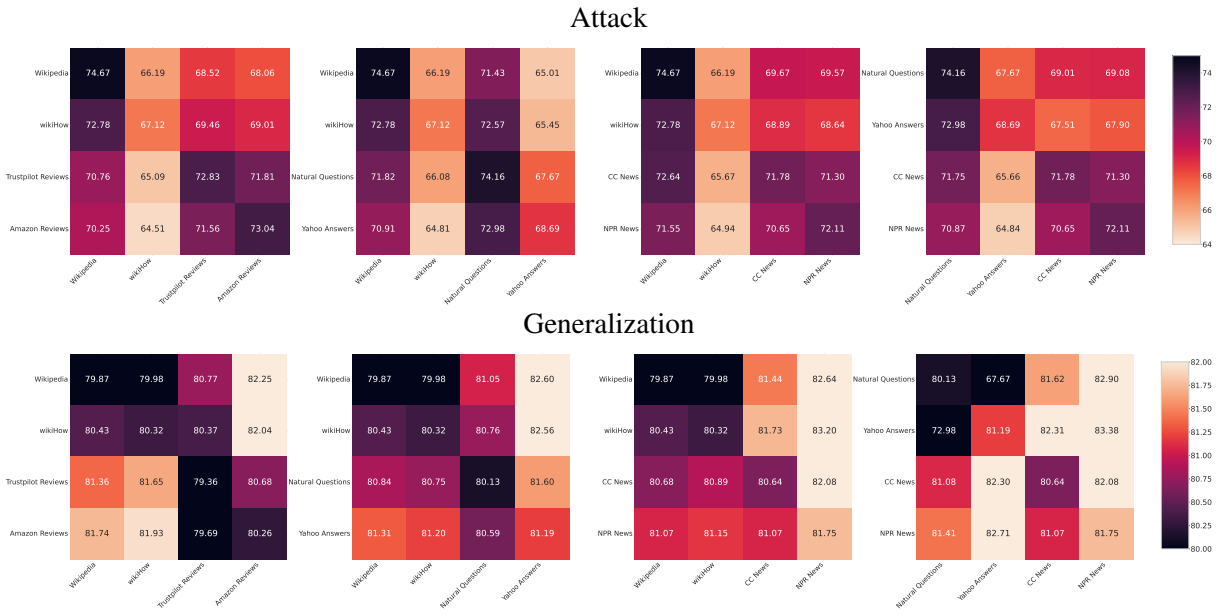

Figure 14: GDARL-CDMR-matrix results. Top: Attack performance; Bottom: Generalization performance.

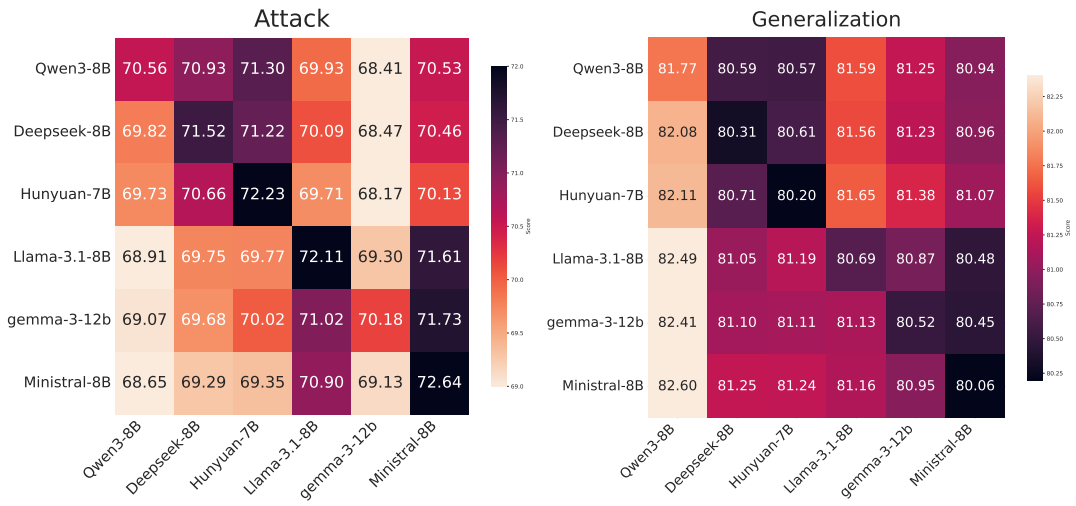

Figure 15: GDARL-CMMR matrix analysis for attack and generalization

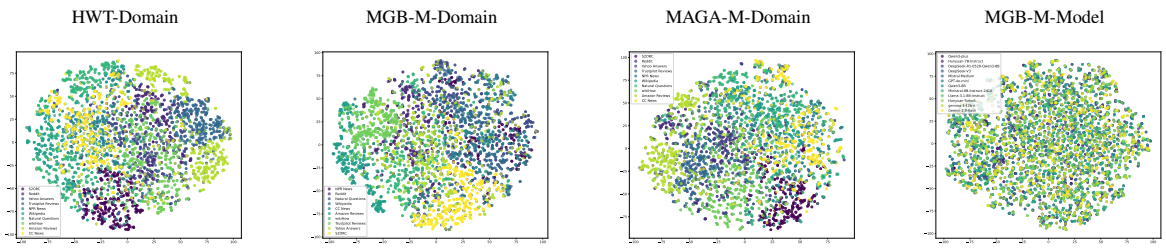

Figure 16: T-SNE visualization results of different domains and models. From left to right: HWT domain, MGB-M domain, MAGA-M domain, MGB-M Model.

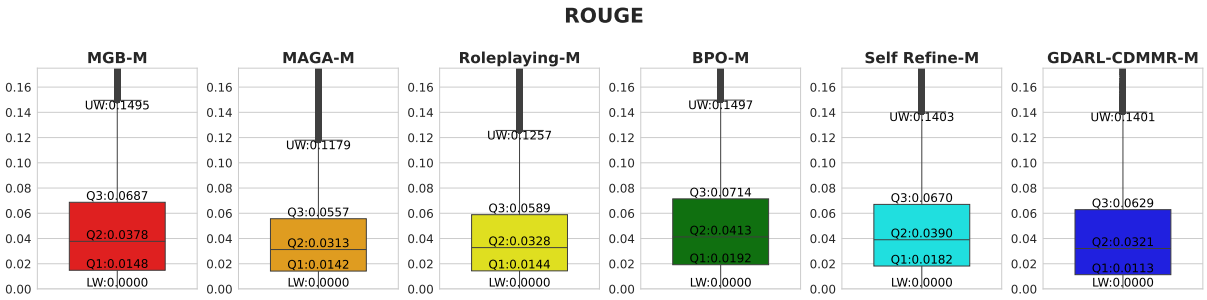

Figure 17: box-rouge

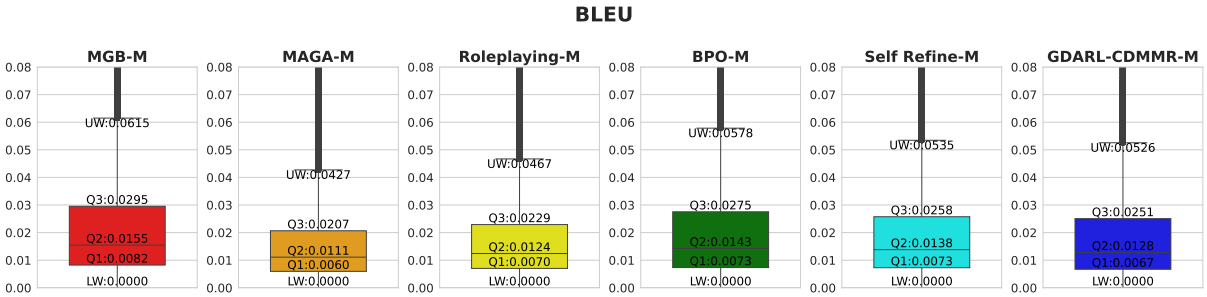

Figure 18: box-bleu

3427

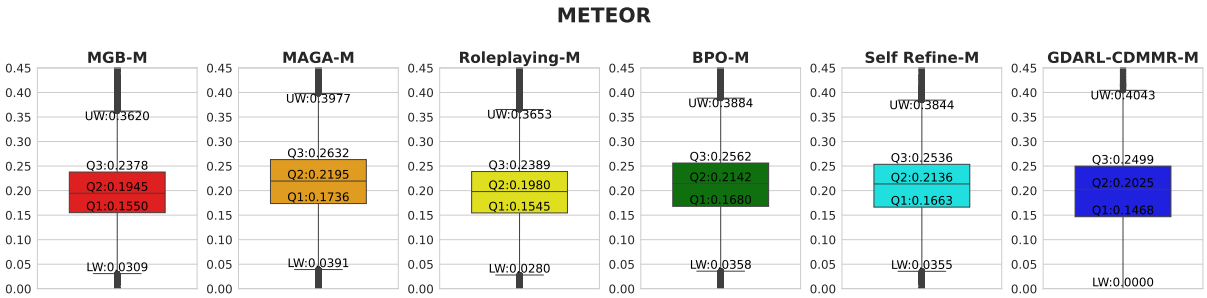

Figure 19: box-meteor

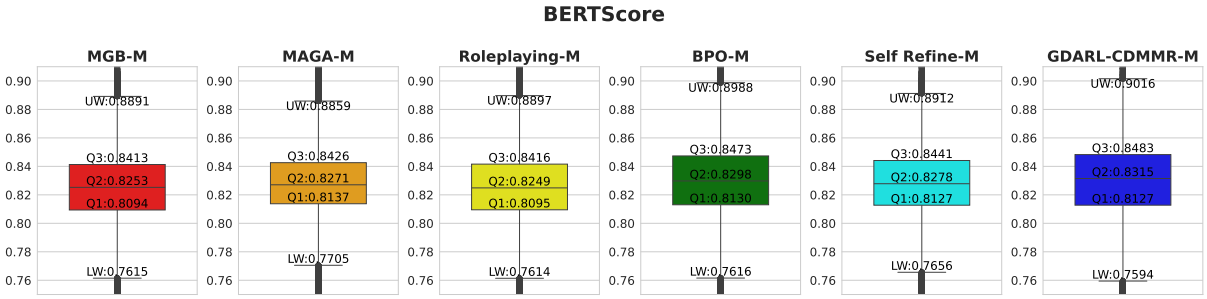

Figure 20: box-bertscore

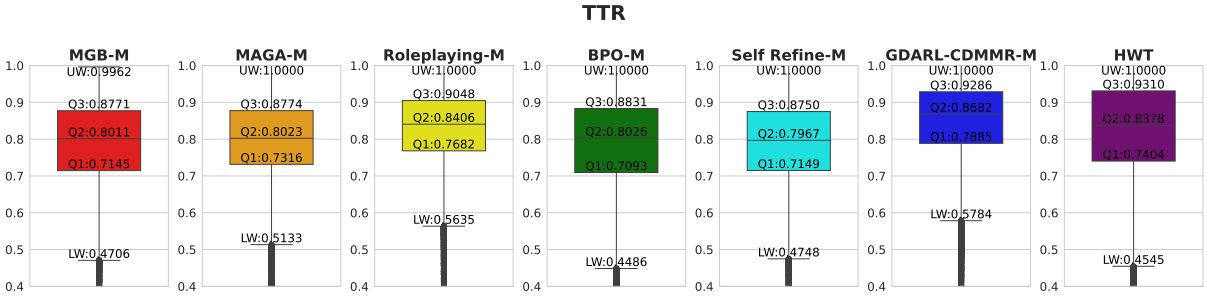

Figure 21: box-ttr

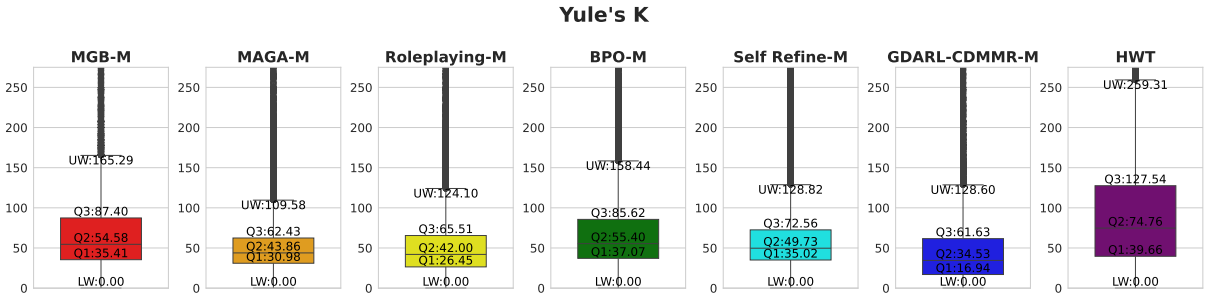

Figure 22: box-yules'k

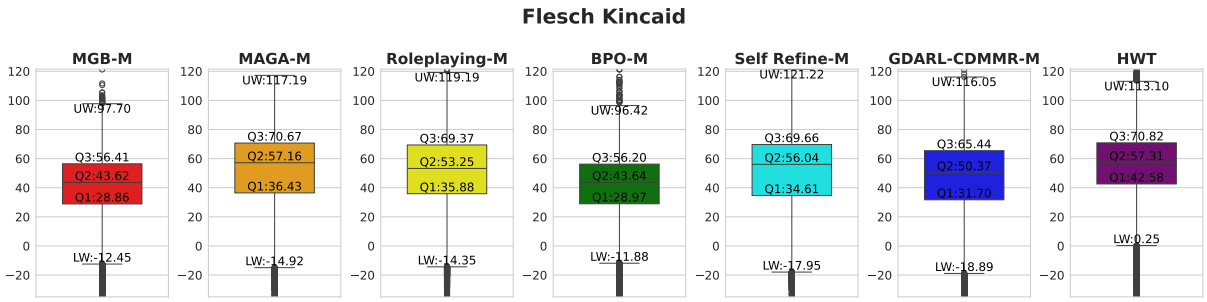

Figure 23: box-flesch-kincaid

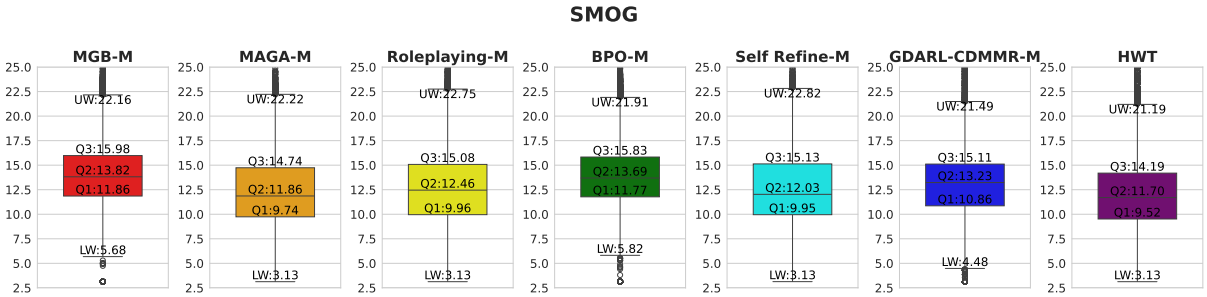

Figure 24: box-smog

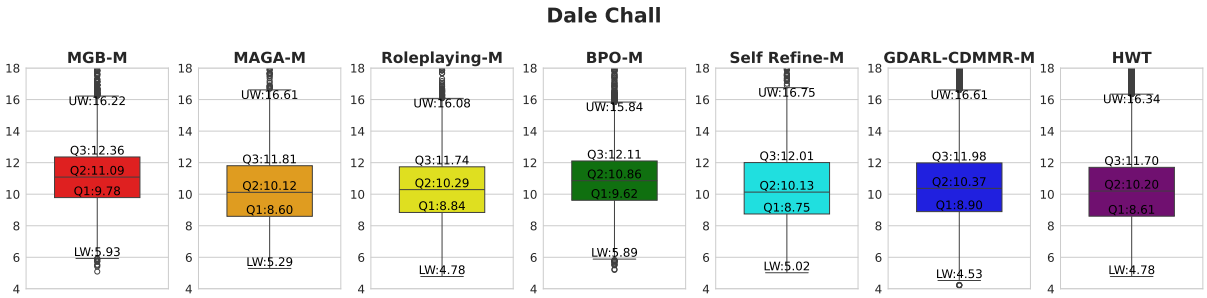

Figure 25: box-dale-chall

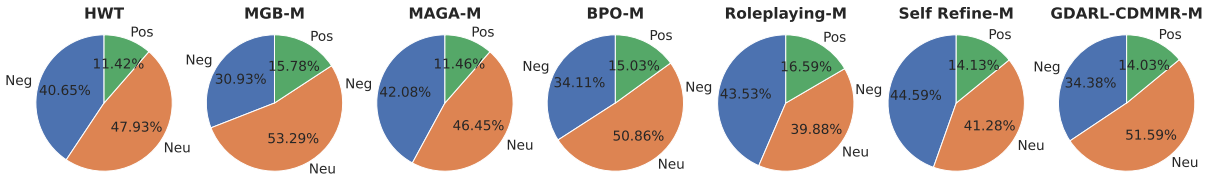

Figure 26: sentiment-detail

	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 30: MAGA-extra-CDMMR 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 31: 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 32: 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 33: 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	93.68	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	94.63	90.58
R-B MAGA (Ours)	87.36 (1.86↑)	**98.72** (26.33↑)	74.79 (-25.20↓)	**98.87** (4.24↑)	**95.41** (4.83↑)

Table 34: 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	98.60	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	80.96	83.31	76.08	86.76	79.65
LLMDet	55.33	47.82	70.92	74.89	53.16
Binoculars	**82.36**	75.03	97.58	89.14	**84.10**
DALD	80.44	83.93	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 35: 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**	47.31	87.30	63.79	**58.89**
DALD	59.82	47.23	82.07	64.96	56.73
GECScore	41.10	7.79	**100.00**	53.67	40.00
R-B MGB (Ours)	44.98	15.44	97.22	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 36: 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**	93.94	94.77	98.45	94.59
R-L GPT2	89.39	69.32	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	99.76	**97.19**
GECScore	81.74	42.86	**100.00**	90.18	86.96
R-B MGB (Ours)	87.09	60.23	99.70	93.14	91.68
R-B MAGA (Ours)	94.47 (7.38↑)	87.12 (26.89↑)	97.92 (-1.78↓)	98.14 (5.00↑)	94.06 (2.38↑)

Table 37: 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 38: MAGA-cn-extra-GDARL-CDMMR 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 39: 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 40: 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 41: MAGA-cn-extra-self-refine Bench.

	ACC	TPR	TNR	AUC	ACC(@FPR=5%)
R-B CGPT CN	95.22	92.30	98.10	99.24	95.19
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	99.63	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 42: 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	94.58	91.86	97.61	98.95	94.29
R-B MPU zhv3	88.08	82.85	93.92	95.40	87.09
Binoculars	93.71	93.58	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	99.08	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 43: 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	97.37	97.02	92.21
GECScore	56.46	0.00	**100.00**	47.94	56.64
R-B MGB CN (Ours)	95.01	94.22	95.62	99.04	94.90
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 44: 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.
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 45: 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 accomadate 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-GDARL-CDMMR-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	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.
extra-self-refine-M	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?
MAGA-M	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.

Table 46: A case study demonstrating a problem that other methods struggle with (highlighted in orange), but which only GDARL can effectively solve. Human-like traces are highlighted in green.

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	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!室友们纷纷建言献策，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？然后这个哥们拿出了一个星球杯，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，为了让我不偷懒，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。但那是因为我在练习呀！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。
MGB-M	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是超乎想象，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。

extra-BPO-M	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剧时，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。
extra-roleplaying-M	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，得看您自己亲身体验了！ 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?
extra-self-refine-M	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！
MAGA-M	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大的，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，顿时觉得生活又有资有味了。 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?

Table 47: A Chinese case study, similar to the previous example.

MGB_train.jsonl

id: 044783ff-8b83-4522-b666-bef847a5fd30
title: My PS4 screen is suddenly dim. Any suggestions?
text: I have a Samsung 4K tv. It's played my PS4 fine for years. Suddenly the screen is dark today. It did this a few months ago but it just restarted and it was fine. I've restarted, changed settings and settings on my tv but no change. Anyone found a solution? Have you replaced the HDMI port and it fixed the issue?
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: None
system_prompt: None
user_prompt: None
model: human
label: 0
temperature: None
top_p: None
top_k: None
repetition_penalty: None

Figure 27: sample-human.

MGB_train.jsonl

id: b49b0d46-5754-46e2-9c2f-dfcad6607f4c
title: My PS4 screen is suddenly dim. Any suggestions?
text: Here is the body of the Reddit post:\n\nHey everyone, I'm reaching out in hopes that someone can help me troubleshoot this issue. I've had my PS4 for a few years now without any major problems, but just today, the screen went suddenly dim. It's still playable, but the brightness is significantly reduced, making it hard to enjoy games, especially those with darker scenes.\n\nHere are some details that might help:\n\n* **TV**: I'm using a Samsung Smart TV (model NU7100), and the issue is only happening when the PS4 is connected. I've checked other inputs, and they look fine.\n* **Cables**: I've tried swapping out the HDMI cable with a different one (HDMI 2.0, 4K compatible) to rule that out, but the problem remains.\n* **Settings**: I've checked both the TV and PS4 settings. The TV's brightness settings are as they were before the issue started, and the PS4's output settings are set to automatic, which has always worked before.\n* **Power Cycle**: I've tried turning both the PS4 and TV off and on several times, but that hasn't resolved the issue.\n\nDoes anyone have any suggestions on what could be causing this sudden dimming? Could it be a PS4 hardware issue, a setting that somehow changed itself, or something else entirely? I'm all ears for any advice or similar experiences. Thanks in advance for your help!
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: 45a50840-d6cd-4446-91aa-b5a559394d47
system_prompt: None
user_prompt: Write just the body of a Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?". Do not repeat the title.
model: Llama-3.1-8B-Instruct
label: 1
temperature: 0.6
top_p: 0.9
top_k: -1
repetition_penalty: 1.0

Figure 28: sample-mgb-m.

MAGA_extra_train_BPO.jsonl

id: 97d8a37a-e48b-44d8-b8dc-3ba4912f27fe
title: My PS4 screen is suddenly dim. Any suggestions?
text: Hey everyone, I recently encountered an issue where my PS4 output has become significantly dimmer than usual without any manual changes to my settings. If you have experienced this or have technical expertise regarding display output issues, I would appreciate your input on the following troubleshooting steps and potential causes.\n\nPotential Causes: An HDMI Handshake Error which is a communication lapse between the console and the display. HDR Misconfiguration where incorrect metadata is sent during High Dynamic Range playback. Cable Degradation or physical wear on the HDMI cable affecting signal integrity. Hardware Failure involving potential issues with the PS4 APU or the HDMI transmitter chip.\n\nTroubleshooting Steps: First, Power Cycle the Hardware by performing a full cold boot. Turn off the console and the monitor or TV, unplug both from the power source for 60 seconds, and then restart. Second, Toggle HDR Settings by navigating to Settings, then Sound and Screen, then Video Output Settings. Try setting HDR and Deep Color Output to Off to see if the brightness normalizes. Third, Adjust RGB Range in the same menu. Change RGB Range from Automatic to Limited or Full. A mismatch between the PS4 and your monitors black levels often causes a washed out or overly dark image. Fourth, perform an HDCP Handshake Reset by disabling HDCP in the System settings temporarily to determine if the encryption protocol is causing a dimming effect. Fifth, Check the Display Mode to ensure your TV has not automatically switched to Eco Mode or Cinema Mode, which significantly lower the backlight.\n\nHardware Checks: Switch HDMI Ports to test the console on a different input on your TV. Replace the Cable using a High Speed HDMI 2.0 cable to rule out signal bottlenecking. Test on Another Display by connecting the PS4 to a different monitor to isolate whether the issue lies with the console or the original screen.\n\nHas anyone successfully resolved this without needing a full factory reset or database rebuild in Safe Mode? Looking forward to your suggestions.
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: ae98b8fb-0976-4d75-b2b3-362bf335a0d9
system_prompt: None
user_prompt: Write the body of a Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?" Provide troubleshooting steps, possible causes, and suggestions for brightening the screen. Do not repeat the title.
model: Llama-3.1-8B-Instruct
label: 1
temperature: 0.6
top_p: 0.9
top_k: -1
repetition_penalty: 1.0

Figure 29: sample-bpo-m.

MAGA_extra_train_GDARL_CDMMR.jsonl

id: b49b0d46-5754-46e2-9c2f-dfcad6607f4c
title: My PS4 screen is suddenly dim. Any suggestions?
text: Has anyone else run into this? I turned on my console today and the display is significantly darker than usual. It's not totally black—I can still see the UI and gameplay—but it looks like the brightness has been cranked down to about 20%.\n\nI've already tried the obvious steps like swapping the HDMI cable and trying a different port on the TV. I also checked the TV settings to make sure it didn't spontaneously switch to an Eco Mode, and I've restarted both the PS4 and the TV multiple times. Everything else plugged into the TV looks perfectly fine, so the issue seems to be coming directly from the console.\n\nIs this a sign of a failing GPU, or is there a specific display setting I might have accidentally toggled? Any suggestions would be a huge help!
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: 92a493b8-eaad-47b8-8271-8447942dd73f
system_prompt: None
user_prompt: Write just the body of a Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?". Do not repeat the title.
model: Llama-3.1-8B-Instruct
label: 1
temperature: 0.6
top_p: 0.9
top_k: -1
repetition_penalty: 1.0

Figure 30: sample-gdarl-CDMRmmr-m.

MAGA_extra_train_role_playing.jsonl

id: 16df1d41-8434-4b13-86d6-66de0050cbe7
title: My PS4 screen is suddenly dim. Any suggestions?
text: I've run into a bit of a technical anomaly with my PS4. Out of nowhere, the display output has become significantly dimmer than usual. I haven't changed any of my physical hardware settings, so it's a bit of a head-scratcher.\n\nI've already performed a basic power cycle, but the luminance levels are still looking pretty low. Before I start deep-diving into the system's safe mode or swapping out HDMI cables, I figured I'd consult the collective wisdom here.\n\nHas anyone dealt with this before? I'm wondering if it's a known firmware bug, a handshake issue with the HDR settings, or if my console's GPU might be starting to sunset. Any troubleshooting steps or insights would be greatly appreciated!
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: d19f3b6b-22cf-4387-a5a9-0dca23ade5ab
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 "My PS4 screen is suddenly dim. Any suggestions?". Do not repeat the title.
model: Llama-3.1-8B-Instruct
label: 1
temperature: 0.6
top_p: 0.9
top_k: -1
repetition_penalty: 1.0

Figure 31: sample-roleplaying-m.

MAGA_extra_train_self_refine.jsonl

id: 1f188cc0-a57d-4252-89c6-bacba78390fe
title: My PS4 screen is suddenly dim. Any suggestions?
text: It was working perfectly fine one minute, then just... dimmed. It's not totally black, but it looks like someone cranked the opacity down to 10% or threw a heavy gray veil over the screen. It's making anything even slightly dark basically unplayable.\n\nI'm honestly stumped. I checked my TV settings first, but my Switch and PC look totally normal on the same screen, so it's definitely a PS4 thing. I've already messed around with the RGB Range and Deep Color settings in the menu, and I even swapped out the HDMI cable for a brand new one, but no luck.\n\nI'm really praying it's not a hardware issue or the GPU dying on me. Has anyone else run into this? If there's some weird setting I'm missing or a specific "reset" trick that actually works, please let me know.\n\nTL;DR: PS4 screen went super dim out of nowhere. Swapped cables and checked settings, but nothing is fixing it. Help!
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: 45a50840-d6cd-4446-91aa-b5a559394d47
system_prompt: None
user_prompt: Write just the body of a Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?". Do not repeat the title.
model: Llama-3.1-8B-Instruct
label: 1
temperature: 0.6
top_p: 0.9
top_k: -1
repetition_penalty: 1.0

Figure 32: sample-self-refine-m.

self refine detail

Original Generation: It happened out of nowhere. I was playing normally, and suddenly the brightness just dropped. It ‘s not completely black, but it looks like the backlight is at 10% or there’s a heavy gray tint over everything.\n\nI ‘ve already tried checking the in-game brightness settings and my TV’s picture mode, but everything there seems fine. Other devices plugged into the same TV look perfectly normal, so I ‘m starting to think it ‘s a console or cable issue.\n\nHere is what I’ve tried so far:\n\nSwapping the HDMI cable.\n\nPower cycling both the TV and the PS4.\n\nChecking the "Deep Color Output" and "RGB Range" settings in the PS4 Sound and Screen menu.\n\nHas anyone dealt with this before? Could it be a failing HDMI port, or is my GPU on its way out? Any help would be appreciated!
Feedback Prompt: 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.
Feedback: The previous draft was structurally sound but a bit too "tech support ticket" in its delivery. It leans slightly robotic due to the formal bullet points and the "Here is what I’ve tried so far" lead-in.\n\nTo make it sound more like a natural Reddit post, consider these **concise improvements**:\n* **Loosen the syntax:** Use more casual phrasing like "It was fine one minute, then just... dimmed" or "I’m stumped."\n* **Ditch the formal list:** Instead of a bulleted "What I’ve tried" section, weave those details into a conversational paragraph (e.g., "Tried the usual stuff like swapping cables and messing with the RGB range, but no luck").\n* **Add "Human" filler:** Use brief observations or expressions of frustration, like "It ‘s making games like *Elden Ring* basically unplayable" or "Praying it’s not the GPU."\n* **Softened formatting:** Redditors often use "Edit:" or "TL;DR:" rather than structured bold headers to organize thoughts.
Refiner Prompt: Please improve your Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?" to make it more conversational and engaging. Here are some specific suggestions:\nThe previous draft was structurally sound but a bit too "tech support ticket" in its delivery. It leans slightly robotic due to the formal bullet points and the "Here is what I’ve tried so far" lead-in.\n\nTo make it sound more like a natural Reddit post, consider these **concise improvements**:\n* **Loosen the syntax:** Use more casual phrasing like "It was fine one minute, then just... dimmed" or "I’m stumped."\n* **Ditch the formal list:** Instead of a bulleted "What I’ve tried" section, weave those details into a conversational paragraph (e.g., "Tried the usual stuff like swapping cables and messing with the RGB range, but no luck").\n* **Add "Human" filler:** Use brief observations or expressions of frustration, like "It ‘s making games like *Elden Ring* basically unplayable" or "Praying it’s not the GPU."\n* **Softened formatting:** Redditors often use "Edit:" or "TL;DR:" rather than structured bold headers to organize thoughts.
Final Generation: It was working perfectly fine one minute, then just... dimmed. It ‘s not totally black, but it looks like someone cranked the opacity down to 10% or threw a heavy gray veil over the screen. It ‘s making anything even slightly dark basically unplayable.\n\nI ‘m honestly stumped. I checked my TV settings first, but my Switch and PC look totally normal on the same screen, so it ‘s definitely a PS4 thing. I ‘ve already messed around with the RGB Range and Deep Color settings in the menu, and I even swapped out the HDMI cable for a brand new one, but no luck.\n\nI ‘m really praying it ‘s not a hardware issue or the GPU dying on me. Has anyone else run into this? If there’s some weird setting I ‘m missing or a specific "reset" trick that actually works, please let me know.\n\nTL;DR: PS4 screen went super dim out of nowhere. Swapped cables and checked settings, but nothing is fixing it. Help!

Figure 33: sample-self-refine-detail.

MAGA_train.jsonl

id: cb3f89e0-a531-4f5e-a2c3-f7605ee67178
title: My PS4 screen is suddenly dim. Any suggestions?
text: I was right in the middle of a game and my PS4 screen straight up went dim for no reason. It's super annoying because I can still barely see the icons, but it's like the brightness is stuck at five percent. Total buzzkill.\n\nI already messed with the TV and swapped the HDMI cord, but no luck. Maybe try digging into the Sound and Screen settings? I also heard HDR can lowkey glitch out and make everything look dark, so toggling that off might work. Someone also mentioned Deep Color Output can be a literal nightmare for some setups.\n\nHonestly, I'm worried the HDMI port is just giving up on life. I'm gonna try a hard reset and leave it unplugged for a bit to see if it just needs a nap. Has anyone else dealt with this lately, or is my console finally cooked?
domain: Reddit
human_source_id: 044783ff-8b83-4522-b666-bef847a5fd30
prompt_id: be97b50b-6098-42ce-8238-e0d614ff81d7
system_prompt: You are American. Please respond in a casual, conversational American English style.
user_prompt: Write the body of a Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?" Provide troubleshooting steps, possible causes, and suggestions for brightening the screen. Do not repeat the title.
model: Llama-3.1-8B-Instruct
label: 1
temperature: 0.6
top_p: 0.9
top_k: -1
repetition_penalty: 1.0

Figure 34: sample-maga-m.

maga self refine detail

Original Generation: I ran into a weird issue today where my PS4 screen just went super dim out of nowhere and I'm trying to figure out what's going on. It's almost like the backlight just gave up, but I can still technically see what's happening on the dashboard.\n\nI've already checked the obvious stuff like the TV's brightness settings and swapped out the HDMI cable just in case it was a handshake issue, but nothing seems to be sticking. One thing you might want to try if this is happening to you is disabling the HDR settings in the console's Sound and Screen menu, since sometimes that metadata gets wonky and messes with the levels. Another possibility is that the "Deep Color Output" is acting up, so toggling that off might bring the light back.\n\nIt could also be a hardware quirk where the HDMI port is starting to fail or the TV itself has a power-saving mode that accidentally kicked in. If none of that works, you could try a full power cycle by unplugging everything for a few minutes to let the capacitors drain. If anyone has dealt with this before and found a weird fix, please let me know because playing games in the dark isn't exactly the vibe I was going for today.
Feedback Prompt: 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.
Feedback: To make the post sound more like a genuine Reddit user and less like an AI assistant, here are a few concise suggestions:\n* **Loosen the structure:** The transition between troubleshooting steps feels a bit too organized. Using phrases like "I also heard..." or "Maybe try..." helps it feel less like a manual.\n* **Add more "filler" or slang:** Incorporating casual American markers like "lowkey," "straight up," or "super annoying" would break the polished feel.\n* **Vary the sentence length:** Most sentences in the post are roughly the same length. Mixing in a few very short, punchy thoughts (e.g., "Total buzzkill.") adds a more human rhythm.\n* **Show more frustration:** Real Reddit posts usually have a bit more "personality" or emotion, like mentioning how it happened right in the middle of a specific game.\n* **End with a specific "Ask":** Instead of a general request for help, asking something like "Is my console finally cooked?" feels more like a community interaction.
Refiner Prompt: Please improve your Reddit post titled "My PS4 screen is suddenly dim. Any suggestions?" to make it more conversational and engaging. Here are some specific suggestions:\nTo make the post sound more like a genuine Reddit user and less like an AI assistant, here are a few concise suggestions:\n* **Loosen the structure:** The transition between troubleshooting steps feels a bit too organized. Using phrases like "I also heard..." or "Maybe try..." helps it feel less like a manual.\n* **Add more "filler" or slang:** Incorporating casual American markers like "lowkey," "straight up," or "super annoying" would break the polished feel.\n* **Vary the sentence length:** Most sentences in the post are roughly the same length. Mixing in a few very short, punchy thoughts (e.g., "Total buzzkill.") adds a more human rhythm.\n* **Show more frustration:** Real Reddit posts usually have a bit more "personality" or emotion, like mentioning how it happened right in the middle of a specific game.\n* **End with a specific "Ask":** Instead of a general request for help, asking something like "Is my console finally cooked?" feels more like a community interaction.
Final Generation: I was right in the middle of a game and my PS4 screen straight up went dim for no reason. It's super annoying because I can still barely see the icons, but it's like the brightness is stuck at five percent. Total buzzkill.\n\nI already messed with the TV and swapped the HDMI cord, but no luck. Maybe try digging into the Sound and Screen settings? I also heard HDR can lowkey glitch out and make everything look dark, so toggling that off might work. Someone also mentioned Deep Color Output can be a literal nightmare for some setups.\n\nHonestly, I'm worried the HDMI port is just giving up on life. I'm gonna try a hard reset and leave it unplugged for a bit to see if it just needs a nap. Has anyone else dealt with this lately, or is my console finally cooked?

Figure 35: sample-maga-self-refine-detail.