Reasoning-Aware AIGC Detection via Alignment and Reinforcement

Zhao Wang[1]*, **Max Xiong**[2]*, **Jianxun Lian**[3], **Zhicheng Dou**[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China,
[2]Duke University, [3]Microsoft Research Asia,
lilin22wz@gmail.com, jianxun.lian@outlook.com, dou@ruc.edu.cn

Abstract

The rapid advancement and widespread adoption of Large Language Models (LLMs) have elevated the need for reliable AI-generated content (AIGC) detection, which remains challenging as models evolve. We introduce AIGC-textbank, a comprehensive multi-domain dataset with diverse LLM sources and authorship scenarios, and propose REVEAL, a detection framework that generates interpretable reasoning chains before classification. Our approach uses a two-stage training strategy: supervised fine-tuning to establish reasoning capabilities, followed by reinforcement learning to improve accuracy, improve logical consistency, and reduce hallucinations. Extensive experiments show that REVEAL achieves state-of-the-art performance across multiple benchmarks, offering a robust and transparent solution for AIGC detection. The project is open-source at https://aka.ms/reveal

1 Introduction

The rapid advancement of Large Language Models (LLMs) has ushered in an era where AI-generated content (AIGC) is increasingly pervasive and often indistinguishable from human writing. As models approach human-level fluency and coherence (Achiam et al., 2023), the ability to reliably discern machine-authored text becomes critical for maintaining integrity across numerous domains. Beyond academic publishing—where undisclosed AIGC threatens to undermine the authenticity of research papers and peer reviews (Perkins, 2023; Su et al., 2023)—AIGC detection is equally crucial in domains like telecommunications fraud prevention, where malicious actors deploy AI to impersonate humans (Ciancaglini et al., 2020). A robust, reliable AIGC detector thus serves as an essential safeguard, enabling verification of authorship and upholding trust in digital communications.

Existing approaches to AIGC detection have largely relied on statistical classifiers (Solaiman et al., 2019; Lavergne et al., 2008) or black-box neural models (Liu et al., 2019), which often exploit surface-level patterns and struggle to generalize as LLMs evolve (Guo et al., 2023). While benchmarks such as M4 (Wang et al., 2024) and LOKI (Ye et al., 2025) have broadened the scope of evaluation, their data scale remains limited compared to real-world requirements and often fails to include outputs from the latest state-of-the-art models. In this work, we aim to consolidate and advance the field of AIGC detection by introducing a more comprehensive benchmark and a reasoning-driven detector that generalizes effectively to evolving generative technologies.

To support this goal, we construct **AIGC-textbank**, a large-scale, multi-domain dataset that includes authentic human writing, fully machine-generated (AI-Native) text, and human-authored text polished by AI (AI-Polish). Our corpus is sourced from 10 diverse domains and generated using 12 different LLMs—including the latest proprietary and open-weight models—providing a realistic and challenging testbed for detecting authorship in both pure and hybrid scenarios. Its parallel structure ensures that each human reference is paired with AI-generated counterparts, enabling controlled comparisons and finer-grained analysis.

We further introduce **REVEAL** (Reasoning-Enhanced Verification and Evaluation for AI Language), a novel framework that shifts detection from opaque classification to transparent, reasoning-based decision-making. REVEAL is trained in two stages: first, supervised fine-tuning (SFT) initializes the model by distilling concise and effective rationales from OpenAI o3, serving as an imitation learning phase; then, reinforcement learning (RL) extends these capabilities by refining reasoning chains to improve logical consistency, reduce hallucinations, and ultimately surpass the

*Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20813–20829

teacher model's performance. By explicitly modeling the *Think-then-Answer* process, our approach not only achieves high accuracy but also provides interpretable evidence for each prediction. Extensive experiments across five benchmarks demonstrate that REVEAL outperforms existing black-box detectors and general-purpose LLMs in both binary and fine-grained settings, while maintaining strong generalization under domain shift and adversarial challenges.

In summary, our contributions are threefold:

- We construct and will release **AIGC-text-bank**, a large-scale, multi-domain dataset featuring state-of-the-art LLM outputs, providing a comprehensive training resource as well as a benchmark for AIGC detection research.
- We propose **REVEAL**, a reasoning-driven detection framework that combines SFT and RL to produce accurate and interpretable authorship judgments.
- We conduct extensive experiments showing that our method sets a new state of the art in generalization and fine-grained discrimination, providing a trustworthy foundation for real-world AIGC detection.

2 Methodology

As LLMs rapidly advance, traditional AIGC detectors relying on superficial statistical cues (e.g., log-likelihood (Solaiman et al., 2019), entropy (Lavergne et al., 2008)) become increasingly inadequate, especially in complex, real-world scenarios like AI-Polished content. To address this, we pursue three goals: developing an *interpretable* LLM-based detector that distinguishes AI from human text with reasoning; extending detection to differentiate *AI-Native*, *AI-Polished*, and *Human content*; and enabling predictive *uncertainty estimation*. Our methodology constructs a comprehensive, multi-scenario dataset and employs a two-stage training framework of supervised fine-tuning and reinforcement learning to build a detector capable of robust, human-readable reasoning.

2.1 Dataset Construction

Existing datasets for AIGC detection often suffer from two critical limitations: they fail to include content generated by the latest state-of-the-art LLMs (e.g., GPT-5), and they overlook the nuanced paradigm of human-AI collaborative writing. To bridge this gap, we construct **AIGC-text-bank**, a comprehensive multi-domain and multi-LLM dataset designed to enhance detector capabilities in real-world human-AI text discrimination. It is structured as a parallel corpus where each human-written document is paired with corresponding AI-generated counterparts. As illustrated in Figure 1, the construction pipeline encompasses three distinct text categories: authentic human writing, fully AI-generated text (**AI-Native**), and human-authored text refined by AI (**AI-Polish**).

Table 1: Statistics of the AIGC-text-bank dataset.

	Samples	Total Tokens	#.LLMs
Human	66,979	22,535,085	-
AI-Native	699,052	195,392,025	12
AI-Polish	732,248	205,644,529	12

Human Data Collection To establish a robust human baseline, we collect 66,979 authentic human-written documents across 10 diverse domains, including academic papers, social discussions, encyclopedic entries and literature. This diversity ensures extensive coverage of various linguistic styles and structural formats. To mitigate the risk of inadvertently including AI-generated text, we source documents published strictly before the release of ChatGPT (November 30, 2022), thereby establishing a temporal cutoff prior to the widespread public use of advanced, human-like language models. More details about the human subset are provided in the Appendix A.1.

Generator Models To mitigate architectural inductive biases and capture a broad stylistic spectrum, we employ a diverse ensemble of LLMs varying in parameter scales and performance profiles. Specifically, our generator pool includes state-of-the-art proprietary models (e.g., GPT-5, Grok-4), representative open-source models (e.g., DeepSeek R1, Llama 3.3, Qwen 3, and Phi-4) and legacy models (e.g., GPT-2) to capture the evolutionary trajectory of generative styles. These models serve as the backbone for synthesizing both AI-Native and AI-Polish subsets. We provide the full list of models and details in Appendix A.2.

AI-Native Generation We propose a semantically aligned reconstruction pipeline to separate intrinsic linguistic signatures from surface-level topical differences. To ensure the synthetic content

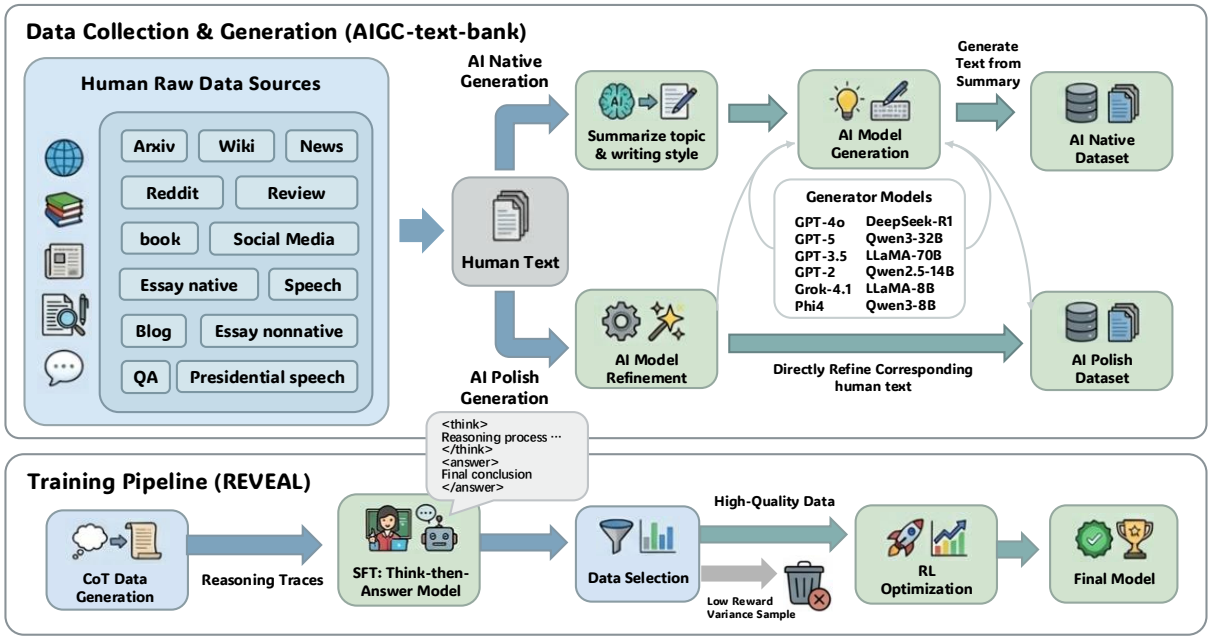

Figure 1: The Overall Framework

remains grounded in real-world contexts, we employ GPT-4o to extract structured meta-attributes from the human reference corpus. For each document, GPT-4o distills a concise thematic summary (e.g., topic, key points) and, where pertinent, a profile of its linguistic style (e.g., formal, narrative, conversational). This transformation preserves the domain diversity of the original data while providing a controlled framework for synthesis. Leveraging these meta-attributes, we task the aforementioned 12 generator models with producing content that adheres strictly to the specified topics and writing styles. In this process, we implement two strategic constraints to ensure both the fairness and the complexity of the dataset. First, we strictly align the output length with human references to eliminate length as a confounding variable. Furthermore, to simulate real-world scenarios and increase the classification difficulty, we introduce a prompt-based intervention on 20% of the data samples. In these cases, generators are instructed to imitate human writing styles, making the AI-Native subset both length-matched and more challenging to distinguish stylistically. Detailed data distributions across models are provided in Appendix A.3.

AI-Polish Generation To address the realistic and nuanced scenario of human-AI collaborative writing—where, in contexts like academic writing, using AI to polish a human-authored draft is often permissible, whereas generating content directly with AI may constitute a violation of integrity—we introduce the AI-Polish subset. This category consists of human-authored texts refined by our generator models to improve fluency and style while strictly preserving the original semantic intent and logical structure. Thus, while the surface presentation may bear AI stylistic signatures, the core ideas remain human-originated, substantively differing from AI-Native content. This subset provides a more challenging detection benchmark, requiring the identification of subtle machine interventions within largely human documents. Examples and data distributions are provided in Appendix A.4.

2.2 Reasoning Initialization via SFT

Conventional detection models treat the task as a discriminative classification problem, often relying on superficial statistical cues or opaque black-box neural models. We argue that a robust, generalizable, and trustworthy detector should instead articulate a human-readable reasoning process before making a decision, rendering the classification transparent and grounded. To this end, we adopt a Think-then-Answer paradigm, which requires the model to base its final verdict on explicit reasoning and concrete evidence extracted from the text.

Since the dataset constructed in Section 2.1 contains only category labels without explanatory rationales, we leverage the advanced reasoning capabilities of a state-of-the-art LLM, OpenAI o3, to augment it with high-quality reasoning trajec-

tories. As illustrated in Figure 1, we employ a hindsight analysis strategy: instead of asking the teacher model to predict the label (which could amplify errors), we provide o3 with the input text $\boldsymbol{x}$ and its ground-truth label $\boldsymbol{y}$. The model is then instructed to reconstruct a plausible decision-making process, explicitly articulating why $\boldsymbol{x}$ belongs to category $\boldsymbol{y}$. For example, given an AI-polished text, o3 is prompted to pinpoint the subtle tensions between the underlying human-authored logic and the surface-level AI stylistic artifacts.

This yields a reasoning-augmented dataset $\mathcal{D}_{\text{sft}}$. Each training instance is formatted as a sequence <think>$\boldsymbol{r}$</think><answer>$\boldsymbol{y}$</answer>, where $\boldsymbol{r}$ is the generated reasoning trace and $\boldsymbol{y}$ is the label. Direct fine-tuning on lengthy reasoning sequences can, however, disperse the model's attention away from the final prediction. To address this, we employ a **Outcome-Weighted Objective** that decouples the generation loss:

$$\begin{aligned}\mathcal{L}_{\text{sft}} = &-\sum_{i=1}^{m} \log P(r_i|\boldsymbol{x}, \boldsymbol{r}_{<i}) \\ &-\lambda \sum_{j=1}^{n} \log P(y_j|\boldsymbol{x}, \boldsymbol{r}, \boldsymbol{y}_{<j}),\end{aligned} \tag{1}$$

where $\lambda > 1$ is a coefficient that increases the weight of the answer loss. By emphasizing the second term, the model is encouraged to use the reasoning path $\boldsymbol{r}$ as supporting context, while focusing optimization on the final prediction $\boldsymbol{y}$.

2.3 Reasoning Refinement via RL

While SFT establishes an initial capacity for generating reasoning chains, the model remains prone to subtle hallucinations, reasoning-answer inconsistencies, and its capabilities are ultimately bounded by those of the teacher model used for data augmentation. To overcome these limitations and further refine the model's reasoning fidelity, we employ RL for direct preference alignment. This stage aims to reduce errors and improve the overall robustness and logical consistency of the generated reasoning.

To maximize the efficiency of RL training, we first construct a high-quality training set through variance-based data selection. After SFT, the model can confidently classify most samples in $\mathcal{D}_{\text{sft}}$, which provide minimal learning signal. We therefore focus on uncertain, borderline cases where the model's predictions are inconsistent. For each prompt x, we perform K stochastic rollouts using the SFT model and compute a binary correctness score $s_k \in \{0, 1\}$ for each rollout. The RL dataset $\mathcal{D}_{\text{rl}}$ is then constructed by selecting only those samples where the model exhibits prediction variance:

$$\mathcal{D}_{\text{rl}} = \left\{ x \in \mathcal{D}_{\text{sft}} \;\middle|\; 0 < \sum_{k=1}^{K} s_k(x) < K \right\}. \tag{2}$$

By excluding samples with deterministic success or failure, this filtering ensures RL training concentrates on high-uncertainty instances, thereby maximizing the informative gradient signal.

For policy optimization, we adopt the DAPO algorithm (Yu et al., 2025) which employs decoupled clipping thresholds to independently constrain policy updates from below and above, effectively preventing entropy collapse and promoting stable convergence. The objective is to maximize the expected return over groups of G sampled outputs:

$$\mathcal{J}(\theta) = \mathbb{E}_{\substack{x\sim\mathcal{D}_{\text{rl}} \\ \{r_i\}\sim\pi_{\theta_{\text{old}}}}} \left(\frac{1}{G} \sum_{i=1}^{G} \mathcal{L}_i \right), \tag{3}$$

where $\mathcal{L}_i$ is the decoupled clipped objective for the i-th sample:

$$\mathcal{L}_i = \min \Big(\rho_i \hat{A}_i, \text{clip}\big(\rho_i, 1-\epsilon_l, 1+\epsilon_h\big) \hat{A}_i \Big). \tag{4}$$

Here, $\rho_i = \pi_\theta(r_i|x)/\pi_{\theta_{\text{old}}}(r_i|x)$ is the importance sampling ratio, and ϵ_l, ϵ_h are lower and upper clipping hyperparameters. Following GRPO, the advantage $\hat{A}_i$ is computed by normalizing rewards within the sampled group, eliminating the need for a separate value function:

$$\hat{A}_i = \frac{R(r_i, y) - \mu_R}{\sigma_R}, \tag{5}$$

with μ_R and σ_R being the mean and standard deviation of rewards $\{R(r_j, y)\}_{j=1}^{G}$ for the group.

To effectively guide policy learning, we design a composite reward function $R(r, y)$ that balances final answer accuracy with the structural and logical quality of the reasoning chain:

$$R(r, y) = R_{\text{acc}}(r, y) + R_{\text{fmt}}(r) + R_{\text{cons}}(r, y). \tag{6}$$

The component $R_{\text{acc}} \in \{0, 1\}$ provides a primary outcome signal, awarding $+1$ for a correct final prediction. R_{fmt} acts as a hard constraint on output format, imposing a penalty of -1 if the required output structure is violated. Finally, R_{cons} is a fine-grained score provided by GPT-4o that evaluates both the internal logical coherence of the reasoning chain and its consistency with the final prediction, preventing the model from exploiting format rewards without genuine reasoning.

3 Experimental Setup

We design experiments to answer three core research questions:

RQ1: How does our reasoning-driven detector compare to traditional black-box methods and state-of-the-art LLMs on standard benchmarks?

RQ2: Can our model generalize robustly to unseen downstream detection tasks, particularly those featuring new or evolving label taxonomies?

RQ3: What is the contribution of each key component—the two-stage training, the weighted loss in SFT, and the data selection in RL—to the overall performance?

3.1 Dataset and Metrics

We use five diverse benchmarks to verify the accuracy and generalization capabilities of models:

(1) **AIGC-bench**: As detailed in Section 2.1, we utilize the held-out test set of our proposed dataset as one benchmark.

(2) **DetectRL** (Wu et al., 2024): A benchmark that includes multiple specific attack methods (e.g., perturbation attacks), which allows us to assess whether our reasoning framework maintains stability under malicious interference.

(3) **M4** (Wang et al., 2024): A large-scale multi-generator corpus covering diverse sources like Wikipedia and Reddit, serving as a standard baseline for distributional generalization.

(4) **Pan** (Bevendorff et al., 2025): Focuses on human-AI collaboration with a fine-grained 6-class taxonomy (e.g., *Human-written then Machine-polished*), representing complex mixed authorship.

(5) **LOKI** (Ye et al., 2025): A comprehensive benchmark encompassing broad text domains (e.g., news, creative writing) designed to evaluate detection capabilities in real-world scenarios.

For these datasets, we primarily use *Accuracy* and *Macro F1* as the metrics.

3.2 Tasks

To answer the research questions, we design two type of distinct tasks:

Task I: General Detection This task evaluates the model's reasoning performance under two protocols: (1) **Binary Classification:** For M4, LOKI, DetectRL, Pan, and AIGC-bench, we unify the label spaces into *Human* vs. *AI* to benchmark generalized detection capabilities; and (2) **Fine-grained Reasoning:** Exclusively on AIGC-bench, we conduct a 3-class classification task (*Human*, *AI-Native*, *AI-Polished*) to verify the model's sensitivity to subtle polishing artifacts.

Task II: Transfer Learning To evaluate our model's potential as a foundation for complex tasks, we initialize with our pre-trained weights and fine-tune the model on the target benchmarks. We evaluate on three tasks requiring high-level adaptation: (1) **M4 (Domain Adaptation):** Adapting the model to the specific distributions of the multi-generator M4 corpus for robust binary detection; (2) **DetectRL (Attack Identification):** Distinguishing between specific adversarial attack types (e.g., paraphrasing); and (3) **Pan (Collaborative Analysis):** Classifying the precise 6-class human-AI collaborative patterns.

3.3 Baseline

Discriminative Baseline: We select supervised RoBERTa-SFT (Liu et al., 2019) and three zero-shot detectors based on token probabilities: Fast-DetectGPT (Bao et al., 2023), Binoculars (Hans et al., 2024), and ImBD (Chen et al., 2025).

General LLMs with Reasoning Prompts: We evaluate representative proprietary and open-source general LLMs by using Think-then-Answer prompt. We employ GPT-5 (Singh et al., 2025), GPT-4o (Hurst et al., 2024), GPT-4o-mini (OpenAI, 2024b) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as baselines.

Reasoning LLMs: This category includes three strong reasoning models: OpenAI o3 (OpenAI, 2024c), QwQ-32B (Team, 2025), and Qwen3-8B (Yang et al., 2025).

3.4 Implementation Details

We implement our framework based on the HuggingFace Transformers (Wolf et al., 2020) and TRL (von Werra et al., 2020) Library, using Qwen3-8B as the backbone. For the Imitation Learning stage, we finetune the model on the constructed 24k reasoning dataset for 3 epochs with a global batch size of 128 and a learning rate of 1e-5. For Preference Alignment, we filter 10k high-uncertainty samples for training; during optimization, we sample 8 outputs per prompt with a temperature of 1.0 and update the policy with a learning rate of 1e-5. For the transfer learning experiments in Setting II, we initialize the model with our aligned weights and apply the same SFT configuration to adapt to downstream benchmarks. All experiments are conducted on four NVIDIA A100-80GB GPUs.

Table 2: Overall performance comparison on different benchmarks. The best results are in **bold** and the second are underlined.

Method	In-Domain		Out of Domain								Avg.	
	AIGC-bench		DetectRL		M4		Pan		LOKI			
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DISCRIMINATIVE BASELINE												
RoBERTa-SFT	**97.80**	**97.80**	93.50	93.50	73.10	72.25	85.50	85.50	**95.70**	91.94	89.12	88.20
ImBD	74.80	74.47	86.20	86.07	73.80	73.29	86.00	85.88	85.50	76.40	81.26	79.22
Binoculars	67.60	67.17	88.60	88.60	85.60	85.49	87.90	87.84	71.50	62.92	80.04	78.40
Fast-DetectGPT	63.90	62.81	83.90	83.88	78.10	77.96	82.60	82.48	57.80	51.75	73.26	71.78
GENERAL LLMS												
Llama3.1-8B	44.59	16.22	39.78	16.78	44.58	18.29	42.51	16.39	23.80	10.32	39.05	15.60
GPT-4o-mini	51.95	42.24	55.47	45.39	56.84	55.55	47.66	40.46	16.60	16.06	45.70	39.94
GPT-4o	52.34	46.28	57.06	52.31	58.12	58.12	48.63	47.44	27.93	27.81	48.82	46.39
GPT-5	72.64	70.17	96.52	96.52	**89.33**	**89.31**	82.57	82.57	89.45	82.18	86.10	84.15
REASONING LLMS												
Qwen3-8B	48.90	10.45	63.86	25.14	58.84	11.80	56.34	11.31	42.17	11.19	54.02	13.98
QwQ-32B	52.75	33.13	69.87	69.16	67.94	67.88	59.20	59.15	57.40	50.81	61.43	56.03
OpenAI o3	75.81	75.06	95.74	95.73	88.03	88.03	88.59	88.58	88.24	77.92	87.28	85.06
REVEAL (Ours)	96.30	96.30	**97.20**	**97.20**	77.86	77.25	**88.80**	**88.79**	95.60	**91.98**	**91.15**	**90.30**

4 Results

4.1 General Detection Performance (RQ1)

We report experimental results for binary detection (Table 2) and fine-grained classification (Table 3), demonstrating that REVEAL achieves state-of-the-art performance with superior stability. First, REVEAL effectively counters the prediction bias observed in smaller models like Llama 3.1 and Qwen3-8B, which show a large Accuracy–F1 gap by systematically misclassifying fluent AI text as human. Our model closes this gap, achieving metric alignment comparable to larger proprietary models. Second, REVEAL exhibits stronger generalization and robustness: while the supervised RoBERTa-SFT suffers a significant drop on out-of-distribution benchmarks (e.g., from 97.80% in-domain to 73.10% on M4), and zero-shot detectors (Fast-DetectGPT, Binoculars, ImBD) struggle to exceed an average accuracy of 81.3%, our model maintains consistently high cross-domain performance (averaging 91.15% overall). This indicates that the reasoning-driven paradigm captures more transferable characteristics of AI-generated text.

Distinguishing AI-polished content presents a particular challenge, as shown in Table 3. While strong proprietary models (GPT-5, OpenAI o3) perform near chance (48% accuracy), REVEAL attains 70.74% accuracy. The confusion matrix (Figure 2) reveals that GPT-5 exhibits a strong bias towards predicting the AI-Native class, failing to disentangle human logic from AI polish. In contrast, REVEAL demonstrates a more balanced prediction profile, confirming its capability to identify the stylistic artifacts of collaborative writing.

Table 3: Fine-grained reasoning results (3-class classification) on AIGC-bench.

Method	Acc	F1
GPT-5	48.30	41.00
OpenAI o3	47.49	38.62
REVEAL (Ours)	**70.74**	**70.99**

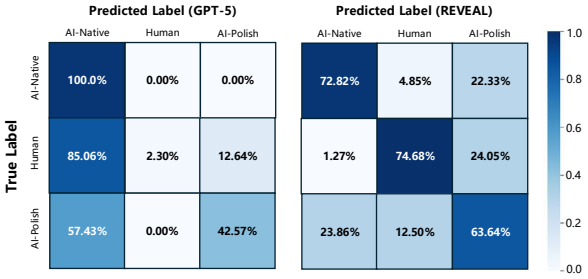

Figure 2: The confusion matrix of GPT-5 and REVEAL.

Table 4: Transfer Learning results.

Method	M4 2 classes	DetectRL 5 classes	Pan 6 classes
Qwen3-8B-SFT	96.61	74.80	45.93
OpenAI o3	89.53	63.69	36.75
REVEAL (Ours)	**97.33**	**75.20**	**49.07**

Table 5: Performance impact of removing individual components in REVEAL.

Method	In-Domain	OOD Avg.
REVEAL	**96.30**	**89.87**
w/o SFT	58.24	61.35
w/o RL	96.16	88.69
w/o Weighted	93.60	87.46
w/o Selection	96.18	89.31
w/o CoT	91.00	85.20

4.2 Transfer analysis (RQ2)

Table 4 presents the results of transfer learning experiments, where fine-tuning from REVEAL consistently outperforms initializing from both the standard Qwen3-8B-SFT baseline and the general-purpose OpenAI o3. Across three benchmarks with varying label taxonomies (2, 5, and 6 classes), REVEAL achieves the highest accuracies (e.g., 97.33% on M4 and 49.07% on the 6-class Pan dataset). This demonstrates that our model, pre-trained with reasoning-driven detection objectives, serves as a superior parameter initialization that effectively transfers to new tasks and complex, unseen label spaces.

4.3 Ablation Study (RQ3)

To evaluate the contribution of each component in our framework, we conduct an ablation study on the setting of binary classification task (the same setting with Table 2): (1) **w/o SFT**: removes the whole SFT phase (Section 2.2); (2) **w/o RL**: removes the whole RL phase (Section 2.3); (3) **w/o Selection**: replaces the uncertainty-based data filtering strategy with random sampling; (4) **w/o Weighted**: uses standard next-token prediction instead of the re-weighted loss objective. (5) **w/o CoT**: disables the reasoning process, forcing the model to predict the final label directly.

Table 5 presents the results of our ablation study, which validates the necessity of each component. The removal of the SFT phase (*w/o SFT*) severely degrades performance, as the model lacks an initial reasoning structure and struggles to converge efficiently during RL. While SFT establishes the reasoning format, the RL phase is crucial for refining it, as shown by the drop in OOD performance for *w/o RL*. The significant declines observed for *w/o Selection* and *w/o Weighted* highlight the importance of our data and optimization strategies: uncertainty-based filtering forces the model to learn from challenging samples, and the re-weighted loss ensures the final prediction remains the optimization focus. Finally, the substantial drop for *w/o CoT* confirms that explicit reasoning is essential, forcing the model to rely on logical derivation rather than spurious correlations.

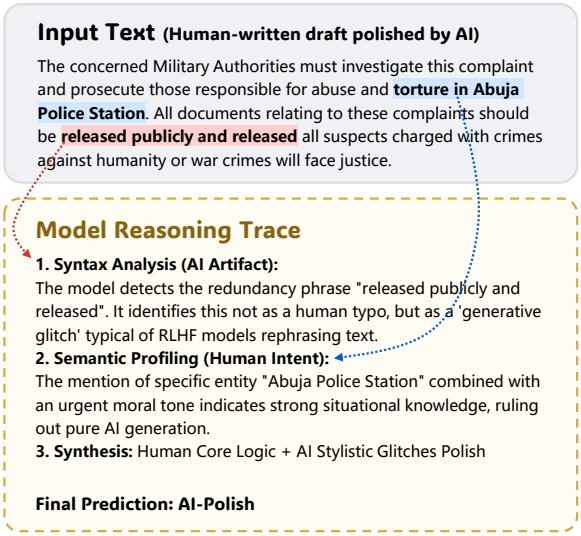

Figure 3: A case study on interpretability in reasoning

4.4 Case Study

Figure 3 illustrates how REVEAL distinguishes complex *AI-Polished* content by reasoning based on explicit linguistic evidence rather than statistic cues. In this case, the model correctly identifies the specific entity "Abuja Police Station" and the urgent moral tone as evidence of human authorship. Simultaneously, it flags the redundancy "released publicly and released" as a specific generative glitch of AI rephrasing rather than a typo. By weighing these conflicting signals, namely human core logic versus machine syntax, REVEAL derives a transparent and verifiable verdict.

4.5 Linguistic Analysis

Based on the reasoning process generated by REVEAL, we conduct a qualitative analysis to identify specific feature sets that distinguish human writing

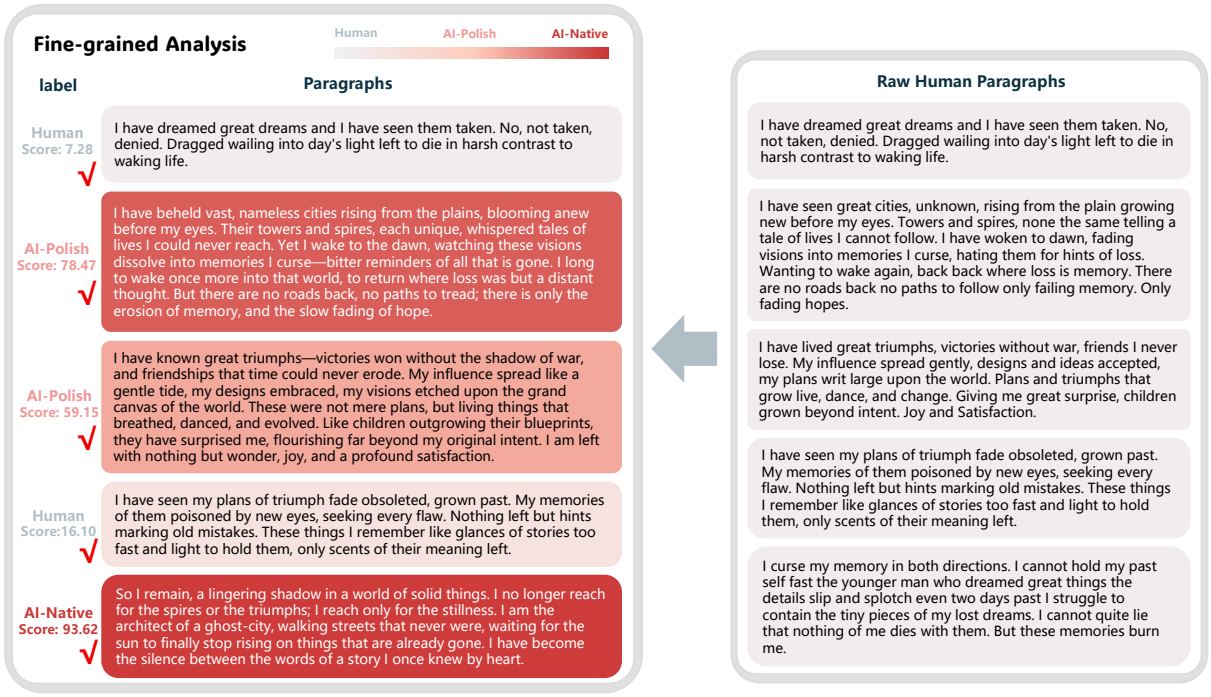

Figure 4: An example on block-wise detection

from AI-Native or AI-Polish content.

Human-Written: "Messy Reality" Human text is primarily defined by its spontaneity and lack of standardization. (1) **Mechanical Irregularities:** Human text frequently contains organic errors, such as comma splices, inconsistent capitalization, and colloquial abbreviations (e.g., "idk", "u"). Such patterns are rarely produced by LLMs without explicit prompting. (2) **Structural Fluidity:** Human narratives often lack rigid structure, exhibiting meandering thoughts, abrupt topic shifts, or sudden endings without formal conclusions. (3) **Hyper-Specificity and Emotion:** The Human content also includes unverifiable but vivid details (e.g., specific prices, distinct sensory descriptions) and raw, unfiltered emotions (anger, confusion).

AI-Native: "Flawless Vacuity" AI-Native text is defined by a high degree of polish but a low degree of specific semantic weight. (1) **Algorithmic Symmetry:** Sentences tend to be balanced in length and rhythm, while grammar and punctuation are invariably perfect. (2) **Template Adherence:** The text often follows a strict rhetorical structure (e.g., a balanced pros-and-cons list or a standard five-paragraph essay format) and overuses transitional phrases (e.g., "Furthermore," "In conclusion"). (3) **Generic Content:** The text relies on clichés, safe metaphors, and broad generalizations. Even when hallucinating facts or quotes, the AI tends to generate plausible but fundamentally generic statements that lack idiosyncratic character.

AI-Polished: "Hybrid" AI-Polished text is the most complex, as it combines human intent with algorithmic execution, making it difficult to distinguish. (1) **The Human Core:** These texts retain the high information density, specific proper nouns, and unique logical leaps of the original human author. The intent remains specific rather than generic. (2) **The Machine Surface:** Although the content originates from a human author, the syntax is stripped of natural irregularities. The resulting text often exhibits an unusual smoothness relative to the specificity of its content, combining expert-level domain knowledge with the rhythmic uniformity of a language model.

4.6 Application Discussion

Practical applications often require both fine-grained uncertainty estimation and block-wise classification, as lengthy documents may comprise a mixture of human-authored, AI-polished, and AI-generated paragraphs (see Figure 4 for an illustrative case). To meet this need, we train another variant **REVEAL-Fast** based on AIGC-text-bank. REVEAL-Fast bypasses the reasoning-generation step to output classification results directly. We found that the full REVEAL model, after produc-

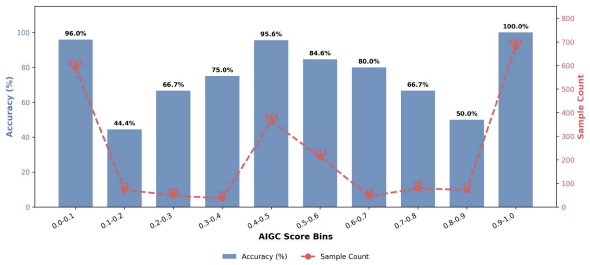

Figure 5: Confidence calibration and correlation with accuracy

ing a reasoning trajectory, yields extremely skewed label probabilities (e.g., >99%), making its output poorly calibrated for confidence estimation. REVEAL-Fast allows us to derive a well-calibrated "AIGC score" by normalizing the logits of the token preceding the final prediction. We map this score such that 0 indicates high confidence in "Human" origin, 0.5 in "AI-Polish", and 1.0 in "AI-Native", with intermediate values reflecting lower certainty. As validated in Figure 5, the score exhibits a strong positive correlation with empirical accuracy, confirming its reliability for segmenting documents and assessing the provenance of individual paragraphs with calibrated uncertainty. Further implementation details can be found in Appendix C.2.

5 Related Works

5.1 AIGC Benchmarks and Datasets

The AIGC benchmarks has shifted from single-source datasets to comprehensive, multi-dimensional evaluations. Early benchmarks like TURINGBENCH (Uchendu et al., 2021) and HC3 (Guo et al., 2023) focused on binary classification across diverse domains. Recently, M4 expanded this scope by introducing a multi-generator and multi-lingual corpus to assess detection generalization in the wild (Wang et al., 2024). To evaluate robustness against adversarial threats, DetectRL constructed datasets simulating real-world scenarios (Wu et al., 2024). Furthermore, LOKI extended detection into the multimodal domain, offering fine-grained annotations for video, image, and text anomalies (Ye et al., 2025). Despite these advancements, most existing benchmarks treat detection as a document-level binary task, failing to reflect the nature of real-world human–AI writing.

AI-Generated Text Detection. Existing detection strategies are generally categorized into white-box and black-box approaches. White-box methods typically require access to the model's internal states or rely on watermarking techniques injected during generation (Kirchenbauer et al., 2023). In contrast, black-box scenarios, which assume access only to the generated text, are more practical for applications. These approaches can be divided into zero-shot methods and supervised classifiers. Zero-shot methods exploit statistical disparities to distinguish AI text from human writing (Mitchell et al., 2023; Bao et al., 2023). Meanwhile, supervised methods fine-tune Pre-trained Language Models (PLMs) like RoBERTa on large-scale corpora to capture semantic patterns (Solaiman et al., 2019). While these detectors achieve high in-domain accuracy, both zero-shot and supervised methods degrade substantially under real-world adversarial settings, such as AI-polished text or perturbation attacks (Wu et al., 2024; Krishna et al., 2023).

6 Conclusion

In this work, we construct **AIGC-text-bank**, a comprehensive dataset for AI-generated content detection, and propose **REVEAL**, a reasoning-driven framework that replaces black-box classification with interpretable, chain-of-thought analysis. REVEAL is trained in two stages: SFT to initialize reasoning, followed by RL to refine consistency and accuracy. Experiments across five benchmarks show that our approach achieves robust performance with strong generalization. This work bridges high-accuracy detection with human-verifiable explainability, providing a trustworthy foundation for real-world AIGC identification.

Acknowledgments

The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance.

Limitations

While our work advances AIGC detection through reasoning-driven methods, several limitations merit consideration.

First, the *Think-then-Answer* paradigm introduces higher inference latency compared to conventional discriminators, posing challenges for real-time applications. Future work may explore model distillation techniques to compress reasoning pathways, parallel processing of reasoning and classification steps, or early-exit mechanisms that adap-

tively shorten the reasoning chain when confidence is high.

Second, REVEAL currently operates only on textual content and cannot process multimodal inputs such as images, audio, or video. Extending the framework to support multimodal detection would require integrating visual or auditory encoders and designing cross-modal reasoning mechanisms. This direction would allow the model to identify AI-generated content in richer, mixed-modality contexts, better aligning with real-world content consumption.

Finally, the rapid evolution of LLMs presents a persistent challenge, as detectors must continuously adapt to new generator architectures and emerging synthetic patterns. Future research could investigate continual learning strategies that enable detectors to incrementally update with minimal retraining, or develop synthetic data generation pipelines that simulate forthcoming model behaviors. Collaboration with model developers for access to early model outputs could also facilitate more proactive detector adaptation.

Ethics Statement

This work aligns with the ACL Code of Ethics. The **AIGC-text-bank** dataset is curated from publicly available sources (e.g., arXiv, Reddit) in strict compliance with their terms of use, intended only for academic research. We recognize the ethical risks inherent in AIGC detection, particularly the potential for false positives that could result in unjust accusations of misconduct. To address this, our **REVEAL** framework follows a "Think-then-Answer" paradigm, generating interpretable reasoning chains that allow human users to verify evidence rather than relying on opaque automated decisions. We emphasize that this detector should serve strictly as an assistive tool for human oversight, not as an autonomous decision-maker in high-stakes scenarios.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Janek Bevendorff, Daryna Dementieva, Maik Fröbe, Bela Gipp, André Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, and 1 others. 2025. Overview of pan 2025: Generative ai detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection. In *European Conference on Information Retrieval*, pages 434–441. Springer.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Chen Xinhui, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Long Tang, Lei Zhang, and 1 others. 2025. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23559–23567.

Vincenzo Ciancaglini, Craig Gibson, David Sancho, Odhran McCarthy, Maria Eira, Philipp Amann, and Aglika Klayn. 2020. Malicious uses and abuses of artificial intelligence. *Trend Micro Research*, pages 4–79.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv 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. *arXiv preprint arXiv:2301.07597*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International conference on machine learning*, pages 17061–17084. PMLR.

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

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4.

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.

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

OpenAI. 2024a. GPT-3.5 Turbo fine-tuning and API updates.

OpenAI. 2024b. GPT-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024c. Introducing OpenAI o3 and o4-mini.

Mike Perkins. 2023. Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond. *Journal of University Teaching and Learning Practice*, 20(2):1–24.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. *arXiv preprint arXiv:2309.02731*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

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.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and 1 others. 2024. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S Chao. 2024. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 37:100369–100401.

xAI. 2024. Grok 4.1.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, and 1 others. 2025. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *ICLR*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

A Dataset Construction Details

In this section, we provide more details about our dataset **AIGC-text-bank**. Table 6 shows the comparison of our dataset with other AIGC detection datasets.

Table 6: The comparison of AIGC detection dataset.

Dataset	Samples	#.LLMs
DetectRL	235,200	4
LOKI	3,359	6
M4	147,895	6
PAN	361,579	3
AIGC-text-bank	**1,498,279**	**12**

Table 7: Generator model List. Specific versions include DeepSeek-R1 (Guo et al., 2025), Grok 4.1 (xAI, 2024), Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct (Grattafiori et al., 2024), GPT-5 (Singh et al., 2025), GPT-4o (Hurst et al., 2024), GPT-3.5 turbo (OpenAI, 2024a), GPT-2 XL (Radford et al., 2019), Phi-4 (Abdin et al., 2024), Qwen3-8B, Qwen3-32B (Yang et al., 2025), and Qwen2.5-14B-Instruct (Qwen et al., 2025).

Model Family	**Specific Version**
DeepSeek	DeepSeek-R1
Grok	Grok 4.1
Llama	Llama-3.3-70B-Instruct Llama-3.1-8B-Instruct
OpenAI	GPT-5 GPT-4o GPT-3.5 turbo GPT-2 XL
Phi	Phi-4
Qwen	Qwen3-8B Qwen3-32B Qwen2.5-14B-Instruct

A.1 Human Data Collection

To ensure the diversity and high quality of the human baseline, we curated data from 10 distinct domains. As mentioned in Section 2.1, all human texts are published before November 2022 to prevent potential interference from LLM-generated content. Table 8 presents the detailed statistics and sources for each domain, and Figure 6 illustrates the token distribution of the human data.

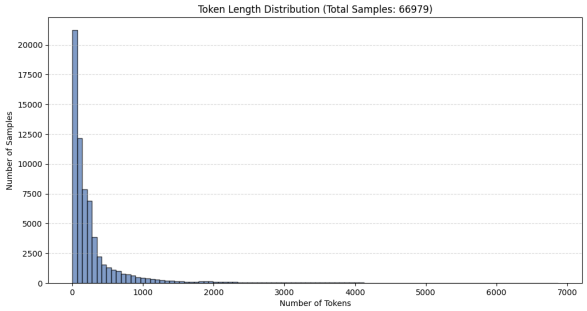

Figure 6: The token distribution of human data.

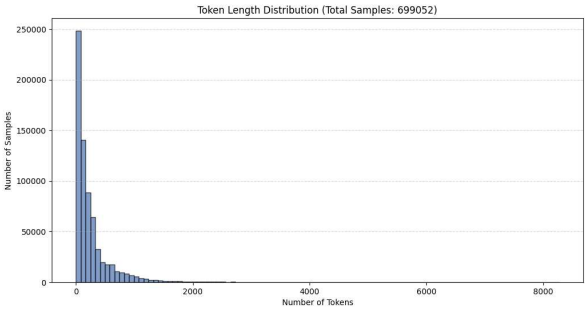

Figure 7: Token distribution of the AI-Native data.

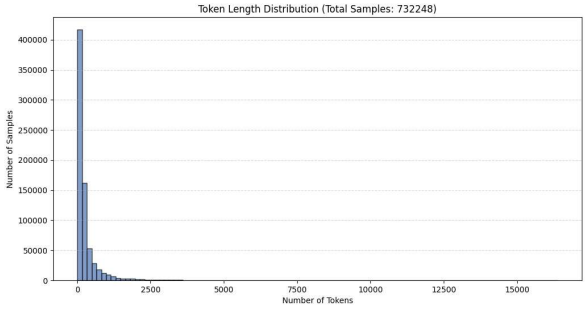

Figure 8: The token distribution of AI-Polish data.

A.2 Generator Models

We utilize a diverse set of 12 LLMs to generate the AI-Native and AI-Polish subsets. Table 7 summarizes the specific model versions used.

A.3 AI-Native Generation

Table 9 presents the detailed statistics of the AI-Native subset across 10 domains and 12 LLMs, and Figure 7 illustrates the token length distribution, which closely aligns with the human dataset to minimize length-based bias.

A.4 AI-Polish Generation

Table 10 presents the detailed statistics of the AI-Polish subset, while Figure 8 illustrates the token length distribution across different samples, providing further insight into its structural characteristics.

Table 8: Detailed breakdown of Human data sources and statistics across domains.

Domain	Samples	Description	Source
Academic	9,894	arXiv papers	https://www.kaggle.com/datasets/Cornell-University/arxiv
Blog	9,986	Blog posts	https://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
Encyclopedic	558	Wikipedia	https://www.wikipedia.org/
Essay	1,375	Native Speaker Essay	https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data
	3,282	Non-Native Speaker Essay	https://language.sakura.ne.jp/icnale/
Literature	41	Classic books	https://www.gutenberg.org/
News	5,000	News articles	https://huggingface.co/cnn_dailymail/datasets
Q&A	9,991	Yahoo Answers	https://www.kaggle.com/datasets/soumikrakshit/yahoo-answers-dataset
Reviews	4,999	Product reviews	https://huggingface.co/amazon_polarity/datasets
Social Media	9,831	Twitter posts	https://huggingface.co/datasets/enryu43/twitter100m_tweets
	9,446	Reddit posts	https://developers.reddit.com/docs/capabilities/server/reddit-api
Speeches	2,450	TED Talks	https://www.kaggle.com/datasets/rounakbanik/ted-talks
	126	Presidential Speech	https://www.kaggle.com/datasets/kboghe/presidentialspeeches

B Prompt Engineering

In this section, we provide the exact prompt templates used in our framework. We detail the prompts for three distinct stages: (1) Reasoning Data Synthesis (Teacher Model), (2) Standard Detection (REVEAL), and (3) Consistency Evaluation (Reward Model).

B.1 Reasoning Data Synthesis

To construct the reasoning-augmented dataset $\mathcal{D}_{\text{sft}}$, we employ a **hindsight prompting** strategy. Specifically, We provide the teacher model (OpenAI o3) with both the input text and its ground-truth label, instructing it to reconstruct the reasoning process leading to the correct classification. The template is presented in Table 12.

Instruction: A conversation between User and Assistant. The Assistant first thinks in <think>...</think> tags then answers in one word (Human or AI) in <answer>...</answer> tags.
Your task: You are given a human-written or AI-generated/edited piece of text. You must determine whether the piece was written/edited by AI or human-written.
Let's think step-by-step. Describe inconsistencies/AI artifacts or any clues that this text may be human/written, summarize your analysis, then answer with Human or AI.

Text: {input_text}

Table 11: The Inference Prompt used for REVEAL training and baseline evaluation.

B.2 Standard Detection Prompt

For both the Supervised Fine-Tuning (SFT) of REVEAL and the inference evaluation of baseline models, we use a uniform **Think-then-Answer**

Table 9: Detailed statistics of the AI-Native subset.

Model	Aca.	Blog	Enc.	Essay	Lit.	News	Q&A	Rev.	Soc.	Spch.	Total
DeepSeek-R1	8539	1676	20	1787	0	1344	1851	955	3451	38	19661
Grok 4.1	9847	9737	502	4413	38	4308	9591	4987	19142	2526	65091
Llama-3.3-70B-Instruct	8077	9060	543	4656	28	4984	8474	4375	14534	1524	56255
Llama-3.1-8B-Instruct	9894	9867	337	4648	11	4781	9978	4998	19053	1043	64610
GPT-5	9894	9532	363	4635	1	4780	9986	4998	18879	442	63510
GPT-4o	9894	9970	495	4656	37	4971	9989	4999	19243	2264	66518
GPT-3.5 turbo	7344	7360	386	3499	9	3721	7438	3758	14,217	1058	65777
GPT-2 XL	8690	7584	219	4525	2	4579	5269	3,295	11,372	637	46172
Phi-4	9285	9032	392	4651	18	4850	9285	4914	18405	920	62318
Qwen3-8B	9806	9850	326	4652	31	4880	9905	4863	18859	1719	64891
Qwen3-32B	9891	9973	554	4657	35	4999	9970	4998	19217	2319	66613
Qwen2.5-14B-Instruct	8052	9182	324	4626	14	4670	8976	4100	16843	849	57636

Table 10: Detailed statistics of the AI-Polish subset.

Model	Aca.	Blog	Enc.	Essay	Lit.	News	Q&A	Rev.	Soc.	Spch.	Total
DeepSeek-R1	9824	8427	141	4211	2	573	9009	4791	18032	167	55227
Grok 4.1	9876	9619	416	4613	19	4247	9757	4971	19001	2336	64855
Llama-3.3-70B-Instruct	9759	9691	516	4634	35	4763	9392	4741	18541	1705	63777
Llama-3.1-8B-Instruct	9892	9837	536	4638	37	4969	9928	4998	19101	1337	65273
GPT-5	9888	1533	97	4649	5	4872	9988	1017	10799	358	43206
GPT-4o	9888	9955	556	4657	18	5000	9955	4996	19218	1572	65806
GPT-3.5 turbo	9884	9977	555	4657	41	4997	9969	4998	19250	2350	66678
GPT-2 XL	8261	6910	82	3282	2	1781	6918	3627	13465	94	44422
Phi-4	9860	9964	547	4547	39	4987	9962	4995	19202	1563	65666
Qwen3-8B	9741	9898	554	4348	41	4990	9458	4781	18929	2575	65315
Qwen3-32B	9886	9973	557	4654	41	5000	9924	4979	19242	2563	66819
Qwen2.5-14B-Instruct	9852	9928	545	4319	41	4982	9881	4935	19101	1621	65204

prompt. This ensures that the model generates an explicit reasoning chain before predicting the final label. The template is presented in Table 11.

B.3 Consistency Reward Prompt

In the Reinforcement Learning stage, we employ GPT-4o as a reward model to assess the logical consistency of the generated reasoning chain (R_{cons}). To guide its evaluation, we provide the model with **2 examples** to illustrate how to evaluate the reasoning process. The evaluation template is presented in Table 13.

C Further Analysis and Discussion

C.1 Ablation Study

To provide deeper insight into the contribution of each component, we visualize the reward curves during the Reinforcement Learning phase.

Figures 9 display the progression of Total Reward, Answer Reward, Consistency Reward, and Format Reward, respectively. These curves reveal three critical observations regarding the stability and efficiency of our framework:

Impact of Data Selection Strategy. As shown in Figure 9, the *w/o Selection* variant (orange curve) exhibits slightly higher rewards than the Full model on the validation set, largely because the validation set follows the same distribution as the training data. Random sampling in *w/o Selection* overrepresents high-confidence, statistically abundant easy samples, inflating validation rewards. In contrast, uncertainty-based filtering biases the Full model toward ambiguous, boundary cases. Although this reduces average rewards on an easy validation set, it mitigates overfitting to trivial patterns and leads to substantially improved OOD robustness, as evi-

Instruction: You're a forensic writing analyst trained to detect whether a piece of text was written by a human or generated by AI.

Below is a passage of text and a known label indicating whether it is Human-written or AI-generated.

Your job is to:
1. Analyze the text step by step.
2. Identify concrete evidence that supports the given label.
3. Contrast it with why the opposite label is less likely.
4. Write your reasoning in natural language inside <think> tags.
5. Conclude with the final label (One word: "Human" or "AI") in <answer> tags.
6. Do not use any other tags or formatting.
7. Do not explicitly mention the ground truth label in your reasoning. Assume you do not yet know the label.

Always ground your analysis in specific stylistic, structural, or semantic features of the text. Avoid generic summaries or descriptions.

Text: {input_text}
Ground Truth Label: {label}

<think>

Table 12: The Hindsight Prompt used for generating reasoning traces from the Teacher Model (OpenAI o3).

denced in Table 5.

The Necessity of SFT Initialization. Removing the SFT stage (*w/o SFT*) leads to a substantial degradation in training efficiency, as reflected by the blue curves. The format reward reveals that, without SFT, the early RL phase is largely consumed by learning basic syntactic structures (e.g., reasoning tags) rather than improving reasoning quality. This cold-start issue propagates to downstream learning: inadequate formatting prevents the model from forming coherent reasoning trajectories, leading to persistently low Consistency Rewards. These results confirm that SFT provides an essential structural prior, enabling the RL phase to concentrate on refining reasoning consistency instead of recovering basic output structure.

Effectiveness of the Full Framework. The Full REVEAL model (red curve) demonstrates consistent improvement across all reward dimensions. In contrast to *w/o SFT*, it maintains strong format adherence from the outset, and compared to *w/o Weighted* (green curve), it achieves superior asymptotic performance in both Answer and Consistency rewards. By integrating SFT initialization, weighted loss, and hard-sample mining, the Full model ensures that gains in reward metrics correspond to genuine improvements in reasoning capability.

Instruction: You are an expert evaluator tasked with assessing whether a model's final classification is consistent with its reasoning. The model's objective is to determine whether a given piece of text was written by a human or generated by a large language model (LLM).

You will be provided with:
- An input text: the passage under evaluation.
- A reasoning trace, enclosed in <think>...</think> tags, representing the model's chain-of-thought.
- A final classification, enclosed in <answer>...</answer> tags, indicating the model's predicted label (Human or AI).

Your task is to return a list of three float scores, each ranging from 0.0 to 1.0, corresponding to the following criteria:
1. Answer–Reasoning Alignment: Does the reasoning logically support the final answer? This should be binary (1.0 or 0.0) based on whether the reasoning is consistent with the final classification.
2. Groundedness: Is the reasoning grounded in the input text and internally coherent?
3. Specificity (Genericness): How specific, informative, and non-generic is the reasoning?

Respond strictly with a Python-style list of floats in this format:
[alignment_score, groundedness_score, genericness]
Do not include any explanations, comments, or extra output.

Examples: {2 examples}
Text:{original_text}
Model Output:{model_output}

Table 13: The Reward Prompt used for evaluating reasoning consistency in RL.

C.2 Detailed Application Discussion

While our main model focuses on reasoning, practical scenarios often require fast, fine-grained scanning. To address this, we utilize **REVEAL-Fast**, trained directly on 3-class labels (Human, AI-Native, AI-Polish) for paragraph-level detection.

To quantify the model's confidence, we extract the raw logits associated with the final token before the generated label. Let z_h, z_p, z_n represent the output logits for Human, AI-Polish, and AI-Native respectively. We first apply a standard Softmax function to normalize these logits into a probability distribution:

$$P_c = \frac{\exp(z_c)}{\sum_j \exp(z_j)}, \quad c \in \{h, p, n\}. \tag{7}$$

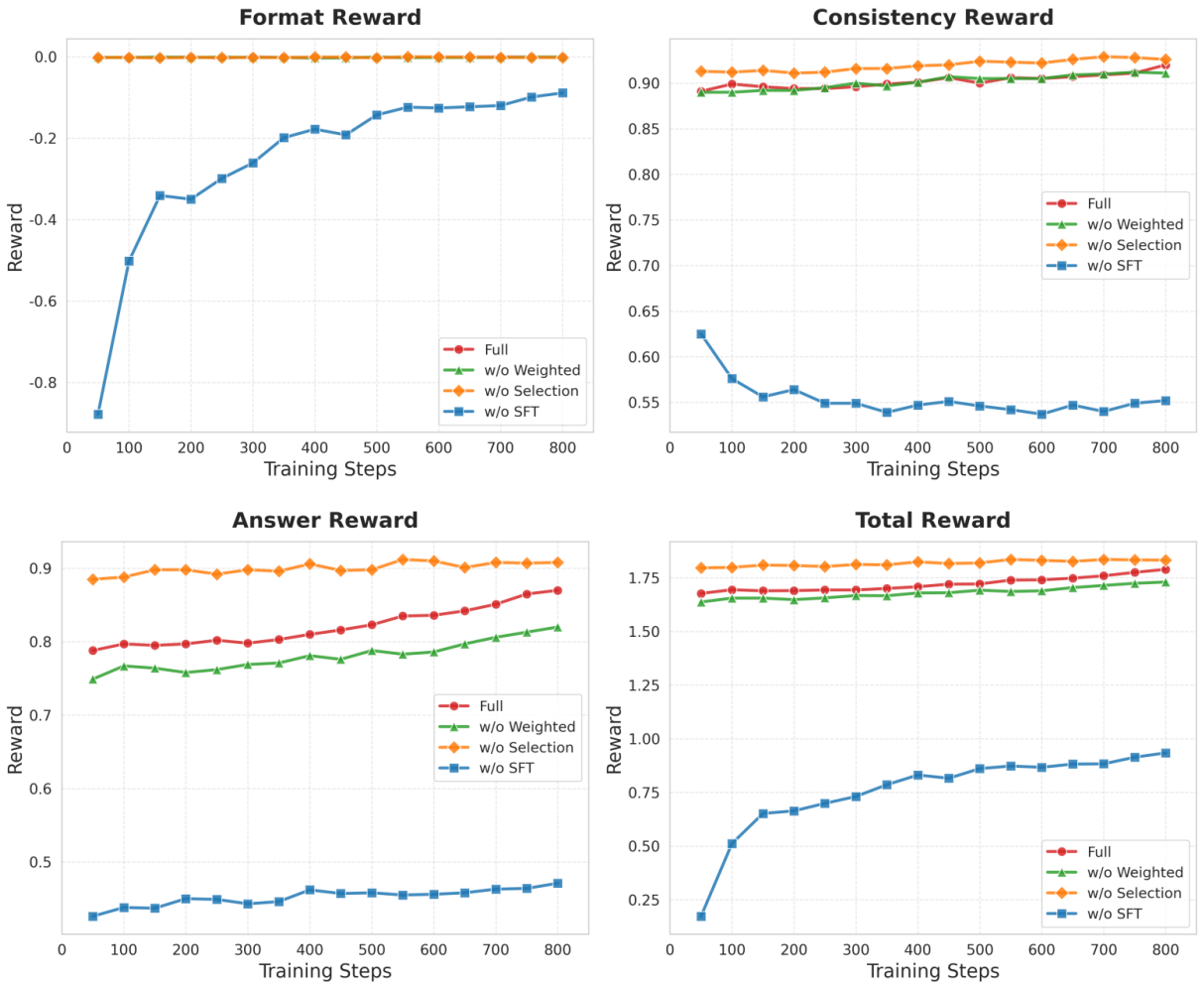

Figure 9: The reward curves during Reinforcement Learning.

To map these discrete probabilities onto a continuous spectrum $S \in [0, 1]$, we formulate the score as the mathematical expectation of the "AI-Generation Degree". We assign discrete quantization values to each category: 0 for Human, 1 for AI-Native, and 0.5 for AI-Polish. The final AIGC Score S can be calculated as:

$$\begin{aligned} S &= \mathbb{E}[\text{AI Degree}] \\ &= 0 \cdot P_h + 0.5 \cdot P_p + 1 \cdot P_n \\ &= P_n + 0.5 \cdot P_p \end{aligned} \tag{8}$$

This expectation-based formulation is theoretically sound as it accounts for the inherent uncertainty between categories. For instance, if the model is uncertain between Human and AI-Polish (e.g., $P_h = 0.5, P_p = 0.5$), the score naturally converges to 0.25, correctly indicating a low-risk segment. Conversely, uncertainty between AI-Polish and AI-Native yields a score around 0.75, flagging high-risk content.

To confirm the validity of this scoring mechanism, we conduct a calibration experiment on the 3-class test set of **AIGC-bench**. We partition the samples into 10 distinct bins based on their predicted AIGC Score (e.g., $[0.0, 0.1), \ldots, [0.9, 1.0]$). For each bin, we calculate the accuracy.

As shown in Figure 5, AIGC scores near 0, 0.5, and 1 indicate high confidence in the respective classes, while accuracy is also the highest in these regions. This pattern demonstrates that the AIGC Score offers a fine-grained and well-calibrated measure of the model's output confidence.

To illustrate the model's capabilities in real-world scenarios, we provide a fine-grained detection case study in Figure 4. Starting with a raw human-written text, we segmented it into distinct paragraphs and manually constructed a hybrid test case: some paragraphs were left as original human text, others polished by an LLM to simulate AI-Polish content, and the concluding section fully generated to represent AI-Native content. As illustrated in the figure, our model effectively scans

the document paragraph-by-paragraph, and provide AIGC score for each section. The visualization highlights the model's ability to differentiate between subtle polishing and complete generation, enabling precise localization of AI content within mixed-source documents.