"I know myself better, but not really greatly": How Well Can LLMs Detect and Explain LLM-Generated Texts?

Anonymous ACL submission

Abstract

Distinguishing between human- and LLM-generated texts is crucial given the risks associated with misuse of LLMs. This paper investigates detection and explanation capabilities of current LLMs across two settings: binary (human vs. LLM-generated) and ternary classification (including an "undecided" class). We evaluate 6 close- and open-source LLMs of varying sizes and find that self-detection (LLMs identifying their own outputs) consistently outperforms cross-detection (identifying outputs from other LLMs), though both remain suboptimal. Introducing a ternary classification framework improves both detection accuracy and explanation quality across all models. Through comprehensive quantitative and qualitative analyses using our human-annotated dataset, we identify key explanation failures, primarily reliance on inaccurate features, hallucinations, and flawed reasoning. Our findings underscore the limitations of current LLMs in self-detection and self-explanation, highlighting the need for further research to address overfitting and enhance generalizability.

1 Introduction

The rise of large language models (LLMs) has brought remarkable advancements in natural language processing (NLP) tasks (Matarazzo and Torlone, 2025), including text generation. Models such as GPT-4o (OpenAI, 2024), LLaMA (Touvron et al., 2023), and Qwen (Team, 2024) have blurred the boundaries between LLM-generated (LGTs) and human-generated texts (HGTs), posing new challenges in distinguishing between the two. While these capabilities of LLMs open new possibilities, they also bring concerns in areas such as misinformation, academic dishonesty, and automated content moderation (Hu, 2025). As a result, detecting LGTs has become an increasingly important research area (Dugan et al., 2024; Lee et al., 2023; Bhattacharjee and Liu, 2024a).

Prior research has mainly focused on developing classifiers to distinguish HGTs and LGTs, including open-source detectors (Hans et al., 2024) and online close-source detection systems (Tian et al., 2023). However, most detection systems have been limited to binary classification, which has several inherent issues. Recently, some works (Lee et al., 2024b) have attempted ternary classification by introducing a "mixed" category, which represents texts originating from mixed sources. However, this approach does not fundamentally resolve the issue. We further adopt the definition of an "Undecided" category based on other studies (Ji et al., 2024) and conduct ternary classification experiments for different LLMs, as certain texts are inherently indistinguishable between LGTs and HGTs. Furthermore, many studies treat the detection task as a black box, offering little insight into the decision-making process. Explainability, a critical aspect of trustworthy AI, has received less attention, but it is essential for building systems that users can trust (Weng et al., 2024; Zhou et al., 2024). This paper presents an analysis of current LLMs in detecting LGTs and HGTs, with a particular emphasis on *evaluating and improving the clarity of the explanations provided by LLM-based detectors*. By investigating how LLMs make predictions and offer explanations for their decisions, we aim to enhance their transparency and provide deeper insights into their reasoning processes.

This paper explores the explainability of LLM-based detectors, addressing two central questions: (1) *How accurately can current LLM detectors identify origins of texts*, and (2) *How reliable are their explanations?* Our study highlights that in a ternary setting compared to traditional binary classification, the average detection performance improves by 5.6%, which demonstrates the necessity of ternary rather than binary setting to detect HGTs and LGTs. We further discovered that explanations are often flawed even when binary pre-

dictions are correct. Based on our comprehensive human-annotators' feedback, we summarize three common issues with explanations: reliance on inaccurate features (e.g., vague or irrelevant characteristics), hallucinations (e.g., non-existent or contradictory features), and incorrect reasoning (e.g., logical errors in attributing text origin). These explanation errors are quantified and categorized, with their distributions analyzed across different LLMs. Consequently, the proportion of explanation errors decreases by 13.3% when we switch to ternary classification setting, which further supports the necessity of ternary classification for LGTs detection.

We evaluated 6 state-of-the-art (SOTA) LLM-based detectors, such as GPT-4o, GPT-4o mini, LLaMA3.3-70B, LLaMA3.3-7B, Qwen2-72B, and Qwen2-7B, on our created dataset comprising LGTs and HGTs. Moreover, our human annotators provided feedback based on correctness of predictions and explanations for this benchmark. Our results show that GPT-4o achieved the highest detection accuracy. In addition, LLMs performed better in self-detection than cross-detection, and ternary classification outperformed binary classification. Finally, explanation quality also improved under ternary setting, with fewer hallucinations and incorrect reasoning observed.

The main contributions of this work are:

- **Comprehensive Evaluation of Detection and Explanation:** We systematically assess current LLMs' ability to detect and explain human- and LLM-generated texts using both binary and ternary classification tasks, demonstrating the advantages of ternary classification for both detection accuracy and explanation quality.
- **Human-Annotated Dataset:** We present a new human-annotated dataset of LLM- and human-generated texts, enabling evaluation of LLM explanations and improved detector training.
- **Analysis of Explanation Errors:** We quantitatively and qualitatively characterize key explanation failures, reliance on inaccurate features, hallucinations, and flawed reasoning, offering insights for LLM detection and self-explanation.

2 Related Work

LGT and HGT Detection. Past efforts to identify LGTs often relied on binary classification systems that distinguish HGTs from LGTs using surface-level features. While these methods were initially effective, they are prone to errors when encountering adversarial attacks or domain shifts, which limit their overall robustness (Bhattacharjee and Liu, 2024a; Dugan et al., 2024). To address these limitations, researchers have explored strategies that integrate external knowledge, such as combining internal and external factual structures, to boost detection against diverse content and styles (Internal and Structures, 2024). Recent studies also highlight the promise of using LLMs themselves for text detection: approaches like self-detection and mutual detection can outperform traditional classifiers, as illustrated by GPT-4's success in tasks like plagiarism detection (Lee et al., 2024a). Notably, smaller models sometimes excel in zero-shot scenarios, offering adaptable solutions across varying architectures (He et al., 2024). Furthermore, Lee et al. (2023) demonstrated that LLMs can reliably identify their own outputs, providing a more nuanced framework for content verification. Despite these advances, the continuing challenges of domain adaptation and adversarial resistance underscore the need for more versatile and robust detection systems.

Explainability in Detection Models. Recent work on LGT detection has focused on improving explainability. Zhou et al. (2024) proposed to incorporate factual consistency into detection models to enhance their interpretability, while Weng et al. (2024) explored mixed-initiative approaches that combine human expertise with automated models for better detection. These studies have made significant contributions to the field; however, they either depend heavily on expert input (Weng et al., 2024) or lack integration of explanation generation within the model itself (Zhou et al., 2024). Our approach, in contrast, enables LLMs to autonomously generate both predictions and detailed explanations, making it a more scalable and transparent solution for detecting machine-generated content.

Ternary Classification. Traditional binary classification methods face limitations when texts exhibit ambiguous characteristics. Introducing an "Undecided" category addresses this by capturing three distinct scenarios (see Appendix E for concrete illustrations): (1) *Mixed texts* co-authored by humans and LLMs, where stylistic blending creates classification challenges; (2) *Inherently ambiguous texts* that could plausibly originate from either source despite single authorship; and (3) *Fragile indicators*, where subtle distinguishing fea-

Datasets \ Models	LLM Detectors					
	GPT-4o	GPT-4o mini	LLaMA3.3-70B	LLaMA3.3-8B	Qwen2-72B	Qwen2-7B
GPT-4o	**71.39**	59.38	57.31	48.41	64.14	59.76
GPT-4o mini	65.71	61.03	53.75	51.73	**67.27**	60.09
LLaMA3.3-70B	**67.26**	60.92	68.10	53.65	58.96	51.57
LLaMA3.3-8B	60.74	55.77	**62.29**	59.09	59.87	49.88
Qwen2-72B	62.66	61.92	57.79	49.20	**68.15**	61.36
Qwen2-7B	62.45	59.06	59.12	48.57	**65.24**	63.44
Average	**65.03**	59.68	59.73	51.78	63.94	57.68

Table 1: F1 scores of LLM-based detectors in binary classification. The first column indicates different LLMs used for text generation, and the first row indicates different LLMs acting as detectors. The highest column-wise F1 score for each LLM detector to classify LGTs and HGTs across six datasets is highlighted in blue. The highest row-wise F1 score for each LLM-generated text dataset across different LLM detectors is marked in **blue**.

tures exist but lack robustness against behavioral evolution of either LLMs or human writers. This approach advances beyond previous methods that primarily addressed mixed texts (Lee et al., 2024b). The complexity is evidenced by Turing tests showing human difficulty in binary classification (Frank et al., 2024), and by studies demonstrating detector limitations with evolving writing patterns (Bhattacharjee and Liu, 2024b). The ternary framework improves both accuracy and explainability, particularly for these edge cases.

3 LLM-based Binary Classification on LGTs and HGTs

3.1 Experimental Design

We selected six SOTA LLMs for text generation and subsequent detection: GPT-4o, GPT-4o mini (Hurst et al., 2024), Qwen2-72B, Qwen2-7B (Yang et al., 2024), LLaMA3.3-70B, and LLaMA3.3-8B (Dubey et al., 2024). These LLMs were chosen for two main reasons. First, they represent the latest advancements in LLM development, demonstrating strong generation and detection capabilities. Second, selection spans different series and model sizes, enabling a comparative analysis of performance across architectures and scales.

To construct the dataset, we first selected 1,000 HGTs from publicly available M4GT-Bench dataset (Wang et al., 2024), ensuring a diverse range of topics, styles, and formats. Based on these selected HGTs, we designed 1,000 prompts that align with themes, structure, and style of the HGTs. Each LLM subsequently generated a corresponding response for each of these prompts. Together, these LGTs and HGTs formed the benchmark used in this study. For each text, the LLMs were tasked to determine its source (LGTs or HGTs) and provide an explanation, as illustrated in Table 2.

Prompt: Please determine whether the following text is generated by large language models or by a human, and provide a clear judgment. Additionally, please offer a detailed explanation for your decision. Please structure your answer in JSON format as follows: {"answer": , "explanation": }.

Table 2: A prompt for LLMs to determine text origin and provide an explanation under a binary setting.

Manual Annotation. To assess LLMs' ability to explain text origins and identify distinguishing features, 3 co-authors, who are undergraduate computer science students, manually evaluated correctness of LLM-generated explanations. They determined accuracy of each explanation. From 7 datasets (6 SOTA LLMs + Human), 100 texts with corresponding explanations per dataset were randomly selected for human evaluation. All annotators assessed explanations provided by each model, which achieved a Fleiss' kappa (Fleiss, 1971) of 0.8387, indicating near-complete agreement. Annotation guidelines are detailed in Appendix D.

Evaluation Metrics. For evaluating the classification performance of the LLMs, the primary metric we used is the F1 score. To assess the quality of explanations, human evaluators reviewed the LLM-generated explanations and classified them as correct or incorrect. The F1 score was also used as the evaluation metric for explanation quality.

3.2 Binary Classification Results

We evaluated the performance of six LLMs across six datasets, as detailed in Table 1, which systematically compares the detection capabilities of various LLMs for both LGTs and HGTs. The results

Datasets \ Models	LLM Detectors					
	GPT-4o	GPT-4o mini	LLaMA3.3-70B	LLaMA3.3-8B	Qwen2-72B	Qwen2-7B
GPT-4o	**73.48** / **67.04**	58.41 / 54.17	57.65 / 54.17	48.29 / 51.32	64.11 / 59.60	59.57 / 62.13
GPT-4o mini	63.72 / 60.95	63.43 / 60.15	53.85 / 52.46	50.01 / 47.75	**66.91** / **61.08**	60.13 / 57.28
LLaMA3.3-70B	68.13 / 63.96	62.12 / 60.32	**68.33** / **64.47**	53.13 / 51.78	58.16 / 59.22	52.41 / 48.82
LLaMA3.3-8B	58.97 / 61.11	56.19 / 55.98	**63.24** / **63.72**	58.73 / 56.29	59.83 / 59.17	49.66 / 48.97
Qwen2-72B	62.70 / 61.09	62.91 / 62.84	58.71 / 56.99	49.12 / 47.26	**70.47** / **67.98**	61.24 / 58.18
Qwen2-7B	63.78 / 61.54	58.11 / 57.84	60.15 / 58.60	48.72 / 49.17	**65.44** / **63.58**	63.83 / 61.91

Table 3: F1 scores of LLM-based detectors on human-annotated texts for **binary** classification. Each dataset contains 100 LGTs and 100 HGTs with human-annotated explanations. Each cell indicates classification/explanation F1, where the highest column-wise F1 of each LLM detector for binary classification and explanations across different generated texts are highlighted with blue and red, respectively. In addition, the highest row-wise F1 among different LLM detectors for each LLM-generated text datasets are indicated with **blue** and **red** in bold, respectively.

demonstrate that GPT-4o achieves the best average detection performance across all datasets, showing relatively strong generalization capabilities. Larger parameter models generally exhibit significantly better detection performance than smaller ones, which suggests that these models are not merely making random guesses but are effectively identifying distinctive textual features.

The F1 scores in Table 1's diagonal direction show that LLMs within same series consistently detect their own outputs more effectively than those from other LLM families. For example, LLaMA3.3 70B achieves the highest F1 score in its generated dataset, which indicates a heightened sensitivity to its own text distribution compared to other LLMs. However, this specialization reduces cross-detection performance, as seen in Qwen2-7B's lower F1 on LLaMA-generated texts. While larger LLMs generally achieve better detection across different LLMs, such as GPT-4o, GPT-4o mini, LLaMA3.3-70B and Qwen2-72B, their outputs are also more difficulty to distinguish by smaller LLMs, such as LLaMA3.3-8B and Qwen2-7B.

Additionally, based on the human annotations of sampled 100 LGTs and 100 HGTs with explanations from each dataset, we observed that the detection and explanation results across different LLMs are not entirely consistent, as shown in Table 3. We noted that in some cases, the F1 score for explanations was higher than that for classification. This is because, in these cases, the explanation correctly identified the reasoning for attribution, but the final classification was incorrect. For instance, the difference in F1 scores between explanation and classification was particularly noticeable for LLaMA3.3-8B and Qwen2-7B, suggesting that these models struggle to truly comprehend the textual features necessary for correctly determining the origin of generated texts, which results in lower detection performance.

As shown in Table 4, analysis of the annotators' results revealed that models are generally more accurate in attributing HGTs compared to LGTs. For example, while GPT-4o demonstrates higher accuracy (78 out of 100) in classifying HGTs, the false explanations account for more than 47%.

4 LLM-based Ternary Classification on LGTs and HGTs

4.1 Experimental Setup

Using the same benchmark in § 3, we prompted the LLMs for ternary classification and the prompt template is demonstrated in Table 6. The ground truth for the ternary classification was determined based on annotators' votes, where the three annotators were aware of the text's origin (LLMs or human) and were asked to distinguish between the ground truth and the "Undecided" category. This allowed for the evaluation of both the LLM's classification results and the explanations provided by the LLMs. The Fleiss' kappa (Fleiss, 1971) for the ternary classification annotations among the three annotators was calculated as 0.7629, which indicates substantial agreement.

4.2 Ternary Classification Results

Table 5 presents the F1 scores of LLMs in the ternary classification setting. Comparing it with Table 3, we observe that introducing the "Undecided" category leads to overall performance improvements across both classification and explanation tasks. Specifically, GPT-4o exhibits the most notable gains, improving from 73.48/67.04 to 79.73/72.04, indicating that a finer-grained classification allows stronger models to better capture nuanced differences between LGTs and HGTs.

Model	MGTs						HGTs					
	TC	TE	FE	FC	TE	FE	TC	TE	FE	FC	TE	FE
GPT-4o	64	51$_{:79.7\%}$	13$_{:20.3\%}$	36	8$_{:22.2\%}$	28$_{:77.8\%}$	78	41$_{:52.6\%}$	37$_{:47.4\%}$	22	2$_{:9.1\%}$	20$_{:90.9\%}$
LLaMA3.3-70B	56	35$_{:62.5\%}$	21$_{:37.5\%}$	44	10$_{:22.7\%}$	34$_{:77.3\%}$	60	37$_{:61.7\%}$	23$_{:38.3\%}$	40	7$_{:17.5\%}$	33$_{:82.5\%}$
Qwen2-72B	60	36$_{:60.0\%}$	24$_{:40.0\%}$	40	7$_{:17.5\%}$	33$_{:82.5\%}$	69	50$_{:72.5\%}$	19$_{:27.5\%}$	31	7$_{:22.6\%}$	24$_{:77.4\%}$

Table 4: Performance of LLMs on LLM-generated and human-generated texts for ternary classification and explanation tasks. It includes results for classification and explanation tasks, where TC represents true classification, FC represents false classification, TE represents true explanation, and FE represents false explanation. Note: TC=TE+FE and FC=TE+FE.

Datasets \ Models	GPT-4o	GPT-4o mini	LLaMA3.3-70B	LLaMA3.3-8B	Qwen2-72B	Qwen2-7B
			LLM Detectors			
GPT-4o	**79.73**/**72.04**	64.62/61.87	62.19/59.04	58.06/57.78	71.62/68.86	63.81/62.72
GPT-4o mini	**70.11**/**68.75**	67.39 / 65.18	58.88/52.95	54.43/51.16	69.65/65.95	65.15/62.60
LLaMA3.3-70B	**74.41**/**75.26**	65.16/64.75	72.11 / 71.83	57.05/57.34	64.94/62.44	56.46/55.32
LLaMA3.3-8B	**71.99**/**70.80**	60.18/61.10	64.82/63.93	63.96 / 62.85	63.12/60.52	54.08/53.01
Qwen2-72B	67.28/66.74	65.12/64.73	61.81/61.74	53.24/52.87	**76.05** / **75.56**	65.26/64.72
Qwen2-7B	68.91/67.42	60.15/59.31	62.06/61.57	52.41/52.30	**70.30**/**68.44**	**66.61** / 65.17

Table 5: F1 scores of LLM-based detectors on the **ternary** classification of LGTs and HGTs. The highest column-wise and row-wise F1 scores are highlighted and marked following the same scheme as in Table 3.

Prompt: Please classify the following text into one of three categories based on its source: LLM-generated, human-generated, or undecided. The "Undecided" category refers to texts that exhibit characteristics of both LLM-generated and human-generated content, making it impossible and inappropriate to distinguish between the two. Provide a clear classification and a detailed explanation for your decision. Structure your answer in JSON format as follows: {"classification": , "explanation": }.

Table 6: A prompt for LLMs to determine text origin and provide an explanation under a ternary setting.

Moreover, Fig 1 reveals how different models distribute predictions across three categories. GPT-4o demonstrates a more balanced distribution, with relatively lower misclassification rates for both HGTs and LGTs. In contrast, LLaMA3-70B shows a stronger tendency to label texts as "human-generated", leading to a higher false positive rate. Meanwhile, Qwen2-72B exhibits a more cautious classification approach, assigning a larger proportion of texts to "Undecided" category, particularly for LGTs.

A closer comparison between binary and ternary classifications in Tables 3 and 5 suggests that added "Undecided" category benefits models differently. While large models like GPT-4o and LLaMA3-70B leverage this additional flexibility to improve both classification and explanation F1 scores, smaller models such as Qwen2-7B show more mixed results, with only marginal improvements. This suggests that high-capacity models may be better equipped to handle ambiguous cases, while smaller models struggle with added complexity.

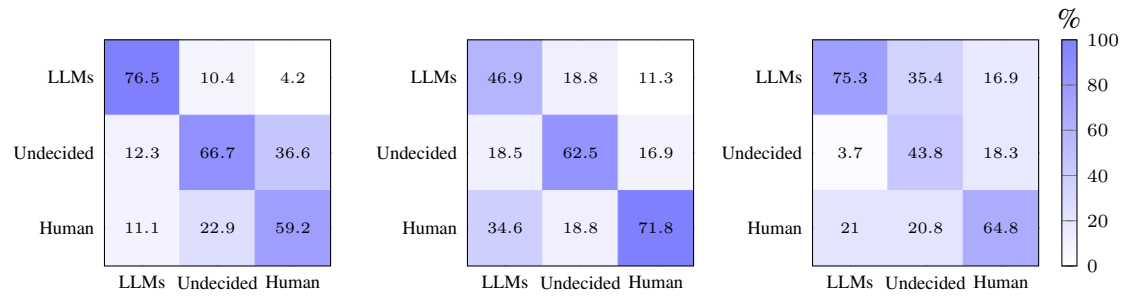
Overall, these findings indicate that ternary classification not only refines detection performance but also enhances the LLMs' ability to generate more meaningful explanations. The improvements are particularly evident in large-scale LLMs, which benefit from a more nuanced decision space.

5 Explainability of LLM-based Detectors

5.1 Incorrect Explanation Attribution

Although LLMs can distinguish LGTs and HGTs, especially in self-detection settings, there are explanations that are incorrect via human evaluation. Normally, incorrect explanations in correctly classified cases fall into three types: inaccurate features (misidentifying key attributes), hallucinations (citing nonexistent or contradictory features), and flawed reasoning (faulty logic despite a correct outcome). For misclassified texts, errors typically involve inaccurate features or hallucinations, which highlights the need to prioritize explanation accuracy alongside detection performance to enhance trust in LLM-based detectors.

Inaccurate Features. Incorrect explanation attribution is often caused by relying on ambiguous, superficial, or misinterpreted features, as shown in the examples in Tables 7 and 12 of Appendix A. In Ex-



(a) GPT-4o (b) LLaMA3.3-70B (c) Qwen2-72B

Figure 1: Confusion matrices showing the performance of LLM-based detectors (GPT-4, LLaMA3.3-70B, Qwen2-72B) in the ternary classification task, where the bottom row represents human-annotated ground truth labels (LGTs, Undecided, and HGTs), and the left column represents classification results predicted by LLM-based detectors.

Example 1: Ambiguous Features
Original Text: "Quantum entanglement enables faster-than-light communication, allowing for instantaneous data transfer across vast distances. This phenomenon has vast implications for future communication technologies, making it a subject of growing interest in both academic and technological circles."
LLM Judgment: LLM-generated
LLM Explanation: "The text discusses highly technical content, filled with specialized scientific terms, which is characteristic of machine-generated text. Such advanced topics are often produced by AI systems rather than human experts."

Example 4: Incorrectly Perceived Repetition
Original Text: "The discovery of gravitational waves was a turning point in the history of physics. This breakthrough confirmed Einstein's theory of relativity, providing compelling evidence for the existence of gravitational waves, which opened new doors in astrophysics research."
LLM Judgment: LLM-generated
LLM Explanation: "The text exhibits repetitive sentence structures, a hallmark of AI-generated content. The repeated mention of 'gravitational waves' and 'breakthrough' indicates a pattern commonly seen in machine-generated texts."

Example 6: Misinterpretation of Formal Tone
Original Text: "One must consider the economic ramifications of such policies, as they extend beyond immediate financial concerns into broader societal impacts. The implications of these policies are far-reaching, touching on employment, income inequality, and long-term societal well-being."
LLM Judgment: LLM-generated
LLM Explanation: "The formal tone and structured language initially suggest human authorship, as such features are often attributed to human experts. However, LLMs can replicate this style with high fidelity, leading to the final classification as LLM-generated."

Table 7: Analysis of LLM vs. Human Writing Attribution Based on Various Features. The table categorizes examples where LLM-generated and human-written texts were incorrectly attributed or analyzed, providing explanations and analyses of these misattributions.

ample 1 "Ambiguous Features", the model misclassifies text on quantum entanglement as LGTs due to technical jargon usage. However, advanced topics can also be written by human experts, not just LLMs. This text was actually human-generated. Similarly, Example 2 "Surface Features" shows how the model links grammatical errors to the machine. Such mistakes are common among both native and non-native writers and should not be sole indicators of LGTs. In fact, HGTs are more likely to contain grammatical errors. Example 3 illustrates a misjudgment where emotional complexity is falsely attributed exclusively to human writing. The model assumes nuanced emotional contrasts inherently reflect human authorship, overlooking modern LLMs' capability to simulate such depth. This case underscores the unreliability of using emotional sophistication alone as a criterion to differentiate between HGTs and LGTs.

Hallucinations. Hallucinations occur when the model incorrectly attributes features to the text that either do not exist or are contrary to the actual content. In Example 4: Incorrectly Perceived Repetition, the model misinterprets the repetition of ideas about gravitational waves as a sign of LLM authorship. The text does not exhibit excessive repetition, and the claim of a repetitive structure is a false attribution, likely due to biases in the model's training data. In Example 5: Fictitious Absence of Domain Knowledge, the model mistakenly claims that a text about RNA interference lacks technical depth, suggesting it is more likely

Models	Ambiguous Features (%)	Surface Features (%)	Logic&Emotion (%)	Vocabulary (%)	Hallucinations (%)	Incorrect Reasoning (%)
GPT-4o	32.7	12.4	43.2	7.2	2.2	2.3
LLaMA3.3-70B	40.1	25.7	18.9	4.1	6.1	5.1
Qwen2-72B	32.1	23.7	25.4	8.9	6.3	3.6

Table 8: Attribution differences among LLMs when the judgment is correct but the explanation is incorrect.

Models	Ambiguous Features (%)	Surface Features (%)	Logic&Emotion (%)	Vocabulary (%)	Hallucinations (%)
GPT-4o	13.9	30.7	23.8	20.7	10.9
LLaMA3.3-70B	26.8	10.1	40.1	4.1	18.9
Qwen2-72B	33.7	20.1	9.9	26.4	9.9

Table 9: Attribution differences among LLMs when both the judgment and the explanation are incorrect.

to be human-written. In reality, the text contains domain-specific biological content, and the model fails to recognize the technical knowledge present.

Incorrect Reasoning. Incorrect reasoning occurs when relevant features are correctly identified but are misinterpreted, leading to incorrect conclusions. Example 6 highlights a classification error rooted in inconsistent reasoning. The model correctly identifies formal stylistic features but misapplies their significance. Enforcing a binary classification may lead to inconsistent reasoning in model's inference process, as it forces an erroneous LLM label despite ambiguity that could be better captured in a ternary framework.

5.2 Human Evaluation

The reasons for incorrect explanations from human annotators are categorized into two scenarios: correct predictions with incorrect explanations and incorrect predictions with incorrect explanations. The results are summarized in Tables 8 and 9.

For cases where the model made correct predictions but provided incorrect explanations, Table 8 shows that the most prevalent reasons were inaccurate features and hallucinations. Inaccurate features, such as attributing the decision to vague or irrelevant characteristics, accounted for a significant portion of errors across all LLMs. Hallucinations were also frequent, particularly for models like Qwen2-7B and GPT-4o. Faulty reasoning, though less common, contributed to the proportion of incorrect explanations, highlighting inconsistencies in reasoning despite identifying correct features. For cases involving both incorrect predictions and incorrect explanations, Table 9 indicates a similar distribution of error types, but with a higher prevalence of hallucinations. Annotators noted that models hallucinated key features, attributing the decision to features not present in the text, which compounded the issue of misclassification.

Overall, the analysis reveals that hallucinations and reliance on inaccurate features are dominant sources of error in explanations, regardless of prediction accuracy. Addressing these issues requires further refinement of the interpretability mechanisms in LLMs, with a focus on grounding explanations in verifiable and relevant textual evidence.

6 Can we improve LLM-based LGT detection and explanation?

Supervised Fine-Tuning and Reinforcement Learning. To investigate effectiveness of supervised fine-tuning (SFT) on improving cross-LLM detection, we conducted a series of experiments using datasets generated from different LLMs, with or without explanations. Details of data construction are provided in Appendix B, and results are summarized in Table 10. When using answer-only data for SFT, we observe a clear advantage when training data is generated by GPT-4o compared to LLaMA3.3-8B. This likely reflects higher quality and linguistic diversity of GPT-4o generations, whereas LLaMA3.3-8B tends to produce simpler and less informative outputs, offering limited supervision signals. Furthermore, although models fine-tuned on GPT-4o-only data generally achieve higher F1 scores when evaluated on GPT-family outputs, the mixed-source dataset provides stronger generalization across all target LLMs. This indicates that exposure to varied generation styles during training helps the model better capture cross-LLM decision boundaries. Additionally, incorporating explanations into training data consistently yields higher performance. The improvement suggests that explanation-augmented SFT encourages model to internalize task-relevant reasoning patterns and enhances its ability to focus on discriminative linguistic cues indicative of text origin, rather than merely memorizing surface features.

Motivated by recent advances in reward-

Model	Datasets					
	GPT-4o	GPT-4o mini	LLaMA3.3-70B	LLaMA3.3-8B	Qwen2-72B	Qwen2-7B
Qwen2-7B (base model)	59.57	60.13	52.41	49.66	61.24	63.83
SFT (only answer)						
w/ LLaMA3.3-8B	58.28	60.17	52.57	49.85	55.41	59.52
w/ GPT-4o	63.70	62.18	53.90	52.91	63.28	62.86
w/ mixed dataset (ALL)	62.87	63.10	53.28	54.33	64.17	67.91
SFT (answer & explanation)						
w/ GPT-4o	70.09	**72.03**	58.19	54.17	67.39	68.55
w/ mixed dataset (ALL)	68.14	71.09	60.41	54.59	70.72	70.90
GRPO						
w/o cold-start	63.31	64.17	57.96	56.09	63.60	63.02
w/ cold-start	**73.29**	69.01	**63.34**	**65.22**	**74.89**	**73.47**

Table 10: F1 score comparison of various training strategies across different evaluation models. **Qwen2-7B (base model)** represents the raw performance without fine-tuning. **SFT (only answer)** and **SFT (answer & explanation)** denote models supervised-fine-tuned using answer-only or answer-plus-explanation data, respectively. In "SFT w/ GPT-4o" or "w/ LLaMA3.3-8B", the training dataset was generated using the outputs of the specified LLM. "Mixed dataset (ALL)" combines training data from multiple sources. **GRPO** refers to the Group Relative Policy Optimization, evaluated with and without cold-start initialization. **Bold** indicates the best score per column; underline indicates the second-best.

optimized reasoning, such as OpenAI-o1(OpenAI, 2024) and DeepSeek-R1(DeepSeek-AI et al., 2025), we further explore the impact of RL on cross-LLM detection. Specifically, we adopt a GRPO-style reward optimization framework inspired by DeepSeek-R1, with implementation details in Appendix C. As shown in Table 10, applying GRPO without any SFT initialization already brings consistent gains over base model across all test LLMs, with the most notable improvement observed on LLaMA3.3-8B. However, performance remains generally lower than the best SFT configurations. When we initialize GRPO with a model that has been fine-tuned using explanation-augmented, mixed-source data (i.e., the cold-start setting), we observe substantial performance boosts across all evaluation datasets. This combination effectively leverages reasoning capacity learned during SFT and further refines it through reward-driven preference learning. Results demonstrate that GRPO with a well-informed initialization can significantly enhance detection accuracy, enabling the model to better align its scoring behavior with human-intuitive criteria for text provenance.

Enhancing LLM Detection through LLM Collaboration. We further explore whether the performance of LLM-based detection can improve via LLM's collaboration. Table 11 shows that the performance of LLM-based detectors improves significantly when their judgments and explanations are complemented by another LLM counterpart. Specifically, the cross-detection of GPT-4o on Qwen-2 72B dataset has noticeable improvements in the classification and explanation F1 with the support of Qwen2-72B. We also find a similar trend, where Qwen2-72B benefits from GPT-4o's support on the cross-detection settings. These findings indicate that LLM's collaboration can further improve the classification and explanation performance on the LLM counterpart's dataset, i.e., cross-detection.

Models \ Datasets	GPT-4o	Qwen2-72B
GPT-4o (+Qwen2-72B)	+1.43% / -0.61%	+3.79% / +2.33%
Qwen2-72B (+GPT-4o)	+2.01% / +0.45%	+0.39% / -1.24%

Table 11: F1 score differences based on LLM collaboration with judgments and explanations integration. The supplemental LLMs in () will generate the judgments and explanations first to further support the detection and explanations of main LLMs.

7 Conclusion

We evaluated how well LLM-based detectors differentiate human- from LLM-generated text, focusing on detection accuracy and explanation clarity. Self-detection by LLM-based detectors reliably outperforms cross-detection, especially within the same model family. Yet their explanations remain flawed, hinging on spurious features, hallucinations, and unsound reasoning, with GPT-4o trading higher accuracy for frequent hallucinations and Qwen2-7B offering a more balanced but vague rationale. These results underscore the imperative for more interpretable, trustworthy detectors in critical applications such as academic integrity and content moderation.

Limitations

This study is subject to several limitations. First, due to the limited number of API calls available for closed-source LLMs, the datasets used for generating and detecting LLM-generated texts were constructed at a scale of 1,000 samples. As a result, the types and variety of texts involved in the analysis may not be fully comprehensive, potentially introducing bias. Additionally, because the generation of explanations requires manual annotation, which is time-consuming, only a random sample of 100 texts per dataset could be selected for evaluation. This sample size may lead to biases in the evaluation of LLM-generated explanations. Finally, given the rapid advancements in LLM technology, the detection and explanation capabilities of models are continually evolving. Therefore, it is crucial to periodically update our research focus and the models under study to ensure the results remain relevant and accurate.

Ethic Statements

All experiments were conducted using publicly available LLMs and datasets. For the datasets we constructed for the work, no any personal or private information is included. All the three human annotators are co-authors, so an research ethics review was not considered necessary. More details on how we used the human annotators can be found in Appendix D.

References

A. Bhattacharjee and H. Liu. 2024a. Fighting fire with fire: Can chatgpt detect ai-generated text? In SIGKDD Explorations Newsletter, volume 25, pages 1–12.

A. Bhattacharjee and H. Liu. 2024b. Limitations of human identification of automatically generated text.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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

Liam Dugan, Alyssa Hwang, Filip 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. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), page 12463–12492.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378–382.

J. Frank et al. 2024. The turing test of online reviews: Can we tell the difference between human-written and gpt-4-written online reviews? Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP), pages 55–73.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. arXiv:2401.12070 [cs.CL].

Zhiyuan He, Rui Liu, Xiaojun Liu, and Yiming Huang. 2024. Smaller language models are better zero-shot machine-generated text detectors. arXiv preprint arXiv:2400.00000.

B. Hu. 2025. Unveiling ambiguity: Chatgpt vs. human writing. Applied Economics Letters, pages 1–9.

Edward Hu, Xiang Peng, Xuehai Liu, Le Song, Furu Zhang, and Xihong Wang. 2021. Lora: Low-rank adaptation of large language models. In Proceedings of the 38th International Conference on Machine Learning (ICML).

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

IDEATE: Detecting AI-Generated Text Using Internal and External Factual Structures. 2024. Lrec-coling 2024.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting machine-generated texts: Not just "ai vs humans" and explainability is complicated. arXiv preprint arXiv:2406.18259.

J. Lee et al. 2023. Do language models plagiarize? In Proceedings of the ACM Web Conference 2023 (WWW '23), pages 3637–3647.

Jooyoung Lee, Toshini Agrawal, Adaku Uchendu, Thai Le, Jinghui Chen, and Dongwon Lee. 2024a. Plagbench: Exploring the duality of large language models in plagiarism generation and detection. arXiv preprint arXiv:2406.16288.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2024b. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In NAACL 2024 Findings.

Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. arXiv preprint arXiv:2501.04040. 174 pages, to be submitted to a journal in a shorter version. Includes figures taken from papers by other authors. All sources have been referenced.

OpenAI. 2024. GPT-4O. Online.

OpenAI. 2024. Openai o1. https://openai.com/o1/. Accessed: 2025-05-18.

Qwen2 Team. 2024. Qwen2 technical report. arXiv preprint arXiv:2407.10671.

Edward Tian, Alex Cui, Olivia Kusio, et al. 2023. GPTZero. Online.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. arXiv:2302.13971 [cs.CL].

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. 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 12493–12523. Association for Computational Linguistics.

L. Weng, S. Liu, et al. 2024. Towards an understanding and explanation for mixed-initiative artificial scientific text detection. volume 23, pages 272–291.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.

Y. Zhou, B. He, and L. Sun. 2024. Humanizing machine-generated content: Evading ai-text detection through adversarial attack. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 8427–8437. ELRA and ICCL.

A Misattributions Examples

To better illustrate the types of errors and their attributions, Table 12 provides examples of common mistakes made by the models in classifying text origins.

B Supervised Fine-Tuning (SFT)

B.1 Dataset Construction

For each language model (e.g., GPT-4o, LLaMA3.3-8B), we first generate synthetic outputs and then perform self-detection using the same LLM. Based on whether the self-detection is correct, we apply the following rejection sampling procedure to construct our SFT datasets:

- If the model correctly identifies its own output, we retain the sample.
- If the model misclassifies, but a stronger model (e.g., GPT-4o) can identify the source

Example 2: Surface Features
Original Text: "The book were interesting. It show a deep exploration of human emotions and relationships. The author's use of metaphors and descriptive language provided rich insights into the human psyche, but the writing style seemed awkward and unpolished in places."
LLM Judgment: LLM-generated
LLM Explanation: "The grammatical errors such as 'were' instead of 'was' and 'show' instead of 'shows' suggest that the text is AI-generated. Machines typically generate text with such mistakes."

Example 3: Misattributed Logical or Emotional Elements
Original Text: "His words carried a strange contradiction—he spoke of hope but with an undeniable sadness. It was as though the optimism in his tone clashed with his deeply melancholic gaze, creating a complex emotional paradox."
LLM Judgment: Human-written
LLM Explanation: "The nuanced contradiction of hope and sadness reflects deeply human emotional complexity, which LLMs typically struggle to replicate authentically. Such layered emotional depth is often unique to human expression."

Example 5: Fictitious Absence of Domain Knowledge
Original Text: "RNA interference plays a crucial role in gene regulation, allowing for targeted suppression of specific genetic sequences. This mechanism is being studied for its potential therapeutic applications in cancer treatment and genetic disorders."
LLM Judgment: Human-written
LLM Explanation: "The text lacks the level of technical depth and domain-specific knowledge typically found in professional scientific writing, suggesting that it is more likely to be written by a human author who is not an expert in the field."

Table 12: Analysis of LLM vs. Human Writing Attribution Based on Various Features. The table categorizes examples where LLM-generated and human-written texts were incorrectly attributed or analyzed, providing explanations and analyses of these misattributions.

correctly and provide an explanation, we include this explanation-augmented sample.

- If all models fail to detect the text's origin correctly, the sample is discarded, as it offers limited training value.

This strategy ensures that our training data are high-quality and informative, containing either answers alone or answers with explanations. It helps the SFT model learn from both model-generated cues and external reasoning signals.

B.2 Training Configuration

Each SFT experiment is conducted using a balanced dataset of 10,000 LLM-generated and 10,000 human-written texts. We perform LoRA-based(Hu et al., 2021) fine-tuning for two epochs using four NVIDIA A100 80GB GPUs with FP16 precision.

- **Answer-only setting:** Each training session takes approximately 18 hours.
- **Answer + explanation setting:** Training requires around 28 hours due to longer input sequences and richer supervision.

We use a batch size that fully utilizes available GPU memory, a learning rate of 2e-5, and the AdamW optimizer. Model performance is evaluated after each epoch based on macro F1 score, and the best-performing checkpoint is selected.

C Reward Optimization with GRPO

C.1 Dataset Construction

For GRPO training, we construct a high-quality, moderately difficult dataset. We filter out both overly simple and excessively hard samples from the LLM-generated pool:

- Samples that are correctly classified by all detectors are excluded as they lack training challenge.
- Samples that are misclassified by all detectors are removed because they may be inherently ambiguous.

From the remaining samples, we randomly select 5,000 LLM-generated texts and mix them with 5,000 human-written texts to form a balanced dataset for reward learning.

C.2 Training Configuration

GRPO training is conducted using the same infrastructure as SFT. We initialize the model either from a base Qwen2-7B checkpoint or from a previously SFT-trained model (cold-start setup). Training is performed using a PPO-style loop with KL-divergence regularization.

- **Without SFT:** Training takes approximately 24 hours.
- **With SFT:** Training requires about 30 hours due to improved convergence and longer sequences.

All training is done with 4 NVIDIA A100 80GB GPUs, using FP16 precision, a reward model learning rate of 1e-5, a policy learning rate of 5e-6, and gradient accumulation for stability. The total training loop runs for 1 epoch with a replay buffer size of 10k examples.

D Annotation Guidelines

This appendix provides detailed instructions for the manual annotation tasks conducted in our study. The annotation process consists of two tasks: (1) a binary classification task to evaluate the accuracy of LLM-generated explanations, and (2) a ternary classification task where annotators evaluate both the correctness of the LLM's ternary classification judgments and the accuracy of its explanations based on known text sources.

E Examples of Undecided Text Categories

E.1 Mixed Human-LLM Co-authored Texts

This category includes texts collaboratively written by humans and large language models (LLMs), where stylistic or structural transitions reflect a shift in authorship. Such texts often begin with nuanced and context-sensitive human input and later transition to more uniform, templated, and encyclopedic LLM-generated content. These transitions are not always clearly marked, making authorship attribution difficult. Table 14 presents two representative examples.

E.2 Inherently Ambiguous Single-source Texts

This category includes texts produced entirely by either humans or LLMs, but whose stylistic and rhetorical features align with both sources. Such texts often exhibit emotionally neutral tones, fact-heavy content, and well-structured reasoning, making it difficult to distinguish their origin. Human-written texts may appear too polished, while LLM outputs may mimic human nuance. Table 15 illustrates two such ambiguous examples.

E.3 Fragile Indicator Cases

This category includes texts that trigger LLM detectors based on subtle linguistic patterns or statistical anomalies. However, these indicators tend to be unstable. Minor changes in wording, paraphrasing, or model sampling parameters often cause the features to disappear. As a result, these texts demonstrate the brittleness of current detection methods. Table 16 presents two examples with unstable features that resist robust attribution.

Task 1: Explanation Accuracy Evaluation
Annotators will assess whether the explanation provided by the model correctly justifies its classification decision. Each explanation should be judged based on its logical consistency, relevance to the text, and whether it accurately reflects distinguishing features.
Columns:

- **Text**: The text sample to be classified.
- **Detection Result**: The model's classification of the text as "LLM-generated" or "Human-generated."
- **Model's Explanation**: The explanation provided by the model for its classification decision.
- **Annotation**: Label the explanation as "Accurate" or "Inaccurate" based on its reasoning quality.

Example:
Text: "In recent years, artificial intelligence has demonstrated remarkable progress, influencing numerous industries, including healthcare, finance, and creative writing. Many experts believe that this rapid advancement will continue to reshape the workforce and redefine human-machine collaboration."
Detection Result: LLM-generated
Model's Explanation: "The highly structured argumentation and precise yet impersonal tone indicate that this text was likely machine-generated rather than composed by a human writer."
Annotation: Inaccurate (While structured argumentation is common in LLM-generated text, human authors can also produce similarly structured and objective writing.)

Task 2: Ternary Classification with Explanation Evaluation
Annotators will classify each text as "LLM-generated," "Human-generated," or "Undecided," and evaluate whether the model's explanation correctly justifies the classification. The "Undecided" label applies when the text lacks sufficient distinguishing features.
Columns:

- **Text**: The text sample to be classified.
- **Detection Result**: LLM' judgment of whether the text is "LLM-generated," "Human-generated," or "Undecided."
- **Model's Explanation**: The explanation provided by the model.
- **Classification Annotation**: Label whether the model's classification is "Correct" or "Incorrect."
- **Explanation Annotation**: Label the explanation as "Accurate" or "Inaccurate" based on its reasoning quality.

Example:
Text: "Quantum mechanics, a field of physics that describes the behavior of particles at the atomic and subatomic levels, has led to groundbreaking discoveries such as quantum entanglement and superposition. These principles have paved the way for advancements in quantum computing and cryptography, revolutionizing modern technology."
Detection Result: Undecided
Model's Explanation: "The text presents factual scientific content in a neutral tone, making it difficult to distinguish between human and machine authorship."
Classification Annotation: Correct
Explanation Annotation: Accurate (The explanation correctly justifies why the text is indistinguishable.)

Table 13: Human Annotation Instructions

Example 1
"In examining urban resilience frameworks, we find that grassroots organizations play a pivotal role in ensuring community adaptability. Interviews with local leaders in Jakarta revealed bottom-up innovation and resource sharing as key drivers of resilience. However, literature on climate-adaptive infrastructure increasingly emphasizes machine learning for real-time flood prediction and decentralized data systems for disaster response. A systematic review of recent publications demonstrates that hybrid models integrating sensor-based monitoring with socio-political data yield more actionable insights. These models offer scalable solutions for cities facing climate uncertainty, as evidenced by pilot projects in Southeast Asia and Latin America."

Example 2
"The first wave of digital humanities emphasized textual digitization and basic metadata annotation, grounded in traditional philological practices. Scholars argued for methodological transparency and historical fidelity. In recent years, however, there has been a shift toward large-scale computational analysis. Transformer-based models are now trained on digitized archives to identify latent narrative patterns across centuries. This methodological turn, while powerful, raises questions about interpretability and disciplinary boundaries. Current debates focus on integrating humanistic inquiry with algorithmic inference in ways that preserve epistemic integrity."

Table 14: Examples of Mixed Human-LLM Co-authored Texts

Example 1 "The monitor offers a 2560×1440 resolution, a 165Hz refresh rate, and a color accuracy rating of Delta E < 2. These specifications make it suitable for both competitive gaming and semi-professional design work. Its adjustable stand and blue-light reduction features enhance long-term usability. In testing, response times remained consistent across refresh rates, and input lag was minimal. While the built-in speakers are underwhelming, the overall design reflects thoughtful engineering. Prospective buyers should note that firmware updates may be required to access advanced color profiles."
Example 2 "The exhibition explores post-colonial identity through mixed media installations that juxtapose industrial debris with archival imagery. Each piece is accompanied by minimal curatorial framing, allowing for open-ended engagement. The spatial arrangement avoids linear narratives, instead emphasizing fragmented temporality and layered symbolism. Visitor responses ranged from emotional introspection to conceptual confusion. Whether the ambiguity is intentional or a result of aesthetic overreach remains debatable, yet the collection undeniably provokes sustained reflection."

Table 15: Examples of Inherently Ambiguous Single-source Texts

Example 1 "The second quarter's economic indicators reflect a modest uptick in consumer confidence, despite lagging wage growth and persistent inflationary pressures. Analysts note that housing starts have stabilized, though regional variation remains high. Meanwhile, the energy sector showed unexpected resilience due to global supply chain recalibrations. Although many forecasts anticipated stagnation, revised models suggest a delayed soft landing. The Federal Reserve's policy stance continues to oscillate between caution and proactive intervention, with uncertainty surrounding long-term bond yields."
Example 2 "In the novel's final chapter, the protagonist revisits the childhood home, now transformed by decay and silence. The narrative perspective shifts from third-person to free indirect discourse, blurring the boundary between memory and present perception. Sentence rhythms slow, with nested subordinate clauses evoking emotional weight. Yet, this stylistic density is briefly interrupted by abrupt declaratives, mirroring the character's psychological fragmentation. Such microstructural choices are atypical but could be easily altered in paraphrased variants, rendering authorship signals imperceptible to automated systems."

Table 16: Examples of Fragile Indicator Cases