Detecting Machine-Generated Texts: Not Just "AI vs Humans" and Explainability is Complicated

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) rapidly advance, increasing concerns arise regarding risks about the actual authorship of texts we see online and in the real world. The task of distinguishing LLM-authored texts is complicated by the nuanced and overlapping behaviors of both machines and humans. In this paper, we challenge the current practice of considering the LLM-generated text detection a binary classification task of differentiating human from AI. Instead, we introduce a novel ternary text classification scheme, adding an "undecided" category for texts that could be attributed to either source, and we show that this new category is crucial to understand how to make the detection result more explainable to lay users. This research shifts the paradigm from merely classifying to explaining machine-generated texts, emphasizing the need for detectors to provide clear and understandable explanations to users. Our study involves creating four new datasets comprised of texts from various LLMs and human authors. Based on the new datasets, we performed binary classification tests to ascertain the most effective state-of-the-art (SOTA) detection methods and identified SOTA LLMs capable of producing harder-to-detect texts. Then, we constructed a new dataset of texts generated by the two top-performing LLMs and human authors, and asked three human annotators to produce ternary labels with explanation notes. This dataset was used to investigate how three top-performing SOTA detectors behave in the new ternary classification context. Our results highlight why the "undecided" category is much needed from the viewpoint of explainability. Additionally, we conducted an analysis of explainability of the three best-performing detectors and the explanation notes of the human annotators, revealing insights about the complexity of explainable detection of machine-generated texts. Finally, we propose guidelines for developing future detection systems with improved explanatory power.

1 Introduction

With the rapid evolution of Large Language Models (LLMs) such as ChatGPT-4 (OpenAI, 2023), the sophistication and human-like quality of texts generated by these models have notably increased, enabling them to produce diverse content in response to specific prompts. These advancements bring not only numerous practical applications but also raise significant challenges including potential academic fraud and actual authorship. Extensive research has been undertaken to differentiate between machine-generated texts (MGTs) and human-generated texts (HGTs), primarily employing model-based approaches (Wang et al., 2023; Bhattacharjee et al., 2023) and statistical methods that analyze inherent text characteristics (Hans et al., 2024; Bao et al., 2024; Zhang et al., 2024). Several online platforms such as GPTZero (Tian et al., 2023) and Sapling (Sapling AI Team, 2023) have also demonstrated robust capabilities in differentiating MGTs from HGTs.

Traditionally, the detection of MGTs has relied on a binary classification framework that discerns between MGTs and HGTs. However, the boundaries between MGTs and HGTs are increasingly ambiguous due to the rapid enhancements in LLMs, thereby complicating the effectiveness of simple binary classification systems. For instance, in statistical detection, the characteristics of a given MGT might deviate significantly from typical MGTs patterns and mirror those of HGTs, leading to a misclassification. Model-based methods often struggle with generalization as they tend to learn features that are specific to the data they are trained on (usually limited to one or more specific LLMs), which may not necessarily work as new models emerge. Moreover, many existing detection systems lack an explainability component. Although some detection methods attempt to consider explanatory features, their effectiveness in providing meaningful insights appears limited according to our evaluations of one such methods (an online closed-source detector) GPTZero (Tian et al., 2023). This shortfall emphasizes a critical gap: the need for enhanced explainability in MGT detectors to improve end users' trust in such systems.

In order to address these limitations, our study introduces a novel ternary classification system for analyzing texts. Recognizing that some texts may simultaneously share characteristics of both MGTs and HGTs, we have added an "undecided" category to our classification framework. We developed a ternary classification dataset and designed experiments to test the validity of this approach. Our methodology not only includes rigorous statistical and model-based analyses, but also incorporates detailed human evaluations to provide a nuanced understanding of the new ternary text classification task and the complexity of producing human-understandable explanations. By comparing the explanatory power of human assessments with that of automated detectors, we highlight the current explanatory limitations faced by MGT detectors.

Through some binary classification experiments based on four new datasets covering multiple state-of-the-art (SOTA) LLMs, we established that the most advanced LLMs currently available are ChatGPT-4 and ChatGPT-3.5, in terms of defeating multiple SOTA MGT detectors. The detectors that performed the best in our experiments are GPTZero (Tian et al., 2023), Sapling (Sapling AI Team, 2023) and Binoculars (Hans et al., 2024). Building on these findings, we crafted a ternary classification dataset using texts from the aforementioned top-performing LLMs. We organized human coders to annotate these texts, applying the ternary classification framework and providing detailed explanations for their decisions. Subsequent experiments with the top three detectors proved the limitations of binary classification so that the new "undecided" category should be seriously considered in future research on MGT detection. Our comparative analysis between the human-provided explanations and those offered by the detector GPTZero illuminated significant gaps in current automated explanations. While human explanations provide valuable insights, they also exhibit inherent limitations and imply the complexity and challenges behind developing more explainable MGT detectors.

In conclusion, our research not only challenges existing paradigms in MGT detection but also sets a foundation for future innovations in detector design, particularly in enhancing explainability. This work suggests new directions for the development of detection systems that are not only effective but also transparent and interpretable to users.

2 Related Work

2.1 Open-Source Detectors

Zero-shot detection. This approach leverages some unique statistical properties distinguishing MGTs from HGTs. Past studies have employed various linguistic model-derived characteristics, such as entropy (He et al., 2023), average log-probability scores (Solaiman et al., 2019) and perplexity (Wu et al., 2023), as useful statistical properties for detection. With the evolution of LLMs that generate increasingly sophisticated texts, more recent zero-shot detection strategies (Gehrmann et al., 2019; Mitchell et al., 2023; Su et al., 2023; Wu and Xiang, 2023; Bao et al., 2024; Kumari et al., 2024) have adapted to discern high-order features of advanced text generators. Notably, the Binoculars model (Hans et al., 2024) leverages LLMs to perform next-token predictions at each text position, utilizing the log perplexity ratio compared to the baseline text as a distinguishing statistic.

Model-based detection. This approach involves adapting existing models to learn from specific datasets for MGT detection (OpenAI, 2021; He et al., 2023). It often includes sentence-level detection and analyses different LLM outputs (Wang et al., 2023; Bhattacharjee et al., 2023; Antoun et al., 2023). However, these methods can suffer from overfitting and generally exhibit limited effectiveness in detecting texts across various domains.

Other approaches. There are also other approaches based on watermarking, adversarial learning based training, and human assistance (Wu et al., 2024). These approaches are more complicated and are often a mixture of different approaches, so in this paper we consider two basic approaches only to make our work more focused.

2.2 Online Close-Source Detection Systems

Despite their closed-source nature, online detectors are of significant interest in academic research (Yang et al., 2023). For instance, GPTZero (Tian et al., 2023) integrates several analytical components that predict if a piece of text is generated by machine or human with a confidence score, together with a sentence-by-sentence analysis capability. Similarly, Sapling (Sapling AI Team, 2023) utilizes a transformer-based architecture akin to those found in generative AI systems. Moreover, various platforms offer an online MGT detection tool for all to use (Originality.AI, 2024; Copyleaks Technologies Ltd., 2023; Inspera, 2023; ZeroGPT.com, 2023).

2.3 Explainability in Current Detectors

According to its official documentation (Tian et al., 2023), GPTZero uses the following six features to achieve explainability: readability, percent SAT, simplicity, perplexity, burstiness, and average sentence length. However, it does not provide clarity on how these features influence its final judgments. Other efforts have focused on integrating explanatory modules into detectors. One study (Mitrović et al., 2023) implemented Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017), which assigns importance values to each feature, enhancing the interpretability of decisions in text source detection. Another investigation (André et al., 2023) computed textual attributes such as perplexity, grammar, and n-gram distributions to measure their effects on detection outcomes. Despite these advancements, the current state of detector explainability remains challenging for lay users to comprehend.

3 Binary Classification Evaluation of Detectors on MGTs and HGTs

This section outlines the assessment of state-of-the-art (SOTA) MGT detectors through binary classification tests on datasets containing both MGTs and HGTs. Our objective is to identify the most effective and consistently accurate detectors across various datasets and to pinpoint LLMs that exhibit the strongest generative abilities and human-like output. This process will involve binary classification trials using custom-built datasets. The selected detectors and LLMs will then be utilized in further experimental investigations.

3.1 Experimental Design

We conducted our experiments using four datasets specifically constructed for this study. It is crucial to carefully select LLMs for text generation and appropriate sources of HGTs to assemble the dataset. We chose a mix of open-source and closed-source SOTA MGT detectors for evaluation and used standard performance metrics for the binary classification tests.

LLMs. For text generation, we have opted for widely recognized models including the closed-source ChatGPT-3.5 (OpenAI, 2022) and ChatGPT-4 (OpenAI, 2023), known for their robust performance. Additionally, we selected Google's Gemini Pro (Hassabis and the Gemini Team, 2023), renowned for its ability to produce coherent and high-quality natural language outputs. From the open-source domain, we have chosen the LLaMA series (Touvron et al., 2023), specifically using LLaMA-13B for generating machine texts based on our computation resources.

HGT sources. To ensure a diverse and representative collection of HGTs, we included selections from public datasets such as the HC3 dataset (Guo et al., 2023), which contains texts from four other public Q&A datasets and data crawled from Wikipedia. Notably, it includes a category of texts, similar to the ELI5 ("Explain Like I'm Five") format (Fan et al., 2019), where complex issues are explained in simple terms. We also extracted short texts from the IDMGSP dataset (Abdalla et al., 2023), which comprises titles, abstracts, introductions and conclusions of human-authored scientific papers, alongside texts manually selected from X/Twitter using tag searches to cover topics of everyday discourse. This blend of sources provides a broad spectrum of topics and writing styles in the human-generated texts within our dataset.

MGT Detectors. Initially, we chose GPTZero and Sapling as the leading commercial (online and closed-source) detectors from the proprietary sector. We then extended our selection to include several notable open-source detectors such as Binoculars (Hans et al., 2024), Fast-DetectGPT (Bao et al., 2024), MMD-MP (Zhang et al., 2024), DEMASQ (Kumari et al., 2024), and Detect-GPT (Mitchell et al., 2023).

Custom-built Datasets. Four datasets were built using the selected LLMs and HGT sources, as detailed in Table 1. To control variables in subsequent analyses, the HGTs within these datasets were maintained consistently across all experiments. This standardization can help isolate the variable effects of different LLM outputs on detector performance.

Dataset	MGTs	HGTs
D1	100 (generated by ChatGPT-4)	100
D2	100 (generated by ChatGPT-3.5)	100
D3	100 (generated by LLaMA-13B)	100
D4	100 (generated by Gemini Pro)	100

Table 1: Composition of the four datasets. The texts cover a wide range of topics including economics, healthcare, science, literature, sports, and daily life.

Evaluation metrics. The detectors are expected to maximize MGT detection accuracy while minimizing false positives among HGTs. Therefore, Precision, Recall, and F1 scores for MGTs are selected as primary evaluation metrics. Other metrics, such as the macro F1 score across two classification situations (MGTs and HGTs as positive samples, respectively), have also been used to provide a comprehensive assessment of detector performance.

3.2 Results

We evaluated various detectors on datasets, as detailed in Table 2, focusing on the dataset generated by ChatGPT-4. This table highlights the performance of detectors using both humans and machines as the positive label. The results indicate that online detectors, GPTZero and Sapling, significantly outperform local open-source counterparts. Specifically, DEMASQ effectively identifies MGTs but struggles with HGT detection. Conversely, DetectGPT shows limited capability in detecting MGTs while performing adequately with HGTs. See Appendix A for extended results for other datasets.

Figure 1 visually compares F1 scores of all tested MGT detectors across all four datasets, confirming the superior performance of GPTZero and Sapling over local models. Among the latter, Binoculars ranks the highest, demonstrating a consistent performance across all datasets, suggesting its being less susceptible to overfitting compared to other local models. Further analysis reveals that texts generated by ChatGPT-3.5 and ChatGPT-4 are generally more challenging to classify across all detectors, compared to those generated by LLaMA-13B and Gemini Pro, implying that ChatGPT-3.5 and ChatGPT4 can produce texts that more closely resemble human writing. Based on these findings, for the further experiments and discussions about the new ternary classification framework and the complexity of explainability, we chose to use a mixed dataset with texts generated by ChatGPT-3.5 and ChatGPT-4, and HGTs. Similarly, on the selection of MGT detectors, we focused on three top-performing ones, GPTZero, Sapling, and Binoculars.

4 Ternary Classification Tests for Selected MGT Detectors

4.1 Manual Annotation and Explanations

Following the outcomes from binary classification experiments, we compiled a new dataset containing texts from ChatGPT-4, ChatGPT-3.5, and human authors. The dataset consists of 200 texts, with 50 from ChatGPT-4, 50 from ChatGPT-3.5, and 100 from human authors. Three co-authors of this paper, who are all Computer Science undergraduate students, annotated the 200 texts to categorize each text into one of three groups: human, machine, and undecided. They also provided explanation notes to justify their annotation results. Each annotator first independently annotated the 100 texts and also indicated their level of confidence for each label. After all the three annotators finished their work, we calculated Fleiss' kappa (Fleiss, 1971), which was 0.1377, indicating a low level of agreement among the annotators. To address the disagreements, all authors entered into a collaborative discussion on the texts with different opinions, without revealing the ground truth to the annotators, and the annotators were asked to refine their annotations. After the annotations were updated, we calculated Fleiss' kappa again, which increased to 0.9438, reflecting a near-complete consensus among all annotators. Any texts that remain to have no consensus were labeled as "undecided". The explanation notes of the three annotators were merged and refined to be more consistent after the first author discussed with the three annotators and other co-authors. More details of the dataset can be found in Table 3, which shows that all human annotated MGTs and HGTs are 100% correct according to the ground truth labels. The high percentage of undecided texts itself is indicative and already shows that the traditional binary classification approaches may be problematic. More information about how the human annotators' work is given in Appendix I.

The human annotation results revealed that, although some automated MGT detectors have achieved very good performance in predicting ground truth labels, human annotators were clearly not convinced by the cases falling into the "undecided" category. This can be partly explained

Models	Accuracy	Machine as Positive			Human as Positive			Macro F1
		Precision	Recall	F1	Precision	Recall	F1	
GPTZero	97.28%	96.84%	97.87%	97.35%	97.75%	96.67%	97.21%	97.28%
Sapling	90.67%	84.96%	98.97%	91.43%	98.75%	82.29%	89.77%	90.60%
Binoculars	86.50%	78.74%	100.00%	88.11%	100.00%	73.00%	84.39%	86.25%
Fast-DetectGPT	73.50%	88.52%	54.00%	67.08%	66.91%	93.00%	77.82%	72.45%
MMD-MP	71.00%	93.75%	45.00%	60.81%	63.82%	97.00%	76.98%	68.90%
DEMASQ	65.50%	59.51%	97.00%	73.76%	91.89%	34.00%	49.64%	61.70%
DetectGPT	52.00%	75.00%	6.00%	11.11%	51.04%	98.00%	67.12%	39.12%

Table 2: Binary classification performance of different detectors on the dataset of ChatGPT4

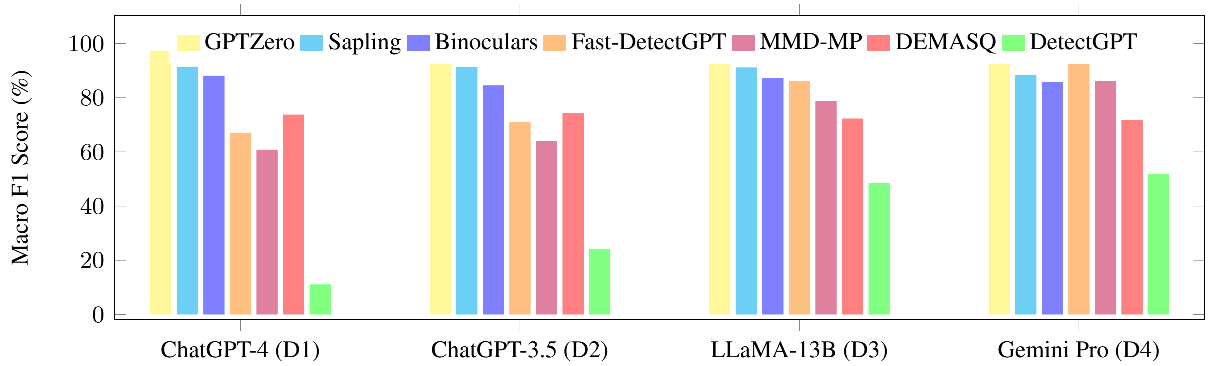


Figure 1: Comparison of detector performance across the four datasets produced by various LLMs, with MGTs as positive samples. The x-axis represents different datasets, while different bars represent different detectors.

Human Annotation	Total	GT: Machine	GT: Human
Machine	21	21	0
Human	85	0	85
Undecided	94	79 (84.04%)	15 (15.96%)

Table 3: Comparison between human annotations and ground truth (GT) labels.

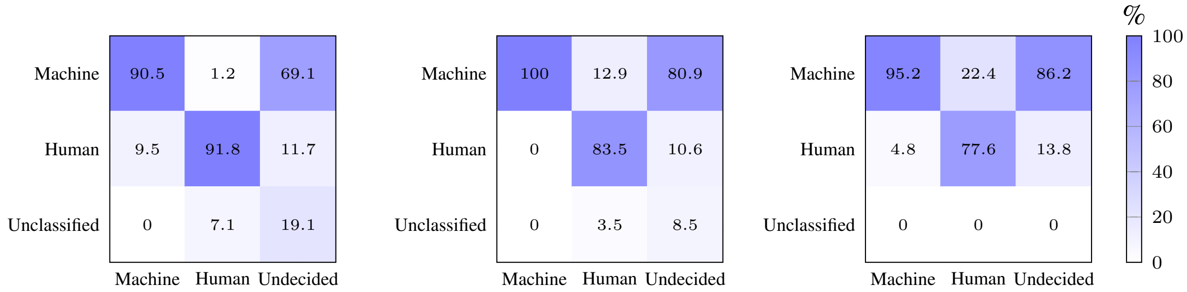
by what an ideal machine-based text generator is supposed to do – to produce texts that are HGTs. Although we may argue that SOTA LLM-based generators are still far from ideal, the human annotators clearly have seen many example MGTs that are sufficiently human-like so that there is no convincing way to label them as just MGTs or HGTs, so "undecided" would be a better class to describe them.

4.2 Method

Using the new dataset with ternary labels, we investigated how the three top-performing binary MGT detectors performed in the context of the ternary classification task. We generated 3×2 confusion matrices to observe how the three different types of texts, particularly those in the new "undecided" category, are classified by the MGT detectors.

4.3 Results

The confusion matrices for the detectors GPTZero, Sapling, and Binoculars, detailed in Fig. 2, reveal that, while the detection accuracy is high for clearly defined MGTs and HGTs (which was expected based on the results of the binary classification experiments reported in the previous section), challenges persist with the "undecided" texts. The most interesting pattern is that all three detectors are clearly biased on texts labeled as "undecided": they all have a clear tendency to classify such texts as MGTs. This bias is largely aligned with the biased percentage of MGTs in the "undecided" category as shown in Table 3. Considering that human annotators considered such texts difficult to judge, it is likely also difficult for the MGT detectors to explain why they consider such texts generated by either machines or humans. Another interesting observation is that, both Sapling and Binoculars have a much higher error rate for HGTs than for MGTs labeled by our human annotators, implying HGTs may be generally harder to detect than MGTs for most detectors. GPTZero does not seem to suffer from this problem, but due to its closed-source nature it is unclear how it achieved such a performance.



(a) GPTZero (b) Sapling (c) Binoculars

Figure 2: Confusion matrices showing how the three binary MGT detectors performed in a ternary classification setting. For GPTZero and Sapling, some texts were considered too short so no classification results were given.

5 Explainability of Detectors

The results in the previous section indicate the importance for binary MGT detectors to explain their results to human users, which is particularly important for texts in the "undecided" category since human users may not agree on binary labels for such texts, not mentioning the results from an automated MGT detector. In this section, we report our analysis of GPTZero, the only MGT detector with an explainability element out of the three we considered in the previous section, and also our analysis of explanation notes given by the three human annotators who constructed the ternary dataset we used.

5.1 Analysis of GPTZero's Explainability

Different from Sapling and Binoculars, which do not provide any explanation to their results, GPTZero offers the following six concrete metrics to offer some level of explainability to their results: readability, percent SAT words, simplicity, perplexity, burstiness, and average sentence length. Other than giving values of the metrics, it does not clarify how they affect its decision-making process.

Table 4 shows an example, comparing the six explainability metrics used by GPTZero and the explanation notes given by our human annotators. As can be seen, the metrics used by GPTZero has limited explanatory power because they are too abstract. For instance, all the six metrics are marked as "Medium", which does not explain why the final judgment is AI. Instead, "Medium" may better fit into the "undecoded" category of our ternary classification framework, as what the human annotators stated in their more human-understandable explanation notes.

A further empirical analysis was performed to study how the six explainability metrics claimed by GPTZero affect the final results. We constructed a dataset using texts in the datasets we used in previous sections, and used the six metrics as the input features and the GPTZero's detection results as the target class labels. We used an 80-20 training-testing split and applied various traditional machine learning models including logistic regression (Cox, 1958), SVC (Cortes and Vapnik, 1995), perceptron (Rosenblatt, 1958), and decision tree (Breiman et al., 1984). The results showed that two metrics, Readability and Perplexity, significantly affect the GPTZero's decision-making, while other metrics played a minor role. Yet, the accuracy rates of all models stayed below 80%, implying that GPTZero uses other features and/or mechanisms to achieve its much higher performance observed in Section 3. For a comprehensive breakdown of these results, refer to Appendix B.

5.2 Explanation Categories Provided by Human Annotators

Our analysis of human annotators' explanation notes revealed eight primary categories, each detailed in Appendix C.

Spelling errors. These involve inaccuracies in word composition such as omissions, insertions, substitutions, transpositions, and phonetic mistakes. Human-written texts tend to include spelling errors more frequently, whereas machine-generated texts seldom display these errors. The difference may be explained by the higher likelihood for human users to misspell words especially in informal writing.

Grammatical errors. These include verb conjugation mistakes, subject-verb disagreements, wrong usage of articles, and misuse or absence of punctuation marks. Such errors are more prevalent in HGTs, likely due to the same reason as spelling

Source: ChatGPT-4 **Text:** Sweating itself does not directly cause colds. Colds are caused by viruses, not by being cold or sweating. However, if you sweat and then get chilled, this might weaken your immune system temporarily, making you more susceptible to catching a cold virus. Additionally, the belief that sweating leads to colds might stem from confusing the symptoms of a cold, which can include sweating, with the cause of the cold.
GPTZero result: AI **GPTZero explanations**: Readability: 72.3 (Medium) | Percent SAT: 1.7 (Medium) | Simplicity: 35.2 (Medium) | Perplexity: 45.3 (Medium) | Burstiness: 37.9 (Medium) | Average sentence length: 22.3 (Medium)
Human labels: undecided **Human explanations**: The text is free from grammatical and spelling errors. This passage elucidates the relationship between sweating and colds, maintaining an objective and rigorous tone. It encompasses both common knowledge and scientific principles. The structure of the text is clear, with adverbial usage enhancing the clarity and fluency of the sentences. The text avoids unnecessary repetition, making it readily comprehensible. Therefore, it should be categorized as "undecided."

Table 4: Comparison between abstract scores from GPTZero and human-readable explanations

errors. On the other hand, it is more sensible that LLMs follow linguistic rules more rigorously, even for informal writing.

Perplexity. This metric evaluates how well a model predicts a text. Higher perplexity indicates more unpredictability and diversity, common in HGTs due to natural creativity. MGTs, on the other hand, tend to adhere to predictable patterns, resulting in lower perplexity.

Logical errors. These occur when there are flaws in reasoning or the logical flow of the text. HGTs might contain occasional logical errors due to typographical oversights, while MGTs can exhibit more obvious and harder-to-explain logical inconsistencies due to limitations in processing complex logical relationships.

Unnecessary repetition. This refers to the excessive repetition of words or phrases without a clear purpose. MGTs often repeat content due to the constraints of their generation models, whereas humans use repetition strategically to emphasize points and maintain a coherent narrative flow.

Readability. This evaluates text clarity based on factors like sentence length and word complexity. HGTs usually score higher for readability because human writers tend to use simpler language and more straightforward sentence structures. MGTs may incorporate more complex vocabulary and sentence structures, often lowering their readability scores.

Text structure. This pertains to how texts organize content using elements like adverbs, prepositions, and bullet points. MGTs tend to use these structural elements more frequently, while HGTs often exhibit a more natural flow with less reliance on rigid structuring.

Bias. This indicates the presence of prejudicial or favoring tendencies in a text. HGTs are more likely to reflect personal or societal biases, while MGTs generally show fewer biases, though they can still mirror biases present in their training data.

These categories helped our human annotators to be more certain on some HGTs and MGTs. However, texts lacking definitive features were categorized as "undecided" based on the absence of clear human or machine indicators.

6 Further Discussions

6.1 Justification for Ternary Classification

The introduction of the "undecided" category has sparked a considerable debate concerning its validity. For instance, a text in Table 4 was categorized as "undecided" by our human annotators, whereas detection tools like GPTZero, Sapling, and Binoculars identified it as MGT – a classification that is technically correct. However, according to our human annotators, these texts were aptly placed in the "undecided" category, arguing that there was no definitive reason to label them strictly as MGTs, suggesting instances where LLMs might merely be mimicking human-like output. More examples of this kind can be found in Appendix G.

Upon reviewing the explanation notes provided by our human annotators, we observed that characteristics of MGTs and HGTs often overlap across several categories. This overlap creates a gray area in determining the origin of the text, as the boundaries between MGT and HGT are not always clear-cut. Moreover, since MGTs are trained on and derive from HGTs, they can produce texts that are indistinguishable from human writings.

Although it is apparent that human annotators struggled with accurately distinguishing the "undecided" category from the other two, this ambiguity also underscores the complexity of text generation origins. Despite these challenges, the ternary classification provides a framework that can guide further refinement in identifying and differentiating these text categories. Future efforts should focus on establishing more precise criteria to discern the unique characteristics and distinctions among these three labels.

6.2 Explainability of Detectors

In our recent experiments, human annotators categorized texts into three groups and provided explanation notes for their classifications. The types of explanation notes identified align with findings in past research, highlighting key factors like errors, perplexity, repetition, and readability as crucial in distinguishing between MGTs and HGTs. For instance, studies such as those by Mindner et al. (2023) and Muñoz-Ortiz et al. (2023) have documented similar observations regarding language usage differences between MGTs and HGTs.

Human annotators' explanation notes are predominantly qualitative, yet quantitative measures can also be applied, particularly for aspects like spelling and grammatical errors, perplexity, and readability. For instance, tools such as Grammarly can assist in evaluating spelling and grammatical errors, while NLP tools can be used to calculate text perplexity. Readability can be assessed using existing formalas such as the Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level (Kincaid et al., 1975). Our experiments demonstrate a gradual decline in readability and perplexity scores from texts in the "human" category to the "undecided" category, and finally to the "machine" category. More detailed experimental results can be found in Appendix H.

Despite the robustness of human explanations, which are grounded in common sense and supported by the literature, discrepancies still exist. For example, Hans et al. (2024) introduced the "capybara problem", where both prompts and responses with high perplexity can lead to misjudgments about text origin, both by humans and automated detectors, particularly when prompt details are unknown. Addressing the "capybara problem" involves creating prompts that encourage LLMs to produce features typical of HGTs, as detailed by our annotators. Effective strategies for this are outlined in Appendix D. Moreover, advancements in LLMs like the reduction of unnecessary repetition from ChatGPT-3.5 to ChatGPT-4 demonstrate ongoing improvements, as discussed in Appendix E.

Currently, detector explainability is very limited, and there are instances where provided explanations do not accurately reflect the underlying reasoning of decisions. Future research should aim to enhance the credibility and transparency of detectors by incorporating explainability modules or integrating explainable AI (XAI) components into existing and future MGT detectors.

Future studies should also focus on a better understanding of the nuances between HGTs and MGTs, possibly through user studies that assess perception and comprehension. Technologically, efforts could be directed towards improving the user interfaces of MGT detectors to provide more user-friendly explanations, potentially in an interactive, personalized and contextualized manner. For example, models could indicate whether sentences are derived from what training data or newly generated, potentially using a confidence scale to differentiate between entirely new creations and slight modifications of existing data. Such transparency could greatly enhance the explainability of AI-generated content.

7 Conclusion

This paper explores the effectiveness and challenges associated with current text detection systems. We initially set up a binary classification experiment to identify the top-performing detectors and LLMs that excel in resisting such top-performing detectors. The study was then extended to include a ternary classification framework involving datasets from ChatGPT-4, ChatGPT-3.5, and human sources, where human annotators assessed and explained their classification decisions. The results affirm the relevance of our ternary classification approach, particularly as LLMs continue to advance and produce increasingly human-like texts, making traditional binary classification approaches less meaningful. Our analysis indicates that while current detectors are lacking in explainability, the insights provided by human annotators are valuable for guiding future researcher on MGT detection. These outcomes lead us to recommend enhancements for future detection systems and their explanatory components.

Limitations

This study is subject to several limitations. First, the relatively small sample size restricts us to a primarily qualitative analysis. Second, while the human-produced explanations from our study contribute valuable perspectives, they predominantly serve as recommendations and pointers for further research on improving detection systems. Lastly, given the ongoing advancements in LLM technology, new research opportunities and directions are likely to emerge, necessitating continual updates and revisions to our approach.

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 I.

References

Mohamed Hesham Ibrahim Abdalla, Simon Malberg, Daryna Dementieva, Edoardo Mosca, and Georg Groh. 2023. A benchmark dataset to distinguish human-written and machine-generated scientific papers. *Information*, 14(10):522:1–522:33.

Christopher M. J. André, Helene F. L. Eriksen, Emil J. Jakobsen, Luca C. B. Mingolla, and Nicolai B. Thomsen. 2023. Detecting AI authorship: Analyzing descriptive features for AI detection. In *Proceedings of the 7th Workshop on Natural Language for Artificial Intelligence (NL4AI 2023)*.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023. From text to source: Results in detecting large language model-generated content. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7531–7543.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *Proceedings of the 12th International Conference on Learning Representations*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. ConDA: Contrastive domain adaptation for AI-generated text detection. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 598–610.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

Copyleaks Technologies Ltd. 2023. AI detector | ChatGPT detector | AI checker - CopyLeaks. Online.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. arXiv:1907.09190 [cs.CL].

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

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical detection and visualization of generated text. arXiv:1906.04043 [cs.CL].

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:2301.07597 [cs.CL].

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].

Demis Hassabis and the Gemini Team. 2023. Introducing Gemini: Google's most capable AI model yet. Online.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. MGTBench: Benchmarking machine-generated text detection. arXiv:2303.14822 [cs.CR].

Inspera. 2023. AI content detector. Online.

J. Peter Kincaid, Robert P. Jr Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command Millington TN Research Branch.

Kavita Kumari, Alessandro Pegoraro, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2024. DEMASQ: Unmasking the ChatGPT wordsmith.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4768–4777.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human-and AI-generated texts: Investigating features for ChatGPT. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices – Proceedings of 2023 4th International Conference on Artificial Intelligence in Education Technology*, pages 152–170. Springer.

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

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or human? detect and explain. explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv:2301.13852 [cs.CL].

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and LLM-generated text. arXiv:2308.09067 [cs.CL].

OpenAI. 2021. GPT-2 output detector demo. Online.

OpenAI. 2022. Introducing ChatGPT. Online.

OpenAI. 2023. GPT-4. Online.

Originality.AI. 2024. Originality.AI. Online.

Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Sapling AI Team. 2023. Sapling. Online.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. arXiv:1908.09203 [cs.CL].

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. arXiv:2306.05540 [cs.CL].

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].

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A survey on LLM-generated text detection: Necessity, methods, and future directions. arXiv:2310.14724v3 [cs.CL].

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

Zhendong Wu and Hui Xiang. 2023. MFD: Multi-feature detection of LLM-generated text.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A survey on detection of LLMs-generated content. arXiv:2310.15654 [cs.CL].

ZeroGPT.com. 2023. ZeroGPT. Online.

Shuhai Zhang, Yiliao Song, Jiahao Yang, Yuanqing Li, Bo Han, and Mingkui Tan. 2024. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. In *Proceedings of The 12th International Conference on Learning Representations*.

A Detailed Results of Binary Classification Experiments

In the binary classification experiments, the performance of various detectors on datasets consisting of texts generated by ChatGPT-4 and humans is presented in Table 2. Table 5, Table 6 and Table 7 respectively show the specific performance of different detectors on texts generated by ChatGPT-3.5, LLaMA-13B, and Gemini Pro.

B More about Explanatory Power of the Six Metrics of GPTZero

In the explanations provided by GPTZero, six explainability metrics are identified: Readability, Percent SAT, Simplicity, Perplexity, Burstiness, and Average Sentence Length. For all texts evaluated by GPTZero and their corresponding six feature

Models	Accuracy	Machine as Positive			Human as Positive			Macro F1
		Precision	Recall	F1	Precision	Recall	F1	
GPTZero	97.25%	96.77%	97.83%	97.30%	97.75%	96.67%	97.21%	97.25%
Sapling	90.62%	84.82%	98.96%	91.35%	98.75%	82.29%	89.77%	90.56%
Binoculars	83.00%	77.50%	93.00%	84.55%	91.25%	73.00%	81.11%	82.83%
FAST-DETECTGPT	76.00%	89.39%	59.00%	71.08%	69.40%	93.00%	79.49%	75.29%
MMD-MP	73.00%	96.00%	48.00%	64.00%	65.33%	98.00%	78.40%	71.20%
DEMASQ	66.00%	59.76%	98.00%	74.24%	94.44%	34.00%	50.00%	62.12%
DetectGPT	56.00%	87.50%	14.00%	24.14%	53.26%	98.00%	69.01%	46.58%

Table 5: Binary classification performance of different detectors on the dataset of ChatGPT-3.5

Models	Accuracy	Machine as Positive			Human as Positive			Macro F1
		Precision	Recall	F1	Precision	Recall	F1	
GPTZero	92.43%	96.55%	88.42%	92.31%	88.78%	96.67%	92.55%	92.43%
Sapling	90.26%	85.22%	98.00%	91.16%	97.50%	82.11%	89.14%	90.15%
Binoculars	85.94%	77.31%	100.00%	87.20%	100.00%	73.00%	84.39%	85.80%
FAST-DETECTGPT	87.00%	92.05%	81.00%	86.17%	83.04%	93.00%	87.74%	86.95%
MMD-MP	82.00%	95.71%	67.00%	78.82%	74.62%	97.00%	84.35%	81.59%
DEMASQ	64.00%	58.75%	94.00%	72.31%	85.00%	34.00%	48.57%	60.44%
DetectGPT	65.00%	91.67%	33.00%	48.53%	59.15%	97.00%	73.48%	61.01%

Table 6: Binary classification performance of different detectors on the dataset of LLaMA-13B

values, we created a new dataset to analyze the explainability provided by GPTZero. The ground truth is based on GPTZero's evaluation results. We partitioned the dataset into training and test sets with an 8:2 ratio. We trained four classifiers: Logistic Regression, SVC, Perceptron, and Decision Tree. The weights and accuracy of the different features obtained from these classifiers are presented in Table 8. From the weights, it is evident that the two most effective metrics in GPTZero's explainability are perplexity and readability scores. The remaining metrics contribute minimally to the final results. Additionally, the trained classifier exhibits relatively low accuracy, suggesting that GPTZero employs more complex calculations or utilizes additional sophisticated features that are not disclosed.

C Examples of Different Types of Explanations Given by Human Annotators

Regarding the different types of explanations provided by human annotators, here are some typical examples and analyses. These examples and analyses confirm the validity and rationality of these explanations.

Spelling errors. In the text in Table 9, the word "pisss" contains a spelling error and should be spelled as "piss." This mistake appears to be a typographical error, likely caused by an inadvertent extra keystroke by a human, thus resembling HGT.

Grammatical errors. The text in Table 10 contains several types of grammatical errors. First, there are capitalization mistakes, as "i" should be capitalized in various instances ("I agree with you", "I mean unless", "I think" and "Maybe I'm"). There are also spelling errors: "Eienstien" should be corrected to "Einstein" and "yhat's" should be "that's." Punctuation and spacing need attention too, particularly missing spaces after commas and incorrect use of quotation marks around "homo smarticus." Lastly, the phrase "much more easier" incorrectly uses a double comparative and should be simplified to "much easier."

Perplexity. The text in Table 11 highlights the complexity of our minds, which are not monolithic but composed of multiple processes. While one part may be daydreaming about a giant cheeseburger, another part worries about being fat. This constant shift in awareness and attention illustrates high perplexity.

Logical errors. The text in Table 12 contains logical errors primarily in its misrepresentation of the relationship between processing costs and pricing: it suggests that white rice is cheaper because it undergoes more processing, but in reality, more processing generally increases production costs.

Models	Accuracy	Machine as Positive			Human as Positive			Macro F1
		Precision	Recall	F1	Precision	Recall	F1	
GPTZero	96.77%	96.88%	96.88%	96.88%	96.67%	96.67%	96.67%	96.77%
Sapling	87.30%	85.19%	92.00%	88.46%	90.12%	82.02%	85.88%	87.17%
Binoculars	83.85%	78.99%	94.00%	85.84%	91.78%	72.83%	81.21%	83.53%
FAST-DETECTGPT	92.50%	92.93%	92.00%	92.46%	92.08%	93.00%	92.54%	92.50%
MMD-MP	87.50%	96.30%	78.00%	86.19%	81.51%	97.00%	88.58%	87.39%
DEMASQ	63.50%	58.49%	93.00%	71.81%	82.93%	34.00%	48.23%	60.02%
DetectGPT	66.50%	92.31%	36.00%	51.80%	60.25%	97.00%	74.33%	63.06%

Table 7: Binary classification performance of different detectors on the dataset of Gemini Pro

Classifier	Feature Importances						Accuracy (%)
	Readability	PSAT	Simplicity	Perplexity	Burstiness	ASL	
LR	3.094	-0.857	1.821	-2.517	0.036	0.713	75.76
SVC	2.637	-0.671	2.677	-2.189	0.051	0.654	77.27
Perceptron	4.109	-0.991	8.148	-4.437	0.417	1.039	78.79
Decision Tree	0.289	0.016	0.199	0.205	0.183	0.109	75.76

Table 8: Weights and accuracy of different classifiers using GPTZero's six explainability metrics as features. LR stands for Logistic Regression. PSAT stands for Percent SAT. ASL stands for Average Sentence Length.

Additionally, the comparison of the costs between white rice and brown rice in the analogy contradicts market trends, as brown rice is usually more expensive due to its higher nutritional content, not cheaper.

Unnecessary repetition. The text in Table 13 repeatedly uses concepts like "persistence", "effort", and "success" excessively. It employs synonyms and phrases such as "keep persevering", "continuous effort", and "repeatedly put in effort", which redundantly stress the idea that ongoing effort leads to success. This overuse of the same concept could weaken the impact of the message by not providing new information or perspectives.

Readability. This text in Table 14 exhibits high readability due to its conversational tone, use of simple language, and relatable content. It effectively communicates the speaker's opinion and personal experience with a movie featuring Keanu Reeves. The informal language, straightforward sentence structure, and personal touch make the message easy to understand and engaging. The text also includes an emotional appeal and humor, which further enhance its readability.

Text structure. This text in Table 15 is a well-structured overview of deepfake technology, clearly divided into sections that define the technology and outline its potential harms. Each harm is categorized under a descriptive subheading, making the content easy to follow. This response was generated by an AI.

Bias. The text in Table 16 discusses confirmation bias, where individuals tend to notice and remember information that supports their beliefs and overlook contradicting information. The last sentence, suggesting that someone might just want to engage romantically, uses a casual and potentially flippant tone, which might be seen as inappropriate in more formal or sensitive contexts. This doesn't necessarily show bias against a specific group but indicates a casual style of expression.

D Counterexamples to Explanations Provided by Human Annotators

Due to the overlapping feature distributions of texts generated by humans and AI, and the increasingly blurred boundaries resulting from the rapid development of LLMs, the explanations given by human annotators are based on their understanding of past MGTs and HGTs, which may lead to situations where they cannot accurately explain the origins of text generation. This appendix provides examples and analyses that highlight the limitations of human-produced explanations and presents corresponding counterexamples.

Spelling errors. In the text in Table 17, "coking" and "reeding" are spelling errors; the correct spelling should be "cooking" and "reading". Although the sentence contains spelling errors, it is machine-generated.

Text: Listen to be honest with you it's not always ok to act happy. You need to get things off your chest once in a while. Mankind was not made to be able to sustain happiness. So, I suggest finding a way to relieve your anger, stress, slowly and not bottle it up. Tell someone when they pisss you off, don't just let it pass and smile.

Table 9: An example of human annotators utilizing spelling errors in the interpretation of textual sources. This text is actually generated by human, and human annotators' label is "human".

Text: i agree with you.since the modern man would be aware of the root of various sciences, it would be much more easier & quicker for them to find things like light & all that.but i'm not sure about the homo smarticus"" .i mean unless some more people like Eienstien & Newton are born ,i think we'll be in the same level as of right now.but hey,yhat's just my opinion!maybe i'm so totally wrong.Cheers!""

Table 10: An example of human annotators utilizing grammatical errors in the interpretation of textual sources. This text is actually generated by human, and human annotators' label is "human".

Grammatical errors. In the text in Table 18, there are several common English grammatical errors including verb conjugational mistakes (e.g., "lives" instead of "live" and "goes" instead of "go"), subject-verb agreement issues ("we likes" should be "we like"), and some awkward phrasing ("I very love my family" should be "I love my family very much"). Although the text contains grammatical errors, it is machine-generated.

Perplexity. The text in Table 19 is highly perplexing due to its use of complex vocabulary like "quantum foam" and "entanglement", intricate sentence structures, and abstract concepts that blend physics and metaphysics. This complexity demands a higher level of understanding and familiarity with advanced scientific theories, thereby increasing the text's perplexity. Although high-perplexity texts are more associated with HGTs, this text is machine-generated.

Logical errors. In the text in Table 20, it is mentioned that alcohol weakens your inhibitory control, such as restraining impulses to spend excessive money or speaking loudly, which are typically restrained. Yet, it also suggests "be more friendly", contradicting both the context and common knowledge, thus indicating a deep-seated logical fallacy. Nevertheless, this text was indeed generated by a human writer.

Unnecessary repetition. The text in Table 21 contains repetition in the final sentence, "The result is often a flood, where water levels rise quickly and cause significant damage." This restates the concept already explained earlier, that floods happen when water accumulates faster than it can disperse, leading to significant consequences.

Readability. The readability of the text in Table 22 is high because it uses short, simple sentences that are easy to understand. The words chosen have few syllables, making them straightforward for readers of all levels, including beginners or those learning English as a second language. However, it's important to note that this text was generated by a machine, designed to specifically use simple language and clear constructions to ensure it is accessible and easy to follow.

Text structure. The text in Table 23 lacks coherence and clear transitions. It jumps between examples (music app and self-driving cars) without smooth connections and mixes explanations with examples, making it harder to follow. The mention of data biases and fairness feels abrupt and disconnected from the previous points.

Bias. The text in Table 24 offers a satirical portrayal of a region's sanitary conditions. Replacing this generic reference with the name of any specific region would amount to discrimination against that region. Nonetheless, it is important to note that this biased commentary was generated by a machine.

E Addressing Issues in ChatGPT-3.5: Improvements in ChatGPT-4

ChatGPT-3.5 still exhibits some inherent machine limitations, such as potential logical errors and unnecessary repetitions. However, these issues have been effectively addressed in ChatGPT-4 to some extent. When given with the same prompt, the problems evident in the text generated by ChatGPT-3.5 have been perfectly resolved in ChatGPT-4.

Logical errors. The text in Table 12 was generated by ChatGPT-3.5, and the aforementioned analysis has identified the presence of several logical

Text: Because our minds are n't some solid monolithic thing but a collection of processes . While you 're aware of one , occupied with daydreaming about a giant cheeseburger , another part is churning away on another subject like worrying about being fat . Awareness and attention change constantly .

Table 11: An example of human annotators utilizing perplexity in the interpretation of textual sources. This text is actually generated by human, and human annotators' label is "human".

Text: "Alright, imagine rice grains are like little houses, and they have a protective outer layer called the ""husk."" Brown rice is like a house that still has its outer walls (the husk), while white rice is like a house where the outer walls have been taken off. Now, to make white rice, you have to remove that outer layer (husk) and a few more layers inside to get to the soft part we like to eat. It's like taking off the protective walls of the house to get to the cozy inside. The reason white rice is often cheaper is because it goes through a bit more processing to remove those outer layers, but this process also makes it last longer and gives it a softer texture. Brown rice is less processed, so it keeps more of its natural goodness, but it's also a bit tougher and has a shorter shelf life. Think of it like this: when you buy a ready-made toy, it might be more expensive because someone worked on it more. But if you buy the pieces to build your own toy, it could be cheaper even though you're doing a bit more work yourself. The same goes for rice – white rice is like the ready-made toy, and brown rice is like the pieces you put together at home."

Table 12: An example of human annotators utilizing logical errors in the interpretation of textual sources. This text is actually generated by machine, and human annotators' label is "machine".

errors. For the same prompt, the text in Table 25 obtained from ChatGPT-4 show a complete absence of logical errors.

Unnecessary repetition. The text in Table 13 was generated by ChatGPT-3.5. The analysis has identified some unnecessary repetitions within it. For the same prompt, the text in Table 26 generated by ChatGPT-4 shows no unnecessary repetition, exhibiting clear structure and precise expression.

F Examples of Changed Judgments in GPTZero Evaluations

Regarding the test results of GPTZero versions from December 1, 2023, and May 1, 2024, the judgment outcomes for the two texts have changed. Both texts were machine-generated but were labeled as "undecided" by our human coders. Initially, GPTZero classified these texts as "human", but in the updated version, the classification has changed to "AI".

The feature values of the two texts in Tables 27 and 28 remained completely consistent across both tests. However, the evaluation results were entirely opposite. This indicates that GPTZero operates with a more complex mechanism, and the explanations provided may not be highly interpretable.

G Examples of "Undecided" and Human Annotators' Explanations

Due to space limitations in the main text, we provide additional examples here to illustrate texts categorized as "undecided" along with the explanation notes given by human annotators. Additional examples are provided in Table 29 and Table 30.

H Quantitative Representation of Explanations from Human Annotators

We quantified the perplexity and readability of explanations provided by human coders. Text perplexity was computed using scripts from the Natural Language Toolkit (NLTK) with the GPT-2 model, while readability was measured using the Flesch Reading Ease and Flesch-Kincaid Grade Level formulas.

Table 31 presents the average values of these measures for texts in each category. It is evident that texts classified as "AI" by human coders exhibit lower perplexity and lower readability scores. Specifically, a higher Flesch-Kincaid Grade Level value indicates a higher required English proficiency level, which corresponds to a lower readability score. Texts classified as "undecided" fall between the "human" and "AI" categories. Thus, our annotations by human coders are validated.

I More Details about Human Annotators' Work

The three human annotators we used are all coauthors of the work. We did not recruit any other human participants for the annotation task because the construction of the ternary dataset required two rounds of iterations, including a second round of discussions among all authors to help the three annotators to reach a consensus. We considered such an approach more appropriate for our study than

Text: Success, many say, is the result of perseverance. Indeed, perseverance, or the act of persisting in an effort, is essential to success. One must keep persevering, keep persisting, and keep pushing forward. This persistence, this continuous effort, is what truly defines the pathway to success. Repeated actions, consistent efforts, and ongoing endeavors are crucial. To succeed, one must continuously and repeatedly put in effort, demonstrating persistence again and again. This idea of repeating the same persistent effort is not just a minor detail—it's central, it's critical, and it's absolutely fundamental to achieving success. Through persistent effort, persistent dedication, and persistent work, success becomes achievable. Thus, the secret to success lies heavily in the repetitive, persistent pursuit of one's goals.

Table 13: An example of human annotators utilizing unnecessary repetition in the interpretation of textual sources. This text is actually generated by machine, and human annotators' label is "machine".

Text: "Umm YES..it has Keanu Reeves in it..he's an awesome actor. And its a very romantic and sad movie. If you like tear jerkers you'll love it..I only watched it cus I am a fan of Keanu Reeves and It was better then I thought...I hope you get to watch it soon. Hey I should get the points for being the only guy with the balls to say I like this movie."

Table 14: An example of human annotators utilizing readability score in the interpretation of textual sources. This text is actually generated by human, and human annotators' label is "human".

using recruited human participants. Since some co-authors conducted the annotation work involving human participants, the work did not need to go through our relevant institution's research ethics review process. We did not financially compensate the time of any of the annotators since their work is part of their technical contribution to the work.

To ensure that we did not introduce any bias into the annotation results, the three annotators did their work independently first and then met to resolve disagreements. For the independent annotation step, we provided the annotators with the unlabeled dataset along with the detailed instructions outlined in Table 32. For the later discussion phase, the first author of this paper was the facilitator, and he also discussed with other co-authors on some borderline cases to help inform his facilitation work with the three annotators.

Text: "Deepfake Technology Overview: Definition: Deepfake technology uses artificial intelligence to create highly realistic fake videos or audio recordings, often replacing the likeness or voice of a person with someone else's. Harms of Deepfake Technology: Misinformation:** Deepfakes can be used to spread false information, leading to confusion and public distrust. Political Manipulation:** Deepfakes can be employed to create fabricated content that manipulates political narratives and influences elections. Privacy Concerns:** Individuals can become victims of malicious deepfakes, compromising personal and professional reputations. Security Risks:** Deepfakes pose a threat to national security by potentially creating fake videos of public figures or officials, causing panic or diplomatic tensions. Erosion of Trust:** The widespread use of deepfakes can erode trust in media and make it challenging to distinguish between real and fake content."

Table 15: An example of human annotators utilizing text structure in the interpretation of textual sources. This text is actually generated by machine, and human annotators' label is "machine".

Text: Confirmation bias. Since it's now on your mind, every time it happens to you it starts building up as an idea in your head and you think it's always happening. Your memory starts to ignore all the times it doesn't happen and only focuses on the times it does. Think about how many people you look at every day, and how many times they actually happen to be looking at you. That, or they just want to bone you.

Table 16: An example of human annotators utilizing bias in the interpretation of textual sources. This text is actually generated by human, and human annotators' label is "human".

Prompt: Please generate a text in which an English beginner describes his hobbies. It should exhibit some spelling errors throughout the text.
AI-generated text: Hello! My name is Tom. I really enjoy coking and reeding. I like to make food from my country because it's tasty. My best dish is fried rice. Also, I love to read books, especially about history and science. I read every night to improve my English. Sorry for the mistakes in my writing. Thank you!

Table 17: An example text generated by ChatGPT-4 containing spelling errors

Prompt: Please generate a text in which an English beginner describes his family. It should exhibit some grammatical errors throughout the text.
AI-generated text: Hello! My name is Li Wei. I lives in a small family. My father he is a teacher and my mother works in a hospital. I goes to school every day. I liking to read books and playing games after school. My sister, she young and goes to kindergarten. In the weekend, we likes to go to the park. Sometimes, my father take us to eat out. I very love my family and I happy to share about them.

Table 18: An examples text generated by ChatGPT-4 containing grammatical errors

Prompt: Please generate a text segment characterized by a relatively high level of perplexity.
AI-generated text: In the penumbra of the quantum foam, time and space convolute into a symphony of probabilities, where each note resonates with the ephemeral dance of quarks and leptons. Here, amidst the cacophony of cosmic strings vibrating through the multiverse, the observer stands, a solitary witness to the entanglement of dimensions that both exist and do not exist simultaneously. This liminal space, fraught with the enigmatic whispers of the universe's nascent murmurs, defies the pedestrian laws of physics, inviting a chorus of theoretical conundrums that ripple through the substrates of reality, challenging the very fabric of existential comprehension.

Table 19: An examples text generated by ChatGPT-4 demonstrating a high level of perplexity

Human-generated text: This is definitely not universal. But alcohol hampers your inhibition. Your urge to moderate yourself is lessened, so you are more likely to spend more money, talk louder, be more friendly, etc. This happens because many of your personal traits are really a tug of war between two or more traits. You want to be social, but you also are shy, but you are also lonely, but you are also etc etc etc. This tug of war is toppled over by the restraining forces when you drink.

Table 20: An examples text generated by a human writer containing logical errors

Human-generated text: Floods typically happen when there's more water than what the drainage (natural or man-made) can handle. Man-made being dykes or slopes to direct water, or sewer systems, natural being mostly the permeability of the rock (how fast the water can squeeze between the rocks) or natural slopes. Water does naturally want to be level, but also wants to move to low points, so if it rains a lot, the water all collects at a low point and disperses like you think, but disperses slower than it accumulates, which is why it happens only during very heavy rainfall or snowmelt. The other case is a reservoir, like a dam breaking and releasing a large amount of water. The result is often a flood, where water levels rise quickly and cause significant damage.

Table 21: An example text generated by a human writer containing unnecessary repetitions

Prompt: Can you generate a text with short words and low amount of syllables about food?
AI-generated text: Food is fun. We eat to live. We can munch on nuts, sip soup, or bite a big, red apple. We can mix eggs, milk, and flour to make a cake. A good meal can fill us up and make us feel glad. Some food is hot, some is cold. We eat with a fork, a spoon, or our hands. Every day, we pick what we want to eat. Food is good for all.

Table 22: An example text generated by ChatGPT-4 with a high readability score

Prompt: Please generate a text segment on the topic of machine learning. The text structure should be informal, with minimal use of transitional phrases.
AI-generated text: Machine learning is all about teaching computers to learn from data. Imagine a music app that gets better at recommending songs the more you use it. That's machine learning in action. It looks at what you've listened to and suggests new tunes. These systems keep getting smarter with more data. Self-driving cars, for example, start with basic road rules and improve with every drive. But there's a catch. If the data has biases, the computer might make unfair decisions. That's why developers need to keep an eye on the data and tweak things to stay fair.

Table 23: An example text generated by ChatGPT-4 containing a poor text structure.

Prompt: Please generate a passage using exaggerated language to describe a region's poor sanitary conditions, to the extent that even wildlife cannot survive.
AI-generated text: In this region, the sanitary conditions are so deplorable that even wildlife struggles to survive. The rampant pollution and accumulation of waste have created an environment that is inhospitable to both humans and animals. The air is thick with noxious fumes, and the waterways are choked with contaminants, leaving no safe refuge for any living creature. The degradation of the environment has reached such an extreme that even the hardiest of species find it impossible to endure the harsh conditions.

Table 24: An example text generated by ChatGPT-4 displaying bias

Prompt: Why is white rice cheaper than brown rice ? I don't understand why white rice is cheaper than brown rice . Isn't there more work involved in making white rice ? AND more demand for it ? Please explain like I'm five.
AI-generated text: Sure! Imagine you have two types of toys, a plain toy and a painted toy. The plain toy is like brown rice, and the painted toy is like white rice. When rice is harvested, it starts as brown rice, which has a layer of bran and germ. To make white rice, you need to remove this outer layer, which takes a bit more work, like painting a toy. So yes, it involves more processing. But here's the trick: lots of people prefer the painted toy (white rice) because they think it looks nicer and tastes better. Because so many people want the painted toy, toy makers (rice producers) make a lot of them, and when you make a lot of something, you can often do it more cheaply. Also, the bran and germ from brown rice can be sold for other uses, adding extra value. So, even though it takes more work to make white rice, the high demand and the ability to sell the parts that are removed help keep the price lower than you might expect.

Table 25: An example text generated by ChatGPT-4 that addresses logical errors present in ChatGPT-3.5.

Prompt: Please generate a short discussion on the relationship between success and effort.
AI-generated text: Effort and success are closely related, with sustained effort often leading to incremental improvements, skill development, and resilience, which collectively contribute to achieving success. However, success is also influenced by opportunity, talent, and luck. Access to resources and opportunities, natural ability, and chance events all play significant roles. The synergy between effort and these factors creates a powerful combination, where being prepared through consistent effort enables individuals to seize opportunities and amplify their inherent talents, leading to greater success.

Table 26: An example text generated by ChatGPT-4 that addresses unnecessary repetitions present in ChatGPT-3.5.

Source: ChatGPT-4 **Text**: Sweating itself does not directly cause colds. Colds are caused by viruses, not by being cold or sweating. However, if you sweat and then get chilled, this might weaken your immune system temporarily, making you more susceptible to catching a cold virus. Additionally, the belief that sweating leads to colds might stem from confusing the symptoms of a cold, which can include sweating, with the cause of the cold.
GPTZero result: AI **GPTZero explanations**: Readability: 61.8(Medium) Percent SAT: 2.5(Medium) Simplicity: 40.0(Medium) Perplexity: 34.4(Medium) Burstiness: 36.0(Medium) Average sentence length: 17.3(Medium)

Table 27: Comparison between abstract scores from GPTZero and human-readable explanations

Source: ChatGPT-4 **Text**: "Imagine sending a toy camera tied to a super long string into a deep, dark well (like a black hole). If you try to pull it back, the string would probably break because the well is so strong it can even pull light inside and not let it escape! So, the camera wouldn't come back. Also, the camera would stop working as it gets closer to the black hole because the black hole's super-strong pull (gravity) would break it. Even if the camera somehow kept working and came back, the pictures or videos it took would be all strange and stretched, not like anything we see around us. This is because black holes bend light and time in weird ways. Right now, this idea is more like a fun science fiction story because we don't have the technology to do it, and black holes are really, really far away from us."
GPTZero result: AI **GPTZero explanations**: Readability: 72.3(High) Percent SAT: 1.7(Medium) Simplicity: 35.2(low) Perplexity: 45.3(Medium) Burstiness: 37.9(Medium) Average sentence length: 22.3(Medium)

Table 28: Comparison between abstract scores from GPTZero and human-readable explanations

Text: Listen, I've been in your shoes before, and the best advice I can give you is to embrace change. Life is unpredictable, and sometimes we get comfortable in our routines, but growth happens when we step out of our comfort zones. Don't be afraid to take on new challenges, explore different opportunities, and learn from every experience, even if it seems daunting at first. Remember, the magic happens outside your comfort zone. So, be open to change, embrace the unknown, and trust in your ability to adapt. You'll be amazed at the personal and professional development that follows.
Human explanations: The text is free from grammatical and spelling errors. It earnestly encourages others to step out of their comfort zones, with a tone that is sincere and language that is clear and fluent. As the expressions pertain to everyday discourse, the use of conjunctions is seamless, and the structure aligns with typical conversational patterns. Consistency in style is maintained throughout, without any unnecessary repetitions. Therefore, the text should be categorized as "undecided."

Table 29: Examples of third-category texts and human annotators' explanations

Text: To conclude, we empirically show that a significant number of later layers of CNNs are robust to the absence of the spatial information, which is commonly assumed to be important for object recognition tasks. Modern CNNs are able to tolerate the loss of spatial information from the last 30% of layers at around 1% accuracy drop; and the test accuracy only decreases by less than 7% when spatial information is removed from the last half of layers on CIFAR100 and Small-ImageNet-32x32. Though depth of the network is essential for good performance, the later layers do not necessarily have to be convolutions.
Human explanations: This passage contains no grammatical or spelling errors. It is a summary related to CNN (Convolutional Neural Networks), presented in a scientific and rigorous manner. The data is thoroughly and comprehensively understood, with no logical errors or unnecessary repetition. This passage could have been generated by either an experienced scholar or a machine. Therefore, it should be classified as "undecided."

Table 30: Examples of third-category texts and human coders' explanations

Category	Perplexity	Flesch Reading Ease	Flesch-Kincaid Grade Level
Human	52.72	69.42	7.95
Undecided	34.21	57.44	9.28
AI	21.62	48.02	10.72

Table 31: Average perplexity and readability scores for different classes labeled by human coders. A higher Flesch Reading Ease score indicates greater readability, while a higher Flesch-Kincaid Grade Level score indicates lower readability.

The file in the experiment folder is a spreadsheet where we record the text, detection results, confidence measure, and explanation of detection results.
Our experiment involves detecting the source of the text in the first column and providing the confidence of your judgment along with an explanation.
The file consists of four columns.
The first column is “text”, which contains the text to be detected. The text sources will be divided into the following three categories: “human”, “machine” and “undecided.”
The second column is “detection results”, where you need to fill in your judgment regarding the source of the text in the first column.The text you think to be generated by a human should be labeled as "human." The text you think to be generated by machine should be labeled as "machine." The text for which you cannot decide should be labeled as "undecided."
The third column is “confidence measures”, where you needs to indicate your confidence level in your judgment regarding the source of the text in the first column. The confidence levels are categorized into five grades: “very low”, “low”, “moderate”, “high” and “very high.”
The fourth column is “explanation”, where you should provide your reasoning for the annotation in “detection results.”
Below is an example for illustration:
Text: A fan is an electrical appliance used for cooling and air circulation. It operates by rotating blades, which create a breeze to cool down a room or space. Fans come in various types, including ceiling fans, table fans, and pedestal fans, each designed for specific needs. They are energy-efficient and provide a cost-effective way to stay cool, especially during hot weather. Fans also help in ventilating areas by moving stale air and introducing fresh air.
Detection result: undecided
Confidence measure: moderate
Explanation: This text primarily discusses the topic of electric fans. The content is straightforward and free of grammatical or logical errors. Additionally, the text lacks any apparent emotional bias and features a relatively simple logic. The text exhibits characteristics of both human-generated and machine-generated content. Therefore, I categorize it as “undecided.” Due to the aforementioned reasons, my confidence level in this matter is moderate.
Note: The labeled results are for academic research purposes only.

Table 32: Human Annotation Instructions