Anti-ZeroGPT

Jeffrey Liu and **Guadalupe Aliseda-Canton** and **Anshay Saboo**

Abstract

In an era where AI-created content is populating the internet, this study explores different paraphrasing methods to lower the AI detection scores of AI-generated text, as measured by detection tools such as ZeroGPT (Zer, 2023) and DetectGPT (Mitchell et al., 2023). The study encompasses five strategies: the use of an alternative large language model (LLM) for paraphrasing, characterizing the AI agent, fine-tuning a T5 model to paraphrase the text, few-shot prompting with targeted examples to steer GPT towards desired writing styles, and fine-tuning the GPT model to mimic user's personal writing styles. We then analyze a dataset comprising human-written responses and corresponding GPT-generated responses on same questions. Notably, fine-tuning the GPT model to mimic a user's writing style emerges as the most effective method to lower AI detection score. These results highlight inherent challenges in current AI detection tools and offer vital insights for advancing the robustness of AI text detection in various domains.

1 Introduction

Since 2022, Generative Pre-training Transformer (GPT) and large language model (LLM) have been hot topics. As more and more people, especially students, start to use this groundbreaking technology, academic plagiarism resulting from use of artificial intelligence (AI) has become a rising problem. Fortunately, GPT detection tools like ZeroGPT have been developed (Shrivastava, 2023) and have shown to be a great tool for detecting GPT generated text on the fly (Sample use of ZeroGPT is shown at Figure 1). Given a text input, ZeroGPT generates a score that indicates the likelihood that the text was generated by artificial intelligence, with a higher score indicating a higher likelihood. Other than ZeroGPT, there are also other tools like GPTZero that do similar tasks. These detection tools assist educators in determining whether a document was written by a student or AI.

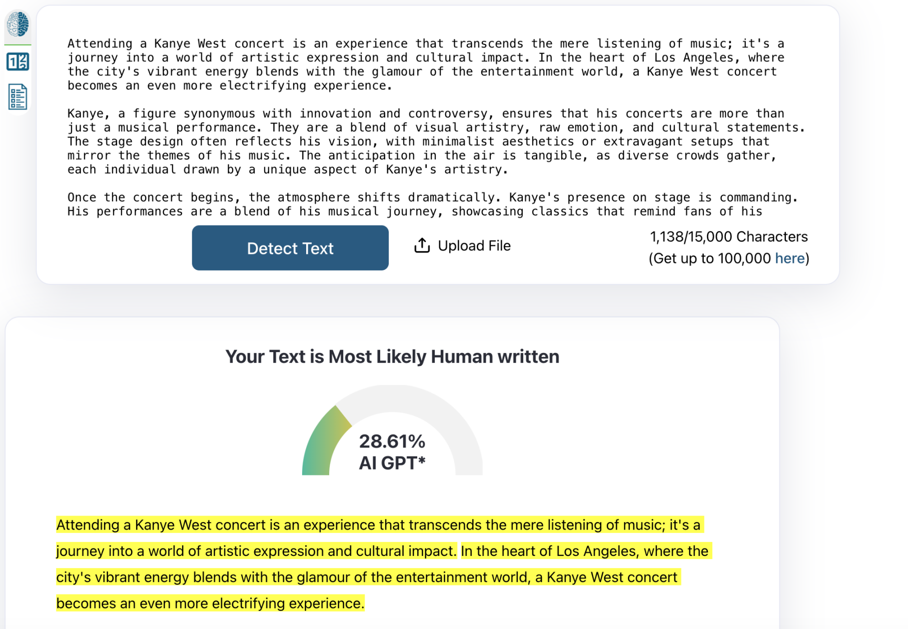
Similarly, in academia, research on AI-generated text detection has advanced with the development of DetectGPT (Mitchell et al., 2023). The new approach introduces a curvature-based criterion to discern if a text is produced by a large language model. DetectGPT operates by analyzing log probabilities computed by the language model in question, alongside random perturbations of the text from another generic pre-trained language model. This method is more discriminative than existing zero-shot methods for model sample detection, offering another robust tool to detect AI plagiarism.

However, are AI detection tools [1] like ZeroGPT or DetectGPT competent enough to detect all AI-generated text? Are there ways to use AI-generated text while avoiding detection by these tools? In this paper, we will look into multiple ways in which GPT generated responses can be paraphrased, and see how these methods can lower AI detection scores. By doing so, we will expose the current issues and drawbacks with existing AI detection tools and provide insights on how they can be improved.

2 Related Work

The field of natural language processing has witnessed transformative advancements (Radford et al., 2019; Brown et al., 2020; Zhang et al., 2022) with the advent of increasingly large language models (LLMs). These large-scale models, including the latest GPT-4 Model developed by OpenAI (OpenAI, 2023), have not only achieved remarkable performance on various language-related benchmarks but also demonstrated an unparalleled ability to generate text that is both convincing and contex-

[1] In our research, for ease of reference, we refer to ZeroGPT and DetectGPT together as the "**AI detection tools**", and refer to the scores obtained from ZeroGPT and DetectGPT as "**AI detection scores**".



Attending a Kanye West concert is an experience that transcends the mere listening of music; it's a journey into a world of artistic expression and cultural impact. In the heart of Los Angeles, where the city's vibrant energy blends with the glamour of the entertainment world, a Kanye West concert becomes an even more electrifying experience.

Kanye, a figure synonymous with innovation and controversy, ensures that his concerts are more than just a musical performance. They are a blend of visual artistry, raw emotion, and cultural statements. The stage design often reflects his vision, with minimalist aesthetics or extravagant setups that mirror the themes of his music. The anticipation in the air is tangible, as diverse crowds gather, each individual drawn by a unique aspect of Kanye's artistry.

Once the concert begins, the atmosphere shifts dramatically. Kanye's presence on stage is commanding. His performances are a blend of his musical journey, showcasing classics that remind fans of his

Your Text is Most Likely Human written

Attending a Kanye West concert is an experience that transcends the mere listening of music; it's a journey into a world of artistic expression and cultural impact. In the heart of Los Angeles, where the city's vibrant energy blends with the glamour of the entertainment world, a Kanye West concert becomes an even more electrifying experience.

Figure 1: Screenshot of ZeroGPT

tually relevant. As these models continue to grow in complexity and sophistication, they redefine the boundaries of AI's potential in mimicking and understanding human language.

Differentiating between text generated by artificial intelligence and human-written content therefore presents a considerable challenge. The study led by Jakesch et al. illuminates the complexities inherent in human processing of AI-generated language (Jakesch et al., 2023). They posit that human evaluators often rely on flawed heuristics, leading to predictable and potentially manipulable judgments regarding AI-produced text. This underscores the necessity for robust detection mechanisms. Prior attempts in this domain, as noted by (Bakhtin et al., 2019) and (Uchendu et al., 2020), primarily centered on creating model-specific classifiers. These classifiers, however, demonstrated a tendency to overfit to their training datasets and the particular models they were designed to scrutinize. This overfitting resulted in a narrowed focus, limiting their effectiveness to the specific type of machine-generated content they were trained on, thereby undermining their broader applicability in distinguishing AI-generated text from human writing.

In response to the challenge of differentiating AI-generated text from human writing, researchers have been moving towards model-agnostic detection methods. This shift is underscored by the development of advanced zero-shot detection techniques, exemplified by the works of (Solaiman et al., 2019) and the subsequent DetectGPT paper (Mitchell et al., 2023).

Solaiman et al.'s methodology leveraged the average log probability under a generative model as a key metric for detection. This approach provided a robust baseline in zero-shot machine-generated text detection, utilizing intrinsic textual characteristics rather than external classifiers (Solaiman et al., 2019). Building upon this foundational work, DetectGPT introduced a more refined strategy. It diverges from solely analyzing raw log probabilities and instead estimates the local curvature of these probabilities around a given text sample (Mitchell et al., 2023). This nuanced approach allows for a deeper insight into the probabilistic framework of language models, enhancing the precision of AI detection mechanisms.

In this paper, we endeavor to identify and refine strategies that effectively counter AI detection tools. Prior research (Sadasivan et al., 2023; Krishna et al., 2023) highlights the potential of paraphrasing as a method to circumvent detection, marking it as a critical area for further exploration. Building on this foundation, our work extends beyond mere paraphrasing. We integrate a range of techniques, including fine-tuning language models and training our own models for enhanced paraphrasing capabilities. Furthermore, we investigate the application of characterizing the AI agent and few-shot learning to modify language model outputs. Based on these multifaceted methods, our research aims to identify the most effective approach in reducing the AI detection scores.

3 Hypothesis

Our project will examine the hypothesis that paraphrasing the text using different methods can lower the AI detection score of AI-generated text in both industry-grade classifier (ZeroGPT) and research-grade classifier (DetectGPT). We specifically choose ZeroGPT since it is one of the most popular AI detection platforms and have similar results with other AI detection tools on sample text input. And it is also one of the few AI detection platforms without complex web scraping defenses, allowing us to easily scrape the AI detection score without paying a large API cost. Also, we specifically choose DetectGPT since it is the most up-to-date (July 2023) research-grade classifier that outperforms existing zero-shot methods for model sample detection.

4 Data

We will use the Hello-SimpleAI/HC3 dataset which includes a set of questions, human answers, and ChatGPT answers. From this dataset we will use a subset of the data called wiki_csai, which contains a collection of human-written Wikipedia articles as the human response and GPT generated answers on the same topic as the ChatGPT generated response. Out of all subdatasets which include: reddit_eli5, medicine, finance, open_qa, and wiki_csai, we have chosen to use wiki_csai due to the fact that other datasets have a large number of questions that ChatGPT cannot answer. Instead of an actual answer to the question, the ChatGPT response recorded in those datasets is something like "I'm sorry, but I cannot provide a response to your request." Detection tools can very easily classify these responses as AI-generated so if we include them in our data the average AI-detection score of the ChatGPT responses will be higher than it would be with normal ChatGPT responses. Additionally, other datasets include questions that are overly specific and thus ChatGPT will also give an answer that can easily be detected.

5 Methods

We will approach verifying this hypothesis by either 1) generating new ChatGPT responses with added parameters (i.e. characterization/few shot) or 2) rephrasing a GPT-generated response to be more "human-like". Then, we will check the AI Scores of generated/paraphrased responses from ZeroGPT/DetectGPT, which indicates how confident these AI detection tools believe that the text was generated by AI.

5.1 Use another large language model to paraphrase original GPT response

In this method, we will use another industry-grade large language model (command-nightly) by Cohere to paraphrase the original GPT response. After running the paraphrasing by command-nightly model, we will check the AI detection scores for paraphrased response from ZeroGPT/DetectGPT.

5.2 Fine-tune GPT-3.5 model

In this method, we will fine-tune the gpt-3.5-turbo model with 100 of researchers' writing samples, including text from academic essay, short Wikipedia-style answers, and long tweets. Then, we check the AI detection scores of the response returned from the fine-tuned personalized model.

5.3 Characterize the AI Agent

In this method, we will change the prompt to add a command to gpt-3.5-turbo model to respond in a certain style. For example, instead of asking "What is the function of the liver?", we change the prompt to be "What is the function of the liver? Explain like you are a college professor". Additionally we also have added prompt to ensure that GPT does not mention anything about the role, but only answer the provided questions. Therefore, using the example above, the final prompt would be: "Your role: Answer like you are a college professor. JUST answer the following question (do not say anything else) and do not explicitly state your role in your response. What is the function of the liver?" After running the new prompt through gpt-3.5-turbo we will run the responses through AI detection tools and check their AI detection scores.

5.4 T5 to paraphrase the text

By using the Hello-SimpleAI/HC3 dataset, we will train and fine-tune a Seq2Seq model to effectively "translate" a ChatGPT response into a human response as shown in Figures 3 and 4. We chose to fine-tune the t5-small language model with the dataset of GPT-generated and corresponding human responses, due to its computational efficiency. Then, we will check the AI detection scores of the response returned from the trained t5-small model.

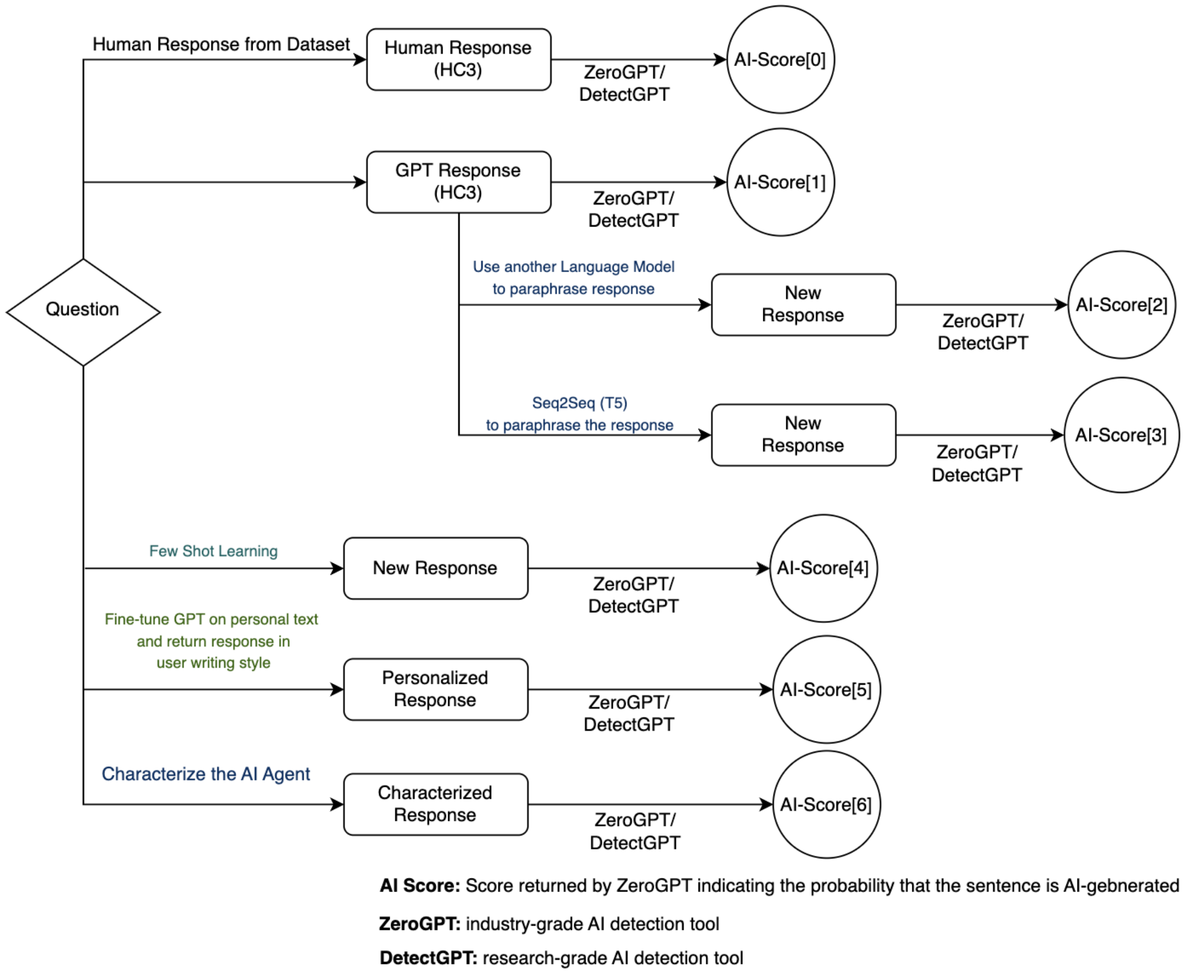

Human Response from Dataset
Human Response (HC3)
ZeroGPT/ DetectGPT
AI-Score[0]
GPT Response (HC3)
ZeroGPT/ DetectGPT
AI-Score[1]
Question
Use another Language Model to paraphrase response
New Response
ZeroGPT/ DetectGPT
AI-Score[2]
Seq2Seq (T5) to paraphrase the response
New Response
ZeroGPT/ DetectGPT
AI-Score[3]
Few Shot Learning
New Response
ZeroGPT/ DetectGPT
AI-Score[4]
Fine-tune GPT on personal text and return response in user writing style
Personalized Response
ZeroGPT/ DetectGPT
AI-Score[5]
Characterize the AI Agent
Characterized Response
ZeroGPT/ DetectGPT
AI-Score[6]
AI Score: Score returned by ZeroGPT indicating the probability that the sentence is AI-gebnerated
ZeroGPT: industry-grade AI detection tool
DetectGPT: research-grade AI detection tool

Figure 2: Diagram of different methods

5.5 Few-shot Learning

In this method, we will change the prompt given to gpt-3.5-turbo to include three samples of questions and human answers as well as instructions to let GPT answer a new question following the writing style of the sample human answers. For example, instead of asking "What is the function of the liver?" we change the prompt to be "You will be given a set of question and response pairs then, using the writing style, grammar choices, and sentence structure from the those question and response pairs, answer the question I give you. Here is the set of questions and responses:" followed by three examples and then followed with "Now answer this question: What is the function of the liver?". After running the new prompt through ChatGPT we will run the response through ZeroGPT/DetectGPT and check their AI detection scores.

5.6 Verify Methods

By passing in the original human and ChatGPT response to ZeroGPT/DetectGPT, we would compute following AI detection scores: **human_score**, and **original_gpt_score**. And then we will use the 5 methods mentioned above to generate new/paraphrased text and pass the text to ZeroGPT/DetectGPT to compute **paraphrased_score**, **fine-tuned_score**, **characterized_score**, **t5_score**, and **fewshot_score**.

paraphrased_score refers to the AI detection score of the text by using another language model to paraphrase original GPT response.

fine-tuned_score refers to the AI detection score of the text returned from the fine-tuned GPT-3.5 model trained on researchers' own writing samples.

characterized_score refers to the AI detection score of the text returned from GPT after assigning it a specific role.

t5_score refers to the AI detection score of the

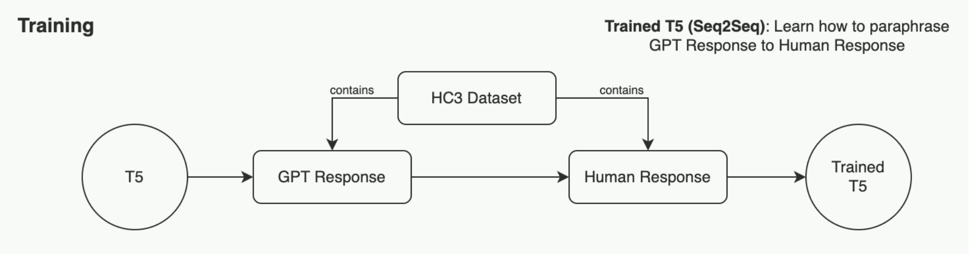

Training
Trained T5 (Seq2Seq): Learn how to paraphrase GPT Response to Human Response
HC3 Dataset
contains
contains
T5
GPT Response
Human Response
Trained T5

Figure 3: Training process of T5 Model

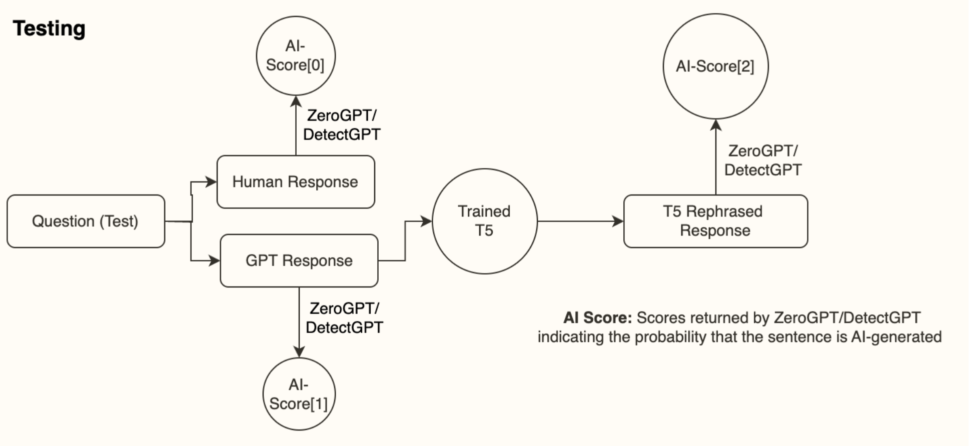


Figure 4: Testing process of T5 Model

paraphrased text returned by the trained t5-small model after it is given the original GPT response as an input.

fewshot_score refers to the AI detection score of the text returned from GPT after giving GPT sample questions and human responses in the prompt.

Then we use the following 2 metrics to evaluate the performance of a method: 1) Examine how much each method reduces the AI detection score compared to the **original_gpt_score**. 2) Examine how closely the AI detection scores from each method align with the **human_score**.

6 Sample Results

6.1 Use another large language model to paraphrase original GPT response

In this method, we use command-nightly large language model (by Cohere) (Coh, 2023), an industry-grade large language model, to rephrase the 500 GPT responses in the dataset. (Examples of paraphrased responses from another large language model are shown in Table 1)

6.2 Fine-tuned GPT-3.5 response with user writing style

In this method, we used one of our main researchers previous writing samples to write question-answering pairs that simulates the question-answer situation of communicating with a chatbot. We in total collected 100 writing samples from the researcher, each roughly 100-200 words, because it is similar to the length of GPT response in the dataset. We then fine-tuned the gpt-3.5-turbo model with the question-answering pairs to learn the user's writing style. And used the fine-tuned model to paraphrase 500 GPT responses in the dataset. (Examples of fine-tuned gpt-3.5-turbo responses are shown in Table 1)

6.3 Characterize the AI Agent

In this method, we assign a role to ChatGPT before asking the question. And use the characterized GPT agent to answer the selected 500 questions in the dataset. (Examples of characterized responses are shown in Table 1)

6.4 T5 Paraphrasing

In this method, we trained a Seq2Seq model (t5-small) based on the Hello-SimpleAI/HC3 dataset. We trained the model to translate ChatGPT-generated answers to "human" answers with the aim of producing responses that retain the same meaning and content, but has a lower AI detection score. And during testing, we sampled 500 GPT-generated responses from the dataset, and use the model to translate into human-like answers to see

Methods	**Prompt**	**GPT Response**	**Modified Prompt**	**Modified Response**
Rephrased by another large language model	Why do some sounds sound so irritating to some people?	Certain sounds, like metal scraping or whiteboard squeaking, can be irritating to some people due to a combination of factors, including their pitch, intensity, and associations with discomfort or stress. Individual sensitivity and personal experiences also play a role in determining which sounds are perceived as annoying.	Same as initial prompt	Some people find certain sounds, such as metal scraping or whiteboard squeaking, to be very unpleasant. This is due to a number of factors, including the pitch and volume of the sound, as well as any negative associations that the person may have with the sound. For example, someone who...
Characterized	Please explain what is "Animal cognition".	Animal cognition refers to the mental capacities of non-human animals. It includes the ability to process information, perceive and attend to stimuli, learn and remember, solve problems, and make decisions...	Your role: Answer like you are a college professor. JUST answer the following question (do not say anything else) and do not explicitly state your role in your response. Please explain what is "Animal cognition"	Animal cognition is a branch of psychology that focuses on studying and understanding the mental processes and abilities of animals. It explores how animals perceive, think, learn, and problem-solve, aiming....
Fine-tuned GPT-3.5	Please explain what is "Amazon Alexa"?	Amazon Alexa is a virtual assistant developed by Amazon, which is designed to be used with smart home devices such as smart speakers and smart displays. It can be activated by saying the "wake word" "Alexa," and can be used to play music, set alarms and timers, provide news and...	Same as initial prompt	Amazon Alexa is a virtual assistant developed by Amazon. Its interaction model allows users to do a series of tasks including playing music, setting an alarm, and weather forecast by speaking on a microphone or typing command on the Amazon Alexa app. Moreover. . .
T5	What is a sloth?	A sloth is a unique and fascinating creature native to the lush rainforests of Central and South America. It is renowned for its exceptionally slow and deliberate movements, making it one of the world's slowest mammals. Characterized by its long, curved claws, sloths are perfectly adapted to their arboreal existence...	Same as initial prompt	The sloth is a unique and fascinating creature native to Central and South America. They are perfectly adapted to their arboreal existence, living the majority of their lives in trees. hey sustain themselves on a diet primarily composed of leaves, twigs, and fruits, which provides limited nutritional value...
Few-shot Learning	Please explain what is "Distributed artificial intelligence"	Distributed artificial intelligence (DAI) refers to the use of multiple artificial intelligence (AI) agents or systems that work together to achieve a common goal or solve a problem...	You will be given a set of question and response pairs then, using the writing style, grammar choices, and sentence structure from the those question and response pairs, answer the question I give you. Here is the set of questions and responses: ... Now answer this question: Please explain what is "Distributed artificial intelligence"	Distributed artificial intelligence (DAI) refers to the field of artificial intelligence that focuses on creating intelligent systems composed of multiple interconnected agents. These agents work collaboratively and interact with each other in order to achieve a common goal or solve complex problems...

Table 1: Comparison of different methods and responses
(GPT/Modified response include first 50 words, full example can be found at the appendix)
Modified Prompt: Question with added instructions/specifications.
Modified Response: Response from corresponding method used.

their new AI detection scores. (Examples of T5 paraphrased responses are shown in Table 1)

6.5 Few-shot Learning

In this method, we prompt GPT with three sets of questions and corresponding humans answers and then ask it to a answer a new question using the writing style from the given sample human answers. We use the added prompt to ask gpt-3.5-turbo to answer the selected 500 questions in the dataset.(Examples of few-shot learning responses are shown in Table 1)

7 Results

To test how effective each method is we used ZeroGPT and DetectGPT. They both return a score from 0 to 100 based on how likely it is that the text is AI-generated, with 0 being least likely to be AI-generated and 100 being the most likely. To set the baseline we ran both tools on 500 of the original human responses and the unmodified GPT-generated responses from our dataset from Hello-SimpleAI/HC3. Then we used the methods described above to generate a modified GPT response and then ran both tools on those responses. The results can be seen in Table ?? and Table ??.

From these results we can see that the methods we used were largely successful in decreasing the AI detection score. For ZeroGPT, the average detection score for the unmodified GPT response was 95.06, in comparison, the highest score for modified responses was 58.58 and the lowest score for modified responses was 20.41. This shows a score decrease between 38.38% and 78.53%. Meanwhile, for DetectGPT, the average detection score for the unmodified original GPT-generated response was 49.04, in comparison the highest score for modified was 40.05 and the lowest score for modified responses was 22.85, this shows a score decrease between 18.33% and 53.41%.

Another important thing to note is the ZeroGPT AI detection score distribution is bimodal (see appendix A.2). A significant portion of the responses, regardless of method, are either given an exceptionally low score (less than 10) or an exceptionally high score (above 90). Notably, the bimodal distribution becomes most evident in t5 paraphrased responses and characterized responses.

The fine-tuning approach applied to gpt-3.5-turbo has proven to be the most effective in reducing the AI detection score, evidenced by achieving an average score of 20.41 with ZeroGPT and 21.85 with DetectGPT. This approach led to a significant decrease in AI detection scores: a 78.53% reduction for ZeroGPT and a 53.41% decrease for DetectGPT, when compared to the AI detection scores obtained from the original GPT response.

Additionally, the fine-tuned gpt-3.5-turbo model had AI detection scores of 20.41 closely mirroring those of human-written responses which had an average score of 16.27, as assessed by ZeroGPT. This indicates a near parity in AI detectability between the fine-tuned AI outputs and human writing. In the case of DetectGPT, the model's responses achieved an even lower AI detection score of 21.85, surpassing the score of 28.24 for human-generated text. This demonstrates the model's enhanced capability to generate more "human-like" content.

	Avg	*Std*	Total
# Human	**16.27**	**27.93**	500
# Unmodified GPT	95.06	15.68	500
# Characterized GPT	34.55	44.04	500
# Multi-LLM Rephrased	58.58	34.94	500
# Fine-tuned GPT-3.5	**20.41**	**31.35**	500
# T5 Paraphrased	53.13	45.91	500
# Few Shot	76.13	31.68	500

Table 2: Summary of the average ZeroGPT detection scores for each type of response

	Avg	*Std*	Total
# Human	**28.24**	**12.64**	500
# Unmodified GPT	49.04	10.41	500
# Characterized GPT	40.05	12.19	500
# Multi-LLM Rephrased	31.19	8.39	500
# Fine-tuned GPT-3.5	**21.85**	**11.34**	500
# T5 Paraphrased	34.00	12.93	500
# Few Shot	37.15	9.26	500

Table 3: Summary of the average DetectGPT detection scores for each type of response

8 Discussion

From these results, it's evident that all the methods employed are effective in reducing the AI detection score. As indicated in Table ?? and Table ??, the most successful approach in lowering the AI detection scores for both ZeroGPT and DetectGPT is personalizing the response to mimic the user's writing style through fine-tuning the gpt-3.5-turbo model. A key factor contributing to this success

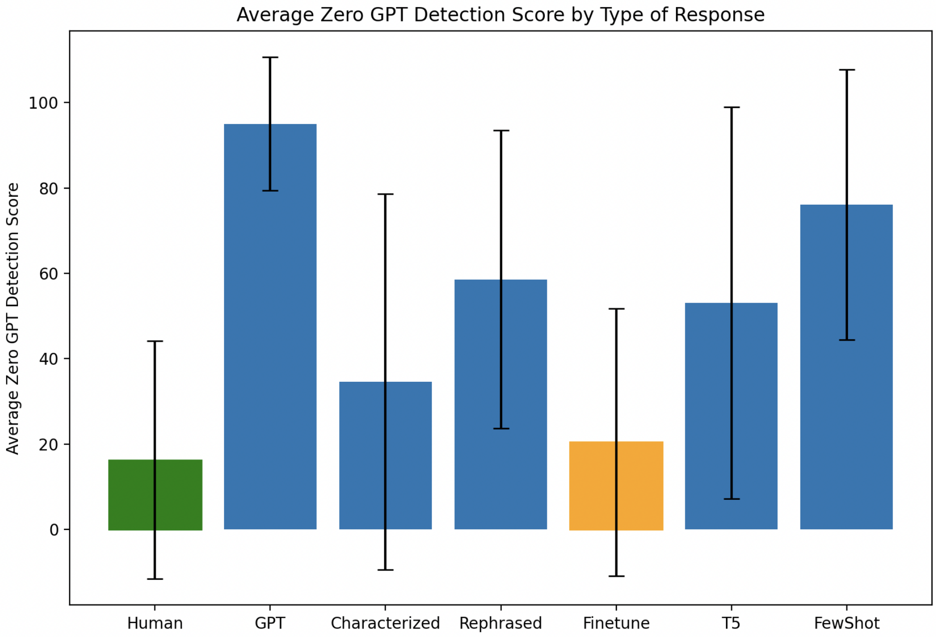


Figure 5: ZeroGPT scores for different methods

is likely the primary objective of rendering the AI output as human-like as possible. By utilizing a substantial volume of actual human-written samples, the GPT model can learn and replicate specific writing styles more accurately. This method leverages the nuanced characteristics of individual writing patterns, enabling the AI to produce responses that not only resonate more closely with human writing but also effectively evade detection by AI classifiers.

Consequently, it becomes apparent that a notable limitation of contemporary AI classifiers is their inability to accurately identify AI-generated text that has undergone minor modifications. This deficiency is particularly evident in industry-standard classifiers such as ZeroGPT, which exhibit a bimodal distribution in their scoring mechanisms. The absence of a more uniform distribution of scores can significantly impact the efficacy of AI detection. Rather than assigning a comprehensive score based on the overall characteristics of the text, these classifiers may default to a bimodal scoring approach, potentially compromising the nuanced evaluation of the text's origin.

This shows the current limitations of AI detection tools. While these tools are adept at identifying original, unmodified AI responses, they lack the sophistication required to detect modified GPT-generated content.

We suggest that further efforts be dedicated to enhance the performance of current AI classifiers. A potential method for improvement is the application of watermarking which involves restricting GPT-generated responses from utilizing specific phrases (Kirchenbauer et al., 2023). Consequently, the use of these restricted phrases would indicate human authorship. This approach could be effective in identifying outputs from methods like our "Characterized GPT" and fine-tuning techniques, as the entirety of these responses are GPT-generated. In instances where blocked phrases are present, detection tools could effectively discern human-written segments.

Nonetheless, the watermarking approach has limitations, particularly when the text is not exclusively GPT-generated. For instance, in methods where we employ a trained T5-small model for paraphrasing, the paraphrased text may inadvertently incorporate the blocked phrases, potentially leading AI detection tools to erroneously classify the text as human-written.

Another method for improvement can be training with a more comprehensive dataset that encompasses outputs from various large language models. Such a dataset would offer a wider range of AI-generated text samples, thereby improving the detection capabilities of these tools.

Ultimately, it is evident that ongoing research and development are imperative for ensuring the accuracy and adaptability of AI detection tools. This is especially crucial in contexts like educational settings, where GPT might be utilized for activities

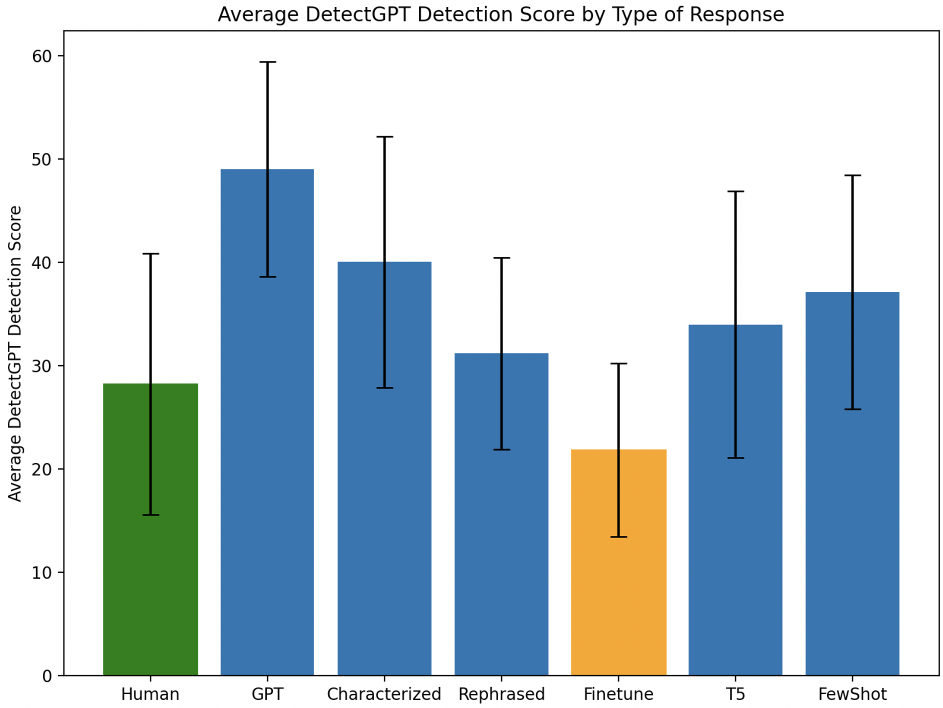


Figure 6: DetectGPT scores for different methods

like plagiarizing assignments.

9 Future Works

Moving beyond the research presented in this paper, we hope to expand upon the method of training our own seq2seq model to "translate" a GPT-generated response into a human response. This involves using external computational resources such as cloud-based compute to train and fine-tune larger models, such as T5-large or GPT-2. Training on these more developed base models will allow us to build more nuance and parameters into the model, and potentially result in a better human response generation than the current trained T5-small model.

Furthermore, we also plan to systematically fine-tune additional models using different individuals' writing samples. This approach aims to ascertain whether the effectiveness of the method is consistent across various personal writing styles.

10 Conclusion

For our research findings, it is evident that current AI detection tools like ZeroGPT and DetectGPT, while proficient in identifying unaltered outputs from advanced large language models, face significant challenges when presented with paraphrased text.

Our studies indicate that paraphrasing methods makes AI-generated text less distinguishable from human-written content, decreasing its detectability by AI detection tools. This effect is further amplified when large language model (gpt-3.5-turbo) is fine-tuned with an individual's unique writing style. The simplicity of implementing such methods raises concerns about their potential misuse for AI-assisted plagiarism. Therefore, we strongly advocate for the enhancement of AI detection tools to recognize and adapt to these more nuanced forms of AI-generated text. The advancement of these tools is vital to uphold the integrity of academic and intellectual standards by effectively mitigating AI plagiarism. And that advancement aligns with the primary objective of our study: to prompt the development of more sophisticated AI text detection tools, ensuring the continued effectiveness of these tools in the face of evolving AI capabilities.

References

2023. Cohere.

2023. ZeroGPT.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie

Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

M. Jakesch, J. T. Hancock, and . . 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11).

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected?

Rashi Shrivastava. 2023. With Seed Funding Secured, Ai Detection Tool GPTZero Launches New Browser Plugin.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jong Wook Wang. 2019. Release strategies and the social impacts of language models.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Ming Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models.

A Appendix

A.1 Full GPT/Modified Responses

Sample question/human response/original response/modified responses pair can be found at our project GitHub Repository.

A.2 ZeroGPT Score distribution for different methods

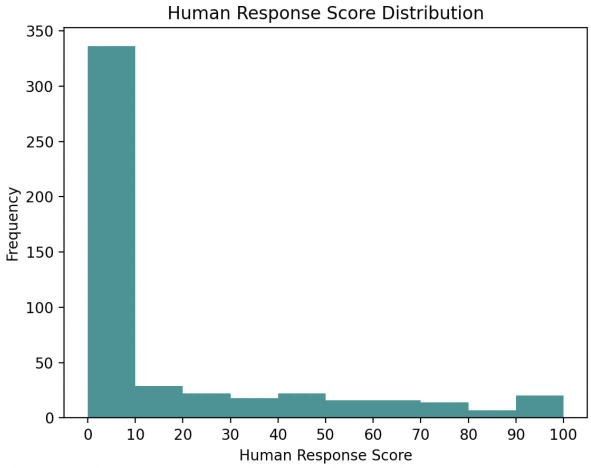


Figure 7: Human Response Score Distribution

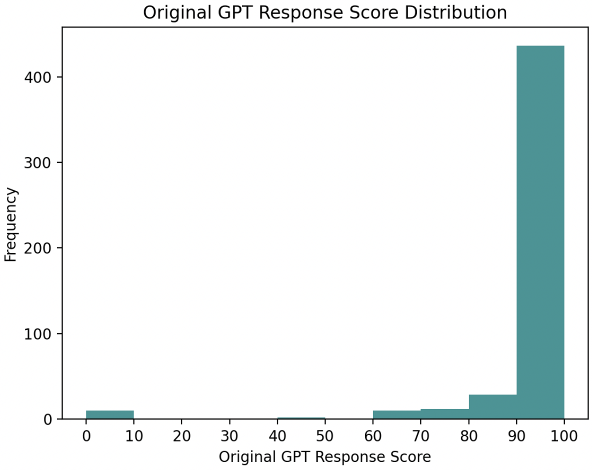


Figure 8: Original GPT Response Score Distribution

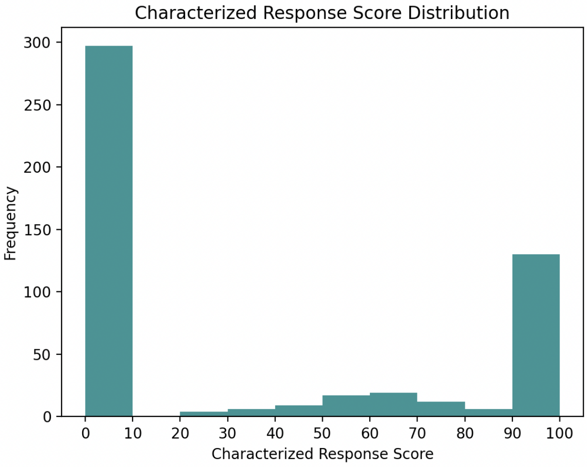


Figure 9: Characterized Response Score Distribution

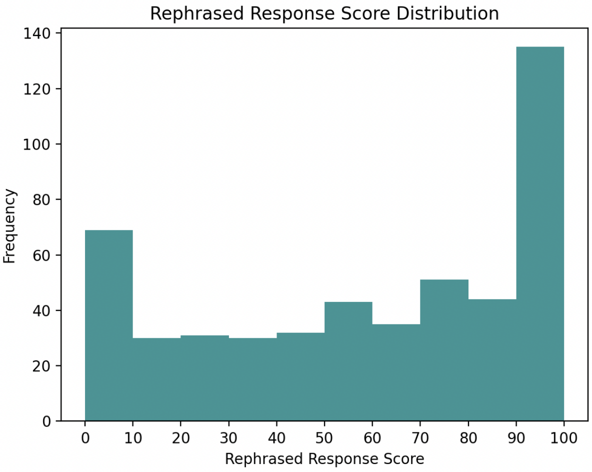


Figure 10: Rephrased Response Score Distribution

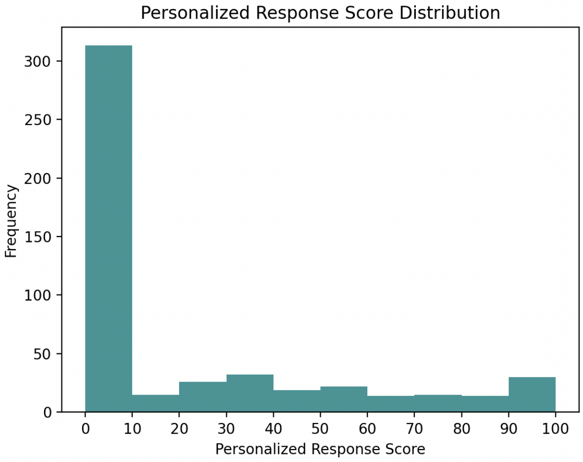


Figure 11: Fine-tuned GPT-3.5 Response Score Distribution

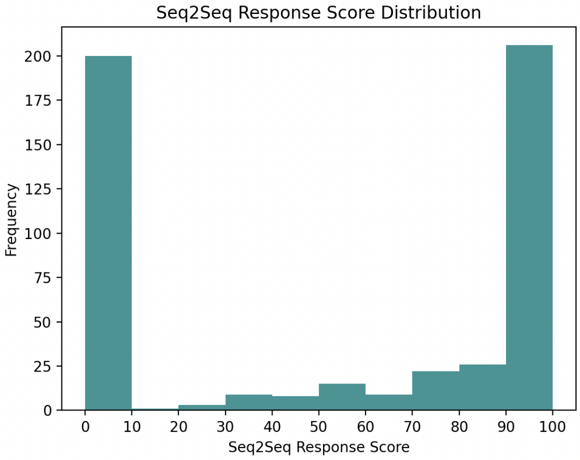


Figure 12: T5 Rephrased Score Distribution

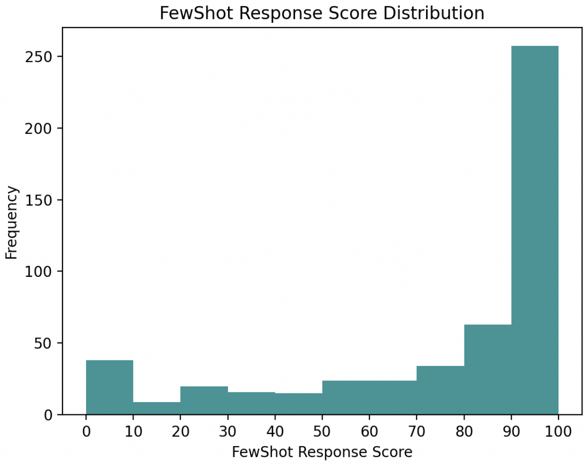


Figure 13: Few Shot Response Score Distribution

A.3 DetectGPT Score distribution for different methods

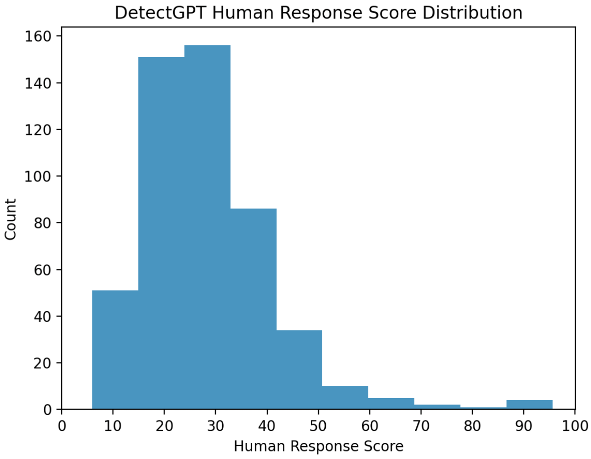


Figure 14: Human Response Score Distribution

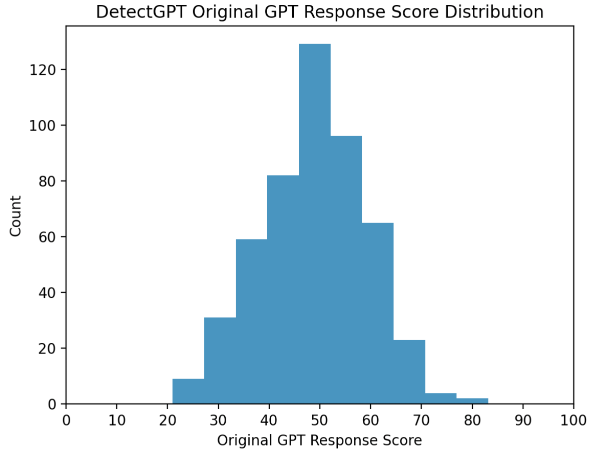


Figure 15: Original GPT Response Score Distribution

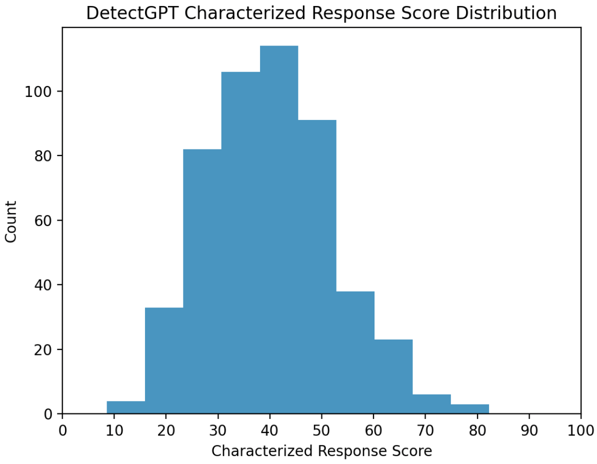


Figure 16: Characterized Response Score Distribution

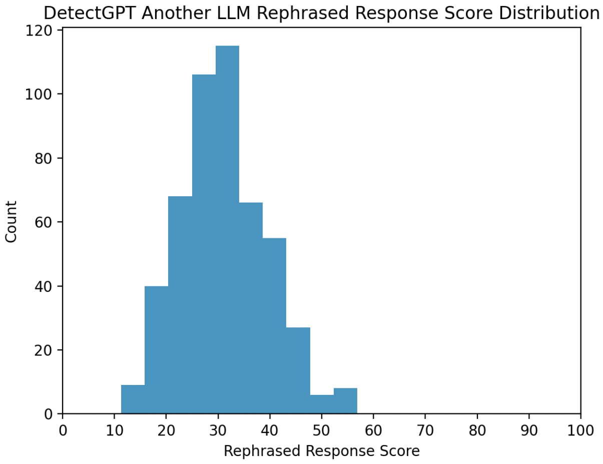


Figure 17: Rephrased Response Score Distribution

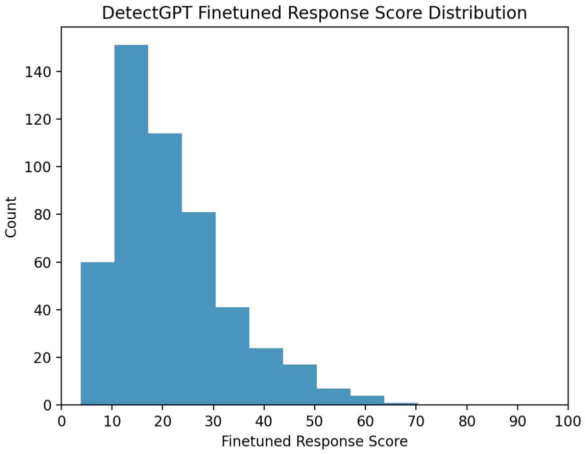


Figure 18: Fine-tuned GPT-3.5 Response Score Distribution

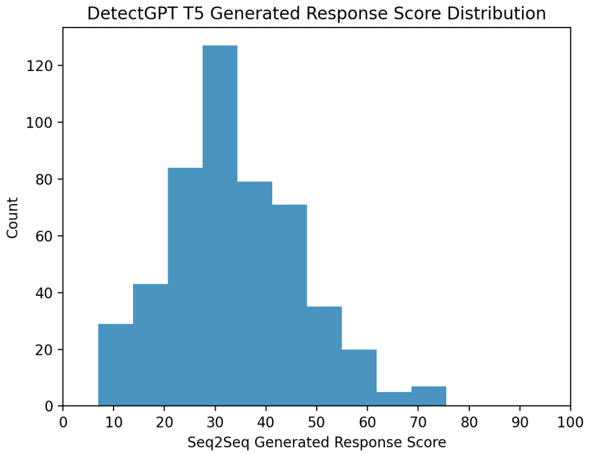


Figure 19: T5 Rephrased Score Distribution

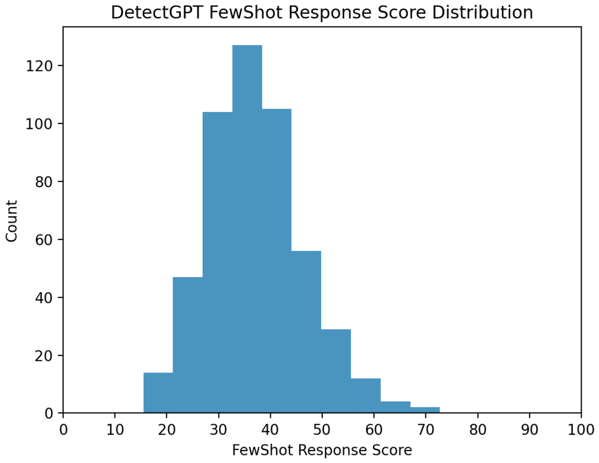


Figure 20: Few Shot Response Score Distribution