Deloitte at #*SMM4H* 2024: Can GPT-4 Detect COVID-19 Tweets Annotated by Itself?

Harika Abburi[1], Nirmala Pudota[1], Balaji Veeramani[2], Edward Bowen[2], Sanmitra Bhattacharya[2]

[1]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited India
[2]Deloitte & Touche LLP, USA
{abharika, npudota, bveeramani, edbowen, sanmbhattacharya}@deloitte.com

Abstract

The advent of Large Language Models (LLMs) such as Generative Pre-trained Transformers (GPT-4) mark a transformative era in Natural Language Generation (NLG). These models demonstrate the ability to generate coherent text that closely resembles human-authored content. They are easily accessible and have become invaluable tools in handling various text-based tasks, such as data annotation, report generation, and question answering. In this paper, we investigate GPT-4's ability to discern between data it has annotated and data annotated by humans, specifically within the context of tweets in the medical domain. Through experimental analysis, we observe GPT-4 outperform other state-of-the-art models. The dataset used in this study was provided by the SMM4H (Social Media Mining for Health Research and Applications) shared task. Our model achieved an accuracy of 0.51, securing a second rank in the shared task.

1 Introduction

The field of Natural Language Generation (NLG) is undergoing a significant transformation driven by the emergence of LLMs like GPT-4 (OpenAI, 2023), and many other Large Language Models. These models are capable of generating text of human-level quality for a wide range of applications, such as data annotation (Tan et al., 2024), medical question answering (Kung et al., 2023), conversation response generation (Mousavi et al., 2023), and code auto-completion (Tang et al., 2023). Notably, the ability of these models to learn without extensive training data (zero-shot learning) or with just a few examples (few-shot learning), simplifies their integration into various language generation applications.

While the LLMs demonstrate the ability to understand the context and generate coherent human-like responses, they do not have a true understanding of what they are producing (Li et al., 2023; Turpin et al., 2023). This could potentially lead to adverse consequences when used in downstream applications. For example, consider an application of a LLM tasked with summarizing a medicinal drug datasheet inadvertently produces wrong dosage information. This generation of plausible but false content (referred as *hallucination* (Bang et al., 2023; Ji et al., 2023)), can unintentionally spread misinformation, false narratives, fake news, and spam. Similarly, the use of LLMs in data annotation has sparked a debate within the research community due to potential issues like inherent biases and hallucinations associated with these models (Yao et al., 2024; Bogdanov et al., 2024). Motivated by these challenges, the automatic detection of AI-generated outputs has emerged as an active area of research.

The detection of data annotations generated by LLMs closely resembles the process of identifying AI-generated text, which aims to distinguish between human-authored and machine-generated content (Abburi et al., 2023b,a). In this paper, we explore methodologies utilized for AI-generated text detection, with a particular emphasis on zero-shot detection techniques, and examine the synergies between these two areas. The AI-generated text detection methods predominantly involve the analysis of outputs from LLMs utilizing features such as entropy, log-probability scores, and perplexity (Wu et al., 2023; Yang et al., 2023a; Bao et al., 2023; Hans et al., 2024) to distinguish between human-written and machine-generated content. Building upon this foundation, DetectGPT (Mitchell et al., 2023) introduced the concept of analyzing negative log probability curvature, identifying a distinct pattern in AI-generated text. Subsequent advancements, such as DNA-GPT (Yang et al., 2023b), improved performance by analyzing the divergence of n-grams between the original text and LLM-prompted versions.

While zero-shot detection methods are effective,

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 79–82

their success often rely on direct access to the internal mechanisms of the specific LLM that generated the text. However, the underlying architecture and weights of many LLMs, including OpenAI's GPT-4, are not publicly available. As a result, these techniques often depend on a substitute LLM, presumed to have mechanisms similar to the proprietary model. The reliance on a proxy LLM may limit the robustness and generalizability of zero-shot detection in various scenarios.

2 GPT-4 as AI-annotated detector

While GPT-4 has exhibited proficiency in various Natural Language Processing (NLP) tasks, its potential for distinguishing between human-annotated and AI-annotated data remains largely untapped. In this study, we explore the effectiveness of GPT-4 in distinguishing between tweets it annotated and those annotated by humans in the medical domain. Our experimental analysis indicates that GPT-4 outperform other state-of-the-art models in detecting AI-annotated data.

2.1 Dataset

We use a dataset provided by the SMM4H shared task. The dataset consists of both human annotated (human) and GPT-4 annotated (generated) tweets detailing COVID-19 symptoms written in Latin American Spanish. In total 3,682 tweets were available for training and 2,110 tweets were available for testing. More details about the dataset can be found in the SMM4H overview paper (Xu et al., 2024).

2.2 Choice of prompt and experimental settings

We experimented with various prompts to identify the most effective one for label prediction. Our findings indicated that complex prompts often caused GPT-4 to generate incorrect labels. Therefore, we chose a simpler prompt, which resulted in better performance. The chosen prompt is as follows:

"Imagine you're a data annotation expert. You're presented with a tweet containing a description of potential COVID-19 symptoms. This tweet has already been annotated as containing COVID-19 symptoms, but you don't know if that is annotated by a human expert or AI model.
Your task is to analyze the tweet to determine whether the initial annotation of 'contains COVID-19 symptoms' was made by a human expert or by an AI model. Can you determine the source of the label (human or generated)? Answer in a single word, predicted label should be one of 'human' or 'generated'.
text : {text}
prediction:{}"

Each tweet is passed to the prompt as *{text}*. Since this task involves classifying whether the given text is annotated by human or AI in a zero-shot setting, we do not fine-tune the model. Instead, we directly prompt GPT-4 to classify the test samples annotated by 'human' or 'generated'. To limit randomness in the model's output, which is crucial for classification accuracy, we set the *temperature* parameter to 0.

Evaluation sets	Number of samples annotated by human	Number of samples annotated by AI
Set1	368	322
Set2	350	340
Set3	357	333

Table 1: Statistics of evaluation sets

Models	Set 1		Set 2		Set 3	
	Acc	F_{mac}	Acc	F_{mac}	Acc	F_{mac}
(Hans et al., 2024)	**0.47**	0.36	0.49	0.38	0.48	0.37
(Bao et al., 2023)	**0.47**	0.33	0.49	0.33	0.48	0.33
Our approach	0.46	**0.40**	**0.49**	**0.42**	**0.49**	**0.41**

Table 2: Performance comparison of zero-shot detection models on 3 evaluation sets. Acc: Accuracy, F_{mac}: F1-macro

3 Results

To assess the performance of our approach in the zero-shot setting, we curated three distinct evaluation sets – Set 1, Set 2, and Set 3 as shown in Table 1. Each set consisted of 690 samples, randomly selected from the training data, to avoid overlap.

We conducted a comparative evaluation of our GPT-4-based approach against multiple zero-shot detection models. The two most effective baselines methods we identified were: 1) Binoculars (Hans et al., 2024), which compares the perplexity and cross-perplexity of two closely related language models to identify AI-generated text; and 2) Fast-DetectGPT (Bao et al., 2023), which utilizes conditional probability curvature to efficiently detect AI

content, particularly from models like GPT.

Table 2 provides a comparative performance analysis of the baselines and our approach across the three distinct evaluation sets. Despite our GPT-4 prompting-based approach consistently achieving the highest Acc and F_{mac} across each of the sets, the top scores did not exceed 0.5. This highlights the significant challenge posed by AI annotation detection, especially for tweets in the medical domain. Nevertheless, we used GPT-4 prompting to generate the predictions on the test set (predictions are uploaded to codalab), resulting in accuracy 0.51, which placed our team in second place in the shared task.

4 Conclusion

In this paper, we investigated the potential of GPT-4 to detect tweets annotated by itself in zero-shot setting. Our experiments highlighted the complexity of this task, particularly for short texts in the medical domain, achieving an accuracy of 0.51. For future work, we aim to improve the model's performance by incorporating the reasoning behind its predictions. By using this reasoning as additional input, we aim to enhance the model's ability to differentiate between data annotations predicted by LLM and those annotated by human experts.

References

Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023a. A simple yet efficient ensemble approach for AI-generated text detection. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 413–421, Singapore. Association for Computational Linguistics.

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023b. Generative ai text classification using ensemble llm approaches. In *IberLEF@ SEPLN*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. *arXiv preprint arXiv:2402.15343*.

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 preprint arXiv:2401.12070*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps?

OpenAI. 2023. Gpt-4 technical report.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey.

Ze Tang, Jidong Ge, Shangqing Liu, Tingwei Zhu, Tongtong Xu, Liguo Huang, and Bin Luo. 2023. Domain adaptive code completion via language models and decoupled domain databases. *arXiv preprint arXiv:2308.09313*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A large language models detection tool.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health (#SMM4H) research and applications workshop and shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023a. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023b. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*.

Yuxuan Yao, Sichun Luo, Haohan Zhao, Guanzhi Deng, and Linqi Song. 2024. Can llm substitute human labeling? a case study of fine-grained chinese address entity recognition dataset for uav delivery. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1099–1102.

A List of prompts explored

1. *"Imagine you're a data annotation expert specializing in identifying the origin of annotations on social media posts. You are presented with a tweet that has been annotated as "contains COVID-19 symptoms. Your task is to analyze the tweet and determine whether this annotation was made by a human expert or generated by an AI model.*
   Please follow these steps to make your determination:
   1.Content Analysis:
   Examine the language and structure of the tweet. Consider the complexity, coherence, and nuance in describing COVID-19 symptoms.
   Evaluate the specificity and accuracy of the symptoms mentioned. Human experts tend to provide precise and medically accurate descriptions, whereas AI models might be more general or formulaic.
   2.Annotation Style:
   Assess the style and quality of the annotation itself. Human annotations often reflect domain expertise and may include subtle contextual understanding.
   Look for patterns typical of AI-generated annotations, such as repetitive phrasing, lack of deep contextual insight, or overly broad categories.
   3.Consistency and Commonality:
   Compare the tweet with common patterns and characteristics known from human annotations versus AI-generated annotations. Humans may show more variability and adaptability in their descriptions.
   Based on your analysis, provide your prediction on whether the annotation was made by a human expert or generated by an AI model. Answer in a single word, predicted label should be one of 'human' or 'generated'.
   text : {text}
   prediction:{}"

2. *"Imagine you're a data annotation expert specializing in identifying the origin of annotations on social media posts. You are presented with a tweet that has been annotated as 'contains COVID-19 symptoms.' Your task is to analyze the tweet and determine whether this annotation was made by a human expert or generated by an AI model. Answer in a single word, predicted label should be one of 'human' or 'generated'.*
   text : {text}
   prediction:{}"

3. "You are a GPT4 data annotation expert. Given a text, predict who annotated the text: human or AI. Predicted label should be one of 'human' or 'generated'.
   text : {text}
   prediction:{}"