JOURNAL OF
EMERGING INVESTIGATORS

Article

Comparing and evaluating ChatGPT's performance giving financial advice with Reddit questions and answers

Sathvik Samant[1], Aditya Dhar[2], Shreya Kochar[3], Aneesha Sreerama[4], Andrew Wang[5], Anirudh Sreerama[6]

[1] The Lawrenceville School, Lawrenceville, New Jersey
[2] Fairview High School, Boulder, Colorado
[3] Columbia University, New York, New York
[4] Northeastern University, Boston, Massachusetts
[5] Stony Brook University, Stony Brook, New York
[6] University of California Berkeley, Berkeley, California

SUMMARY

As artificial intelligence (AI) and particularly Large Language Models (LLMs) rapidly advance, there is a growing interest in AI in the financial industry while understanding its impact on the future of financial advising. To evaluate the performance of such LLMs in the role of a financial advisor, this experiment utilized financial questions asked on the Reddit forum "r/Financial Planning". We hypothesized that Chat-GPT would offer commonly observed, yet concise feedback related to typical financial behaviors rather than delivering personalized financial advice. We compared the GPT-4 outputs to actual Reddit comments, assessing the model's response content, length, and advice. By evaluating the model's advisory competency, this study explored the role of AI in financial forums, its ethical consequences, and potential threat to employment and existing systems. We found that while AI can present accurate information, it failed in its delivery, clarity, and decisiveness. This study further analyzed the implications of GPT-4's performance and its impact on future financial forum systems. More broadly, this study revealed that at its current capabilities GPT-4 does not pose a direct threat to traditional financial forums but has the potential in the future to shift financial forums and advisories to more AI-based systems.

INTRODUCTION

Given the rapid development of artificial intelligence (AI), particularly within the past year, a growing focus has been given to the applications of these technologies in various interdisciplinary fields. Due to this advancement, there has been a growing call for innovation within the financial sector, including financial advisory (1). In particular, growing research seeks to explore and develop these AI technologies to help users develop financial literacy and give financial recommendations (1).

One such content creation tool, 'Chat Generative Pre-Trained Transformer' (ChatGPT), was released globally in late 2022. Using various data inputs, ChatGPT generates explanatory sentences in conversational-style settings, which greatly increases efficiency in meeting user needs (2). ChatGPT is a significant tool in generating financial recommendations, acting in a similar capacity as an advisor. Although this tool does not utilize specific raw financial data, it can respond to specific situational text blocks and provide options and responses (2).

ChatGPT, which is based on Large Language Model (LLM) technology, uses large sets of data and deep learning models to generate content (2). Although prior research has been conducted in the field of LLMs in finance, many focus on financial markets or market trends, instead of personal finance and financial advisory (3).

Prior studies in this field have also explored the usage of ChatGPT in financial advisory contexts. A study by Neilson found that while ChatGPT can generate relevant insights for investors who are looking for simple financial recommendations, it struggles with complex advice and requires oversight to meet regulatory policy standards (4). Similarly, Cao highlighted the broader benefits of AI in the financial sector, emphasizing its potential to handle complex challenges and create an intelligence-drive economy (5). However, Biswas et al. investigated more specific, significant limitations, including GPT-4's reliance on outdated data and concerns over data security and model transparency, which could compromise recommendation reliability (6). These findings underscore the need for more detailed research to address the limitations of GPT-4's advisory capabilities and their use cased in the financial sector.

While each of these studies address the benefits and limitations of AI models such as ChatGPT in giving financial advice, the goal of this research is to analyze the effectiveness of ChatGPT's financial advice compared to finance advice given on social media platforms, such as Reddit. This broader investigation can also be broken into multiple, smaller core investigations: comparing GPT-4's financial advice to the advice given on "r/FinancialPlanning" forum; using quantitative analysis to determine the similarities and differences between them; understanding ChatGPT's strengths and weaknesses in giving financial advice; and finally, assessing the role that ChatGPT plays in financial forums, advisory, and the future.

We automated this comparison using natural language processing (NLP) techniques in order to quantitatively understand similarities between responses (7). Primarily, this study focused on content, or what feedback the model suggested, as well as structure, length, and potential ethical implications. We used a variety of quantitative measures to

https://doi.org/10.59720/23-296

compare AI-generated responses to Reddit user comments.

We initially hypothesized that Chat-GPT would provide succinct, yet general feedback that referred predominantly to common financial habits rather than individualized financial goals. However, the analyses revealed that Chat-GPT generally aligned with Reddit responses in terms of word content, demonstrating Chat-GPT's ability to generate similar financial content, yet in prolific wording, allowing us to further analyze its impact on financial advisory in the future. This investigation revealed that Chat-GPT was more generalized and repetitive in responding to Reddit questions, suggesting that it is currently unable to replace traditional financial advisory.

RESULTS

We used multiple quantitative analyses to compare GPT-4 response performance and Reddit user comments. We first scraped 100 posts from the "r/Financial Planning" sub-Reddit and generated GPT-4 responses to those questions. First, we collected average word counts from both GPT-4 and Reddit comment responses to ascertain and compare the amount of information given in the response and its conciseness (**Figure 1**). To compare word counts between GPT responses and Reddit comments, we used a two-sample *t*-test. The corresponding p-value of 1.57e-48 suggested a significant distinction in word counts between the two communication mediums.

In order to determine the presence of a correlation between word count of Reddit input vs word count of GPT-4 response, we performed Pearson correlation analysis **(Figure 2)**. We used the Pearson correlation coefficient (PCC) to determine linear correlation between the two, and returned a coefficient of 0.28, as well as a p-value of 0.005.

We also collected the Reddit response word counts according to the word counts of their respective questions using a scatter plot and returned a PCC of 0.27 and a p-value of 0.008. At the 0.05 significance level, this indicates a weak yet statistically significant correlation between Reddit input and output word counts (**Figure 3**). We observed this correlation with both GPT and Reddit outputs, suggesting that both respond similarly to responses of the same word count.

Figure 1: Comparison of word counts between GPT-4 response and Reddit Comments. Box plots with error bars represent 2 standard errors indicate average word counts observed in GPT-4 and Reddit comment responses. Out of a total sample size of 100, outliers were plotted with hollow circles, and word counts were derived from previous analysis. With a two-sample *t*-test, the corresponding *p*-value was 1.57e-48.

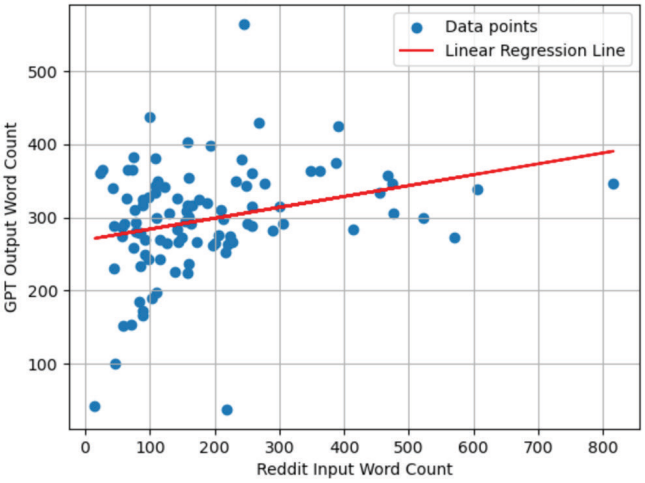

Figure 2: Reddit question word count in comparison to GPT-4 output word count with linear regression line. Scatter plot of the GPT-4 output lengths. Linear regression line: y = 0.15x + 269.13. Pearson's correlation coefficient: 0.28.

Moreover, one of the methods of comparison, the spaCy similarity algorithm, is a natural language processing (NLP) technique to quantify sentence similarity based on the similarity of word vectors and structures (8). This algorithm produces similarity scores, where a similarity score of 100% indicates perfect word vector similarity, while 0.0% indicates no similarity at all. The mean of the output (similarity score) was 80.3%, while the median was 83.4% (**Figure 4**). Additionally, 97.0% of similarity scores were above 50.0% similar, and 74.00% were above 75.0% similar (**Figure 4**). These findings revealed a high average similarity between the two texts, showing similarity in their word vector structure.

Based on empirical observation of the Reddit responses, we observed many instances of negative, neutral, and positive sentiment. Due to this observation, we collected the distribution of sentiment labels between GPT-4 responses and Reddit comment responses to determine the tone

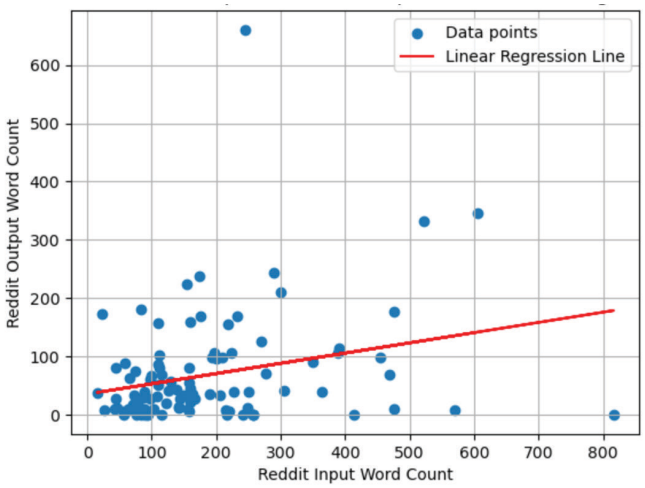

Figure 3: Scatter plot mapping Reddit question word count in comparison to Reddit comment word count. Linear regression line: y = 0.18x + 36.23. Correlation coefficient: 0.27.

https://doi.org/10.59720/23-296

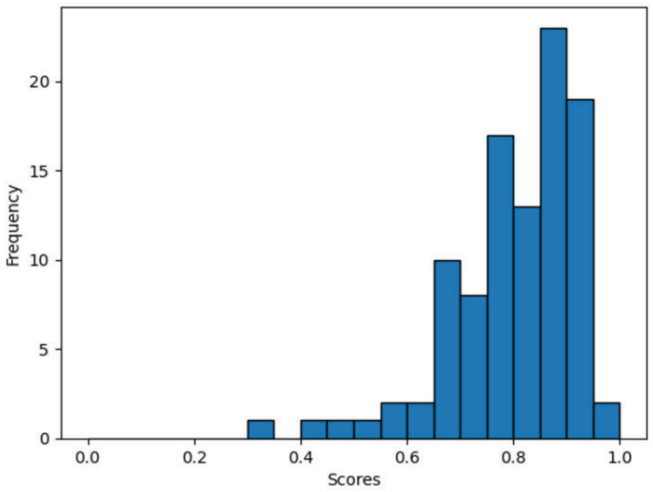

Figure 4: Distribution of similarity scores between GPT-4 Output and Reddit comments. Histogram showing spaCy similarity score frequencies, ranging between 0.0-1.0. SpaCy library was used to collect similarity results of 100 GPT and Reddit response samples each.

and sentiment that the responses have, to see the attitude with which the information is presented (**Figure 5**). This investigation used the Twitter RoBERTa model for sentiment analysis because it is a model trained on social media text interactions. Sentiment analysis is a computational process that identifies the attitude of a text as positive, negative, or neutral. Both GPT and Reddit had a similar distribution of neutral text sentiment, although Reddit comments had a higher proportion of negative sentiment and GPT responses had a higher proportion of positive sentiment (**Figure 5**). However, there was no significant statistical differences between the sentiment (*p*-value = 0.93, chi-squared test).

Additionally, we analyzed our n-gram (9) word patterns found throughout the texts to determine common phrasing patterns and identify structural similarities to determine (**Tables 1 and 2**). An n-gram is a series of n successive items across a body of text used to reveal common text sequences and word patterns. This analysis revealed repetition and patterns employed by GPT-4 and how this impacts the content, quality, and length of the response. We collected the top 10 most frequently found bigrams and trigrams within the GPT-4 response, alongside their frequency, in the below tables. The trigram "consult, financial, adviser" appeared most frequently, while the bigram containing "financial, adviser" appeared 107 times (**Table 1).** However, we observed very few bi- and tri-grams in Reddit responses, and their highest frequency of appearance was only 4, while in comparison N-grams appeared with significantly higher frequency in GPT-4 responses (**Table 2**).

DISCUSSION

This study evaluated the capabilities of GPT-4 in providing financial advice by comparing it to Reddit questions and responses. We used various quantitative measures to compare the accuracy of this LLM. These measures revealed multiple core differences in the word structure, length, and repetition between GPT-4 and Reddit responses.

The first method of analysis we employed, the spaCy similarity algorithm, found similarity levels between the word vectors of the GPT-4 responses and Reddit Comments. The results of this algorithm indicated a clear similarity in word structure and vocabulary between the two texts, demonstrating a clear connection in the advice and financial content that they are recommending. However, the GPT-4 response is notably more prolific in comparison to Reddit comments. Reddit comments are substantially shorter and more focused on the question, while GPT-4 responses are

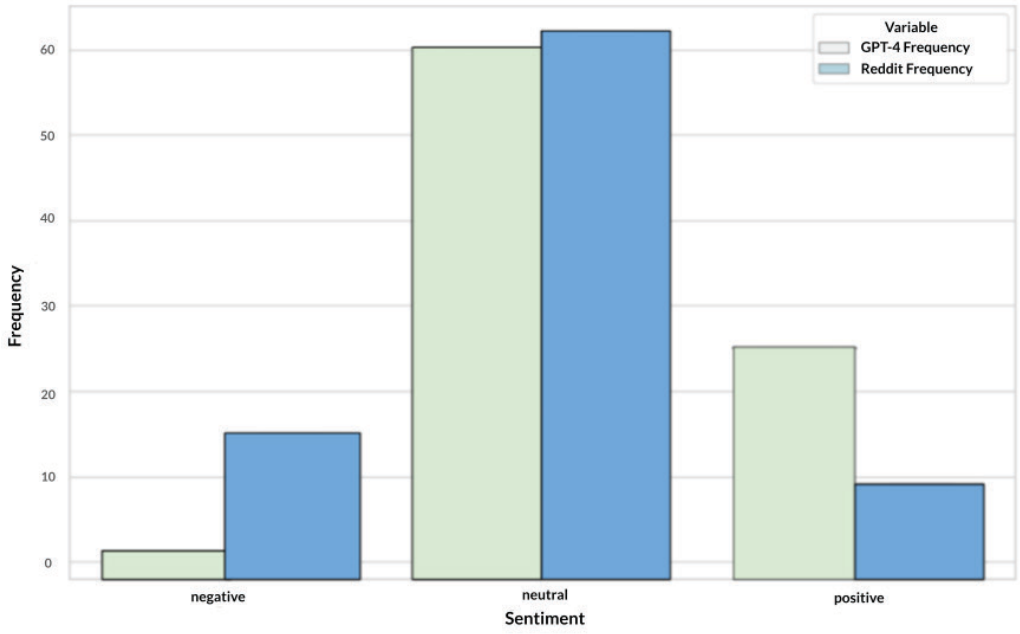

Figure 5: Frequency of sentiments for GPT responses and Reddit comments. Bar chart offers a direct comparison of sentiment of GPT responses and Reddit comments, light green indicating GPT and blue indicating comment responses. Sentiment labels collected, represented in a bar chart. The sample size was n = 100 responses, and chi-squared test resulted in *p*-value = 0.93.

https://doi.org/10.59720/23-296

GPT Response Bigram		GPT Response Trigram	
2-gram	frequency	3-gram	frequency
(financial, advisor)	107	(consult, financial, advisor)	34
(emergency, fund)	88	(provide, personalized, advice)	20
(financial, goals)	72	(financial, advisor, provide)	19
(roth, ira)	54	(based, information, provided)	18
(risk, tolerance)	50	(goals, risk, tolerance)	16
(financial, situation)	49	(consulting, financial, advisor)	16
(interest, rate)	43	(make, informed, decision)	15
(real, estate)	41	(help, make, informed)	14
(credit, card)	40	(personalized, advice, based)	14
(retirement, savings)	37	(credit, card, debt)	14

Table 1: Bi- and tri- n-gram analyses of GPT responses. Tables demonstrate the results of n-gram analysis, showing word sequence patterns throughout the GPT and comment responses collected using Python n-gram library after data pre-processing. Collected from 100 GPT responses.

longer. Instead of directly providing an opinion or response, which AI cannot do, GPT-4 suggested many avenues that are potentially viable options.

In fact, when asked why it cannot provide financial advice, GPT-4 states that it cannot provide personal finance advice without "understanding of an individual's unique financial situation, goals, [and] risk tolerance," but can provide general information and answer non-specific questions. While GPT-4 is able to provide a multitude of strong suggestions, it rarely recommends a specific financial strategy, showing a substantial difference between its advice and Reddit responses. As seen in the "Response structure" section of the results, the response provides multiple, individual steps rather than one compact answer. Moreover, the response ends with a recommendation to consult with a financial advisor–something that few Reddit user comments do, as seen through the n-gram analysis. Additionally, the high frequency of n-grams in GPT-4 responses compared to very few n-gram responses in Reddit comments may also suggest that ChatGPT responses are less personalized compared to Reddit and are instead more general. This indicates that GPT-4 generated advice is much less specific to the question asked, while Reddit is more targeted to the individual content of the question. Analysis of the differences in structure reveal GPT-4's current inability to provide opinion or specific financial advice, potentially because of safeguards implemented by

Comment Response Bigram		Comment Response Trigram	
2-gram	frequency	3-gram	frequency
(roth, ira)	20	(credit, card, debt)	4
(emergency, fund)	15	(max, roth, ira)	3
(discretionary, income)	8	(taxable, brokerage, account)	3
(brokerage, account)	7	(selling, car, buying)	2
		(carrying, credit, card)	2

Table 2: Bi- and tri- n-gram analyses of Reddit comment responses. Tables demonstrate the results of n-gram analysis, showing word sequence patterns throughout the GPT and comment responses collected using Python n-gram library after data pre-processing. Collected from 100 Reddit responses.

OpenAI (10). GPT-4 is currently not a substitute for peer-to-peer financial advice, but is capable of providing general advice, information, and feedback.

We conducted sentiment analysis to determine the emotional tone of GPT-4 responses and Reddit comments (11). There are no significant differences in sentiment, and both are predominantly neutral; however, GPT-4's responses are more positive than Reddit comments, while Reddit has a higher proportion of negative sentiment comments (Figure 3). GPT-4 primarily responds with neutral sentiment but may offer words of encouragement or praise instead of harsher criticism. In contrast, some Reddit responses have negative sentiment, due to occasional critical nature of user comments. Responses with negative sentiment can convey disapproval, which can be important in financial advice settings (12). Because GPT-4 does not report any instances of negative sentiment, GPT-4 can fail to emphasize financial risks or critique current financial decisions, showing a significant difference in its capabilities. GPT-4 responses, however, can present a multitude of different suggestions in a manner that is similar to Reddit responses.

Although this study does not explicitly evaluate GPT-4 responses for accuracy, it is able to evaluate the performance of GPT-4 in peer-to-peer forums in financial spaces. GPT-4, while competent in presenting substantial financial advice, is not able to condense and deliver responses in the same structure and format as other users on the forum. At its current stage, GPT-4 does not directly present a risk towards such peer-to-peer advice forums due to its inability to provide direct advice, instead of referring to financial advisors, products, etc.

Moreover, the responses collected from the GPT-4 demonstrate a clear pattern: GPT-4 consistently refers to financial advisors to discuss and make financial decisions. At its current capability, GPT-4 cannot access raw financial data, directly suggest recommended paths, or necessarily pose a direct threat to employment of financial professionals. However, for those with general questions, GPT can provide a starting point with strong ideas and pathways to explore.

Other similar research that has been conducted often compares GPT-4 responses to real financial advisor advice, while ours particularly focuses on other Reddit users to explore how such LLMs might shape the future of such forums. While this research does not evaluate the accuracy

https://doi.org/10.59720/23-296

and content of GPT's responses, it instead uses multiple quantitative metrics to gauge GPT's ability to emulate human responses in regards to finance related questions. This study uses multiple comparative metrics to gauge GPT's ability to emulate human responses in financial-related questions.

Furthermore, it is important to consider the confounding factor that Chat-GPT's response can be altered by the command's specificity. The specificity and content of the scraped Reddit questions during this study could result in an altered GPT response. Therefore, an interesting course of further research would be to understand how a command's structure, language, and specificity, could affect the strength of GPT-4's financial advice.

To extend and improve the scope of this study, it would be interesting to evaluate the factual accuracy of GPT-4 recommendations to understand the accuracy of content and whether the model hallucinates, or generates incorrect data. At its current scope, this study evaluates GPT-4's performance in giving peer-to-peer, forum-style financial advice; however, a possible future direction would be to compare GPT-4's performance to that of a professional financial advisor, to understand its pitfalls and potential impact on the field of financial advisory. Furthermore, this study uses a pre-trained GPT-4 model, which is not specifically targeted to generate financial advice. In a future study, it would be interesting to understand how specific, financially-trained generative models perform under the same comparisons with Reddit comments.

Based on the findings of this research, many comparisons can be drawn between GPT-4's response to financial questions and actual Reddit user comments. Its ability and efficiency to provide financial recommendations is comparable to human comments, but unparalleled in speed, efficiency, and language processing skills. Despite this, the results of this study present potential drawbacks of GPT-4's financial advice including its length, lack of definitive answer, and constant repetition in word structure patterns. This reveals that at this point, GPT-4 cannot be a suitable alternative for financial forums in terms of advice structure.

MATERIALS AND METHODS

Tools

This study used the Jupyter Notebook editor for Python programming. To collect the data used for this research, the Reddit API was used, while OpenAI's GPT-4 API was used to generate LLM responses. The following packages were used for data manipulation, analysis, and presentation: Pandas, Seaborn, spaCy, SciPy, Natural Language Toolkit (NLTK), and MatPlotLib.

Data Collection

To collect the raw user data, the Reddit API was used due to its accessibility, as well as long standing relevance in financial spaces. Due to the focus of the study on financial advisory, the API was used to scrape from the forum "r/FinancialPlanning," a popular forum to ask and provide personal financial advice, which has over 685K users . In this subreddit, users ask for financial advice in posts, while fellow users can respond to them in comments, which receive upvotes based on community perception of their accuracy and quality. Due to time and material constraints, only the 100 most recent posts were scraped, alongside each of their most upvoted comments to gauge the most popular response to the question. Comments that were in foreign languages or were gibberish were removed, from collection, and replaced by the next most upvoted comment. Using the Pandas library, this data was stored as a dataframe for later manipulation and analyses.

Generating GPT-4 Responses

This study utilized the GPT-4 API's Chat Completion API to generate LLM responses to community questions posed on Reddit (13). This API uses GPT-4, an advanced LLM notable for its capabilities, as well as passing the bar exam amongst the top 10% of test takers. The questions posed on the forum were inputted into the API, and their responses were collected within the data frame.

Data Pre-Processing

To ensure that the data was understandable by the SpaCy similarity algorithm, the NLTK library was first implemented to pre-process our data to improve the accuracy of the natural language processing algorithms. This processing included: removing stop words which includes removing articles, pronouns, and conjunctions that don't add significant information to the text; making the data lowercase and removing any symbols; and performing lemmatization or converting words into their meaningful base form (14). The process of lemmatization is important in optimizing the performance in sentiment analysis algorithms (15). For instance, the words 'change,' 'changing,' and 'changes' would all be lemmatized into the base word 'change' to ensure the algorithm understands them. These preprocessing methods helped to improve the performance and accuracy of the NLP conducted.

SpaCy Similarity

The spaCy similarity algorithm was used to quantify the similarity of the two sentences using a score from 0 to 1, semantically. For this analysis, spaCy used a large model containing word vectors. This similarity is by comparing multi-dimensional meanings or word vectors of words within the text. This similarity score was collected within the dataframe, and the MatPlotLib library was used to generate graphs and plots to visualize the results.

Sentiment Analysis with Twitter-RoBERTa:

Sentiment analysis of both the AI response and Reddit user comments was conducted in order to gauge differences and similarities in sentiment between them. While the content of this data surrounded financial matters, its structure and content aligned most closely with models trained on social media data, rather than financial news and markets data. Hence, the "Twitter-roBERTa-base for Sentiment Analysis" model was employed for this analysis due to the similarity between its training content and the collected data used in this study. The labels (negative, neutral, positive) were collected within the dataframe for analysis.

Bi/Tri-gram Analysis:

Often, meaningful phrases or sequences of words are more informative and descriptive than singular words. N-gram analyses allow for deeper word and sentence structure analysis to complement NLP. Hence, bigram and trigram

https://doi.org/10.59720/23-296

analyses were implemented on GPT response and Reddit comment data in order to reveal word sequence patterns that appear across collected data. The NLTK tokenizer and ngrams library was used for this analysis, alongside the collections library (16). These word patterns were sorted by frequency, in order to be analyzed later.

Statistical Analysis

After conducting all of the above quantitative analyses, the chi-square test was used to determine statistical significance in the difference between GPT-4 outputs and Reddit comments in a variety of tests. Furthermore, Pearson Correlation Coefficient was used to calculate correlation (17).

ACKOWLEDGMENTS

We would like to thank Shrinivas Samant for assistance in formatting this paper, as well as our classmates at the Marvel Research Program for their thoughtful collaboration and inspiration. Without them, this paper would not have been possible!

Received: October 18, 2023
Accepted: March 5, 2024
Published: August 13, 2024

REFERENCES

1. Zhu, Hui, *et al.* "Implementing artificial intelligence empowered financial advisory services: A literature review and critical research agenda." *Journal of Business Research,* vol. 174, 2024, p. 114494, https://doi.org/10.1016/j.jbusres.2023.114494.
2. Achiam, Josh, *et al.* "GPT-4 Technical Report." 2023, https://doi.org/10.48550/arXiv.2303.08774.
3. Deng, Xiang, *et al.* "What Do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis." *ArXiv.org,* 21 Dec. 2022. https://doi.org/10.1145/3543873.3587324.
4. Neilson, Ben. "Artificial Intelligence Authoring Financial Recommendations: Comparative Australian Evidence." *Journal of Financial Regulation,* 15 May 2023, Accessed 16 Sept. 2023.
5. Cao, Longbing. "AI in Finance: A Review." *SSRN Electronic Journal,* 2020, https://doi.org/10.2139/ssrn.3647625.
6. Biswas, Sanjib, *et al.* ChatGPT in Investment Decision Making: An Introductory Discussion. 2 May 2023.
7. Chowdhary, K. R. "Natural Language Processing." *Fundamentals of Artificial Intelligence,* 2020, pp.03-649, https://doi.org/10.1007/978-81-322-3972-7_19.
8. "Linguistic Features · SpaCy Usage Documentation." Linguistic Features, spacy.io/usage/linguistic-features.
9. Qin, Han, *et al.* Relation Extraction with Word Graphs from N-Grams. Association for Computational Linguistics, 2868.
10. OpenAI. "Our approach to AI safety." OpenAI, 2023, openai.com/index/our-approach-to-ai-safety/. Accessed 4 June 2024.
11. Khader, Mariam, *et al.* "The Impact of Natural Language Preprocessing on Big Data Sentiment Analysis." *The International Arab Journal of Information Technology,* vol. 16, 1 Jan. 2019, pp. 506-513. Accessed 8 Oct. 2023.
12. Madamba, Anna, *et al.* "The value of advice: Assessing the role of emotions." *Vanguard Research,* Mar. 2020, Accessed 4 June 2024.
13. "GPT-4 API General Availability and Deprecation of Older Models in the Completions API." Openai.com, openai.com/blog/gpt-4-api-general-availability.
14. Sarica, Serhad, and Jianxi Luo. "Stopwords in Technical Language Processing." *PLOS ONE,* vol. 16, no. 8, 5 Aug. 2021, p. e0254937, arxiv.org/abs/2006.02633, https://doi.org/10.1371/journal.pone.0254937.
15. Plisson, Joël, *et al.* A Rule Based Approach to Word Lemmatization.
16. "Collections - Container Datatypes - Python 3.8.3 Documentation." Docs.python.org, docs.python.org/3/library/collections.html.
17. Adler, Jeremy, and Ingela Parmryd. "Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient Is Superior to the Mander's Overlap Coefficient." *Cytometry Part A,* vol. 77A, no. 8, 30 Mar. 2010, pp. 733-742, https://doi.org/10.1002/cyto.a.20896.
18. Loureiro, Daniel, *et al.* TimeLMs: Diachronic Language Models from Twitter. 8 Feb. 2022, Accessed 29 June 2023. https://doi.org/10.18653/v1/2022.acl-demo.25.
19. Milica Miocević, and Van De. Small Sample Size Solutions : A Guide for Applied Researchers and Practitioners. Abingdon, Oxon ; New York, NY, Routledge, 2020.

APPENDIX

GitHub with code: github.com/flytrotter/GPT_Financial_Forums_Code