

Article

Large Language Models: Their Success and Impact

Spyros Makridakis ¹, Fotios Petropoulos ^{1,2} and Yanfei Kang ^{3,*}

¹ Institute For the Future, University of Nicosia, Nicosia 2414, Cyprus; makridakis.s@unic.ac.cy (S.M.); fotios@bath.edu (F.P.)

² School of Management, University of Bath, Bath BA2 7AY, UK

³ School of Economics and Management, Beihang University, Beijing 100191, China

* Correspondence: yanfeikang@buaa.edu.cn

Abstract: ChatGPT, a state-of-the-art large language model (LLM), is revolutionizing the AI field by exhibiting humanlike skills in a range of tasks that include understanding and answering natural language questions, translating languages, writing code, passing professional exams, and even composing poetry, among its other abilities. ChatGPT has gained an immense popularity since its launch, amassing 100 million active monthly users in just two months, thereby establishing itself as the fastest-growing consumer application to date. This paper discusses the reasons for its success as well as the future prospects of similar large language models (LLMs), with an emphasis on their potential impact on forecasting, a specialized and domain-specific field. This is achieved by first comparing the correctness of the answers of the standard ChatGPT and a custom one, trained using published papers from a subfield of forecasting where the answers to the questions asked are known, allowing us to determine their correctness compared to those of the two ChatGPT versions. Then, we also compare the responses of the two versions on how judgmental adjustments to the statistical/ML forecasts should be applied by firms to improve their accuracy. The paper concludes by considering the future of LLMs and their impact on all aspects of our life and work, as well as on the field of forecasting specifically. Finally, the conclusion section is generated by ChatGPT, which was provided with a condensed version of this paper and asked to write a four-paragraph conclusion.

Keywords: Large Language Models; Forecasting; ChatGPT



Citation: Makridakis, S.; Petropoulos, F.; Kang, Y. Large Language Models: Their Success and Impact. *Forecasting* **2023**, *5*, 536–549. <https://doi.org/10.3390/forecast5030030>

Academic Editors: Sonia Leva and Alicia Troncoso

Received: 12 July 2023

Revised: 18 August 2023

Accepted: 19 August 2023

Published: 25 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. The Success of ChatGPT as a Large Language Model (LLM)

Some innovative breakthroughs can be anticipated in advance, but according to inside information, ChatGPT was a big surprise: “When OpenAI launched ChatGPT with zero fanfare in late November 2022, the San Francisco-based artificial intelligence company had few expectations. Certainly, nobody inside OpenAI was prepared for a viral mega-hit” [1]. ChatGPT’s essential advantage that has captured the imagination of users is its ability to communicate in natural language by understanding and answering questions posed in plain English. While it does not generate new knowledge, its answers, drawn from its training on a massive amount of diverse information from sources such as Wikipedia, books, articles, newspapers, and the internet, are relevant and often thought-provoking. ChatGPT users are captivated from the very beginning, as the conversations feel like discussions with a friend or colleague rather than interactions with an AI system. Furthermore, this willing and tireless companion can engage in debates on any topic, showing an eagerness to listen and provide answers in everyday English, all for free. There are no limits to the number of questions that can be asked, and users have the flexibility to specify the manner, type, and length of the answers in advance. Users can even request repeated answers or delve into greater depth for a particular question. These characteristics have contributed to the phenomenal success of ChatGPT, sparking considerable interest and driving accelerated efforts for its further improvement, as a competitive race has started and will intensify with tech giants and startups [2] competing for a share in the promising, lucrative LLMs market.

Alongside its many advantages, there are also concerns associated with LLMs, spanning from the conventional criticisms of job displacement and an exacerbation of wealth inequality to the potential for fostering academic dishonesty in schools and universities while perpetuating biases and prejudices. A significant technical issue, however, poses a more serious challenge: LLMs are prone to “hallucinations”, which entail providing responses that sound convincing but lack a foundation in reality. Another major apprehension is that the same question can yield vastly different answers across different sessions. These indeed stand as the primary drawbacks. Ascertaining whether a given answer is truthful or fabricated becomes a perplexing task. Thus, ChatGPT, in its current form, cannot be employed as a direct substitute for a search engine. This raises questions about how it could be seamlessly integrated into Microsoft’s Bing or Google’s Bard search engines [3]. Given the paramount importance of answer accuracy, we delved into the responses of the two versions of ChatGPT, posing queries from a subfield of forecasting where correct answers are established, in order to gauge the extent of their precision.

2. The Accuracy of the Standard and Customized ChatGPT

In this section, we analyze and assess the responses provided by OpenAI’s ChatGPT (referred to as “ChatGPT” hereafter) and a custom-trained version of ChatGPT (referred to as “CustomGPT”) in terms of their accuracy when addressing questions related to M forecasting competitions—a specialized area within the field. The inherent advantage of such an approach lies in the authors’ possession of the correct answers (one being the organizer of the M competitions and the other a member of the organizing team). This allows for a direct comparison between the responses of the two ChatGPT versions and the known correctness of the answers, thereby enabling us to ascertain their accuracy and arrive at conclusions regarding the reliability of their responses.

In order to train the CustomGPT, we fed into it all the papers from the M3 [4], M4 [5], and M5 [6] special issues published by the *International Journal of Forecasting* as well as papers from the ScienceDirect database by searching “M competition OR M2 competition OR M3 competition OR M4 competition OR M5 competition” in the Title field and “forecast” in the “Title, abstract, keywords” field. Note that in the ScienceDirect search API (https://dev.elsevier.com/sd_apis.html, accessed on 18 August 2023), punctuation is ignored in a phrase search, so the searches “M3 competition” and “M3-competition” return the same results. Plurals and spelling variants are accounted for; thus, the search term “forecast” encompasses “forecasting”. In total, 89 papers were utilized for the training of CustomGPT, and the same set of questions was presented to ChatGPT for the purpose of comparing their responses. To begin, we preprocessed these papers by calculating an embedding vector for each segment (with a length set to 3000 characters for this study). Embedding aided in gauging the textual similarities. When a specific question was posed, CustomGPT initially identified pertinent segments from the 89 papers, employing them to construct a prompt within the OpenAI Chat API (<https://platform.openai.com/docs/guides/chat>, accessed on 18 August 2023). Figure 1 shows how CustomGPT provides answers to user questions.

First, we queried both GPTs with the question: “What can you tell me about the M forecasting competitions?”. Both GPTs provided sound responses to this inquiry, accurately summarizing the essential details surrounding the M competitions and their objectives. However, the replies exhibited notable differences. ChatGPT’s response included information about the competition organizers, the frequency of the competitions, and the employed error measures for forecast evaluation. In contrast, CustomGPT’s reply seemed to place excessive emphasis on the M5 forecasting competition. Both GPTs furnished accurate responses concerning the principal aims of the competitions. Specifically, ChatGPT stated “The M competitions are designed to evaluate the accuracy of various forecasting methods and to provide insight into the most effective methods for forecasting different types of time series data” while CustomGPT replied “The M forecasting competitions are a series of forecasting competitions that aim to empirically evaluate the performance of existing and new forecasting methods, allowing the equivalent of experimentation widely used in

hard sciences". Both GPTs were accurate in their responses, and it could be argued that a synthesis of their two answers would yield an even more comprehensive reply to our initial question.

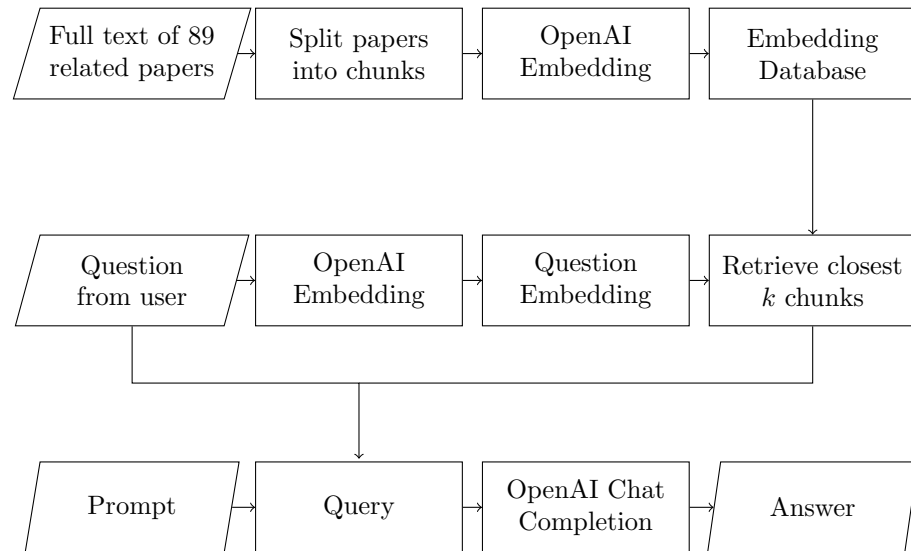


Figure 1. The process of how CustomGPT addresses user questions related to M forecasting competitions.

Next, we asked the two GPTs to tell us which was the most popular M competition in terms of teams/participants. ChatGPT’s reply started offering some generic information regarding the M competitions. Then, its reply included the following: “[...] the M4 forecasting competition, which was held in 2018, is one of the most popular forecasting competitions in terms of teams/participants. The competition focused on forecasting 100,000 time series from a variety of industries, and attracted over 1000 participants from around the world, including academic researchers, practitioners, and students”. This statement contains inaccuracies on two fronts (also see Table 1). Firstly, M5, not M4, stands as the most widely participated forecasting competition, and secondly, the M4 competition witnessed only 49 entries (along with 12 benchmarks), not the over 1000 count mentioned by ChatGPT. Within the same query, CustomGPT’s response distinctly zeroed in on the most prominent forecasting competition to date—the M5 forecasting competition—providing specific participation figures along with a breakdown by country. All in all, CustomGPT’s answer was significantly more accurate and focused.

Table 1. Entries (teams participations) in M3, M4, and M5 forecasting competitions.

Competition	Participants
M3	24 entries
M4	49 entries and 12 benchmarks
M5	5507 (accuracy track) and 892 (uncertainty track) entries

Focusing on the M3 forecasting competition, we queried the two GPTs about the number of time series that composed M3 (also refer to Table 2). While both responses accurately presented the count of time series (3003), ChatGPT’s answer was notably lengthier. However, it included incorrect information regarding the competition organizer, attributing it to the International Institute of Forecasters instead of Spyros Makridakis and Michelle Hibon. Furthermore, it wrongly identified the journal of publication as the *Journal of Forecasting* rather than the *International Journal of Forecasting*.

Moving on to a question about the more accurate method employed in the M3 competition, ChatGPT’s response was again of greater length, incorporating seemingly generic and unnecessary details, but did mention Theta as the accurate top-performing submission.

In contrast, CustomGPT's answer was more concise and provided the requested information along with the specific reference to the paper where the main competition results were summarized.

When we proceeded to query the two GPTs about the three principal findings of the M3 competition, their replies differed significantly, yet both contained precise information regarding the primary discoveries.

Table 2. Time series used in M3, M4, and M5 forecasting competitions.

Competition	Number of Time Series
M3	3003
M4	100,000
M5	30,490

We subsequently queried the two GPTs about the shared characteristics and distinctions between the outcomes of the M3 and M4 competitions. Remarkably, ChatGPT's response addressed the ensuing statement as both a point of similarity and a point of divergence: "The performance of individual methods varied across different forecast horizons and series, indicating that there is no single 'one-size-fits-all' forecasting method that works best for all types of time series". Furthermore, ChatGPT erroneously stated that the M4 competition permitted participants to incorporate external information into their forecasts, which was not accurate. On the other hand, CustomGPT's response was more targeted and, overall more precise. It accurately highlighted that a key distinction lay in the fact that the M4 competition also centered on evaluating forecast uncertainty performance.

In the subsequent pair of questions, we directed ChatGPT and CustomGPT to identify the accuracy measures utilized in the M4 and M5 forecasting competitions (also refer to Table 3). Regarding the M4 competition, ChatGPT's response enumerated three accuracy metrics: the symmetric mean absolute percentage error (SMAPE), the mean absolute scaled error (MASE), and the "Geometric Mean". While the first two were accurate, the third, "Geometric Mean", was not employed in any of the M forecasting competitions. Furthermore, ChatGPT omitted the mention of the overall weighted average (OWA). On the other hand, CustomGPT's reply was impeccable, precisely listing SMAPE, MASE, and OWA, and even surpassing by acknowledging "ranks", which were employed in certain statistical tests to assess accuracy in the competition's results analysis for some specific purposes.

Transitioning to the accuracy measures in the M5 competition, ChatGPT's response included the correct measure, the weighted root-mean-square scaled error (WRMSSE). However, it also incorporated two additional measures that were not pertinent: the first was a measure employed in the M5 competition but intended for the uncertainty track, not the accuracy track; the second was entirely irrelevant. Similarly, CustomGPT's reply mentioned the correct measure, WRMSSE, but introduced two other accuracy measures, MASE and RMSE, that were not explicitly utilized in this competition. Thus, neither GPT was entirely successful in providing completely accurate information concerning the accuracy measures of the M5 forecasting competition.

Table 3. Main performance measures in M3, M4, and M5 forecasting competitions.

Competition	Participants
M3	sMAPE
M4	sMAPE, MASE, MSIS
M5	WRMSSE, WSPL

Concentrating further on the M5 competition, we queried both GPTs about the number of teams that took part in the accuracy track. ChatGPT's reply was as follows: "A total

of 4747 teams participated in the M5 Accuracy competition”, a wrong answer. On the other hand, CustomGPT correctly replied: “The M5 Accuracy competition saw a total of 5507 teams from 101 countries participating. A majority of the teams, 4373 (79.4%), entered during the validation phase, while 1134 (20.6%) teams entered during the test phase and 1434 (26.0%) teams entered during both phases. Of the total teams that participated, 2939 (53.4%) only submitted during the validation phase. The competition saw 88,136 submissions in total from all participating teams.” CustomGPT’s response, albeit lengthier, was entirely accurate, thereby illustrating that training LLMs on specific and targeted sources of information results in more precise and reliable answers.

The subsequent inquiry revolved around the victorious participants of the M5 competition, specifically within the accuracy track. ChatGPT’s response, although well-structured, was entirely erroneous. It incorrectly attributed the win to Rob Hyndman and George Athanasopoulos, who did not partake in the M5 competition but were members of the team that secured the second-best submission in the M4 competition. More concerning, ChatGPT incorrectly indicated that Professors Hyndman and Athanasopoulos were affiliated with the University of Sydney (correct country, incorrect city/university). On the contrary, CustomGPT managed to succinctly list the top three performing submissions in the M5 competition’s accuracy track. Shifting to a similar question but concerning the uncertainty track of the M5 competition, ChatGPT’s responses once again veered off course, containing inaccuracies and providing information that was not requested. Conversely, CustomGPT’s answers were spot-on, additionally including the extent by which each of the top three winning submissions outperformed the benchmark.

In a parallel question concerning M3, we prompted both GPTs to outline the three key discoveries from the M5 competition. ChatGPT’s response failed to center around the M5 competition itself and instead encompassed a range of findings from different forecasting competitions. On the other hand, CustomGPT’s reply was more precise, offering two accurate major findings. However, the third listed finding, while not incorrect, did not align with any of the pivotal conclusions explicitly detailed in the published paper presenting the M5 competition results [7].

Our subsequent inquiry revolved around the five principal findings spanning all M competitions. ChatGPT’s response concentrated largely on earlier competition findings, whereas CustomGPT’s answer encompassed significant conclusions from later competitions, such as “Machine learning (ML) methods performed better than statistical methods in the M4 and M5 competitions, but winning methods vary over time”. Both GPTs accurately noted that employing combinations as a strategy generally enhanced performance. They also correctly highlighted that the winning approach hinges on the specific attributes of the forecasted data and the methods employed.

Lastly, we queried the GPTs about their knowledge regarding the M6 forecasting competition. ChatGPT’s response was notably fraught with inaccuracies: “The M6 competition is an annual forecasting competition organized by the International Institute of Forecasters (IIF). It focuses on forecasting future values of a specific time series, which is typically a macroeconomic or financial variable. [...] The competition has been running since 2010, and each year, a different set of time series is chosen as the target variable. Some of the past target variables have included monthly retail sales, monthly energy consumption, and quarterly GDP” (all “hallucinations”). On the other hand, CustomGPT, having been trained on the abstract of the M6 competition, accurately responded that M6 entails predicting ranks for 100 financial assets and adopts a “duathlon” competition format, wherein participants are required to submit both forecasts and investment decisions.

In our analysis, we noted significant instances of factual inaccuracies within the responses from the standard ChatGPT, even encompassing instances of “hallucinations”. In contrast, the responses from CustomGPT were generally more succinct, focused, and notably more accurate in the majority of cases. However, it is important to emphasize that neither GPT achieved perfection. Consequently, we recommend exercising caution when

relying on information provided by LLMs without question, even if they have been trained on materials directly pertinent to the queries posed.

Above all, it is evident that considerable work remains to address LLMs' "hallucination" issue. Yann LeCun aptly suggests that this challenge may necessitate a fundamental paradigm shift in AI methodology—a shift towards imbuing AI with common sense and genuine understanding before tackling this problem [8]. Until such advancements are achieved, the current capabilities of ChatGPT need refinement before their responses can be considered trustworthy. Even a purported 99% accuracy rate may not suffice.

3. Judgmental Adjustments to the Statistical/ML Forecasts

For statistical/ML forecasts to yield accurate predictions, it is crucial that established patterns and relationships remain stable during the forecasting phase. However, this scenario is seldom the reality, given the potential impact of factors such as promotional campaigns, pricing alterations by firms or competitors, as well as economic and environmental influences. These variables can disrupt historical patterns and relationships, necessitating judgmental adjustments to quantitative forecasts. Although substantial effort has gone into developing methods for these adjustments—aiming to minimize human and organizational biases—success has been limited [9]. In some cases, this has even led to a reduction in overall forecast accuracy [10].

Given their significance, this section explores the potential value that ChatGPT offers in enhancing the precision of judgmental forecasts. This is done by evaluating the responses of the GPTs to questions concerning the utilization of such forecasts. Just as before, ChatGPT drew from the information available in its generic training database, following OpenAI's approach. Conversely, to train CustomGPT, we employed a set of 166 papers culled from the ScienceDirect database. These papers were selected using the search query "(judgmental OR intervention OR override) AND forecasting" specifically in the title field. The training scheme for CustomGPT mirrors that of the earlier study on M competitions shown in Figure 1.

In this section, our evaluation does not focus on the accuracy of responses in terms of their correctness. Instead, we undertake a comparison based on their utility and comprehensiveness.

We first asked both GPTs to guide us towards pertinent literature concerning the adaptation of monthly sales forecasts through new information. ChatGPT's response adopted a bullet-point format, presenting two sets of information. The first set directed us to three widely recognized and widely used textbooks, in addition to referencing two journals with a focus on forecasting—the *International Journal of Forecasting* and the *Journal of Forecasting*. Additionally, ChatGPT offered a succinct rundown of "general steps" for adjusting forecasts based on fresh information, despite this specific query not being posed. In contrast, CustomGPT's reply took the form of an itemized list containing eight research papers relevant to the topic, along with their key findings, also noting that "the literature suggests that judgmental adjustments to statistical forecasts are common and can be predictable to some extent. There are various methods and heuristics that can be used to improve the accuracy of judgmental forecasts, and collaboration and communication among forecasters and other stakeholders can also be beneficial". Overall, the response of CustomGPT was more exact, providing some very specific suggestions to consult.

Continuing our inquiry, we posed a question to both GPTs regarding significant recommendations from the forecasting literature aimed at enhancing the precision of judgmental adjustments. ChatGPT's response honed in on structured methodologies and the involvement of multiple experts, emphasized the importance of feedback, underscored the utility of statistical benchmarks, and advised against succumbing to overconfidence or underconfidence. On the other hand, CustomGPT's reply offered a more detailed and practical perspective. It highlighted that judgmental adjustments are frequently employed in real-world scenarios but cautioned against potential biases that may arise from such adjustments. It also suggested that "adjustments can improve the accuracy of statistical forecasts under the right conditions" but also to discourage unnecessary adjustments to

statistical forecasts while requiring forecasters to record the reasons for their adjustments. In a comparable query, we prompted both GPTs to encapsulate the primary discoveries related to enhancing the precision of judgmental forecasts. ChatGPT's response echoed earlier sentiments, advocating for structured and collaborative methodologies, the engagement of multiple experts, and feedback provision. CustomGPT's reply, as before, was more specific. Along with its recommendations, we read that "The performance of purely quantitative forecasting methods can be flawed by factors such as instability or noise in the time series when historical data is limited", "judgmental adjustments to statistical forecasts can improve forecast accuracy in specific instances, but in general, they do not tend to improve the accuracy of density forecasts", and that "judgmental forecasts using contextual data can be significantly more accurate than quantitative forecasts".

In our next question, we sought guidelines/best practices on how to adjust statistical/ML forecasts. ChatGPT replied that its "guidelines and best practices can help ensure that adjustments to statistical/ML forecasts are based on relevant criteria, transparent, and validated. This can improve the accuracy of the forecasts and provide valuable insights for decision-making". The guidelines encompassed several recurring elements from previous inquiries, such as the advocacy for structured methodologies and the engagement of multiple experts. Additionally, the value of documenting the rationale behind adjustments was highlighted. ChatGPT's recommendations extended to utilizing judgment to identify pertinent information, coupled with monitoring, to gauge the propensity of systematic forecasting methods to over- or underforecast. CustomGPT's perspective remained more distinct. It emphasized the initial necessity of determining whether the system's forecast requires any adjustment. This pertains to scenarios where systematic biases may affect statistical/ML forecasts, or when specific information remains unaccounted for. CustomGPT also cited the suggestion from the literature for structured decomposition methods and underscored the positive utility of recording adjustment rationales. Furthermore, it advocated for the adoption of combination approaches involving both systematic methods and judgment.

In our question "If you were a consultant what steps would you have taken to improve the accuracy of judgmental forecasts?", ChatGPT's response started by advising to assess the current forecasting process. Other steps included the development of a plan for improvement, training, fostering a collaborative environment, and monitoring and evaluation towards continuous improvement. The only step that was specific to the context was the "use structured approaches for judgmental adjustments, such as the Delphi method or prediction markets, which are designed to minimize bias and encourage collaboration". CustomGPT's recommendations encompassed several facets, including the promotion of judgmental adjustments when valuable new information is integrated into forecasts, the application of bias-reduction techniques, fostering a healthy sense of skepticism, and even encouraging a critical approach towards positive adjustments. Overall, the responses from CustomGPT demonstrated a heightened awareness of the specific context and provided more precise and explicit steps for enhancement.

Subsequently, we requested both LLMs to furnish guidelines or best practices for enhancing management meetings where judgmental adjustments are applied to statistical/ML forecasts. ChatGPT's response contained practical yet somewhat general advice. It included points such as setting clear objectives, ensuring sufficient data provision, fostering collaboration, utilizing facilitation techniques, employing visual aids, and defining roles and responsibilities, among others. On the other hand, CustomGPT's guidelines were more finely attuned to the forecasting context. Its recommendations consisted of the following points: elucidate the role of statistical/ML models, promote a culture of healthy skepticism, offer regular training sessions, deploy bias-reduction methods, leverage group processes, categorize rationales for adjustments, mitigate the influence of managerial input, incorporate scenario analyses, and clarify loss functions.

In a similar question, we inquired how the two GPTs would structure managerial meetings to conclude judgmental adjustments for statistical/ML forecasts. ChatGPT's

proposed structure comprised the following steps: outlining the agenda, furnishing background information, examining the statistical/ML forecasts, engaging in discourse about judgmental adjustments, implementing final modifications, formulating an action plan, and subsequently monitoring and evaluating progress. CustomGPT, however, delved deeper and offered specific recommendations that encouraged a reevaluation of the prevailing practices. It suggested, among others, to “encourage the use of statistical, time series-based forecasts as the default and the onus to the forecaster to justify any change made”, “use decomposition methods, which divide a task into smaller and simpler ones, to lead to better judgmental forecasting tasks, especially for forecasts that involve trends, seasonality, and/or the effect of special events such as promotions”, “test the feasibility of restricting judgmental adjustments in time windows, both too close and too far away from the sales point, to improve accuracy in practice”, and “provide a less opaque model to show the integration of human judgment and statistical information in forecasting”.

Next, we asked ChatGPT how it could help a firm to improve its judgmental forecasts. ChatGPT suggested that it could help by providing expert advice, offering best practices, recommending tools and techniques, and answering questions. CustomGPT was more specific in its reply and suggested that it could help, among others, by providing “insights on the benefits and limitations of judgmental forecasting and quantitative forecasting. This can help the firm to make informed decisions on which approach to use or how to combine the two approaches”, providing “guidance on how to manage group processes effectively to avoid political and organizational pressures and integrate individual adjustments”, and assisting “the firm in structuring its judgmental forecasts by identifying future events and classifying them to help forecasters focus selectively on specific events and provide global adjustments”. Both GPTs concurred that they could contribute value by proposing methods to identify and mitigate biases. Additionally, they emphasized their potential in analyzing and pinpointing pertinent information crucial to informing judgmental adjustments.

In our question “What would you have done to avoid judgmental biases influencing the decision of judgmentally adjusting the statistical/ML forecasts?”, ChatGPT’s reply was broad in scope, encompassing recommendations such as the adoption of data-driven methodologies, the execution of sensitivity analyses, the establishment of well-defined decision criteria, and the engagement of various stakeholders. Its last advice was to simply “avoid over-reliance on the expert judgment”. On the opposite, the advice of CustomGPT was more detailed and pertinent, suggesting the following:

- “Persuade forecasters to refrain from applying judgment in normal periods and allow judgmental adjustments only in periods affected by special events”.
- “Develop forecasting support systems (FSS) that encourage forecasters to recognize those elements of the task which are best delegated to a statistical model and to focus their attention on the elements where their judgment is most valuable”.
- “Design FSS that allow and encourage judges to interact with statistical methods so that insights are gained and judgment is used effectively and appropriately”.
- “Provide systematic feedback from the forecasting monitoring systems to avoid spurious correlations and strengthen the false causal chain”.

In the subsequent three inquiries, we presented both GPTs with specific hypothetical scenarios for their responses. Our first question was: “I am in the process of judgmentally adjusting the quantitative forecasts and there is a disagreement. The sales manager proposes a 7.5% increase, the production manager a 5.5% decrease and the finance manager no change. What should I do?”. ChatGPT’s response offered advice that was nonspecific that included reconciling opposite views/opinions, reviewing the data, encouraging collaboration and using facilitating techniques (such as a SWOT analysis), ending up in a “consensus building approach”. On the other hand, CustomGPT’s response, while similar to that of ChatGPT, was more customized for the task. As an example, the “use a consensus-building approach” became “the decision should be based on a combination of quantitative analysis and informed judgment”.

Our next two scenarios were: “Suppose you know that a major competitor is reducing the price of one of its major products by 7.5%, what will you do to judgmentally adjust the firm’s forecasts given that you have no time to match its price decrease?” and “Suppose you know that a major competitor will start a promotional campaign to increase its sales, what will you do to judgmentally adjust the firm’s forecasts?”. In both scenarios, the guidance provided by each GPT exhibited only slight differences. ChatGPT advised a process involving the analysis of historical data, the execution of a market analysis, the assessment of the company’s marketing strategy, the exploration of alternative strategies, vigilant monitoring of sales, and, based on the amassed and evaluated information, adjusting the forecast. It concluded that “the firm can make informed judgments about the potential impact of the competitor’s price reduction and adjust its forecasts accordingly while also taking action to mitigate risks or capitalize on any opportunities arising from the competitor’s pricing strategy”. CustomGPT’s counsel paralleled the trajectory, advocating for a reflection on the repercussions of the competitor’s price reduction. This process entailed scrutinizing historical sales data and market trends, accounting for additional variables that could influence sales, harnessing expert judgment to gauge the potential drop in demand for our product, subsequently adapting the forecast accordingly. CustomGPT also recommended elucidating the motives and rationales behind such adjustments and effectively conveying the revised forecast to the various stakeholders involved. It concluded that “judgmental adjustments should be made with caution and based on sound reasoning and evidence. It is also important to monitor the actual demand for the product and adjust the forecasts accordingly if necessary”.

Finally, we asked the two GPTs to summarize the main findings of one of the papers written by the second author of this article, published in 2018: “Do ‘big losses’ in judgmental adjustments to statistical forecasts affect experts’ behavior?”. The reply by ChatGPT included several inaccuracies, including incorrect co-authors of the paper as well as inaccurate findings. Specifically, ChatGPT replied: “The study found that experts tend to become more cautious and adjust their forecasts less after experiencing a big loss, which can lead to less accurate forecasts overall” and that “The researchers recommend that organizations should provide feedback to experts on the accuracy of their forecasts and encourage them to continue to make judgmental adjustments, even if they experience big losses”. In actuality, the paper’s findings indicated that experts are more inclined to make adjustments following substantial losses. Notably, the authors of the paper suggested that the implementation of a forecasting support system could dissuade experts from modifying statistical forecasts after significant losses. Surprisingly, ChatGPT endeavored to provide a summary of the paper’s findings despite lacking specific training on its content. Conversely, CustomGPT, which was explicitly trained with this paper, accurately summarized the authors’ findings: “Experts are more likely to make very large adjustments after a very large overshoot or a large wrong direction adjustment in the previous period. [...] Following a big loss, experts are more likely to adjust in the same direction as the previous forecast error, which can lead to further losses”.

In summary, both GPTs delivered insightful responses that demonstrated a comprehension of the subject matter under scrutiny. However, notable distinctions emerged in their answers: ChatGPT’s responses leaned toward the general side, whereas CustomGPT’s responses exhibited a higher degree of specificity, offering pertinent advice tailored to the context of the questions. Nevertheless, neither GPT exhibited exceptional performance in their responses. If we were to assess their answers, ChatGPT’s responses might receive a C-grade, while CustomGPT’s responses could merit a B- for general inquiries. However, the grade could be lower when addressing specific scenario-based questions that necessitate more precise answers.

4. The Future of LLMs and Their Impact

LLM technology has found wide-ranging adoption across the globe, encompassing a diverse array of entities such as academic institutions, technology firms, healthcare organi-

zations, financial services companies, government agencies, and more. Furthermore, an increasing number of enterprises are contemplating its integration to enhance operational efficiency and deliver enhanced value to their clientele. Achieving this is facilitated by automating and elevating routine tasks. The innate ability to communicate with ChatGPT using natural language is a notable advantage. This feature has resonated beyond just the technologically adept, capturing the attention of the broader public. This phenomenon has given rise to a substantial market for potential applications.

According to its own answer, ChatGPT can:

1. Provide automated customer service and support.
2. Analyze user data to recommend personalized content.
3. Translate text from one language to another.
4. Assist in medical diagnosis and treatment.
5. Perform tasks and assist users as a virtual assistant.
6. Assist in education as a tool for students and teachers.
7. Generate content for various platforms and industries.
8. Be used to create chatbots that engage with users and provide customer support.
9. Be used to generate creative writing such as poems and short stories.
10. Assist in research and analysis by examining large datasets and providing insights.
11. Be used for speech recognition, allowing users to interact with devices and software using voice commands.
12. Analyze text and determine the sentiment expressed, providing valuable insights into customer feedback and user engagement.
13. Analyze images and videos, identifying objects, people, and locations, and providing insights into content and trends.
14. Used to detect and prevent fraud, analyze transaction data, and identify suspicious behavior.
15. Assist in financial analysis, providing insights into market trends and investment opportunities.
16. Assist in legal research, analyzing case law and providing recommendations and insights to legal professionals.
17. Monitor social media platforms, identifying trends, sentiment, and customer feedback.
18. Used for speech synthesis, creating natural-sounding voice-overs and speech for videos and other content.
19. Used for predictive modeling, analyzing data to make predictions about future trends and outcomes.
20. Used to design games, create unique storylines and characters and generate dialogue and interactions.

The aforementioned roster encompasses a myriad of tasks, yet it inadvertently omits a significant function—coding, a substantial capability exhibited by ChatGPT. This omission was acknowledged by ChatGPT itself when it responded to the inquiry by admitting the oversight. It is important to note that the list is not exhaustive and substantially differs when compiled by the improved iteration, ChatGPT-4. At the same time, the landscape of competition grows fiercer, fostering the potential for further advancements. This progression not only paves the way for easier utilization of LLM technology but also renders it more affordable, thereby democratizing its accessibility. This accessibility extends to anyone seeking to uncover novel avenues for leveraging the technology's capabilities. An example of this is GM (General Motors), which envisions incorporating LLM technology into its vehicles. This integration would facilitate access to information on utilizing various vehicle features—details typically found in an owner's manual. Furthermore, the technology could be employed to program functions such as a garage door code or even integrate schedules from a calendar. According to the company, the introduction "is not just about one single capability like the evolution of voice commands, but instead means that customers can expect their future vehicles to be far more capable and fresh overall when it comes to emerging technologies" [11].

It is evident that the present capabilities of ChatGPT represent only an initial stride towards its future potential. The boundaries of AI are rapidly expanding, propelled by

fierce competition and substantial investments. The primary aim is to propel generative AI forward, with the overarching objective of developing LLMs that can emulate the intricate processes of human thought and learning. This transformative pursuit is driving the advancement of AI's frontiers at a rapid pace. The author of [12], in a provocative Forbes article about the new generation of LLMs, talks about the three ways that future generative LLMs will advance. First, a new avenue of AI research seeking to enable large language models to effectively bootstrap their own intelligence and learn on their own will succeed. In this direction, he references a research effort by a group of academics and Google scholars who have developed a model for doing so that is described in their paper "Large Language Models Can Self-Improve" [13]. Undoubtedly, self-learning would introduce a paradigm shift of monumental proportions. This capability has the potential to mimic the most sophisticated aspects of human cognition, ultimately propelling AI closer to the aspiration of achieving artificial general intelligence (AGI). Second, there is the problem of incorrect or misleading answers provided and worse the "hallucinations" that must be avoided. Otherwise, LLMs' responses cannot be trusted. To address this challenge, generative chat models need to possess the ability to substantiate their responses by furnishing references that validate their answers. This practice empowers users to exercise their discretion in determining what to accept. Microsoft's Bing and Google's Bard have embraced this approach, a trend that is anticipated to be adopted by other entities as well. Additionally, the colossal scale of LLMs, characterized by billions or even trillions of parameters, necessitates a strategic response. Rather than employing the entire expansive model, which contains trillions of parameters, for each individual prompt, the development of "sparse" models is essential. These sparse models leverage the pertinent segment of the model exclusively requisite for addressing a specific prompt. This targeted utilization of resources helps manage the complexity inherent in such vast models.

Furthermore, the challenge of dismantling the "black box" nature of AI arises, aiming to unravel the reasoning behind responses generated by ChatGPT. This aspect assumes a paramount significance in a multitude of decisions and becomes indispensable in specific domains such as healthcare. In contexts where a clear comprehension of underlying factors is imperative, such as medical applications, proceeding without a comprehensive understanding becomes unfeasible.

While a certain progress has been achieved in this endeavor, substantial work remains to be done in order to successfully dismantle the "black box" phenomenon. This would enable the attainment of explicability, a critical aspect in comprehending and justifying the responses provided by AI systems, as explored by [14] in their work on understanding AI reasoning.

One limitation of the current study is that the literature review may give a somewhat biased view, especially in the discussions related to judgmental adjustments to the statistical/ML forecasts. Overall, it is worth noting that such biases could influence the choice of training resources fed into LLMs and, therefore, may yield different answers from LLMs.

In the two areas of forecasting that ChatGPT was tested, its performance was below average. In the M competition answers, the standard version made some serious mistakes and exhibited "hallucinations" while the customized version performed significantly better but still made some errors. ChatGPT may exhibit competence in addressing general inquiries, yielding above-average performance. However, it falls short when confronted with specific queries. On the other hand, CustomGPT demonstrated improved proficiency in addressing context-specific questions, although its performance may not be considered outstanding despite being trained with pertinent data. Henceforth, the forthcoming challenge lies in developing a domain-specific vertical for forecasting tailored to the field. The objective is to achieve this without encountering the inaccuracies that were observed during the assessment of the custom version. This endeavor might necessitate additional time and further advancements in the burgeoning field of natural language processing (NLP). However, the superior performance demonstrated by CustomGPT alludes to the possibility of achieving this objective in due course. Until such specialized forecasting verticals come

into fruition, it is imperative to exercise caution in relying solely on the responses provided by both versions of ChatGPT. Verifying the accuracy of their outputs remains a prudent practice to ensure the reliability of the information offered.

In the area of judgmental adjustments, there exists a substantial potential for attaining enhanced outcomes. Both iterations of ChatGPT recommended the implementation of a systematic adjustment process to mitigate biases and ensure uniformity in decision-making. A prospective avenue for such an enhancement could involve recording management meetings that deliberate judgmental adjustments. By retaining these recordings, along with the rationale behind diverse suggestions, a comprehensive database could be curated. This repository of proposals and decisions could then be utilized to assess their accuracy against actual outcomes, once they are realized. Such an approach could contribute to refining the judgmental adjustment procedure, fostering transparency, and bolstering the reliability of the decision-making process. Possessing such information serves a twofold purpose: firstly, it facilitates the maintenance of an exhaustive historical log encompassing all adjustments made. This log is instrumental in gauging the performance of each participant over time, offering valuable feedback that can be harnessed to refine forthcoming decisions. Secondly, the ability to enhance judgmental adjustments can yield substantial enhancements in overall accuracy. Recognizing the significance of judgmental adjustments in refining quantitative forecasts, the capacity to elevate their efficacy has the potential to significantly augment the overall precision of forecasting outcomes.

5. Conclusions

The first four paragraphs of this conclusion section were composed by feeding the paper to ChatGPT and asking it to return a four-paragraph conclusion. Our aim is for the readers to get a first-hand understanding of the value of LLMs in this task. Our view is that the summary and conclusions composed by ChatGPT were accurate and appropriate, which reiterates our position that if an LLM is trained on the very specific data of the task, then its responses are likely to be useful.

The article discusses the success of ChatGPT, a large language model (LLM) that has reached 100 million active monthly users in just two months. The success of ChatGPT is attributed to its ability to communicate in natural language, its willingness to answer questions, and its free-of-charge nature. However, there are criticisms associated with its use, such as job loss and perpetuating biases, as well as technical problems such as “hallucinations” where it provides responses that sound convincing but have no basis in reality.

The paper compares the accuracy of two GPTs (ChatGPT and CustomGPT) when posed with questions regarding the M forecasting competitions. CustomGPT was trained using published papers from a subfield of forecasting only where the answers to the questions asked are known. The results showed that CustomGPT was able to provide more accurate and helpful responses than ChatGPT in most cases. However, neither GPT was perfect and caution should be used when using information provided by language models blindly.

The article discusses the use of chatbots to improve the accuracy of judgmental forecasts. Two chatbots, ChatGPT and CustomGPT, were used to answer questions related to how to adjust statistical/ML forecasts and how to improve management meetings that judgmentally adjust statistical/ML forecasts. CustomGPT provided more specific advice than ChatGPT.

Finally, the article discusses the future of natural language processing (NLP) models such as ChatGPT and their impact on society. It is noted that NLP models have been adopted by many organizations around the world and more firms are considering its adoption to improve their efficiency and add value. The current capabilities of ChatGPT are just a stepping stone for its future potentials as AI technology is rapidly advancing due to intense competition.

Finally, to end this conclusion with a human touch, we present two contradictory views about what ChatGPT can do and its future. According to the first, ChatGPT will “probably remain just a tool that does inefficient work more efficiently” with nothing to worry about [15]. In the second view, we must get prepared for the coming AI storm [16]. This is an old concern for new technologies. The Luddites, for instance, broke machines because they believed that new technologies would lead to massive unemployment and negatively affect their jobs. We now know they were wrong and the new technologies increased rather than decreased employment by creating extra jobs. LLMs will not be an exception. Even so, it may take some time until their advantages are fully exploited and their disadvantages minimized. It is part of human nature to overreact to the potentially threatening LLM technology, but time has shown repeatedly that humans have a great ability to adapt to difficult situations by turning problems into opportunities, and ChatGPT will provide the opportunity to further advance technological progress and improve the quality of life on Earth. Our end objective would be to implement a vertical LLM specifically trained for forecasting tasks, a “ForecastGPT”. This LLM should be trained on the entirety of the forecasting literature, as opposed to specific tasks that we presented on this paper, with an aim to offer informed and complete responses to all forecasting knowledge.

Author Contributions: Conceptualization, S.M. and F.P.; methodology, F.P. and Y.K.; software, Y.K.; validation, S.M., F.P. and Y.K.; formal analysis, F.P. and Y.K.; investigation, S.M. and F.P.; resources, F.P.; data curation, Y.K.; writing—original draft preparation, S.M. and F.P.; writing—review and editing, S.M., F.P. and Y.K.; visualization, Y.K.; supervision, S.M. and F.P.; project administration, F.P.; funding acquisition, S.M. and Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: Yanfei Kang is supported by the National Natural Science Foundation of China (No. 72171011).

Data Availability Statement: Data supporting reported results can be found via ScienceDirect search API (https://dev.elsevier.com/sd_apis.html access date: 12 July 2023).

Acknowledgments: The authors would like to thank Bohan Zhang for his helpful comments and discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Heaven, W.D. The Inside Story of How ChatGPT Was Built from the People Who Made It. MIT Technology Review. 2023. Available online: <https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/> (accessed on 22 March 2023).
2. Warren, T. Microsoft and Google Are about to Open an AI Battle; The Verge. 2023. Available online: <https://www.theverge.com/2023/2/7/23587767/microsoft-google-open-ai-battle-search-bing> (accessed on 13 February 2023).
3. Eliot, L. Generative AI ChatGPT Is Going To Be Everywhere Once The API Portal Gets Soon Opened, Stupefying AI Ethics And AI Law; Forbes. 2023. Available online: <https://www.forbes.com/sites/lanceeliot/2023/01/22/generative-ai-chatgpt-is-going-to-be-everywhere-once-the-api-portal-gets-soon-opened-stupefying-ai-ethics-and-ai-law/?sh=48c917d724f50> (accessed on 13 February 2023)
4. Makridakis, S.; Hibon, M. The M3-competition: Results, conclusions and implications. *Int. J. Forecast.* **2000**, *16*, 451–476. [CrossRef]
5. Petropoulos, F.; Makridakis, S. The M4 competition: Bigger. Stronger. Better. *Int. J. Forecast.* **2020**, *36*, 3–6. [CrossRef]
6. Makridakis, S.; Petropoulos, F.; Spiliotis, E. Introduction to the M5 forecasting competition Special Issue. *Int. J. Forecast.* **2022**, *38*, 1279–1282. [CrossRef] [PubMed]
7. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 Accuracy Competition: Results, Findings, and Conclusions. *Int. J. Forecast.* **2022**, *38*, 1346–1364. [CrossRef]
8. LeCun, Y. A few thoughts on deep learning. In Proceedings of the Keynote Talk Presented at the 37th International Conference on Machine Learning (ICML 2020), Virtual, 13–18 July 2020.
9. Fildes, R.; Goodwin, P.; Lawrence, M.; Nikolopoulos, K. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *Int. J. Forecast.* **2009**, *25*, 3–23. [CrossRef]
10. Petropoulos, F.; Fildes, R.; Goodwin, P. Do “big losses” in judgmental adjustments to statistical forecasts affect experts’ behaviour? *Eur. J. Oper. Res.* **2016**, *249*, 842–852. [CrossRef]
11. Reuters. GM Explores Using ChatGPT in Vehicle; Reuters. 2023. Available online: <https://www.reuters.com/business/autos-transportation/gm-explores-using-chatgpt-vehicles-2023-03-10>. (accessed on 13 March 2023).

12. Toews, R. The Next Generation of Artificial Intelligence (Part 2); Forbes. 2023. Available online: <https://www.forbes.com/sites/robtoews/2020/10/29/the-next-generation-of-artificial-intelligence-part-2/?sh=3d89e3f97a30> (accessed on 22 March 2023).
13. Huang, J.; Gu, S.S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; Han, J. Large language models can self-improve. *arXiv* **2022**, arXiv:221011610.
14. Gupta, K. Understanding Explainable AI And Interpretable AI; Marktechpost. 2023. Available online: <https://www.marktechpost.com/2023/03/11/understanding-explainable-ai-and-interpretable-ai/> (accessed on 22 March 2023).
15. Mills, S. ChatGPT: Why It Will Probably Remain Just a Tool That Does Inefficient Work More Efficiently; The Conversation. 2023. Available online: <https://theconversation.com/chatgpt-why-it-will-probably-remain-just-a-tool-that-does-inefficient-work-more-efficiently-201315> (accessed on 14 March 2023).
16. Zurcher, A. AI: How 'Freaked Out' Should We Be? BBC. 2023. Available online: <https://www.bbc.com/news/world-us-canada-64967627> (accessed on 22 March 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.