
AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy

Philipp Schoenegger

London School of Economics and Political Science

Peter S. Park

Massachusetts Institute of Technology

Ezra Karger

Federal Reserve Bank of Chicago*

Sean Trott

University of California San Diego

Philip E. Tetlock

University of Pennsylvania

Abstract

Large language models (LLMs) match and sometimes exceeding human performance in many domains. This study explores the potential of LLMs to augment human judgement in a forecasting task. We evaluate the effect on human forecasters of two LLM assistants: one designed to provide high-quality ('superforecasting') advice, and the other designed to be overconfident and base-rate neglecting, thus providing noisy forecasting advice. We compare participants using these assistants to a control group that received a less advanced model that did not provide numerical predictions or engaged in explicit discussion of predictions. Participants ($N = 991$) answered a set of six forecasting questions and had the option to consult their assigned LLM assistant throughout. Our preregistered analyses show that interacting with each of our frontier LLM assistants significantly enhances prediction accuracy by between 24% and 28% compared to the control group. Exploratory analyses showed a pronounced outlier effect in one forecasting item, without which we find that the superforecasting assistant increased accuracy by 41%, compared with 29% for the noisy assistant. We further examine whether LLM forecasting augmentation disproportionately benefits less skilled forecasters, degrades the wisdom-of-the-crowd by reducing prediction diversity, or varies in effectiveness with question difficulty. Our data do not consistently support these hypotheses. Our results suggest that access to a frontier LLM assistant, even a noisy one, can be a helpful decision aid in cognitively demanding tasks compared to a less powerful model that does not provide specific forecasting advice. However, the effects of outliers suggest that further research into the robustness of this pattern is needed.

*Any views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1 Introduction

Recent advances in artificial intelligence (AI), and large language models (LLMs) specifically, demonstrate impressive AI capabilities across a large number of complex and economically valuable tasks (Naveed et al. 2023). This development challenges previously held beliefs about the necessity of human cognition for many of these tasks (Bubeck et al. 2023), and raises the possibility of significant negative effects of AI systems on the (human) labor market in large parts of the knowledge economy (George and Baskar 2023). Understanding the current ability of LLMs to interface with economically central tasks requires a broad empirical study across domains. However, most knowledge-work jobs require substantial reasoning capabilities that use data and patterns of observations beyond any model’s training data. This makes finding a suitable study context central in any attempt to understand how LLMs might impact advanced economies in the near future.

Our focus in this paper is on Large Language Models, which represent a significant advance in AI and natural language processing. These models build upon the transformer architectural paradigm (Vaswani et al. 2017) and are characterized by their massive scale, often containing billions or even trillions of parameters, trained on an enormous amount of diverse textual data (Shen et al. 2023). The core capability of LLMs is next-token prediction: the ability to predict the most probable next word or subword (token) given a sequence of preceding tokens. However, this seemingly simple objective, when scaled, results in a wide array of emergent abilities that seem to extend far beyond basic next-token prediction. These advanced AI systems demonstrate proficiency in tasks such as natural language understanding and generation, few-shot learning, and complex reasoning across various domains. Importantly, many of these specialized advanced skills emerge in ways that could not have been fully predicted before training, due to non-linearities in how capabilities scale with model size and data (Wei et al. 2022).

Some areas where LLMs have shown strong performance are marketing (Fraïwan and Khasawneh 2023), translation (Jiao et al. 2023), high levels of reading comprehension (Winter 2023), teaching (Fraïwan and Khasawneh 2023; Sallam et al. 2023), summarization (Goyal, Li, and Durrett 2023), abstract categorization of objects (Atari et al. 2023), programming (Bubeck et al. 2023; Cheng et al. 2024), spear phishing cyber attacks (Hazell 2023; Heiding et al. 2023), human personality (Schoenegger et al. 2024a), robotics (Vemprala et al. 2023), medical reasoning (Bubeck et al. 2023; Nori et al. 2023; Sallam et al. 2023), legal reasoning (Bubeck et al. 2023; Katz et al. 2023), deception (Park et al. 2023), and others. LLMs’ many capabilities substantially increase the amount of money and talent going into LLM research and products (Sutton 2023), suggesting further growth in capabilities in the near future.

Crucially, modern state-of-the-art or frontier language models are not inherently autonomous for most relevant tasks (Xi et al. 2023). While they can be imbued with general autonomy through agent frameworks like AutoGPT (Firat and Kuleli 2023) or other scaffolding approaches, the reliability of such methods remains questionable. Future iterations of models may enable autonomous behavior directly (Kinniment et al. 2023), potentially making agency—the ability to take actions and achieve goals independently—more accessible. However, at present, LLMs are not economically viable as autonomous agents due to significant limitations including inefficiency, forgetting, speed, cost, cultural complexity (McIntosh et al. 2024) and hallucinations (Firat and Kuleli 2023).

Instead, these models are primarily used in combination with human labor, forming a hybrid technology that necessitates human input at various stages (Dell’Acqua et al. 2023). This synergistic approach allows humans to leverage the strengths of LLMs, producing outcomes that can surpass what either humans or machines could achieve independently. For instance, LLM augmentations have demonstrably enhanced the performance of human graders (Xiao et al. 2024) and programmers (Peng et al. 2023), and have also been applied in the context of co-creating visual stories (Antony and Huang 2023), illustrating the potential of human-AI collaboration in diverse fields.

Our study contributes to the growing research on human-AI collaboration in complex decision-making tasks, a key focus in HCI and AI research (Steyvers and Kumar 2023). By examining LLM-augmented forecasting, we extend recent work on human-AI interaction modes (Gao et al. 2024) and address challenges in AI-assisted decision-making (Steyvers and Kumar 2023). Our approach aligns with calls to develop nuanced understandings of human-AI complementarity (Yang 2024) and explores how different LLM prompts affect forecasting outcomes, contributing to discussions on designing AI systems that effectively augment human cognition (Wang et al. 2024a).

In this paper, we study the application of present-era frontier LLMs as a hybrid augmentation technology in the context of forecasting future events. This allows us to test their ability to augment human decision-making in a domain robust to in-sample overfitting of training data, since no one, including LLMs or the experimenters themselves, can know the answer to prospective forecasting questions at the time of data collection. This context is also practically relevant as accurate forecasting is essential to many aspects of economic activity, especially within white-collar occupational domains such as law, business, and policy: fields that may be disrupted by LLM capabilities (Acemoğlu 2023; Park and Tegmark 2023; Summers and Rattner 2023). If the use of present or future AI systems increases the forecasting accuracy of humans and organizations, the efficiency and productivity

gains to the relevant industries’ individuals and businesses are clear, and if there are risks, they ought to be discussed prior to widespread adoption.

Our specific object of interest in this study is human judgment forecasting, where humans provide forecasts of future events, such as the probability that inflation will hit a certain milestone over the next twelve months or the anticipated number of barrels in the Strategic Petroleum Reserve at the end of the year. This context is distinct from the more widely studied topic of time series forecasting (Jin et al. 2023), as the central input are judgements by human forecasters as opposed to machine learning algorithms. The science of judgemental forecasting has found that aggregated forecasts of a crowd of forecasters can be surprisingly accurate (Tetlock and Gardner 2016), can impact policy debates (Tetlock, Mellers, and Scoblic 2017), and can affect businesses (Schoemaker and Tetlock 2016), and that much of this effect is derived from the high accuracy of a subset of forecasters, often called ‘superforecasters’. Previous work on the topic focuses on a variety of other topics, ranging from the identification of these highly skilled forecasters (Himmelstein, Budescu, and Han 2023; Mellers et al. 2015b; Tetlock and Gardner 2016) and novel aggregation methods (Atanasov et al. 2017) to improvements of forecasting accuracy (Chang et al. 2016; Karger, Atanasov, and Tetlock 2022) as well as applications to topics like development economics (Bernard and Schoenegger 2024) or pandemics (McAndrew et al. 2024).

Related to our project, some previous work focuses on human-machine hybrid forecasting in the context of IARPA’s ‘Hybrid Forecasting Competition.’ Benjamin et al. (2023) report the results of ‘SAGE,’ a hybrid forecasting system designed to combine human- and machine-generated forecasts (such as autoregressive integrated moving average (ARIMA) forecast outputs). They find that their hybrid forecasting system outperformed their human-only baseline, suggesting that cost savings and accuracy increases of these hybrid systems may be “a viable approach for maintaining a competitive level of accuracy” (Benjamin et al. 2023, p. 113). Similarly, Atanasov et al. (2017) introduce a ‘Human Forest’ method that enables human forecasters to define custom reference classes, draw on historical databases, and review base rates in their forecasting. They find that these forecasters outperform statistical model predictions. However, both approaches used pre-LLM methods as their machine counterparts. Unlike these systems, LLM-based assistants allow for new systems where humans and models communicate interactively in dynamic settings.

In this paper, we extend this literature on human-AI interactions in light of recent breakthroughs in LLMs. The central advancement for this context is the possibility of a free-flowing exchange between the human and the model via a chat function, in which the human can query the model, receive a response that is often indistinguishable from a human response (Jones and Bergen 2024), and then continue the conversation, with the previous iteration being part of the model’s memory. This back-and-forth on advanced topics necessitating strong model reasoning capabilities is something that previous technologies were not capable of, and is a potential way for humans to learn skills in their interaction with AI systems (Yang 2024). Those interacting with the model can query it to fill their own gaps in knowledge or perceived weaknesses, they can ask it to produce a full forecast for them and provide the reasoning underlying it, they can input their own reasoning and predictions into the model for feedback, or they can do a combination of these and other approaches they might find helpful (Wang et al. 2024a). This is similar to work by Guo et al. (2024) that provides a natural language interface for questions of tabular data. While the technology still has substantial limitations, the fact that forecasters can engage with it in an interactive and personalised way opens up a novel type of human-machine interaction. Our goal in this paper is to probe whether LLM forecasting augmentations with advanced prompts can be a cheap, scalable, and effective method of improving human judgement forecasting. Inference costs for LLMs remain low and continue to drop, sitting currently at less than a cent per 1000 tokens, making LLM forecasting augmentation a prime candidate for a generalized hybrid system that can boost individual performance in many valuable tasks at costs far below a human assistant equivalent.

Current best-practice measures of LLM proficiency often rely on task benchmarks, where models are evaluated against a set of predefined tasks. We argue that evaluating forecast accuracy in real-world scenarios like actual forecasting presents a more comprehensive assessment of reasoning capabilities and reduces risks of overstating model capabilities due to training data memorization. This also increases the likelihood that these results generalize to different—and perhaps out-of-distribution—settings (Arora and Goyal 2023). As such, our approach diverges from conventional task benchmarks, focusing on the LLM’s ability to apply its knowledge and understanding to novel settings, rather than settings that may be represented in some shape or form in its training data or output that may have been training on to perform well on benchmarks. Even if an LLM excels at a given task benchmark, it is unclear whether this reveals a deep understanding of the process behind the task, instead of rote memorization of the task benchmark’s answers in the training data (Bender et al. 2021; Biderman et al. 2023; Carlini et al. 2023; Magar and Schwartz 2022). The difficulty in disentangling true understanding from training data memorization is non-trivial. Deep understanding, after all, also originates from exposure to relevant content within the training dataset. However, the success or failure to generalize outside of the training data appears central to this disentangling (Grove and Bretz 2012). In our study, we analyze human forecasting behavior on a set of prediction questions that resolve in the future such that no human forecaster or AI-based system can access the answer at the time of data collection, avoiding these concerns.

Past work found that the at-the-time frontier model GPT-4, released by OpenAI in March 2023, significantly underperformed the median human-crowd forecast in a real-world forecasting tournament, failing to even