
Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy

Philipp Schoenegger
London School of Economics
and Political Science

Indre Tuminauskaite
Independent Researcher

Peter S. Park
MIT

Philip E. Tetlock
University of Pennsylvania

Abstract

Human forecasting accuracy in practice relies on the ‘wisdom of the crowd’ effect, in which predictions about future events are significantly improved by aggregating across a crowd of individual forecasters. Past work on the forecasting ability of large language models (LLMs) suggests that frontier LLMs, as individual forecasters, underperform compared to the gold standard of a human crowd forecasting tournament aggregate. In Study 1, we expand this research by using an LLM ensemble approach consisting of a crowd of twelve LLMs. We compare the aggregated LLM predictions on 31 binary questions to that of a crowd of 925 human forecasters from a three-month forecasting tournament. Our preregistered main analysis shows that the LLM crowd outperforms a simple no-information benchmark and is not statistically different from the human crowd. In exploratory analyses, we find that these two approaches are equivalent with respect to medium-effect-size equivalence bounds. We also observe an acquiescence effect, with mean model predictions being significantly above 50%, despite an almost even split of positive and negative resolutions. Moreover, in Study 2, we test whether LLM predictions (of GPT-4 and Claude 2) can be improved by drawing on human cognitive output. We find that both models’ forecasting accuracy benefits from exposure to the median human prediction as information, improving accuracy by between 17% and 28%: though this leads to less accurate predictions than simply averaging human and machine forecasts. Our results suggest that LLMs can achieve forecasting accuracy rivaling that of human crowd forecasting tournaments: via the simple, practically applicable method of forecast aggregation. This replicates the ‘wisdom of the crowd’ effect for LLMs, and opens up their use for a variety of applications throughout society.

1 Introduction

In the field of artificial intelligence (AI), the rapidly increasing capabilities of large language models (LLMs) have shown promise and even market-competitiveness in a rapidly increasing number of economically valuable and cognitively demanding tasks (Naveed et al. 2023; Sutton 2023). State-of-the-art LLMs with billions of parameters, built on the Transformer architecture (Vaswani et al. 2017), are trained on a very large amount of internet text data (Shen et al. 2023b), before being fine-tuned. The LLMs are trained on this data to predict the next word or subword (token) when given an input string. This step of next-token prediction—when applied repeatedly—generates a sequence of tokens that form an output string coherently text-completing the input, often at a level of coherence previously thought to be only achievable by human cognition (Anthropic 2023; Gemini Team et al. 2023; OpenAI et al. 2023; Touvron et al. 2023) and at a high level of applicability to chat interfaces and various other settings.

This general training objective of next-token prediction, coupled with fine-tuning, also indirectly results in these LLMs displaying an array of specialized skills, which are often only emergently observed after the fact: in ways that were not—and for all practical purposes, likely could not have been—predicted before the first observation of the given capability (Wei et al. 2022). Such skills include but are not limited to marketing (Fraiwan and Khasawneh 2023), reading comprehension (Winter 2023), teaching (Fraiwan and Khasawneh 2023; Sallam et al. 2023), abstract object classification (Atari et al. 2023), cyberattacks (Heiding et al. 2023), robotics (Vemprala et al. 2023), social-science applications (Abdurahman et al. 2023; Park, Schoenegger, and Zhu 2024), medical analysis (Bubeck et al. 2023; Nori et al. 2023; Sallam et al. 2023), legal analysis (Bubeck et al. 2023; Katz et al. 2023), deception (Park et al. 2023), surgical knowledge (Beaulieu-Jones et al. 2024), and computer graphics assessment (Feng et al. 2024).

When evaluating the capabilities of a given AI system, the predominant traditional method is to measure how well an AI system performs at fixed benchmarks for specific tasks (Kistowski et al. 2015). The significant advancements achieved by transformer-based LLMs in these domains have rendered many previously established benchmarks obsolete (Laskar et al. 2023; Shen et al. 2023a), moving the metaphorical goalposts forward in the form of more challenging and comprehensive benchmarks (Alzahrani et al. 2024). It is plausible that a significant portion of the unprecedented successes that state-of-the-art LLMs have achieved on past task benchmarks is genuinely due to a deep understanding of the task-relevant cognitive skills achieved by the LLMs (Bubeck et al. 2023). Indeed, this argument is corroborated by the economic competitiveness—and even promises of economic superiority—that LLMs are achieving for an increasing array of human occupations (Sutton 2023), such as transcription (Peng et al. 2023), translation (Jiao et al. 2023), and programming (Bubeck et al. 2023).

However, it is also plausible that a significant portion of these successes on task benchmarks is due to a superficial memorization of the task’s solutions: and shallow understanding of training-set patterns in general (Bender et al. 2021; Biderman et al. 2023; Carlini et al. 2023; Magar and Schwartz 2022). Distinguishing between deep understanding and shallow memorization is a complex challenge, and is central to accurate assessments of advanced reasoning capabilities in AI. This is akin to the examiner’s problem of testing their student for deep understanding of the course material, even when many of the potential exam questions can be correctly answered by shallow memorization instead. In fact, just like the student can memorize the answers to exam questions if they see it beforehand, so too can an LLM if its training data contain the questions and answers used in the task benchmark. To resolve this ambiguity, one can exploit the testable presence or absence of the ability to generalize out-of-distribution: to apply learned knowledge beyond the settings represented in the training data (Arora and Goyal 2023). Such a test is arguably key to discerning deep understanding of the task at hand (Grove and Bretz 2012), but is difficult to design when aiming to assess broad LLM capabilities.

In contrast to task benchmarks, where questions and answers are fixed and potentially contained in an LLM’s training data, there are contexts where this concern can be ruled out fully: for example, when predicting the future in real-world settings (Schoenegger and Park 2023; Schoenegger et al. 2024). This test stands out for its high external validity, in that the correct answer to a given real-world forecasting question cannot be in a given LLM’s training set, as not even the human evaluator knows the answer at the time of evaluation. Moreover, the practice of forecasting is omnipresent in the cognitive tasks undertaken by humans, encompassing a wide range of applications from forecasting the trajectory of current events to setting long-term plans. The ubiquity of forecasting—especially

in white-collar occupations where the increasing capabilities of LLMs are predicted to disrupt or even replace human professionals (Acemoğlu 2023; Park and Tegmark 2023; Summers and Rattner 2023)—combined with the intrinsic external validity makes testing the forecasting capabilities of AI systems an ideal test for assessing the real-world applicability of LLMs.

One context where this can be tested directly are forecasting tournaments. These tournaments involve participants who make probabilistic predictions about future occurrences and are then evaluated and rewarded for their accuracy (Tetlock et al. 2014). Across a set of questions, prediction accuracy of these forecasts determines the reputational or monetary rewards, with more precise predictions yielding greater rewards, incentivising forecasters to research the questions and to provide well-informed predictions. Based on the predictions of a crowd of forecasters, their aggregate is a gold-standard for human intelligence gathering. This effectiveness of the aggregate of competitive forecasting endeavors relies on the ‘wisdom of the crowd’ phenomenon, which is the effect that results in the collective accuracy of a set of predictions often surpasses the vast majority of individual judgments that make up the respective crowd. This concept is supported by extensive research across various fields such as prediction markets (Bassamboo, Cui, and Moreno 2018), political forecasting, and more, showing that the combined forecasts of many individuals tend to be remarkably precise (Da and Huang 2020; Lichtendahl Jr, Grushka-Cockayne, and Pfeifer 2013; Surowiecki 2004). This ‘wisdom of the crowd’ effect relies on independent and unbiased judgements, which achieves an error-cancellation effect (Budescu and Chen 2015) and thereby causes the aggregate to outperform randomly selected forecasts from parts of that crowd (Davis-Stober et al. 2014). As Budescu points out, this aggregation mechanism increases information and accounts for extremes (Budescu 2006), with the ‘wisdom of the crowd’ effect also holding in contexts of biased inputs (Koriat 2012) or when there are correlations among judgements (Davis-Stober et al. 2014), showing remarkable robustness. Moreover, there is a large literature on improvements of this aggregation process (Baron et al. 2014; Himmelstein, Budescu, and Han 2023; Himmelstein, Budescu, and Ho 2023), with a central take-away being that a simple median is a surprisingly powerful aggregation mechanism across contexts.

Past work has compared the prediction performance of frontier models against a human crowd. With respect to evaluating a single model, Schoenegger and Park (2023) found that the frontier model GPT-4 performed poorly when comparing its predictions to that of a crowd drawn from a forecasting tournament. In fact, GPT-4 did not even significantly outperform the no-information benchmark strategy of predicting 50% on every question. Also, the work of Halawi et al. (2024) has investigated the prediction capabilities of an LLM system, including a combination of news retrieval and reasoning systems. They replicated the finding of Schoenegger and Park (2023) that individual models show poor prediction accuracy, but also found that their optimised system approach aggregated human accuracy. This suggests that individual LLMs may have poor forecasting accuracy, but can produce accurate predictions if they are set in an advanced system.

A hypothesis worth probing is that the underperformance of individual LLMs in real-time forecasting may be, at least in part, due to not making use of the ‘wisdom of the crowd’ effect. It is reasonable that LLM forecast accuracy may be enhanced by aggregation, as crowd aggregates are known to result in better predictions even high-performing individuals. To test whether such a ‘wisdom of the silicon crowd’ effect exists, we simulate a crowd of diverse LLMs and draw questions from a real-world forecasting tournament, directly comparing the LLM crowd estimate to that of the human crowd, without introducing further additions like retrieval systems.

In Study 1, we test this LLM ensemble approach, aggregating twelve LLMs’ forecasts into a collective crowd forecast, leveraging the diversity inherent in the different models’ training data, parameters, and methodologies (such as idiosyncratic fine-tuning). We test whether this diversity improves machine forecast accuracy by reducing the impact of individual model biases and errors. We first test whether the LLM ensemble, unlike GPT-4 in the study of Schoenegger and Park (2023), will significantly outperform the no-information benchmark in a forecasting tournament. This benchmark provides a minimal benchmark of accuracy that is equivalent to guessing 50% on every question.

Null hypothesis 1, Study 1: The average of median LLM forecasts is neither statistically significantly more nor less accurate than the 50% baseline, $H_{01} : \bar{B}_{LLM} = 0.25$.

We also conduct the stronger test of whether the LLM ensemble will significantly outperform the human crowd drawn from the real-world forecasting tournament. For both studies, we use a three-month tournament run on the platform Metaculus as our human crowd comparison. This provides a