Who is the better forecaster: humans or generative AI?

The ability to forecast and predict future events with a degree of accuracy is central to many professional occupations. Utilising a prediction competition between human and AI forecasters, **Philipp Schoenegger** *and* **Peter S. Park**, *assess their relative accuracy and draw out implications for future AI-society relations.*

OpenAI's GPT-4, on which the well-known ChatGPT chat model is based, is an example of a large language model (LLM). These are AI systems that are made up of an extremely large number of parameters and trained on a massive corpus of text data. LLMs have surprised us with their ostensible capabilities in many economically relevant tasks previously thought to require human cognition. Tasks at which LLMs excel include reading comprehension, summarisation, coding, translation, deception, and even the bar exam.

But there is a caveat. LLMs might be great at task benchmarks like these simply because the benchmarks' questions and the corresponding answers are present in their training data. This is analogous to a student acing an exam by memorising answers from past papers they have seen, rather than by a deep understanding of the task at hand.

For this reason, we decided to put OpenAI's GPT-4 to the test in a different kind of exam: one where the answers are genuinely unknown at the time of testing, even to the human evaluating the answers. The setting? A forecasting tournament. The logic is straightforward: if GPT-4 can predict future events, it demonstrates a deeper understanding of how the world works beyond just regurgitating memorised data. This makes real-world forecasting tournaments an ideal testbed for evaluating the generalised reasoning and prediction capabilities of AI systems.

> if GPT-4 can predict future events, it demonstrates a deeper understanding of how the world works beyond just regurgitating memorised data

The stage for our evaluation of GPT-4's predictive capabilities was the Metaculus platform, where over the course of three months, GPT-4 and 843 human participants tried their hand at predicting various future events. The topics were diverse, spanning Big Tech, U.S. politics, viral outbreaks, and the Ukraine conflict. Comparing GPT-4 to human-crowd forecasting tournaments raises the bar quite high, as such tournaments have remarkable accuracy in predicting future events.

The results of our study? In simple terms, GPT-4 was no Nostradamus. Not only did it underperform compared to the human crowd's median predictions, but its forecasts were also indistinguishable from just guessing 50% for every question. This suggests that while GPT-4 might be an intellectual heavyweight in many areas, when it comes to looking into the crystal ball, humans still have the upper hand for now.

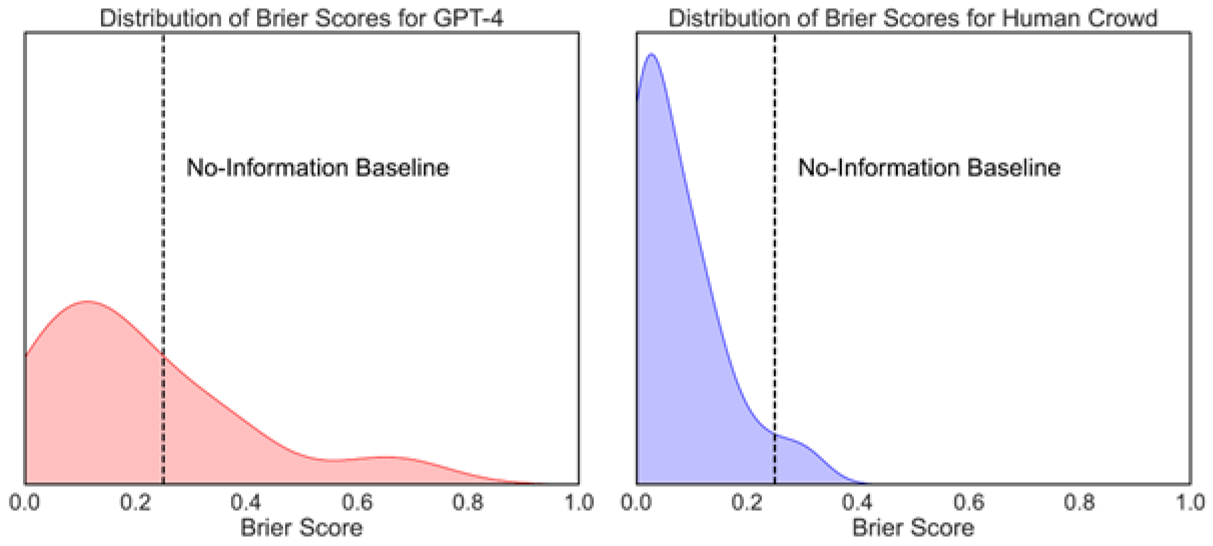

Fig.1: This figure shows accuracy for both GPT-4 and the human crowd. Accuracy is calculated via the Brier score, a metric that quantifies the difference between predicted probabilities and actual outcome, where '0' indicates perfect accuracy and '1' indicates maximal inaccuracy. The left panel shows GPT-4's forecasting accuracy while the right panel shows the human crowd accuracy. The black dotted line is the no-information benchmark of assigning 50% to every question.

One possible reason for GPT-4's subpar performance is its inability to keep up with real-time information. While human forecasters can adjust their predictions based on new information and current events, GPT-4's knowledge has a fixed cut-off point. Even with background information fed into the model, it can't fully account for the dynamic, evolving nature of world events.

In simple terms, GPT-4 was no Nostradamus.

But what do these results mean for the broader field of AI and society at large? One implication is that our results cast doubt on the immediate prospects of AI taking over jobs that rely on predictive decision-making. OpenAI's mission is to create "highly autonomous systems that outperform humans at most economically valuable work." Whether this mission is on track to occur will be largely determined by how capable LLMs—and AI systems in general—turn out to be at economically relevant tasks. Predicting the future is a task of especially high economic relevance, especially given that many white-collar jobs in business, policy, and law rely on the ability to make accurate predictions in various domains. Our results suggest that even state-of-the-art AI systems would not yet be competitive with human expertise in occupations that heavily rely on accurate future predictions.

Another implication of our results pertains to the threat posed by AI systems that are proficient at long-term planning. An AI system that excels at long-term planning would be able to pursue its own goals. This can be concerning if the system's goal happens to be incompatible with the well-being of humans. For example, consider the goal of engineering a pandemic that kills as many people as possible over the long run. An AI system with such a goal would be much more dangerous if it excels at planning. The threat posed by AI systems that can create effective long-term plans in pursuit of their goals is especially concerning, given that AI systems can manifest new goals never intended by their human developers.

In order to make effective long-term plans, it is crucial to accurately forecast and plan for future scenarios. Our finding that GPT-4 has particularly poor forecasting capabilities bolsters the case that the threat of an AI system planning in the long term against human interests remains, thankfully, quite low.

That said, the field of AI is constantly evolving. Today's shortcomings might become tomorrow's breakthroughs. Future research might look into creating ensembles of LLM forecasters, or perhaps into designing models that can actively access and learn from the internet. Another potential research direction is hybrid forecasting models, which aims to combine the comparative strengths of both humans and AI systems and is more akin to most currently deployed AI solutions.

It is important to note that while human forecasters have won this forecasting duel, this should not be interpreted as a reason for complacency. The speed at which AI capabilities have been advancing so far suggests that it remains vital to constantly monitor AI capabilities: including, but not limited to how their forecasting capabilities evolve over time. Such a forward-thinking approach can help us ensure that the development of these AI systems will be a boon, rather than a bane, to us humans.

This post draws on the authors' preprint, Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament, published on arXiv.

The content generated on this blog is for information purposes only. This Article gives the views and opinions of the authors and does not reflect the views and opinions of the Impact of Social Science blog (the blog), nor of the London School of Economics and Political Science. Please review our comments policy if you have any concerns on posting a comment below.

Image Credit: Anton Vierietin on Shutterstock.