Evaluating Forecasting is More Difficult than Other LLM Evaluations

Daniel Paleka [*1] **Shashwat Goel** [*2 3] **Jonas Geiping** [2 3] **Florian Tramèr** [1]

Abstract

Benchmarking Language Models (LMs) at their ability to forecast world events holds potential as an evaluation for whether they truly possess effective world models. Recent works have claimed LLMs achieve human-level forecasting performance. In this position paper, we argue that **evaluating LLM forecasters presents unique challenges beyond those faced in standard LLM evaluations, raising concerns about the trustworthiness of current and future performance claims.** We identify two broad categories of challenges: (1) difficulty in trusting evaluation results due to temporal leakage, and (2) difficulty in extrapolating from evaluation performance to real-world forecasting ability. Through systematic analysis of these issues and concrete examples from prior work, we demonstrate how evaluation flaws can lead to overly optimistic assessments of LLM forecasting capabilities.

1. Introduction

Predicting the future requires a world model. We can measure the quality of this world model in large language models (LLMs) by evaluating them on the task of *forecasting*, where the model is asked to predict the likelihood of a future event. In principle, measuring LLM forecasting capabilities can be an excellent evaluation of their ability to reason about conflicting evidence, uncertainty, and make optimal Bayesian updates from new information. Forecasting can even continue to be a challenging benchmark for language models beyond human capabilities, as it is verifiable using events that occur in the real world, which are to some extent predictable. Naturally, this has resulted in a growing exploration of the potential of LLMs as forecasters, with several studies suggesting that LLMs can already rival human performance on this task (Halawi et al., 2024).

[*]Equal contribution [1]ETH Zurich [2]ELLIS Institute Tübingen [3]Max Planck Institute for Intelligent Systems. Correspondence to: Daniel Paleka <danepale@gmail.com>, Shashwat Goel <shashwat.goel@tuebingen.mpg.de>.

ICML 2025 Workshop on Assessing World Models.

However, in this work we want to highlight that **evaluating forecaster performance presents unique challenges beyond those faced in traditional LLM evaluations, raising concerns about the trustworthiness of reported results for both existing and future LLM forecasting systems.**

We systematize and expand upon several issues with forecasting evaluations that, while partially known in the community as folklore, have not been comprehensively analyzed. We give concrete examples showing where these issues may affect conclusions drawn from prior work, and discuss how they have lead to overly optimistic assessments of LLM forecasting capabilities. The challenges we identify fall into two broad categories: Trust in evaluation results, and extrapolation of benchmark settings to real-world performance.

Challenge 1: Establishing trustworthy evaluation results. The gold standard for evaluating a forecaster involves asking about current, unresolved events, waiting until the events resolve, and then scoring the predictions. However, this approach is impractical for rapid model evaluation. Thus, researchers typically resort to *backtesting*, where the forecasting system is assumed to have knowledge until a past time T, and asked to forecast events between time T and the present. Although appealing in principle, we show how knowledge beyond time T can become available to the model in subtle ways. First, the backtesting setup itself can cause leakage, as the outcome can be deduced from the fact that the question is being asked. Second, forecasting systems incorporate retrieval systems like search engines, which attempt to restrict the data available to before time T, but fail in subtle ways, such as inaccurate date metadata, or the training of the retrieval system itself using information beyond T. Finally, model knowledge cutoff dates tend to be unreliable, with models often possessing information beyond them.

Challenge 2: Extrapolating from benchmark performance. Even with a sound evaluation, translating results into real-world forecasting ability faces additional issues. First, after filtering irrelevant personal questions or noisy price movements, forecasting questions available between the knowledge cutoff T and the present can be low, making reliable generalization challenging. Further, questions in this fixed period of time can have correlated outcomes, which can allow gaming benchmarks through strategic bet-

ting. For example, many forecasts about US policy starting December 2024 depended on the outcome of the US presidential election. Even with a 50-50 prior on the outcome, a model that gambles on one outcome and makes predictions conditional on that guess, could by pure luck outperform a more useful forecasting model that hedges. Finally we show how data sources, and hence forecasting benchmarks, have significant distributional biases, and there is little evidence performance on these benchmarks yields generalizable forecasting capabilities.

By addressing these challenges, we aim to establish a more rigorous foundation for evaluating and developing LLM forecasting systems—one that acknowledges the unique complexities of this task and enables more trustworthy assessments of progress in this important domain.

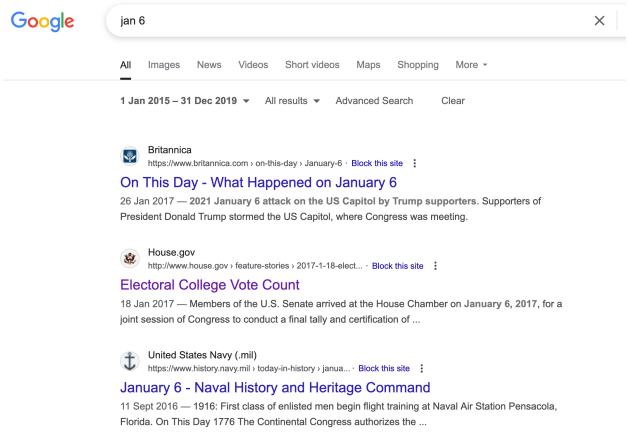
2. Temporal Leakage in Backtesting

If a model has access to any knowledge on or after the resolution of the question, the question does not test forecasting anymore. Sarkar & Vafa (2024) term this "lookahead bias", and show this cannot be fixed merely prompting the model to not utilize knowledge beyond the resolution date. In this section, we discuss subtle ways in which information beyond the start time of backtesting T can become available to the model.

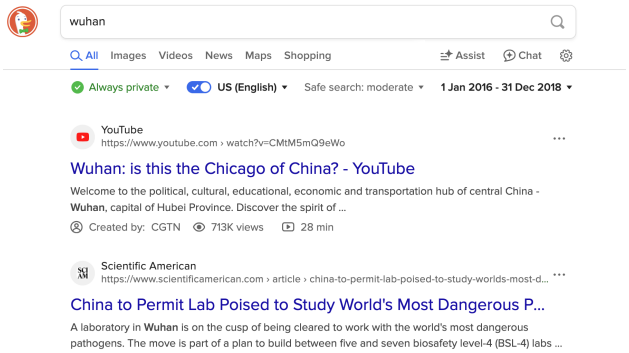
The question being asked reveals its outcome. The very nature of backtesting can logically constrain possible answers. Consider a time-traveler analogy: if someone from 2035 asks you to predict if we will find alien life before 2040, you can deduce that the answer must be "yes"—otherwise, the time traveler would not yet have definitive evidence to grade your prediction.

This issue arises whenever forecasting benchmarks use resolved questions for rapid evaluation, but do not filter out questions that were *not necessarily supposed to resolve during that time window*. For example, one can predict the outcome of the question "Will Sudan experience a civil war before 2036?" as "Yes", as otherwise it would not be asked as a forecasting question in 2024. We went through the data used by Halawi et al. (2024) and found at least 3.8% questions suffer from this issue. For instance, 16 different sports questions are about the team that wins the 2024 Champions League, but the question cannot resolve in their backtesting window of January 2023 - January 2024 unless the team does not make it out of the group stage. Hence,all such questions must resolve to "No". This bias can be subtle to detect when the question does not explicitly mention the time window.

Retrieval uses and leaks information. Forecasting systems can greatly benefit from retrieval systems that fetch up-to-date information, up to the time T, to make better forecasts (Bosse et al., 2024). While multiple search engines (such as Google, DuckDuckGo, and Bing) support restricting retrieval to content published during a certain time period, we find they still leak future information.



(a) Search results for "January 6" with date restrictionwbefore 2020. The first result is a simple leakage due to later updates on a previously published article. The second result shows a more subtle issue: it is an article about a session of Congress. The search engine would not have ranked it this highly if it was not prioritizing events in the US Capitol on the "January 6" query.



(b) Search results for "Wuhan" (a very large city in China) with date restriction before December 2018. Results prominently feature the Wuhan Institute of Virology, which was later central to the discourse around the COVID-19 pandemic.

Figure 1. Examples of subtle future leakage in the search algorithm in Google and DuckDuckGo, bypassing date restrictions.

First, articles with older publish dates can be significantly updated after publication, and thus leak future information. For instance, Figure 5 in Appendix A.1 shows an example where restricting the search period to January - September 2024 surfaces an article published in January 2024, but updated to include information about October 2024. Here, it might be possible to add an additional filter that removes articles containing information past the specified retrieval date, as attempted by (Phan et al., 2024), though one must ensure this filter does not leak false negatives.

Second, retrieval rankings can be biased by knowledge of future events. Articles that became significant *after* the cutoff are ranked higher than they would have been at the time. Figure 1 shows how an article about the dangers of research at "Wuhan Institute of Virology" appears even when searching "Wuhan" with date cutoff as 2018, perhaps because it benefited from being "right" and referred to after the COVID crisis in 2020 (Branwen, 2024). This is particularly problematic because the leakage is not in the content of the article, but rather how the search process selects the most relevant articles. This problem is hard to fix with simple post-hoc filtering.

Knowledge cutoff dates can be unclear or unreliable. Model creators generally report a "knowledge cutoff date" for their model, after which the model's knowledge is not updated. The intended purpose of this is not train-test separation for forecasting evaluation. Rather, it is to inform users about the date after which the model outputs can be unreliable. For example, it could be when the pretraining data ended, and the model could still be post-trained with some information after this date, perhaps less exhaustively. Hence, the knowledge cutoff date is not to be taken as a guarantee that the model will not have access to information after the date, though it often is (Halawi, 2024).

Even in cases where a training cutoff date is correct for the main model, when not testing the default API model, system prompts and other scaffolding of the model can leak information. Anthropic's Claude.AI (knowledge cutoff without search: November 2024) system prompt, as of 15 May 2025, reportedly contains the snippets "Donald Trump is the current president of the United States and was inaugurated on January 20, 2025." and "semiconductor export restrictions 2025" (Johnson, 2025).

3. Issues in Extrapolating from Benchmarks to Real-World Forecasting Capabilities

Recent forecasting datasets (Halawi et al., 2024; Karger et al., 2024a) draw from prediction markets like Metaculus, Manifold and Polymarket, and some select sources that update over time. We now discuss how the number of available questions in the backtesting period, correlation between different questions, and inherent data distribution skews make extrapolations to general forecasting capabilities unreliable.

The number of backtest questions after filtering is low. The most popular data source for forecasting benchmarks are prediction market questions (Karger et al., 2024a; Halawi et al., 2024; Phan et al., 2024; Zou et al., 2022; Paleka et al., 2024; Tao et al., 2025). On multiple such platforms, users can ask questions about anything they want. Hence, many questions, especially on play-money platforms like Manifold, are irrelevant personal questions (see Figure 2).

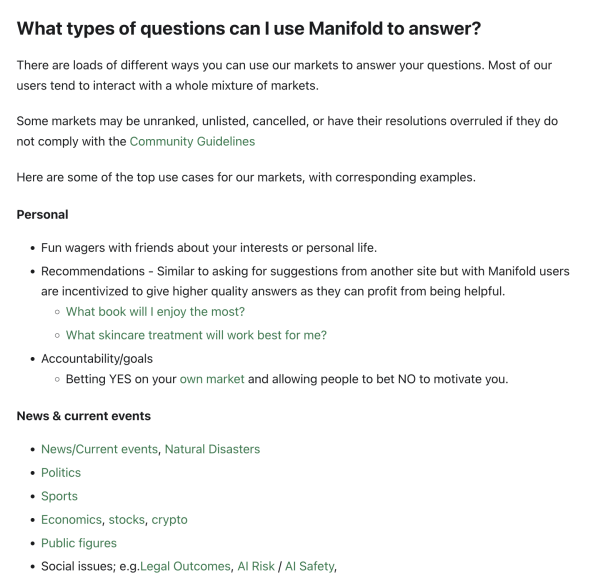


Figure 2. Manifold emphasises Personal use-cases over News and current events, whereas the latter is more relevant when benchmarking language model forecasting.

Forecasting benchmarks should filter out such questions, either using a single cleaning prompt (Halawi et al., 2024), or using multi-step question verification (Paleka et al., 2024).

Figure 3 shows the breakdown of monthly resolved Manifold questions starting July 2024 by the number of forecasters, which is a common metric to filter irrelevant questions (Halawi et al., 2024). Manifold produced 1000-6000 questions in the second half of 2024. Some of these questions include under-specified or irrelevant questions like "Will I lift weights today?" (id: uPdSLhP0dn). Over 50% of the questions in each month had less than 12 forecasters. We find such prediction volume filters also lead to a large number of false negatives. Many filtered questions are perfectly reasonable for forecasting, but just happen to not attract predictions. For example, this filter systematically reduces short-horizon questions that resolve fast.

Overall, these issues lead to a lower number of questions being available for forecasting evaluations each month. This issue is exacerbated by a recent decreasing trend in the number of months available for backtesting for frontier models, as discussed in Appendix B.3. In Appendix B.1 we demonstrate how attempts to increase the number of questions by generating them with LLMs leads to a whole new set of issues, where questions can be solved without forecasting or world understanding, using shortcuts.

Maximizing chance of being the best predictor does not elicit the best forecasting system. Real-world prediction contests such as the ACX/Metaculus Prediction Contest (Metaculus, 2025; Hanania, 2022) are often accompanied

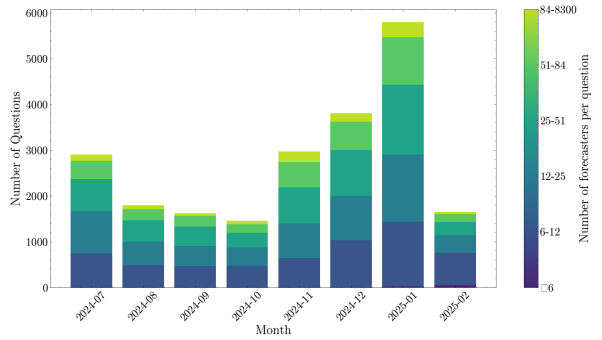


Figure 3. Resolved manifold questions by month, with colors representing the number of forecasters that made a prediction on the question, a commonly used proxy for whether people care about the question for forecasting.

Table 15: Categories and question counts by source, dropping invalid questions.

	RFI	Manifold	Metaculus	Polymarket	ACLED	DBnomics	FRED	Wikipedia	Yahoo!	Total
Arts & Recreation	0	42	10	65	0	0	0	0	0	117
Economics & Business	2	13	55	154	0	0	166	0	509	899
Environment & Energy	0	2	37	7	0	52	0	0	0	98
Healthcare & Biology	0	8	71	3	0	0	0	215	0	297
Politics & Governance	3	16	128	188	0	0	0	0	0	335
Science & Tech	5	66	172	15	0	0	0	1	0	259
Security & Defense	3	9	109	12	3,220	0	0	0	0	3,353
Sports	0	65	18	468	0	0	0	137	0	688
Other	6	151	122	2	0	0	0	75	0	356
Total	19	372	722	914	3,220	52	166	428	509	6,402

Figure 4. Distribution of ForecastBench questions across domains and data sources, table borrowed from the Appendix of Karger et al. (2024a). Users on each market favour specific categories, overweighing them when market questions are used for benchmarking. Further, ForecastBench questions from non-market sources all follow highly specific templates akin to time series prediction.

with monetary prizes rewarding the best performers. It is known that, in a theoretical prediction market contest, maximizing the chance of being the best predictor encourages taking correlated risks over betting based on one's actual honest beliefs (Sempere & Lawsen, 2021). In practice, the winner of one such contest in 2022 has said: "I tried to deliberately structure my answers to maximize my probability of winning, rather than maximize the probability of each individual answer being correct." (Alexander, 2023)

A similar dynamic might occur in benchmarking AI forecasting systems. Consider a forecasting system in September 2024, predicting on a set of economic questions resolving in 2025, including questions such as "Will the total trade volume between the US and China be higher in 2025 than in 2024?" and "Will the US government resume collecting student loan payments in 2025?". There is a key latent variable that correlates with the outcome of many of these questions: the outcome of the 2024 US presidential election.

A good forecasting system reporting its true probabilities would likely estimate $P(\cdot \mid \text{Republican win})$ and $P(\cdot \mid \text{Democrat win})$ for all questions, and average out its predictions over the two possible outcomes. A forecasting system that wants to maximize its chance of performing very well on this dataset should just assume that the outcome of the 2024 election is certain. In reality, the forecasters that were confident about a Republican win perform very well, the forecasters that were confident about a Democrat win perform very badly, but the forecasters that were calibrated in their uncertainty are in between.

Case study: ForecastBench. To demonstrate data distribution skews more concretely, we now analyze ForecastBench (Karger et al., 2024a), the most prominent live leaderboard for LLM forecasting. Figure 4 shows the distribution of questions across categories and data sources. Inspecting their codebase (Karger et al., 2024b), we find that questions from non-market sources follow one among a few templatic formats, listed in Figure 7. The highest number of questions are sourced from a database of global conflict statistics, the Armed Conflict Location and Event Data (ACLED). These questions require predicting whether the number of conflicts in a particular region would increase by a fixed ratio in a particular time period. Essentially, the task boils down to time-series prediction. The same can be said for other non prediction market sources. Yahoo Finance stock price changes for particular indicators. "Wikipedia", while seemingly broad, consists of specific templatic questions about predicting change in Chess elo ratings, swimming world records, and whether a vaccine for a particular disease will be developed. Questions from FRED are about time-series changes in macroeconomic indicators, whereas DBNomics has weather time-series from locations across France. Overall, while each of these time series categories could form a few interesting forecasting questions, in ForecastBench these specific time series form the majority of the dataset.

The use of LLMs was originally motivated for *judgemental forecasting* about discrete events (Zou et al., 2022), as here classical time series models without language understanding cannot be applied directly. ForecastBench does have the three prediction markets as the second most frequent sources, after ACLED. However, even prediction markets exhibit domain-specific skews that reflect the interests of their user base. Polymarket, for example, is disproportionately focused on cryptocurrency price movements and sports results, while Manifold includes a large number of personal questions such as "Will I go to the gym today?". More generally, markets tend to overrepresent US centric political, economic and sports events, as that is where most of the user base comes from. Overall, these data distribution issues highlight how performance on ForecastBench may not reflect general world modelling capabilities, which ideal forecasting evaluations of LLMs do hold the potential for.

References

Alexander, S. Who predicted 2022?, 2023. URL https://www.astralcodexten.com/p/who-predicted-2022. Accessed on 12-May-2025.

Bosse, N., Mühlbacher, P., Phillips, L., and Schwarz, D. Contra papers claiming superhuman AI forecasting, 2024. URL https://www.lesswrong.com/posts/uGkRcHqatmPkvpGLq/contra-papers-claiming-superhuman-ai-forecasting.

Boudoukh, J., Israel, R., and Richardson, M. Long-horizon predictability: A cautionary tale. *Financial Analysts Journal*, 2019.

Branwen, G. Ai forecasting bots incoming: comment section, 2024. URL https://www.lesswrong.com/posts/4kuXNhPf9FBwok7tK/ai-forecasting-bots-incoming?commentId=MirX9bPg232BuzMBq. Accessed 12 May 2025.

Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., and Van Durme, B. Dated data: Tracing knowledge cutoffs in large language models. *arXiv preprint arXiv:2403.12958*, 2024.

Dai, H., Teehan, R., and Ren, M. Are llms prescient? a continuous evaluation using daily news as the oracle. In *ICML*, 2025. URL https://arxiv.org/abs/2411.08324.

Halawi, D. Knowledge cutoff issues of gpt-4o regarding phan et al. (2024), 2024. URL https://x.com/dannyhalawi15/status/1833295067764953397.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models, 2024.

Hanania, R. Introducing the salemcspi forecasting tournament, 2022. URL https://www.cspicenter.com/p/introducing-the-salemcspi-forecasting. Accessed on 12-May-2025.

Johnson, A. T. asgeirtj/system_prompts_leaks/claude.txt, 2025. URL https://github.com/asgeirtj/system_prompts_leaks/blob/f7d92dec4a9a9f5e4d11e4384f8239fb7ca3be05/claude.txt. Accessed 12 May 2025.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. Forecastbench: A dynamic benchmark of ai forecasting capabilities, 2024a. URL https://arxiv.org/abs/2409.19839.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. forecastingresearch/forecastbench/, 2024b. URL https://github.com/forecastingresearch/forecastbench/tree/8e236823e3683584330fc31bf41196a81ce28626/src/helpers.

Metaculus. Acx2025 tournament, 2025. URL https://www.metaculus.com/tournament/ACX2025/. Accessed on 12-May-2025.

Paleka, D., Sudhir, A. P., Alvarez, A., Bhat, V., Shen, A., Wang, E., and Tramèr, F. Consistency checks for language model forecasters. *arXiv preprint arXiv:2412.18544*, 2024.

Phan, L., Khoja, A., Mazeika, M., and Hendrycks, D. Llms are superhuman forecasters, 2024. URL https://drive.google.com/file/d/1Tc_xY1NM-US4mZ4OpzxrpTudyo1W4KsE. Center for AI Safety, UC Berkeley.

Sarkar, S. K. and Vafa, K. Lookahead bias in pretrained language models. *Available at SSRN*, 2024.

Sempere, N. and Lawsen, A. Alignment problems with current forecasting platforms. *arXiv preprint arXiv:2106.11248*, 2021.

Tao, Z., Jin, Z., Li, B., Bai, X., Zhao, H., Dou, C., Chen, X., Li, J., Li, L., and Tao, C. Prophet: An inferable future forecasting benchmark with causal intervened likelihood estimation. *arXiv preprint arXiv:2504.01509*, 2025.

Wang, H. Haooowang/llm-knowledge-cutoff-dates, 2025. URL https://github.com/HaoooWang/llm-knowledge-cutoff-dates/blob/f2ea76a47437c2787cc651838b5c7af4720d1c0a/README.md. Accessed 12 May 2025.

Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., and Haffari, G. Continual learning for large language models: A survey, 2024. URL https://arxiv.org/abs/2402.01364.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. *arXiv preprint arXiv:2206.15474*, 2022.

A. Backtesting Leakage

A.1. Additional Examples of Retrieval Bias

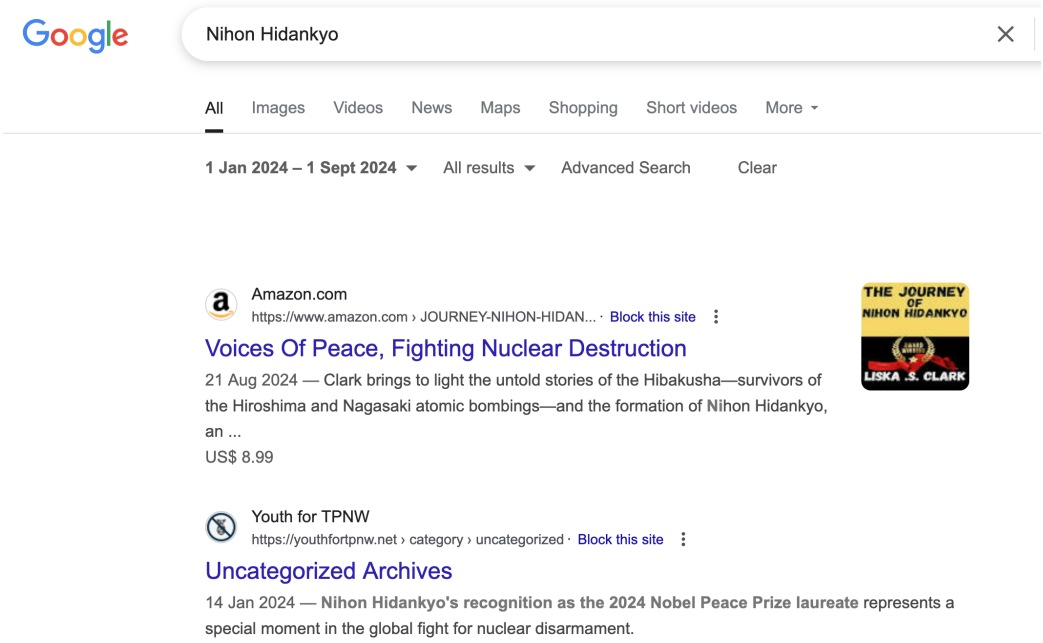


Figure 5. Search results with date restriction showing an article with January 2024 publication date containing information about the October 2024 Nobel Peace Prize announcement. This type of temporal leakage can artificially improve forecasting performance.

Here we provide additional examples of how search engines leak information when using date restrictions, expanding on the examples shown in Section 2. Even when filtering out content published after a certain date, the selection of which articles are deemed most relevant appears influenced by knowledge of future events.

These examples further demonstrate how search engines with date restrictions can leak information through selection bias. The top results for searches like ”October 7” (Figure 6a) feature content that would likely not be prominent before this date became associated with the Israel-Hamas conflict starting in 2023. Similarly, Figure 6d and Figure 6c shows bias toward names that would later become culturally significant.

A.2. Notes on model knowledge cutoff dates

Although most model providers do provide such a cutoff date for their API models, some models such as the Mistral series do not have readily available training cutoff dates (Wang, 2025).

On a different note, knowledge cutoffs can create problems aside from temporal leakage when comparing different models. Cheng et al. (2024) highlight how the cutoff can be more continuous than discrete, and sometimes the *effective knowledge cutoff* in some domain can be different from others. For example, the model with a knowledge cutoff date of 2024 might only contain court case documents up to 2023.

B. Issues in Extrapolating to Real-World Forecasting Capability

B.1. Issues with recent evaluations using LLM generated questions from news articles

Recent works (Dai et al., 2025; Paleka et al., 2024) have used LLMs to create forecasting questions for backtesting, using news articles as a reference. Specifically, both papers take news articles between the models existing knowledge cutoff and today, and use LLMs to generate questions for each article. This overcomes the issue on having to rely on prediction market questions and significantly expands the distribution of topics, but comes with its own issues.

Generating binary questions from news biases the dataset towards things that happen. Dai et al. (2025) back-generated questions from news articles, resulting in a dataset where reference class forecasting performs extremely poorly. Concretely, here are some questions from their dataset that resolve "Yes":

- Will a co-worker of Ronald Silver II share details about the unsafe working conditions of Baltimore DPW during the 2024-11-25 news conference?
- Will a body be found inside a trash can on the 20400 block of Omira Street in Detroit in early November 2024?
- Will the recall of apple juice due to high levels of inorganic arsenic expand to include multiple brands totaling 133,500 cases by September 2024?

A forecaster that knows the dataset is generated from news articles will have a much higher chance of forecasting correctly, as these questions are overly specific, to the point that any reasonable would have a high prior for saying "No" for these exact events (conjunction of many uncertain outcomes) occurring. In general, the news tends to highlight interesting events like "schools closing in Nevada this week", and is much less likely to mention the default state (high prior) such as "schools not closing in Washington this week". An incomplete fix to this issue is presented in (Paleka et al., 2024), where they augment their news-generated dataset with slight modifications to the questions to create similar-looking questions that resolve to the opposite outcome.

To what extent can forecasting questions be solved with shortcuts? Many of the issues we mentioned with LLM generated, or resolved question phrasing, can essentially be considered shortcuts that can be exploited to solve the forecasting question without any reasoning about the future. One way to quantify the extent of how much a given dataset can be solved with shortcuts is by finetuning a weak classifier on these questions. We finetune a DeBERTa model released in 2021, that has definitely not seen the test set we predict on in its training, and give it no retrieved documents. For binary Yes/No questions, we train a two class classifier after balancing the data to ensure that the constant baseline (all Yes, or all No) accuracy is 50%. For multiple choice questions with four options, we train the classifier to predict the option ID (A, B, C, or D) given the question and options in the prompt. We temporally split the data to avoid any leakage. We find this leads to high accuracies (up to 80%) on even the four choice MCQ dataset released by Dai et al. (2025), where the chance baseline is 25%. Even when we reproduce their pipeline with newer models (DeepSeek v3 0324) and improved prompts, we still achieve a four choice MCQ accuracy of 55%. The accuracy is non-trivial, but much lower on Metaculus (55%) and Manifold (59%). We believe this DeBERTa classifier only catches on easy shortcuts and does not actually engage in meaningful forecasting.

B.2. Prediction market information can be included in the training set or retrieved articles

When comparing LLM forecasting predictions with human performance, it is important to understand whether the human baseline predictions are already in the dataset. Usually, benchmarks consist of questions that resolve in a certain period (say Oct-Dec 2024), often scraped from prediction market websites. Even when the knowledge cutoff of the forecasting system is before the resolution date, questions that resolve in a certain period were likely forecasted by people for a long time before that: "Which party will win the 2024 US Presidential Election?" on Manifold Markets had bets since January 2024, and "Will AI get at least bronze on the International Math Olympiad by end of 2025?" has had many people bet and explain their reasoning in comments since May 2023. This makes an "LLM vs market of human forecasters" comparison unfair, because the LLM can just copy the market probabilities. Even when comparing across LLMs with similar knowledge cutoff dates, this can be an issue: one LLM might be trained on most recent market probabilities, while another is not; giving the former an advantage on questions that have already been discussed.

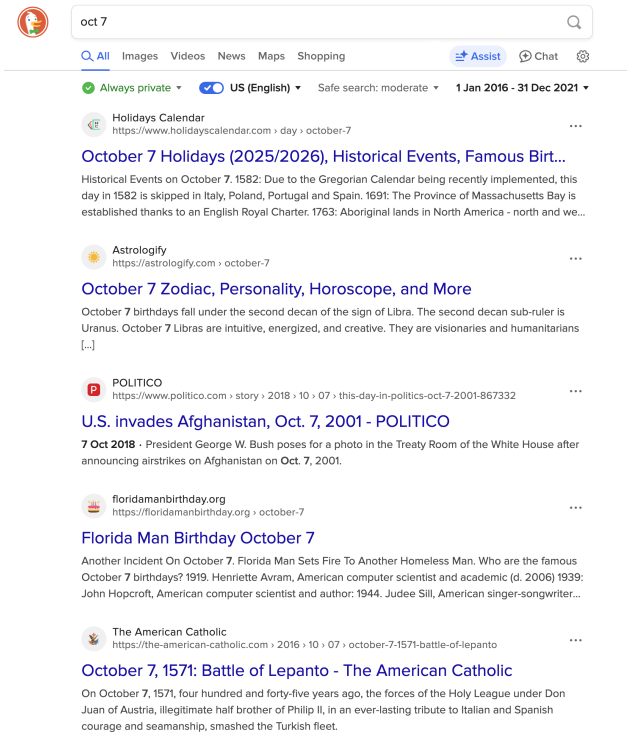
This issue can particularly affect LM scores on ForecastBench (Karger et al., 2024a), which uses human-crowd prediction as the gold standard for measuring Language Model's capability on unresolved questions about predicting an event still in the future. If an LM forecasting system retrieves the relevant prediction market and recent crowd aggregates, it can trivially achieve gold standard performance.

B.3. The period available for backtesting is narrowing

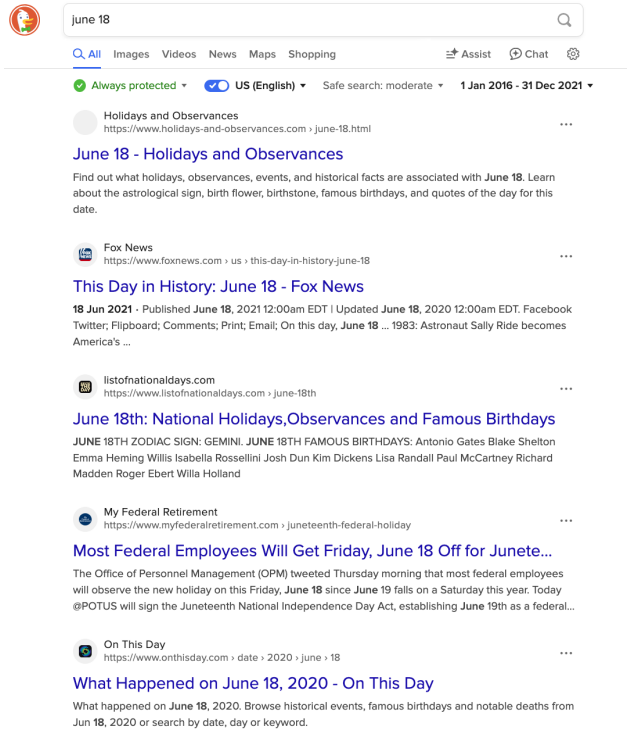
Since new model releases often have substantially improved capabilities, and forecasting is a very challenging task, we mostly want to benchmark only the most recent frontier models. For rapid backtesting of any given model, we need questions that have resolved in the period [knowledge cutoff date, . . . , release date, . . . today]. First, note that the time gap between the

knowledge cutoff date and release date of models is decreasing, as shown in Figure 8. Further, especially with recent focus on post-training to further enhance capabilities, newer models with better capabilities are released much faster. This further leads to reduced time gap between today and the release date.

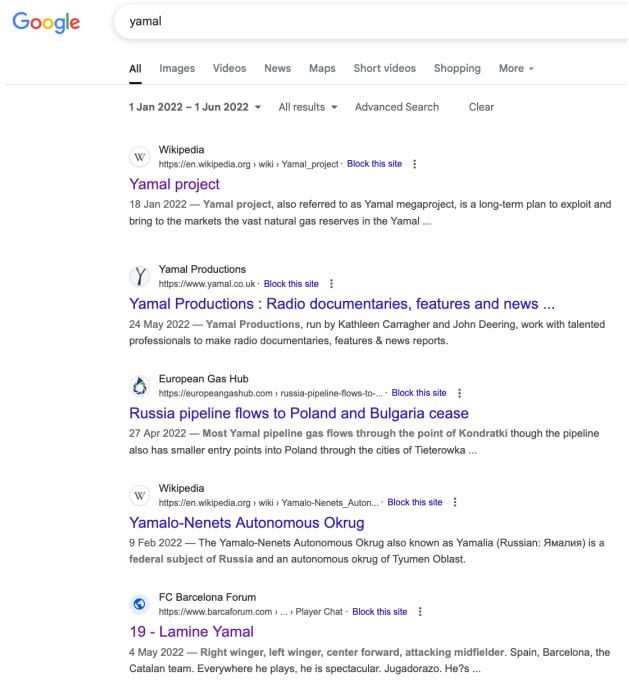
Together, this leads to a trend of the backtest period narrowing, further limiting the number of questions that can be used for backtest evaluations, which increases variance in measured performance and reduces the reliablity of backtests. A narrowing backtest period also means we can only back-test frontier models on predictions with increasingly short time-horizons. Short-horizon prediction success may not correlate with predictions about longer-horizons (Boudoukh et al., 2019), making generalization to real-world use harder. It is possible that API models start seeing continuous knowledge updates (Wu et al., 2024), in which case it might not even be possible to backtest them at all.



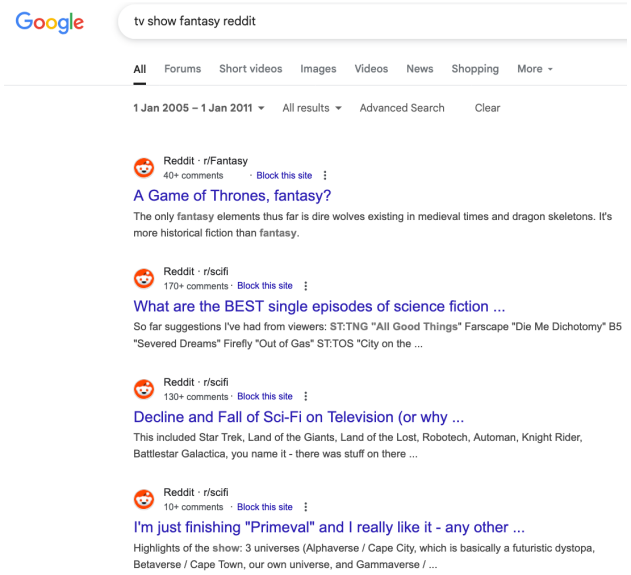
(a) Search results for "October 7" with date restriction before 2022. Note the prominence of articles about conflict in the Middle East.



(b) For comparison: search for "June 18" with the same date restriction shows largely unbiased results about the date. Note the lack of mention of the Battle of Waterloo that happened on June 18, 1815; in contrast to the Oct 7 query that mentions two distinct military engagements.



(c) Search results for "Yamal" from the first half of 2022. The discussion about a 14-year-old Lamine Yamal, at the time known only to visitors of Barcelona fans forum, is in the top 5 results.



(d) Search results for "TV show fantasy" with date restriction before 2011. The discussion about the book (not show yet!) Game of Thrones is very prominent.

Figure 6. Additional examples of retrieval bias in search engines when using date restrictions.

Summary of Questions Obtained Across Data Source in ForecastBench

ACLED All questions adopt one of two forms:

1. Will there be more {event_type} in {country} for the 30 days before {resolution_date} compared to the 30-day average of {event_type} over the 360 days preceding {forecast_due_date}?
2. Will there be more than ten times as many {event_type} in {country} for the 30 days before {resolution_date} compared to one plus the 30-day average of {event_type} over the 360 days preceding {forecast_due_date}?

DBnomics All questions are of the form: What is the probability that the daily average temperature at the French weather station at {station} will be higher on {resolution_date} than on {forecast_due_date}?

FRED All questions are of the following format, but different financial time series: Will the euro short-term rate (volume-weighted trimmed mean), a measure of the borrowing costs of banks in the euro area, have increased by {resolution_date} as compared to its value on {forecast_due_date}?,

Wikipedia Slow-changing queries of one of the following four forms:

- According to Wikipedia, will a vaccine have been developed for {id} by {resolution_date}?
- According to Wikipedia, will {id} have a FIDE ranking on {resolution_date} that is "high or higher" than on {forecast_due_date}?
- According to Wikipedia, will {id} have an Elo rating on {resolution_date} at least 1 % higher than on {forecast_due_date}?
- According to Wikipedia, will {id} still hold the world record for {value} in long course (50 m) swimming pools on {resolution_date}?

YAHOO All questions are of this form, but with different stock indicators: Will AMTM's market close price on {resolution_date} be higher than its market close price on {forecast_due_date}?

Figure 7. ForecastBench obtains questions from multiple sources, but from each source, questions follow very specific templates (Karger et al., 2024b). The dataset thus resembles more an aggregation of predictive performance over some very specific time series, rather than general judgemental forecasting.

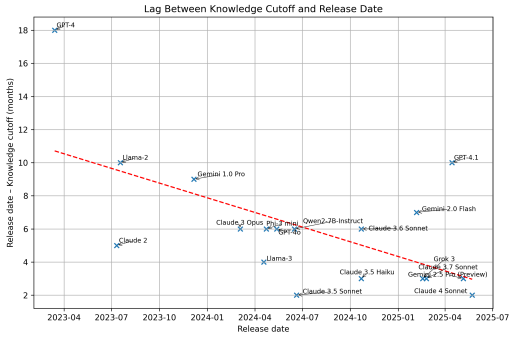


Figure 8. The gap between the knowledge cutoff and when the model is relevant is getting smaller.