Are LLMs Prescient? A Continuous Evaluation using Daily News as the Oracle

Hui Dai [1] **Ryan Teehan** [1] **Mengye Ren** [1]

Abstract

Many existing evaluation benchmarks for Large Language Models (LLMs) quickly become outdated due to the emergence of new models and training data. These benchmarks also fall short in assessing how LLM performance changes over time, as they consist of a static set of questions without a temporal dimension. To address these limitations, we propose using future event prediction as a continuous evaluation method to assess LLMs' temporal generalization and forecasting abilities. Our benchmark, Daily Oracle, automatically generates question-answer (QA) pairs from daily news, challenging LLMs to predict "future" event outcomes. Our findings reveal that as pre-training data becomes outdated, LLM performance degrades over time. While Retrieval Augmented Generation (RAG) has the potential to enhance prediction accuracy, the performance degradation pattern persists, highlighting the need for continuous model updates. Code and data are available at https://agenticlearning.ai/daily-oracle.

1. Introduction

Traditional Large Language Model (LLM) benchmarks are often static, and do not reflect real-world information that evolves over time. This presents two significant challenges. First, as LLMs are updated, there is a risk that static benchmarks become outdated and more vulnerable to data leakage, where their content might end up in the training data of newer models. This undermines the reliability of performance assessments on these benchmarks (Sainz et al., 2023; Xu et al., 2024; McIntosh et al., 2025; Li & Flanigan, 2024). Second, static benchmarks often lack temporal information, making it difficult to track models' performance variations over time (McIntosh et al., 2025). This creates a need for evaluation methods that always remain relevant and incorporate temporal dynamics.

Daily news provides a natural setting for continuous evaluation of LLMs. Since the world is constantly changing, a benchmark designed around forecasting the next day's news will never be out of date by construction. In addition to enabling continuous evaluation, forecasting is itself a longstanding challenge with significant implications across various domains, including healthcare, finance, and policy-making (Tetlock & Gardner, 2016; Dempsey et al., 2017; Gillingham et al., 2018; Lopez-Lira & Tang, 2023). While human experts have traditionally made such forecasts, machine learning models, particularly LLMs, have emerged as promising alternatives due to their ability to learn from vast and diverse corpora (Halawi et al., 2024; Ye et al., 2024; Yan et al., 2024). Several recent forecasting question-answer (QA) datasets have been developed (Jin et al., 2021; Zou et al., 2022; Zhang et al., 2024), however, they are limited in either size, scope, or they do not continuously keep pace with the rapidly changing world. More critically, the extent to which LLMs' predictive abilities change over time remains understudied.

In this work, we propose Daily Oracle—a continuous evaluation benchmark that uses automatically generated QA pairs from daily news to assess how the future prediction capabilities of LLMs evolve over time. The QA pairs are generated on a daily basis, consisting of True/False (TF) and Multiple Choice (MC) questions across various categories such as business, politics, and arts. Unlike traditional reading comprehension tasks, these QA pairs are designed to challenge LLMs to predict future events based on their own existing knowledge, effectively evaluating their temporal generalization and forecasting abilities.

We continuously evaluate various LLMs, both with and without access to a limited archive of news articles. Our experiments reveal that LLMs experience **significant performance degradation** between January 2020 and December 2024, with degradation becoming more pronounced before and after the models' knowledge cutoff dates. On average, performance drops by 21.55% on TF questions

[1]New York University. Correspondence to: Hui Dai <hd2584@nyu.edu>, Ryan Teehan <rst306@nyu.edu>, Mengye Ren <mengye@nyu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

and 11.33% on MC questions. While model performance can be improved with more recent news articles using Retrieval Augmented Generation (RAG) (Lewis et al., 2020), the **downward** trend persists, suggesting the challenge in maintaining its prediction ability over time.

To summarize, our key contributions are two-fold:

- **Continuous Forecasting Evaluation Benchmark**: We present Daily Oracle, the largest and most up-to-date forecasting dataset, composed of automatically generated QA pairs. This benchmark continuously evaluates LLMs' temporal generalization and future prediction abilities using daily news, ensuring relevance over time and offering a challenging evaluation framework.

- **Empirical Findings on Performance Degradation**: Since our benchmark provides new questions each day, we can study how model performance changes along the temporal axis. Our work effectively reveals a clear performance degradation pattern in LLMs' forecasting accuracy over time. Additionally, we study how this pattern changes as the LLMs are given access to updated knowledge up to different times. Surprisingly, we find that, **even when the model has access to recent information in an "open-book" setting,** it still experiences performance degradation. Moreover, the sheer degree of decline, along with its smoothness over time, was unexpected. On the one hand, this highlights the problems with outdated LLM pre-training data and on the other hand underscores the need for continuous model updating.

2. Related Work

Temporal Generalization of LLMs. Lazaridou et al. (2021) define temporal generalization as the ability of Language Models to generalize well to future data from beyond their training period. They demonstrate that Transformer-XL's performance deteriorates over time, evidenced by increasing perplexity when evaluated on post-training data. However, perplexity-based metrics have two main limitations: they cannot be applied to closed-source models lacking accessible logits, and increased perplexity does not necessarily indicate degraded performance on downstream tasks (Röttger & Pierrehumbert, 2021; Agarwal & Nenkova, 2022). Zhu et al. (2025) investigate temporal generalization using the Bits Per Character (BPC) metric. Similar to perplexity, BPC fails to capture higher-level performance on downstream tasks. In contrast, our work focuses on the downstream forecasting task, evaluating how well models understand world knowledge and make predictions. This approach offers a more reliable evaluation of temporal generalization with direct relevance to real-world applications and public interest.

Dynamic QA Datasets. While static QA datasets evaluate models on fixed knowledge snapshots, dynamic QA datasets incorporate a temporal dimension, allowing assessment of how models adapt to evolving information. Several dynamic QA datasets are proposed. Chen et al. (2021) construct TimeQA by using time-sensitive facts in WikiData with aligned Wikipedia passages to synthesize QA pairs. Zhang & Choi (2021) introduce SituatedQA by manually annotating temporally and geographically dependent questions. StreamingQA (Liska et al., 2022) and RealtimeQA (Kasai et al., 2024) are both dynamic benchmarks with QA pairs answerable from news articles. StreamingQA, however, does not provide continuous evaluation with always-relevant data. RealTimeQA does not address forecasting and is more like a plugin for a search engine, in the sense that it tests whether a model has updated its knowledge as facts change, rather than testing whether it can predict what will change given its knowledge of the past. FreshQA (Vu et al., 2024) contains a fixed set of human-written open-ended questions whose answers by nature can change based on new developments in the world, but is smaller and does not address forecasting. It is also updated weekly rather than daily. While all these datasets have some form of time-sensitivity like the Daily Oracle, they either do not provide continuous evaluation or do not evaluate forecasting capabilities, or neither.

Forecasting Datasets. Forecasting questions aim to assess a model's ability to predict the outcomes of future events based on its existing knowledge. Several datasets in the event forecasting field have been introduced. ForecastQA (Jin et al., 2021) used crowdworkers to collect 10,392 QA pairs from news articles. Zou et al. (2022) argue that the QA pairs from ForecastQA are often nonsensical or ambiguous since they are written by humans without forecasting expertise. They further introduce AutoCast, a forecasting dataset from popular human forecasting tournaments containing 6,707 QA pairs. While ForecastQA and AutoCast remain static, ForecastBench (Karger et al., 2025) regularly updates a set of 1,000 forecasting questions either sourced from forecasting markets or generated via fixed templates based on real-world event datasets. However, it still depends on users actively submitting new forecasting questions or maintaining the underlying datasets. In contrast, our Daily Oracle dataset is generated automatically from daily news articles, which means that it is never out of date, can easily grow its size automatically without additional inputs from human forecasters, and provides more comprehensive event coverage than human forecasting tournaments.

Similar to our generation method, TLB-forecast (Zhang et al., 2024) has an automatic forecasting QA generation framework using news articles. However, their dataset is constrained both temporally and topically, only containing cooperation and conflict events in Middle-Eastern countries

Dataset	Continuous?	Interval	Forecast?	Size	Latest Update
TimeQA (Chen et al., 2021)	✗	None	✗	20,000	2021
SituatedQA (Zhang & Choi, 2021)	✗	None	✗	4,757	2021
StreamingQA (Liska et al., 2022)	✗	None	✗	36,800	2021
RealTimeQA (Kasai et al., 2024)	✗	None	✗	1,470	2023
FreshQA (Vu et al., 2024)	✓	Weekly	✗	600	2024
ForecastQA (Jin et al., 2021)	✗	None	✓	10,382	2019
AutoCast (Zou et al., 2022)	✗	None	✓	6,707	2022
ForecastBench (Karger et al., 2025)	✓	Biweekly	✓	1,000	2024
TLB-forecast (Zhang et al., 2024)	✗	None	✓	6,604	2022
FreshBench (Zhu et al., 2025)	✓	Unknown	✓	2,769	2024
Daily Oracle (Ours)	✓	Daily	✓	31,510	2024*

* Our experiments use the subset generated until December 2024. Daily Oracle itself remains active, continuing to generate new questions daily from 2025 onward.

Table 1. We compare Daily Oracle with existing benchmarks in the literature. For continuously updated datasets (e.g. Daily Oracle, FreshQA, FreshBench, and ForecastBench), "Interval" refers to the dataset update interval, and "Size" and "Latest Update" refer to the fixed data currently available. Our Daily Oracle benchmark is the only forecasting benchmark which is **continuously updated every day** using questions generated from daily news.

from 2015 to 2022. This restricts the dataset from evaluating more general event-prediction abilities. Furthermore, considering most of the powerful LLMs have been developed after 2020, the portions of the dataset covering earlier years may contain answers already seen during training. This prior exposure compromises the dataset's effectiveness as a forecasting benchmark. In contrast, our dataset spans a broader timeframe and covers more topics, offering a more comprehensive forecasting benchmark.

Note that none of the aforementioned datasets provide insights into how prediction ability changes over time. Zhu et al. (2025) introduce FreshBench, a forecasting dataset scraped from the Good Judgment Open platform, and also study temporal generalization. However, they report accuracy in a relatively short time window (from January 2023 to August 2024) with only 2,769 questions. While we observe a gradual performance decline in our dataset, they report significant fluctuations in model accuracy shortly after release. A closer look reveals key limitations of forecasting market–based questions for studying temporal generalization: they suffer from limited early coverage, inconsistent distribution over time, and reduced dataset size after filtering due to a high proportion of low-quality questions, making it difficult to reliably analyze temporal performance trends. In contrast, our automatically generated dataset has broad event coverage, consistent growth, and more uniform question quality over time.[1]

In order to clearly showcase the differences between our dataset and prior work, we highlight a few key features in Table 1. The Daily Oracle is the only benchmark which is **continuously updated** on a daily basis and evaluates forecasting ability. Additionally, at the fixed size we use for analysis we provide significantly more evaluation examples than the other automatically updated benchmarks.

3. The Daily Oracle Dataset

In this section, we present Daily Oracle, a continuously updated QA benchmark of forecasting questions that are automatically generated from daily news. For our current analysis of LLM performance, we utilize a subset of the data consisting of 16,783 TF questions and 14,727 MC questions, covering a diverse range of forecasting topics, which are generated using daily news articles from January 2020 up until December 2024. However, our QA generation framework is continuous and updates daily. In Section 3.1, we describe our LLM-based dataset construction pipeline, detailing the data sources and the four-step construction process. Section 3.2 provides an analysis and general overview of the dataset. Lastly, in Section 3.3, we conduct a human evaluation, similar to our QA filtering process, to verify the quality of the generated QA pairs.

3.1. Dataset Construction

Data Source. Following Zou et al. (2022), we collect a large corpus of news articles from the daily-updated Common Crawl News Dataset (Nagel, 2016) with the `news-please` package (Hamborg et al., 2017). To further enrich our news dataset, we supplement it with daily scraped news using the `Newspaper3k` package.[2] We filter for mainstream sources—CBS News, CNBC, CNN, Forbes, and NPR. While our data collection and evaluation are performed daily, for this study we utilize a static news corpus

[1] See Appendix C for details on comparing LLM-generated and forecasting market datasets.

[2] https://newspaper.readthedocs.io/en/latest/