with 1,246,973 English articles spanning January 2019 to December 2024. This corpus is also used for the constrained open-book evaluation setting in Section 4.1.

LLM-based Construction Process. QA pairs are generated from articles published between January 2020 and December 2024.[3] For each day, we select six articles for QA generation: three are chosen randomly, and three are selected from hot topics.[4] For each selected article, we then use LLM to generate two TF QA pairs and two MC QA pairs with the few-shot prompting technique.[5]

We adopt the methodology of Zhang et al. (2024), as their prompt design largely suits our setting. To further filter the QA pairs, we establish seven key criteria to ensure they qualify as valid forecasting questions, incorporating these into a *QA Filtering step*. The QA construction follows four steps, as illustrated in Figure 7:

(1) *Article Summary.* We generate a summary for each article, focusing on new events from the publishing date, instead of opinion articles discussing events from the past. This approach allows us to use the publication date as the resolution date of the generated question. Questions can then be regarded as valid forecasting questions since they are prior to the resolution date.

(2) *QA Generation.* After filtering out the articles that do not introduce new events, two TF questions and two MC questions are generated together with the answers per article. To ensure balance in the TF questions, we instruct the LLM to generate the first question with a "Yes" answer and the second with a "No."

(3) *Misleading Choices Generation.* For MC, we provide the article, its publishing date, and the QA pair to the LLM, which then generates three misleading choices.

(4) *QA Filtering.* We prompt the LLM to check seven principles: correctness of answers, non-answerability before the publication date, absence of information leakage, objectivity, inclusion of a clear temporal element, public interest, and non-obviousness of the answer. Each principle is scored with 0, 1, or 2 points, and we selected the questions that received at least 13 points in total. These principles are detailed in Appendix A.3.

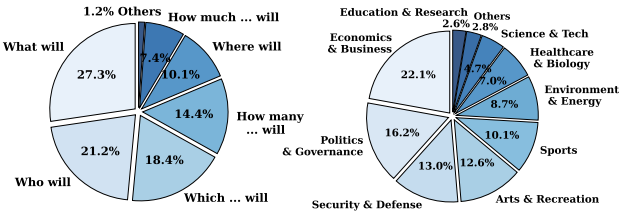
We use GPT-3.5 (OpenAI, 2024a) or GPT-4o-mini (OpenAI, 2024b) for steps (1) and (4), while GPT-4 (OpenAI, 2023) or GPT-4o (OpenAI, 2024b) is utilized for steps (2) and (3) to ensure high data quality.[6]

[3]For news articles in 2019, we use them as the corpus for the constrained open-book setting.

[4]See Appendix A.2 for hot-topic selection details.

[5]See Appendix D for all the prompts we use.

[6]We use GPT-3.5 and GPT-4 until September 2024. After October 2024, we switch to GPT-4o-mini and GPT-4o, as they are both more cost-effective and more powerful.



(a) MC question types (b) Category breakdown

Figure 1. Pie charts showing (a) MC question type distribution and (b) question category distribution in Daily Oracle.

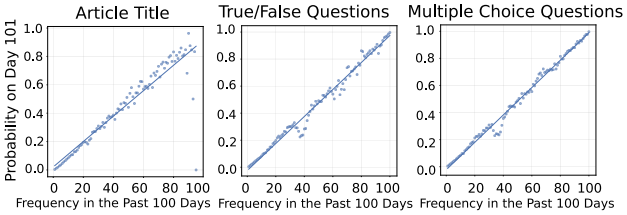


Figure 2. Following Anderson & Schooler (1991), we plot the probability of a word occurring in an **(left)** article title, **(middle)** True/False question, or **(right)** Multiple Choice question given how frequently it had appeared in one over the past 100 days, computed over our entire dataset. We fit a linear regression and show a linear relationship in each case ($R^2 = 0.843$, 0.982, and 0.990 for left, middle, and right respectively).

3.2. Dataset Analysis

Summary Statistics. At the time of writing this paper, the subset dataset we use from Daily Oracle consists of 16,783 TF and 14,727 MC QA pairs, covering the period from January 1st, 2020, to December 31st, 2024, with an average of 17.2 questions per day. Figure 1(a) shows that our dataset covers various MC question types, mainly starting with "What will" (27.3%), "Who will" (21.2%), and "Which ... will" (18.4%). Figure 1(b) provides a breakdown of the categories, highlighting our dataset's broad coverage. The categorization of each question is determined using GPT-3.5, based on the prompt from Halawi et al. (2024). Examples of QA pairs are shown in Table 2.

Past and Future Information Usage. Each question in Daily Oracle implicitly requires the model to retrieve relevant knowledge. How do these requirements change day by day over the course of our benchmark? Anderson & Schooler (1991) explored similar patterns in human information environments. Inspired by their work, Figure 2 examines whether a word's frequency over the past 100 days

Table 2. Daily Oracle Example Questions and Answers.

Type		Category	Question and Answer
TF		Politics & Governance	– Will the prosecution's key witness in the New York hush money trial in April 2024 be someone other than Michael Cohen? –**No.**
TF		Politics & Governance	– Will the House Energy and Commerce Committee vote unanimously to advance a bill that could potentially ban TikTok if ByteDance does not sell the app by March 2024? –**Yes.**
MC	What	Science & Tech	– What will be the starting price range for the Google Pixel 8a as of May 2024? A.$599–$649 B. $199–$249 C. $750–$800, D. $499–$559. –**D.**
MC	Who	Sports	– Who will go on the injured list before the New York Mets' game on May 29, 2024? A. Pete Alonso B. Edwin Diaz C. Jeff McNeil D. Francisco Lindor –**B.**
MC	Which	Arts & Recreation	– By May 2024, on which streaming service will "The First Omen" become available for subscribers? A. Disney+, B. Hulu, C. Amazon Prime Video, D. Netflix –**B.**
MC	How many	Science & Tech	– How many U.S. states will the path of totality cross during the total solar eclipse on April 8, as reported by February 2024? A. 15 B. 10 C. 20 D. 6 –**A.**
MC	Where	Healthcare & Biology	– Where will the second known U.S. case of bird flu in a human be reported by March 2024? A. California, B. Texas, C. New York, D. Florida –**B.**
MC	How much	Economics & Business	– How much will Apple, Inc. (AAPL) be up year-to-date by the end of June 2024? A. Up 149.5% B. Just over 19% C. 9.7%. D. 27%. –**C.**

predicts its occurrence the next day—e.g., if many questions concern the unemployment rate, will this trend continue?

We analyze this relationship for words in the titles of the articles we use to generate questions as well as in the text of the TF and MC questions themselves. Past frequency is computed by checking, for each day in the 100 day window, if a word has occurred in any article title (so, the maximum frequency is 100). We find that there is a linear relationship between the frequency of usage in the past 100 days and the probability of occurrence on the 101st day in all cases, replicating Anderson & Schooler's findings for *New York Times* headlines. This indicates that past information usage strongly predicts future retrieval needs, suggesting a temporal structure in the knowledge demands of our benchmark.

3.3. Human Evaluation

To assess the quality of our LLM-based filtering method, we randomly sample 30 TF and 30 MC QA pairs and ask 4 human annotators to evaluate them using the seven principles outlined in the *QA Filtering* step in Section 3.1. We evaluate the consistency among annotators using Fleiss' Kappa (Fleiss, 1971), which yields an average inter-rater agreement score of 0.26, indicating fair agreement. We then compute the human consensus score as the average of human scores and compare it to LLM-assigned scores, finding an average accuracy of 89.52% across the 7 principles. For final QA pair acceptance (i.e., threshold above 13 points), the LLM and human consensus scores demonstrated an accuracy of 85.00%, further supporting the reliability of our LLM-based filtering approach. A detailed breakdown of human evaluation metrics is provided in Appendix A.4.

4. Experiments

We first introduce three evaluation settings in Section 4.1: 1) no access to external information, 2) access to retrieved recent news articles, and 3) access to gold articles. Section 4.2 presents the results, and Section 4.3 provides deeper insights into the observed degradation patterns.

4.1. Experimental Setup

Closed-Book Setting. We evaluate various LLMs on Daily Oracle to assess their understanding of real-world events and temporal generalization abilities, i.e., how accurately LLMs can answer forecasting questions based on the knowledge they learned from their training data. Our evaluation differentiates between two scenarios based on the question's resolution date and model's knowledge cutoff date: (1) *Pre-Knowledge Cutoff Questions:* These questions have resolution dates before the model's knowledge cutoff, testing the model's understanding of past events. (2) *Post-Knowledge Cutoff Questions:* These have resolution dates after the knowledge cutoff, requiring models to predict future events and test their forecasting and temporal generalization abilities.

Constrained Open-Book Setting. In addition to a closed-book evaluation, we explore the constrained open-book setting: how access to news articles up to different time cutoffs influences LLM performance using RAG (Lewis et al., 2020). We introduce the concept of the RAG cutoff (*R-Cutoff*), which limits the latest accessible date for retrieving articles. To prevent the models from leveraging information beyond the resolution date, for any question with a resolution date (d_{res}), the accessible articles span from January 1st, 2019 (the start of our news corpus) up to whichever comes first between the day before the resolution date and the RAG cutoff date ($d_{\text{R-Cutoff}}$). Formally, the accessible date range is $[01/01/2019, \min(d_{\text{res}} - 1, d_{\text{R-Cutoff}}))$. Following prior work (Jin et al., 2021; Zou et al., 2022; Zhang et al., 2024), we employ BM25 (Robertson et al., 1995) as the retriever and select the top 5 articles relevant to each question. We truncate each retrieved article to a maximum length of 512 words. These articles are then incorporated into the input prompts to serve as additional information.

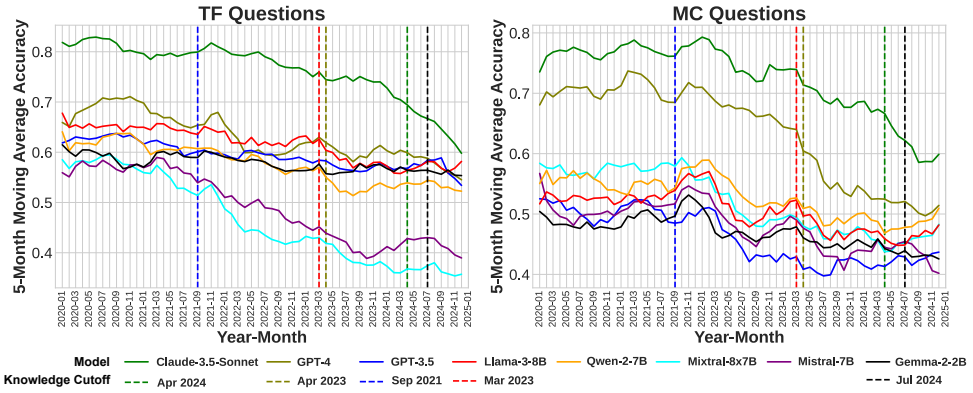


Figure 3. Results for the closed-book setting. We plot the 5-month moving average accuracy for TF and MC questions across various models, showing LLMs' performance degradation in future event prediction.

Table 3. For evaluating various LLMs with different knowledge cutoffs (K-Cutoffs), we show the yearly average accuracy (calculated as the average across months) from 2020 to 2024, along with the average YoY accuracy change (%) before the knowledge cutoff date (Pre-Cutoff), after the knowledge cutoff date (Post-Cutoff), and the overall average YoY accuracy change across all months (Avg).

	LLM	K-Cutoff	Average Yearly Accuracy (%)					Average YoY Accuracy Change (%)		
			2020	2021	2022	2023	2024	Pre-Cutoff	Post-Cutoff	Avg
TF	Claude-3.5-Sonnet	Apr 2024	81.21	79.88	78.05	74.38	64.29	-4.77	-12.41	-5.58
	GPT-4	Apr 2023	69.68	66.41	60.36	60.54	56.90	-5.83	-1.96	-4.75
	GPT-3.5	Sept 2021	62.86	60.12	59.36	57.11	56.09	-4.33	-3.43	-2.84
	Mixtral-8x7B	Unknown	57.83	52.69	43.09	39.34	36.01	–	–	-10.78
	Mistral-7B	Unknown	57.57	54.65	48.22	41.35	40.92	–	–	-7.75
	Llama-3-8B	Mar 2023	65.06	64.24	62.35	58.68	56.99	-1.95	-6.50	-2.97
	Qwen-2-7B	Unknown	62.42	60.18	57.67	53.39	53.04	–	–	-3.75
	Gemma-2-2B	Jul 2024	58.71	59.31	57.64	56.61	55.79	-1.41	-3.68	-1.04
MC	Claude-3.5-Sonnet	Apr 2024	76.86	77.67	74.32	69.37	61.84	-6.26	-11.78	-5.03
	GPT-4	Apr 2023	70.60	70.62	66.76	56.36	51.63	-4.23	-18.54	-7.04
	GPT-3.5	Sept 2021	50.27	50.40	44.38	41.45	43.09	0.14	-0.31	-3.08
	Mixtral-8x7B	Unknown	57.38	56.97	50.76	47.10	46.31	–	–	-4.68
	Mistral-7B	Unknown	50.07	52.36	48.06	44.40	42.99	–	–	-2.82
	Llama-3-8B	Mar 2023	52.44	54.18	50.66	47.94	46.95	-2.21	-1.25	-2.30
	Qwen-2-7B	Unknown	55.28	55.93	53.44	49.77	49.37	–	–	-2.35
	Gemma-2-2B	Jul 2024	47.87	50.71	46.81	45.20	43.28	-4.46	-4.07	-1.98

Gold Article Setting. We further include a setting where models are provided direct access to the gold article, from which the question is generated.[7] This transforms the forecasting questions into reading comprehension ones, which can also access LLMs' general question-answering capabilities. Achieving high accuracy here ensures that the questions from our Daily Oracle dataset are answerable.

Metrics. Accuracy score is used as the evaluation metric. Though LLMs are tested daily, to show clearer trends, we plot the monthly performance in Figure 3, and apply a 5-month moving average to smooth the curve. We also report yearly averages and average year-over-year (YoY) accuracy change before and after models' knowledge cutoff dates in Table 3. Additionally, despite prompting the models to avoid responses like "I cannot predict the future" and instead provide definitive answers, there are cases where such refusals still occur. The refusal rates are provided in the Appendix B.2, and these cases are counted as incorrect to ensure comparability across model results.

[7]See Appendix B.5 for a case example of evaluating LLMs under all three different settings.

4.2. Main Results

Results for the Closed-Book Setting. Figure 3 and Table 3 present our primary results for the closed-book setting. The "Avg" column in Table 3 shows the average YoY accuracy change of all months, revealing a clear degradation in performance over time across all models on both TF and MC questions. When comparing accuracies from the beginning to the end of the evaluation period, we observe that, on average, the models' performance declines by 21.55% on TF questions (from 64.68% to 50.74%) and by 11.33% on MC questions (from 58.30% to 51.69%). This indicates that while LLMs demonstrate certain abilities to understand real-world events and make predictions, they struggle to