Appendix

A. Dataset Details

A.1. Illustration of the dataset construction process.

Figure 7 shows how Daily Oracle is automatically generated as discussed in Section 3.1.

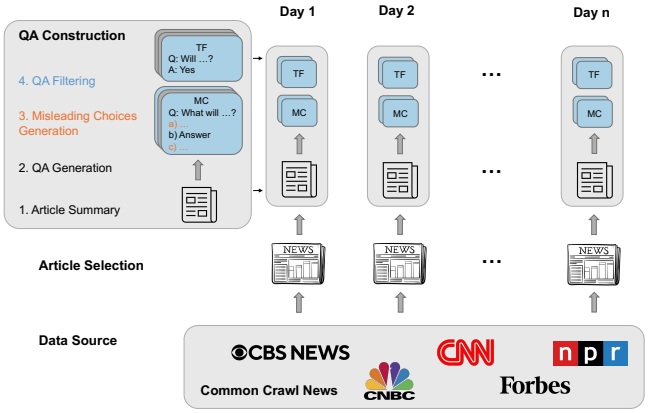

Figure 7. Data Construction Process of Daily Oracle.

A.2. Details for Article Selection

We select daily articles that generate the QA pairs in two ways: (1) *Random Selection:* We randomly sample three articles each day. (2) *Hot Topic Selection:* To better capture daily events and reduce noise, we select three articles from the top three hot topics of the day. We identify these hot topics by applying the density-based clustering algorithm DBSCAN (Ester et al., 1996) to the new articles based on TF-IDF (Term Frequency-Inverse Document Frequency) representations, forming clusters of news articles for each day. We filter out chaotic clusters by removing those with low average in-cluster cosine similarity scores, which typically correspond to clusters containing a large number of diverse articles. The top three clusters, determined by size, are assumed to represent the most discussed events, i.e. hot topics, since larger clusters indicate more articles covering the same event. One article is picked randomly from each of the top three clusters.

A.3. QA Filtering Principles

During the design stage of QA pair generation, we manually review the questions and identify seven key criteria to ensure the QA pairs qualify as valid forecasting questions. These principles guide the QA filtering step in the data construction process:

(1) *Correctness of Answers:* The answer must be factually accurate and fully aligned with the information in the given article.

(2) *Non-answerability Before the Publication Date:* Since we treat the article's publication date as the question's resolution date, the question should not be definitively answerable based on information available before the article's publication.

(3) *Absence of Information Leakage:* Questions must avoid revealing information that became known only after the article's publication, maintaining fairness for pre-publication evaluation.

(4) *Objectivity:* Both questions and answers must rely on objective facts, avoiding subjective ideas from the authors.

(5) *Inclusion of a Clear Temporal Element:* Questions must contain a specific and clear reference to time, avoiding vague phrases like "in the future" or "soon."

(6) *Public Interest:* The questions should address topics of broad public concern.

(7) *Non-obviousness of the Answer:* The answer should not be immediately predictable from the question and must provide new or non-trivial insights.

A.4. Details for Human Evaluation

We assess the quality of our dataset by evaluating the effectiveness of our LLM-based evaluator in the *QA Filtering* step. Four human annotators independently review a randomly sampled subset of Daily Oracle, consisting of 30 TF and 30 MC QA pairs. They follow the same instructions used to prompt the LLM and evaluate each QA pair based on the seven filtering principles listed in Appendix A.3.

Table 4 presents the inter-rater agreement among human annotators and the agreement between human and LLM evaluators. The average Fleiss' Kappa of 0.26 indicates fair agreement among annotators. Among the seven principles, *Objectivity* exhibits the highest agreement (0.66), while *Non-Answerability Before the Publication Date* has the lowest (0.02).

Comparing human-assigned and LLM-assigned scores, the exact-match accuracy between the human consensus and LLM evaluations averages 89.52% across the seven principles, showing the effectiveness of our LLM-based filtering method. *Non-Answerability Before the Publication Date* shows the lowest agreement (83.33% accuracy), suggesting it is the most challenging principle for both humans and the LLM to evaluate consistently.

Table 4. The inter-rater agreement among four human annotators is evaluated using Fleiss' Kappa, while the agreement between human and LLM evaluators is measured through accuracy scores. We report metrics across seven *QA Filtering* principles using a sample of 60 randomly selected QA pairs.

Metric	**Human Agreement** Fleiss' Kappa	**Human vs. LLM** Accuracy (%)
Correctness of Answers	0.11	96.67
Non-answerability Before the Publication Date	0.02	83.33
Absence of Information Leakage	0.33	86.67
Objectivity	0.66	98.33
Inclusion of a Clear Temporal Element	0.21	90.00
Public Interest	0.18	88.33
Non-obviousness of the Answer	0.29	83.33
Average	0.26	89.52

A.5. Distribution of Question Categories Over Time

In Figure 8, we provide the distributions of question categories for both TF and MC questions.

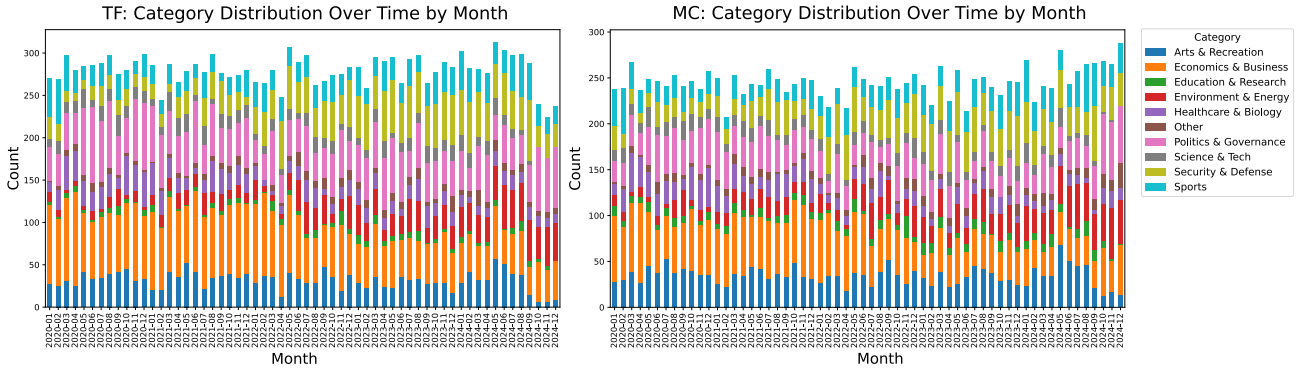

Figure 8. Question category distribution over time.

B. Experiment Details

B.1. Baseline Models Information

Table 5 lists the LLM model versions used in our experiments.

Table 5. Baseline model versions.

Model	Model Version
Claude-3.5-Sonnet	claude-3-5-sonnet-20240620
GPT-4	gpt-4-1106-preview
GPT-3.5	gpt-3.5-turbo-0125
Mixtral-8x7B	Mixtral-8x7B-Instruct-v0.1
Mistral-7B	Mistral-7B-Instruct-v0.3
Llama-3-8B	Meta-Llama-3-8B-Instruct
Qwen-2-7B	Qwen2-7B-Instruct
Gemma-2-2B	gemma-2-2b-it

B.2. Refusal Rates

Although the models are prompted to provide definitive answers rather than responding like "I cannot predict the future," some models still occasionally refuse to do so. Figures 9(b), 10(b), 11(b) show the refusal rates for the closed-book, constraint open-book, and gold article settings. In closed-book evaluation (Figure 9(b)), we can see that the refusal rates increase throughout the time for Mistral-7B in TF questions and Mixtral-8x7B in both TF and MC questions. Additionally, these two models exhibit notable refusal rates, with approximately 10–30% on TF questions and 1.5–8% on MC questions, resulting in their closed-book performances dropping below the random baseline of 50% in certain months, as shown in Figure 3. In comparison, Qwen-2-7B and Gemma-2-2B show relatively low refusal rates—$<5\%$ for TF and $<2\%$ for MC—while all other models have near-zero refusal rates for TF and $<1\%$ for MC.

The refusal behavior is likely influenced by alignment techniques, which discourage uncertain responses in the post-training stage. Although refusal rates contribute to lower accuracies for certain models, our results show that performance degradation trends persist even when refusals are excluded (Figures 9(a), 10(a), 11(a)). We consider refusal to answer an indicator of performance limitations in forecasting tasks, as it reflects the model's lack of actionable knowledge. When models are supplied with more up-to-date and relevant information, their refusal rates decrease (Figure 10(b)). This suggests that refusal is one example of the broader challenge of temporal generalization and reinforces the need for continual model updates or improved external knowledge integration.

B.3. Results for GPT-3.5 in the Gold Article Setting

To more effectively illustrate the trends of other models at a suitable scale, we display GPT-3.5's performance in the gold article setting separately. As shown in Figure 12, this outdated model performs relatively poorly throughout. While its accuracy could improve with chain-of-thought prompting (Wei et al., 2022), we report its performance using the same prompt format as the other models for consistency in comparison. Nevertheless, the degrading trend can still be observed.

B.4. More Results in the Constraint Open-Book Setting

Figures 13, 14, 15, 16, 17, 18, and 19 show the constrained open-book evaluation results for more models. Similar patterns are observed as discussed in Section 4.2. Specifically, for Claude-3.5-Sonnet, the constrained open-book performance lags behind its closed-book performance, likely because it already has robust representations of world events, suggesting that irrelevant or confounding retrieved information may degrade performance. This highlights the need for more careful RAG design in models that already possess robust world knowledge. GPT-3.5 is not included in the constrained open-book setting due to its unexpectedly poor performance in the gold article setting (Figure 12) and budget limitations. Additionally, due to budget constraints, open-book evaluations of proprietary LLMs (Claude-3.5-Sonnet, GPT-3.5, GPT-4) are conducted only up to September 2024, whereas other LLMs are evaluated through December 2024.

B.5. An Example of Evaluating LLMs Under Different Settings

Figure 20 presents a case study demonstrating how Mixtral-8x7B responds to a question under different experimental settings. The model provides an incorrect answer in the closed-book setting. However, when supplemented with retrieved relevant articles or the gold article, it produces the correct answer.