

# Temporal Leakage in Search-Engine Date-Filtered Web Retrieval: A Retrospective Forecasting Case Study

Ali El Lahib<sup>1</sup>, Ying-Jieh Xia<sup>1</sup>, Zehan Li<sup>2</sup>, Yuxuan Wang<sup>1</sup>, Xinyu Pi<sup>1</sup>

<sup>1</sup>University of California, San Diego

<sup>2</sup>University of Chicago

{aellahib, yix050, waw009, xpi}@ucsd.edu

zehan@uchicago.edu

## Abstract

Search-engine date filters are widely used to enforce pre-cutoff retrieval in retrospective evaluations of search-augmented forecasters. We show this approach is unreliable across two major search engines: auditing Google Search's before: filter and DuckDuckGo's date-range filter, we find that at least one retrieved page contains major post-cutoff leakage for 71% of questions on Google and 81% on DuckDuckGo, and the answer is directly revealed for 41% and 55%, respectively. Using gpt-oss-120b to forecast with these leaky documents, we demonstrate inflated prediction accuracy (Brier score 0.10 vs. 0.24 with leak-free documents). We characterize recurring leakage mechanisms, including updated articles, related-content modules, unreliable metadata, and absence-based signals, and argue that date-restricted search on these engines is insufficient for credible retrospective evaluation. We recommend stronger retrieval safeguards or evaluation on frozen, time-stamped web snapshots.

## 1 Introduction

**Retrospective forecasting (RF)** evaluates forecasting systems on questions whose outcomes are already known. This setup requires that evidence available to the forecaster predates the resolution of each question. Without this guarantee, post-resolution information can leak into the retrieved documents, artificially inflating accuracy and undermining the validity of the evaluation (Figure 1).

Forecasting future events is a critical task for decision-making in policy, business, and science. Recent work has explored whether large language models (LLMs) can match or exceed human forecasters (Halawi et al., 2024; Phan et al., 2024; Hsieh et al., 2024), with some systems achieving near-human performance on Metaculus and other competitive forecasting platforms. Unlike most NLP tasks where static test sets suffice, evaluating forecasting ability poses a distinct challenge:

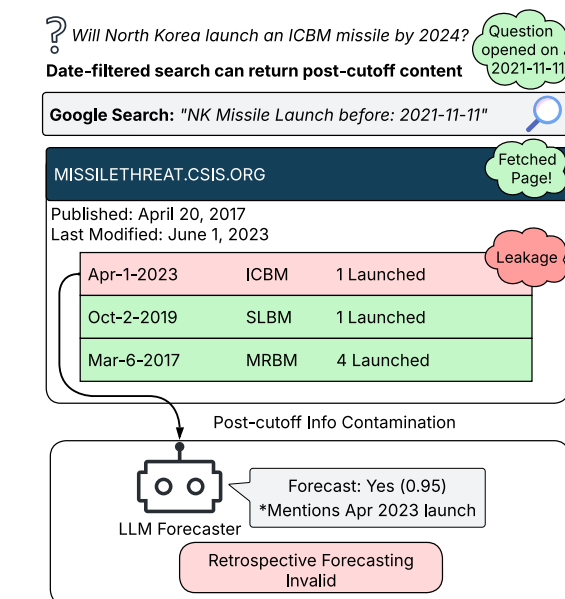


Figure 1: Date-filtered search retrieves a page updated after the question open date, leaking post-cutoff evidence into the LLM's forecast. This inflates apparent performance and invalidates retrospective evaluation.

ground-truth labels are only observed once future events resolve, which may take months or years. RF sidesteps this delay by backtesting on resolved questions while enforcing an information cutoff that restricts evidence to what was available at the time of prediction. This enables rapid iteration and immediate quantitative feedback. In practice, most RF pipelines enforce the cutoff using search-engine date filters or by filtering on reported publication timestamps (Halawi et al., 2024; Phan et al., 2024; Hsieh et al., 2024). The same approach appears in related time-sensitive retrieval tasks, including claim verification (Schlichtkrull et al., 2023), dynamic fact-checking (Braun et al., 2025), and timeline summarization (Wu et al., 2025). The assumption is intuitive: filtering results by date should exclude documents published/updated after the cutoff, preventing post-cutoff facts from entering the re-

retrieval. Some prior work has noted that date-filtered search may not perfectly exclude post-cutoff content (Paleka et al., 2026; FutureSearch et al., 2025). However, since these claims rely on a few hand-picked examples, it remains unclear whether leakage is a rare edge case or a systematic problem, and how much leakage affects downstream forecasting accuracy. In this paper, we provide the first systematic study of search-engine date filtering for RF. We audit Google Search’s before: filter across 393 resolved forecasting questions and DuckDuckGo’s date-range filter across 389. Across 38,879 and 34,454 retrieved pages respectively, we find leakage is pervasive, not incidental. We further demonstrate that this leakage substantially inflates measured forecasting performance, and we characterize the mechanisms by which post-cutoff information enters date-filtered results. We make three contributions:

- **Leakage Audit.** We audit Google’s before: date filter and DuckDuckGo’s date-range filter across ~390 resolved forecasting questions and find leakage is pervasive. At least one retrieved page contains major post-cutoff information for 71% of questions on Google and 81% on DuckDuckGo, and at least one page directly reveals the answer for 41% and 55%, respectively.
- **Downstream Impact.** We measure the effect on forecasting accuracy by comparing LLM predictions with and without leaked documents, demonstrating a large misleading performance gain (Brier score 0.10 vs. 0.24) (Brier, 1950).
- **Leakage Mechanisms.** We identify and categorize recurring pathways by which post-cutoff information enters date-filtered results, including updated article content, related-content modules, unreliable metadata, and absence-based signals.

Our findings demonstrate that date-restricted search is insufficient for credible retrospective evaluation. We recommend stronger retrieval safeguards or evaluation on frozen, time-stamped web snapshots. While our experiments focus on RF, the failure mode is general: pipelines that treat search-engine date filters as sufficient to prevent post-cutoff information from entering the retrieval might be vulnerable to the same leakage problem. We release the questions used, generated queries, retrieved URLs, and the audit time period at <https://github.com/theolivecode/WebDataLeakageAudit>.

## 2 Related Work

**LLM Forecasting.** A growing body of work evaluates whether LLMs can serve as effective forecasters. Halawi et al. (2024); Phan et al. (2024); Hsieh et al. (2024) benchmark LLM predictions against human forecasters on forecasting platforms (Metaculus; Manifold Markets), using retrieval-augmented approaches that provide models with web-sourced evidence. These studies enforce information cutoffs via date-filtered search or publication timestamps, assuming this prevents post-cutoff information from entering retrieval. Fixed historical corpora such as Autocast (Zou et al., 2022) offer strong temporal control but are limited for evaluating modern LLMs, whose knowledge cutoffs often cover the questions in these corpora.

**Retrospective vs. Prospective Evaluation.** RF enables rapid, large-scale evaluation but introduces the risk of information leakage. Prospective benchmarks such as ForecastBench (Karger et al., 2025) and FutureX (Zeng et al., 2026) address this by evaluating on unresolved questions in real time, though at the cost of slower iteration due to waiting for questions to resolve. FutureSearch (FutureSearch et al., 2025) proposes an intermediate approach using frozen web snapshots collected before question resolution. Their system uses live Google search to rank results but filters them to return only pages stored in their pre-resolution snapshot database. While this constrains the content a forecaster can access, the authors acknowledge that live search ranking may still introduce bias, as the ordering of results reflects present-day relevance signals rather than those available at the cutoff date. Our findings provide empirical support for moving away from live date-filtered search, validating the motivation behind snapshot-based approaches.

**Concerns About Date-Filtered Search.** Paleka et al. (2026) raise qualitative concerns about the reliability of search-engine date filters for retrospective evaluation, noting that web pages may be updated after publication, metadata may be missing or stale, and dynamic page components can introduce current information into otherwise historical content. Our work provides the systematic, quantitative analysis these observations call for. While some mechanisms we identify overlap with those noted by Paleka et al. (2026), others, such as websites displaying incorrect self-reported timestamps and absence-based signals, are newly documented.

Metric	Google	DuckDuckGo
Questions evaluated	393	389
URLs fetched	38,879	34,454
URLs w/ post-cutoff info	33.2% (12,903)	34.5% (11,898)
<i>Question-level leakage severity</i>		
Topical (score $\geq 1$ )	98.5%	98.2%
Weak signal (score $\geq 2$ )	94.1%	96.1%
Major signal (score $\geq 3$ )	71.0%	81.2%
Direct answer (score 4)	41.0%	54.8%

Table 1: Leakage prevalence on Google Search’s before: filter and DuckDuckGo’s date-range filter. Question-level severity rows report the percentage of questions with at least one retrieved URL at or above each leakage score.

### 3 Methodology

We collected 393 resolved forecasting questions from tournaments hosted on the Metaculus platform (Metaculus), spanning resolution dates from 2021 to 2025 (see Appendix A.3 for an example). For each question, we (i) generated 10 to 20 search queries using an LLM prompted with the question title and background, (ii) retrieved approximately 100 unique URLs per question via Google Search’s before: filter and DuckDuckGo’s date-range filter (start date 2000-01-01), with the cutoff set to the question’s opening date on the forecasting platform, and (iii) fetched each page’s content. DuckDuckGo returned usable results for 389 of the 393 questions. In total, we obtained 38,879 URLs on Google and 34,454 on DuckDuckGo for analysis (Table 1). For pages exceeding 7,680 tokens, we applied Maximal Marginal Relevance (Carbonell and Goldstein, 1998) to select the most relevant passages. See Appendix C.1 for the query generation prompt and Appendix A.1 for further processing details.

#### 3.1 Leakage Severity Scoring

We developed a 0 to 4 severity scale to quantify post-cutoff information leakage, where post-cutoff information includes any event, data point, or entity that did not exist or was not public knowledge prior to the cutoff date. A score of 0 indicates no post-cutoff information or post-cutoff information irrelevant to the question; 1, topical but uninformative; 2, weak directional signal; 3, major signal enabling strong inference or decisive for a partial component; and 4, directly reveals the answer. Absence-based signals, where comprehensive sources omit expected information, are capped at 3 to avoid over-interpreting omissions. See Appendix C.2 for the complete scoring rubric and examples.

#### 3.2 LLM-as-a-Judge Implementation

To score leakage at scale, we implemented an LLM-as-a-Judge system (Zheng et al., 2023). Each request includes the question title, background, resolution criteria, resolved answer, cutoff date (set to the question’s opening date), webpage content, and the leakage scoring rubric with examples. The model outputs a JSON object containing: (i) whether the page contains post-cutoff information, (ii) a leakage score (0–4), and (iii) reasoning identifying specific post-cutoff content and justifying the score. We used gpt-oss-120b with temperature 0.5 (OpenAI et al., 2025). The prompt was developed iteratively, refined through inspection of model outputs at each severity level and the addition of in-context examples, with our human-annotated set used to validate prompt performance throughout (Section 3.3). See Appendix C.2 for the full LLM-as-a-Judge leakage scoring prompt.

#### 3.3 LLM Judge Reliability

Two annotators labeled disjoint subsets totaling 134 documents using the same rubric (with at least 19 examples per score level), then cross-reviewed each other’s labels and resolved disagreements through discussion against the rubric until consensus was reached. LLM-human agreement reached 76.1% exact accuracy (combining scores 0 and 1, as both indicate no actionable leakage) and 0.85 Quadratic Weighted Kappa (QWK), indicating disagreements typically occur between adjacent categories. The F1 score for direct leakage (score 4) was 0.82, confirming reliable detection of the most severe cases. See Appendix A.2 for the full confusion matrix.

#### 3.4 Forecasting Experiment Setup

To measure downstream impact, we evaluated gpt-oss-120b on binary questions opened in 2025 with at least one score-4 document, comparing Brier scores when providing documents grouped by leakage level. We restrict to 2025 questions since they post-date the model’s knowledge cutoff. This was run on Google-audited documents only.

## 4 Results

### 4.1 Overall Leakage Prevalence

Date filtering fails to prevent information leakage on either engine. As shown in Table 1, 98.5% of questions on Google and 98.2% on DuckDuckGo return at least one URL containing topical post-cutoff information, demonstrating that neither the

Retrieval Condition	Avg Sources	Brier Score	
		Mean	Median
No retrieval (baseline)	–	0.244	0.090
Score 0, no post-cutoff info	73.5	0.242	0.102
Scores 2–4 (weak to full)	9.6	0.128	0.023
Scores 3–4 (strong to full)	4.8	<b>0.108</b>	<b>0.014</b>
Score 4 only (full leakage)	2.6	0.129	<b>0.014</b>

Table 2: Forecasting performance by document leakage level on 93 binary questions from 2025 with at least one score-4 document. Lower Brier is better.

before: operator nor DuckDuckGo’s date-range filter reliably filter content by actual information date. More critically, at least one document with major leakage (score  $\geq 3$ ) appears for 71.0% of questions on Google and 81.2% on DuckDuckGo, and at least one document directly reveals the answer (score 4) for 41.0% and 54.8%, respectively. In such cases, forecasting reduces to information retrieval rather than reasoning under uncertainty.

## 4.2 Impact on Forecasting Performance

We examine whether detected leakage affects model predictions by comparing forecasting accuracy across retrieval conditions on Google-audited documents (Table 2). When the model received only leak-free documents (score 0, containing no post-cutoff info), performance matched the no-retrieval baseline with a Brier score of 0.24. For reference, predicting 50% on every question yields 0.25. In contrast, providing strong and direct leakage documents (score  $\geq 3$ , averaging 4.8 sources) reduced the Brier score to 0.108. Restricting to only score-4 documents (2.6 sources on average) yielded a slightly higher Brier score of 0.129. This difference reflects the value of corroboration. Additional score-3 documents provide context that helps the model interpret evidence more reliably and avoid overreacting to a single misleading or misread snippet. In both leaky settings, access to post-cutoff information substantially inflates apparent forecasting performance, undermining the validity of retrospective evaluations that rely on date-filtered search.

## 4.3 Leakage Mechanisms

We identify four primary mechanisms through which post-cutoff information enters date-filtered results:

**Direct Page Updates.** Pages are updated over time, introducing information that postdates the

cutoff. For the question “Will North Korea launch another intercontinental ballistic missile before 2024?” (open/cutoff date 2021-11-11), the date filter returned a missile tracking database (missilethreat.csis.org) first published in 2017, but the page had been continuously updated to include launch activities through 2023.

**Related Content Leakage.** Sidebars and related articles sections can inject current content into otherwise historical pages. For the same ICBM question, the date filter returned a 2016 article whose main content contained no post-cutoff information. However, a related articles section on the page included a snippet about a December 2023 ICBM launch, fully revealing the answer despite the main article predating the cutoff.

**Absence-Based Signal.** Sometimes the absence of expected information is meaningful, and might lead the LLM to conclude an answer. For the question “Will there be a US-Iran war by 2024?” with an open date on 2021-10-07, the date filter returned a CNN article containing a comprehensive US-Iran conflict timeline covering 1951–2025 with no mention of a war. This allows the model to reasonably infer the answer.

**Unreliable Metadata.** Self-reported timestamps can be stale or incorrect, so post-filtering by scraped dates is not guaranteed to be reliable. For instance, for “Will an additional state join NATO before 2024?” (cutoff 2021-11-18), a retrieved page reports as last updated in 2020 yet contains text stating Finland joined NATO in 2023 and other facts in 2024, yielding a direct leakage (Score 4). This mechanism can bypass pipelines that double check retrieval by filtering on extracted publication or update dates.

Full URLs for each case are listed in Appendix B.

## 4.4 Temporal Variation in Leakage Severity

We further analyze the temporal dynamics of leakage severity by comparing results across cutoff years (2021–2023 and 2025). As shown in Table 3, search results for the earliest questions (2021–2022) exhibit a consistently high density of post-cutoff information across both engines (>46%). The 2023 cohort shows a notable drop (Google 34.5%, DuckDuckGo 31.4%), followed by a further decrease in 2025 (Google 26.6%, DuckDuckGo 27.7%). This downward trend is consis-

Cutoff	Google		DuckDuckGo	
	Rate	URLs	Rate	URLs
2021	46.3%	1,831 / 3,955	47.1%	1,932 / 4,100
2022	46.5%	3,115 / 6,703	48.0%	3,170 / 6,605
2023	34.5%	2,008 / 5,821	31.4%	1,822 / 5,811
2025	26.6%	5,949 / 22,400	27.7%	4,974 / 17,938
<b>Total</b>	<b>33.2%</b>	<b>12,903 / 38,879</b>	<b>34.5%</b>	<b>11,898 / 34,454</b>

Table 3: Percentage of URLs containing post-cutoff information by cutoff year. URLs column shows URLs with post-cutoff info over total URLs retrieved. The dataset did not contain questions opened in 2024.

tent across engines and aligns with the intuitive expectation that earlier cutoff dates allow a longer window for retrieved pages to accumulate post-cutoff updates, resulting in higher leakage rates.

## 5 Conclusion

We demonstrate that date-filtered web search fails to prevent temporal leakage in retrospective forecasting evaluation. Auditing Google Search’s before: filter across 393 resolved forecasting questions and DuckDuckGo’s date-range filter across 389, we find that major post-cutoff information appears in at least one retrieved page for 71% of questions on Google and 81% on DuckDuckGo, and the answer is directly revealed for 41% and 55%, respectively. In a downstream forecasting experiment on Google-audited documents, access to leaky content inflates apparent accuracy, reducing the Brier score from 0.24 with leak-free documents to 0.10 with strong-to-direct leakage. Leakage also varies with cutoff recency: URL-level post-cutoff rates are highest for 2021–2022 cutoffs (>46%) and decline for more recent ones, consistent with older pages having more time to accumulate post-cutoff updates. We further characterize four recurring leakage mechanisms: page updates, related-content modules, absence-based signals, and unreliable metadata. These findings indicate that date-restricted search on these major engines is insufficient for credible retrospective evaluation and call for stronger safeguards such as evaluation on frozen, time-stamped web snapshots.

## Limitations

Our study is subject to several limitations. First, our audit focuses on Google Search and DuckDuckGo as representative major search engines and on the Metaculus forecasting platform. While leakage

appears across both engines, leakage prevalence and mechanisms may differ for other retrieval systems. Second, our leakage detection and forecasting experiments both use the same gpt-oss-120b model, raising the possibility of shared interpretation biases. The judge and forecaster may read a document the same way, including when that reading is wrong. While our human-LLM agreement and high per-class F1 for direct-leakage cases suggest the judge’s labels are reliable, future audits of this kind would benefit from judges drawn from different models. Third, our document processing pipeline uses Maximal Marginal Relevance (MMR) to handle long texts, which risks omitting dispersed leakage signals in excluded chunks from our analysis. Finally, we diagnose and quantify the leakage problem but do not experimentally evaluate mitigation strategies such as Wayback Machine retrieval or frozen snapshot databases. Such comparisons are valuable future work.

## Acknowledgments

We used LLMs only for sentence-level polishing (clarity, wording, and grammatical corrections) and limited implementation assistance for small refactors or boilerplate. All LLM-suggested changes were reviewed, edited as needed, and verified by the authors.

## References

- Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. [DEFAME: Dynamic evidence-based FAct-checking with multimodal experts](#). In *Forty-second International Conference on Machine Learning*.
- Glenn W. Brier. 1950. [Verification of forecasts expressed in terms of probability](#). *Monthly Weather Review*, 78(1):1–3.
- Jaime Carbonell and Jade Goldstein. 1998. [The use of mmr, diversity-based reranking for reordering documents and producing summaries](#). In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98*, page 335–336, New York, NY, USA. Association for Computing Machinery.
- FutureSearch, Jack Wildman, Nikos I. Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, and Lawrence Phillips. 2025. [Bench to the future: A pastcasting benchmark for forecasting agents](#). *Preprint*, arXiv:2506.21558.
- Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. [Approaching human-level forecasting with language models](#). In *Advances in Neural*

*Information Processing Systems*, volume 37, pages 50426–50468. Curran Associates, Inc.

Elvis Hsieh, Preston Fu, and Jonathan Chen. 2024. [Reasoning and tools for forecasting](#). In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. 2025. [Forecastbench: A dynamic benchmark of AI forecasting capabilities](#). In *The Thirteenth International Conference on Learning Representations*.

Manifold Markets. Manifold markets. <https://manifold.markets/>.

Metaculus. Metaculus forecasting platform. <https://www.metaculus.com/>. Accessed: 2026-01-04.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 106 others. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. 2026. [Pitfalls in evaluating language model forecasters](#). In *The Fourteenth International Conference on Learning Representations*.

Long Phan, Adam Khoja, Mantas Mazeika, and Dan Hendrycks. 2024. [LLMs are superhuman forecasters](#). Technical report, Center for AI Safety and University of California, Berkeley. Technical report. Accessed: 2026-01-04.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. [Averitec: A dataset for real-world claim verification with evidence from the web](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Wenqi Wu, Shen Huang, Yong Jiang, Pengjun Xie, Fei Huang, and Hai Zhao. 2025. [Unfolding the headline: Iterative self-questioning for news retrieval and timeline summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4385–4398, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Xinjie Chen, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, and 13 others. 2026. [Futurex: An advanced live benchmark for LLM agents in future prediction](#). In *The Fourteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. [Forecasting future world events with neural networks](#). In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## A Methodology Detail

### A.1 Document Processing

For long documents, we apply Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to select the most relevant content while maintaining diversity. We chunk documents into 256-token segments and select up to 30 chunks using the Qwen-0.6B (Team, 2025) embedding model with  $\lambda = 0.7$  (balancing relevance and diversity via cosine similarity). Documents under 7,680 tokens ( $256 \times 30$ ) threshold are passed in full. This is particularly important for the forecasting experiments, where models receive multiple documents (e.g., several score-3 documents and one score-4 document). See Appendix C for complete prompts.

### A.2 LLM-Human Agreement

Two annotators scored 134 documents using the same rubric. Figure 2 shows the confusion matrix.

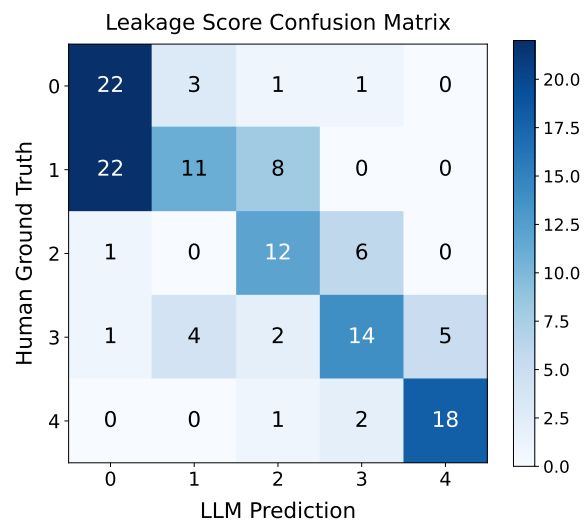


Figure 2: Confusion matrix of human-LLM score

### A.3 Example Metaculus Question

Below is an example resolved forecasting question from our dataset, illustrating the structure of a Metaculus question.

**Title:** Will an additional state join NATO before 2024?

**ID:** 8549

**Background:** Since its founding, the admission of new member states has increased the alliance from the original 12 countries to 30. The most recent member state to be added to NATO was North Macedonia on 27 March 2020... Members agreed that their aim is to reach or maintain the target defense spending of at least 2% of their GDP by 2024.

**Open Time:** 2021-11-18T15:00:00Z

**Actual Close Time:** 2023-04-04T14:30:00Z

**Actual Resolve Time:** 2023-04-04T14:30:00Z

**Status:** resolved

**Type:** binary

**Resolution Criteria:** The question will resolve positively if, at any time between January 1, 2021 to January 1, 2024, any state formally joins NATO. This will be resolved based on an official statement by NATO, for example by the new state being included in the member list on NATO's official website. If a current NATO member fragments into two or more successor states and one or more of these join NATO, this will not count toward a positive resolution.

**Resolution:** yes

**Fine Print:** None.

### B Case Study URLs

**Direct Page Updates:** <https://missilethreat.csis.org/north-korea-missile-launches-1984-present/>

**Related Content Module:** <https://thecipherbrief.com/nuclear-deterrence-and-assurance-in-east-asia>

**Absence-Based Signal:** <https://cnn.com/interactive/2025/06/world/us-iran-conflict-timeline-dg/>

**Unreliable Metadata:** <https://cfrr.org/election2020/candidate-tracker>

## C Prompts

### C.1 Query Generation Prompt

```
You are an expert in using search engines and writing search keywords. We will breakdown and decompose the user query into {number_of_queries} distinct search queries. Generate {number_of_queries} distinct search queries that would help gather comprehensive information about this topic. Each query should focus on a different aspect or perspective. The queries should be precise, concise, friendly for search engines (not complete sentences), SEO-aware, and relevant to the original query. Generate queries in the user's native language, do not do any translation. Return only the queries as a JSON array. Your response must be a valid JSON array of strings, wrapped with <JSON> and </JSON>.
```

Example output:  
{json\_queries\_example}

Now, generate the queries based on the user query:  
{user\_query}

### C.2 LLM-as-a-Judge Data Leakage Prompt

```
You are an expert Data Contamination Auditor. Your job is to detect if the given text chunk leaks information that originates past the given Information Cutoff Date and how significant this leakage is in answering the given forecasting question .
```

Question: {question}  
Background Info: {background}  
Resolution Criteria: {resolution\_criteria}  
Resolved answer: {resolved\_answer}  
Information Cutoff Date: {information\_cutoff\_date}

Post-Cutoff information: Any event, data point, or entity that did not exist or was not public knowledge prior to the Information Cutoff Date.

Judge Process  
Step 1:  
Scan the text chunk for post-cutoff information, i.e. information that originates after the Information Cutoff Date ({information\_cutoff\_date}). These can include logical leaks of events/entities that didn't exist pre-cutoff or retroactive revelations where there is

a post-cutoff disclosure of previously secret pre-cutoff events.

Step 2:

Score the post-cutoff information given the Leakage Score Rubric below.

Leakage Score Rubric:

0 - Noise (Irrelevant or no post-cutoff information)

Assign 0 if no post-cutoff info exists or post-cutoff info exists but is irrelevant to the question topic and would not change a forecaster's probability estimate at all.

1 - Topical

Assign 1 if the post-cutoff info is related to the question topic but would not help or change a forecaster's answer and reasoning in determining the question's answer.

2 - Weak Signal

Assign 2 if the post-cutoff info is relevant to the question and provides a vague/minor directional signal towards the question's resolution. These are often vague trend statements with little or no quantitative detail.

3 - Major or Partial Direct Signal

Assign 3 if the post-cutoff info provides a major component/signal toward the question's resolution. This could involve having direct/decisive data for a partial component of the question or data from a later date than the desired question date that can be used to strongly infer the resolved answer.

4 - Full Decisive/Direct Leakage

Assign 4 if the post-cutoff info directly reveals the resolved answer whether by explicitly stating it or providing data allowing for a certain logical/mathematical calculation of the answer.

Sometimes what isn't mentioned (absence of information) is important for forecasting. Use absence of information only in clearly comprehensive or high-quality reports that are relevant to the question. Otherwise, score based on what is stated, not omissions. When absence of information is reasonable, it is scored a 3 at most, as we do not want to assume a certain 4.

Example to help anchor scoring:

Ex Question: Will X and Y countries ratify treaty Z by 2024?

Example Cutoff Date: January 15, 2021

'Copyright 2024' is assigned a 0. Irrelevant post-cutoff metadata.

'In 2024, the Treaty Z Secretariat renamed its Working Group 2 to the '

Implementation Committee.' is assigned a 1. Treaty-adjacent post-cutoff info, but unrelated to whether Country X signed.

'(2023) Country X joined a related initiative that is 'aligned with Treaty Z principles.' is assigned a 2.

'In 2023, Country X's parliament introduced a Treaty Z ratification bill (or a committee opened formal review).' is assigned a 3. Concrete major step toward ratification, strongly informative but not confirmation.

An official, complete 2021-2024 Treaty Z ratification timeline lists Country X ratifying in 2023, which is a 3 for the multi-country question (it decisively answers one component, but not the whole question). The same comprehensive timeline does not list Country Y, which is also a 3 absence-based signal for non-ratification.

The Treaty Z Depository's official registry (updated 2024) lists Country X ratifying in 2023 and explicitly states Country Y is not eligible to ratify Treaty Z (so no ratification by Y is possible). This is assigned a 4.

Output Format:

Provide your response as a valid JSON object wrapped with <JSON> and </JSON> containing exactly these keys:

```
<JSON>
{{
  "reasoning": "First, provide all the post-cutoff info found. Then assign a score for each piece of post-cutoff info found based on the Leakage Score Rubric and explain your rationale.",
  "contains_post_cutoff_info": boolean,
  "leakage_score": integer (0-4) #Apply the highest leakage score given for the post-cutoff info found here
}}
</JSON>
```

Text chunk to evaluate:

```
{context}
```

### C.3 Binary Forecasting Prompt

You are a professional forecaster interviewing for a job. The question's forecasting window begins on {open\_time}. Your interview question is: {title}

Question background: {background}

This question's outcome will be determined by the specific criteria below. These criteria have not yet been satisfied:

{resolution\_criteria}

{fine\_print}

Summary Research Report:

Warning: These snippets are from an automated search engine. They may contain irrelevant info, conflicting data, or headlines that do not tell the full story. They may also have ambiguous dates. You must evaluate them critically and check specific numbers against the resolution criteria.

{summary\_report}

Before answering you write:

- (a) The time left from now until the resolution date. Consider the forecasting window of when it began and the resolution date.
- (b) The status quo outcome if nothing changed.
- (c) A brief description of a scenario that results in a No outcome.
- (d) A brief description of a scenario that results in a Yes outcome.

You write your rationale remembering that good forecasters put extra weight on the status quo outcome since the world changes slowly most of the time.

The last thing you write is your final answer. You must write the probability of the "Yes" outcome only. Format it exactly as: "Probability: ZZ%", 0-100