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

Anonymous ACL submission

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: auditing Google Search with a `before:` filter, 71% of questions return at least one page containing strong post-cutoff leakage, and for 41%, at least one page directly reveals the answer. Using a large language model (LLM), `gpt-oss-120b`, to forecast with these leaky documents, we demonstrate an inflated prediction accuracy (Brier score 0.10 vs. 0.24 with leak-free documents). We characterize common leakage mechanisms, including updated articles, related-content modules, unreliable metadata/-timestamps, and absence-based signals, and argue that date-restricted search is insufficient for temporal evaluation. We recommend stronger retrieval safeguards or evaluation on frozen, time-stamped web snapshots to ensure credible **retrospective forecasting**.

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.

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; Schoenegger et al., 2024; Phan et al., 2024; Hsieh et al., 2024), with some systems achieving near-human performance on competitive forecasting platforms (Metaculus, 2025). Unlike most NLP tasks where static test sets suffice, evaluating forecasting ability poses a distinct challenge: 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; Schoenegger et al., 2024; Phan et al., 2024; Hsieh et al., 2024). The same approach appears in related time-sensitive retrieval tasks, including 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 retrieval. Some prior work has noted that date-filtered search may not perfectly exclude post-

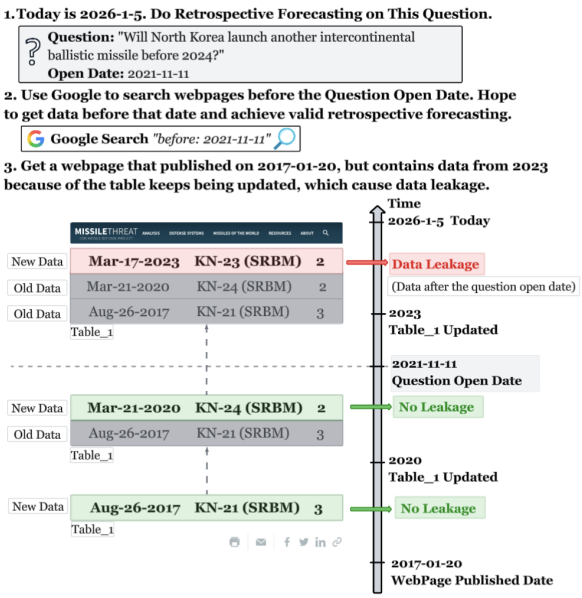

Figure 1: Example of leakage from live page updates. Despite using `before:2021-11-11` (question opened 2021-11-11), a retrieved page includes data from 2023 that reveals the answer and is considered a leakage.

cutoff content (Paleka et al., 2025; FutureSearch et al., 2025). However, since their claim only relies on several hand-picked examples, it is still unclear whether the issue is a rare edge case or a systematic problem, and how much any resulting 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 nearly 39,000 retrieved pages, finding that 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 Search with `before:` date filters and find that 71% of questions return at least one result containing relevant post-cutoff information, and 41% return at least one result that directly reveals the answer.
- **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, misleading self-reported timestamps, 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: pipeline that treats search-engine date filters as sufficient to prevent post-cutoff information from entering the retrieval might be vulnerable to the same leakage problem.

2 Related Works

LLM Forecasting. A growing body of work evaluates whether LLMs can serve as effective forecasters. Halawi et al. (2024); Schoenegger et al. (2024); Phan et al. (2024); Hsieh et al. (2024) benchmark LLM predictions against human forecasters on platforms such as Metaculus (2025), using retrieval-augmented approaches that provide models with web-sourced evidence. These studies typically rely on retrospective evaluation with date-filtered search to enforce information cutoffs.

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., 2025) 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. (2025) 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. (2025), others, such as websites displaying incorrect self-reported timestamps and absence-based signals, are newly documented.

3 Methodology

3.1 Dataset Construction

We collected 393 resolved forecasting questions from tournaments by Metaculus (2025) spanning resolution dates from 2021 to 2025.

For each question, we: *(i)* Generated 10 search queries using an LLM prompted with the question title and background.*(ii)* Retrieved approximately 10 URLs per query using Google Search API with the `before:` operator set to the question's opening date. *(iii)* Scraped full page content for analysis.

Statistic	2021–2024	2025
Questions	169	224
Total URLs	16479	22400

Table 1: Dataset statistics by resolution period.

This yielded 38,879 unique URLs for analysis. See Appendix A.2 for the details of document processing, and Appendix B for the complete prompt of query generation.

3.2 Leakage Severity Scoring

We developed a 0–4 severity scale to quantify leakage: **0** (noise), **1** (topical but uninformative), **2** (weak directional signal), **3** (major signal enabling strong inference), and **4** (direct answer stated). See Appendix D for the complete rubric with examples.

3.3 LLM-as-Judge Implementation

We implemented an LLM-based judge to score leakage at scale. Each scoring request includes the question title, resolution criteria, question open date (used as the search cutoff date), webpage content, and scoring rubric with examples.

The model outputs a JSON object with a score in it. We used `gpt-oss-120b` with a temperature of 0.5 to allow some variability while maintaining consistency (OpenAI et al., 2025).

See Appendix B.2 for the LLM-as-Judge prompt.

3.4 LLM Judge Reliability

For testing human-llm agreement, two human annotators manually scored 134 using the same scoring rubric, with at least 19 examples for each score.

We measured LLM–human agreement with 76.12% accuracy (combined 0 and 1 because they both indicate no related leakage) and 0.852 Quadratic Weighted Kappa (QWK). The high QWK indicates that the LLM judge produces consistent scores, with disagreements typically occurring between adjacent categories. Notice that the f1-score for score 4 is 0.82, which means the model is very reliable at identifying "Direct Leakage".

See Appendix A.1 for the confusion matrix of LLM-human scoring agreement and the LLM reliability metrics.

Score Range	% of questions
Score ≥ 1	98.5%
Score ≥ 2	94.1%
Score ≥ 3	71.0%
Score 4	41.0%

Table 2: Percentage of forecasting questions containing at least one URL with the specified leakage level. Date-filtered search fails catastrophically: 98.5% of questions contain some post-cutoff information, and 41% contain explicit outcome statements.

4 Results

4.1 Overall Leakage Prevalence

Date filtering fails to prevent information leakage. As shown in Table 2, 98.5% of questions contain at least one URL with topical post-cutoff information, demonstrating that the `before:` operator does not reliably filter content by actual information date.

More critically, 71.0% of questions contain major leakage (Score ≥ 3) that can strongly constrain the outcome, and 41.0% contain at least one direct-answer leak (Score $= 4$). In these cases, forecasting becomes more about information retrieval rather than reasoning under uncertainty.

4.2 Leakage Mechanisms

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

Direct Leakage In Main Article. Pages are updated from time to time, thus contains information after the cutoff date. For example, for a question "Will North Korea launch another intercontinental ballistic missile before 2024" with an open date on 2021-11-11, a scraped page tracking database at `missilethreat.csis.org` displays a list of launch activities by North Korea through 2024.

Related Articles Leakage. Web pages inject current content through sidebars and "related articles,". For example, for the same quesiton "Will North Korea launch another intercontinental ballistic missile before 2024" with an open date on 2021-11-11, the main article in the returned page was published in 2016 with no leakage. However, the page has a related article referencing a December 2023 ICBM launch, which fully leaks the answer.

Leakage via Absence of Event. Sometimes the absence of expected information is meaningful, and might lead the LLM to conclude an answer. For

Retrieval Condition	Brier Score	
	Mean	Median
No retrieval (baseline)	0.244	0.090
Score 0, no post-cutoff info	0.242	0.102
Scores 2–4 (weak to full)	0.128	0.023
Scores 3–4 (strong to full)	**0.108**	**0.014**
Score 4 only (full leakage)	0.129	**0.014**

Table 3: Forecasting performance by document leakage level on 93 binary questions from 2025 that each had at least one score-4 retrieved document. Lower Brier is better.

example, for the question "Will there be a US-Iran war by 2024?" with an open date on 2021-10-07, a CNN article contains a comprehensive US-Iran conflict timeline (1951-2025) without mentioning any war having happened between them. This gives the model enough information to reasonably conclude that no US-Iran war by 2024.

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 displays "Last updated May 2020" yet contains text stating Finland joined NATO in 2023 and mentions 2024 events, yielding direct leakage (Score 4). This mechanism can bypass pipelines that double check retrieval by filtering on extracted publication or update dates. See Appendix E.4 for more case studies.

See Appendix E for real-world case studies.

4.3 Impact on Forecasting Performance

Critically, we examine whether detected leakage actually affects model predictions. We compare forecasting accuracy when models are provided with documents at different leakage levels.

The results in Table 3 demonstrate that on questions where the model only saw leak-free documents (contains no post-cutoff information), performance was close to chance overall. The average Brier score was 0.24, indicating poor calibration and accuracy on these mostly binary, non-trivial questions.

When we allowed post-cutoff evidence into the context, performance improved sharply. Providing documents labeled as strong leakage (score ≥ 3) reduced the average Brier score to 0.10. Restricting evidence to only the most extreme leaks (score 4), such as pages that explicitly state the outcome, produced a slightly worse Brier score of 0.12. This

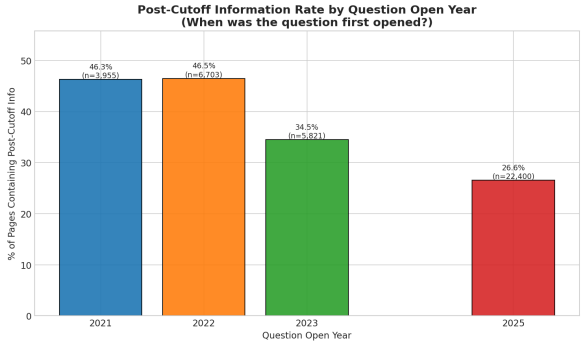

Figure 2: Percentage of pages containing post-cutoff information, stratified by question open year (2021–2025). The question set had no questions open in 2024.

difference is consistent with score-4 leaks being sparse for many questions: additional score-3 documents often provide corroboration and context that helps the model interpret the 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 relative to the leak-free condition.

4.4 Temporal Variation in Leakage Severity

We further analyze the temporal dynamics of leakage severity by comparing results across question open years (2021–2023 and 2025). As illustrated in Figure 2, search results for the earliest questions (2021–2022) exhibit a consistently high density of post-cutoff information (>46%). However, we observe a notable drop in the 2023 cohort (34.5%), followed by a further decrease in 2025 (26.6%). This downward trend aligns with the intuitive expectation that older cutoff dates allow for a longer accumulation window, increasing the probability of contemporary web content being indexed into historical search queries.

5 Conclusion

We demonstrate that date-filtered web search fails to achieve temporal isolation in LLM forecasting evaluation. Through analysis of 393 Metaculus questions and 38879 retrieved URLs, we find pervasive leakage that significantly affects forecasting performance. Our LLM-as-judge methodology, validated against human annotations, provides a scalable approach for auditing retrieval contamination. These findings call for new evaluation paradigms that do not rely on the assumption that date filtering creates clean temporal boundaries.

Limitations

- We only tested Google Search; other search engines may exhibit different behavior.
- Our question set is limited to certain high-salience topics (geopolitics, technology, nuclear policy).
- Our LLM judge was developed on this dataset and may not generalize to all question types.
- We did not evaluate mitigation strategies such as archival retrieval.

Use of LLMs. We wrote the paper ourselves. We used LLMs only for sentence-level polishing (clarity, wording, and grammatical corrections) and limited implementation assistance (i.e., "vibe coding" for small refactors or boilerplate). All LLM-suggested changes were reviewed, edited as needed, and verified by the authors, and the final manuscript and code reflect human author decisions.

References

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. Defame: Dynamic evidence-based fact-checking with multimodal experts. *Preprint*, arXiv:2412.10510.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

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, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. In *NeurIPS*.

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*.

Metaculus. 2025. Metaculus forecasting platform. 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, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. 2025. Pitfalls in evaluating language model forecasters. *Preprint*, arXiv:2506.00723.

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.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. 2024. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Preprint*, arXiv:2402.19379.

Weiqi 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, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, and 12 others. 2025. Futurex: An advanced live benchmark for llm agents in future prediction. *Preprint*, arXiv:2508.11987.

A Methodology Detail

A.1 LLM-Human Agreement

Table 4 shows LLM-human agreement metrics for leakage scoring. Figure 3 shows the confusion matrix of LLM-human agreement on the 134 annotated questions.

Metric	Value
LLM–human agreement (0 and 1 combined)	76.12%
Quadratic Weighted Kappa	0.852

Table 4: LLM reliability metrics for leakage scoring.

A.2 Document Processing

Document Processing. For long documents, we apply Maximal Marginal Relevance (MMR) 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 embedding model with $\lambda = 0.7$ (balancing relevance and diversity via cosine similarity). Documents under 7,680 tokens (256×30) are passed in full. This is particularly important for the forecasting experiments, where models receive multiple documents (e.g., several score-3 documents

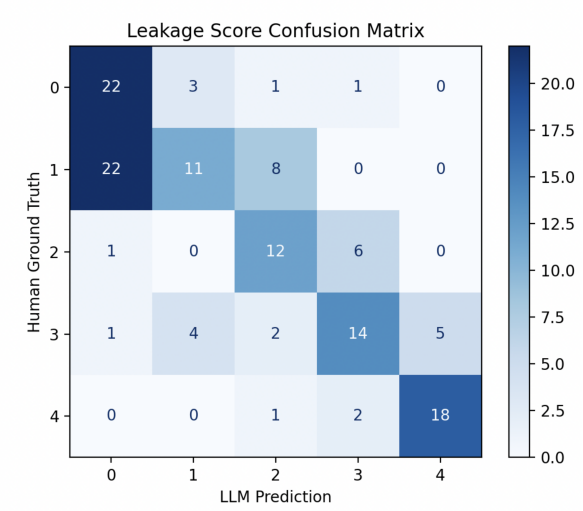

Figure 3: Confusion matrix of human-LLM score

and one score-4 document). See Appendix B for complete prompts.

Score	Precision	Recall	F1-score	Support
0	0.48	0.81	0.60	27
1	0.61	0.27	0.37	41
2	0.50	0.63	0.56	19
3	0.61	0.54	0.57	26
4	0.78	0.86	**0.82**	21
Accuracy			0.57	134
Macro Avg	0.60	0.62	0.58	134
Weighted Avg	0.59	0.57	0.55	134

Quadratic Weighted Kappa: 0.852
LLM–human agreement (0 & 1 combined): 76.12%

Table 5: LLM reliability metrics for leakage scoring. The high f1-score of 4 means the model is reliable at identifying "Direct Leakage."

B Prompts

B.1 Query Generation Prompt

```
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}
```

B.2 LLM-As-Judge Data Leakage Prompt

```
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 Depositary'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}
```

B.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
     ambigious 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
```

C Result Statistics

C.1 Cumulative Leakage

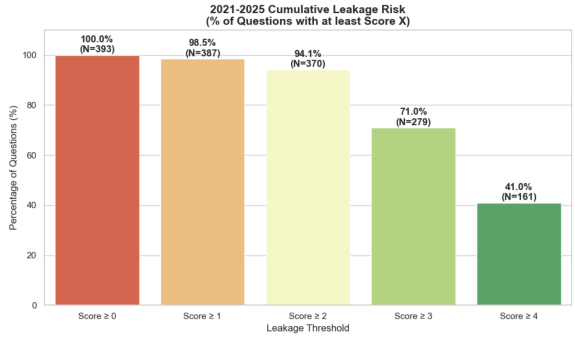

Figure 4: 2021-2025 Cumulative Leakage Risk

C.2 Distribution of Question Max Scores

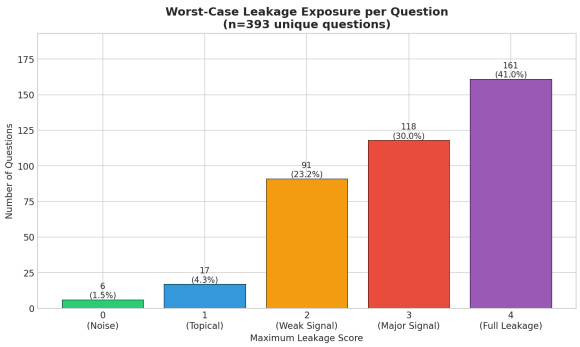

Figure 5: Distribution of Leakage Scores Across All Retrieved Web Content

C.3 Global Leakage Score Distribution

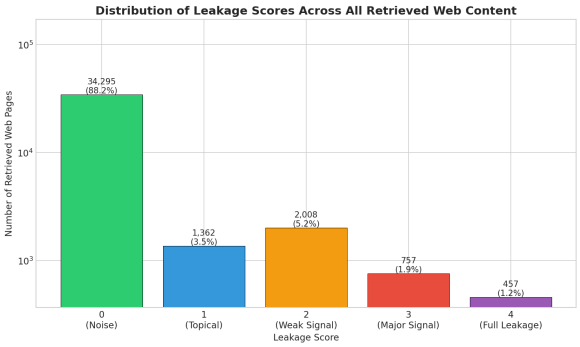

Figure 6: The number of retrieved pages grouped by leakage score.

D Scoring Rubric Details

Table 6 shows the complete scoring rubric.

E Case Studies

E.1 Question Information Example

Table 7 shows an example of the question information from Metaculus.

Score	Description and Example
0	**Noise.** No post-cutoff information, or post-cutoff info exists but is irrelevant to the question. *Example: "Copyright 2024" footer on an otherwise historical page.*
1	**Topical.** Post-cutoff info relates to the question topic but does not help determine the answer. *Example: A 2023 article about missile technology in general, unrelated to the specific launch question.*
2	**Weak Signal.** Post-cutoff info provides a vague directional signal toward resolution, often lacking quantitative detail. *Example: "Country X joined an initiative aligned with Treaty Z principles in 2023."*
3	**Major Signal.** Post-cutoff info provides a major component toward resolution, or data from a later date enabling strong inference. *Example: "In 2023, Country X's parliament introduced a Treaty Z ratification bill."*
4	**Direct Leakage.** Post-cutoff info explicitly states the resolved answer or provides data for certain calculation of the answer. *Example: "North Korea launched an ICBM on December 18, 2023."*

Table 6: Leakage severity scoring rubric. Scores reflect how much post-cutoff information helps answer the forecasting question.

E.2 Case 1: Silent Update (Score 4)

Question: Will North Korea launch another ICBM before 2024?

Cutoff: 2021-11-11

URL: `missilethreat.csis.org/north-korea-missile-launches-1984-present/`

Finding: The page displays a continuously updated table of missile launches, including post-2021 entries, despite appearing in pre-2021 date-filtered results.

E.3 Case 2: Related Articles (Score 4)

Question: Will North Korea launch another ICBM before 2024?

Cutoff: 2021-11-11

URL: `thecipherbrief.com/nuclear-deterrence-and-assurance-in-east-asia`

Finding: Main article from 2016, but a dynamically injected sidebar states: "On December 18, 2023, North Korea successfully launched a solid fuel, road mobile ICBM."

Field	Information
Title	Will the USDA-posted recall of Michael Foods Inc.'s Fair Meadow Foundations Liquid Egg Products issued June 30, 2024 be closed before October 1, 2024?
ID	28244
Background	According to the USDA: "June 30, 2024 – M.G. Waldbaum dba Michael Foods Inc., a Gaylord, Minn. establishment, is recalling approximately 4,620 pounds of liquid egg products due to misbranding and undeclared allergens..."
Open Time	2024-09-17T14:30:00Z
Actual Close Time	2024-09-18T14:30:00Z
Scheduled Resolve Time	2024-10-01T13:05:00Z
Actual Resolve Time	2024-10-01T13:05:00Z
Status	resolved
Type	binary
Resolution Criteria	This question resolves as Yes if the status of the recall posted by the U.S. Department of Agriculture's Food Safety and Inspection Service (FSIS) is changed from Active to Closed...
Resolution	yes
Fine Print	No other resolution source will be considered.

Table 7: Example question and its information.

E.4 Case 3: Unreliable Metadata (Score 4)

Question: Will an additional state join NATO before 2024?

Cutoff: 2021-11-18

URL: cfr.org/election2020/candidate-tracker

Finding: Page metadata claims "last updated 2020" while the content references events from 2023–2024.

E.5 Case 4: Absence of Information (Score 3)

Question: Will there be a US-Iran war by 2024?

Cutoff: 2021-10-07

URL: cnn.com/interactive/2025/06/world/us-iran-conflict-timeline-dg/

Finding: A comprehensive timeline updated through 2025 contains no "war" entry, enabling inference that no war occurred.