Subset	Hypothesis	Decision	p-value	$\bar{d}$
L1	$\mathcal{H}_0^{\leq}: \mu(Z) \leq 0.077$	Rejected	0.0264	0.0835
	$\mathcal{H}_0^{\geq}: \mu(Z) \geq 0.027$	Not rejected	1.0000	
L2	$\mathcal{H}_0^{\leq}: \mu(Z) \leq 0.077$	Not rejected	1.0000	0.0176
	$\mathcal{H}_0^{\geq}: \mu(Z) \geq 0.027$	Rejected	0.0474	

Table 2: Hypothesis-testing results for subsets L1 and L2. $\bar{d} = \bar{x} - \bar{y}$ denotes the mean improvement of *Top CIL* over *without RAG.*

The statistical test results reveal that, under the L1 setting, the null hypothesis $\mathcal{H}_0^{\leq}$ is rejected ($p = 0.0264$), indicating that forecasts constructed from the top-20 CIL scored articles achieve a significantly larger improvement (over 7.7) compared with directly answering the question. In contrast, under the L2 setting, $\mathcal{H}_0^{\leq}$ cannot be rejected ($p = 1.0000$), suggesting that the improvement over direct answers is not statistically significant. This pattern demonstrates that CIL can successfully identify highly supportive news articles. And the construction of PROPHET is valid based on this metric.

4.2 Reasoning Performances

We select the top-10 CIL articles of each question for prediction, and compare to performances without RAG. The results are shown in Figure 4. From the experimental results, it is clear that the Brier Scores of the top-10 CIL selection are significantly better than those achieved without RAG for all tested models. This further demonstrates the effectiveness of the CIL metric in identifying high-quality articles that are capable of boosting forecasting performance.

Moreover, the strong performance observed in the top-10 CIL setting suggests that the degree of answer leakage within the dataset is minimal. If substantial leakage were present, RAG-assisted prediction would not exhibit such improvements over the baseline, as the retrieved content would simply repeat the ground-truth answers rather than genuinely aiding reasoning. Therefore, these results provide additional evidence that our dataset maintains integrity while allowing meaningful performance gains through retrieval.

4.3 Evaluation On Naive RAG Baselines

We evaluate a set of naive Retrieval-Augmented Generation (RAG) baselines over the constructed PROPHET dataset to establish a performance reference and to examine the practical challenges of the task. For the generator component, we select several representative Large Language Models: Claude-sonnet-4, Doubao-1.5, GPT-4o-mini, DeepSeek-v3, and Gemini-2.5-flash. The retrieval component employs seven popular open-source embedding models: all-MiniLM-L6-v2 (AM), msmarco-distilbert-cos-v5 (MDC), msmarco-MiniLM-L6-cos-v5 (MM), msmarco-distilbert-dot-v5 (MDD), msmarco-bert-base-dot-v5 (MBD), instructor-large (IL), and instructor-base (IB). We evaluate each LLM–embedding pair with retrieval sizes $n = 10$ and $n = 20$, reporting Brier Scores (lower is better). The results are shown in Table 3. From the results, we make the following takeaways:

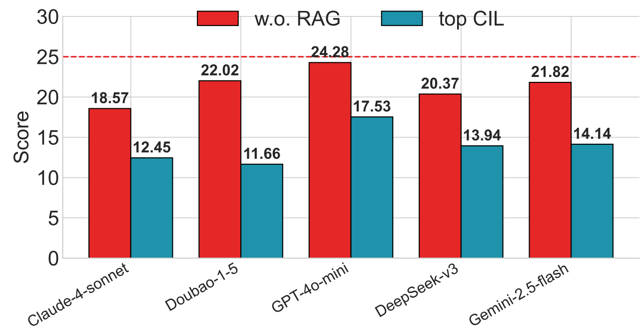

Figure 4: Reasoning ability evaluation. The red line stands for random results which is 25.0.

Limited capability of Naive RAG. Across all evaluated configurations, introducing naive RAG does not consistently outperform the "w.o. RAG" (no retrieval) baseline, and in many cases even leads to performance degradation. This indicates that simply appending retrieved documents to the LLM input, without any filtering, temporal alignment, or causal reasoning, is insufficient for the PROPHET forecasting task. The retrieved context often contains redundant or irrelevant information, and may conflict with the model's internal knowledge, which can confuse the generator and inflate the Brier Score. Moreover, the challenges of identifying truly predictive evidence from historical data suggest that naive RAG lacks mechanisms to reason about event timelines, domain-specific causal links, or uncertainty, all of which are essential for accurate forecasting.
Small differences across current embedding models. Changing among the seven embedding models yields only modest performance variation for a given LLM. This suggests that current general-purpose retrievers are not well suited for temporally and causally grounded forecasting, and that more task-specific embedding models may be needed.
Increasing retrieval size has limited benefit. Increasing from $n = 10$ to $n = 20$ retrieved documents seldom improves performance and sometimes even degrades it. Merely adding more documents can introduce noise and increase reasoning difficulty for the LLM. High-quality selection of truly relevant evidence is more important than raw retrieval quantity.

Overall, naive RAG configurations underperform on PROPHET, and the results highlight the necessity of causally aware retrieval strategies and reasoning methods tailored to forecasting.

4.4 Agentic RAG Performances

We also conduct experiments on Agentic RAG methods. AM and MDC are different embedding models used in the tool. From Table 4, Agentic RAG achieves consistent gains over the w.o. RAG baseline and outperforms naive RAG (Section 4.3) for several LLMs, with models like GPT-4.1, and Gemini-2.5-pro showing Brier Score reductions exceeding -2.0. Unlike naive RAG's simple document concatenation, Agentic RAG allows the LLM to iteratively retrieve, inspect, and integrate evidence, improving temporal and causal relevance while reducing noise. Performance gains are often linked to deeper reasoning, as indicated by higher *Step* counts and longer

	w.o. RAG	AM	MDC	MM	MDD	MBD	IL	IB
Claude-4-sonnet	18.57	19.00/19.49	19.15/18.72	18.12/18.00	19.15/19.64	18.41/18.87	18.24/19.45	19.69/19.14
Doubao-1-5	22.02	20.73/21.01	21.45/21.00	22.72/21.09	20.53/20.25	21.89/21.23	20.33/21.18	21.85/22.05
GPT-4o-mini	24.28	27.17/27.20	28.32/28.92	28.34/27.89	29.41/28.95	29.08/30.34	27.52/28.06	29.10/28.46
DeepSeek-v3	20.37	21.56/21.48	21.94/21.74	22.22/21.04	21.33/22.60	22.10/22.27	21.60/22.96	23.08/22.55
Gemini-2.5-flash	21.82	21.63/21.48	23.69/22.19	22.58/23.63	21.11/22.68	22.06/20.15	21.54/21.16	21.99/22.06

Table 3: Naive RAG on the PROPHET dataset. Lower is better. Each cell reports the score with $n = 10/n = 20$ articles.

	w.o. RAG	AM			MDC		
	BS	BS (δ)	Step	Thought	BS (δ)	Step	Thought
Claude-4-sonnet	18.57	18.65 (+0.08)	3.67	195.65	17.89 (-0.68)	3.58	211.99
Claude-3.5-sonnet	22.22	21.53 (-0.69)	1.05	169.75	21.54 (-0.68)	1.02	202.83
Doubao-1.5	20.20	20.40 (+0.20)	1.02	129.97	24.53 (+4.33)	1.08	166.08
Gemini-25-pr0	21.41	20.51 (-0.90)	0.95	139.64	19.26 (-2.15)	0.95	151.01
GPT-4.1	21.64	21.14 (-0.50)	0.61	91.24	18.44 (-3.20)	0.18	101.94
GPT-4o	22.31	23.33 (+1.02)	1.02	102.64	22.47 (+0.16)	1.04	125.38

Table 4: Performances on Agentic RAG. Lower is better. δ stands for differences between Agentic RAG and w.o. RAG. Thought is the average thought length of each tool call.

Thought lengths. These results highlight Agentic RAG as a promising direction for forecasting, combining adaptive retrieval with structured reasoning to better exploit external knowledge.

4.5 Temporal Analysis

Future forecasting is a continuous process that begins when the question is posed and ends when the question is answered. The earlier the answer can be predicted, the more valuable it is. We investigate the system's forecasting at different times. Similar to Section 3.5, we compute the progress in the whole forecasting. We represent the progress of each news by the percentage of its date in the forecasting. We run different models based on the top-10 CIL articles in various prediction progress. The results are in Figure 5. As the prediction process progresses, the difficulty of prediction decreases, but early predictions still face great difficulties.

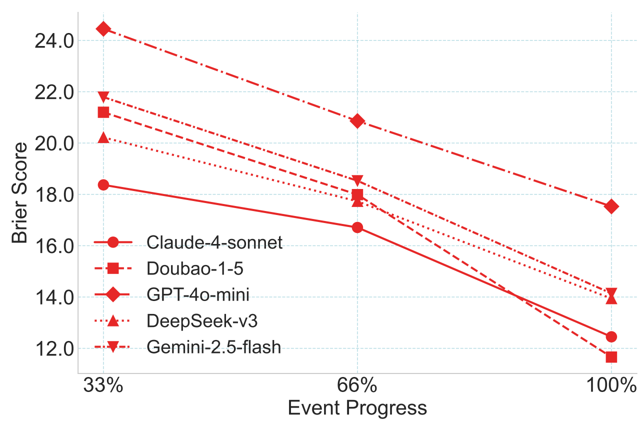

Figure 5: Temporal analysis. Results are on top-10 CIL articles on various forecasting progress.

5 Related Work

5.1 Future Forecasting and Benchmarks

Previous research on future forecasting benchmarks has evolved in different paradigms, each addressing different aspects of the task. Early benchmarks, such as MCNC [7], SCT [19], and CoScript [34], focused on script learning and common sense reasoning in synthetic scenarios. Although these data sets facilitated structured reasoning, they lacked real-world applicability and grounding in factual news. Time series datasets such as GDELT [14] and ICEWS [24] introduced real-world event tracking but did not formalize prediction as a retrieval-augmented reasoning task or ensure answerability. Later works, such as ECARE [4] and EV2 [28], advanced event reasoning has made significant progress in understanding abstract or synthetic scenarios but remains largely confined to settings without real-world grounding, limiting its applicability to practical forecasting or causal inference tasks.

With the rise of LLMs, recent benchmarks such as Halawi et al. [9], OpenEP [8], and ForecastBench [13] shifted the focus to real-world questions and news-based search. However, these datasets suffer from two critical limitations: (1) they lack explicit validation of inferability, allowing questions with insufficient supporting evidence to persist, and (2) they prioritize dynamic data sources over reproducibility, risking inconsistent evaluations due to evolving news archives. PROPHET addresses these gaps by filtering via the introduced Causal Intervened Likelihood estimation.

5.2 RAG and Benchmarks

Foundational QA Datasets for RAG: Traditional QA datasets, including MMLU [10], StrategyQA [6], ASQA [25], Multi-HopQA [16], and 2WikiMultiHopQA [16], are adapted to evaluate RAG systems.

These datasets, grounded in knowledge bases like Wikipedia, form the basis for RAG evaluation.

Domain-Agnostic: RAGBench [5] is a multi-domain benchmark across biomedical, legal, customer support, and finance domains. CRAG [29] provides a factual QA benchmark across five domains, simulating web and knowledge graph search.

Domain-Specific: Domain-specific benchmarks include LegalBench-RAG [30], WeQA [18], PubHealth [36], and MTRAG [27]. These benchmarks address niche applications and improve evaluation precision in domains.

Capability-Oriented: RGB [17] evaluates four RAG capabilities: noise robustness, negative rejection, information integration, and counterfactual robustness. TRIAD [38] assesses retrieval quality, fidelity, and utility through a three-dimensional framework.

In this work, we focus on the inferability of RAG benchmarks, a key property for domain-specific and real-world scenarios. Our method can be generalized to other domains.

6 Conclusion

We address the challenge of building the inferable RAG benchmark for evaluating future forecasting systems by introducing PROPHET. It is rigorously validated for inferability by our Causal Intervened Likelihood (CIL) estimation. By leveraging causal inference to quantify the inferability of prediction questions based on their associated news articles, PROPHET ensures that questions are answerable through retrieved rationales, thereby providing a more accurate assessment of the model capabilities. Experimental validation confirms the effectiveness of CIL in correlating with system performance, while evaluations of state-of-the-art systems on PROPHET reveal key strengths and limitations, particularly in retrieval and reasoning. This work establishes a basis for the development of more nuanced models. With ongoing updating, PROPHET ensures the inferable evaluation in driving progress towards AI-powered forecasting.

References

[1] Nitin Aravind Birur, Tanay Baswa, Divyanshu Kumar, Jatan Loya, Sahil Agarwal, and Prashanth Harshangi. 2024. VERA: Validation and Enhancement for Retrieval Augmented systems. *arXiv preprint arXiv:2409.15364* (2024).

[2] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[3] Lucius EJ Bynum and Kyunghyun Cho. 2024. Language Models as Causal Effect Generators. *arXiv preprint arXiv:2411.08019* (2024).

[4] Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a New Dataset for Exploring Explainable Causal Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 432–446.

[5] Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005* (2024).

[6] Mor Geva, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. StrategyQA: A Question Answering Benchmark Requiring Strategy and Planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online, 5835–5847. doi:10.18653/v1/2021.emnlp-main.466

[7] Granroth-Wilding. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[8] Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2024. OpenEP: Open-Ended Future Event Prediction. *arXiv preprint arXiv:2408.06578* (2024).

[9] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching Human-Level Forecasting with Language Models. *arXiv preprint arXiv:2402.18563* (2024).

[10] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Transactions of the Association for Computational Linguistics* 9 (2021), 479–498. doi:10.1162/tacl_a_00357

[11] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060* (2020).

[12] Elvis Hsieh, Preston Fu, and Jonathan Chen. 2024. Reasoning and tools for human-level forecasting. *arXiv preprint arXiv:2408.12036* (2024).

[13] Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. 2024. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839* (2024).

[14] Kale Leetaru and Philip A Schrodt. 2013. GDELT: Global Data on Events, Location, and Tone, 1979-2012. *The GDELT Project* (2013).

[15] Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024. AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 773–783. https://aclanthology.org/2024.lrec-main.69/

[16] Chin-Yew Lin, Xi Victoria Lin, and Jimmy Lin. 2020. 2WikiMultiHopQA: A Dataset for Multi-Hop Question Answering on Wikipedia. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7380–7391. doi:10.18653/v1/2020.acl-main.654

[17] Nianzu Liu, Tianyi Zhang, and Percy Liang. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, Washington, DC, USA, 17754–17762. doi:10.1609/aaai.v38i16.29728

[18] Rounak Meyur, Hung Phan, Sridevi Wagle, Jan Strube, Mahantesh Halappanavar, Sameera Horawalavithana, Anurag Acharya, and Sai Munikoti. 2024. WeQA: A Benchmark for Retrieval Augmented Generation in Wind Energy Domain. *arXiv preprint arXiv:2408.11800* (2024).

[19] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. 46–51.

[20] Judea Pearl. 2010. An introduction to causal inference. *The international journal of biostatistics* 6, 2 (2010).

[21] Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. 2024. Can Language Models Use Forecasting Strategies? *arXiv preprint arXiv:2406.04446* (2024).

[22] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[23] Victor Rotaru, Yi Huang, Timmy Li, James Evans, and Ishanu Chattopadhyay. 2022. Event-level prediction of urban crime reveals a signature of enforcement bias in US cities. *Nature human behaviour* 6, 8 (2022), 1056–1068.

[24] Philip A Schrodt, David J Gerner, Peter W Foltz, Moon-Soo Cho, and Young Joon Park. 2012. The Integrated Crisis Early Warning System (ICEWS). *Conflict Management and Peace Science* 29, 4 (2012), 432–450.

[25] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid Questions Meet Long-Form Answers. doi:10.48550/ARXIV.2204.06092

[26] Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

[27] Yixuan Tang and Yi Yang. 2024. MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. *arXiv preprint* (2024). arXiv:2401.15391 https://arxiv.org/abs/2401.15391

[28] Zhengwei Tao, Zhi Jin, Yifan Zhang, Xiancai Chen, Haiyan Zhao, Jia Li, Bing Liang, Chongyang Tao, Qun Liu, and Kam-Fai Wong. 2024. A comprehensive evaluation on event reasoning of large language models. *arXiv preprint arXiv:2404.17513* (2024).

[29] Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2024. CRAG: Corrective Retrieval-Augmented Generation for Robust Knowledge Grounding. *arXiv preprint* (2024). arXiv:2406.04744 https://arxiv.org/abs/2406.04744

[30] Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks.