Are LLMs Prescient? A Continuous Evaluation using Daily News as the Oracle

Hui Dai, Ryan Teehan, and Mengye Ren
New York University
{hd2584,rst306,mengye}@nyu.edu

Abstract

Existing evaluation benchmarks for Large Language Models (LLMs) quickly become outdated due to model updates and an evolving information landscape. Moreover, they often lack the ability to assess how model performance evolves over time, as they consist of static questions without a temporal dimension. To address these, we propose using future event prediction as a continuous evaluation method to assess LLMs' temporal generalization and forecasting abilities. Our benchmark, Daily Oracle, automatically generates question-answer (QA) pairs from daily news, challenging LLMs to predict "future" events based on pre-training data. Our findings reveal that as pre-training data becomes outdated, LLM performance degrades over time. While Retrieval Augmented Generation (RAG) can enhance prediction accuracy, the degradation persists, highlighting the need for ongoing model updates.

1 Introduction

Traditional Large Language Model (LLM) benchmarks are often static, and do not reflect real-world information that evolves over time. This presents two significant challenges. First, as LLMs are updated, there is a risk that the static benchmarks become outdated and more vulnerable to data leakage, where their content might end up in the training data of newer models. This undermines the reliability of performance assessments on these benchmarks [1–4]. Second, the static benchmarks often lack the temporal information to track the model's performance variations over time [3]. This creates a need for evaluation methods that stay relevant over time and incorporate temporal dynamics.

Daily news provides a natural setting for continuous evaluation of LLMs. Since the world is constantly changing, a benchmark designed around forecasting the next day's news will never be out of date by construction. In addition to enabling continuous evaluation, forecasting is itself a longstanding challenge with significant implications across various domains, including healthcare, finance, and policymaking [5–7], and LLMs are increasingly seen as valuable tools for this task [8–10]. While several recent forecasting question-answer (QA) datasets have been developed [11–13], they are limited in either size, scope, or do not continuously keep pace with the rapidly changing world.[1] More critically, the extent to which LLMs' predictive abilities change over time remains underexplored.

We propose a continuous evaluation benchmark Daily Oracle, using automatically generated QA pairs from daily news to assess LLMs' temporal generalization and forecasting abilities. We continuously evaluate various LLMs, both with and without access to a limited archive of news articles. Our experiments reveal that LLMs experience an average performance decline of 20.14% on True/False (TF) questions and 23.26% on Multiple Choice (MC) questions between January 2020 and September 2024, with degradation becoming more pronounced before and after the models' knowledge cutoff dates. Although models utilizing Retrieval Augmented Generation (RAG) [14] can demonstrate

[1] See Appendix A for more related work discussion.

NeurIPS 2024 Workshop on Adaptive Foundation Models.

improved prediction performance, the downward trend persists, suggesting an ongoing challenge in maintaining temporal generalization. Overall, our benchmark highlights the challenges posed by outdated pertaining data in LLMs, and underscores the necessity for continuous model updating to keep up with the constantly evolving stream of real-world information.

2 The Daily Oracle Dataset

2.1 Dataset Overview

We present Daily Oracle, a continuously updated QA benchmark of forecasting questions that are automatically generated from daily news. For our current analysis of LLM performance, we utilize a subset of the data consisting of 16,082 TF questions and 13,906 MC questions, covering a diverse range of forecasting topics, which are generated using daily news articles from January 2020 up until September 2024.[2] However, our QA generation framework is continuous and updates daily.

2.2 Dataset Construction

Data Source. Following Zou et al. [12], we collect a large corpus of news articles from the daily-updated Common Crawl News Dataset [15] with the `news-please` package [16]. We filter for reputable sources – CBS News, CNBC, CNN, Forbes, and NPR. For this study, the static version of news corpus we use consists of 1,216,925 English articles spanning January 2019 to September 2024. This corpus is also used for the constrained open-book evaluation setting in section 3.1.

LLM-based Construction Process. QA pairs are generated from articles published between January 2020 and September 2024.[3] For each day, we select six articles for QA generation: three are chosen randomly, and three are selected from hot topics–due to budget constraints. Details for how hot topics are chosen can be found in Appendix C.1. For each selected article, we then use few-shot prompting with LLM to generate two TF QA pairs and two MC QA pairs.[4]

Inspired by Zhang et al. [13], our QA construction follows four steps, with GPT-3.5 [17] used for steps (1) and (4), and GPT-4 [18] for steps (2) and (3) to ensure high data quality: (1) *Article Summary.* We generate a summary for each article, focusing on new events from the publishing date, instead of opinion articles discussing events from the past. This approach allows us to use the publication date as the resolution date of the generated question. Questions can then be regarded as valid forecasting questions since they are prior to the resolution date. (2) *QA Generation.* After filtering out the articles that do not introduce new events, two TF questions and two MC questions are generated together with the answers per article. To ensure balance in the TF questions, we instruct the LLM to generate the first question with a "Yes" answer and the second with a "No." (3) *Misleading Choices Generation.* For MC, we provide the article, its publishing date, and the QA pair to the LLM, which then generates three misleading choices. (4) *QA Filtering.* In the final step, we prompt the LLM to check seven principles: correctness of answers, non-answerability before the publication date, absence of information leakage, objectivity, inclusion of a clear temporal element, public interest, and non-obviousness of the answer. Each principle is scored with 0, 1, or 2 points, and we selected the questions that received at least 13 points in total. These principles are detailed in Appendix C.2.

3 Experiments

3.1 Experimental Setup

Closed-Book Setting. We evaluate various LLMs on Daily Oracle to assess their understanding of real-world events and temporal generalization abilities. Our evaluation differentiates between two scenarios based on the question's resolution date and model's knowledge cutoff date: (1) *Pre-Knowledge Cutoff Questions:* These questions have resolution dates before the model's knowledge cutoff, testing the model's understanding of past events. (2) *Post-Knowledge Cutoff Questions:* These

[2] See Appendix B for dataset details such as summary statistics, examples, and information usage analysis.

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

[4] See Appendix F for all the prompts we use.

have resolution dates after the knowledge cutoff, requiring models to predict future events and test their forecasting and temporal generalization abilities.

Constrained Open-Book Setting. In addition to a closed-book evaluation, we explore the constrained open-book setting: how access to news articles up to different time cutoffs influences LLM performance using RAG [14]. We introduce the concept of the RAG cutoff (*R-Cutoff*), which limits the latest accessible date for retrieving articles. To prevent the models from leveraging information beyond the resolution date, for any question with a resolution date (d_{res}), the accessible articles span from January 1st, 2019 (the start of our news corpus) up to whichever comes first between the day before the resolution date and the RAG cutoff date ($d_{\text{R-Cutoff}}$). Formally, the accessible date range is $[01/01/2019, \min(d_{\text{res}} - 1, d_{\text{R-Cutoff}}))$. Following prior work [11–13], we employ BM25 [19] as the retriever and select the top 5 articles relevant to each question. We truncate each retrieved article to a maximum length of 512 words. These articles are then incorporated into the input prompts to serve as additional information.

Metrics. Accuracy score is used as the evaluation metric. Though LLMs are tested daily, to show clearer trends, we plot the monthly performance in Figure 1, and apply a 5-month moving average to smooth the curve. We also report yearly averages and average year-over-year (YoY) accuracy change before and after models' knowledge cutoff dates in Table 1. Additionally, despite prompting the models to avoid responses like "I cannot predict the future" and instead provide definitive answers, there are cases where such refusals still occur. The rejection rates are provided in the Appendix D.3, and these cases are counted as incorrect to ensure comparability across model results.

3.2 Main Results

Results for the Closed-Book Setting. Figure 1 and Table 1 demonstrate our primary results for the closed-book setting, revealing a clear degradation in performance over time across all models on both TF and MC datasets. When comparing accuracies from the beginning to the end of the evaluation period, we observe that, on average, the models' performance declines by 20.14% on TF questions (from 64.68% to 51.65%) and by 23.26% on MC questions (from 58.30% to 44.74%). This indicates that while LLMs demonstrate certain abilities to understand real-world events and make predictions, they struggle to maintain these abilities.

Notably, the average YoY accuracy declines provide further insight. Before the knowledge cutoff, the average YoY decline across all models was relatively moderate. However, post-knowledge cutoff, we observe steeper declines in many models, with GPT-4 showing the most drastic drop in MC performance, declining by 18.47%, compared to just 4.23% before the cutoff. This contrast highlights that while LLMs manage to retain a baseline of past knowledge with small degradation, their ability to forecast future events deteriorates much more rapidly as they move beyond their training data, struggling with temporal generalization.

Among different models, Claude-3.5-Sonnet [20] significantly outperforms all others, while GPT-4 excels in MC questions but its performance in TF is not as remarkable as in MC. GPT-3.5, Qwen-2-7B [21] and Llama-3-8B [22] show smaller temporal declines in TF questions than GPT-4 and similar trends in MC. Interestingly, Mistral-7B [23] and Mixtral-8x7B [24] show the most pronounced drops in TF accuracy, with scores falling below the random baseline due to increased answer refusals, as shown in the Appendix D.3. Gemma-2-2B [25] exhibits the most consistent performance with the smallest average YoY decline, likely due to its more recent knowledge cutoff date.

Results for the Constrained Open-Book Setting. We present the constrained open-book setting on Mixtral-8x7B with TF questions and Llama-3-8B with MC questions across different RAG cutoff dates.[5] Figure 2 shows that, for Mixtral-8x7B, as the RAG cutoff dates extend to closer to the resolution dates, we observe a clear improvement in performance, indicating the model benefits from increasingly updated information retrieval. However, there are noticeable performance drops immediately after each RAG cutoff date when compared to providing information up to the day before the resolution date (with the exception of the March 2024 cutoff). This highlights the importance of keeping up-to-date information for optimal RAG performance. Interestingly, RAG does not uniformly enhance performance. Llama-3-8B may perform worse than the closed-book setting when the RAG

[5] Refer to Appendix **??** for results of other models in the constrained open-book setting.

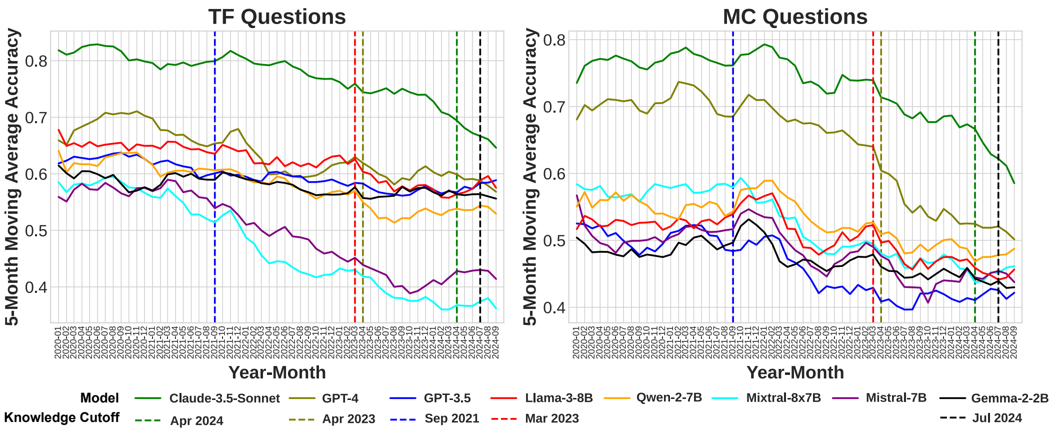


Figure 1: Results for the closed-book setting. We plot the 5-month moving average accuracy for True/False (TF) and Multiple Choice (MC) questions across different LLMs.

Table 1: For various LLMs with different knowledge cutoffs (K-Cutoffs), we show the yearly average accuracy (calculated as the average across months) from 2020 to 2024, along with the average year-over-year (YoY) accuracy change (%) before the knowledge cutoff date (Pre-Cutoff), after the knowledge cutoff date (Post-Cutoff), and the overall average YoY accuracy change across all months (Avg) on the Daily Oracle dataset.

	LLM	**K-Cutoff**	**Average Yearly Accuracy (%)**					**Average YoY Accuracy Change (%)**		
			2020	**2021**	**2022**	**2023**	**2024**	**Pre-Cutoff**	**Post-Cutoff**	**Avg**
TF	Claude-3.5-Sonnet	Apr 2024	81.21	79.88	78.05	74.38	66.25	-4.77	-11.97	-4.47
	GPT-4	Apr 2023	69.68	66.41	60.36	60.54	57.88	-5.83	-1.96	-4.21
	GPT-3.5	Sept 2021	62.86	60.12	59.36	57.11	57.80	-4.33	-3.43	-2.11
	Mixtral-8x7B	Unknown	57.83	52.69	43.09	39.34	36.29	–	–	-10.93
	Mistral-7B	Unknown	57.57	54.65	48.22	41.35	41.89	–	–	-7.67
	Llama-3-8B	Mar 2023	65.06	64.24	62.35	58.68	56.44	-1.95	-6.5	-3.23
	Qwen-2-7B	Unknown	62.39	60.15	57.67	53.39	53.14	–	–	-3.86
	Gemma-2-2B	Jul 2024	58.71	59.31	57.64	56.61	55.87	-1.41	-5.28	-0.97
MC	Claude-3.5-Sonnet	Apr 2024	76.86	77.67	74.32	69.37	61.79	-6.26	-12.82	-4.83
	GPT-4	Apr 2023	70.59	70.62	66.75	56.40	50.96	-4.23	-18.47	-7.48
	GPT-3.5	Sept 2021	50.27	50.36	44.43	41.43	42.32	0.14	-0.46	-4.25
	Mixtral-8x7B	Unknown	57.38	56.97	50.76	47.10	45.09	–	–	-5.37
	Mistral-7B	Unknown	50.07	52.36	48.06	44.40	44.08	–	–	-2.56
	Llama-3-8B	Mar 2023	52.44	54.18	50.66	47.94	45.97	-2.21	-1.25	-3.01
	Qwen-2-7B	Unknown	55.28	55.93	53.44	49.77	47.94	–	–	-3.04
	Gemma-2-2B	Jul 2024	47.87	50.71	46.81	45.20	43.65	-4.46	-2.33	-1.73

cutoff is prior to the knowledge cutoff dates, suggesting outdated information may negatively impact performance. Conversely, for more recent RAG cutoff dates that extend beyond the knowledge cutoff, significant performance improvements are observed (as illustrated by the curves with cutoffs in September 2023 and March 2024). Notably, across all different RAG cutoffs, the overall performance decline pattern persists, likely due to outdated internal representations and the model's inherent knowledge limitations.

3.3 Discussion

LLMs' Performance Evolution Across Time. We observe several LLMs' performance evolution patterns in Figure 1: (1) *Gradual Decline in the Recent Past:* In the months before the knowledge cutoff date, which we call the *recent past*, we observe a gradual decline in model performance, as seen in Llama-3-8B, GPT-4, and Claude-3.5-Sonnet, likely due to a lack of representation of recent news in the training data. (2) *Rapid Decline in the Near Future:* In the *near future*, which we define as the months following a model's knowledge cutoff date, sharp performance drops are observed in several models in MC questions. For instance, the decline in Claude-3.5-Sonnet and GPT-4 accelerates soon

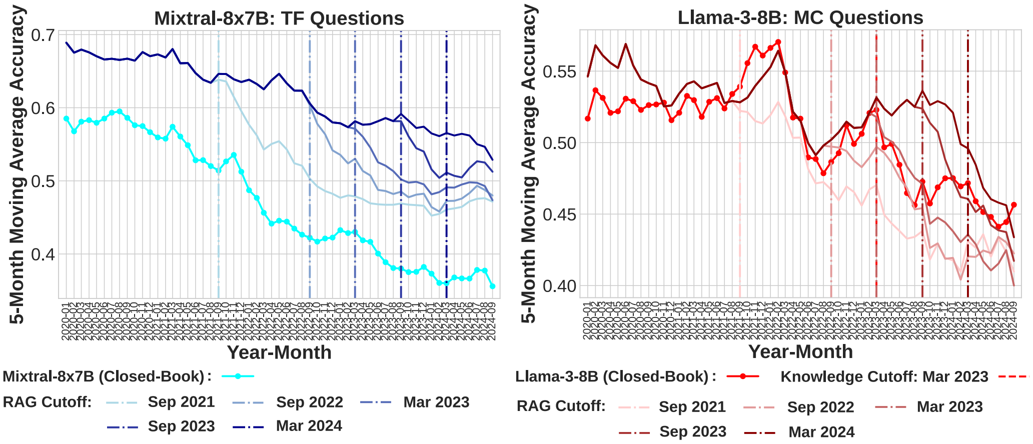


Figure 2: Results for the constrained open-book setting, evaluating Mixtral-8x7B on TF questions and Llama-3-8B on MC questions with different RAG cutoff dates.

after their knowledge cutoffs. Most of the models, however, do not lose all the predictive power at once, as evidenced by the further decline into the farther future.[6]

Consistent Performance Decline After September 2021. Interestingly, Figure 1 reveals a higher rate of performance decline around September 2021, which is the knowledge cutoff date of GPT-3.5, across all models, particularly for MC questions. In contrast, performance remains relatively stable prior to this date. We hypothesize that this trend arises because the period up to September 2021 may be overrepresented in many pretraining corpora [26–33], compared to more recent periods. Another potential cause of this imbalance is an increasing number of websites restricting access to web crawlers [34] after the rise of ChatGPT.

Limitations. Our paper proposes the continuous evaluation benchmark but at the time of the writing there isn't long enough time horizon especially after the cutoff dates on each model for a thorough analysis. Ideally, we would like to analyze the relation between the effect of knowledge and RAG cutoff dates but the trend seems to be weak within the time horizon available. Moreover, the generated questions as well as the distractor answers could contain biases from an outdated LLM, making the benchmark less reliable in the long run unless we upgrade the models. Lastly, running the constrained open-book setting on commercial LLMs is still costly due to the large number of tokens for retrieved articles and we leave it as another future item.

4 Conclusion and Future Work

We introduce Daily Oracle, a continuously updated QA benchmark leveraging daily news to evaluate the temporal generalization and future prediction capabilities of LLMs. Our experiments reveal that while LLMs maintain a degree of predictive power over future events, their prediction accuracy exhibits a gradual decline over time across various models. Notably, while stronger models such as Claude-3.5-Sonnet outperform others significantly, it still exhibits around 12% performance drop in its post-knowledge cutoff period. Although RAG mitigates the effect of outdated knowledge, a noticeable decline in performance remains. Our findings underscore the necessity for ongoing model updates with more current information and emphasize the importance of disentangling missing knowledge from the lack of up-to-date representations.

We hope this work will draw attention to the need for more practical applications of the continuous training of LLMs, driving advancements in adapting models to real-time data changes. In the future, alongside maintaining Daily Oracle, we plan to incorporate a broader range of models and investigate how continuous pre-training and efficient adaptation can address the performance degradation challenges presented in our work.

[6] We provide evidence by analyzing the accuracy slope over time in Appendix E. Before the knowledge cutoff, a slightly negative and stable slope reflects a gradual decline in performance. After the cutoff, a sharply negative slope signals a rapid performance decline.

Acknowledgments and Disclosure of Funding

We thank the Microsoft Accelerating Foundation Models Research program for providing Azure cloud compute credits for the GPT APIs. The compute was also supported by the NYU High Performance Computing resources, services, and staff expertise.

References

[1] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[2] Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

[3] Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*, 2024.

[4] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[5] Walter H. Dempsey, Alexander Moreno, Christy K. Scott, Michael L. Dennis, David H. Gustafson, Susan A. Murphy, and James M. Rehg. iSurvive: An interpretable, event-time prediction model for mHealth. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[6] Kenneth Gillingham, William Nordhaus, David Anthoff, Geoffrey Blanford, Valentina Bosetti, Peter Christensen, Haewon McJeon, and John Reilly. Modeling uncertainty in integrated assessment of climate change: A multimodel comparison. *Journal of the Association of Environmental and Resource Economists*, 5(4):791–826, 2018.

[7] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

[8] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*, 2024.

[9] Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

[10] Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. *arXiv preprint arXiv:2310.01880*, 2023.

[11] Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. ForecastQA: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[12] Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[13] Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. *arXiv preprint arXiv:2406.02472*, 2024.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[15] Sebastian Nagel. Common crawl news dataset. https://data.commoncrawl.org/crawl-data/CC-NEWS/index.html, 2016.

[16] Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, 2017.

[17] OpenAI. New embedding models and api updates, 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

[18] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[19] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[20] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[21] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[22] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[23] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[24] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[25] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[27] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[28] Sosuke Kobayashi. Homemade bookcorpus. https://github.com/soskek/bookcorpus, 2018.

[29] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[30] Yukun Zhu. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*, 2015.

[31] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

[32] Jörg Tiedemann. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[33] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

[34] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. *arXiv preprint arXiv:2407.14933*, 2024.

[35] Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, 2021.

[36] Paul Röttger and Janet Pierrehumbert. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[37] Oshin Agarwal and Ani Nenkova. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 2022.

[38] Chenghao Zhu, Nuo Chen, Yufei Gao, and Benyou Wang. Is your llm outdated? evaluating llms at temporal generalization. *arXiv preprint arXiv:2405.08460*, 2024.

[39] Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[40] Michael Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[41] Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, 2022.

[42] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. Realtime qa: what's the answer right now? In *Advances in Neural Information Processing Systems*, 2024.

[43] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, 2024.

[44] John R. Anderson and Lael J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

[45] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, 2022.

[47] Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*, 2022.

[48] Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[49] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. In *International Conference on Learning Representations*, 2022.

[50] Zixuan Ke, Yijia Shao, Haowei Lin, Hu Xu, Lei Shu, and Bing Liu. Adapting a language model while preserving its general knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[51] Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. Investigating continual pretraining in large language models: Insights and implications. *arXiv preprint arXiv:2402.17400*, 2024.

A Related Work

Temporal Generalization of LLMs. Lazaridou et al. [35] define temporal generalization as the ability of Language Models to generalize well to future data from beyond their training period. They demonstrate that Transformer-XL's performance deteriorates over time, evidenced by increasing perplexity when evaluated on post-training data. However, perplexity-based metrics have two main limitations: they cannot be applied to closed-source models lacking accessible logits, and increased perplexity does not necessarily indicate degraded performance on downstream tasks [36, 37]. Zhu et al. [38] investigate temporal generation using the Bits Per Character (BPC) metric. Similar to perplexity, BPC fails to capture higher-level performance on downstream tasks. In contrast, our work focuses on evaluating how well models acquire and utilize real-world event knowledge in downstream forecasting tasks, providing a more nuanced assessment of temporal generalization.

Dynamic QA Benchmark. Several dynamic QA benchmarks are proposed. Chen et al. [39] construct TimeQA by using time-sensitive facts in WikiData with aligned Wikipedia passages to synthesize QA pairs. Zhang and Choi [40] introduce SituatedQA by manually annotating temporally and geographically dependent questions. StreamingQA [41] and RealtimeQA [42] are both dynamic benchmarks with QA pairs answerable from news articles. StreamingQA, however, does not provide continuous evaluation with always-relevant data. RealTimeQA does not address forecasting and is more like a plugin for a search engine, in the sense that it tests whether a model has updated its knowledge as facts change, rather than testing whether it can predict what will change given its knowledge of the past. FreshQA [43] contains a fixed set of human-written open-ended questions whose answers by nature can change based on new developments in the world, but is smaller and does not address forecasting. It is also updated weekly rather than daily. While all these benchmarks have some form of time-sensitivity like the Daily Oracle, they either do not provide continuous evaluation or do not evaluate forecasting capabilities, or neither.

Forecasting Dataset. Several datasets in the event forecasting field have been introduced. ForecastQA [11] used crowdworkers to collect 10,392 QA pairs from news articles. Zou et al. [12] argue that the QA pairs from ForecastQA are often nonsensical or ambiguous since they are written by humans without forecasting expertise. They further introduce AutoCast, a forecasting dataset from popular human forecasting tournaments containing 6,707 QA pairs. In contrast, our Daily Oracle dataset is generated automatically from daily news articles, which means that it is never out of date, can easily grow its size automatically without additional inputs from human forecasters, and provides more comprehensive event coverage than human forecasting tournaments.

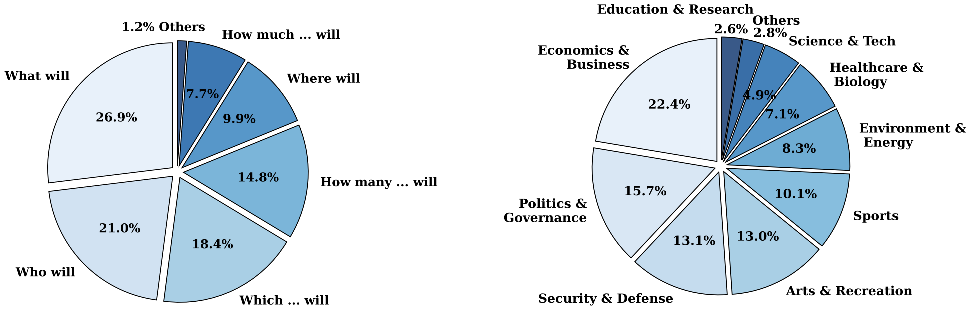
Similar to our generation method, TCELongBench [13] has an automatic forecasting QA generation framework using news articles. However, their TLB-forecast dataset is constrained both temporally and topically, and only contains cooperation and conflict events in Middle-Eastern countries from 2015 to 2022. This restricts the dataset from evaluating more general event-prediction abilities. Furthermore, considering most of the powerful LLMs have been developed after 2020, the portions of the dataset covering earlier years may contain answers already seen during training. This prior exposure compromises the dataset's effectiveness as a forecasting benchmark. In contrast, our dataset spans a broader timeframe and covers more topics, offering a more comprehensive and reliable forecasting benchmark.

Note that none of the aforementioned datasets provide insights into how prediction ability changes over time. Zhu et al. [38] introduce Freshbench, a forecasting dataset scraped from the GoodJudgmentOpen platform, and also study temporal generalization. However, they report accuracy in a relatively short time window (from January 2023 to April 2024) with only 2,532 questions. While we observe a gradual performance decline in our dataset, they report significant fluctuations in model accuracy shortly after release. We argue that our dataset presents a more challenging task, as evidenced by the continuous degradation in model performance over time, making it a robust benchmark for assessing LLMs' temporal generalization.

In order to clearly showcase the differences between our dataset and prior work, we highlight a few key features in Table 2. The Daily Oracle is the only benchmark which is continuously updated on a daily basis and evaluates forecasting ability. Additionally, at the fixed size we use for analysis we provide significantly more evaluation examples than the other automatically updated benchmarks.

Dataset	Continuous?	Interval	Forecast?	Size	Latest Update
TimeQA	✗	None	✗	20,000	2021
SituatedQA	✗	None	✗	4,757	2021
StreamingQA	✗	None	✗	36,800	2021
RealTimeQA	✗	None	✗	1,470	2023
FreshQA	✓	Weekly	✗	600	2024
ForecastQA	✗	None	✓	10,382	2019
AutoCast	✗	None	✓	6,707	2022
TCELongBench	✗	None	✓	88,821	2022
FreshBench	✓	Unknown	✓	2,532	2024
Daily Oracle (Ours)	✓	Daily	✓	29,988	2024

Table 2: We compare Daily Oracle with existing benchmarks in the literature. For continuously updated datasets (e.g. Daily Oracle, FreshBench, and FreshQA), "Interval" refers to the dataset update interval, and "Size" and "Latest Update" refer to the fixed data currently available. Our Daily Oracle benchmark is the only forecasting benchmark which is continuously updated every day using questions generated from daily news.



(a) MC question types for Daily Oracle.

(b) Category Breakdown for Daily Oracle.

Figure 3: Comparison of question types and categories.

B Dataset Details

B.1 Summary Statistics

At the time of writing this paper, the subset dataset we use from Daily Oracle consists of 16,082 TF and 13,906 MC QA pairs, covering the period from January 1st, 2020, to September 30th, 2024, with an average of 17.3 questions per day. Figure 3a shows that our dataset covers various MC question types, mainly starting with "What will" (26.9%), "Who will" (21.0%), and "Which ... will" (18.4%). Figure 3b provides a breakdown of the categories, highlighting our dataset's broad coverage. The categorization of each question is determined using GPT-3.5, based on the prompt from Halawi et al. [8]. Examples of QA pairs are shown in Table 3.

B.2 Dataset Information Usage Analysis

Past and Future Information Usage. Each question in Daily Oracle implicitly requires the model to retrieve relevant knowledge. How do these requirements change day by day over the course of our benchmark? [44] study a similar relationship in human information environments, specifically in *New York Times* headlines, children's verbal interactions, and emails. In Figure 4, we take inspiration from their work and analyze whether a word's frequency of occurrence in the past 100 days predicts its occurrence on the next day. In other words, if over the past 100 days we have frequently required the model to retrieve specific knowledge, e.g. if there are many questions about the unemployment rate, is it likely it will have to retrieve this knowledge in the future?

We analyze this relationship for words in the titles of the articles we use to generate questions as well as in the text of the TF and MC questions themselves. Past frequency is computed by checking,

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

Table 3: Daily Oracle Example Questions and Answers.

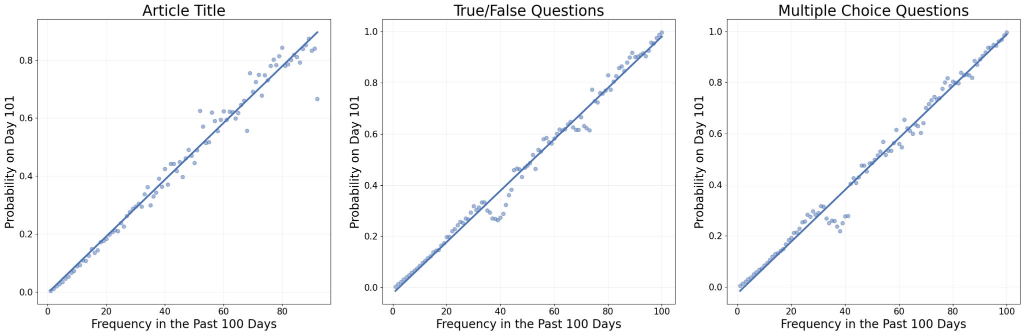


Figure 4: Following Anderson and Schooler [44], we plot the probability of a word occurring in an **(left)** article title, **(middle)** True/False question, or **(right)** Multiple Choice question given how frequently it had appeared in one over the past 100 days, computed over our entire dataset. We fit a linear regression in each case and show a linear relationship in each case ($R^2 = 0.978$, 0.986, and 0.985 for left, middle, and right respectively).

for each day in the 100 day window, if a word has occurred in any article title (so, the maximum frequency is 100). We find that there is a linear relationship between the frequency of usage in the past 100 days and the probability of occurrence on the 101st day in all cases, replicating Anderson & Schooler's findings for *New York Times* headlines. Interestingly, there is a drop in probability for both TF and MC questions for words with a frequency of 40, though it is unclear why. There is also some clustering at lower frequencies, particularly in the TF and MC question plots. Many words appear less than 20 times during the 100 day window. The temporal structure exhibited in the daily news stream may be of a future point of interest from a modeling perspective.

C Dataset Construction Details

Figure 5 shows the flowchart of Daily Oracle's data construction process.

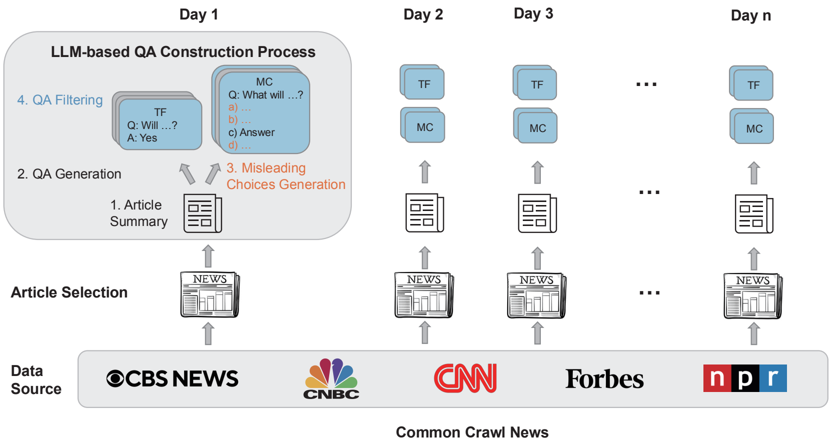


Figure 5: Data Construction Process.

C.1 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 [45] 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.

C.2 QA Filtering Principles

Seven principles of 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.

D Experiment Details

D.1 Baseline Models Information

Table 4 lists the LLM model versions used in our experiments.

D.2 Gold Article Setting

Besides the closed-book setting and the constrained open-book setting, we further include a setting where models are provided direct access to the gold article, from which the question is generated. This transforms the forecasting questions into reading comprehension ones. Achieving high accuracy here ensures that the questions from our Daily Oracle dataset are answerable.

Table 4: 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

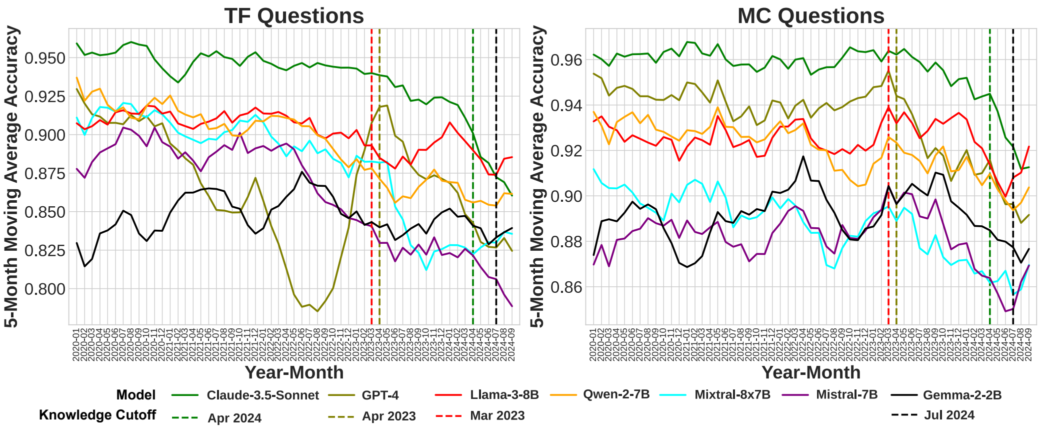


Figure 6: Results for the Gold Article Setting. Most of the models struggle with temporal generalization, even when provided with gold articles containing the answers.

Results for the Gold Article Setting. Figure 6 shows that when given access to the gold articles from which the questions are generated, LLM performance can be improved significantly to around 90%, demonstrating the answerability of Daily Oracle. However, most of the models still show declining trends. This indicates that even when treating forecasting as reading comprehension, LLMs still struggle with temporal generalization—their out-of-date representations prevent them from generating the correct answer even when the up-to-date knowledge from the gold article is provided.

GPT-3.5's performance in the gold setting is shown separately from the other models. As shown in Figure 7, this outdated model performs relatively poorly throughout. While its accuracy could improve with chain-of-thought prompting [46], we report its performance using the same prompt format as the other models for consistency in comparison. Nevertheless, a clear downward trend is observed in MC questions.

D.3 Rejection Rates in the Closed-Book Setting

Figure 8 shows the rejection rates for the closed-book setting. We can see that the rejection rate increases throughout the time for Mistral-7B and Mixtral-8x7B. The high rejection rate of TF in these two models leads to the performance below than the random baseline in Figure 1.

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

Figures 9, 10, 11, 12, 13, and 14 show the constrained open-book evaluation results for more models. Similar patterns are observed as in Section 3.2.

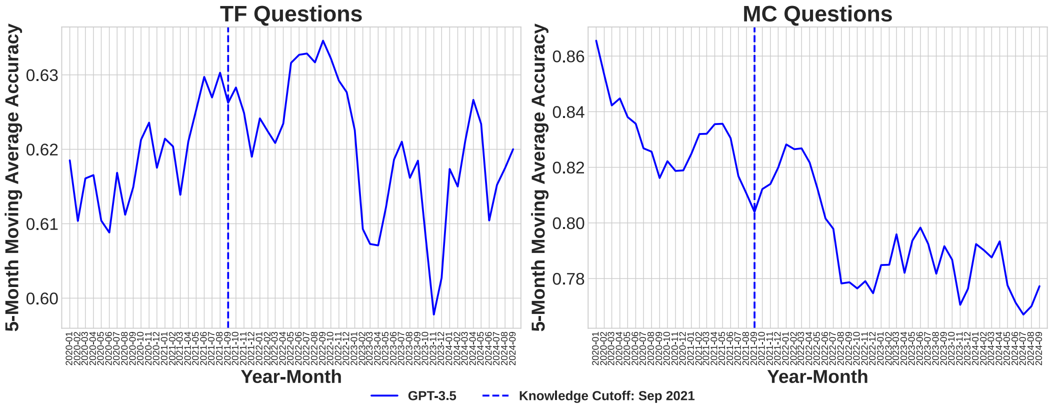


Figure 7: Results for GPT-3.5 in the Gold Article Setting. Compared to other models achieving around 0.9 accuracy, GPT-3.5 performs worse in both MC and, more notably, in TF.

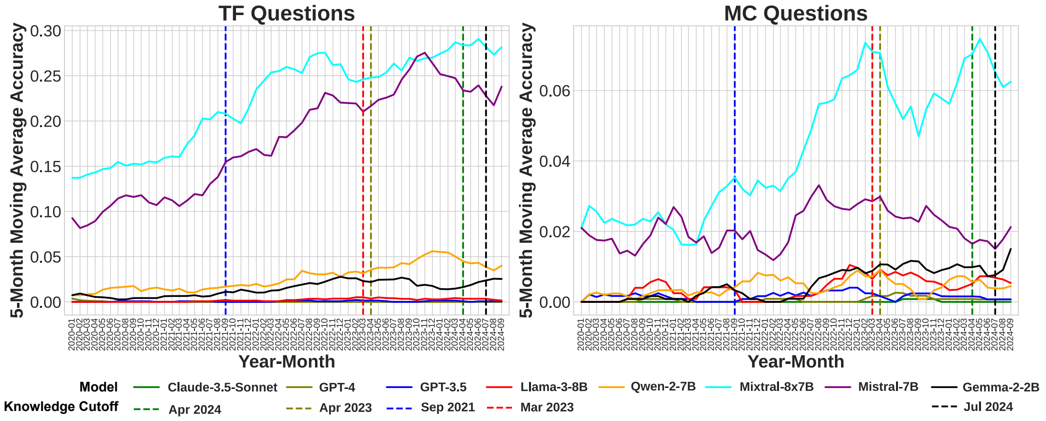


Figure 8: Rejection rates for the closed-book setting. We plot the 5-month moving average rejection rates for True/False (TF) and Multiple Choice (MC) questions across different LLMs.

E Discussion

Regression Coefficient Analysis for LLMs' Performance Evolution. We explore the LLMs' performance evolution patterns further by analyzing the slope of accuracy as a function of time. In Figure 15, we show how the slope changes as we fit a regression to an increasingly larger window of data, until we reach the full set of accuracies. Specifically, using the 5-month moving average of each model's accuracy on MC questions (visualized in Figure 1), we start by fitting a linear regression line on the first 10 months of data. We then add an additional month and compute a new regression on the larger window, repeating until we reach the final month, and applying an exponential decay weighting to past data to reduce the influence of distant observations. With this, we can analyze how the slope of our regression line changes as each month is added to the data. The slope in each case is negative after the cutoff data and for Claude-3.5-Sonnet, GPT-4, and Llama-3-8B, the slope eventually or immediately becomes more negative than it was at any point preceding the cutoff. Both Claude-3.5-Sonnet and Llama-3-8B have a crossover from positive to negative slope in late summer 2022, July and August, respectively, while GPT-4's seems to occur slightly earlier, in March of 2022. For GPT-3.5, GPT-4, and Llama-3-8B, the slope becomes increasingly negative not long after the knowledge cutoff, giving evidence for a rapid decline in the near future. Likewise, the period preceding the cutoff shows a less negative slope and in some cases, most obviously in the case of GPT-3.5 and Llama-3-8B, a slightly negative but consistent slope, in other words, a gradual decline.

Need for Continuous Pretraining. The overall decline trend may come from two sources, the missing knowledge of future and a lack of up-to-date language representation. While the lack of

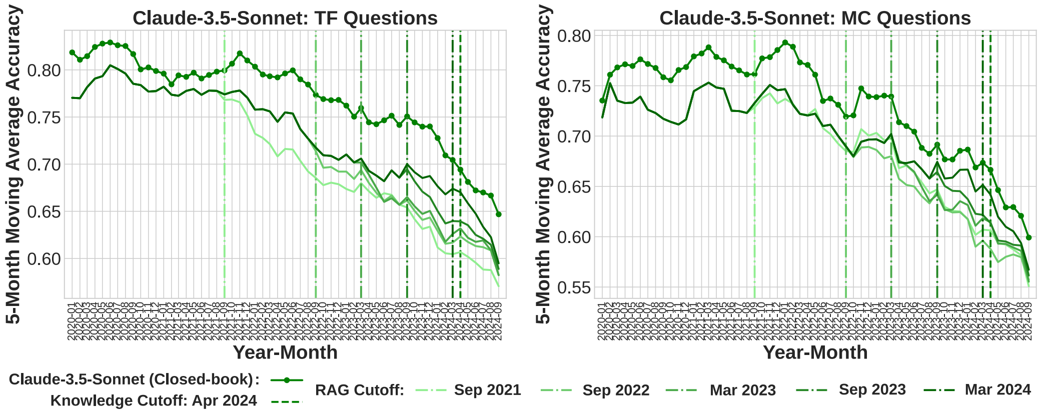


Figure 9: Results for Claude-3.5-Sonnet in the constrained open-book setting.

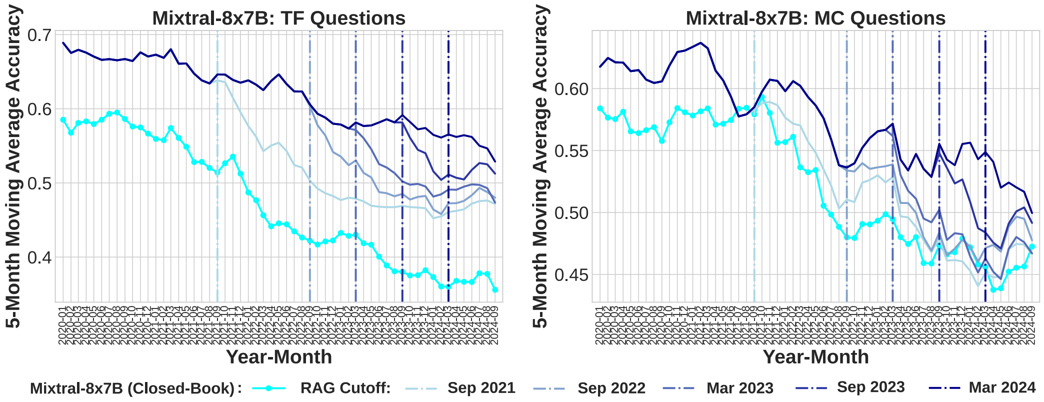


Figure 10: Results for Mixtral-8x7B in the constrained open-book setting.

knowledge can be partially recovered with the constrained open-book setting, the gold article setting provides an "upper bound" of open-book. When provided the gold article, the remaining decline in the model's performance suggests that continuous pre-training of LLMs [47–51] is still needed in the context of news event forecasting.

TF & MC Comparison. All models except for Claude-3.5-Sonnet struggle with TF questions, where the degradation trends towards the random baseline accuracy of 50%, indicating that predicting if a future event will happen or not can be sometimes challenging for LLMs. In contrast, on MC questions, models tend to perform much better than the random baseline at 25%. There are two potential reasons that can explain the disparity. First, TF questions can be considered more open-ended than MC because the "No" answer contains other possible open-ended outcomes. Second, since the distractor choices are created by an LLM, they may not be as likely to happen as the true answer.

F Prompts

We show all the prompts we use in this section. The QA generation prompts and evaluation prompts are adapted from Zhang et al. [13], and the prompt to categorize our generated questions is taken from Halawi et al. [8].

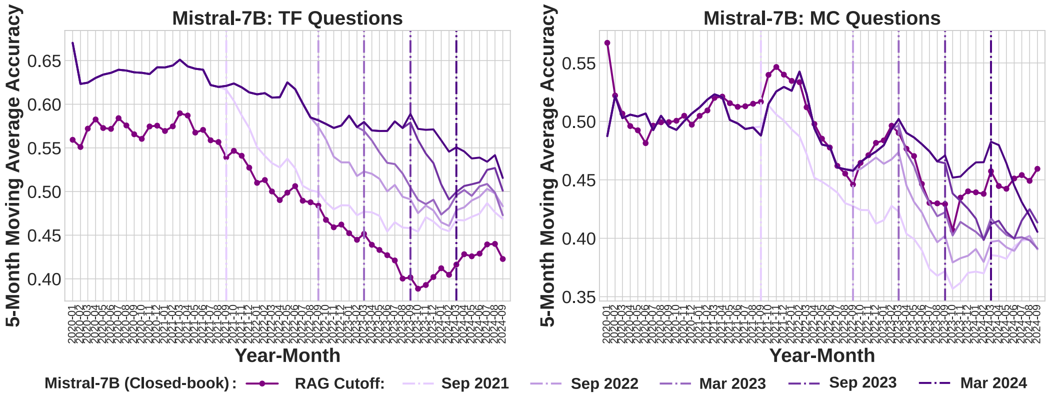


Figure 11: Results for Mistral-7B in the constrained open-book setting.

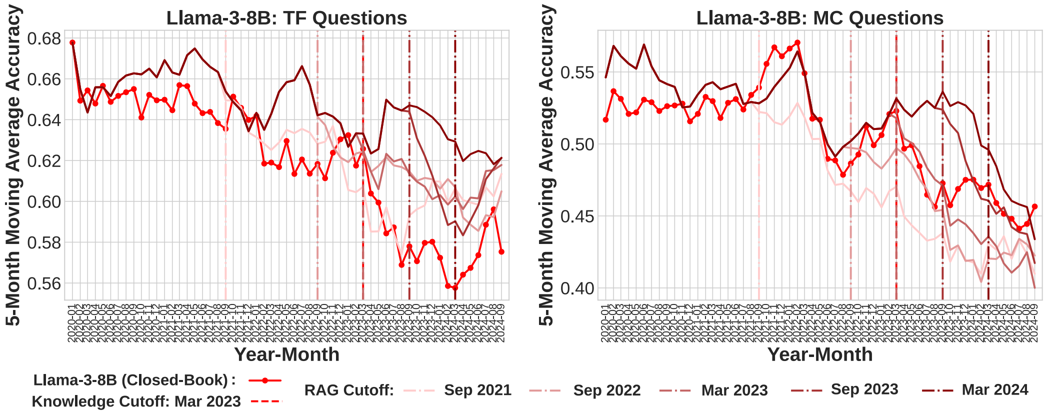


Figure 12: Results for Llama-3-8B in the constrained open-book setting.

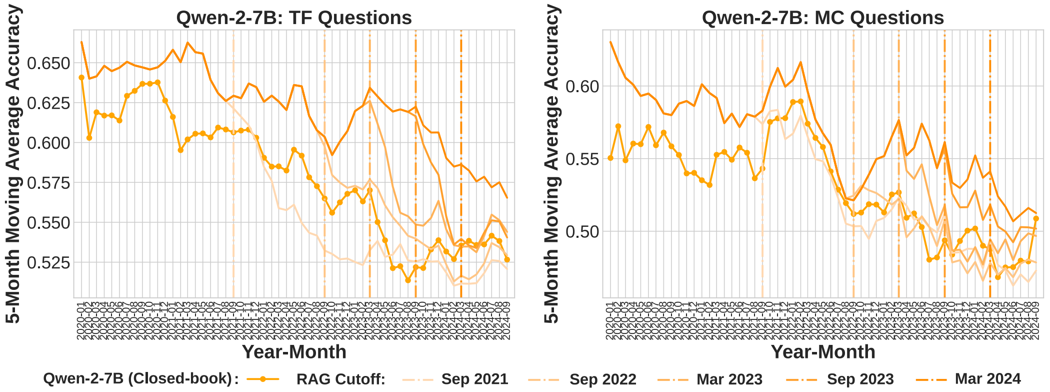


Figure 13: Results for Qwen-2-7B in the constrained open-book setting.

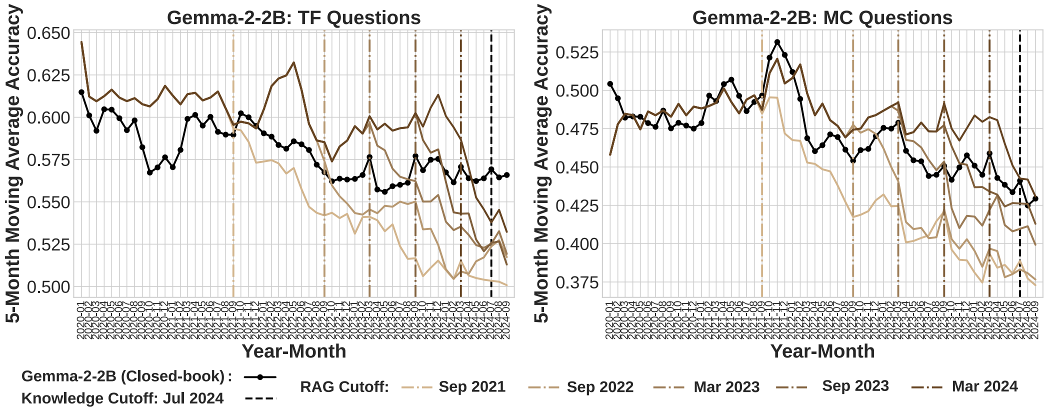
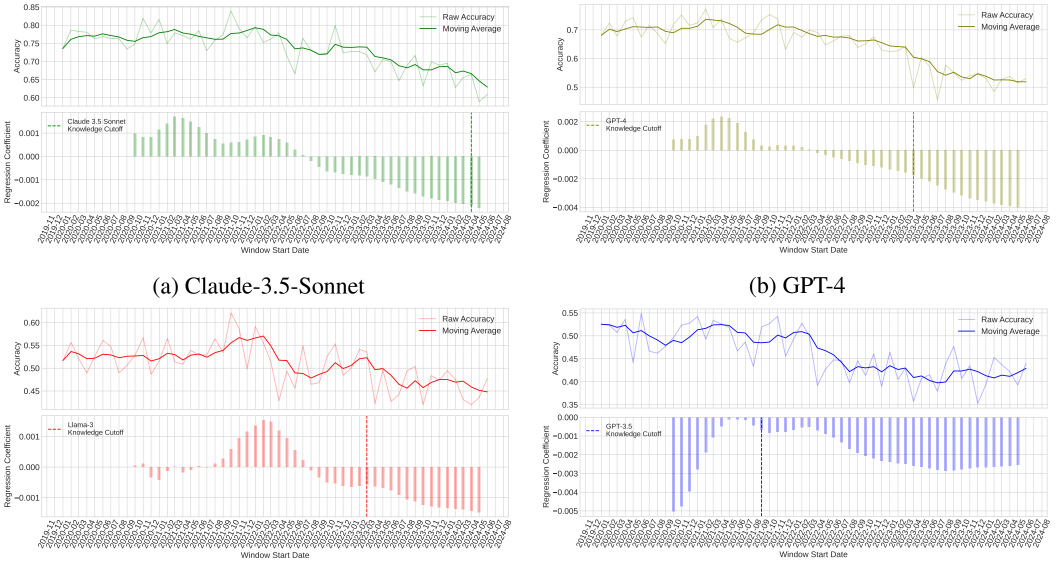


Figure 14: Results for Gemma-2-2B in the constrained open-book setting.



(a) Claude-3.5-Sonnet

(b) GPT-4

(c) Llama-3-8B

(d) GPT-3.5

Figure 15: Coefficients for regressing accuracy on the MC questions against time, as the number of months grows. Using an initial window of 10 months, we progressively add data for additional months to our regression and plot the coefficient (slope) for the regression of accuracy against time. For our regression, we use the moving average of accuracy and apply exponentially decaying weights to older months' accuracies (i.e. given a window of data from k months, we weight x_t with λ^{k-t}; in this case $\lambda = 0.995$). In each case, the slope becomes more negative following the knowledge cutoff period compared to the months immediately preceding it.

```
You are an expert in extracting summary and keypoint from articles.

# Rules
1. Provide a comprehensive summary of the entire article in one paragraph, ensuring that all essential aspects are addressed. Your summary should include key statistics, notable dates, and any significant statements to fully convey the context and content of the news story.
2. Please provide one keypoint that summarizes the new event from the article with the following rules:
- Focus specifically on events that are newly occurring on the publication date of the article. If the article does not introduce a new event but instead discusses ongoing topics or is about non-news content like advertisements, state 'No new event reported.'
- The point should be concise, accurate and complete, especially for numbers, names and dates.
- Basically NO "he, she, they, it, them, etc" are allowed. Please clearly write out the entity you are referencing in the point.
- You are not allowed to start with any of the phrases: the article discusses, the article shows, the article emphasizes, the article discusses, the speaker says, the speaker discusses, the author mentions, etc.

# Examples
Here are several examples of extracting keypoints from articles. Note that the articles in different examples are irrelevant.
## Example 1:
Article: Professional golfer Lexi Thompson has announced her retirement from professional golf at the end of the 2024 season at the age of 29. Thompson, an 11-time LPGA Tour champion, made the announcement ahead of her 18th consecutive US Women's Open appearance. She turned professional in 2010 and won her first major at the 2014 Kraft Nabisco Championship. Despite enduring injuries that led to a drop in her world ranking, Thompson continued to compete at a high level. In her retirement announcement, Thompson expressed gratitude for the opportunities golf provided her and highlighted her excitement for the next chapter of her life.
Publishing Date: 2024-05-29, Wednesday
Keypoint: Lexi Thompson announced her retirement from professional golf at the end of the 2024 season at the age of 29.

## Example 2:
Article: India's capital territory of Delhi sweltered to its highest-ever temperature of 49.9 degrees Celsius (121.8 degrees Fahrenheit) on Tuesday, as an oppressive heat wave forced authorities to impose water rationing.

The Indian Meteorological Department (IMD) said the new record was measured in the suburb of Mungeshpur, surpassing Delhi's previous high of 49.2 degrees Celsius (120.5 degrees Fahrenheit), observed in May 2022.
Publishing Date: 2024-05-29, Wednesday
Keypoint: Delhi experienced its highest-ever temperature of 49.9 degrees Celsius on 2024-05-28, Tuesday.

# New Article
Based on the provided rules and examples, please summarize the article and identify one key point that concludes the new event on the publishing date.
Article: {}
Publishing Date: {}

# Output: Output should follow the format of
Summary:
Keypoint:
```

Figure 16: Prompt in the *Article Summary* step, adapted from Zhang et al. [13].

[Setup:] Today is {}. You will be testing people who are from the past, i.e. a week or a month ago. A recently published article will serve as the basis for your questioning. Your objective is to ask them questions that assess the accuracy and plausibility of their predictions about events.

You will write question-answer pairs:
1. The question should challenge the person to predict an outcome or development related to the article's content as if they were being asked one week or one month ago. Please provide the question that can be answered on {}, but only guessable not answerable before {}.
2. The answer MUST be based on factual information from the article. Ensure that the answers do not predict outcomes that have not been explicitly stated in the article.

[Rules:]
Article: {}.
Publishing date: {}

Please generate four questions about the above article, along with answers. You should follow the instructions below:
1. Please turn the key point "{}" into the question, with focusing more on whether the event will happen.
2. The question should NOT be designed for reading comprehension. Please focus more on what happened rather than the implications after the event.
3. The question MUST be in future tense.
- Start the first question with "Will", with the answer as "Yes".
- Start the second question with "Will", with the answer as "No".
- Start the third and fourth questions with a phrase like "What will", "Who will", "Where will", "Which xxx will", "How much will", or "How many will".
4. There must be a time element in the question. It can be phrases like "In {} ...", "By {}, ...", "... in {}?".
5. You MUST NOT use unclear implicit time element phrases like "in the future" or "in the upcoming weeks".
6. You should avoid: questions that require numerical reasoning; questions that require substantial world knowledge.
7. The answer MUST be short and concise, avoiding using redundant words or repeating the information in the question.
8. The question must be grammatically correct and contain the information required to answer. NO "he, she, they, it, them, etc" allowed. Please clearly write out the entity you are referencing in the question.
9. The question MUST be able to be answered by the article.
10. The question MUST NOT include the information that came out just now. It should be understandable to people from the past. Avoid using "How will" or "Why will" questions, as they imply that the event has already occurred.

[Suggested questions and questions to avoid are detailed below:]

- Keypoint: Delhi experienced its highest-ever temperature of 49.9 degrees Celsius on Tuesday, leading to water rationing due to the oppressive heat wave.
- Suggested Question: Will Delhi break the highest temperature record again by May 2024?
- Avoid This Question: Will extreme heat events continue to pose a threat to India's development in the upcoming years?
- Reason to Avoid: The time constraint "in the upcoming years" is vague and the question can not be answered based on today's knowledge.

Figure 17: Prompt in the *QA Generation* step (part 1), adapted from Zhang et al. [13].

- Keypoint: Owners of nearly 84,000 older Nissan vehicles in the United States equipped with recalled, unrepaired Takata air bags, including models such as the 2002-2006 Nissan Sentra, are advised by NHTSA to immediately stop driving them due to safety concerns.
- Suggested Question: Will the older Nissan vehicles such as the 2002-2006 Nissan Sentra exhibit quality issues by May 2024?
- Avoid This Question: Will owners of the 2002-2006 Nissan Sentra, 2002-2004 Nissan Pathfinder, and 2002-2003 Infiniti QX4 heed the NHTSA's advice to immediately stop driving their vehicles in late May 2024?
- Reason to Avoid: This question is overly specific. People from the past would not have known the "NHTSA's advice".

- Keypoint: Children's sketches of violent scenes, likely made by children aged 5-7 before the eruption of Mt. Vesuvius in 79 AD, have been uncovered at the archaeological park of Pompeii.
- Suggested Question: Will children's sketches of violent scenes be discovered at the archaeological park of Pompeii by May 2024?
- Avoid This Question: Will the newly discovered children's sketches at the archaeological park of Pompeii be available for public viewing by May 2024?
- Reason to Avoid: This question includes future events about newly discovered children's sketches in Pompeii, which wouldn't be known to a past audience

- Keypoint: North Korea has been sending "filth and garbage" across the border to South Korea using giant balloons as a new strategy, prompting South Korean authorities to warn of the objects landing in residential areas. The move, according to North Korean state media KCNA, was to retaliate against South Korean activists who often send materials to the North.
- Suggested Question: What will North Korea do to retaliate against South Korean activists who often send materials to the North by May 2024?
- Avoid This Question: Will North Korea continue using balloons to send items across the border to South Korea by May 2024?
- Reason to Avoid: The word "continue" should not be used here. The question MUST NOT include the information that came out just now.

[Output:] Now please write four clear and concise question-answer pairs following the instructions and examples above. Once again the question should NOT be designed for reading comprehension but of forecasting interests. Also, vague and implicit time elements like "in the future", "in the upcoming weeks" or "in the coming years" should NOT be used. The question should be able to answer on {}, but only guessable not answerable before {}. You should output the question along with its answer, in the format of
"""
Question 1: "Will xxx?"
Answer 1: Yes.

Question 2: "Will xxx?"
Answer 2: No.

Question 3: Either "What will xxx?", "Who will xxx?", "Where will xxx?", "Which xxx will", "How much will xxx?", or "How many will xxx?"
Answer 3: xxx.

Question 4: Either "What will xxx?", "Who will xxx?", "Where will xxx?", "Which xxx will", "How much will xxx?", or "How many will xxx?"
Answer 4: xxx.
"""

Figure 18: Prompt in the *QA Generation* step (part 2), adapted from Zhang et al. [13].

```
# Rules
Article: {}
Given the article, please generate three noising answers to the given questions, whose correct answers can be obtained from the article. Name the three noising answers as (b), (c) and (d) respectively. While (b), (c) and (d) should all be unambiguously incorrect, they should also make sense and be plausible.

# Examples
Here are examples showing the output format. This example is NOT related to the noising answers you will generate.

Question: What will be the annual change in the UK's Consumer Prices Index (CPI) for November 2021?
Correct Answer: 'Less than 1.7%'
Noising Answers:
(b) 'Between 1.7% and 2.2%, inclusive'
(c) 'More than 2.2% but less than 2.9%'
(d) '2.9% or more'

Question: Who will win the 2020 Georgia Democratic primary?
Correct Answer: 'Joe Biden'
Noising Answers:
(b) 'Michael Bloomberg'
(c) 'Pete Buttigieg'
(d) 'Someone else'

Question: Before July 2020, will it be officially announced that the Tokyo 2020 Summer Olympics and/or Paralympics will be postponed, canceled, and/or relocated?
Correct Answer: Yes, the Olympic Games only
Noising Answers:
(b) 'Yes, the Paralympic Games only'
(c) 'Yes, both'
(d) 'No'

# Input:
Question 1: {}
Correct Answer 1: {}

Question 2: {}
Correct Answer 2: {}

# Output: Now please generate three noising answers to the question, given the above article, instructions and examples. DO NOT output the backgrounds, the question or any other explanations.
Noising Answers 1:
(b) xxx.
(c) xxx.
(d) xxx.

Noising Answers 2:
(b) xxx.
(c) xxx.
(d) xxx.
```

Figure 19: Prompt in the *Misleading Choices Generation* step, adapted from Zhang et al. [13]..

# Task
Please help evaluate the quality of question-answer pairs derived from the given news article. The questions will be presented to someone who has not seen the corresponding news article, in order to evaluate the accuracy and plausibility of the event prediction ability.

# Inputs
Article: {}
Publishing Date: {}
Question 1: {}
Answer 1: {}
Question 2: {}
Answer 2: {}
Question 3: {}
Answer 3: {}
Question 4: {}
Answer 4: {}

# Scoring Categories
## Correctness: Given the above article, please check if the answer is correct to the question with 100% certainty.
- 2 points: There is evidence in the article that the answer is correct with 100% certainty.
- 1 point: The answer generally aligns with the news facts but has minor inaccuracies or missing details.
- 0 point: Significantly misaligned with the news facts.

## Only Answerable on Publishing Date: Imagine traveling back in time to one week before the article's publishing date ({}). At that time, you are asked the question without having seen this specific article, but you do have access to all earlier news articles. The question should ideally be only guessable—not definitively answerable—based on the information available at that time. That is, the answer should be able to be found in the given article, but it should not be obtainable from earlier articles. Note that past tense descriptions in the article DO NOT INFLUENCE this assessment.
- 2 points: The question is answerable on {}, but only guessable not answerable before {}.
- 1 point: Could be somewhat predicted before {}, but not with complete certainty.
- 0 point: A person (could be anyone, even an expert in the field) would be able to find an article (or many) published before {} that answers the question with 100% certainty.

### 0 point examples
Example 1:
Question: What will be one of Lexi Thompson's career highlights in professional golf?
Answer: Winning 11 LPGA Tour titles.
Reasoning: This question is answerable with prior knowledge and does not test predictive ability related to the publishing date.

## No New Information: Ensure the question does not include new information that only became known on the publishing date, making it understandable for a past audience.
- 2 points: No new information from the publishing date are included.
- 1 point: Minor new information from the publishing date might be inferred but are not explicitly stated.
- 0 point: Includes clear new information from the publishing date, unsuitable for past understanding.

Figure 20: Prompt in the *QA Filtering* step (part 1).

### 0 point examples

Example 1:

Question: Will owners of the 2002-2006 Nissan Sentra, 2002-2004 Nissan Pathfinder, and 2002-2003 Infiniti QX4 heed the NHTSA's advice to immediately stop driving their vehicles in late May 2024?

Reasoning: This question contains new information on the publishing date. People from the past would not have known the "NHTSA's advice".

Example 2:

Question: "What will Lexi Thompson's ranking be at the time of her retirement announcement in May 2024?"

Reasoning: This question contains the information that Lexi will annouce her retirement, which is not known to the people from the past.

Example 3:

Question: "Will the newly discovered children's sketches at the archaeological park of Pompeii be available for public viewing by May 2024?"

Reasoning: This question includes future events about newly discovered children's sketches in Pompeii, which wouldn't be known to a past audience

## Objectiveness: The answer should not rely more on the author's personal views than on objective facts.

- 2 points: Completely objective, based strictly on reported facts.
- 1 point: Primarily objective, with minor subjective interpretations.
- 0 point: Largely subjective or opinion-based, lacking a factual basis.

## Clear Time Element: This category checks if the question has a clear element in it, without having vague phrases like "in the future" or "in the upcoming weeks".

- 2 points: The question has clear time elements, like "by May 2024" or "in July 2023".
- 1 point: The question includes a general timeframe, like "next month" or "this winter," which allows for some estimation but lacks precise dates.
- 0 point: The question includes vague time phrases like "in the future" or "in the upcoming weeks," which do not specify a clear or precise timeframe.

### 0 point examples

Example 1:

Question: Will extreme heat events continue to pose a threat to India's development in the upcoming years?

Reasoning: The time constraint "in the upcoming years" is vague.

Example 2:

Question: "What will Illinois require from parents who monetize their children's online activities starting in July?"

Reasoning: The mention of "July" specifies only the month and lacks the necessary detail of the year.

## Public Interest: Determine if the question addresses a topic of public concern.

- 2 points: The question covers a topic that widely affects or interests the public.
- 1 point: The question is of moderate interest, relevant to specific groups.
- 0 point: The topic is overly personal or localized, lacking relevance to the broader public.

### 0 point examples

Example 1:

Question: Will the exhibition 'Fragile Beauty' at London's Victoria & Albert Museum include both midcentury and contemporary works in May 2024?

Reasoning: The specific details of an personal art exhibition's contents are generally of limited public interest.

Figure 21: Prompt in the *QA Filtering* step (part 2).

## Answer Not Too Obvious: This category evaluates whether the answer to a question is too predictable or straightforward based on the question itself.
- 2 points: The answer provides new or non-obvious insights, requiring additional context or understanding not explicit in the question.
- 1 point: The answer is somewhat predictable but includes minor additional information or a slight twist.
- 0 point: The answer directly restates or closely mirrors the question, offering no new details or insights.

### 0 point examples
Example 1:
Question: What will New York officials do to ensure safety for the ICC Men's T20 Cricket World Cup following global threats from ISIS-K?
Answer: New York officials will implement increased safety precautions for the event. Reasoning: The answer is straightforward and expected, as it directly restates the premise of the question without providing any new or specific details on how the safety precautions will be implemented or what they might entail.

# Instructions
Evaluate each question-answer pair by assigning points in each of the categories based on the criteria provided. Please be strict on giving points. If the requirements of a category are not fulfilled, assign a point of 0.

# Please strictly follow this output template:
Question 1
## Correctness
- Reasoning:
- Point:
## Only Answerable on Publishing Date
- Reasoning:
- Point:
## No New Information
- Reasoning:
- Point:
## Objectiveness
- Reasoning:
- Point:
## Clear Time Element
- Reasoning:
- Point:
## Public Interest
- Reasoning:
- Point:
## Answer Not Too Obvious
- Reasoning:
- Point:

Question 2
## Correctness
- Reasoning:
- Point: ## Only Answerable on Publishing Date
- Reasoning:
- Point:
## No New Information
- Reasoning: - Point:

Figure 22: Prompt in the *QA Filtering* step (part 3).

```
## Objectiveness
- Reasoning:
- Point:
## Clear Time Element
- Reasoning:
- Point:
## Public Interest
- Reasoning:
- Point:
## Answer Not Too Obvious
- Reasoning:
- Point:

*Question 3*
## Correctness
- Reasoning:
- Point:
## Only Answerable on Publishing Date
- Reasoning:
- Point:
## No New Information
- Reasoning:
- Point:
## Objectiveness
- Reasoning:
- Point:
## Clear Time Element
- Reasoning:
- Point:
## Public Interest
- Reasoning:
- Point:
## Answer Not Too Obvious
- Reasoning:
- Point:

*Question 4*
## Correctness
- Reasoning:
- Point:
## Only Answerable on Publishing Date
- Reasoning:
- Point:
## No New Information
- Reasoning:
- Point:
## Objectiveness
- Reasoning:
- Point:
## Clear Time Element
- Reasoning:
- Point:
## Public Interest
- Reasoning:
- Point:
## Answer Not Too Obvious
- Reasoning:
- Point:
```

Figure 23: Prompt in the *QA Filtering* step (part 4).

System Prompt: You're an expert in forecasting events. You will NEVER refuse to answer a forecasting question by saying "I cannot predict the future", even if without 100% certainty.

User Prompt: You should output your answer as either 'Yes' or 'No' WITHOUT anything else.

Question: {}
Choices: 'Yes' or 'No'
[Output:] Your answer:

Figure 24: Closed-book evaluation prompt for TF questions, adapted from Zhang et al. [13].

System Prompt: You're an expert in forecasting events. You will NEVER refuse to answer a forecasting question by saying "I cannot predict the future", even if without 100% certainty.

User Prompt: You should output your answer as either '(a)', '(b)', '(c)' or '(d)' WITHOUT anything else.

Question: {}
Choices:
(a) {}
(b) {}
(c) {}
(d) {}
[Output:] Your answer:

Figure 25: Closed-book evaluation prompt for MC questions, adapted from Zhang et al. [13].

System Prompt: You're an expert in forecasting events. You will NEVER refuse to answer a forecasting question by saying "I cannot predict the future", even if without 100% certainty.

User Prompt: You should output your answer as either 'Yes' or 'No' WITHOUT anything else. Below are the top 5 relevant news article fragments retrieved for the question, which may or may not assist you in making a forecast.
Article 1: {}
Article 2: {}
Article 3: {}
Article 4: {}
Article 5: {}

Question: {}
Choices: 'Yes' or 'No'
[Output:] Your answer:

Figure 26: Constrained open-book evaluation prompt for TF questions, adapted from Zhang et al. [13].

System Prompt: You’re an expert in forecasting events. You will NEVER refuse to answer a forecasting question by saying "I cannot predict the future", even if without 100% certainty.

User Prompt: You should output your answer as either ’(a)’, ’(b)’, ’(c)’ or ’(d)’ WITHOUT anything else. Below are the top 5 relevant news article fragments retrieved for the question, which may or may not assist you in making a forecast.
Article 1: {}
Article 2: {}
Article 3: {}
Article 4: {}
Article 5: {}

Question: {}
Choices:
(a) {}
(b) {}
(c) {}
(d) {}
[Output:] Your answer:

Figure 27: Constrained open-book evaluation prompt for MC questions, adapted from Zhang et al. [13].

System Prompt: You’re an expert in forecasting events. You will NEVER refuse to answer a forecasting question by saying "I cannot predict the future", even if without 100% certainty.

User Prompt: You should output your answer as either ’Yes’ or ’No’ WITHOUT anything else. Below is the updated news article relevant to the question, which may help you in providing an answer.
Article: {}

Question: {}
Choices: ’Yes’ or ’No’
[Output:] Your answer:

Figure 28: Gold article evaluation prompt for TF questions, adapted from Zhang et al. [13].

System Prompt: You’re an expert in forecasting events. You will NEVER refuse to answer a forecasting question by saying "I cannot predict the future", even if without 100% certainty.

User Prompt: You should output your answer as either ’(a)’, ’(b)’, ’(c)’ or ’(d)’ WITHOUT anything else. Below is the updated news article relevant to the question, which may help you in providing an answer.
Article: {}

Question: {}
Choices:
(a) {}
(b) {}
(c) {}
(d) {}
[Output:] Your answer:

Figure 29: Gold article evaluation prompt for MC questions, adapted from Zhang et al. [13].

```
Question: {}
Options:
- Science & Tech
- Healthcare & Biology
- Economics & Business
- Environment & Energy
- Politics & Governance
- Education & Research
- Arts & Recreation
- Security & Defense
- Social Sciences
- Sports
- Other
Instruction: Assign a category for the given question.
Rules:
1. Make sure you only return one of the options from the option list.
2. Only output the category, and do not output any other words in your response.
3. You have to pick a string from the above categories.
Answer:
```

Figure 30: Prompt to categorize the generated questions, taken from Halawi et al. [8].