When Benchmarks Age: Temporal Misalignment through Large Language Model Factuality Evaluation

Xunyi Jiang **Dingyi Chang** **Xin Xu***
University of California, San Diego
{xuj003, dic006, xinxucs}@ucsd.edu

Abstract

The rapid evolution of large language models (LLMs) and the real world has outpaced the static nature of widely used evaluation benchmarks, raising concerns about their reliability for evaluating LLM factuality. While substantial works continue to rely on the popular but old benchmarks, their temporal misalignment with real-world facts and modern LLMs, and their effects on LLM factuality evaluation remain underexplored. Therefore, in this work, we present a systematic investigation of this issue by examining five popular factuality benchmarks and eight LLMs released across different years. An up-to-date fact retrieval pipeline and three metrics are tailored to quantify benchmark aging and its impact on LLM factuality evaluation. Experimental results and analysis illustrate that a considerable portion of samples in the widely used factuality benchmarks are outdated, leading to unreliable assessments of LLM factuality. We hope our work can provide a testbed to assess the reliability of a benchmark for LLM factuality evaluation and inspire more research on the benchmark aging issue[2]

1 Introduction

New large language models have been released at an unprecedented pace recently [Zhao et al., 2023, Minaee et al., 2024]. Accompanying this proliferation is the rise of numerous benchmarks that aim to compare diverse LLMs across a wide range of tasks [Liang et al., 2023, Chang et al., 2024, Ma et al., 2025]. Many benchmarks are static [Vu et al., 2024], meaning that the factual information they contain remains unchanged in response to real-world updates. For example, the answer to the question "What is the most populated country in the world?" is India[3] nowadays in 2025. However, the gold answer from SelfAware [Yin et al., 2023a] released in May 2023 is still China. As a result, LLMs that provide up-to-date and factually correct answers may be unfairly penalized when evaluated against outdated benchmarks. Despite this issue, numerous studies continue to rely on these static benchmarks to assess LLM factuality (§3.3). Although some works [Kasai et al., 2023, Vu et al., 2024] have introduced real-time updated benchmarks or methods to help LLMs obtain real-world information, the effects of using old benchmarks to evaluate present LLMs have not been systematically investigated (See Appendix A for a detailed review of related works).

Motivated by this gap, we conduct a comprehensive empirical study on the temporal misalignment with five popular factuality benchmarks across different years and explore their implications for evaluating modern LLMs. Our study focuses on two key research questions:

- **RQ1**: To what extent do widely used static benchmarks contain outdated factual answers compared to current real-world facts?

*Corresponding Author.

[2]Codes are available in https://github.com/JiangXunyi/BenchAge.

[3]https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle.

- **RQ2**: How does benchmark aging affect the factuality evaluation of modern LLMs?

To answer these questions, we focus on time-sensitive questions (§2.1). We introduce a retrieval pipeline to obtain current real-world facts for temporal comparison (§2.2). Three metrics, including **Dataset Drift Score, Evaluation Misleading Rate, and Temporal Alignment Gap**, are tailored to quantify temporal misalignment and measure the impacts of benchmark aging on LLM factuality evaluation (§2.3). Extensive experiments are implemented across five commonly used benchmarks and eight LLMs released over different years. The results illustrate a considerable portion of samples in the old but popular benchmarks are outdated (RQ1). The systematic analysis reveals that the temporal misalignment of the benchmarks with modern LLMs and the real-world facts will lead to unreliable assessments, raising concerns about the trustworthiness of LLM factuality evaluation (RQ2). We hope that our work can provide a testbed to evaluate the benchmark reliability for factuality evaluation and inspire more considerations about benchmarking aging in future work.

2 Experimental Setup

To investigate temporal misalignment, we design a three-stage experiments to compare benchmark gold labels, LLM responses, and real-world facts.

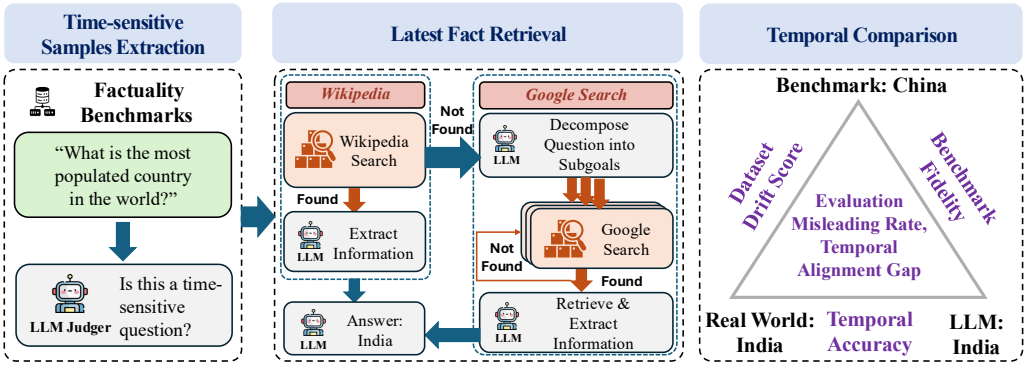


Figure 1: Experimental setups. We first extract time-sensitive samples and then collect the corresponding real-world fact (with the latest fact retrieval pipeline), LLM response, and the gold label in the benchmark for each sample. Finally, we apply the proposed metrics to measure the temporal misalignment among them.

2.1 Time-sensitive Samples Extraction

Focusing on temporal misalignment, we first extract *time-sensitive samples*, which have verifiable factual answers that will change over time [Wei et al., 2024], from the benchmarks. The time-sensitive questions are identified for each benchmark by an LLM with human evaluations, achieving a 100% recall and 90% accuracy. The details are described in Appendix B.1 and C.2.

2.2 Latest Fact Retrieval

To obtain the real-world facts, we retrieve the up-to-date answer from the Internet for each time-sensitive question. Our approach combines Wikipedia-focused retrieval and iterative web search, as depicted in Figure 1. For each time-sensitive question, we first retrieve related information from Wikipedia, a widely regarded source of reliable factual information for popular topics and recent events [McDowell, 2024], using Brave Search[4]. Secondly, GPT-4o-mini[5] is deployed to extract final answers from the retrieved information. If Wikipedia lacks suitable coverage, we use the Google Search API. Following ReAct and Chain-of-Action [Yao et al., 2023, Pan et al., 2025], we combine iterative reasoning with evidence retrieval. The system (1) decomposes questions into subgoals; (2) runs targeted searches for subgoals; (3) extracts key facts and temporal metadata; and (4) uses

[4] https://brave.com/search/api/

[5] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Qwen2.5-14B-Instruct [Yang et al., 2024] to decide if further refinement and search are needed. Detailed workflow is discussed in Appendix C.3. To assess the retrieval quality, each annotator with instruction in Figure 8 manually reviewed 105 samples to determine whether the retrieved answers accurately reflect the searched evidence. The process achieves 89.52% retrieval accuracy with moderate inter-annotator agreement ($\kappa = 0.6$).

2.3 Temporal Comparison

After obtaining the latest fact and the LLM response (Appendix C.1) for each time-sensitive question, we conduct a thorough analysis to explore temporal misalignment and its effects on LLM factuality evaluation. We tailor the following metrics to help with analysis. Specifically, given the query x_i and the gold answer y_i in each sample from the time-sensitive subset $\mathcal{D}_{\text{ts}}$ of a benchmark $\mathcal{D}$, the corresponding LLM response $\hat{y}_i$, and the real-world answer y_i^* searched from the Internet, we compute two binary alignment scores: $s_i^{\text{gold}} = \mathbb{1}[\hat{y}_i = y_i]$ (the agreement between y_i and $\hat{y}_i$), and $s_i^{\text{search}} = \mathbb{1}[\hat{y}_i = y_i^*]$ (the agreement between $\hat{y}_i$ and y_i^*).

RQ1 To quantify how the gold answers of time-sensitive samples in a benchmark have diverged from up-to-date real-world facts, we propose **Dataset Drift Score** (DDS):

$$DDS = \frac{1}{|D_{ts}|} \sum_{i=1}^{|D_{ts}|} \mathbb{1}[y_i \neq y_i^*], \tag{1}$$

RQ2 We introduce two metrics to capture how benchmark aging affects LLM evaluation. **Evaluation Misleading Rate** (EMR) reflects how often benchmark aging results in misleading evaluation results, which is the proportion of cases where an LLM gives an up-to-date answer, but is penalized by outdated benchmark labels:

$$EMR = \frac{1}{|D_{ts}|} \sum_{i=1}^{|D_{ts}|} \mathbb{1}[\hat{y}_i = y_i^* \wedge \hat{y}_i \neq y_i] \tag{2}$$

Temporal Alignment Gap (TAG) measures the discrepancy between LLM–world (Temporal Accuracy, TA) and LLM–benchmark (Benchmark Fidelity, BF):

$$TAG = TA - BF \tag{3}$$

$$= \frac{1}{|D_{ts}|} \sum_{i=1}^{|D_{ts}|} s^{\text{search}} - \frac{1}{|D_{ts}|} \sum_{i=1}^{|D_{ts}|} s^{\text{gold}} \tag{4}$$

$$= \frac{1}{|D_{ts}|} \sum_{i=1}^{|D_{ts}|} (s^{\text{search}} - s^{\text{gold}}) \tag{5}$$

A positive TAG indicates that the LLM responses align more with the real-world facts than with the benchmark gold labels.

Based on these metrics, we investigate 8 diverse LLMs on commonly used LLM factuality benchmarks (Figure 2b), including TriviaQA [Joshi et al., 2017], BoolQ [Clark et al., 2019], Natural Questions [Kwiatkowski et al., 2019], TruthfulQA [Lin et al., 2022], and SelfAware [Yin et al., 2023b], which were released in different years. The DDS and EMR are shown in Table 1. More details are shown in Appendix C.1 and C.4.

3 Experimental Results and Analysis

3.1 RQ1: A Considerable Portion of the Benchmarks Are Outdated

DDS values in Table 1 show that at least 24.19% (even up to 63.78%) of time-sensitive samples are outdated as of July 19, 2025. Among all benchmarks, the relatively old benchmark, BoolQ, exhibits the highest DDS. In contrast, newer benchmarks such as SelfAware show relatively less misalignment, reflecting shorter temporal distance from their release dates. These results suggest that a considerable portion of the existing static and old benchmarks, though valid at release time, have become outdated over time.

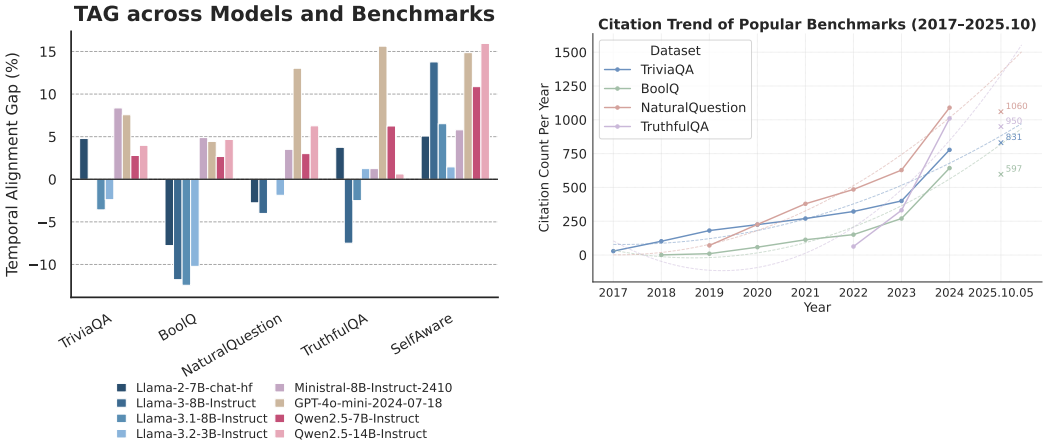
Table 1: Dataset Drift Score (%) of five widely-used LLM factuality benchmarks released along with time and Evaluation Misleading Rate (%) of the modern LLMs (>20%).

Dataset Release Time	**TriviaQA** July 2017	**BoolQ** May 2019	**NaturalQuestion** July 2019	**TruthfulQA** May 2022	**SelfAware** July 2023
Dataset Drift Score (%)	37.05	63.78	24.19	36.88	28.26
LLM (Release Time)			**Evaluation Misleading Rate (%)**		
Llama-2-7B-chat-hf (Jul 2023)	14.74	9.11	10.28	11.25	15.22
Llama-3-8B-Instruct (Apr 2024)	11.16	8.22	10.28	8.13	19.57
Llama-3.1-8B-Instruct (Jul 2024)	12.35	7.56	11.40	9.38	14.49
Llama-3.2-3B-Instruct (Sep 2024)	9.16	8.67	9.52	10.63	10.51
Ministral-8B-Instruct-2410 (Sep 2024)	18.33	16.67	14.04	14.38	15.22
GPT-4o-mini-2024-07-18 (Jul 2024)	19.92	17.11	24.06	23.13	22.10
Qwen2.5-7B-Instruct (Sep 2024)	10.76	14.44	12.41	19.38	16.67
Qwen2.5-14B-Instruct (Sep 2024)	13.55	16.00	16.04	16.88	22.46

3.2 RQ2: Benchmark Aging Affects the Reliability of LLM Evaluation

The outdated benchmarks can mislabel factually correct model responses. According to Table 1, more than half of the EMR is larger than 10%, indicating that a non-trivial fraction of LLM outputs are factually correct with respect to the real-world facts but judged as incorrect by stale benchmark labels. GPT-4o-mini and Qwen2.5-14B-Instruct exhibit relatively higher EMR across all datasets than other LLMs, suggesting that newer LLMs are more vulnerable to evaluation bias, as they more frequently produce up-to-date answers that conflict with outdated references. Overall, these results highlight that benchmark aging introduces systematic misalignment between factual model performance and reported evaluation scores.

The present LLMs are more aligned with real-world facts than with gold answers in the benchmarks. We further analyze the temporal consistency between LLM outputs, benchmark labels, and real-world facts through TAG in Figure 2a, Table 4, and 5. Overall, 70% of TAG scores are positive, indicating that the LLMs mostly align more with up-to-date real-world facts than with outdated benchmark labels. This pattern is observed consistently across the five benchmarks, especially for SelfAware, whose data are from relatively old datasets, such as SQuaD [Rajpurkar et al., 2016], TriviaQA [Joshi et al., 2017], and HotpotQA [Yang et al., 2018]. Figure 11 shows Cohen's Kappa κ [McHugh, 2012] between each other among LLM responses, searched real-world information, and gold benchmark answers (Appendix D.1). The generally low κ of LLM vs Gold



(a) TAG across the LLMs and benchmarks.

(b) Annual citation growth of the benchmark.

Figure 2: (a) TAG. (b) Benchmark Citations.

(<0) and Gold vs Search (mostly <0.6) emphasizes the significant temporal misalignment between the gold labels in the benchmarks and the others.

3.3 Dataset Analysis

The usage of static benchmarks with outdated information is increasing. The release time in Table 1 suggests that the benchmarks we investigated are very old, and there is a significant time gap among the benchmarks, present LLMs, and real-world facts. The upward trend in Google Scholar citations for these benchmarks is shown in Figure 2b. In the single year of 2024, the citations of Natural Questions and TruthfulQA surpassed 1,000, demonstrating their popularity for LLM evaluation. These benchmarks have not been systematically updated to reflect evolving real-world facts. Nevertheless, they have been widely adopted in prior work and are likely to remain in use. This persistent reliance highlights the need for more attention to the unreliable use of the outdated benchmarks.

The outdated contexts amplify the temporal misalignment. In open-book QA tasks, outdated information in the provided context can worsen factually temporal drift. BOOLQ [Clark et al., 2019], for instance, includes a supporting passage before a query. As shown in Table 6, models consistently exhibit more negative TAG when performing passage-grounded inference. For example, Qwen2.5-7B-Instruct's TAG drops from 2.67% without the passage to -12.22% with it. This indicates that the passages often encode outdated facts and anchor the model toward obsolete answers instead of correcting them since LLMs rely more on contexts instead of memorized knowledge [Li et al., 2023, Zhou et al., 2023, Xie et al., 2024], which suggests temporal degradation is not limited to open-ended generation but also affects passage-grounded evaluations.

4 Conclusion

In this work, we conduct a comprehensive empirical study and provide a testbed to investigate the temporal misalignment of the existing static LLM benchmarks with present LLMs and the real world, and its impacts on LLM factuality evaluation. The results and analysis reveal that a considerable portion of samples in the widely used factuality benchmarks are outdated. Reliance on these aging benchmarks will lead to unreliable LLM factuality evaluation. We hope this work can suggest future research to consider temporal misalignment in the benchmark design and LLM factuality evaluation.

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15 (3):39:1–39:45, 2024. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1300. URL https://doi.org/10.18653/v1/n19-1300.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: what's the answer right now? In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey

Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9941624ef7f867a502732b5154d30cb7-Abstract-Datasets_and_Benchmarks.html.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL https://doi.org/10.1162/tacl_a_00276.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL https://aclanthology.org/D19-1598/.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.112. URL https://aclanthology.org/2023.findings-acl.112/.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.435. URL https://aclanthology.org/2022.naacl-main.435/.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.425. URL https://aclanthology.org/2025.acl-long.425/.

Zachary J McDowell. Wikipedia and ai: Access, representation, and advocacy in the age of large language models. *Convergence*, 30(2):751–767, 2024. doi: 10.1177/13548565241238924. URL https://doi.org/10.1177/13548565241238924.

Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282, 2012. doi: 10.11613/BM.2012.031. URL https://pubmed.ncbi.nlm.nih.gov/23092060/.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *CoRR*, abs/2402.06196, 2024. doi: 10.48550/ARXIV.2402.06196. URL https://doi.org/10.48550/arXiv.2402.06196.

Seyed Mahed Mousavi, Edoardo Cecchinato, Lucia Hornikova, and Giuseppe Riccardi. Garbage in, reasoning out? why benchmark scores are unreliable and what to do about it, 2025. URL https://doi.org/10.48550/arXiv.2506.23864.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=1BdPHbuimc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1264. URL https://doi.org/10.18653/v1/d16-1264.

Revanth Gangi Reddy, Tanay Dixit, Jiaxin Qin, Cheng Qian, Daniel Lee, Jiawei Han, Kevin Small, Xing Fan, Ruhi Sarikaya, and Heng Ji. WINELL: wikipedia never-ending updating with LLM agents. *CoRR*, abs/2508.03728, 2025. doi: 10.48550/ARXIV.2508.03728. URL https://doi.org/10.48550/arXiv.2508.03728.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728, 2019. URL http://arxiv.org/abs/1904.09728.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.663. URL https://aclanthology.org/2023.findings-acl.663/.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13697–13720. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.813. URL https://doi.org/10.18653/v1/2024.findings-acl.813.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *CoRR*, abs/2411.04368, 2024. doi: 10.48550/ARXIV.2411.04368. URL https://doi.org/10.48550/arXiv.2411.04368.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

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, Tingyu Xia,

Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551/.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.FINDINGS-ACL.551. URL https://doi.org/10.18653/v1/2023.findings-acl.551.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023. doi: 10.48550/ARXIV.2303.18223. URL https://doi.org/10.48550/arXiv.2303.18223.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

A Related Works

The evaluation of large language models (LLMs) relies heavily on standardized benchmarks, which provide a common ground for comparing model performance across tasks and over time. Benchmarks such as Natural Questions [Kwiatkowski et al., 2019] and TriviaQA [Joshi et al., 2017] are widely used for evaluating LLMs, as they enable standardized comparisons, offer an objective measure of model performance, and help track advances in model capabilities. However, most existing benchmarks are static, capturing a snapshot of knowledge at a particular point in time, and thus may not reflect the evolving nature of real-world information.

As factual knowledge evolves, static benchmarks can quickly become outdated, leading to differences between what a benchmark evaluates and the current state of the world. This temporal misalignment is highlighted in recent studies showing that LLMs may provide correct answers according to up-to-date information yet be penalized by benchmarks anchored to outdated facts [Kasai et al., 2023, Vu et al., 2024]. To address this, several works propose dynamic or regularly updated benchmarks, such as RealTimeQA [Kasai et al., 2023] and FreshQA [Vu et al., 2024], which are designed to evaluate models on time-sensitive questions and recent events. In a complementary direction, Luu et al. [2022] investigate temporal drift from a model-centric perspective, analyzing how model performance

Table 2: Overview of QA datasets with time-sensitive (TS) questions.

Property	TriviaQA	BoolQ	NaturalQuestion
Year	2017	2019	2019
# QA Pairs	11,313	3,270	7,830
TS %	2.22	13.76	10.19
Type	Open QA	Multi-choice	Open QA

Property	TruthfulQA	SelfAware
Year	2022	2023
# QA Pairs	817	2,475
TS %	19.58	11.15
Type	Open QA	Open QA

changes when trained and tested on temporally shifted datasets. Recently, WiNELL [Reddy et al., 2025] tackles temporal misalignment from the data side by maintaining up-to-date content. These efforts emphasize the importance of incorporating temporal dynamics into benchmark design to ensure accurate and meaningful LLM evaluations.

Beyond temporal misalignment, recent research has revealed fundamental flaws in benchmark construction and evaluation methodologies that compromise the reliability of performance assessments. Systematic audits of popular reasoning benchmarks, including SocialIQa [Sap et al., 2019], FauxPas-EAI [Shapira et al., 2023], and ToMi [Le et al., 2019], have uncovered issues across many dimensions [Mousavi et al., 2025]. These studies show that a substantial portion of LLM errors are attributable not to the model, but to flaws in benchmark design, including duplicated items, ambiguous wording, implausible answers, and scoring procedures that prioritize output format over reasoning process. These findings challenge the validity of current benchmark-based claims about reasoning in LLMs and highlight the need for evaluation protocols that assess intended capabilities rather than unfairly penalize models due to benchmark design flaws or outdated factual content.

B Dataset Information

B.1 Dataset Creation Year and Time-sensitive Percentage

Table 2 summarizes the statistics of QA datasets used in this paper, including release year, total QA pairs, percentage of time-sensitive (TS) questions, and answer types. The time-sensitive questions extraction details are shown in Appendix C.2. Aside from TriviaQA, which only contains 2.22% time-sensitive questions, other datasets contains more than 10% of time-sensitive data. This shows that a non-negligible proportion of time-sensitive data exists in these popular benchmarks.

B.2 Google Scholar Citation Trend

To estimate the influence of each benchmark, we measure its citation trend using Google Scholar [6]. The citation data we collect is as of Oct 3, 2025. Specifically, we record the number of "cited by" results with year-specific filters from 2017 to 2025. In order to measure the future prediction trend, we use a polynomial to predict the citations at the end of 2025. In 2024 single year, the summation of citations for these 4 datasets is 3,521, revealing consistent scholarly interest in natural language processing and factuality question answering task.

C Experiment Details

C.1 Experiment Setups

1. **Computation environment**: All experiments were conducted on an Ubuntu 22.04 workstation equipped with four NVIDIA RTX A6000 GPUs (48 GB each), running CUDA 12.6 and

[6] https://scholar.google.com/

PyTorch 2.7.0+cu126. The software stack includes Python 3.10.18, Transformers 4.53.1, and OpenCompass 0.5.0 for evaluation orchestration, together with vLLM 0.9.2 for optimized inference.

2. **Model architectural analysis**: Models from different architectures released in a similar timeframe (June-October 2024): Qwen2.5-7B-Instruct [7], Ministral-8B-Instruct-2410 [8], Llama-3.1-8B-Instruct [9], and GPT-4o-mini [10] . Controlling for release date isolates the effects of architectural differences on temporal knowledge retention.
3. **Model scale analysis**: Qwen2.5 models of varying sizes (1.5B, 3B, 7B, and 14B) released simultaneously in September 2024 [Yang et al., 2024], isolating the effect of model scale.

C.2 Time-sensitive Samples Extraction

We use Qwen-2.5-14B-Instruct to extract time-sensitive samples from existing QA benchmarks. Specifically, we serve Qwen-2.5-14B-Instruct using the vLLM framework[11] for efficient inference. To reduce the randomness in LLM responses when identifying time-sensitive questions, we apply the prompt shown in Figure 3 three times independently and determine the final label by majority voting. We then conduct a grid search over the model's temperature and the number of voting rounds. Our results show that using three votes with a temperature of 1.0 yields the highest accuracy 90% while maintaining 100% recall for time-sensitive questions.

Prompt

You are a helpful assistant. Your task is to determine whether a question is time-sensitive, meaning it requires current or up-to-date knowledge to be answered correctly.
A question is considered **time-sensitive** ONLY IF both of the following are true:
1. It has a **verifiable factual answer**, AND
2. That answer can **change over time** due to events, leadership, scientific progress, or changing data.

A question is NOT time-sensitive if:
- It is subjective or opinion-based (e.g., "What is the best medicine?")
- It is hypothetical or open-ended (e.g., "If it's cold outside, what does that tell us about global warming?")
- It has **no specific factual answer** or depends on personal/local context

You must reason in **two steps**:
Step 1: Reasoning
Start with "Reasoning:" and explain whether the question meets both time-sensitivity criteria.
Step 2: Final Decision
Start with "Answer:" and respond only with "Yes" or "No."

Figure 3: Prompt for Determining the Time-Sensitivity of Dataset Questions.

To further validate the labeling quality, we conduct a human evaluation. Given the instruction in Figure 4, three domain experts manually annotate 150 questions, with results presented in Table 3. All

[7] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[8] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410
[9] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[10] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[11] https://github.com/vllm-project/vllm

Human Evaluation Guidelines: Time-sensitive Questions

Here is the instruction for annotation:

A question is TIME-SENSITIVE only if BOTH of the following are true:
1. It has a verifiable factual answer.
2. The correct answer can change over time (due to events, updated data, leadership change, etc.).

For each question in the form, please answer whether the question is time sensitive. You can type: y (represent yes)/ n (represent no)

The result is like:
```
{
   "question": "How has poverty changed over time in Africa?",
   "human": "y"
}
```

Figure 4: Instructions for Human Evaluation of Time-sensitive Questions

annotations were performed by graduate-level NLP researchers from our institution. These annotators are fluent in English and have prior experience evaluating QA datasets. Since the annotation task involved only publicly available benchmark data, no new human subject data was collected. Importantly, no crowd-sourcing platforms were involved; instead, the annotators participated voluntarily without any financial compensation. As a result, issues of participant recruitment, payment fairness, or data consent do not apply. Nonetheless, the annotation process and data usage were reviewed internally to ensure ethical compliance.

Table 3: Human evaluation of time-sensitive question detection.

Metric	Recall	F1 Score	Accuracy	Cohen's Kappa
Score	1.000	0.909	0.9	0.83375

C.3 Web Search Pipeline

All web search results were collected during a fixed time window from July 18 to July 19, 2025, ensuring consistency and temporal alignment across all queries. We utilize both the Google Search API and Brave Search (which includes access to Wikipedia content) to retrieve supporting evidence from the open web. To ensure robustness, our system is designed to tolerate transient network errors and incomplete results. In practice, we implement a retry mechanism: for Brave search, we retry up to three times in the event of failure. For Google search, as shown in Figure 5, we repeat the search process adaptively until either sufficient information is found (as judged by the LLM) or a hard limit of 15 search attempts is reached. These search engines are chosen for their broad coverage, freshness, and reliability—especially valuable for capturing real-world updates that static benchmarks fail to reflect.

This design reflects the reality that many benchmark answers cannot be verified from a single source like Wikipedia. Our logs reveal that only 22.3% of our selected questions were retrieved using Wikipedia as the source. The remainder required external evidence. To address this, our pipeline combines search engines, evidence consolidation, and LLM-based filtering, ensuring higher precision. We report the sources and methods used for transparency, and emphasize that our pipeline is intended for analysis of benchmark staleness, not as a replacement for routine benchmark updating. Notably, FreshQA also relies on web search to retrieve the latest answers for their benchmark questions, underscoring that web-based retrieval is a practical and accepted strategy for keeping benchmarks aligned with real-world facts.

To support the retrieval and reasoning process, we design a set of LLM prompts tailored to each stage of the pipeline. These prompts guide the model through subgoal planning, evidence extraction, fact sufficiency evaluation, and final answer generation. Visualizations of the four prompt templates are shown in Figures 6a–7b.

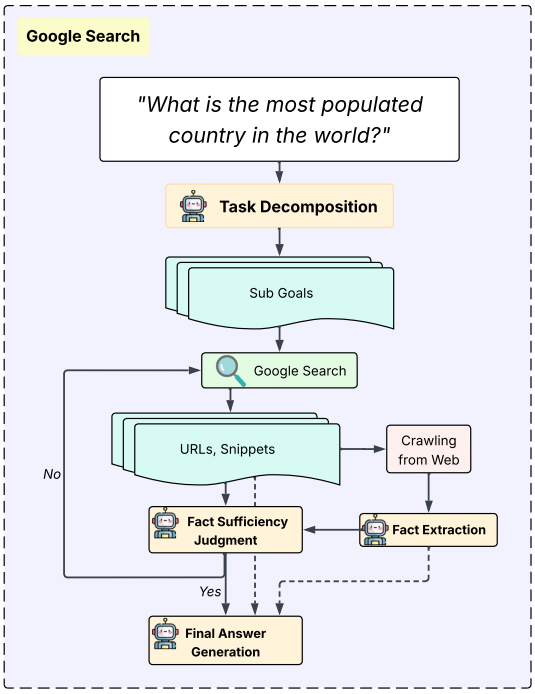


Figure 5: Workflow of Google Search and Fact Retrieval.

Task Decomposition Prompt
You are a reasoning assistant.
Your job is to break down the following question into a sequence of smaller information retrieval questions.
Original Question: "{question}"
Please list 2-4 sub-questions needed to answer this.
Output format:
- Subgoal 1: ...
- Subgoal 2: ...

(a) Task Decomposition Prompt

Fact Extraction Prompt
Given the following documents, extract the most relevant facts and evidence that could help answer a factual question.
Do NOT generate an answer. Just list facts.
Document1:
{text1}
Document2:
{text2}

(b) Fact Extraction Prompt

Figure 6: Prompts used in the fact retrieval pipeline: (a) task decomposition and (b) fact extraction.

To assess the quality of web search output, we adopt two complementary methods:

Human Evaluation: We randomly sample 105 questions from the dataset and ask three domain experts to manually assess whether the web search outputs provide factually correct answers. The overall accuracy reaches 89.52%, indicating a high degree of factual consistency. Furthermore, inter-annotator agreement, measured using Cohen's Kappa, is 0.58, which reflects moderate agreement. The detailed annotation instruction is provided in Figure 8.

Cohen's Kappa Analysis: To further evaluate the alignment between web search results and the dataset's gold answers, we calculate Cohen's Kappa scores across all time-sensitive questions. As illustrated in Figure 11, the green polygon representing this agreement lies between 0.5 and 0.8, suggesting a relatively strong consistency between search-derived answers and dataset labels. This level of agreement is expected, as only a small portion of questions involve fast-changing knowledge. Therefore, we infer that the retrieved web search results are generally reliable and can serve as a valid approximation of current factual information.

C.4 LLM-as-a-judge Prompt

To evaluate how well the model responses, benchmark answers, and real-world information agree with each other, we use Cohen's Kappa coefficient. This metric measures how consistently different

Fact Sufficiency Judgment
You are helping evaluate whether the current facts are enough to reasonably answer a given question.
QUESTION:
"{question}"
FACTS:
{list of facts}
SNIPPETS:
{list of snippets}
INSTRUCTION:
- Determine whether the available information is sufficient to give a **reasonable and helpful answer**, even if the data is **not the most recent**, **precise**, or **complete**.
- If the information provides a reasonable estimate or an approximate range that is directly related to the question, consider it **sufficient**.
- Only consider it **insufficient** if the information is clearly irrelevant, outdated by more than a decade without mention, or off-topic.
Respond in **exactly** one of the following two formats:
If sufficient:
Yes
If not sufficient:
No
REASON: <why it's not sufficient>
REVISED QUESTION: <a better follow-up query to help answer the original question>

(a) Fact Sufficiency Judgment Prompt

Final Answer Generation
You are a helpful and intelligent assistant.
Your task is to answer the user's question based on the following evidence gathered from web search.
Main Question:
{question}
The following subgoals were used to break down the question, and for each subgoal, supporting evidence is provided.
Subgoal 1: {question1}
Evidence: {evidence}
Facts: {facts}

Please use the information from the evidence to answer the main question as accurately and clearly as possible.
- You may synthesize facts from multiple sources.
- If the evidence contains conflicting or incomplete information, explain what is uncertain.
- If the answer requires inference (e.g., numerical comparison, trend analysis), explain your reasoning.
- If the answer is straightforward and factual, answer directly.
Final Answer:

(b) Final Answer Generation Prompt

Figure 7: Prompts used in the final stages of the retrieval pipeline: (a) fact sufficiency judgment and (b) final answer generation.

Instructions for Human Evaluation of Web Search Results

You are asked to evaluate the factual accuracy of search results produced by web search algorithms. For each example, please refer to the `search_results` field and determine whether the information provided is factually accurate.
If the search result is factually accurate, enter for the `accurate_or_not` field.
If the search result is factually incorrect or misleading, enter 0 for the `accurate_or_not` field.
Please follow the format below when recording your judgment:
"accurate_or_not": 1 or "accurate_or_not": 0
Thank you for your careful evaluation.

Figure 8: Human Evaluation Instruction for Web Search Results

sources align in their answers. We treat each source—the model, the benchmark, and the web search result—as an independent evaluator. Using the LLM-as-a-judge setup, we apply a clear and interpretable prompt in Figure 9 that asks the LLM to judge whether two answers express the same factual content. This process allows us to convert the answers into simple agreement scores, giving us a reliable and style-independent way to compare factual consistency across sources that may reflect knowledge from different points in time.

Similar to time-sensitive classification and websearch, we perform human evaluation of LLM-as-a-judge with the instructions in Figure 10. Outputs yields the following agreement: the accuracy is 97% and the average Cohen's Kappa between three evaluators is 0.72.

D Experiment Results

D.1 Cohen's Kappa Score

To systematically evaluate the agreement between different information sources, we compute the Cohen's Kappa coefficient, a standard inter-rater reliability metric in statistics and NLP. Formally, Cohen's Kappa is defined as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the observed agreement and p_e the expected agreement by chance. Unlike raw accuracy, Cohen's Kappa adjusts for chance-level agreement and thus provides a more robust and interpretable measure of consistency across different answer sources.

Prompt

If Answer 1 and Answer 2 **state the same fact** (allowing synonyms, abbreviations, or different date formats), reply True. Otherwise reply False.
Reply with exactly one word: True or False.
Q: *question*
Answer 1: *a1*
Answer 2: *a2*

Figure 9: Prompt for determining the time-sensitivity of dataset questions.

Human Evaluation Guidelines: LLM-as-a-Judge

Please review each question and its corresponding prediction and reference (gold or search-based answer). For each entry, complete the `annotator_label` field using the following criteria:
Label as 1 if you believe the prediction and reference convey the same factual meaning, allowing for synonyms, paraphrasing, or minor stylistic differences.
Label as 0 if the prediction and reference differ semantically or factually.
Your judgment should focus on factual consistency, not surface-level similarity.
Thank you for your careful evaluation.

Figure 10: Human Evaluation Instructions for LLM-as-a-judge

Figure 11 presents a radar plot of pairwise Cohen's Kappa scores among model outputs, web search results, and benchmark gold labels, computed across four datasets and ten representative LLMs. The radar shape reveals several insights. First, the agreement between web search and gold answers is generally high, indicating that our retrieval pipeline reliably captures accurate, up-to-date information. Second, the agreement between LLMs and the benchmark is lower, suggesting possible misalignment due to temporal drift or limitations in training data coverage. Finally, the agreement between LLMs and web search tends to be more variable, highlighting the inconsistent ability of models to match real-world facts in time-sensitive contexts.

Overall, this analysis illustrates the discrepancy between static benchmarks, dynamic web content, and model outputs. It motivates the need for time-aware evaluation and fact-checking frameworks that consider real-world knowledge freshness.

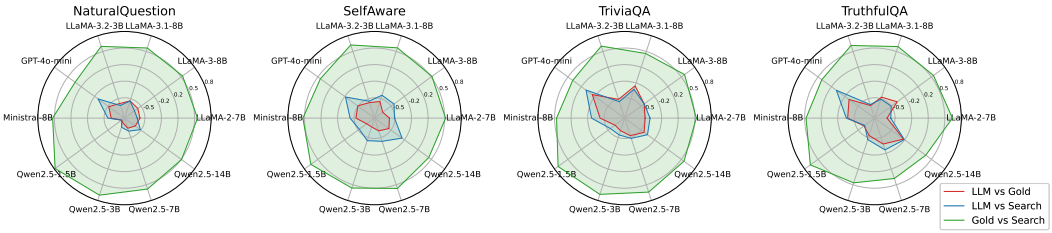


Figure 11: Cohen's Kappa Score between each other among LLMs' responses, searched real-world information, and gold benchmark answers from datasets and LLMs.

D.2 Temporal Accuracy and Benchmark Fidelity

We calculate TAG from Temporal Accuracy and Benchmark Fidelity. TA is shown in Table 4. GPT-4o-mini-2024-07-18 still performs best overall datasets. This represents that GPT-4o-mini-

2024-07-18 contains most up-to-date real world information. As the only close-source model, this observation highlights that currently the commercial model update more frequently.

Table 4: Temporal Accuracy (%): the proportion of time-sensitive questions answered correctly with respect to present-day information, reported across benchmarks.

Model / Dataset (Release Time)	**TriviaQA** July 2017	**BoolQ** May 2019	**NaturalQuestions** July 2019	**TruthfulQA** May 2022	**SelfAware** July 2023
Llama-2-7B-chat-hf	29.88%	49.78%	17.42%	20.63%	24.64%
Llama-3-8B-Instruct	28.69%	54.22%	17.04%	25.00%	28.62%
Llama-3.1-8B-Instruct	34.66%	57.56%	22.68%	24.38%	28.62%
Llama-3.2-3B-Instruct	21.91%	49.56%	17.92%	18.75%	22.46%
Ministral-8B-Instruct-2410	37.45%	59.56%	21.30%	32.50%	28.99%
GPT-4o-mini 2024-07-18	51.79%	77.78%	36.97%	51.25%	40.22%
Qwen2.5-7B-Instruct	25.90%	56.44%	18.30%	37.50%	28.99%
Qwen2.5-14B-Instruct	32.67%	63.78%	24.19%	40.63%	38.41%

Table 5: Benchmark Fidelity (%) showing model alignment with benchmark gold answers.

Model / Dataset (Release Time)	**TriviaQA** July 2017	**BoolQ** May 2019	**NaturalQuestions** July 2019	**TruthfulQA** May 2022	**SelfAware** July 2023
Llama-2-7B-chat-hf	25.10%	57.56%	20.18%	16.88%	19.57%
Llama-3-8B-Instruct	28.69%	66.00%	21.05%	32.50%	14.86%
Llama-3.1-8B-Instruct	38.25%	70.00%	22.56%	26.88%	22.10%
Llama-3.2-3B-Instruct	24.30%	59.78%	19.80%	17.50%	21.01%
Ministral-8B-Instruct-2410	29.08%	54.67%	17.79%	31.25%	23.19%
GPT-4o-mini 2024-07-18	44.22%	73.33%	23.93%	35.63%	25.36%
Qwen2.5-7B-Instruct	23.11%	53.78%	15.29%	31.25%	18.12%
Qwen2.5-14B-Instruct	28.69%	59.11%	17.92%	40.00%	22.46%

D.3 BoolQ TAG Comparison by Controlling Context

To investigate how outdated context in benchmarks can override updated internal knowledge in LLMs, we conduct controlled experiments on the BoolQ using two prompts, as illustrated in Figure 12. One setting provides both the passage and the question, while the other includes only the question without any supporting passage.

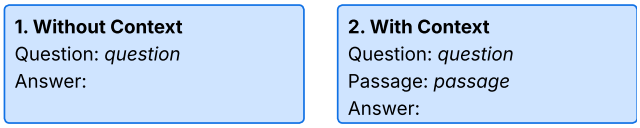


Figure 12: Two prompt formats used in BoolQ experiments: with and without passage context.

Interestingly, we observe a significant increase in Temporal Alignment Gap when the passage is included. This suggests that although models may have internally updated knowledge, the inclusion of outdated passages often causes them to regress toward older information. Quantitatively, this effect is most pronounced in Ministral-8B-Instruct-2410, which shows a TAG increase of 20.67 when conditioned on the passage. Similarly, GPT-4o-mini-2024-07-18 exhibits an increase of 19.33. These large deltas indicate that the models' updated knowledge is not robust against temporally stale input.

While BoolQ is originally constructed for reading comprehension, our analysis reveals that its static passages can contain outdated facts that actively mislead the model. The TAG between the two settings quantifies the vulnerability of LLMs to temporal anchoring by context, and highlights the need for temporal-awareness in prompt construction and model alignment.

Table 6: TAG (%) on BoolQ with and without passage contexts. Positive values (highlighted) indicate alignment between model outputs and current web results, diverging from outdated benchmark gold answers.

Model	w/o Passage	w/ Passage
LLaMA-2-7B-Instruct	-7.78	-7.56
LLaMA-3-8B-Instruct	-11.78	-16.22
LLaMA-3.1-8B-Instruct	-12.44	-21.33
LLaMA-3.2-3B-Instruct	-10.22	-16.44
Ministral-8B-Instruct-2410	4.89	-15.78
GPT-4o-mini-2024-07-18	4.44	-14.89
Qwen2.5-7B-Instruct	2.67	-12.22
Qwen2.5-14B-Instruct	4.67	-13.56

D.4 Model Analysis

We categorize LLMs into two different groups based on isolated factors to analyze their impacts on temporal misalignment, as shown in Appendix C.1. To quantify the impact of time-sensitive questions, we define TAG-adjusted accuracy a_{TAG} and the EMR-adjusted accuracy a_{EMR}.

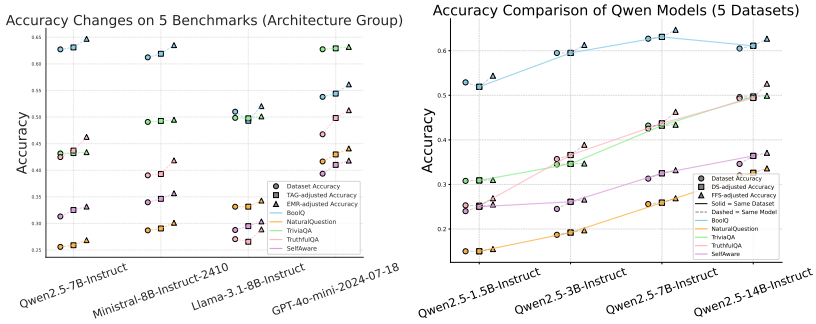
$$a_{TAG} = a_o + TAG \cdot \frac{|\mathcal{D}_{\text{ts}}|}{|\mathcal{D}|} \tag{6}$$

$$a_{EMR} = a_o + EMR \cdot \frac{|\mathcal{D}_{\text{ts}}|}{|\mathcal{D}|} \tag{7}$$

a_o denotes the LLM accuracy on $\mathcal{D}$. These adjusted accuracies measure the overall impact of the temporal change of facts in the benchmarks.

Model Family: The times of memorized facts vary between different LLM families. Despite similar release periods and size, LLMs vary in knowledge recency and accuracy, as shown in Figure 13a. For example, GPT-4o-mini-2024-07-18 shows a larger performance improvement on the searched information while Llama-3.1-8B-Instruct relies more on outdated answers, indicating that different LLM architectures and training data lead to different times of LLM-memorized facts.

Model Size: larger models are more robust to time change. A study of the Qwen models with different sizes (Figure 13b) reveals that as model size increases, LLm responses align more with up-to-date searched answers instead of the outdated benchmark answers, suggesting that larger models are more robust and better at adapting to time changes. We conjecture that more training data for larger models [Yang et al., 2024] will cover more recent information.



(a) Across model architecture families. (b) Across model sizes within the Qwen family.

Figure 13: Performance comparison across model families and sizes. Accuracy is reported in three settings: Dataset Accuracy, TAG-adjusted Accuracy, and EMR-adjusted Accuracy.