Time Waits for No Benchmark: Exploring the Temporal Misalignment between Static Benchmarks, Modern LLMs, and the Real World

Anonymous Author(s)
Affiliation
Address
email

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 the reliability of model assessments. Substantial works used popular but old benchmarks to evaluate modern LLMs. In this work, we explore the temporal misalignment between static benchmarks, modern LLMs, and the real world. We develop a multistage pipeline to implement temporal comparison with 10 LLMs and 5 benchmarks. Three quantitative scores are introduced to measure temporal misalignment. Experimental analysis illustrates that there is significant temporal misalignment between widely-used benchmarks, present LLMs, and the real-world facts. This temporal misalignment leads to untrustworthy LLM evaluations in a great number of existing and future works. These findings underscore a fundamental temporal issue of using static benchmarks to evaluate modern LLMs, suggesting a further exploration of temporally aware strategies to ensure more reliable and robust LLM evaluation.

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[1] 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that are expected to provide up-to-date and factually correct answers may be unfairly penalized when evaluated against outdated benchmarks [Kasai et al., 2023]. This temporal misalignment between static benchmarks, present LLMs, and the real world highlights a critical issue in LLM evaluation: the lack of temporal awareness when evaluating factuality.

In this work, therefore, we comprehensively investigate the temporal misalignment in widely used QA benchmarks and explore its implications for evaluating modern LLMs. Our study focuses on two key research questions: **RQ1**: Is there a temporal misalignment between static benchmarks, present LLMs, and the real world? **RQ2**: If so, how does such a temporal misalignment influence LLM evaluation? To answer these questions, we tailor a three-stage pipeline to conduct a temporal comparison between benchmarks, LLMs, and the real world. Specifically, we first extract time-sensitive questions from popular benchmarks and then search for the corresponding latest answers from the Internet. Finally,

[1] `https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population`

we compare and analyze the temporal misalignment between the gold benchmark answers, LLM responses, and searched answers. Extensive experiments are implemented across 10 LLMs and 5 benchmarks. The results illustrate significant temporal misalignment between static benchmarks, present LLMs, and the real-world facts (RQ1). Three quantitative scores – Drift Score, Cohen's Kappa, and Factual Staleness Score – are introduced to measure the temporal misalignment. The systematic analysis reveals that evaluating LLM factuality using existing benchmarks without updates will lead to unreliable comparisons in substantial past and future works, raising concerns about the trustworthiness of LLM evaluation. We hope that our study can inspire the NLP community to incorporate temporal awareness with LLM evaluation in the future.

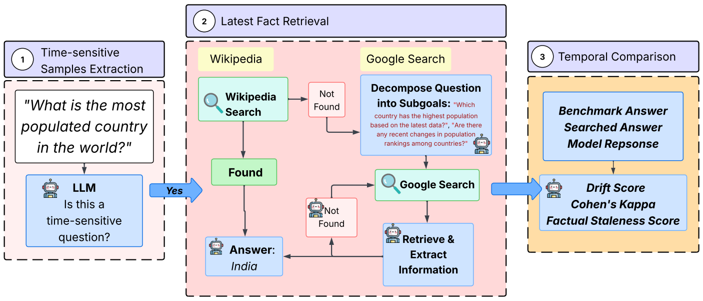

1 Time-sensitive Samples Extraction
"What is the most populated country in the world?"
LLM
Is this a time-sensitive question?
Yes
2 Latest Fact Retrieval
Wikipedia
Google Search
Wikipedia Search
Not Found
Decompose Question into Subgoals: "Which country has the highest population based on the latest data?", "Are there any recent changes in population rankings among countries?"
Found
Google Search
Not Found
Retrieve & Extract Information
Answer: India
3 Temporal Comparison
Benchmark Answer
Searched Answer
Model Repsonse
Drift Score
Cohen's Kappa
Factual Staleness Score

Figure 1: Experimental pipeline with three stages: (1) Time-sensitive Samples Extraction, (2) Latest Fact Retrieval with Wikipedia and Google Search, (3) Temporal Comparison with quantitative metrics

2 Experimental Setup

To comprehensively explore how time affects LLM evaluation, we tailor a multistage pipeline to compare model responses, benchmark answers, and the latest real-world information, in Figure 1.

2.1 Time-sensitive Samples Extraction

To focus on temporal misalignment, we first extract time-sensitive samples from commonly used LLM factuality benchmarks (Figure 3), 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]. Inspired by the SimpleQA collection criteria [Wei et al., 2024], we define a *time-sensitive question* as one that has a verifiable factual answer that changes over time. Time-sensitive questions are identified for each benchmark by an LLM with human evaluation (details in Appendix C.2 and B.1).

2.2 Latest Fact Retrieval

Next, we use the Internet to retrieve up-to-date answers for the time-sensitive questions. 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[2]. Secondly, GPT-4o-mini[3] 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) decides if further refinement and search are needed. Detailed workflow is discussed in Appendix C.3. Three annotators with instruction in Figure 11 reviewed 105 samples to determine whether the answer accurately reflect the searched evidence. The process achieves 89.52% accuracy with moderate inter-annotator agreement ($\kappa = 0.6$).

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

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

2.3 Temporal Comparison

To measure the temporal difference, we define the **Drift Score** ($\mathcal{DS}$). Given the query x_i and the gold benchmark answer $\hat{y}_i$ in each sample from the time-sensitive subset $\mathcal{D}_{\text{ts}}$ of a benchmark $\mathcal{D}$, the corresponding LLM response y_i, and the real-world answer y_i^* searched from the Internet, we compute two binary alignment scores. $s_i^{\text{gold}}(x_i, \hat{y}_i, y_i) \in \{0, 1\}$ is the agreement between y_i and $\hat{y}_i$. $s_i^{\text{search}}(x_i, y_i, y_i^*) \in \{0, 1\}$ represents the agreement between y_i and y_i^*. The **Drift Score** is: $\mathcal{DS} = \frac{1}{|\mathcal{D}_{\text{ts}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{ts}}} \left(s_i^{\text{search}} - s_i^{\text{gold}}\right)$. A positive $\mathcal{DS}$ indicates that the model responses align more with the latest facts than with the benchmark, highlighting temporal fact drift from modern large language models to static benchmarks and the real world. Finally, we evaluate 10 diverse LLMs on 5 commonly used benchmarks. The $\mathcal{DS}$ is shown in Table 1. All details are shown in Appendix C.1 and C.4.

Table 1: $\mathcal{DS}$ of models on 5 benchmarks (%). The $\mathcal{DS}$ above 0 means models align more closely with current information than static benchmark labels.

Model / Dataset (Release Time)	**TriviaQA** July 2017	**BoolQ** May 2019	**NaturalQuestion** July 2019	**TruthfulQA** May 2022	**SelfAware** July 2023
Llama-2-7B-chat-hf (Jul 2023)	4.78	-7.78	-2.76	3.75	5.07
Llama-3-8B-Instruct (Apr 2024)	0.00	-11.78	-4.01	-7.50	13.77
Llama-3.1-8B-Instruct (Jul 2024)	-3.59	-12.44	0.13	-2.50	6.52
Llama-3.2-3B-Instruct (Sep 2024)	-2.39	-10.22	-1.88	1.25	1.45
Ministral-8B-Instruct-2410 (Sep 2024)	8.37	4.89	3.51	1.25	5.80
GPT-4o-mini-2024-07-18 (Jul 2024)	7.57	4.44	13.03	15.62	14.86
Qwen2.5-1.5B-Instruct (Sep 2024)	5.58	-7.56	0.13	-0.62	9.06
Qwen2.5-3B-Instruct (Sep 2024)	4.38	0.00	4.51	4.37	14.13
Qwen2.5-7B-Instruct (Sep 2024)	2.79	2.67	3.01	6.25	10.87
Qwen2.5-14B-Instruct (Sep 2024)	3.98	4.67	6.27	0.62	15.94

3 Results, Analysis and Discussion

3.1 Main Results and Discussion

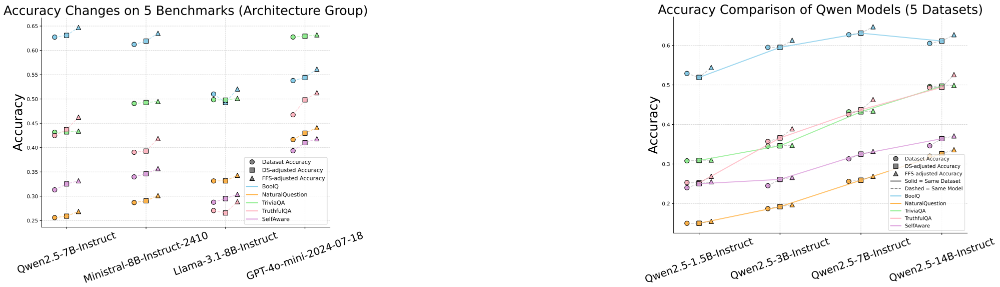
RQ1: There is a significant temporal misalignment between static benchmarks, modern LLMs, and the real world. In Table 1, 70% positive $\mathcal{DS}$ indicate that LLM responses align more with searched information than the gold answers from benchmarks. This pattern is observed consistently across the 5 evaluated datasets, especially for SelfAware, whose data are from SQuaD [Rajpurkar et al., 2016], TriviaQA [Joshi et al., 2017], and HotpotQA [Yang et al., 2018]. Figure 14 shows Cohen's Kappa κ between each other among LLM responses, searched information, and gold benchmark answers (Appendix D.1). The low κ (mostly < 0.6) emphasizes the temporal misalignment. The predominant coverage of blue polygons over red polygons underscores that contemporary LLMs align more closely with up-to-date factual information than with static benchmark datasets.

RQ2: Temporal misalignment introduces unreliability in evaluating time-sensitive samples, raising concerns about the trustworthiness of LLM evaluation. To quantify temporal drift in time-sensitive questions, we define the Factual Staleness Score ($\mathcal{FSS}$) as: $\mathcal{FSS} = \frac{1}{|\mathcal{D}_{\text{ts}}|} \sum_{i=1}^{|\mathcal{D}_{\text{ts}}|} \mathbb{1}[s_i^{\text{search}} = 1 \wedge s_i^{\text{gold}} = 0]$, which captures the fraction of questions where model outputs align with real-world facts but diverge from the dataset. As shown in Table 4, more than half of the $\mathcal{FSS}$ is larger than 10%, and the maximum $\mathcal{FSS}$ reaches up to 24% across datasets. If these static benchmarks continue to be used without updating, the resulting evaluation of LLM factuality will not be reliable.

3.2 Dataset Analysis

Increasing usage of static benchmarks with outdated information. The release time in Table 1 suggests that the benchmarks we investigated are outdated, and there is a significant time gap between the benchmarks, present LLMs, and real-world facts. The upward trend in Google Scholar citations for these benchmarks is shown in Figure 3. 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 attention to the unreliable use of outdated benchmarks.

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 5, models consistently exhibit more negative $\mathcal{DS}$ when performing passage-grounded inference. For example, Qwen2.5-7B-Instruct's $\mathcal{DS}$ 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.



(a) Different Model Architecture Families

(b) Different Model Sizes (Qwen Family)

Figure 2: Performance comparison of LLMs across (a) architecture families and (b) model sizes. Accuracies are: Dataset Accuracy, $\mathcal{DS}$-adjusted Accuracy, and $\mathcal{FSS}$-adjusted Accuracy.

3.3 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 $\mathcal{DS}$-adjusted accuracy $a_{\mathcal{DS}} = a_o + \mathcal{DS} \cdot \frac{|\mathcal{D}_{\text{ts}}|}{|\mathcal{D}|}$, and the $\mathcal{FSS}$-adjusted accuracy $a_{\mathcal{FSS}} = a_o + \mathcal{FSS} \cdot \frac{|\mathcal{D}_{\text{ts}}|}{|\mathcal{D}|}$. 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 2a. 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 2b) 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 [Qwen et al., 2025] will cover more recent information.

4 Conclusion

This study comprehensively investigates the temporal misalignment between static LLM benchmarks, present LLMs, and the real world, which reveals that reliance on outdated benchmarks leads to unreliable LLM factuality evaluation. Our analysis and discussion emphasize the importance of considering the temporal alignment between benchmark construction, model training, and the evolution of real-world facts. We hope this work motivates future research to incorporate temporal awareness into the design and evaluation of benchmarks, thereby enhancing the accuracy and robustness of LLM assessments.

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.

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.

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. 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`.

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`.

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`.

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

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`.

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`.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

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

A.1 Evaluating Large Language Models

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 the capabilities of large language models over time. 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.

A.2 Benchmark Updating and Temporal Misalignment

As factual knowledge evolves, static benchmarks can quickly become outdated, leading to differences between the knowledge the benchmark evaluates for and the current state of the world. This phenomenon, defined as temporal misalignment, is highlighted in recent studies, which demonstrate 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. These efforts emphasize the importance of incorporating temporal dynamics into benchmark design to ensure accurate and meaningful LLM evaluations.

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.

Property	TriviaQA	BoolQ	NQ (dev)
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	FreshQA
Year	2022	2023	2024
# QA Pairs	817	2,475	600
TS %	19.58	11.15	42.83
Type	Open QA	Open QA	Open QA

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

B.2 Google Scholar Citation Trend

To estimate the influence of each benchmark, we measure its citation trend using Google Scholar [4]. The citation data we collect is as of July 28, 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.

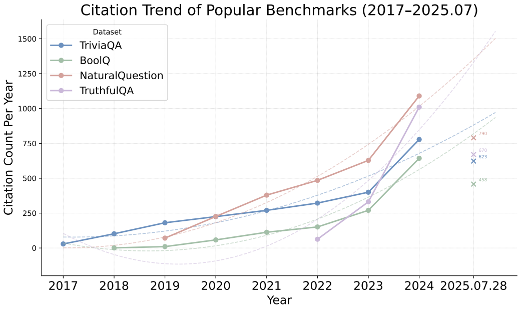


Figure 3: Trends of yearly Google Scholar citation for popular LLMs Benchmarks.

C Experiment Details

C.1 Experiment Groups

1. **Architectural analysis**: Models from different architectures released in a similar timeframe (June-October 2024): Qwen2.5-7B-Instruct [5], Ministral-8B-Instruct-2410 [6], Llama-3.1-8B-Instruct [7], and GPT-4o-mini [8] . Controlling for release date isolates the effects of architectural differences on temporal knowledge retention.

2. **Model scale analysis**: Qwen2.5 models of varying sizes (1.5B, 3B, 7B, and 14B) released simultaneously in September 2024 [Qwen et al., 2025], isolating the effect of model scale.

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

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

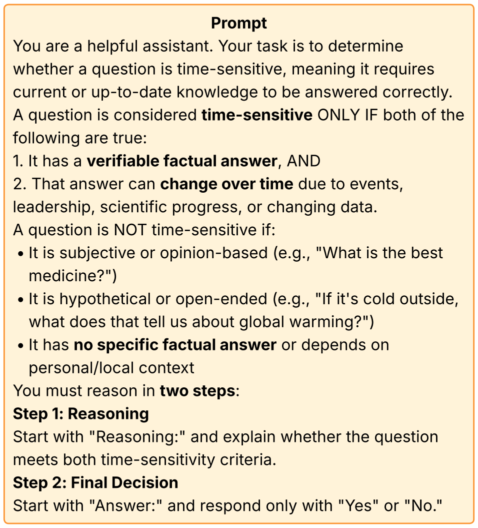
[6] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

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

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[9] for efficient inference. To reduce the randomness in LLM responses when identifying time-sensitive questions, we apply the prompt shown in Figure 4 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 while maintaining 100% recall for time-sensitive questions.



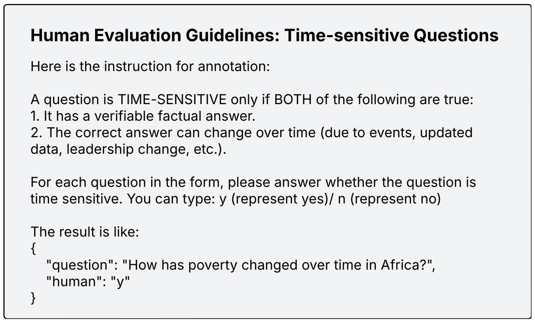
Prompt

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



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 5: Instructions for Human Evaluation of Time-sensitive Questions

To further validate the labeling quality, we conduct a human evaluation. Given the instruction in Figure 5, three domain experts manually annotate 150 questions, with results presented in Table 3. All 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.

[9] https://github.com/vllm-project/vllm

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

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

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 6, 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.

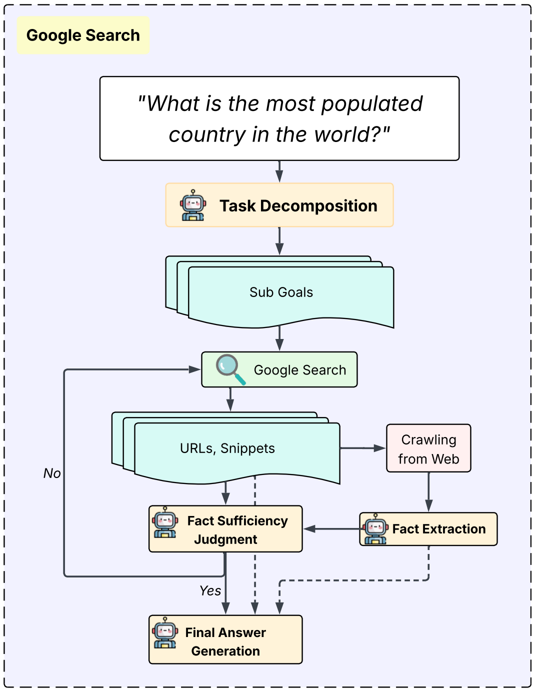
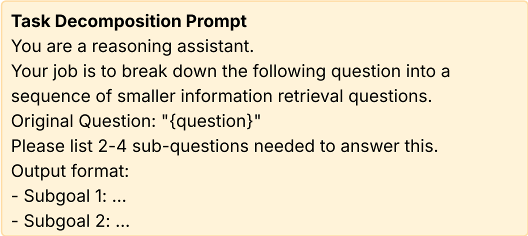


Figure 6: Workflow of Google Search and fact retrieval.

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 7–10.

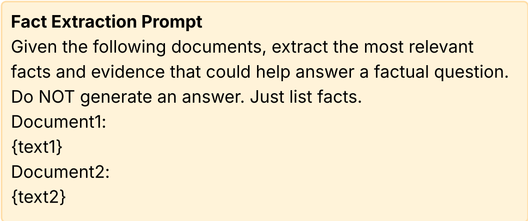
To assess the quality of web search output, we adopt two complementary methods:

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 11.



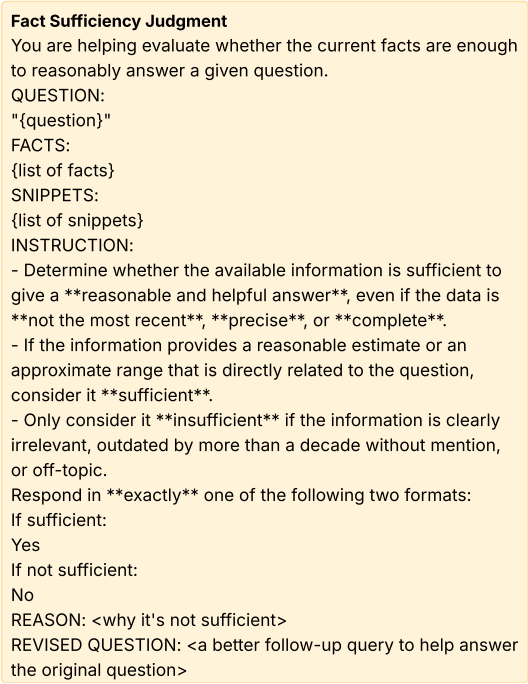
```
Task Decomposition Prompt
You are a reasoning assistant.
Your job is to break down the following question into a
sequence of smaller information retrieval questions.
Original Question: "{question}"
Please list 2-4 sub-questions needed to answer this.
Output format:
- Subgoal 1: ...
- Subgoal 2: ...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

Figure 7: Task Decomposition Prompt



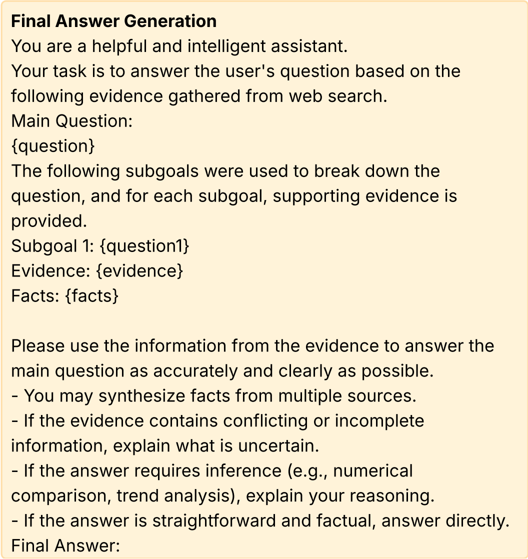
```
Fact Extraction Prompt
Given the following documents, extract the most relevant
facts and evidence that could help answer a factual question.
Do NOT generate an answer. Just list facts.
Document1:
{text1}
Document2:
{text2}
```

Figure 8: Fact Extraction Prompt



```
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>
```

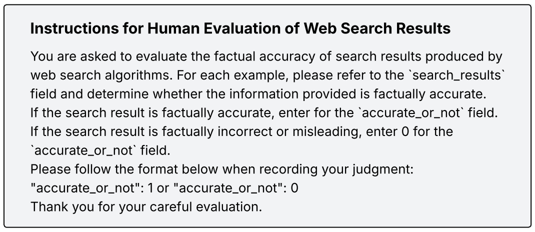
Figure 9: Fact Sufficiency Judgment Prompt



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:

Figure 10: Final Answer Generation Prompt



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.

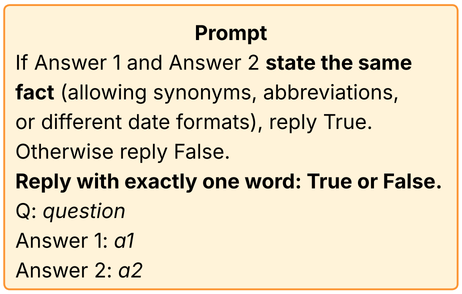
Figure 11: Human Evaluation Instruction for Web Search Results

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 14, 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 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 12 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 13. outputs yields the following agreement: the accuracy is 97% and the average Cohen's Kappa between three evaluators is 0.72.



Prompt

If Answer 1 and Answer 2 **state the same fact** (allowing synonyms, abbreviations, or different date formats), reply True. Otherwise reply False.

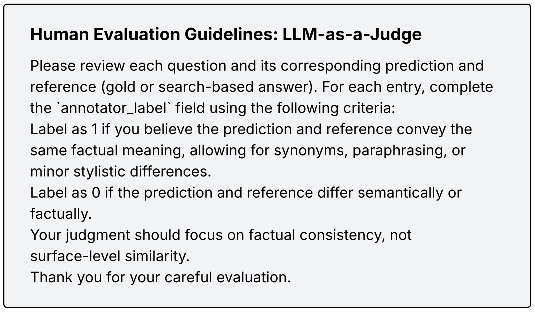
Reply with exactly one word: True or False.

Q: *question*

Answer 1: *a1*

Answer 2: *a2*

Figure 12: 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 13: Human Evaluation Instructions for LLM-as-a-judge

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.

Figure 14 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.

D.2 Factual Staleness Score

To further quantify factual degradation over time, we compute the Factual Staleness Score (FSS), which captures the semantic drift of LLM responses compared to up-to-date web content. Specifically, FSS measures the alignment gap between model predictions and current retrieved facts for each question, with higher values indicating greater staleness.

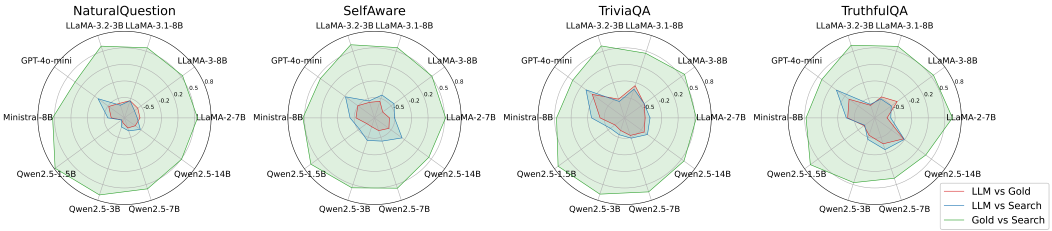


Figure 14: Cohen's Kappa Score between each other among LLMs' responses, searched real-world information, and gold benchmark answers from 4 datasets and 10 LLMs.

Model	TriviaQA	BoolQ	NaturalQuestion	TruthfulQA	SelfAware
Llama-2-7B-chat-hf	14.74	9.11	10.28	11.25	15.22
Llama-3-8B-Instruct	11.16	8.22	10.28	8.13	19.57
Llama-3.1-8B-Instruct	12.35	7.56	11.40	9.38	14.49
Llama-3.2-3B-Instruct	9.16	8.67	9.52	10.63	10.51
Ministral-8B-Instruct-2410	18.33	16.67	14.04	14.38	15.22
GPT-4o-mini-2024-07-18	19.92	17.11	24.06	23.13	22.10
Qwen2.5-1.5B-Instruct	9.96	10.67	5.51	8.13	13.41
Qwen2.5-3B-Instruct	10.36	13.56	10.03	16.25	18.48
Qwen2.5-7B-Instruct	10.76	14.44	12.41	19.38	16.67
Qwen2.5-14B-Instruct	13.55	16.00	16.04	16.88	22.46

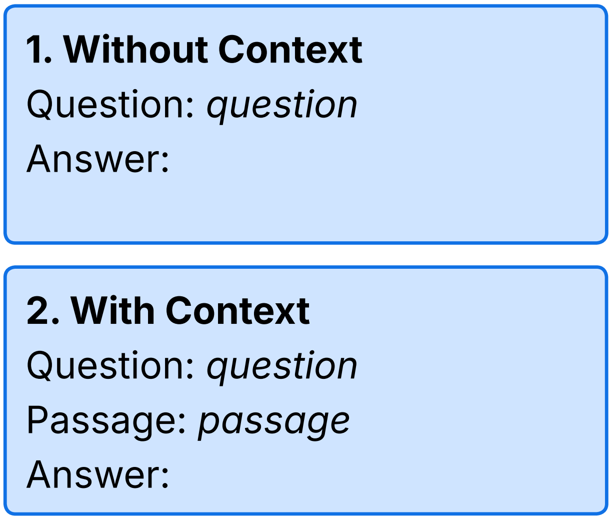
Table 4: Factual Staleness Score (FSS) (%) on five datasets when substituting model generations with web-searched answers. Cells highlighted in green indicate substantial gain (>15%) from incorporating the latest information.

We apply this metric over thousands of QA samples from each dataset, allowing us to track model factuality at scale. Notably, although only a small subset of the questions are highly time-sensitive (e.g., involving politics, events, or leadership changes), the majority of questions evolve more slowly over time. As a result, while the average FSS across all questions remains relatively modest—around 2%—we observe significantly higher staleness among the time-sensitive subset. This discrepancy highlights that benchmarks without temporal annotations may obscure the actual factual degradation that occurs in dynamic contexts.

Together with the Cohen's Kappa analysis, FSS reinforces the importance of time-aware benchmarks that can reveal subtle factual misalignments and content aging in LLMs.

D.3 BoolQ Drift Score 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 15. One setting provides both the passage and the question, while the other includes only the question without any supporting passage.



1. Without Context
Question: *question*
Answer:

2. With Context
Question: *question*
Passage: *passage*
Answer:

Figure 15: Two prompt formats used in BoolQ experiments: with and without passage context.

Interestingly, we observe a significant increase in Drift Score 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 Drift Score 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 Drift Score gap 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.

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
Qwen2.5-1.5B-Instruct	-7.56	-18.89
Qwen2.5-3B-Instruct	0.00	-14.00
Qwen2.5-7B-Instruct	2.67	-12.22
Qwen2.5-14B-Instruct	4.67	-13.56

Table 5: Drift scores (%) on BoolQ with and without passages. Positive values (highlighted) indicate alignment between model outputs and current web results, diverging from outdated benchmark gold answers.