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

Xunyi Jiang, Dingyi Chang, Xin Xu*
University of California, San Diego

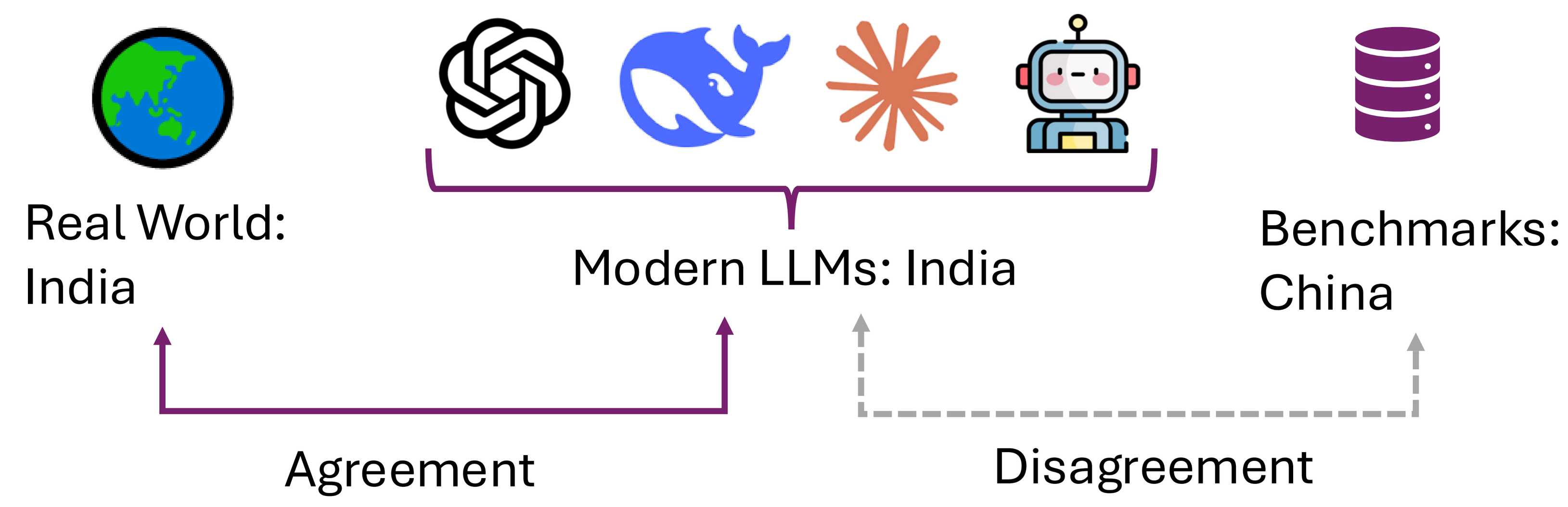
NeurIPS 2025 Workshop
Evaluating the Evolving LLM Lifecycle

Github

Motivation

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.

What is the most populated country in the world?

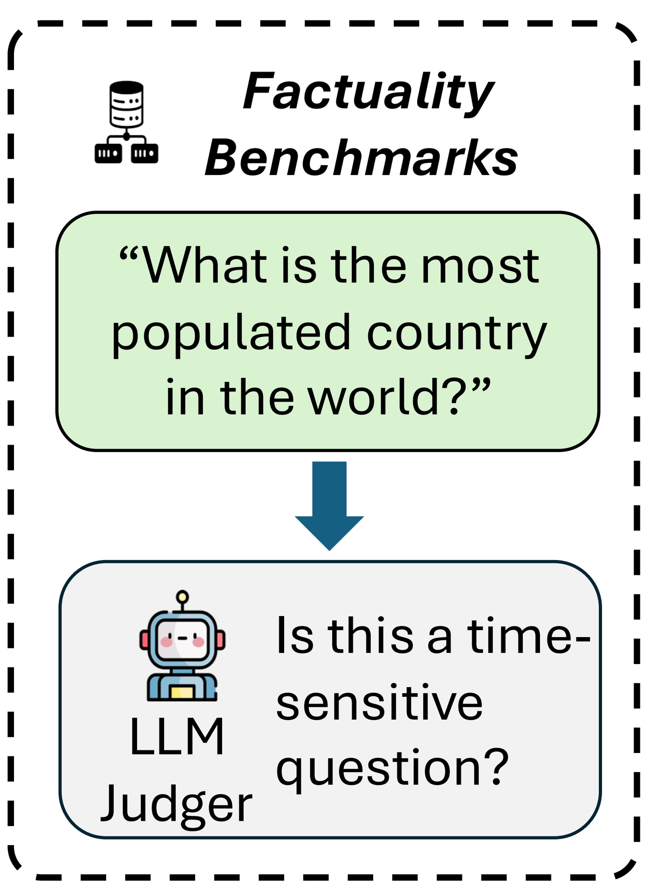


LLMs that provide up-to-date and factually correct answers may be unfairly penalized when evaluated against outdated benchmarks.

Time-sensitive Samples Extraction

Time-sensitive question definition:

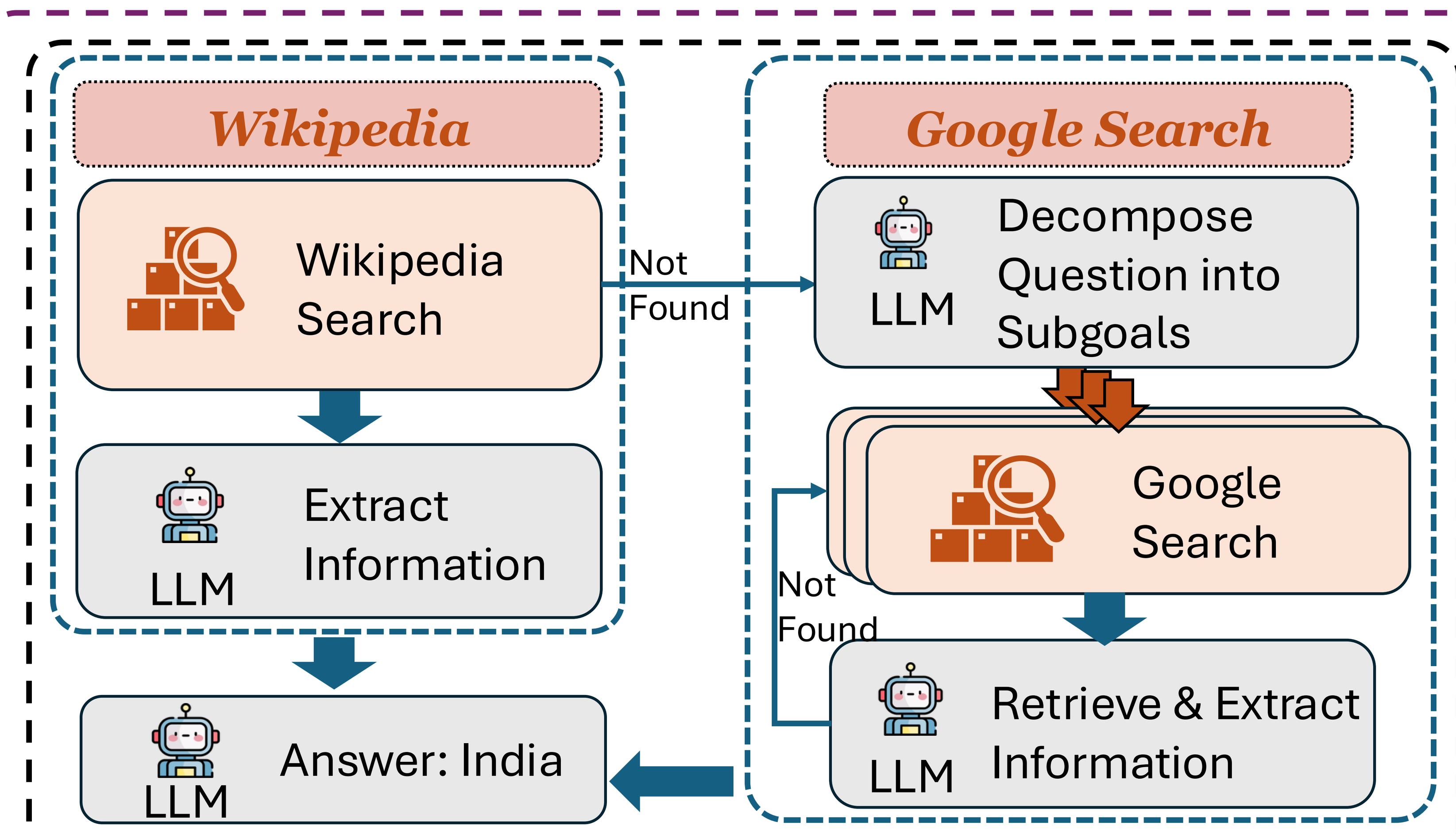
- **Verifiable factual answer**
- **Answer changes over time**



Human evaluation of time-sensitive question detection.

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

Latest Fact Retrival



If Wikipedia search fails, we will switch to Google search

Stage1: Wikipedia Search

- Retrive related information from Wikipedia
- Extract final answers from retrieved information

Stage2: Google Search

- Decompose questions into sub-goals
- Run Google search of sub-goals
- Extract key facts and temporal metadata
- Decide whether need further search

Temporal Comparison

Dataset Drift Score

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

where $|D_{ts}|$ is the number of time-sensitive data

Evaluation Misleading Rate

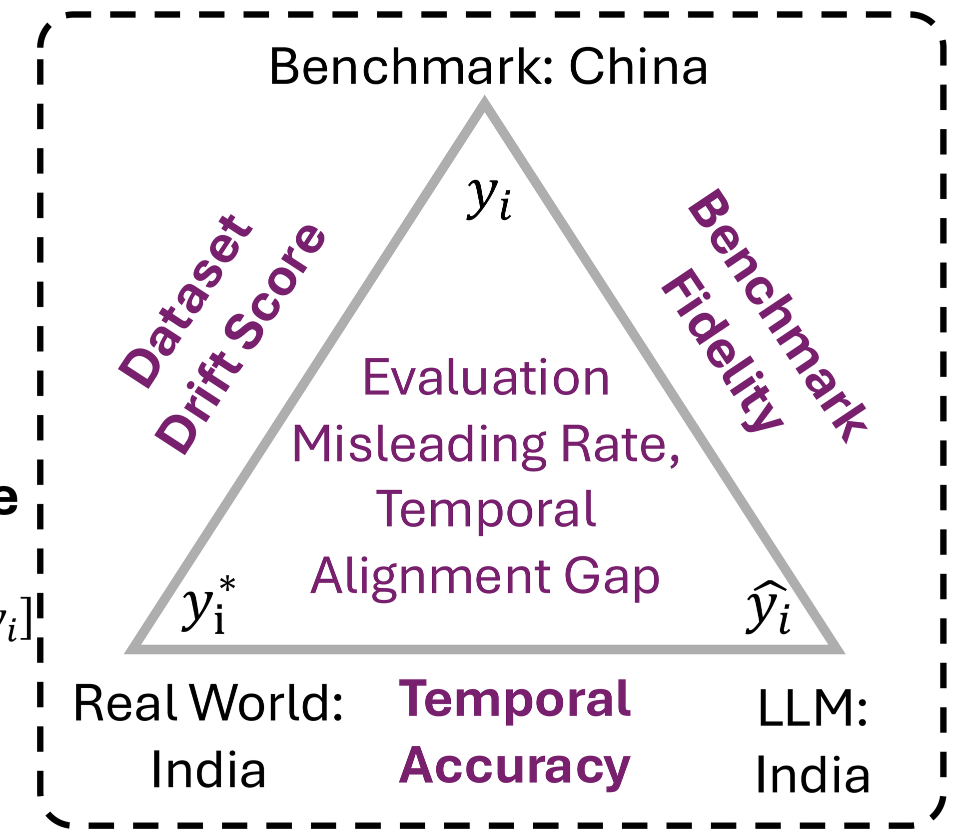
$$EMR = \frac{1}{|D_{ts}|}\sum_{i=1}^{|D_{ts}|} 1[\hat{y}_i = y_i^* \wedge \hat{y}_i \neq y_i]$$

Temporal Alignment Gap

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

where $s_i^{\text{gold}} = 1[\hat{y}_i = y_i]$ is the agreement between y_i and $\hat{y}_i$.

$s_i^{\text{search}} = 1[\hat{y}_i = y_i^*]$ is the agreement between $\hat{y}_i$ and y_i^*.



Experimental Results and Analysis

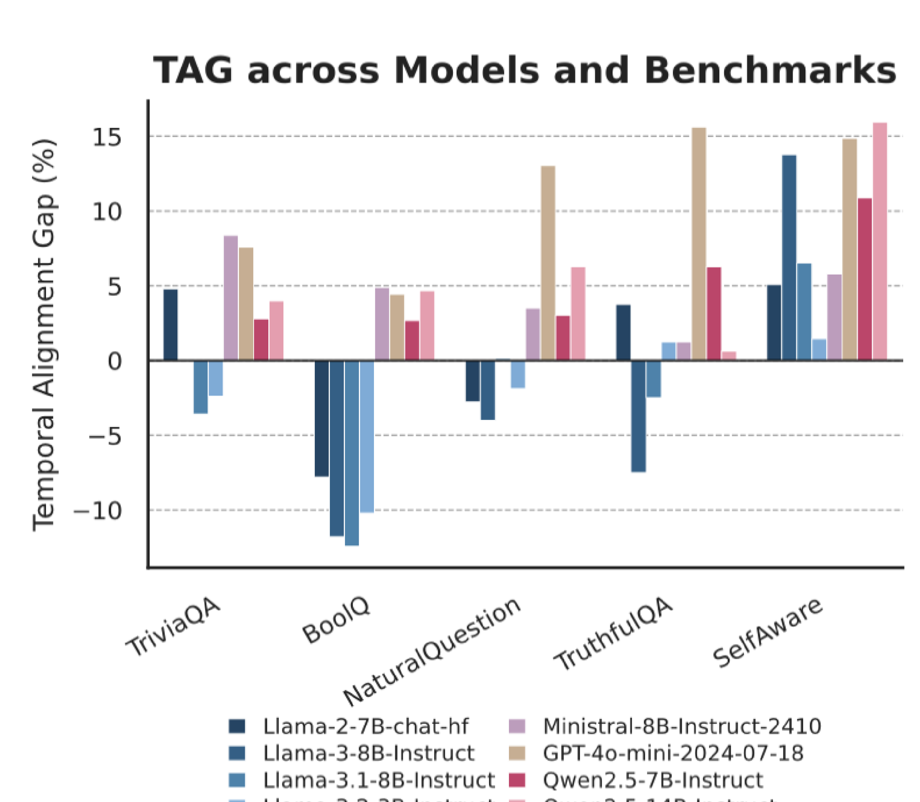
RQ1: To what extent do widely used static benchmarks contain outdated factual answers compared to current real-world facts?

Ans: A Considerable Portion of the Benchmarks Are Outdated

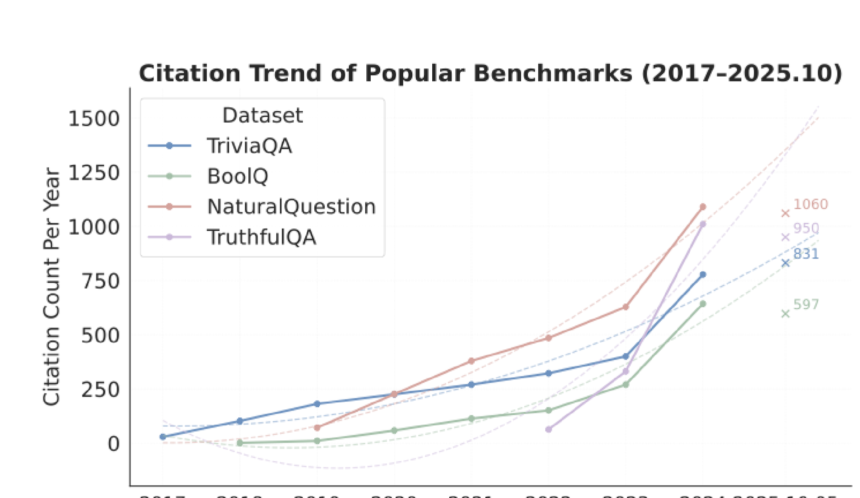
Dataset Release Time	**TriviaQA** July 2017	**BoolQ** May 2019	**NaturalQuestion** July 2019	**TruthfulQA** May 2022	**SelfAware** July 2023
Dataset Drift Score (%)	37.05	63.78	24.19	36.88	28.26
LLM (Release Time)	**Evaluation Misleading Rate (%)**				
Llama-2-7B-chat-hf (Jul 2023)	14.74	9.11	10.28	11.25	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

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

Ans: Benchmark Aging Affects the Reliability of LLM Evaluation



- The outdated benchmarks can mislabel factually correct model responses.
- The present LLMs are more aligned with real-world facts than with gold answers in the benchmarks.



- The usage of static benchmarks with outdated information is increasing.
- The outdated contexts amplify the temporal misalignment.