Custom Project: A Fairer Evaluation for the Time Awareness in Instruction-Tuned LLMs

Ernie Chu schu23@jhu.edu **Junhyeok Lee** **Dengjia Zhang**

Motivation, Related Work and Hypothesis

Large language models (LLMs) have achieved strong performance across a range of natural language tasks [1, 2, 3, 4]. Despite these successes, temporal context—crucial for factual accuracy—is often under-evaluated. For example, answering "Who is the US President?" requires temporal context.

To fill this gap, Herel et al. [5] introduced the TimeAware dataset and the TimeShift framework, which evaluates models by computing log probabilities for time-sensitive declarative statements across different temporal prefixes. While effective for base models, this method may disadvantage instruction-tuned LLMs, which are optimized for instruction-following and question-answering tasks, not sentence completion probabilities.

To address this mismatch, we propose a more natural evaluation format that better aligns with how instruction-tuned LLMs are used in practice. Instead of using log-probability scoring of dated statements, we reframe the task as question-answering over time-sensitive events. Using the same dataset, we generate QA pairs and evaluate model responses for correctness using GPT-4o as a judge.

This approach not only complements log-probability-based methods but also reflects real-world use cases more faithfully. It enables a fairer comparison across model types while providing deeper insight into LLMs' temporal reasoning capabilities.

Dataset and Method

To address the potential biases inherent in log-probability-based evaluations of temporal awareness, particularly for instruction-tuned LLMs, we propose a refined method centered on question-answering (QA) paradigms. This approach aims to more accurately reflect real-world user interactions and provide a fairer assessment of LLMs' ability to recall and utilize time-sensitive factual information. Our method leverages the TimeAware dataset [5], which comprises over 8,000 event descriptions paired with precise timestamps, as illustrated in Table 1.

Our evaluation framework consists of two distinct tasks, each designed to probe different aspects of temporal awareness in LLMs.

Task 1: Precise Date Prediction. This task evaluates the LLM's ability to directly predict the date of an event given its description. We formulate the input as a prompt asking the LLM to provide the date in YYYY/MM/DD format. The specific procedure is laid out as follows:

1. **Prompt construction:** For each event in the TimeAware dataset, we construct a prompt of the form: "When did the following event occur? Provide the date in YYYY/MM/DD format. Event: [Event Description]".
2. **Output Processing:** For open-source models, we analyze the output probability distribution to ensure the generated date adheres to the specified format. For closed-source APIs, we enforce the format through explicit prompting and discard any outputs that do not comply.
3. **Evaluation Metric:** We assess the accuracy of the predicted date by comparing it to the ground truth date from the TimeAware dataset. A prediction is considered correct if it exactly matches the ground truth.

This task offers a direct measure of the LLM's recall of event timestamps, providing insights into its temporal knowledge retention.

Table 1: Examples from the TimeAware dataset [5] showcasing event descriptions with timestamps.

Event	Date
China's government appoints Li Qiang, a close ally of President Xi Jinping ...	2023/03/11
Four people, including the perpetrator, are killed in a vehicle attack in Rochester ...	2024/01/01

Task 2: Time-Conditioned Question Answering This task simulates real-world user interactions by posing factual questions about events within specific temporal contexts. It aims to evaluate the LLM's ability to provide accurate and contextually relevant answers. The detailed procedure is described as follows:

1. **Question Generation:** We employ GPT-4o to generate a factual question based on each event description in the TimeAware dataset. This ensures that the questions are natural and relevant to the event.
2. **Propmt preparation:** We introduce temporal context by specifying a particular date or time frame in the prompt. For instance, we might ask, "What happened on [Date]?" or "What was the major event in [Year]?".
3. **Answer Generation:** The LLM generates an answer to the posed question, considering the provided temporal context.
4. **Answer Evaluation:** We utilize GPT-4o as an evaluator to assess the correctness and relevance of the generated answer against the ground truth event description from the TimeAware dataset. This evaluation considers the semantic meaning and factual accuracy of the response.

For example:

- **Event (Date):** Croatia adopts the euro and joins the Schengen Area. (2023/01/01)
- **Possible question generated by GPT-4o:** What major economic and travel-related changes did Croatia implement on January 1, 2023?
- **Possible answer generated by LLMs:** On January 1, 2023, Croatia adopted the euro as its official currency and joined the Schengen Area.

This task is specifically designed for instruction-tuned models, as it aligns with their training objectives of answering questions and following instructions. The use of GPT-4o as an evaluator mitigates the subjectivity inherent in open-ended question answering, ensuring a more objective and consistent assessment.

Experiments

We plan to run our proposed evaluation on a range of open-sourced models from leading LLMs vendors, including Llama 3 [6], Gemma 2 [7], Phi 4 [8], and Qwen 2.5 [9], as well as close-sourced APIs, including GPT[1], Gemini[2], Grok[3], and Claude[4], for both of their pretrained base models (if available) and the instruction-tuned models.

Expected Outcome

By adopting this QA-based evaluation framework, we aim to provide a more comprehensive and nuanced understanding of LLMs' temporal awareness, addressing the limitations of traditional log-probability-based methods and fostering a more accurate reflection of real-world performance. We expect the instruction-tuned models would be more competitive under this framework.

Midway Goal

We plan to include our preliminary results in the midway progress report, which will cover performance metrics for all open-source models on both Task 1 and Task 2. Following the report, our focus will shift to evaluating the performance of closed-source APIs and compiling our findings into the final paper.

[1] `https://platform.openai.com`

[2] `https://ai.google.dev`

[3] `https://x.ai/api`

[4] `https://www.anthropic.com/api`

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report. Technical report, OpenAI, 2024. URL `https://arxiv.org/abs/2303.08774`.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[4] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

[5] David Herel, Vojtech Bartek, Jiri Jirak, and Tomas Mikolov. Time awareness in large language models: Benchmarking fact recall across time. *arXiv preprint arXiv:2409.13338*, 2024. URL `https://arxiv.org/abs/2409.13338`.

[6] Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[7] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[8] Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[9] An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.