JOHNS HOPKINS UNIVERSITY

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

Ernie Chu, Junhyeok Lee, Dengjia Zhang Johns Hopkins University

Summary

Motivation: Time awareness in LLMs is crucial for real-world factual accuracy. However, existing protocols like **TimeShift** assess temporal knowledge by **comparing log probabilities** of dated statements for all possible dates, which does not align with the actual human-LLM interaction and may unfairly penalize instruction-tuned models optimized for more natural QA.

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

Contribution: We proposed a QA-based evaluation built on the TimeAware dataset that enables a user-centric assessment of temporal understanding. Instead of comparing logprobs, **we ask LLMs to respond with date directly**. In addition, we synthesize **time-dependent factual questions** from event descriptions to test if LLMs know what happened on a specific date.

Method

We introduce a QA-based evaluation to overcome biases in log-probability metrics and more accurately mirror real user interactions, providing a fairer measure of instruction-tuned LLMs' time-sensitive factual recall.

- **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: *"Answer in YYYY/MM/DD, on what date was this news event reported? {Event Description}. Answer: "*.
2. **Controlled Generation:** We analyze the output probability distribution to ensure the generated date adheres to the specified format.
3. **Evaluation Metric:** We assess the multi-level accuracy of the predicted date by comparing it to the ground truth date from the TimeAware dataset.

- **Task 2: Time-Conditioned Question Answering.**

This task simulates real-world user interactions by posing factual questions about events within specific temporal contexts. The detailed procedure is described as follows:

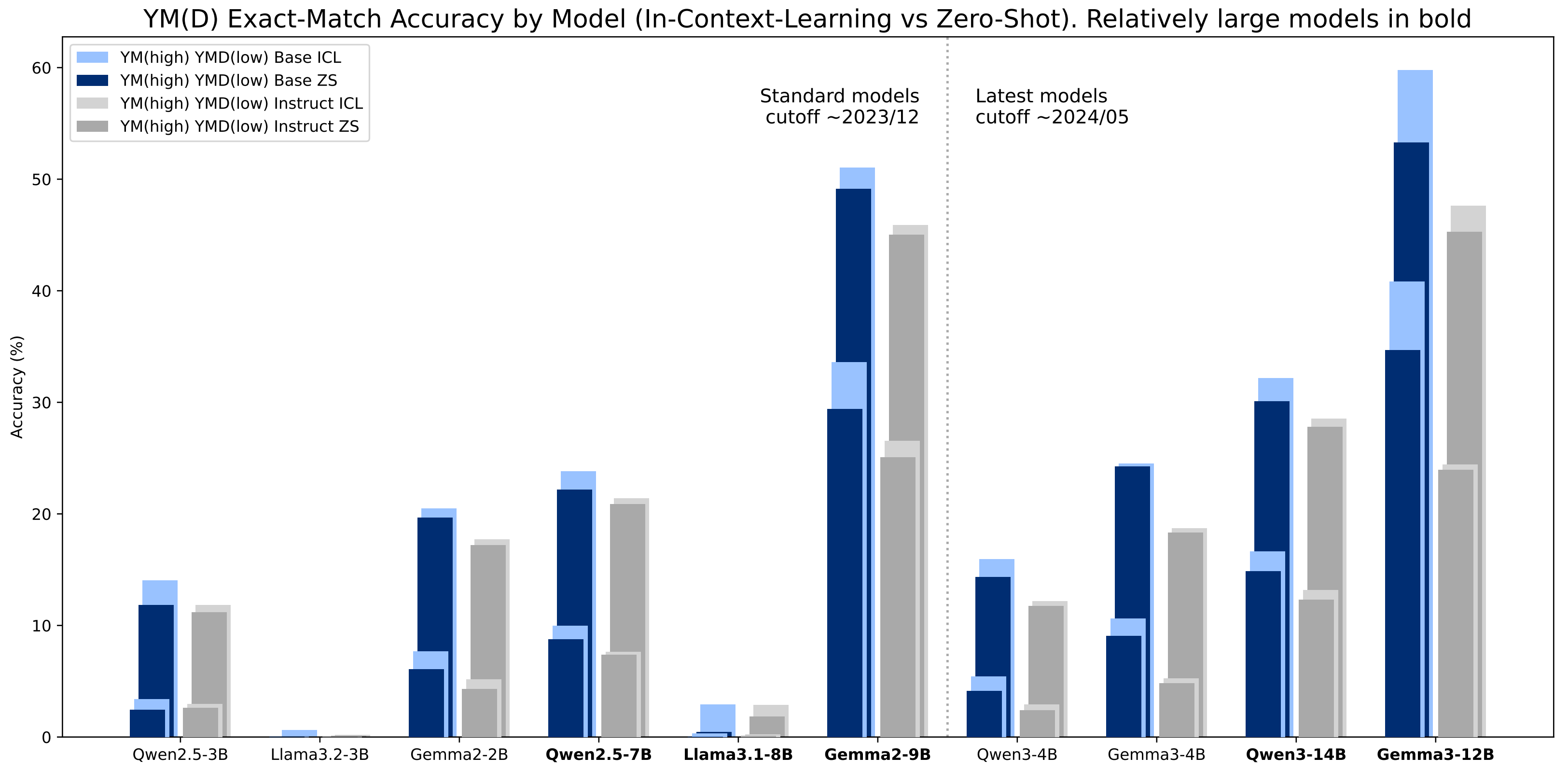
Method (Continued)

1. **Question Generation:** We employ GPT-4o to generate factual questions based on each event description in the TimeAware dataset.
2. **Prompt 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}?".
3. **Answer Generation:** We test LLMs to generate answers to the posed questions, considering the provided timeframe.
4. **Answer Evaluation:** We utilize open-source models as LLM judges to automatically assess whether the generated answers echo the GT event description from the TimeAware dataset. This evaluation considers the semantic meaning and factual accuracy of the response.

We also test ICL in addition to zero-shot for all evaluations.

Task 1: Precise Date Prediction

TL;DR, Gemma > others; ICL > ZS; Base > Instruct consistently.



ZS: zero shot. ICL: In Context Learning. YMD $\pm n$D: YMD with n days tolerance (within the correct month)

Model (ZS/ICL)	Y	YM	YMD	YMD ±3D	YMD ±5D	YMD ±10D
Qwen2.5-3B	39.36/44.30	11.85/14.02	2.46/3.38	5.47/6.66	6.82/8.31	9.19/10.75
Qwen2.5-3B-it	39.81/41.71	11.17/11.83	2.63/2.95	5.31/5.99	6.62/7.27	8.64/9.43
Llama3.2-3B	0.00/13.27	0.00/0.64	0.00/0.04	0.00/0.13	0.00/0.15	0.00/0.33
Llama3.2-3B-it	0.55/3.58	0.12/0.19	0.01/0.03	0.04/0.03	0.04/0.04	0.07/0.07
Gemma2-2B	**47.52/52.56**	**19.65/20.49**	**6.09/7.68**	**10.60/12.30**	**12.22/13.68**	**15.57/16.73**
Gemma2-2B-it	44.50/49.69	17.19/17.72	4.31/5.15	8.73/9.57	10.16/11.04	13.52/13.81
Qwen2.5-7B	48.98/53.96	22.16/23.81	8.78/9.98	14.42/15.07	15.91/16.72	18.74/19.86
Qwen2.5-7B-it	47.96/49.29	20.88/21.41	7.36/7.62	12.66/12.82	14.41/14.72	17.30/17.84
Llama3.1-8B	2.37/28.54	0.44/2.94	0.01/0.32	0.07/0.77	0.10/1.00	0.15/1.47
Llama3.1-8B-it	8.65/26.36	1.86/2.88	0.09/0.25	0.42/0.65	0.61/0.89	0.86/1.53
Gemma2-9B	**66.89/69.18**	**49.15/51.05**	**29.39/33.58**	**41.68/43.14**	**43.62/45.22**	**46.21/48.01**
Gemma2-9B-it	62.74/64.19	45.04/45.89	25.09/26.56	36.17/37.33	38.48/39.60	41.54/42.67
Qwen3-4B	43.09/44.83	14.36/15.95	4.15/5.44	7.86/8.89	9.23/10.44	11.43/13.05
Qwen3-4B-it	37.07/38.51	11.73/12.20	2.41/2.94	5.38/6.15	6.79/7.43	9.15/9.92
Gemma3-4B	53.53/57.17	24.26/24.53	9.07/10.61	15.08/16.16	16.97/18.03	20.56/21.08
Gemma3-4B-it	46.60/48.41	18.34/18.71	4.81/5.27	9.38/9.76	11.14/11.70	14.12/14.71
Qwen3-14B	53.47/58.89	30.11/32.17	14.88/16.65	22.22/24.26	23.77/26.03	26.88/29.13
Qwen3-14B-it	51.75/55.34	27.82/28.55	12.31/13.19	19.43/20.62	21.11/22.36	24.38/25.41
Gemma3-12B	**67.76/75.33**	**53.29/59.77**	**34.67/40.84**	**46.08/52.54**	**48.12/54.64**	**50.82/57.29**
Gemma3-12B-it	64.74/68.72	45.27/47.62	23.96/24.42	34.39/35.88	36.77/38.59	40.74/42.39

Task 2: Time-Conditioned QA

We generate the factual questions with the following setup

Original Prompt	At {year}-{month}-{day}, the event {description} happened in {country}, {continent} which is related to {category}. Can you ask a time-aware question that would lead to this event as the answer?
Formatted Prompt	At 2024-1-1, the event "A magnitude 7.6 earthquake strikes Japan's western coast, killing an estimated 120 people and injuring more than 100." happened in Japan, Asia, which is related to Environment & Ecology. Can you ask a time-aware question that would lead to this event as the answer?
GPT-4o-mini	What significant natural disaster occurred in Japan on January 1, 2024, that resulted in numerous casualties?
GPT-4.1-nano	Which environmental disaster occurred in Japan's western coast on January 1, 2024, resulting in approximately 120 fatalities and over 100 injuries?

We use cross-validation to prevent LLM judges from favoring their own writing style

Model (ZS/ICL)	Qwen Judge		Gemma Judge	
	Base	Instruct	Base	Instruct
Qwen3-4B	61.56/63.69	62.57/61.91	89.01/93.27	96.51/94.42
Gemma-3-4B	64.03/70.77	62.18/66.50	91.78/98.22	96.46/98.61
Qwen3-14B	76.57/77.42	76.78/80.05	91.30/95.06	86.06/95.13
Gemma-3-12B	80.98/81.66	79.63/78.60	93.60/98.50	95.63/97.35

Conclusion

Base models still consistently outperform instruction-tuned models in our framework. Our study introduces a more user-centric time-aware QA for LLMs, and the results reaffirm the need for research on temporal-knowledge-preserving post-training.

Limitations and Future Work

1. **Unreliable Judgment from Smaller LLMs.** We used Qwen3-14B and Gemma3-12B as judges and observed that these models often produce false positives, especially when responses included repeated or overlapping text with the input prompt, leading to overly optimistic and unreliable evaluations. Use proprietary API as judge if budget allows.
2. **Interference from Reasoning Mode.** The Qwen3 series exhibits "reasoning" behavior by default. Although its answer structures are often unexpected, our LLM judges consistently fail to reject them.
3. **Extension to larger models and closed -sourced API.** Due to limited resources, we've only evaluated LLMs with up to 14B parameters. Future work will extend this analysis to larger models and proprietary APIs, enabling a more comprehensive understanding of temporal reasoning across a broader range of capabilities.