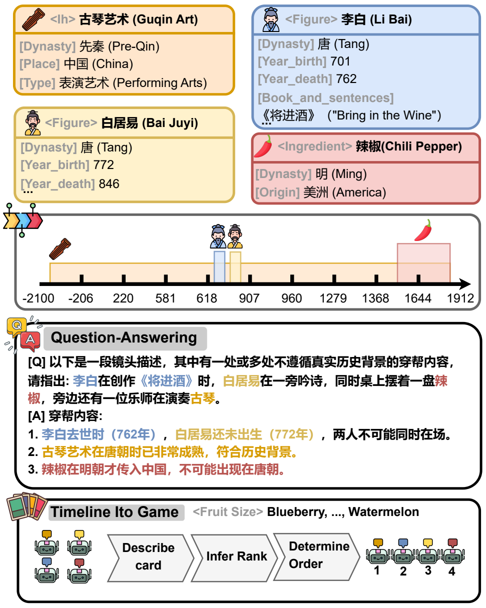
Benchmarking Temporal Reasoning and Alignment Across Chinese Dynasties

Zhenglin Wang[*♡], **Jialong Wu**[*♡◇], **Pengfei Li**[♡], **Yong Jiang**[◇], **Deyu Zhou**[♡†]
[♡] School of Computer Science and Engineering, Key Laboratory of Computer Network
and Information Integration, Ministry of Education, Southeast University, China
[◇] Tongyi Lab, Alibaba Group
`{zhenglin, jialongwu, d.zhou}@seu.edu.cn`

Abstract

Temporal reasoning is fundamental to human cognition and is crucial for various real-world applications. While recent advances in Large Language Models have demonstrated promising capabilities in temporal reasoning, existing benchmarks primarily rely on rule-based construction, lack contextual depth, and involve a limited range of temporal entities. To address these limitations, we introduce **C**hinese **Tim**e Reasoning (CTM), a benchmark designed to evaluate LLMs on temporal reasoning within the extensive scope of Chinese dynastic chronology. CTM emphasizes cross-entity relationships, pairwise temporal alignment, and contextualized and culturally-grounded reasoning, providing a comprehensive evaluation. Extensive experimental results reveal the challenges posed by CTM and highlight potential avenues for improvement.[1]



[Q] 以下是一段镜头描述，其中有一处或多处不遵循真实历史背景的穿帮内容，请指出: 李白在创作《将进酒》时，白居易在一旁吟诗，同时桌上摆着一盘辣椒，旁边还有一位乐师在演奏古琴。

[A] 穿帮内容:

1. 李白去世时（762年），白居易还未出生（772年），两人不可能同时在场。
2. 古琴艺术在唐朝时已非常成熟，符合历史背景。
3. 辣椒在明朝才传入中国，不可能出现在唐朝。

Figure 1: A QA pair from a script error correction task and an instance of the Timeline Ito Game with a "*fruit size*" theme from CTM. [2]

1 Introduction

“究天人之际，通古今之变。”
— 司马迁《史记·报任安书》

Understanding time is fundamental to human cognition and plays a pivotal role in shaping our perception and interaction with the world (Islakoglu and Kalo, 2025). Recently, Large Language Models (LLMs)have shown promising abilities in temporal reasoning (Chu et al., 2024; Su et al., 2024).

Previous benchmarks, which rely on rule-based constructed methods, lack contextualization and involve a limited number of entities in temporal relation evaluation. The core principle in assessing temporal reasoning lies in evaluating whether the model has a clear understanding of the event time within a temporal coordinate system. Compared to other temporal coordinate systems, the Chinese dynastic chronology spans a significantly longer historical scope and encompasses a broader range of culturally-grounded and historical knowledge (Sun et al., 2024; Li et al., 2024b; Yuan et al., 2024). It serves as a well-suitable background for temporal reasoning, as real-world applications can be found in various media, including films, short dramas, and novel writing, all of which rely on it.

Therefore, we introduce **C**hinese **Tim**e Reasoning (**CTM**) benchmark in this study. The comparison between CTM and other benchmarks is shown in Table 1. **CTM** focuses on contextualization, cross-entity relationships, and pair-wise temporal alignment capability. As shown in Figure 1, answering this question requires a clear temporal understanding of four entities, “李白” (701 to 762), “白居易” (772 to 826), “古琴” (Since “Pre-Qin”),

---

[*] Equal Contribution. The work was partially done during Jialong's internship at Alibaba Group.

[†] Corresponding Author.

[1] Code and dataset are available at `https://github.com/Linking-ai/ctm_bench`

[2] The English translation is presented in App. C.2.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 105–120
March 24-29, 2026 ©2026 Association for Computational Linguistics

Table 1: **Comparison** between CTM and other benchmarks. Detailed discussion is presented in Appendix A.

	Language	Construction	Time Scope	Contextualization	Temporal Alignment	Complex Aspects
TimeQA (2021)	En	Rule-based	1367–2018	✗	✗	✗
TempLAMA (2022)	En	Rule-based	2010–2020	✗	✗	✗
TempReason (2023)	En	Rule-based	634–2023	✗	✗	✗
SituatedGen (2023)	En	LLM-based	-	✓	✗	✓
CoTempQA (2024)	En	Rule-based	-	✗	✗	✗
TimeBench (2024)	En	-	-	✓	✗	✓
TRAM (2024)	En	Rule-based	-	✓	✗	✓
ChronoSense (2025)	En	Rule-based	-	✗	✗	✗
CTM	Zh	LLM-based	-2100–1912	✓	✓	✓

and "辣椒" (Since "Ming"). In addition, we develop the Timeline Ito Game to evaluate the LLM's ability to align entities across temporal and other dimensions, requiring pairwise order perception of different entities. The CTM benchmark is built upon a curated and authoritative Chinese cultural entity repository, which encompasses over 4,700 entities, spanning from figures, places, allusions, ingredients, and intangible cultural heritage.

We evaluate the performance of the CTM benchmark using various mainstream LLMs, including both closed-source and open-sourced from diverse perspectives. We conduct experiments under both zero-shot and chain-of-thought (CoT) settings (Wei et al., 2022), respectively. Further analysis shows the challenge of CTM and provides empirical insights into enhancing LLMs' temporal reasoning abilities and alignment across Chinese dynasties.

The contributions of this work are as follows: 1). We construct an interesting and challenging benchmark CTM, comprising **8,750** QA pairs and **60** instances of Timeline Ito Games. 2). We conduct extensive empirical experiments with various LLMs, which show that the proposed tasks are challenging. 3). Analysis of the experiments further reveals potential directions for addressing temporal reasoning tasks.

2 CTM Dataset

2.1 Task Definition

Question-Answering We design the below eight challenging tasks using the Question-Answering format: ***(i)*** *Entity-based Dynasty Determination* (**EDD**): infer the historical dynasty of a given entity based on contextual information. ***(ii)*** *Plausibility Judgment* (**PJ**): assess whether a described historical scenario is plausible by reasoning about temporal and factual consistency. ***(iii)*** *Temporal Order Understanding* (**TOU**): understand and compare the chronological order of historical events or figures. ***(iv)*** *Relation Reasoning* (**RR**): reason about the historical relationships between entities, such as their spatial, temporal, or functional connections. ***(v)*** *Script Error Correction* (**SEC**): identify and correct historical inaccuracies in visual or textual narratives. ***(v)*** *Entity Evolution Understanding* (**EEU**): track and understand the evolution of entity names or attributes across different historical periods. ***(vi)*** *Time Interval Calculation* (**TIC**): calculate the temporal gap between historical entities or events. ***(vii)*** *Temporal Entity Selection* (**TES**): select the correct historical entity based on temporal and contextual constraints. ***(viii)*** *Long Script Error Correction* (**LSEC**): identify and correct complex historical inaccuracies in long narratives by reasoning across extended contexts. The key aspect of these task designs is to examine LLM's ability to accurately **perceive and reason** about temporal relationships in a structured manner.[3]

Timeline Ito Game Our developed Timeline Ito Game is a collaborative reasoning game where agents infer the chronological order of historical entities within a dynasty timeline using thematic metaphors. As shown in Figure 1, the rules can be divided into the following steps:

- **Step1: Describe Card**: Agents describe their assigned historical entity using a given theme without explicit temporal references.
- **Step2: Infer Rank**: Agents collaboratively deduce their relative positions in the timeline based on shared contexts.
- **Step3: Determine Order**: Each Agent sequentially predicts their position in the timeline relative to the others, and the team's final order is based on these individual predictions.

The game ends when the team's predicted order matches the true chronological sequence or when the maximum number of rounds, K, is reached.[4]

2.2 Data Collection

Source We construct a comprehensive entity information repository by collecting diverse data

---

[3] Each task's examples are presented in App. F.

[4] We present a running case in App. F.2.

Table 2: **Main results on QA tasks** within CTM benchmark. The best results among all backbones are **bolded**, and the second-best results are underlined.

Method	**Cross Temp Count**					**Question Type**							**Avg.**
	$= 1$ (EDD)	$= 2$	$= 3$	≥ 4	$\geq 4_L$ (LSEC)	PJ	TOU	RR	SEC	EEU	TIC	TES	
Closed-Sourced LLMs													
GPT-4o	56.52	51.12	44.76	26.10	53.60	58.64	38.42	57.26	36.15	40.58	15.36	59.31	48.08
+ CoT	67.40$_{+10.88}$	58.08$_{+6.96}$	49.24$_{+4.48}$	29.60$_{+3.50}$	31.60$_{-22.0}$	64.10$_{+5.46}$	44.71$_{+6.29}$	59.62$_{+2.36}$	47.09$_{+10.94}$	44.06$_{+3.48}$	17.70$_{+2.34}$	61.68$_{+2.37}$	54.21$_{+6.13}$
Qwen-max	60.48	53.12	50.54	30.80	62.00	**64.39**	42.55	59.10	40.71	46.38	**20.87**	60.22	52.27
+ CoT	**69.56**$_{+9.08}$	**59.32**$_{+6.20}$	**54.48**$_{+3.94}$	31.90$_{+1.10}$	39.60$_{-22.40}$	63.29$_{-1.10}$	**48.58**$_{+6.03}$	**63.75**$_{+4.65}$	**55.77**$_{+15.06}$	**53.91**$_{+7.53}$	15.19$_{-5.68}$	**63.14**$_{+2.92}$	**57.24**$_{+4.97}$
o1-preview	52.80	46.56	49.64	**32.70**	**67.20**	58.28	44.28	53.01	43.16	40.87	11.02	56.02	48.24
Open-Sourced LLMs													
LLaMA3.1$_{8b}$	33.04	16.86	15.60	9.10	10.80	19.66	12.95	18.65	7.37	0.87	2.01	37.04	20.14
+ CoT	35.05$_{+2.01}$	26.44$_{+9.58}$	19.96$_{+4.36}$	10.70$_{+1.60}$	12.40$_{+1.60}$	26.48$_{+6.82}$	19.55$_{+6.60}$	23.20$_{+4.55}$	20.02$_{+12.65}$	15.70$_{+14.83}$	5.51$_{+3.50}$	34.37$_{-2.67}$	24.91$_{+4.77}$
ChatGLM3$_{6b}$	38.40	21.60	16.04	5.80	4.80	21.40	12.28	22.67	12.25	12.75	1.84	35.58	22.52
+ CoT	37.24$_{-1.16}$	22.72$_{+1.12}$	15.28$_{-0.76}$	8.20$_{+2.40}$	4.00$_{-0.80}$	20.32$_{-1.08}$	15.92$_{+3.64}$	20.12$_{-2.55}$	14.98$_{+2.73}$	16.52$_{+3.77}$	3.01$_{+1.17}$	29.74$_{-5.84}$	22.61$_{+0.09}$
InternLM2.5$_{7b}$	60.64	47.32	39.36	21.60	42.00	51.39	30.16	48.64	45.78	42.61	11.19	50.18	45.75
+ CoT	61.44$_{+0.80}$	51.40$_{+4.08}$	39.36$_{+0.00}$	20.20$_{-1.40}$	38.00$_{-4.00}$	51.70$_{+0.31}$	31.45$_{+1.29}$	49.47$_{+0.83}$	52.86$_{+7.08}$	44.19$_{+1.58}$	11.52$_{+0.33}$	48.54$_{-1.64}$	46.90$_{+1.15}$
Qwen2.5$_{7b}$	51.80	39.88	35.96	12.40	30.00	46.28	26.38	46.28	24.14	36.23	7.35	52.01	38.76
+ CoT	59.96$_{+8.16}$	47.60$_{+7.72}$	36.64$_{+0.68}$	18.30$_{+5.90}$	30.80$_{+0.80}$	52.46$_{+6.18}$	29.95$_{+3.57}$	52.18$_{+5.90}$	34.13$_{+9.99}$	40.58$_{+4.35}$	8.18$_{+0.83}$	49.64$_{-2.37}$	44.22$_{+5.46}$
Qwen2.5$_{14b}$	54.36	51.16	42.56	23.80	42.00	57.44	36.86	51.83	36.90	39.07	18.26	58.58	46.32
+ CoT	57.92$_{+3.56}$	45.44$_{-5.72}$	41.24$_{-1.32}$	22.50$_{-1.30}$	30.80$_{-11.20}$	52.73$_{-4.71}$	34.36$_{-2.50}$	46.52$_{-5.31}$	42.57$_{+5.67}$	36.81$_{-2.26}$	10.02$_{-8.24}$	51.82$_{-6.76}$	44.89$_{-1.43}$
Qwen2.5$_{32b}$	56.28	52.78	46.24	26.90	46.40	60.66	38.54	56.79	39.12	43.77	20.10	60.04	48.83
+ CoT	60.80$_{+4.52}$	49.32$_{-3.46}$	45.32$_{-0.92}$	24.80$_{-2.10}$	31.20$_{-15.20}$	50.67$_{-9.99}$	40.65$_{+2.11}$	51.12$_{-5.67}$	43.40$_{+4.28}$	40.29$_{-3.48}$	17.03$_{-3.07}$	57.12$_{-2.92}$	48.14$_{-0.69}$
Qwen2.5$_{72b}$	58.20	48.76	46.84	31.30	60.80	61.38	40.77	54.31	36.62	42.03	11.52	62.23	49.30
+ CoT	69.00$_{+10.80}$	57.24$_{+8.48}$	49.88$_{+3.04}$	32.50$_{+1.20}$	46.00$_{-14.80}$	61.50$_{+0.12}$	45.01$_{+4.24}$	61.51$_{+7.20}$	50.18$_{+13.56}$	49.86$_{+7.83}$	17.53$_{+6.01}$	59.85$_{-2.38}$	55.39$_{+6.09}$
Deepseek-R1	**70.84**	**67.12**	**60.64**	**45.50**	**72.40**	**76.63**	**58.17**	**67.30**	**59.69**	**61.16**	**24.37**	**67.70**	**64.02**

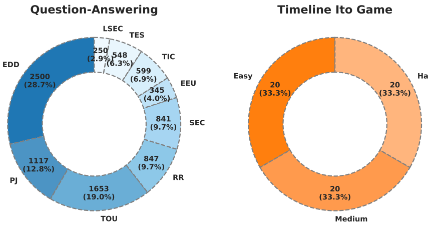


Figure 2: Statistic of CTM.

from multiple authoritative sources, *e.g.*, `Gushiwen`, `CBDB`, `CHGIS`, `Wikipedia`, and `Ihchina`. The historical dynasties are simplified into ten major periods based on `Allhistory` and `CHINA—Timeline of Historical Periods`, specifically: “先秦”, “汉”, “六朝”, “隋”, “唐”, “五代”, “宋”, “元”, “明”, “清”. The entity repository contains 1,652 figures (with attributes such as birth address, birth year, death year, and associated books or sentences), 2,907 places (including 990 primary administrative regions and 1,917 subordinate localities), 93 allusions, 49 ingredients, and 44 intangible cultural heritage items.

Annotation Process The annotation process is structured into three key steps to ensure systematic and high-quality data generation: **seed prompt creation**, **entity-aware data generation**, and **validation and quality control**.[5] The process systematically generates annotated data while aligning with the repository's structured knowledge. The statistics of CTM on the task are shown in Figure 2.

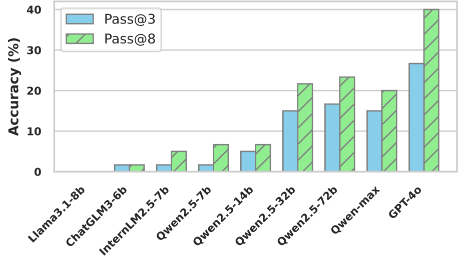


Figure 3: Average performance of Time Ito Game. Detailed results can be found in Appendix I.

2.3 Evaluation

We use the **accuracy** metric to evaluate the QA tasks while **Pass@**K is used to evaluate Time Ito Game. Due to the varying lengths of LLM-generated text, it is challenging to perform exact match evaluation. We use GPT-4o[6] as the evaluator[7], which determines the correctness of responses by comparing the prediction with the ground truth using the CoT (Wei et al., 2022). Pass@K measures whether the sequential alignment is achieved within K attempts, we set K to 3 and 8.

3 Experiments

Backbones We evaluate **twelve** mainstreaming LLMs, the complete list of models is in App. G.

---

[5] The details of each step are provided in the App. B.

[6] Used default parameters (temperature=1.0, max_len).

[7] The prompt for the evaluator is provided in Appendix H.

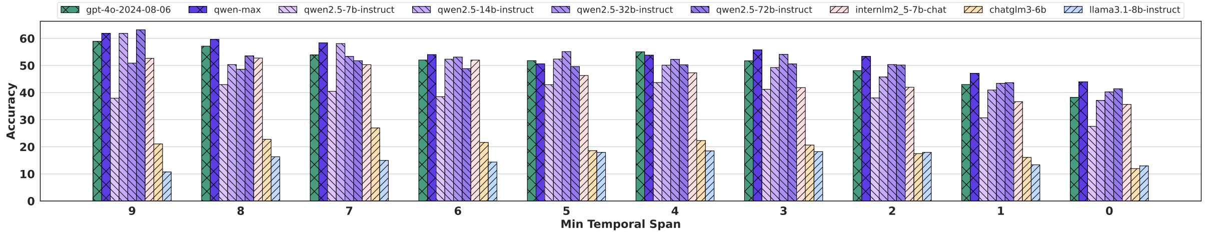


Figure 4: Accuracy across entity inter-dynastic intervals under direct prompting setting. The detailed results are shown in Figure 23, Figure 24 and Figure 25.

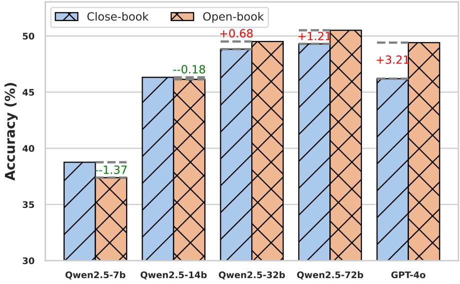


Figure 5: Performance in the close-book and open-book settings. Detailed results can be found in App. J.

3.1 Main Results

Table 2 and Figure 2 present the experimental results of QA and Time Ito Game, respectively. We observe the following empirical findings: **(I)** The more entities considered, the worse the performance, and Time Interval Calculation (TIC) is the most challenging task. The former requires identifying the temporal information of multiple entities, while the latter demands a more precise assessment of specific timestamps. **(II)** CoT generally improves performance but may show diminishing returns for mid-sized models due to reasoning depth trade-offs, and can become ineffective with excessively long contexts. This aligns with the conclusions of work (Chu et al., 2024) and may be attributed to the knowledge sensitivity inherent in temporal reasoning. **(III)** InternLM2.5 demonstrates strong performance among small open-source models, which may be attributed to the quality and composition of its training data. **(IV)** The reasoning model demonstrates remarkably strong performance. **(V)** Temporal alignment is highly challenging, and even powerful model GPT-4o fail to exceed 40 on the Pass@8 metric. **(VI)** Small LLMs cannot align entities across different dimensions, and the Pass@K performance for LLMs smaller than 32B does not exceed 10.

3.2 Analysis

The shorter the time interval between the entities, the greater the difficulty. As illustrated in Figure 4, we evaluate performance across various models based on entity inter-dynastic intervals. For example, an interval of 1 indicates adjacent dynasties, while an interval of 0 represents the same dynasty. As the interval decreases, performance declines. This is because reasoning in QA tasks requires a clear understanding of the temporal relationships between entities, with closer intervals demanding more precise examination.

In the open-book setting, temporal reasoning performance can be moderately improved. To obtain more precise temporal information about entities, we can leverage search engines to retrieve relevant information from the web, enhancing the specificity of entity details (Wu et al., 2025). In the open-book setting, we use the titles and snippets of the Top-10 webpages retrieved via `Google` search as retrieval-augmented information. As shown in Figure 5, it can be observed that performance improves after incorporating the retrieved content, except for Qwen2.5-7B, possibly due to its weaker longe contextual understanding.

4 Conclusion

We introduce CTM, a benchmark designed to evaluate LLMs on temporal reasoning and alignment across Chinese dynasties. CTM benchmark emphasizes contextualization, cross-entity relationships, and temporal alignment. Empirical evaluations on various LLMs reveal the challenges posed by CTM, demonstrating that existing LLMs struggle with nuanced temporal understanding. These findings through analysis suggest the need for improved pretraining, structured knowledge integration, and refined reasoning mechanisms. CTM provides a culturally rich resource for advancing temporal reasoning research.

Limitations

Prompt Design and Evaluation Settings This study evaluates the performance of LLMs on CTM using various prompts, including the most common settings of direct prompting and chain-of-thought (CoT). However, it is acknowledged that the effectiveness of these prompts may vary across different tasks and models. Future work could explore the possibility of dynamically adapting prompt designs to better suit specific temporal reasoning tasks, as well as expanding to more diverse few-shot and zero-shot settings. As LLMs continue to evolve, it will be crucial to periodically update prompt strategies to ensure a robust and comprehensive evaluation.

Dataset Scale and Coverage While CTM currently includes a diverse range of Chinese temporal reasoning tasks, there is significant potential for expanding both its size and coverage. With 8,750 examples already developed, the dataset can be further enriched with larger and more complex temporal scenarios, as well as longer historical events and a broader range of question types. Additionally, the timeline Ito game data could be expanded to incorporate more intricate details and interesting themes, providing greater challenges for models and revealing their strengths and limitations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Ting Bai, Jiazheng Kang, and Jiayang Fan. 2024. Baijia: A large scale role-playing agent corpus of chinese historical charcaters. arXiv preprint arXiv:2412.20024.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. arXiv preprint arXiv:2403.17297.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. Transactions of the Association for Computational Linguistics, 10:257–273.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. arXiv preprint arXiv:2406.09170.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793.

Duygu Sezen Islakoglu and Jan-Christoph Kalo. 2025. Chronosense: Exploring temporal understanding in large language models with time intervals of events. arXiv preprint arXiv:2501.03040.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. CMMLU: Measuring massive multitask language understanding in Chinese. In Findings of the Association for Computational Linguistics: ACL 2024, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024b. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2024. Corecode: A common sense annotated dialogue dataset with benchmark tasks for chinese large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 18952–18960.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? In Proceedings of the 62nd Annual Meeting of the

Association for Computational Linguistics (Volume 1: Long Papers), pages 13014–13033, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking Chinese commonsense reasoning of LLMs: From Chinese-specifics to reasoning-memorization correlations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11205–11228, Bangkok, Thailand. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. Webwalker: Benchmarking llms in web traversal. arXiv preprint arXiv:2501.07572.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10452–10470.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, and Usman Naseem. 2024. Cultural palette: Pluralising culture alignment via multi-agent palette. arXiv preprint arXiv:2412.11167.

Yunxiang Zhang and Xiaojun Wan. 2023. Situatedgen: Incorporating geographical and temporal contexts into generative commonsense reasoning. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

A Related Works

Chinese Cultural Understanding in LLMs Recent advancements in LLMs have shown promise in cultural understanding tasks, with some studies specifically evaluating their performance in Chinese culture, including assessments of commonsense knowledge (Shi et al., 2024; Sun et al., 2024; Li et al., 2024a), foodie culture (Li et al., 2024b), and historical knowledge (Bai et al., 2024). As one of the world's longest-standing cultures, Chinese culture spans a vast historical timeline, with each dynasty rich in historical figures, anecdotes, and cultural narratives. Its strong cultural attributes also allow for effective contextualization. This makes dynastic timelines particularly well-suited for temporal reasoning and alignment in our work.

Temporal Reasoning in LLMs Temporal reasoning is a critical capability for LLMs, with existing benchmarks focusing on factual temporal grounding (Chen et al., 2021; Dhingra et al., 2022), complex temporal logic (Tan et al., 2023; Su et al., 2024), and multi-granular temporal awareness (Chu et al., 2024; Islakoglu and Kalo, 2025). As shown in Table 1, these benchmarks are primarily English-based and rely on rule-based dataset construction, which limits contextualization and diversity. Recently, while existing benchmarks like TGQA (Xiong et al., 2024) provide valuable insights into event ordering, their focus on English datasets leaves open challenges in other linguistic and cultural contexts. Additionally, approaches (Sun et al., 2024; Fatemi et al., 2024) suggest potential benefits from incorporating more open-ended scenarios that better match real-world applications. As shown in Table 1, current benchmarks predominantly employ rule-based construction for English datasets. **CTM** extends this work by (1) anchoring evaluation in Chinese cultural and historical contexts, (2) combining structured and open-ended response formats through LLM-enhanced question generation, and (3) introducing novel task designs, the Timeline Ito Game, that examine multi-agent temporal collaboration.

B Discussions and Clarifications

Benchmark Construction and Quality Control Our dataset construction employed rigorous validation:

- **Initial Manual Verification**: Domain experts manually checked 100 samples per task type to ensure alignment between LLM-generated data and human expectations.
- **Automated Rule-based Validation**: Subsequent data generation incorporated systematic checks (*e.g.*, entity accuracy, temporal consistency) to maintain quality standards.
- **Prompt Design**: Both human annotators and LLMs followed strict guidelines (*e.g.*, requiring all reasoning to be grounded in provided temporal facts).

Annotation

- **Step1: Seed Prompt Creation**: For each entity type, we manually design seed prompts (Taori et al., 2023) to guide the self-instruct-based data generation process. These prompts serve as templates to ensure diversity and relevance in the generated data.
- **Step2: Entity-Aware Data Generation**: During LLM-based generation, the LLMs dynamically incorporate entity descriptions sampled from the pre-constructed entity information repository. This ensures that the generated content is contextually grounded in the repository's structured knowledge, enhancing control over entity-related information.
- **Step3: Validation and Quality Control**: After generation, each data point undergoes a validation step, where the temporal entities **mentioned** in the output are cross-referenced with the repository. This ensures the accuracy and consistency of the entities, aligning the generated data with the repository's constraints.

Data Sources and Processing Our benchmark integrates culturally rich sources, including:

- **Textual Corpora**: Gushiwen (`https://www.gushiwen.cn/`), Chinese Wikipedia (`https://zh.wikipedia.org/`), and Ihchina (`https://www.ihchina.cn/`)
- **Structured Databases**: CBDB (`https://projects.iq.harvard.edu/chinesecbdb`) and CHGIS (`https://gis.harvard.edu/china-historical-gis`) for entity relationships
- **Temporal Standardization**: Historical periods were aligned using Allhistory (`https://www.allhistory.com/`) and the Columbia China Timeline (`https://afe.easia.columbia.edu/timelines/china_timeline.htm`).

C English Translations

C.1 Ten Major Dynasties and Corresponding Period

“先秦(Pre-Qin)” (-2100 to -206), “汉(Han)” (-206 to 220), “六朝” (Six Dynasties) (220 to 589), “隋（Sui）” (581 to 618), “唐(Tang)” (618 to 906), “五代(Five Dynasties)” (907 to 960), “宋(Song)” (960 to 1279), “元(Yuan)” (1279 to 1368), “明(Ming)” (1368 to 1644), “清(Qing)” (1644 to 1912).

C.2 Translated QA Pair

Q. 以下是一段镜头描述，其中有一处或多处不遵循真实历史背景的穿帮内容，请指出:

The following is a scene description containing one or more anachronisms that do not align with historical accuracy. Please identify them:

李白在创作《将进酒》时，白居易在一旁吟诗，同时桌上摆着一盘辣椒，旁边还有一位乐师在演奏古琴。

While **Li Bai** is composing *“Bring in the Wine”*, **Bai Juyi** is reciting poetry beside him. On the table, there is a plate of **chili peppers**, and a musician is playing the **guqin art** nearby.

A. 穿帮内容:

Anachronisms:

1. 李白去世时（762 CE），白居易还未出生（772 CE），两人不可能同时在场。

When **Li Bai** passed away (762 CE), **Bai Juyi** had not yet been born (772 CE), making it impossible for them to be present together.

2. 古琴艺术在唐朝时已非常成熟，符合历史背景。

Guqin Art was already well-developed during the Tang Dynasty, which aligns with the historical context.

3. 辣椒在明朝才传入中国，不可能出现在唐朝。

Chili peppers were not introduced to China until the Ming Dynasty, so they could not have appeared during the Tang Dynasty.

D Statistics of CTM

The statistics of CTM on tasks are shown in Table 3.

Table 3: The statistics of CTM.

Statistic	Question-Answering								
	EDD	**PJ**	**TOU**	**RR**	**SEC**	**EEU**	**TIC**	**TES**	**LSEC**
# Sample	2500	1117	1653	847	841	345	599	548	250
Cross Temp Count	1	2, 3, 4..10							4..15

Statistic	Timeline Ito Game		
	Easy	**Medium**	**Hard**
# Sample	20	20	20
Cross Temp Count	3	4	5
Agent Num	3	4	5

E Entity Repository

Figure 6, Figure 7, Figure 8, Figure 9 and Figure 10 show the case of historical figure, place, event, ingredient and intangible cultural heritage, respectively.

Historical Figure

```
## JSON Format
{
    "屈原": {
        "dynasty": "先秦",
         "address": "楚国丹阳秭归（今湖北宜
昌）",
        "year_birth": "-340",
        "year_death": "-278",
        "book_and_sentences": [
            {
              "sentence": "身既死兮神以灵，
魂魄毅兮为鬼雄。",
                "book": "《国殇》"
            },
            {
              "sentence": "路曼曼其修远兮，
吾将上下而求索。",
                "book": "《离骚》"
            },
            # ...
        ]
    },
    # ...
}
```

Figure 6: A JSON-format case for historical figure entity.

Place

```
## JSON Format
{
  "巴州": {
    "dynasty": "唐",
    "id": "hvd_111423",
    "begin": "758",
    "end": "762",
    "pre_address": "四川省巴中市",
    "subordinate_units": [
      {
        "begin_year": "758",
        "end_year": "762",
        "child_id": "hvd_44640",
        "name": "七盘县",
       "pre_address": "今四川巴中县西北一
百二十里（今旺苍县东南"
      },
      # ...
    ]
  }
  #...
}
```

Figure 7: A JSON-format case for place entity.

Event

```
## JSON Format
{
  "司马迁写《史记》": {
    "id": "070",
    "dynasty": "汉",
    "main_figures": "司马迁"
  },
  # ...
}
```

Figure 8: A JSON-format case for event entity.

Ingredient

```
## JSON Format
{
  "水稻": {
    "dynasty": "先秦",
    "origin": "中国"
  },
  # ...
}
```

Figure 9: A JSON-format case for ingredient entity.

Intangible Cultural Heritage

```
## JSON Format
{
  "昆曲": {
    "dynasty": "明",
      "place": "北京，上海，江苏省，浙江
省，湖南省",
    "type": "表演艺术"
  },
  # ...
}
```

Figure 10: A JSON-format case for intangible cultural heritage entity.

F Cases

F.1 Cases in Question-Answering

Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19 show the Entity-based Dynasty Determination, Plausibility Judgment, Temporal Order Understanding, Relation Reasoning, Script Error Correction, Entity Evolution Understanding, Time Interval Calculation, Temporal Entity Selection and Long Script Error Correction tasks in JSON-format, respectively.

Entity-based Dynasty Determination (EDD)

```
## JSON Format
{
  "type": "根据食材确定传入朝代",
  "question": "假设你穿越到某个朝代，听商
贩说辣椒是最近几年才有的食材，你能推测你
穿越到了哪个朝代吗？",
  "temporal_entities": ["辣椒"],
   "construct_explanation": "这个问题通过
穿越情境让提问者推测辣椒的引入时间。辣
椒最早于明朝从美洲传入中国，因此问题涉
及1个时空实体-朝代关系，辣椒传入中国的朝
代是‘明’。",
  "answer": "明"
}
```

Figure 11: A JSON-format case in EDD type of QA.

Plausibility Judgment (PJ)

```
## JSON Format
{
  "type": "合理性判断",
  "question": "苏轼是眉州人，眉州是今四川
眉山市，而四川以麻辣闻名。那么，苏轼是否
吃辣？",
 "temporal_entities": ["苏轼", "眉州", "辣
椒"],
   "construct_explanation": "苏轼生活在宋
代（约公元1037年-1101年），辣椒最早于明朝
传入中国。因此问题涉及3个时空实体-朝代关
系，‘苏轼-宋’、‘眉州-宋’和‘辣椒-明’，通过
时间线推理，苏轼生活的时代不可能接触到辣
椒。",
  "answer": "否"
}
```

Figure 12: A JSON-format case in PJ type of QA.

Temporal Order Understanding (TOU)

```
## JSON Format
{
  "type": "时间顺序理解",
   "question": "请将以下实体按时间顺序排
列：屈原，李白，白居易，辣椒，古琴艺术，
石榴？",
 "temporal_entities":["屈原","李白","白
居易","辣椒","古琴艺术","石榴"],
   "construct_explanation": "屈原生活在先
秦（约公元前340年-公元前278年），李白生
活在唐代（约公元701年-762年），白居易生活
在唐代（约公元772年-846年），辣椒在明朝传
入中国（约公元16世纪），古琴艺术起源于先
秦（约公元前11世纪），石榴在汉代传入中国
（约公元前2世纪）。因此问题涉及6个时空实
体-朝代关系，‘屈原-先秦’、‘李白-唐’、‘白
居易-唐’、‘辣椒-明’、‘古琴艺术-先秦’和‘石
榴-汉’，按时间顺序排列为古琴艺术、屈原、
石榴、李白、白居易、辣椒。",
  "answer": "古琴艺术、屈原、石榴、李白、
白居易、辣椒"
}
```

Figure 13: A JSON-format case in TOU type of QA.

Relation Reasoning (RR)

```
## JSON Format
{
  "type": "关系判断",
  "question": "虞州和河东郡在历史上有什么
关系？",
 "temporal_entities": ["虞州", "河东郡"],
  "construct_explanation": "虞州和河东郡
在历史上的行政区划大致对应现在的山西运
城。河东郡是古代的一个重要行政区划，虞州
则是后来的行政区划。因此问题涉及2个时空实
体-朝代关系，‘虞州-隋’和‘河东郡-先秦’，两
者在地理位置上与现代的山西运城有较高的重
合度。",
  "answer": "都为现山西运城"
}
```

Figure 14: A JSON-format case in RR type of QA.

Entity Evolution Understanding (EEU)

```
## JSON Format
{ "type": "实体名称变迁",
  "question": "在唐朝的浈州内，今日的四会
县在当时名为何？",
 "temporal_entities": ["浈州", "四会县"],
  "construct_explanation": "唐朝时期存在
的浈州（公元634年至638年）包含今日的四
会县。根据历史资料，此时四会县作为浈州
的下属地方，在唐朝已经是现名四会县。因
此问题涉及2个时空实体-地方名称关系，‘浈
州-唐’ 及‘四会县-唐’，从历史上来看，当时
的四会县没有变更过名称。",
  "answer": "四会县"}
```

Figure 15: A JSON-format case in EEU type of QA.

Script Error Correction (SEC)

```
## JSON Format
{ "type": "穿帮镜头指正",
   "question": "以下是一段镜头描述，其中
有一处不遵循真实历史背景的穿帮内容，请指
出：在唐代的朗州，一位表演者正在用昆曲演
绎盛唐的繁华，而高粱酒则是当场招待贵宾的
饮品。",
 "temporal_entities":["昆曲","朗州","高
粱"],
   "construct_explanation": "1. 朗州存在
于唐代从621年到741年之间。2. 昆曲起源于明
代，与唐代不重叠。3. 高粱传入中国的时间更
为久远，早在先秦时期便已存在，因此可用于
唐代饮品制作。因此，使用昆曲演绎显然是历
史穿帮。",
  "answer": "昆曲在唐代的朗州出现是穿帮内
容，因昆曲起源于明代。"}
```

Figure 16: A JSON-format case in SEC type of QA.

Time Interval Calculation (TIC)

```
## JSON Format
{ "type": "时间差计算",
  "question": "从屈原投江到李白出生，再到
苏轼出生，中间经历了多少年？",
 "temporal_entities":["屈原","李白","苏
轼"],
   "construct_explanation": "屈原投江发生
在约公元前278年，李白出生于公元701年，苏
轼出生于公元1037年。因此问题涉及3个时空实
体-朝代关系，‘屈原-先秦’、‘李白-唐’和‘苏
轼-宋’，通过计算时间差，屈原到李白相差
约979年，李白到苏轼相差约336年，总共相差
约1315年。",
  "answer": "约1315年"}
```

Figure 17: A JSON-format case in TIC type of QA.

Temporal Entity Selection (TES)

```
## JSON Format
{
  "type": "选出对应时空实体",
    "question": "以下四个实体中，哪个属于唐朝？\n(A) 李白 \n(B) 苏轼 \n(C) 屈原 \n(D) 曹操",
 "temporal_entities": ["李白", "苏轼", "屈原", "曹操"],
    "construct_explanation": "李白生活在唐代（约公元701年-762年），苏轼生活在宋代（约公元1037年-1101年），屈原生活在先秦（约公元前340年-公元前278年），曹操生活在汉代（约公元155年-220年）。因此问题涉及4个时空实体-朝代关系，‘李白-唐’、‘苏轼-宋’、‘屈原-先秦’和‘曹操-汉’，通过朝代背景推理，李白属于唐朝。",
  "answer": "李白"
}
```

Figure 18: A JSON-format case in TES type of QA.

F.2 Running Example of Timeline Ito Game

A Timeline Ito Game running example given the "*fruit size*" theme is below.

A Running Example of Timeline Ito Game

```
----------------------------------------
真实顺序：{"屈原": 1, "李白": 2, "苏轼": 3}
初始化：
Agent P1: "李白"
Agent P2: "屈原"
Agent P3: "苏轼"
----------------------------------------
Agent P3 prediction in Round 1:
{
  "分析": "根据对xxx的了解，xxx是宋朝时期的一位著名诗人。因此，时间实体对应的朝代是宋。根据历史朝代时间线，宋朝在十个朝代中居中偏后的位置。结合水果大小的主题，应该选择稍大一些但不是最大的水果。在已提供的选择中，西瓜是最大的，蓝莓是最小的，对应最前的朝代。选择朝代在居中偏后的主题实体是桃子。",
   "理由": "我的时间实体对应的朝代居中偏后",
  "主题实体": "桃子"
}
----------------------------------------
###=== Round 1 ===
当前主题: 水果大小
Agent P1: 我的时间实体对应的朝代居中偏前. 因此我选择橙子
Agent P2: 我的时间实体对应的朝代最前. 因此我选择蓝莓
Agent P3: 我的时间实体对应的朝代居中偏后. 因此我选择桃子
本轮排序结果: {'P1': 2, 'P2': 1, 'P3': 2}

本轮排序错误，错误的玩家：P3

###=== Round 2 ===
当前主题: 书本厚度
Agent P1: 我的时间实体对应的朝代居中. 因此我选择字典
Agent P2: 我的时间实体对应的朝代最前. 因此我选择书签
Agent P3: 我的时间实体对应的朝代居中偏后. 因此我选择百科全书

本轮排序结果: {'P1': 2, 'P2': 1, 'P3': 3}
游戏结束！所有玩家的排序正确。
```

G LLM Backbone List

We validate the total number of twelve models, including both closed-sourced and open-sourced ones (Achiam et al., 2023; Dubey et al., 2024; Yang et al., 2024; Cai et al., 2024; GLM et al., 2024). The complete list of evaluated LLMs is shown in Table 4.

Long Script Error Correction (LSEC)

```
## JSON Format
{  "type": "剧本穿帮问题指正_长上下文",
   "question": "以下是一段剧本描述，请指出其中不符合历史背景的穿帮内容：\n\n背景设定在东汉末年，名将曹操正在洛阳的一座文人聚会酒楼中与手下谋士们讨论政治大计。酒楼内弥漫着美酒的香气，桌上的菜肴色香俱全，包括蜜饯杏仁、红烧鹿肉以及一盘山药炖鸡。曹操端起酒杯，起身与众人敬酒，话题转向了当前国家的局势。突然，酒楼的门外传来一阵欢声笑语，几位文人相伴走入酒楼，他们是当时知名的诗人王之涣和杜牧。王之涣手捧一卷《登鹳雀楼》的诗稿，热衷地与杜牧讨论诗句的对仗工整。两位诗人落座后，曹操也邀请他们一起品酒作诗。\n\n就在此时，一位穿着朴素的商人走了进来，自称是商贸使者张衡。他向酒楼的主人要求提供一些来自西南的特产，并推荐了其中的一种热带水果—草莓。酒楼老板随即呈上了这一新鲜水果，几人尝过后纷纷表示口感非常独特，直言这是一种他们从未品尝过的美味。与此同时，另一位年轻的学者走进酒楼，他正是历史学者班固，他从洛阳的学府赶来，听到酒楼内的喧闹声，不禁走入与众人交谈。\n\n随着酒楼里的谈话逐渐深入，众人讨论的议题转向了国家的未来。王之涣对曹操表示，若有朝一日可以恢复汉朝的光辉，他愿意为之创作更多诗篇。而杜牧则提到，诗词创作对于平定民心具有重要作用。他们的对话引发了在座每个人的深思。\n\n此时，酒楼外突然传来一阵马蹄声，一位身穿铠甲的将军匆匆走进酒楼，向曹操报告前线战况。这位将军正是夏侯，他刚从许都赶来，带来了最新的军情。曹操听后眉头紧锁，立即召集众人商议对策。夏侯提议联合孙权共同对抗袁绍，但曹操却认为应当先稳固内部，再图外敌。\n\n就在众人争论不休时，酒楼的门再次被推开，一位身穿道袍的老者走了进来。他自称是华佗，手中捧着一瓶刚刚炼制的‘麻沸散’。华佗向曹操献上药物，声称服用后可缓解头痛。曹操接过药物，若有所思地看了看众人，随后将药物放入怀中。\n\n酒楼的氛围逐渐热烈起来，众人一边饮酒一边讨论着国家的未来。突然，一位年轻的乐师走到酒楼中央，开始演奏一曲《短歌行》。琴声悠扬，众人纷纷停下手中的酒杯，静静地聆听。乐师演奏完毕后，曹操起身鼓掌，称赞其技艺高超，并邀请他加入自己的幕府。\n\n随着夜幕降临，酒楼内的灯火逐渐点亮，众人继续畅谈。曹操举起酒杯，高声说道：‘今日与诸位相聚，实乃幸事！愿我们共同努力，恢复汉室荣光！’众人纷纷举杯响应，酒楼内充满了欢声笑语。",
  "temporal_entities": ["曹操", "王之涣", "杜牧", "张衡", "草莓", "班固", "夏侯", "华佗"],
    "construct_explanation":  "剧本中出现的多个历史人物和食材存在不符合历史背景的穿帮内容：\n\n1. 王之涣和杜牧不可能同时出现在东汉末年。王之涣生活在唐朝，杜牧则生活在唐朝晚期，二人不可能与曹操同时存在。\n2. 张衡生活在东汉时期，他是著名的天文学家和文学家，但剧本中将他错误地设定为商贸使者，这一身份与历史背景不符。\n3. 草莓直到清朝才传入中国，剧本中的草莓出现在东汉末年不符合历史事实。\n4. 班固作为东汉时期的历史学者，他比曹操年长不少，因此不可能与曹操同时出现在剧本中。班固应当是年老之际，无法与年轻的曹操同时活跃。",
  "answer": "1. 王之涣和杜牧不可能同时出现在东汉末年。\n2. 张衡不可能作为商贸使者与曹操同时出现。\n3. 草莓在东汉末年未传入中国，属于不合时宜的食材。\n4. 班固的年纪和身份不应与曹操的同时期相关。"
}
```

Figure 19: A JSON-format case in LSEC type of QA.

Models	Full Name	Open Source?	Model Size
GPT-4o	gpt-4o-2024-08-06	✗	-
Qwen-max	qwen-max	✗	-
o1-preview	o1-preview	✗	-
LLaMA3.1$_{8b}$	Meta-Llama-3.1-8B-Instruct	✓	8B
ChatGLM3$_{6b}$	chatglm3-6b	✓	6B
InternLM2.5$_{7b}$	internlm2_5-7b-chat	✓	7B
Qwen2.5$_{7b}$	qwen2.5-7b-instruct	✓	7B
Qwen2.5$_{14b}$	qwen2.5-14b-instruct	✓	14B
Qwen2.5$_{32b}$	qwen2.5-32b-instruct	✓	32B
Qwen2.5$_{72b}$	qwen2.5-14b-instruct	✓	72B
DeepSeek-R1	deepseek-r1	✓	671B

Table 4: LLMs evaluated in our experiments

H Prompt

Prompt for Direct Prediction

```
请回答以下问题：
{question}
```

Figure 20: Prompt for Direct Prediction

Prompt for CoT Prediction

```
请按照以下步骤一步一步思考并回答问题：
1. {question}
2. 思考并详细分析问题，然后得到结论及简要
理由，如果理由比较复杂请分条简要列出。
3. 请以 JSON 格式返回结果，格式如下：
{
  "思考": "请在这里填写详细的思考过程",
  "回答": "请在这里填写结论及理由"
}
```

Figure 21: Prompt for CoT Prediction.

I Timeline Ito Game Performance

The detailed performance across difficulty levels is shown in Table 5. The difficulty level is determined based on the number of entities, where 3 corresponds to easy, 4 to medium, and 5 to hard. This number also represents the number of agents.

Prompt of Evaluator

```
你是一个专业的问答系统评估员。请根据以下信息评估答案的质量，并输出详细的思考过程：

**问题类型**：{question_type}
**问题**：{question}
**参考答案**：{reference} {answer}
**待评估答案**：{prediction}

**评估步骤**：
1. 判断问题类型：
- 如果问题是“是否问题”，进入步骤 2。
- 如果问题不是“是否问题”，进入步骤 3。
2. 对于“是否问题”：
- 检查待评估答案是否正确回答了“是/否”。
- 检查待评估答案的原因是否与参考答案一致。
- 如果两者都正确，评估结果为1；否则评估结果为0。
3. 对于非“是否问题”：
- 检查待评估答案是否**完全覆盖**参考答案中的所有关键点。
- 检查待评估答案中的每个点是否与参考答案**完全一致**，包括事实、逻辑和时间线等。
- 检查朝代是否宽松匹配：
  - 比较待评估答案中的朝代范围与参考答案的朝代范围，允许评估答案使用细分的朝代划分。
    - 参考答案中的朝代范围：先秦（-2100～-206），汉（-206～220），六朝（220～589），
隋（581～618），唐（618～906），五代（907～960），宋（960～1279），元（1279～1368），明
（1368～1644），清（1644～1912）。
  - 秦朝以前的朝代范围夏、商、周、春秋、战国等都属于先秦。六朝和五代的朝代范围分别为东晋、宋、
齐、梁、陈、后周和后梁、后唐、后晋、后汉、后周。
  - 允许待评估答案中的细分朝代在参考答案的朝代范围内进行宽松匹配。
- 如果待评估答案**完全覆盖且完全一致**，评估结果为1；否则返回评估结果为0。

**思考过程**：
- 详细分析待评估答案与参考答案的异同。
- 重点关注跨时空推理的历史时间线和逻辑一致性。
- 判断多个时空实体之间及与朝代之间的关系。

**请以 JSON 格式返回结果**：
{
  "思考": "请在这里填写详细的思考过程",
  "评估结果": "请在这里填写0或1"
}
```

Figure 22: A JSON-format case in intangible cultural heritage entity.

Table 5: **Main results on Timeline Ito Game** within CTM benchmark.

Method	Easy		Medium		Hard		Overall	
	Pass@3	Pass@8	Pass@3	Pass@8	Pass@3	Pass@8	Pass@3	Pass@8
GPT-4o	55.00	80.00	20.00	30.00	5.00	10.00	26.67	40.00
Qwen-max	25.00	35.00	10.00	10.00	10.00	15.00	15.00	20.00
LLaMA3.1$_{8b}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGLM3$_{6b}$	5.00	5.00	0.00	0.00	0.00	0.00	1.67	1.67
InternLM2.5$_{7b}$	5.00	15.00	0.00	0.00	0.00	0.00	1.67	5.00
Qwen2.5$_{7b}$	0.00	15.00	5.00	5.00	0.00	0.00	1.67	6.67
Qwen2.5$_{14b}$	15.00	20.00	0.00	0.00	0.00	0.00	5.00	6.67
Qwen2.5$_{32b}$	40.00	50.00	5.00	15.00	0.00	0.00	15.00	21.67
Qwen2.5$_{72b}$	40.00	55.00	10.00	10.00	0.00	5.00	16.67	23.33

J Open-Book Performance

Detailed results across tasks and entity numbers ars shown in Table 6.

Prompt For Step1 of Timeline Ito Game

你是一个玩家，参与了一个叫做“命悬一线（ito）”的游戏，现在你收到一个时间实体：'{self.entity}'。
1．首先需要推理得出该时间实体所属的朝代，朝代时间线为：
先秦（-2100～-206），汉（-206～220），六朝（220～589），隋（581～618），唐（618～906），五代（907～960），宋（960～1279），元（1279～1368），明（1368～1644），清（1644～1912）

2．主题与朝代的对应规则：如果主题是”水果大小“，则水果越小，对应朝代越早；如果是“书本厚度”，则书本越薄，对应朝代越早；如果是“船只大小”，则船只越小，对应朝代越早；如果是“电子产品”大小，则电子产品越小，对应朝代越早；如果是“行星大小”，则行星越小，对应朝代越早；如果是动物大小，则动物越小，对应朝代越早；如果是“建筑高度”，则建筑越矮，对应朝代越早；如果是“水的温度”，则水越冷，对应朝代越早。
本轮的主题是{theme}，主题实体的个数与朝代个数对应。你需要在{theme_entities}中选择1个对应“主题实体在主题的顺序”，来代表你的“时间实体的朝代在朝代时间线中的位置”，返回你认为最能代表你的“时间实体的朝代”的1个“主题实体”

3．以下是之前的交流记录，排序相关判断可作为参考，其余的不作参考。(如果这里是空的，就不参考)
{public_memory_str}

4．分析你的时间实体在朝代时间线中的位置，输出分析过程，分析内容中你的“时间实体”和你推理出的“朝代”用“xxx”代替，只能出现“主题实体”的名称，可以用“朝代偏前或偏后”等表达。请注意，你的分析过程需要符合主题与朝代的对应规则。
结合分析给出一句话表达你的推理理由，“我的时间实体对应的朝代最前/偏前/居中/偏后/最后等”（理由中不要出现名称）。

5．请根据上述信息，返回如下JSON格式：
```
{
  "分析":"请在这里填写分析",
  "理由":"我的时间实体对应的朝代最前/偏前/居中/偏后/最后等",
  "主题实体": "请在这里填写主题实体"
}
```

Prompt For Step2 of Timeline Ito Game

基于之前的交流记录：{public_memory_str}
我的ID是：Agent {self.agent_id}，请分析我选择的实体在所有Agent中的排序。
如果在上一轮中，“本轮排序错误，错误的玩家：”后不包含我的ID，则我的排序值维持上一轮的排序值。

1．如果主题是”水果大小“，则水果越小，排序越小；如果是“书本厚度”，则书本越薄，排序越小；如果是“船只大小”，则船只越小，排序越小；如果是“电子产品”大小，则电子产品越小，排序越小；如果是“行星大小”，则行星越小，排序越小；如果是动物大小，则动物越小，排序越小；如果是“建筑高度”，则建筑越矮，排序越小；如果是“水的温度”，则水越冷，排序越小。
本轮的主题是{theme}，返回我的“主题实体”在该主题下在所有Agent中的排序”，排序值为整数，最小为1。

2．之前轮次我的个人记忆（这部分不公开，大家互相猜测对齐）：
{self.memory}

3．请根据上述信息，返回如下JSON格式：
```
{
  "我的排序": "请在这里填写排序"
}
```

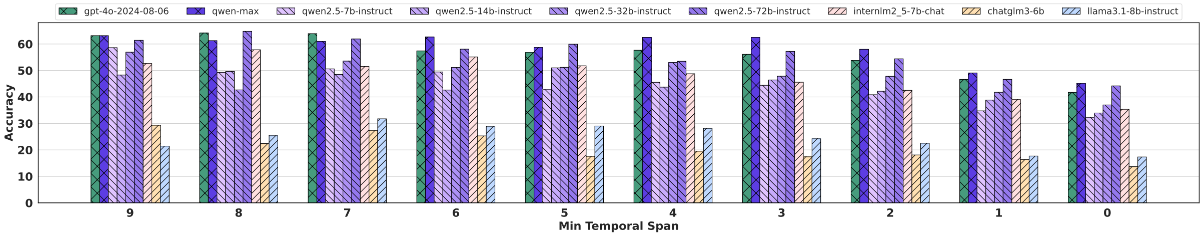


Figure 23: Accuracy across entity inter-dynastic intervals under CoT prompting setting.

Table 6: Detailed results under the open-book setting.

Method	**Cross Temp Count**					**Question Type**							**Avg.**
	$= 1$ (EDD)	$= 2$	$= 3$	≥ 4	$\geq 4_L$ (LSEC)	PJ	TOU	RR	SEC	EEU	TIC	TES	
GPT-4o	56.52	51.12	44.76	26.10	53.60	58.64	38.42	57.26	36.15	40.58	15.36	59.31	46.20
+ Openbook	$57.76_{+1.24}$	$53.40_{+2.28}$	$45.52_{+0.76}$	$26.90_{+0.80}$	$56.80_{+3.20}$	$59.00_{+0.36}$	$38.72_{+0.30}$	$54.66_{-2.60}$	$45.30_{+9.15}$	$42.61_{+2.03}$	$17.20_{+1.84}$	$58.39_{-0.92}$	$49.41_{+3.21}$
Qwen2.5$_{7b}$	51.80	39.88	35.96	12.40	30.00	46.28	26.38	46.28	24.14	36.23	7.35	52.01	38.76
+ Openbook	$48.64_{-3.16}$	$39.92_{+0.04}$	$31.88_{-4.08}$	$17.90_{+5.50}$	$31.60_{+1.60}$	$47.63_{+1.35}$	$27.89_{+1.51}$	$42.15_{-4.13}$	$26.04_{+1.90}$	$31.88_{-4.35}$	$5.84_{-1.51}$	$44.53_{-7.48}$	$37.39_{-1.37}$
Qwen2.5$_{14b}$	54.36	51.16	42.56	23.80	42.00	57.44	36.86	51.83	36.90	39.07	18.26	58.58	46.32
+ Openbook	$54.32_{-0.04}$	$51.28_{+0.12}$	$41.76_{-0.80}$	$23.60_{-0.20}$	$44.40_{+2.40}$	$58.82_{+1.38}$	$36.48_{-0.38}$	$51.83_{+0.00}$	$39.95_{+3.05}$	$39.71_{+0.64}$	$13.86_{-4.40}$	$52.92_{-5.66}$	$46.14_{-0.18}$
Qwen2.5$_{32b}$	56.28	52.78	46.24	26.90	46.40	60.66	38.54	56.79	39.12	43.77	20.10	60.04	48.83
+ Openbook	$57.92_{+1.64}$	$53.32_{+0.54}$	$46.16_{-0.08}$	$26.80_{-0.10}$	$50.80_{+4.40}$	$61.15_{+0.49}$	$39.93_{+1.39}$	$55.61_{-1.18}$	$40.67_{+1.55}$	$45.22_{+1.45}$	$16.86_{-3.24}$	$58.21_{-1.83}$	$49.51_{+0.68}$
Qwen2.5$_{72b}$	58.20	48.76	46.84	31.30	60.80	61.38	40.77	54.31	36.62	42.03	11.52	62.23	49.30
+ Openbook	$57.96_{-0.24}$	$52.00_{+3.24}$	$48.04_{+1.20}$	$30.60_{-0.70}$	$63.60_{+2.80}$	$62.67_{+1.29}$	$42.86_{+2.09}$	$54.07_{-0.24}$	$41.26_{+4.64}$	$44.64_{+2.61}$	$18.03_{+6.51}$	$56.75_{-5.48}$	$50.51_{+1.21}$

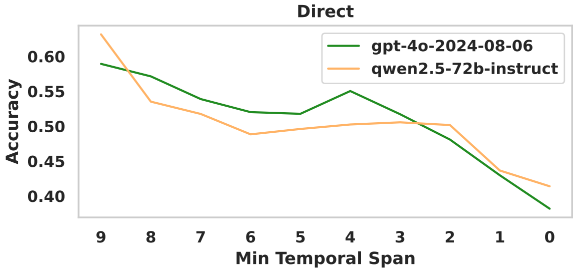


Figure 24: Accuracy across entity inter-dynastic intervals under direct prompting setting on GPT-4o and Qwen2.5-7B.

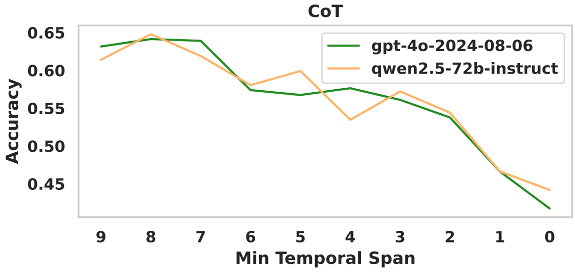


Figure 25: Accuracy across entity inter-dynastic intervals under CoT prompting setting on GPT-4o and Qwen2.5-7B.