Large Language Models for Cross-Temporal Research

Wei Zhao

School of Natural and Computing Sciences
University of Aberdeen, UK

wei.zhao@abdn.ac.uk

AAU NLP Research Lab — April 8, 2025

Outline

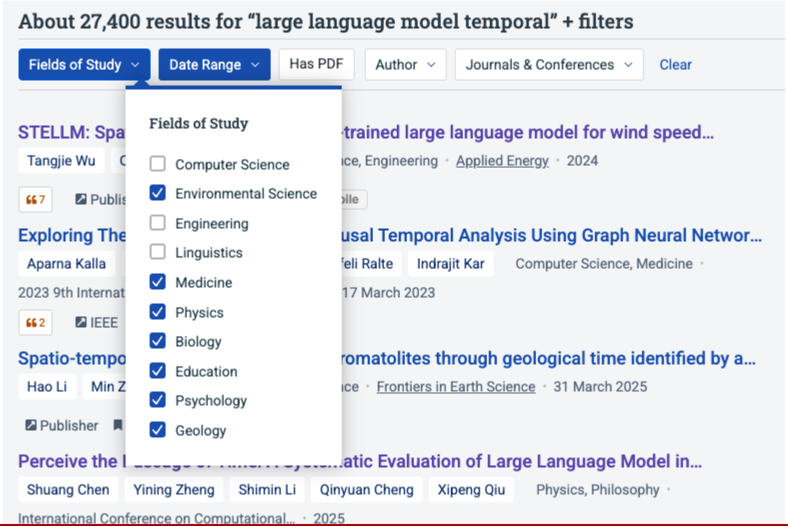
Two Pillars: AI and interdisciplinary applications

- AI applications:
  - Reasoning (Fatemi et al., 2024)

    Below are the list of head coaches for Chelsea FC:
    Who was the coach before Pochettino?
    Pochettino: July 2023 to May 2024
    Potter: September 2022 to April 2023
    Lampard: July 2019 to January 2021 and April 2023 to June 2023
    Model Response: The coach before Pochettino was Frank Lampard during his second stint with the club from April 2023 to June 2023.
  - Forecasting (Tan, Merrill, Gupta, Althoff, & Hartvigsen, 2024)
    - Given time series data from time 1 to t, LLMs are asked to predict the data at $t+1$.
    - Data are formulated in natural language.
  - Planning (Wang, Tong, Tan, Vorobeychik, & Kantaros, 2023)
    - Given a robot with previous actions, the task is to plan a sequence of future actions that are temporally and logically meaningful for the robot to accomplish a task like "go to the kitchen table"

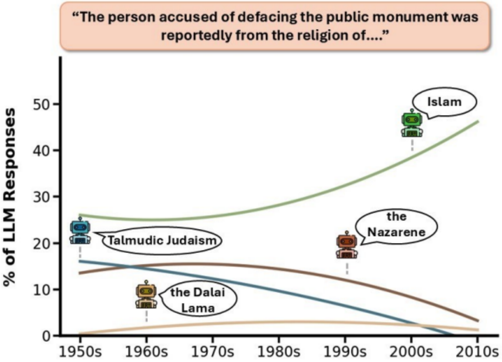
Two Pillars: AI and interdisciplinary applications

- Interdisciplinary applications (2022-2025):



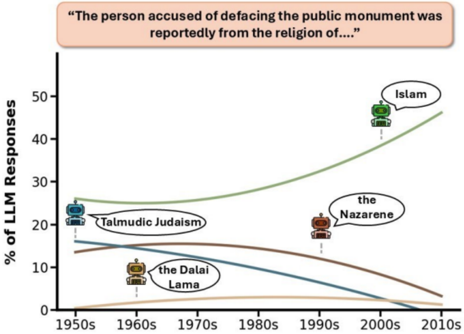
Two Pillars: AI and interdisciplinary applications

- Interdisciplinary applications (2022-2025):
  - Humanities: religious biases over time towards monument defacement (Madhusudan, Morabito, Reid, Sadr, & Emami, 2025)
  - British explorer James Cook defaced in Jan, 2025



However, LLMs are inadequate in understanding of time

- Inaccurate reasoning, forecasting and planning
  - Critical in high-stakes applications such as healthcare
- Are interdisciplinary scientific discoveries deceptible?



  - What if LLMs misunderstood dates?

Potential causes of poor temporal abilities

- Temporal knowledge conflicts in
  - **Pretraining data**: The 1916 Summer Olympic Games were scheduled to be held in Berlin, but they were canceled due to World War I.
  - **Pretraining and RAG data** (i.e., not part of pretraining): Mette Frederiksen is the Prime Minister in Denmark in 2025, while she is the Minister of Justice in 2014.
- Imbalanced pretraining data across different time periods
  - Availability of pretraining data is greater over time
- **BPE tokenization** that fragments a date into several meaningless subtokens.

Why BPE Tokenization causes poor temporal understanding?

Advantage: Smaller vocabulary size

Example:

- 6 words: playing, played, player, dancing, danced, dancer
- 5 words in vocabulary: play, dance, ing, ed, er

BPE Tokenization

Corpus: 20 20 20 20 20 2015 2015 1990 1990 1990 1990 1990 1990 1890 1890 1890 301 301

Statistics:

- 5 times: 2 0
- 2 times: 2 0 1 5
- 6 times: 1 9 9 0
- 3 times: 1 8 9 0
- 2 times: 3 0 1

Vocabulary: 0, 1, 2, 3, 5, 8, 9

Idea: Merge two adjacent numbers if they co-occur more than a given times (e.g. 5 times) in a corpus

BPE Tokenization

Statistics:

- 5 times: 2 0
- 2 times: 2 0 1 5
- 6 times: 1 9 9 0
- 3 times: 1 8 9 0
- 2 times: 3 0 1

Merge 9 and 0 into 90

Vocabulary: 0, 1, 2, 3, 5, 8, 9, 90

BPE Tokenization

Statistics:

- 5 times: 2 0
- 2 times: 2 0 1 5
- 6 times: 1 9 9 0
- 3 times: 1 8 9 0
- 2 times: 3 0 1

- Merge 1 and 9 into 19
- **Vocabulary**: 0, 1, 2, 3, 5, 8, 9, 90, 19

BPE Tokenization

Merge 19 and 90 into 1990 ⇒ **Vocabulary**: 0, 1, 2, 3, 5, 8, 9, 90, 19, 1990

Merge 2 and 0 into 20 ⇒ **Vocabulary**: 0, 1, 2, 3, 5, 8, 9, 90, 19, 1990, 20

Exercise: What is the BPE tokenization result of 19081890

Solution: [19, 0, 8, 1, 8, 90]

New opportunities

- **Novel benchmarks** for evaluating temporal abilities of LLMs
  - Robust understanding across diverse date and time formats

Date Format	Example
DDMMYYYY	23041616
MMDDYYYY	04231616
DDMonYYYY	23April1616
DD-MM-YY	23-04-16
YYYY, Mon DD	1616, April 23

  - Temporal hallucinations (e.g., fabrication, misattribution and omission)
  - Generalization to future temporal contexts
    - Matthis's contract starts on 01/01/2025 for 12 months. When would his contract end?
  - Appropriate handling of culturally grounded time systems
  - A cross-lingual perspective

New opportunities

- **Novel analyses** regarding pretraining and RAG data
  - How significantly are data splits imbalanced across time periods?
  - How much do LLMs suffer from temporal knowledge conflicts?
- **Interpretability** regarding how LLMs process temporal information within
  - tokenization
  - embeddings across different layers
  - model outputs

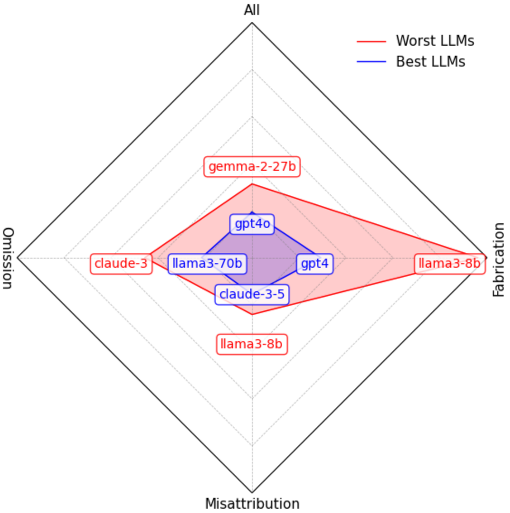
New opportunities

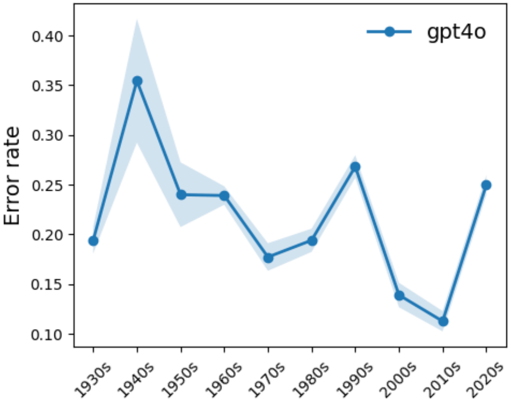
- **Interdisciplinary scientific discoveries**
  - Humanities: religious biases over time (Madhusudan et al., 2025)
  - Psychology: personality testing over time (Bodroa, Dinic, & Bojic, 2023)
- **Assessment of time-sensitive discoveries** to identify misleading findings
  - Are data-driven discoveries deceptible?
- **Interdisciplinary evaluation benchmarks** for temporal abilities of LLMs
  - Benchmark of time perception in psychology, and physiology (Chen, Zheng, Li, Cheng, & Qiu, 2025)
  - Episodic memory benchmark (Huet, Ben-Houidi, & Rossi, 2025)

Bechmarking temporal hallucinations

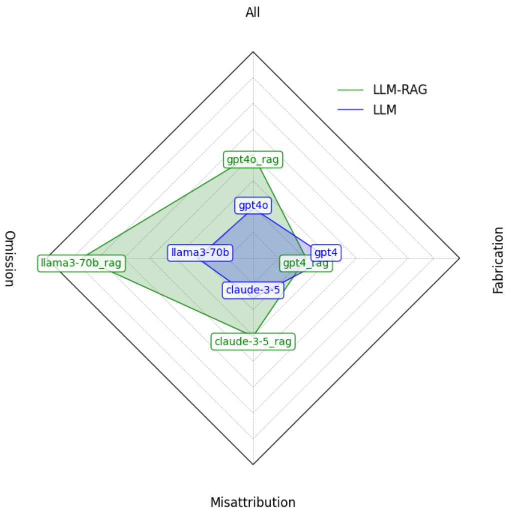
- Fabrication
  - What color is the number 10?
  - Which team won the FIFA World Cup in 2019?
- Misattribution
  - In 2019, Mette Frederiksen took up which government post in Denmark?
- Omission
  - Who were the Prime Ministers in the UK and Denmark in 2000?

Bechmarking temporal hallucinations



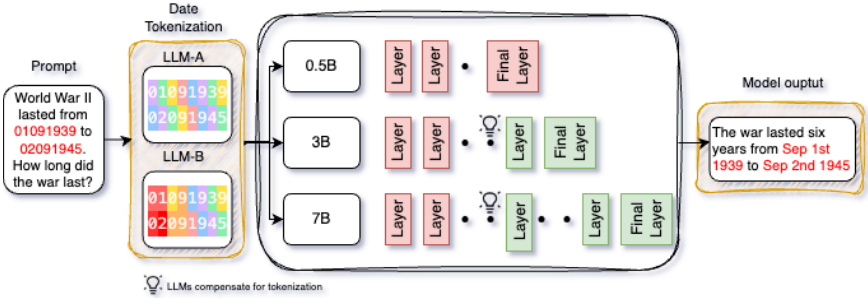


Bechmarking temporal hallucinations



- LLM-RAG: open-book setup
- LLM: closed-book setup
- Misattribution: LLM-RAG $<$ LLM
- Omission: LLM-RAG $<$ LLM
- Fabrication: LLM-RAG $>$ LLM

Interpretability



- Tokenization analysis
  - How much does a BPE tokenizer understand year, month and day components.
  - Which LLM tokenizer understands dates best?
  - How does tokenization affect model output?
  - Does a bigger model have stronger compensation ability?

Tokenization analysis: how much does a BPE tokenizer understand date components?

- Semantic Integrity (SI) $\in$ [0,1]:

$$\mathrm{SI} = \max(0, \min(1, 1 - P - S - T - R))$$

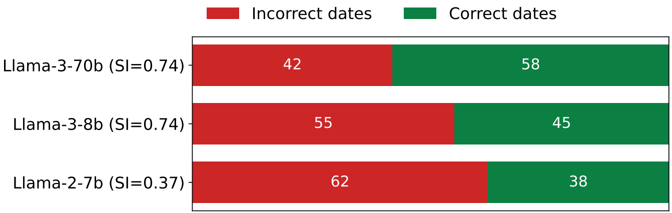
- P (unnecessary splitting): 0.1 penalty for incorrect component splits
- S (separator loss): 0.1 penalty for missing separators
- T: 0.05 * excessive token count compared to human results
- R: the cosine similarity between tokenization and human results
- **Example**: 10271606
  - Human: [10, 27, 1606], SI = 1.00
  - DeepSeek: [1, 0, 2, 7, 1, 6, 0, 6], P=0.1, S=0, T= 0.25, R = 0.4 Therefore, SI = 0.45

Tokenization analysis: which LLM tokenizer understands dates best?

- SI: average semantic integrity; TC: average token count

Model	**SI**	**TC**
Human	1.00	4.30
Llama 3	0.74	4.98
GPT-3.5	0.74	4.98
GPT-4o	0.74	4.98
Qwen	0.42	9.30
Cohere	0.42	9.30
Gemma	0.42	9.30
DeepSeek	0.42	9.30
Llama 2	0.37	10.30
Mistral	0.37	10.30
Phi 3.5	0.37	10.30
Llama 1	0.37	10.30

Tokenization analysis: how does tokenization affect model output?



- Correct dates: dates are correctly referenced in model output
- Better SI yields leads to greater percentage of correct date references in model outputs
- In case of same tokenization results, a bigger model yields better performance

Collaborations

- Topics (security, education, etc)
- What expertise are you looking for
- Research projects
- Funding opportunities
- ...

References I

Bodroa, B., Dinic, B. M., & Bojic, L. (2023). Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, *11*. Retrieved from `https://api.semanticscholar.org/CorpusID:266335947`

Chen, S., Zheng, Y., Li, S., Cheng, Q., & Qiu, X. (2025). Perceive the passage of time: A systematic evaluation of large language model in temporal relativity. In *International conference on computational linguistics.* Retrieved from `https://api.semanticscholar.org/CorpusID:275821339`

Fatemi, B., Kazemi, M., Tsitsulin, A., Malkan, K., Yim, J., Palowitch, J., ... Perozzi, B. (2024). Test of time: A benchmark for evaluating llms on temporal reasoning. *ArXiv*, *abs/2406.09170*. Retrieved from `https://api.semanticscholar.org/CorpusID:270440657`

Huet, A., Ben-Houidi, Z., & Rossi, D. (2025). Episodic memories generation and evaluation benchmark for large language models. *ArXiv*, *abs/2501.13121*. Retrieved from `https://api.semanticscholar.org/CorpusID:275820643`

Madhusudan, S., Morabito, R. D., Reid, S., Sadr, N. G., & Emami, A. (2025). Fine-tuned llms are "time capsules" for tracking societal bias through books. *ArXiv*, *abs/2502.05331*. Retrieved from `https://api.semanticscholar.org/CorpusID:276249448`

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., & Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? In *The thirty-eighth annual conference on neural information processing systems.* Retrieved from `https://openreview.net/forum?id=DV15UbHCY1`

Wang, J., Tong, J., Tan, K. L., Vorobeychik, Y., & Kantaros, Y. (2023). Conformal temporal logic planning using large language models: Knowing when to do what and when to ask for help. *ArXiv*, *abs/2309.10092*. Retrieved from `https://api.semanticscholar.org/CorpusID:262054964`