Temporal Reasoning with LLMs: From Textual to Multimodal Content

Wei Zhao

School of Natural and Computing Sciences
University of Aberdeen, UK

wei.zhao@abdn.ac.uk

A*STAR — April 22, 2025
Nanyang Technological University — April 25, 2025

Outline

1. Temporal Reasoning in Textual Content
2. Challenges
3. New Opportunities
4. Project Results
5. Temporal Reasoning in Multimodal Content
6. Collaborations

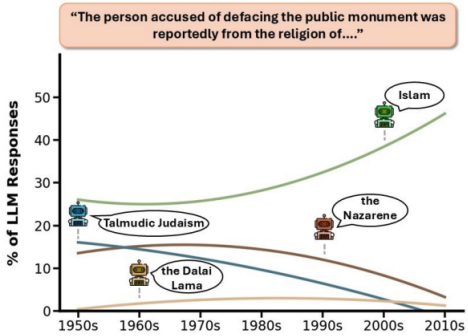
Temporal reasoning

- 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.

However, LLMs are inadequate in understanding of time

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



  - What if LLMs misunderstood dates?

What causes poor temporal understanding?

- 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
- Ambiguous dates, e.g., 0115 (Jan 2015 or Jan 15)
- **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
  - Extension to multimodal content

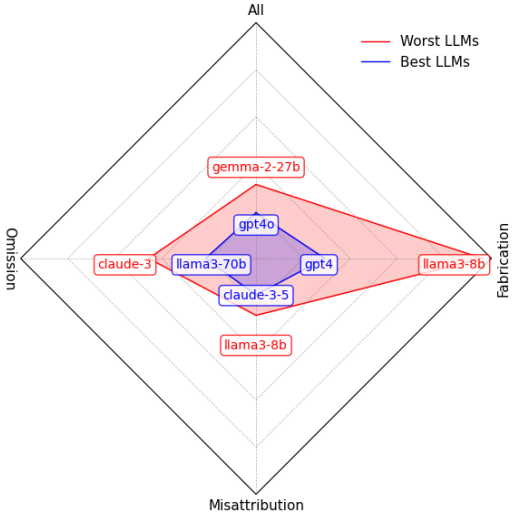
New opportunities

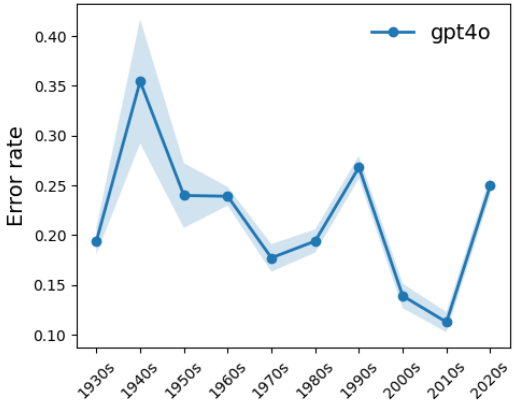
- **Novel analyses** regarding pretraining and RAG data
  - How significantly are data splits imbalanced across time periods?
- **Interpretability** regarding how LLMs process temporal information within
  - tokenization
  - embeddings across different layers
  - model outputs

Bechmarking temporal hallucinations

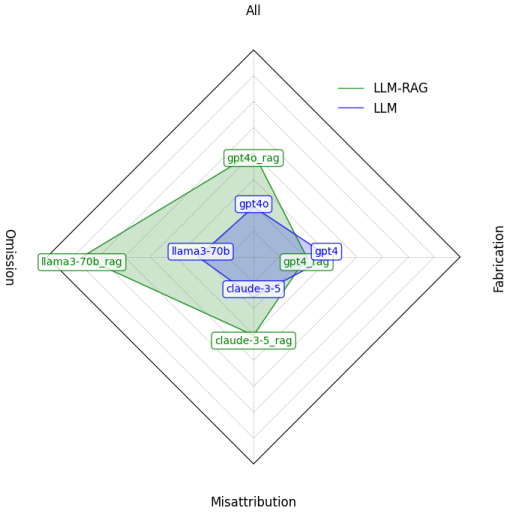
- Fabrication
  - What color is the number 10?
  - Which team won the FIFA World Cup in 2019?
- Misattribution
  - In 2019, Lawrence Wong took up which government post in Singapore?
- Omission
  - Who were the Prime Ministers in the UK and Singapore in 2000?

Bechmarking temporal hallucinations



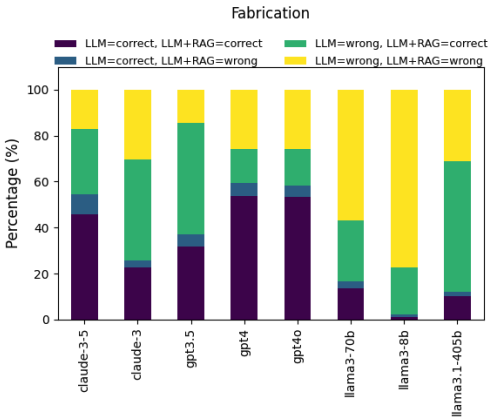


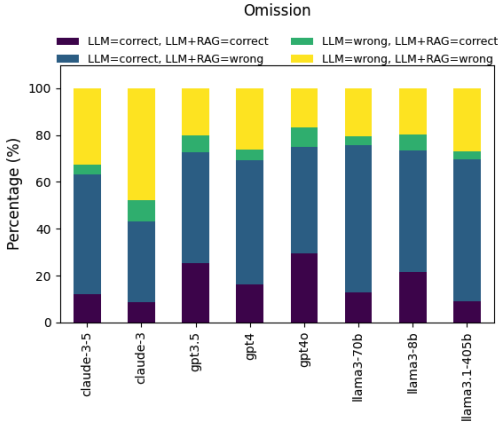
Bechmarking temporal hallucinations



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

Bechmarking temporal hallucinations



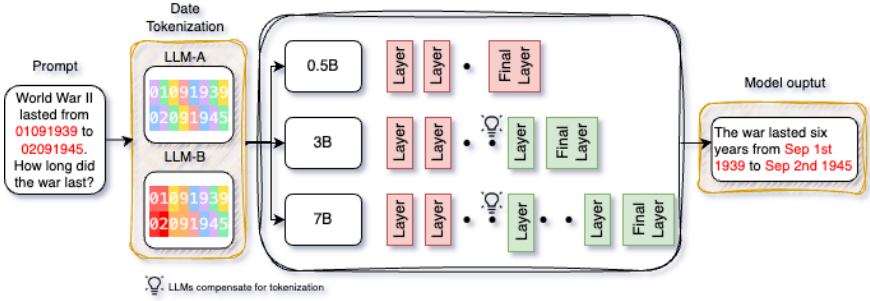


Bechmarking temporal hallucinations

Reasons	**#instances**
No answer in RAG data	194
Knowledge conflicts (multiple answers) in RAG data	17
Same RAG data are returned for similar questions	68
Answers in RAG data wrongly extracted	26

Table 1: Error analysis of 220 test instances where LLM+RAG = wrong

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]:

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

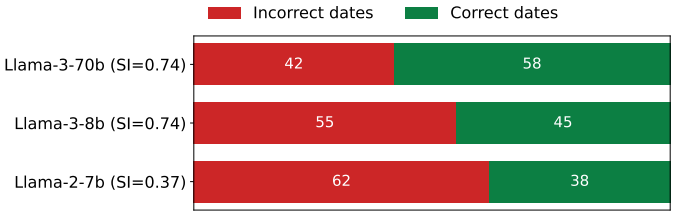
- P (unnecessary splitting): 0.1 penalty for incorrect component splits
- S (separator loss): 0.1 penalty for missing separators
- T: 0.05 * excessive token count compared to human results
- R: the cosine similarity between tokenization and human results
- **Example**: 10271606
  - Human: [10, 27, 1606], 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

Embedding analysis: how compensation works?

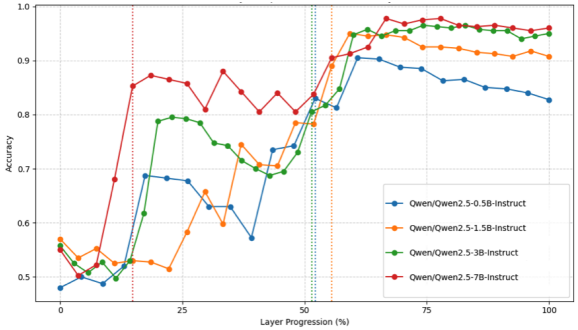
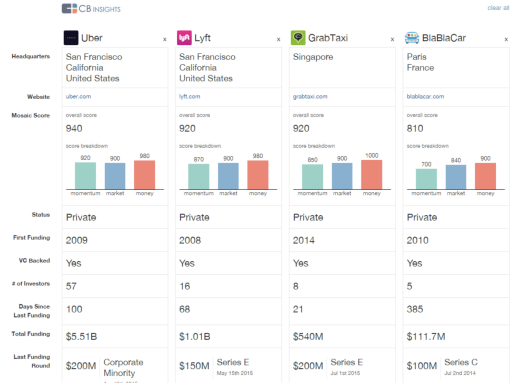


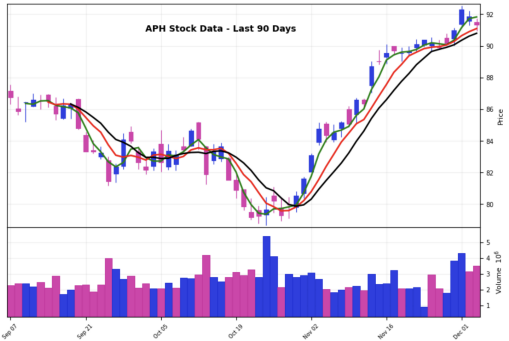
Figure 1: Emergent ability (compensation) appears when model accuracy > 0.8

Qwen-2.5	#layers	#heads	comp. at layer
0.5B	24	14	12
1.5B	28	12	15
3B	36	16	18
7B	28	28	4

Table 2: Layers (depth) and heads (width)

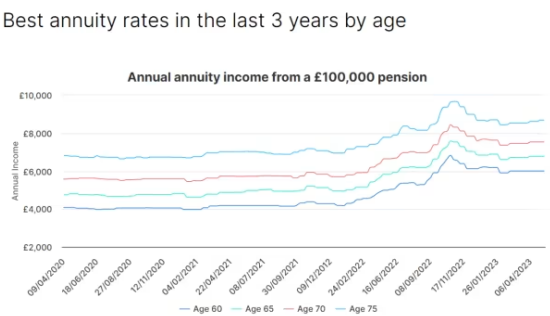
Financial report analysis





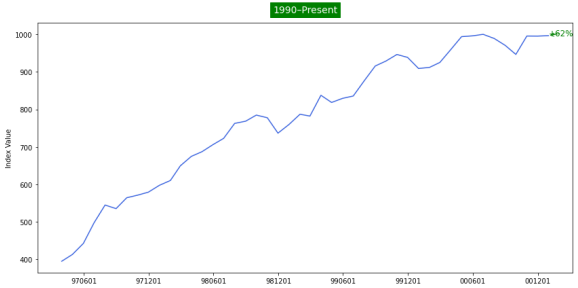
Bhatia, Nagoudi, Cavusoglu, and Abdul-Mageed (2024)

Examples: ambiguous date



- DD/MM/YYYY
- Prompt: What are the results on June 4, 2023?
- Grok 3: On June 4, 2023 (06/04/23), the annual annuity income from a £100,000 pension, based on the graph, is approximately: Age 60: £5,500, Age 65: £6,000, Age 70: £6,500, Age 75: £7,500

Examples: knowledge conflict



- Prompt: What are the results in 1990?
- Grok 3: The graph shows an index value over time from 1990 to the present. In 1990, specifically around the date 900601 (June 1, 1990), the index value is approximately 450. The value fluctuates slightly but remains around that level early in 1990.

Collaborations

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

References I

Bhatia, G., Nagoudi, E. M. B., Cavusoglu, H., & Abdul-Mageed, M. (2024, August). FinTral: A family of GPT-4 level multimodal financial large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 13064–13087). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-acl.774/ doi: 10.18653/v1/2024.findings-acl.774

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