RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Submission ID: 778
Mentor Name: Dr. Felix Muzny
Undergraduate

Northeastern University

Internalized Biases of Time Encoding in Commercial Large Language Models

Yuliya Buturlia

Question

Do Large Language Models (LLMs) have internalized biases for time encoding?
Do they superimpose English standards on other languages?

Background

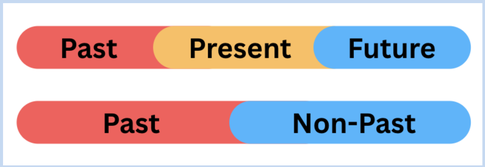
LLM 'understanding' is a salient issue currently due to widespread integration

- 78% of businesses reported using AI in 2024[1]

Most widely spoken languages today have extensive digital support and widespread internet usage

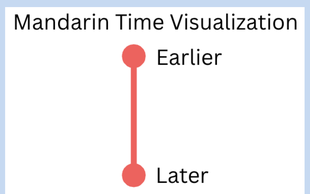
- 40% of the world's languages are endangered, many due to dominance of other languages

The dominant language of an LLM has influence on a critical global scale.



The language you speak can frame how you think about time

- Mandarin speakers are more likely to think about time vertically than English speaker[2]



Temporal reasoning in LLMs is currently being studied in greater detail:

- Moving towards neuro-symbolic temporal reasoning, where LLMs can generate Python code to perform temporal calculations[3]
- Creating benchmarks to evaluate LLM performance in reasoning about relationships between events and time, as well as between events in the real world[4]

As is LLM use in multi-lingual settings:

- Looking to Small Language Models for translating low-resource languages that lack sufficient representation in datasets for high quality translation[5]
- Creating LLM-aided preservation efforts for African indigenous languages with specialized mechanisms for linguistic nuances like tonal patterns[6]
- Investigating temporal generalization, which encompasses an LLM's ability to understand, predict, and generate text relevant to past, present, and future contexts[7]

But what about the intersection of multi-lingual temporal reasoning?

Process and Methods

Temporal reasoning was evaluated with linguistic markers of time like tense, aspect, and mood (TAM) and through cultural/societal references to events tied to specific temporal relations (e.g., breakfast happens in the morning)

- Russian was chosen as the second language of evaluation for its high-resource advantages and ease of qualitative analysis
- Parallel data of English and Russian prompts were created for:
  - Few-shot time sequencing events
  - Targeted aspect-centric prompts to analyze known differences between the languages
- GPT-5.1, Sonnet 4.6, and Gemini 3 were tested with all prompts

1 Time Sequencing

Before **he brushed his teeth, washed his face, and got dressed**, he realized he was hungry. What meal will he eat?

↓

Before **he brushed his teeth and washed his face**, he realized he was hungry. What meal will he eat?

↓

Before **he washed his face**, he realized he was hungry. What meal will he eat?

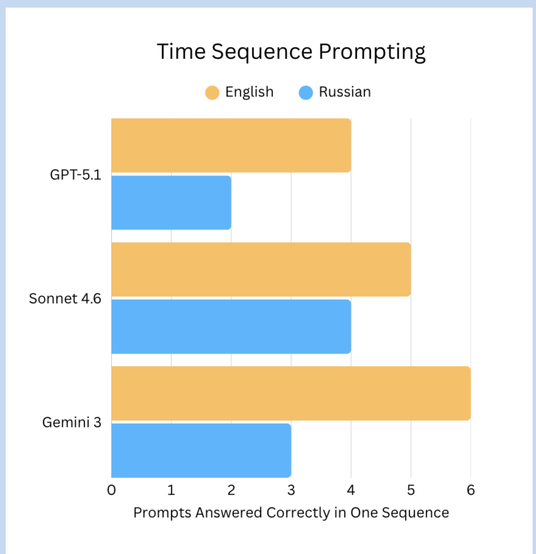
2 English-Russian Time Parallel

Я прочитаю ваше письмо завтра.
I will read your email tomorrow
(the entire email, completely).

Я буду читать ваше письмо завтра.
I will be reading your email tomorrow
(in progress).

3 Direct Comparison of All Translations

Findings



LLMs do well with:

- Temporal adverbs without a multi-prompt sequence
- Numeric times/dates
- Tense time distinctions

LLMs do **not** do well with:

- Habitual aspect, especially in Russian
- Temporal adverbs in sequence

LLM prompts that included "reason in Russian but respond in English" yielded nearly the same result as the standard English prompt

- Prompts relating to breakfast included suggestions referring to common markedly American breakfast foods

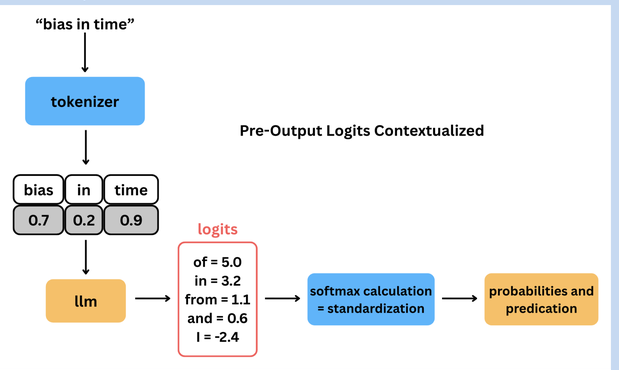
Conclusions

1. LLMs can struggle with vague references to time, although their error is often assuming a time of day when a human would say it is unclear.
2. In Russian, LLMs struggle with habitual time encodings in verbs, especially when the representation would be notably different in English.

Future and Current Work

This analysis was largely qualitative with a group of Russian speakers for validation. All next steps are moving towards forming a robust quantitative analysis of time representation in non-English languages.

1. Adding Mandarin Chinese to the analysis, as it does not have a formal tense system
2. Finding a language with robust realis/irrealis morphological distinctions with enough data for parallel analysis with other target languages. This will result in a complete tense, aspect, and mood analysis, providing a stronger linguistic foundation.
3. Using pre-output logits from an LLM, moving away from commercial models for analysis to medium local models for logit extraction, to assess model uncertainty in final output, forming a quantitative result that can accurately portray multi-lingual time representation



While the conclusion from this analysis supports a need to investigate the potential bias in multi-lingual LLM responses in more depth, a further analysis is necessary to determine the true implications and scope of any English-favoring biases models may exhibit.

Acknowledgments

Thank you to Dr. Felix Muzny and Dr. Meica Magnani for their mentorship and guidance throughout this project, as well as Dr. Terra Blevins for ongoing advising and mentorship in the next steps.

References

1. McKinsey & Company. The state of AI: How organizations are rewiring to capture value. McKinsey & Company. Published March 12, 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
2. Boroditsky L, Fuhrman O, McCormick K. Do English and Mandarin speakers think about time differently? *Cognition*. 2011;118(1):123-129. doi:https://doi.org/10.1016/j.cognition.2010.09.010
3. Ge Y, Romeo S, Cai J, et al. TReMu: Towards Neuro-Symbolic Temporal Reasoning for LLM-Agents with Memory in Multi-Session Dialogues. arXiv.org. Published February 3, 2025. Accessed February 25, 2026. https://arxiv.org/abs/2502.01630
4. Sigang Luo, Yinan Liu, Dongying Lin, Yingying Zhai, Bin Wang, Xiaochun Yang, and Junpeng Liu. 2025. ETRQA: A Comprehensive Benchmark for Evaluating Event Temporal Reasoning Abilities of Large Language Models. In Findings of the Association for Computational Linguistics: ACL 2025, pages 23321–23339, Vienna, Austria. Association for Computational Linguistics.
5. Lothritz C, Cabot J, Bernardy L. Testing Low-Resource Language Support in LLMs Using Language Proficiency Exams: the Case of Luxembourgish. arXiv.org. Published 2025. Accessed February 25, 2026. https://arxiv.org/abs/2504.01667
6. Uchechukwu Ajuzieogu. Multimodal Generative AI for African Language Preservation: A Framework for Language Documentation and Revitalization. ResearchGate. doi:https://doi.org/10.13140/RG.2.2.23144.99842
7. Zhu C, Chen N, Gao Y, Zhang Y, Tiwari P, Wang B. Is Your LLM Outdated? A Deep Look at Temporal Generalization. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (2025) 7433-7457*. Published online January 1, 2025:7433-7457. doi:https://doi.org/10.18653/v1/2025.naacl-long.381