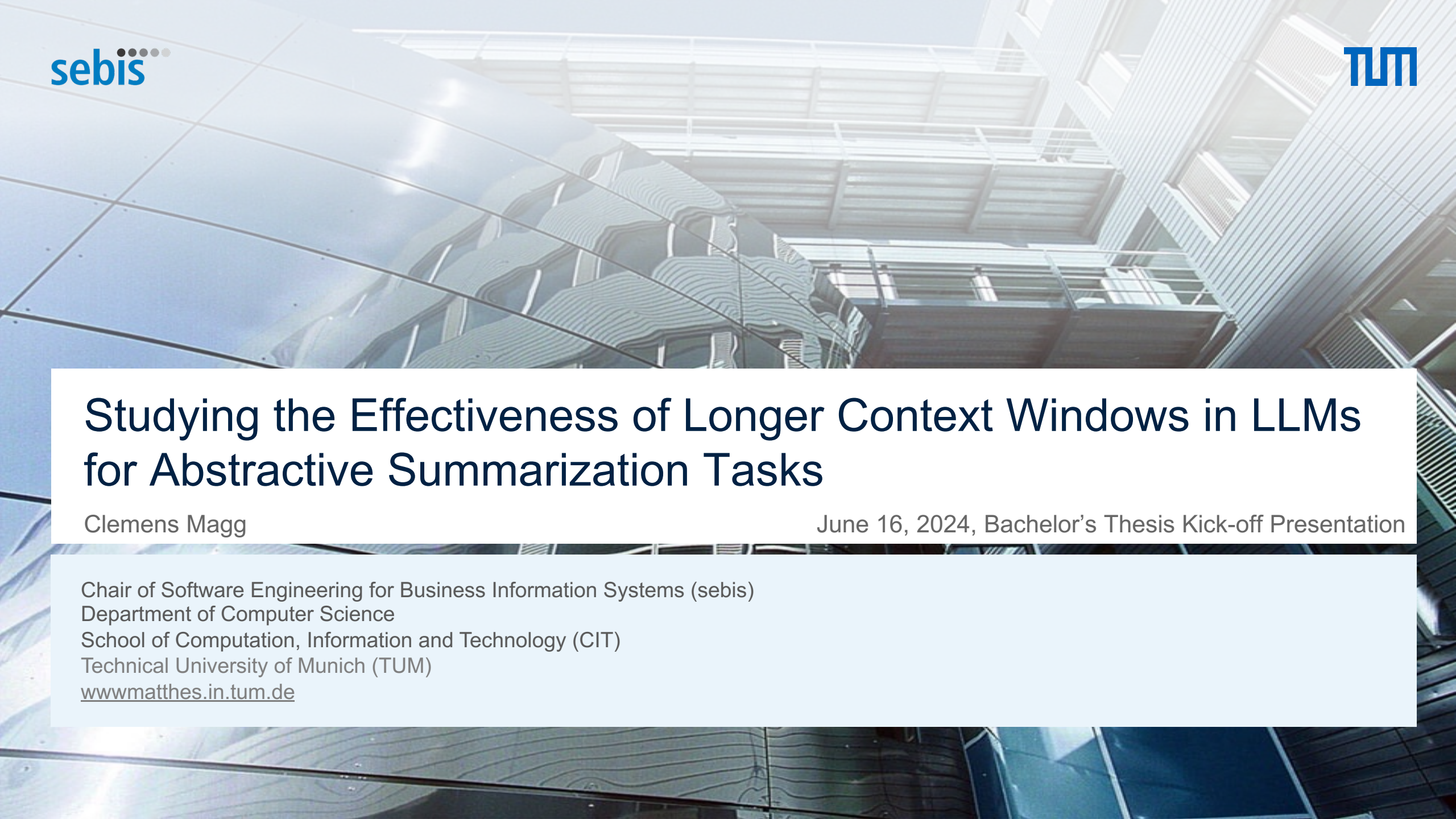


Studying the Effectiveness of Longer Context Windows in LLMs for Abstractive Summarization Tasks

Clemens Magg

June 16, 2024, Bachelor's Thesis Kick-off Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Background & Motivation

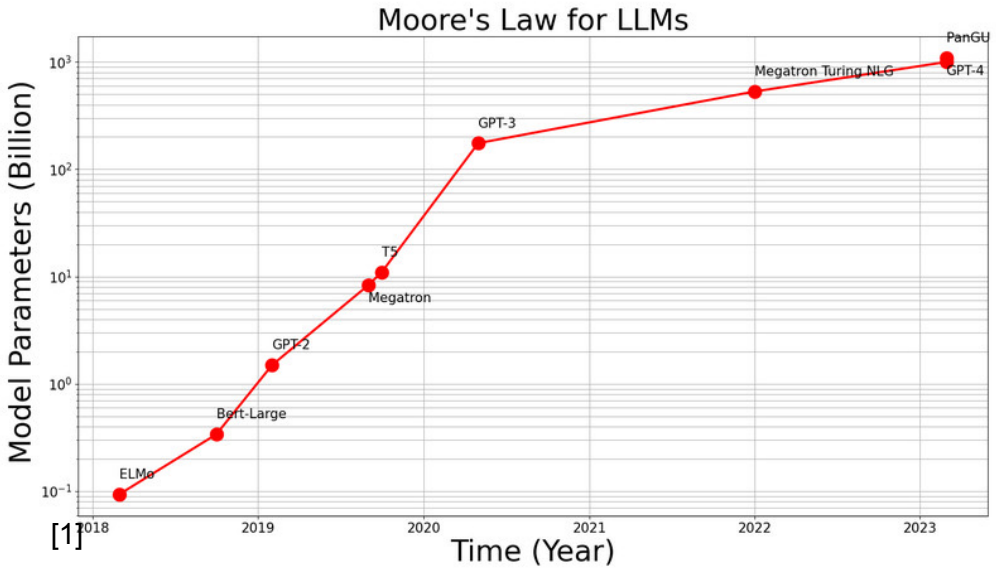
Research Questions

Methodology

Initial Findings

Timeline

Background & Motivation : Complexity and Performance of LLMs



[1]

- Large Language Models (LLMs) have grown drastically in size.
- Transformer architecture coupled with improved hardware.
- Research focuses on enabling long-text comprehension of LLMs.
- Context window extension techniques aim to increase the input size LLMs can process effectively.

Background & Motivation : Generative Text Summarization

- Data is as accessible as never before.
- However, its sheer amount makes it difficult to process information effectively.
- Automatic and accurate text summarization helps mitigate the difficulty of processing large amounts of data.
- From summarizing articles with a few thousand tokens to books with millions of tokens, LLMs are used across various domains to generate summaries.
- Comprehension of extremely long sequences is increasingly important for generating useful summarizations.
- Effective summarization requires long-range dependency comprehension and information retrieval.

Background & Motivation: Context Window Extension Techniques

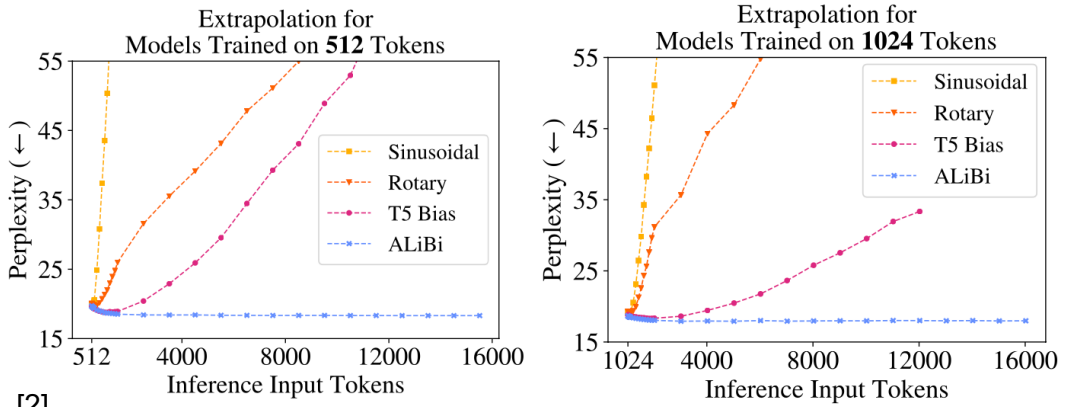
- Researchers developed numerous techniques for extending the context window size of LLMs.
- Increasing the input length enhances semantic understanding and enables LLMs to capture long-range dependencies [2].

Positional embedding techniques:

- ALiBi
- Positional Interpolation and YaRN
- LongRoPE

Specialized Attention Mechanism and Memory Retrieval:

- Focused transformer
- Landmark attention



[2]

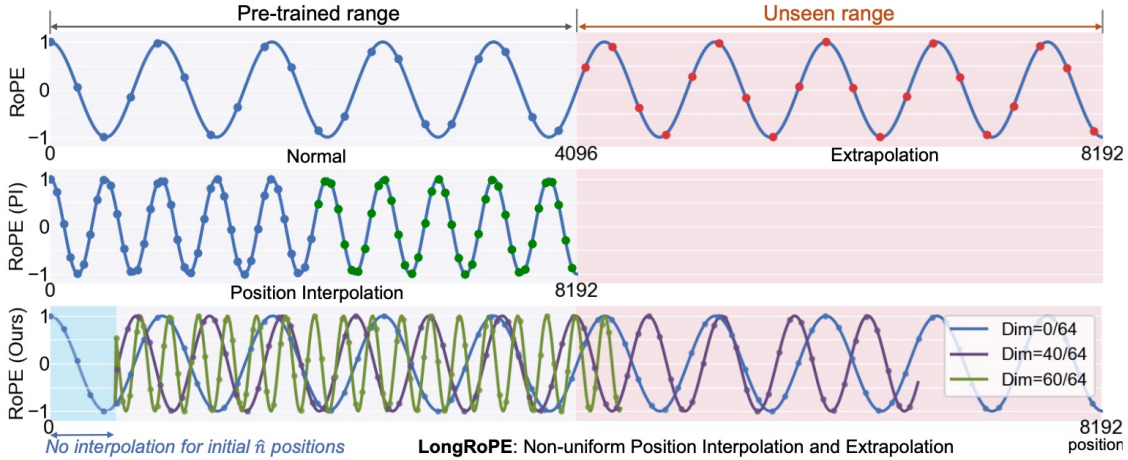


Figure 2. An illustrative example to show RoPE embedding under different interpolation methods. *Upper*: RoPE under direct extrapolation. *Middle*: Rescaled RoPE under linear positional interpolation. *Down*: LongRoPE fully exploits the identified two non-uniformities, leading to varied interpolation and extrapolation across RoPE dimensions at different token positions.

[3]

Background & Motivation: Benchmarking Long Context Window LLMs

- Many summarization benchmarks rely on standard n-gram metrics like ROUGE, F1, or simple perplexity.
- XL2-bench, LongBench, InfinityBench, L-eval, etc., use standard metrics to provide general performance indicators and neglect semantic nuances and contextual details.
- Proven benchmarks can only test models with a limited context window size (max. ca. 10k tokens).
- How and where information is derived to generate the output is not measured.

Background & Motivation: Benchmarking Long Context Window LLMs

- Benchmarking the discussed context window extension techniques on the open source Llama-2/3 model.
- Investigating the effect of incrementally increasing the context window length.
- Exploring the effect of feeding more information to the model.
- Using a relatively small testing corpus enables human evaluation.
- Visualizing performance trends across multiple context window lengths using mock-up tools.
- Mapping information of the generated summarization to their origin in the document. Where does the model derive its information, depending on the input length and the technique.

Background & Motivation: Benchmark Dataset Decision

- Many different datasets are used to evaluate the text summarization performance of LLMs.

Deciding on a dataset:
- Average sequence length
- Domain
- Language

RedPajama-Data-V2:
- Widely used dataset containing 30T documents.
- Multilingual text summarization tasks.
- Includes arXiv, a dataset that includes scientific articles.
- Filter dataset upon sequence length, language, and domain.

Research Questions

RQ1	What are the most effective techniques for extending the context window of LLMs?
RQ2	How do LLMs use the information contained in their context? Do LLMs benefit from a long context window for text summarization? Is the model able to pay attention to all parts of the document, or is it clustered toward some parts?
RQ3	How can we adequately test the quality of text summarization of LLMs? Does the quality of the generated summary improve if more content of the article is passed?

Outline

Background & Motivation

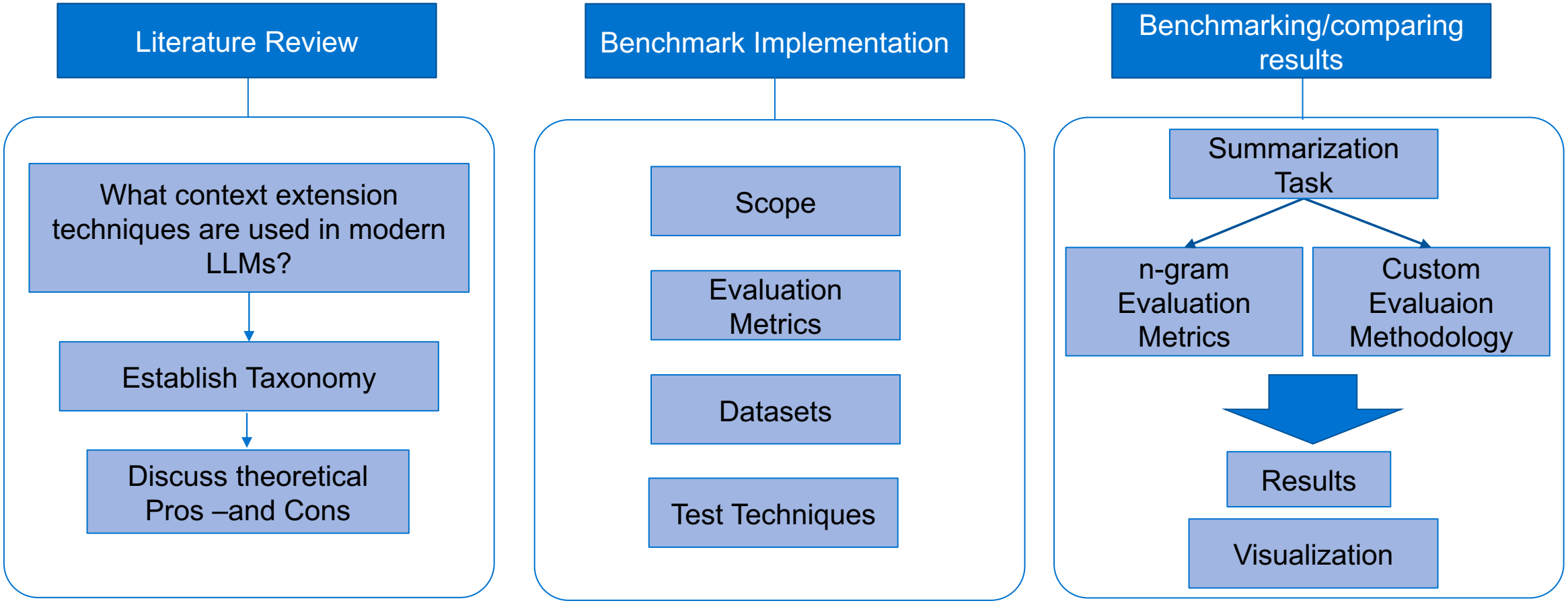
Research Questions

Methodology

Initial Findings

Timeline

Methodology



Outline

Background & Motivation

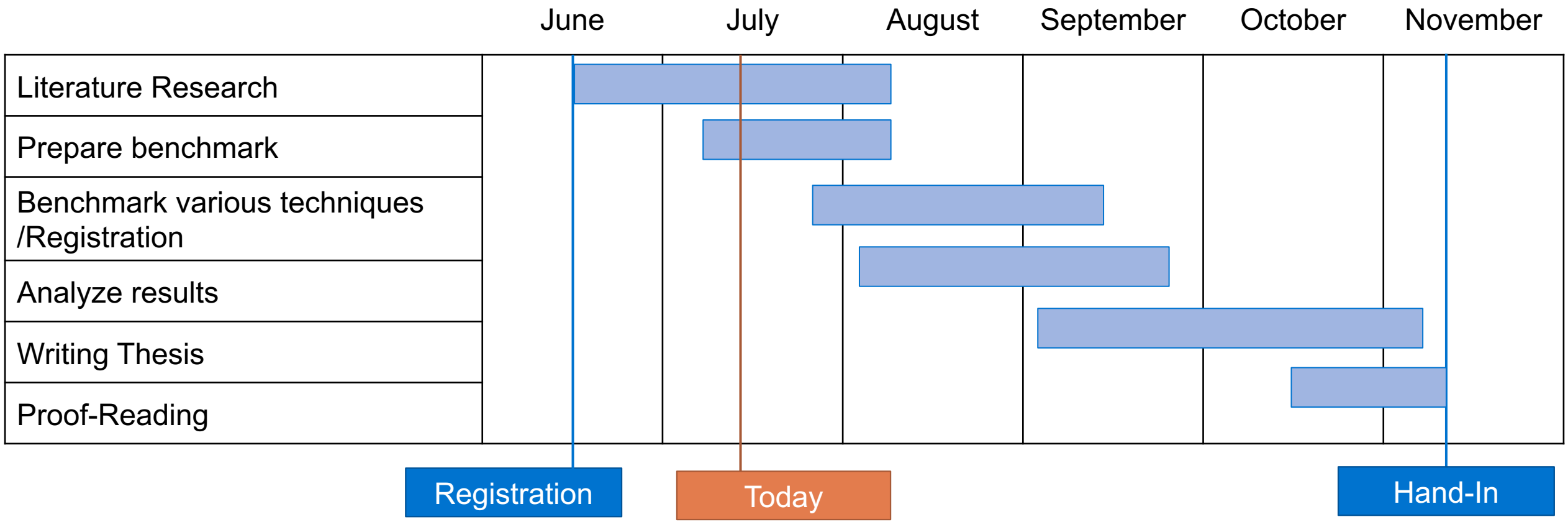
Research Questions

Methodology

Initial Findings

Timeline

Timeline



Prof. Dr.
Florian Matthes

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

Sources

[1] https://www.researchgate.net/figure/LLMs-A-New-Moores-Law-as-presented-the-size-of-models-grows-exponentially-with-time_fig2_372248458

[2] Pawar, Saurav, et al. "The What, Why, and How of Context Length Extension Techniques in Large Language Models--A Detailed Survey." *arXiv preprint arXiv:2401.07872* (2024).

[3] Press, O., Smith, N. A., & Lewis, M. (2021). Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

[4] Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., ... & Yang, M. (2024). Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

[5] Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., ... & Li, J. (2023). Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Prof. Dr.
Florian Matthes

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de