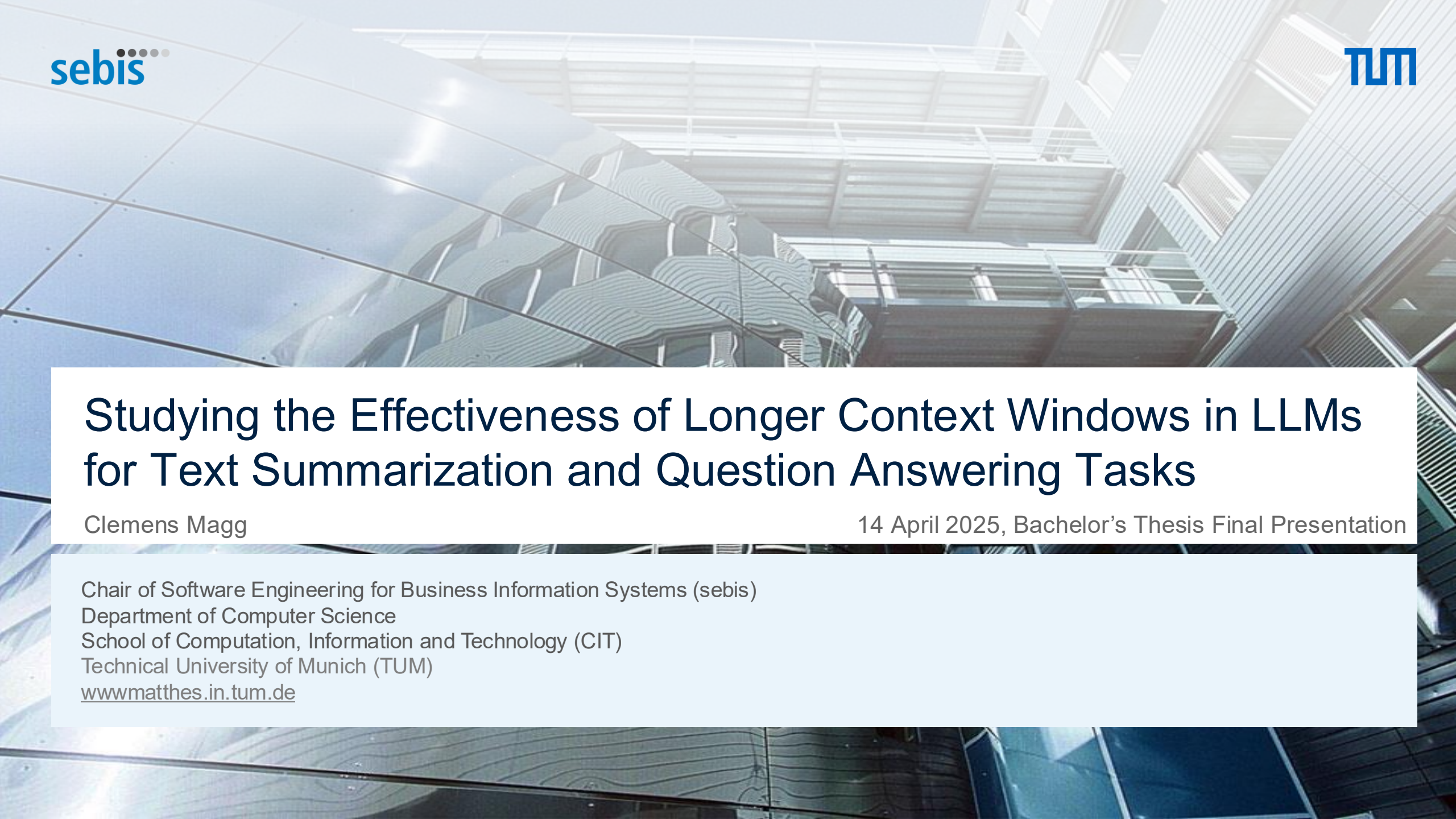


Studying the Effectiveness of Longer Context Windows in LLMs for Text Summarization and Question Answering Tasks

Clemens Magg

14 April 2025, Bachelor's Thesis Final 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

1) Introduction
- Motivation
- Research Questions

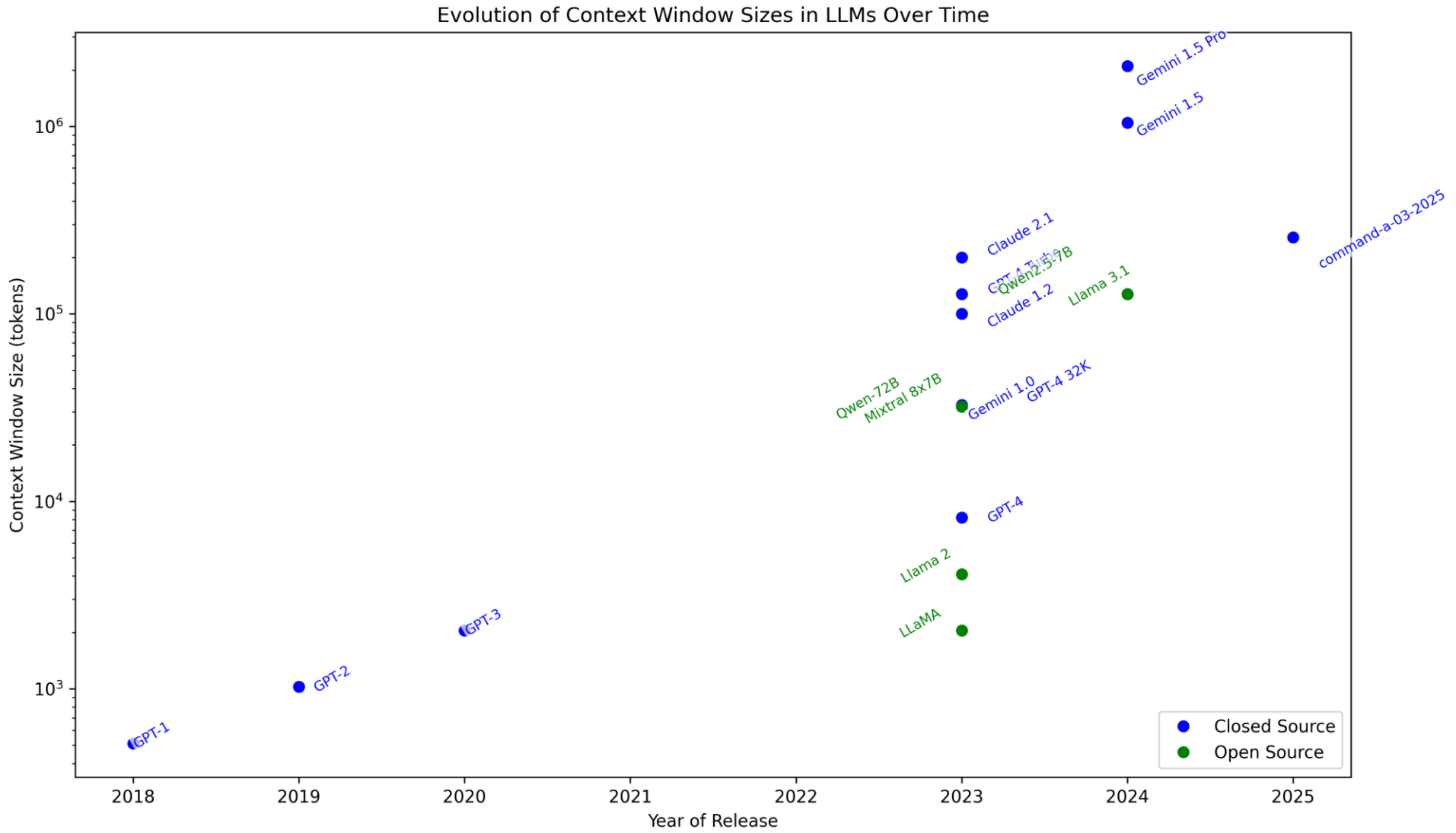
2) Methodology

3) Datasets

4) Results

5) Discussion
- Contributions & Limitations

Motivation



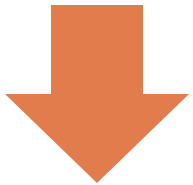
Motivation

Rapid development of NLP systems and increase in computing power resulted in:

- More powerful and bigger models
- Context window length of LLMs increases exponentially

LLMs used as text summarizers and for question answering have potential but have to be tested rigorously

- Most contemporary benchmarks for similar tasks test only for a limited context window length



- Introducing our evaluation pipeline for testing LLMs with long context window lengths (up to 128k tokens)

LLM-Benchmarks

- Benchmarks aim to examine the performance of LLMs on various tasks under different conditions.
- Examples: **SCROLLS, L-Eval, LongBench.**

- Most existing benchmarks offer testing for only a limited context window size.
- Involve shallow evaluation metrics like ROUGE comparison of candidate to reference summary.

Research Questions

How can we adequately test the quality of text summarization and question answering produced by LLMs? Does the quality of the generated summary or answer improve when more content from the article is provided to the model?

What are the most effective techniques for extending the context window of LLMs?

How do LLMs utilize the information within their context window during text summarization and question-answering tasks? Can LLMs distribute their attention uniformly across the entire document, or is their attention clustered towards specific sections?

Outline

1) Introduction
- Motivation
- Research Questions

2) Methodology

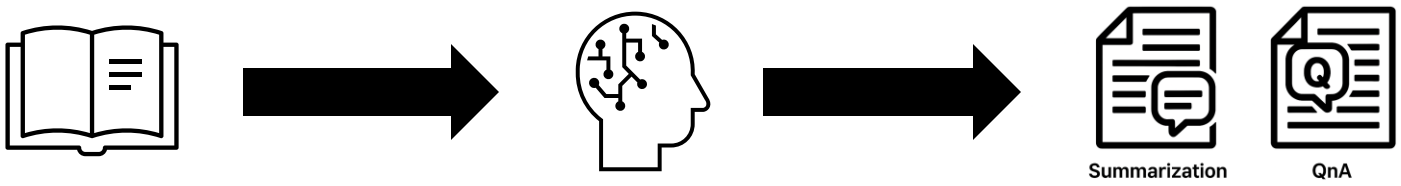
3) Datasets

4) Results

5) Discussion
- Contributions & Limitations

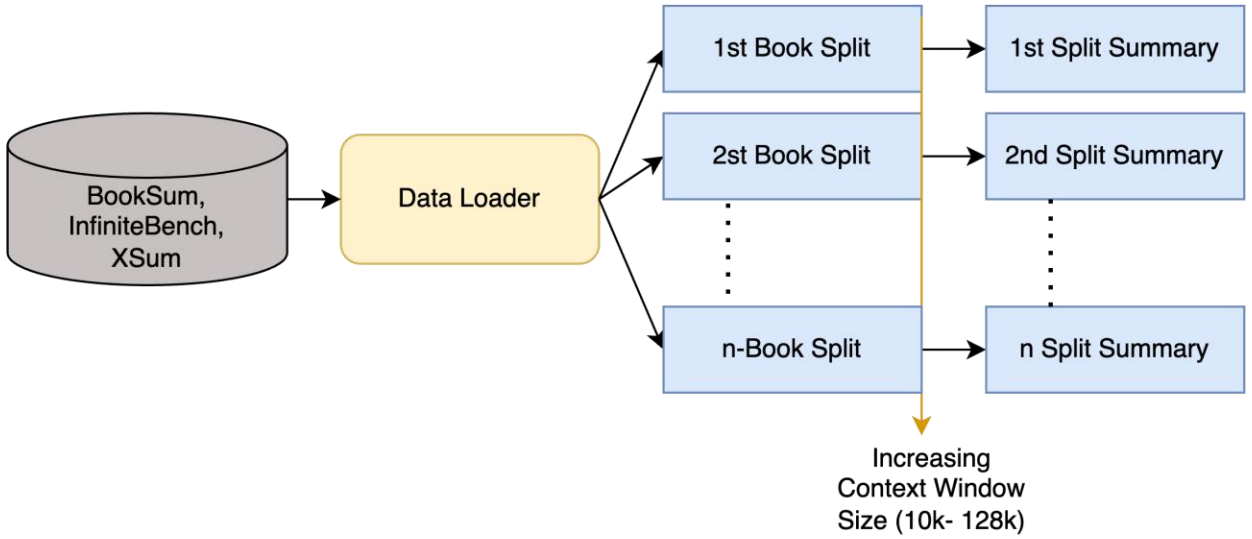
Primary Evaluation Goals (RQ2)

- Can modern LLMs distribute their attention evenly or do they observe positional biases within their context window?



- Do larger context windows lead to better performance? Does their attention distribution change depending on the context window size?

Evaluation Pipeline: Summarization Workflow



- Dataset entries are split into parts of incremental size (from 10k to 128k tokens).
- We use zero-shot prompting for generating abstractive summaries for each book and book parts.

Each summary is then evaluated on two metrics:

- Positional analysis
- Atomic facts analysis

Evaluation Pipeline: Positional Analysis

- Sentence-paragraph matching through two different similarity metrics: Sentence Transformer, TF-IDF
- Tracing information from each summary sentence to its most likely position within the source document

Jennifer is an earnest intelligent woman who makes a serious error in judgment when she chooses to marry Mina Loris, a pompous scholar many years her senior.

But now Terence was really startled at the suspicion which had darted into her mind. She was seldom taken by surprise in this way, her marvellous quickness in observing a certain order of signs generally preparing her to expect such outward events as she had an interest in. Not that she now imagined Mr. Loris to be already an accepted lover: she had only begun to feel disgust at the possibility that anything in Jennifer's mind could tend towards such an issue. Here was something really to vex her about Dodo: it was all very well not to accept Sir Briar Bronwen, but the idea of marrying Mr. Loris! Terence felt a sort of shame mingled with a sense of the ludicrous. But perhaps Dodo, if she were really bordering on such an extravagance, might be turned away from it: experience had often shown that her impressibility might be calculated on. The day was damp, and they were not going to walk out, so they both went up to their sitting-room; and there Terence observed that Jennifer, instead of settling down with her usual diligent interest to some occupation, simply leaned her elbow on an open book and looked out of the window at the great cedar silvered with the damp. She herself had taken up the making of a toy for the curate's children, and was not going to enter on any subject too precipitately.

Evaluation Pipeline: Atomic Facts Analysis

- Utilizes atomic facts extraction from ***FActScore*** [1].
- An **atomic fact** is a concise sentence that conveys a single piece of information
- We adapt Factscore with GPT-4o-mini and change the prompt template to fit the abstract nature of the data better.

- For each book and book part, we extract the atomic facts from each sentence summary.
- Pair-wise comparison of **incremental context window sizes**.
- Categorization into **matching, new,** and **lost facts.**

Few-Shot Prompt Template:

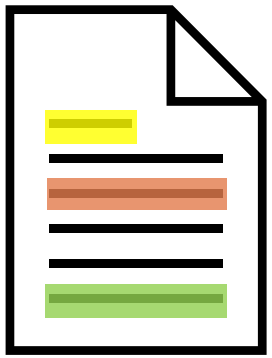
Tom toils in a factory job he loathes to support his aging Southern belle mother, Amanda, and his disabled, crushingly shy sister, Laura.

- Tom works in a factory job he loathes.
- He supports his aging mother, Amanda.
- Amanda is a Southern belle.
- He also supports his disabled, shy sister, Laura.

[1] https://github.com/shmsw25/FActScore

Evaluation Pipeline: Question Answering

- Abstract: In the excerpt, how do the characters perceive Mr. Mya's pride, and what differentiates pride from vanity according to Edie's reflection?
- Extractive: What does Miss Lilia say about the young man's pride in the excerpt?
- Binary: Does Andromeda feel certain about the degree of her own regard for Mr. Adrian after knowing him for a fortnight?
- Unanswerable: What specific illness is Andromeda suffering from during her stay at Netherfield Park?



- **Split the context window into three parts**; each gets four questions of all types assigned.
- The questions are based on a **paragraph in each section of the input document.**
- **Question- answer pair** generated with GPT-4o-mini

Outline

1) Introduction
- Motivation
- Research Questions

2) Methodology

3) Datasets

4) Results

5) Discussion
- Contributions & Limitations

Datasets

BookSum, InfiniteBench, XSum

	book	summary
0	\n "Mine ear is open, and my heart prepared:\n The worst is worldly loss thou canst unfold:\n Say, is my kingdom lost?"\n\n SHAKESPEARE.\n\n\nIt was a feature peculiar to the colonial wars of North America, that\nthe toils and dangers of the ...	Before any characters appear, the time and geography are made clear. Though it is the last war that England and France waged for a country that neither would retain, the wilderness between the forces still has to be overcome first. Thus it is in ...
1	\n "Before these fields were shorn and tilled,\n Full to the brim our rivers flowed;\n The melody of waters filled\n The fresh and boundless wood;\n And torrents dashed, and rivulets played,\n And fountains spouted in the shade."\n\n ...	In another part of the forest by the river a few miles to the west, Hawkeye and Chingachgook appear to be waiting for someone as they talk with low voices. It is now afternoon. The Indian and the scout are attired according to their forest habits...
2	\n "Well, go thy way: thou shalt not from this grove\n Till I torment thee for this injury."\n\n _Midsummer Night's Dream._\n\n\nThe words were still in the mouth of the scout, when the leader of the\nparty, whose approaching footsteps had cau...	When the mounted party from Fort Howard approaches the three men of the woods, Hawkeye addresses first Gamut and then Heyward only to learn that they are lost because their Indian guide has taken them west instead of north toward Fort William Hen...
3	\n "In such a night\n Did Thisbe fearfully o'ertrip the dew;\n And saw the lion's shadow ere himself."\n\n _Merchant of Venice._\n\n\nThe suddenness of the flight of his guide, and the wild cries of the\npursuers, caused H...	The pursuit of Magua is unsuccessful, but Hawkeye feels that he has wounded him slightly and is certain of it when they find bloodstains on the sumach leaves. Heyward wants to continue the chase, but the scout fears an ambush, particularly since ...

- Novels and Articles
- Highly abstractive and non-redundant reference summaries
- Padded or truncated to 128k tokens
- Even distribution of all relevant information

- BookSum: 405 novels and plays
- InfiniteBench: 103 books
- Xsum: 204.045 news articles

Outline

1) Introduction
- Motivation
- Research Questions

2) Methodology

3) Datasets

4) Results

5) Discussion
- Contributions & Limitations

Model Selection (RQ2)

Cohere Command-R7B

Proprietary Baseline Model

→ RoPE

LLaMA-3.1-8B

→ RoPE

Qwen2.5-7B

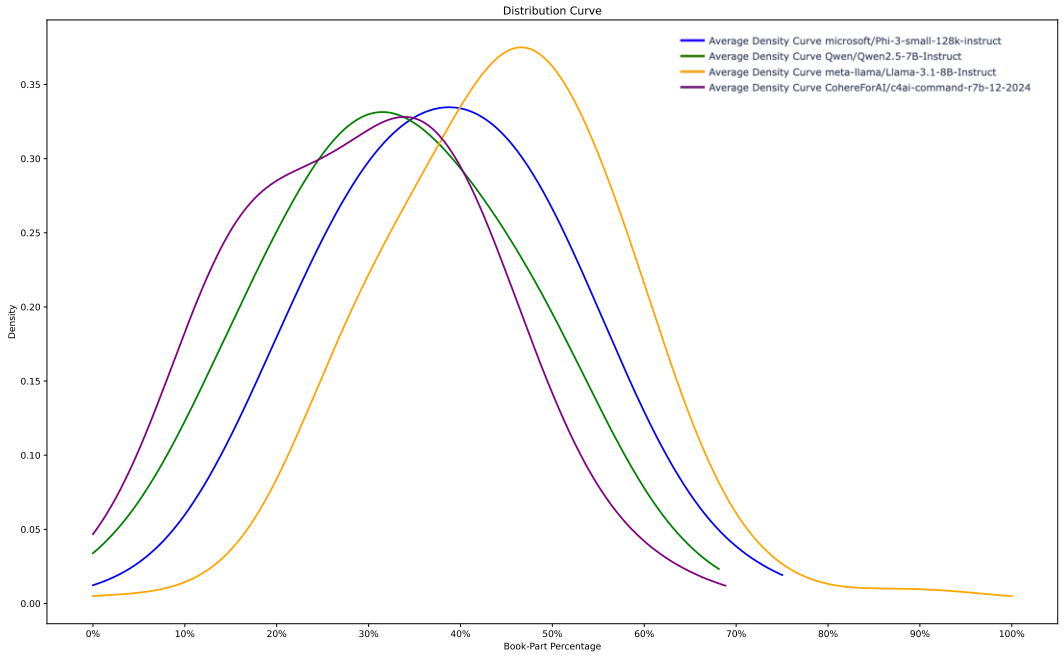
→ YaRN

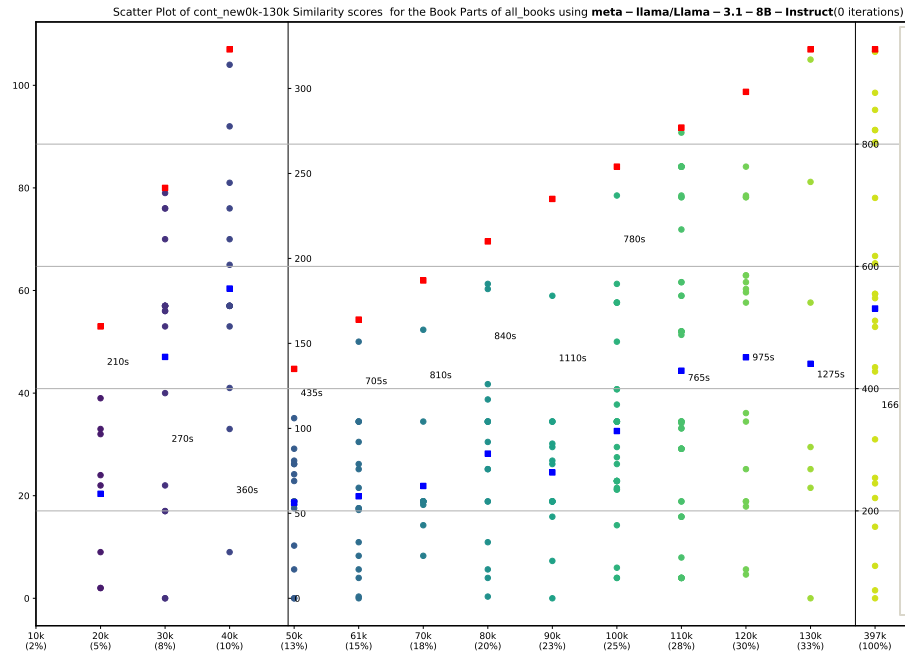
Phi-3-medium

→ LongRoPE

Results: Positional Analysis (RQ3)

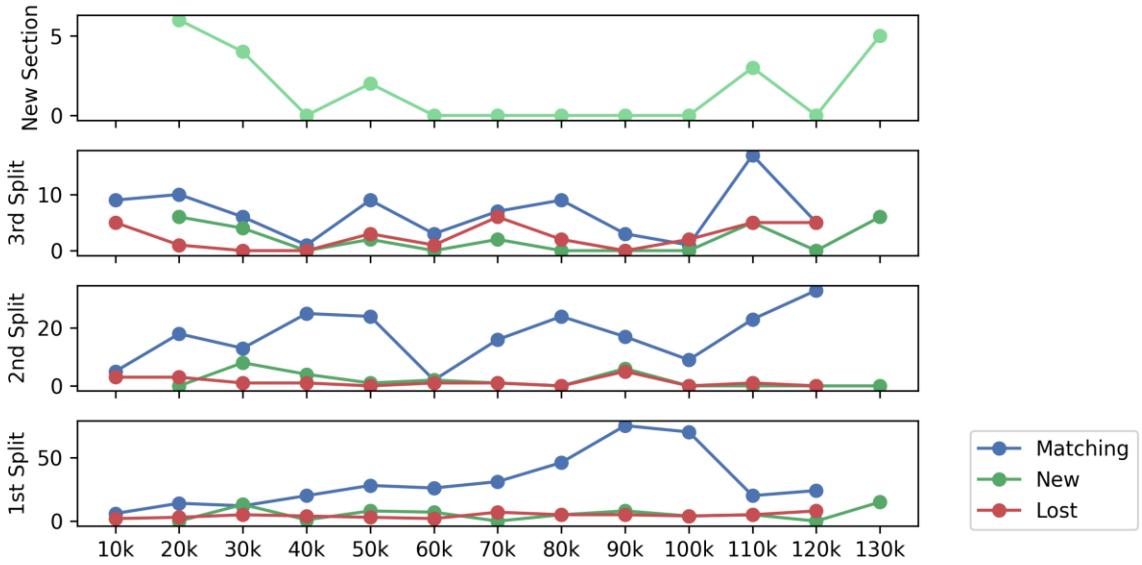
- Skewed attention for all tested models.
- Bias towards early parts of the context window.
- this trend is exaggerated for all models if we gradually increase the context window size.



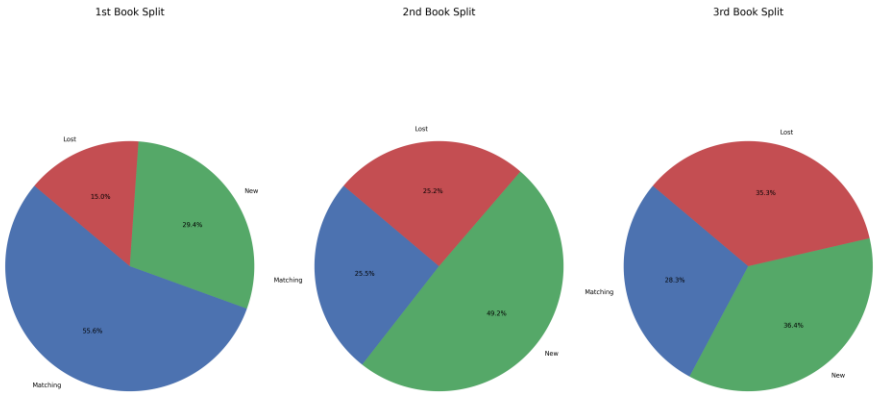


Results: Atomic Facts Analysis (RQ3)

- The models can retain information from earlier parts of their context window.
- They have difficulty extracting new information and retaining details found in the later parts of the context window.



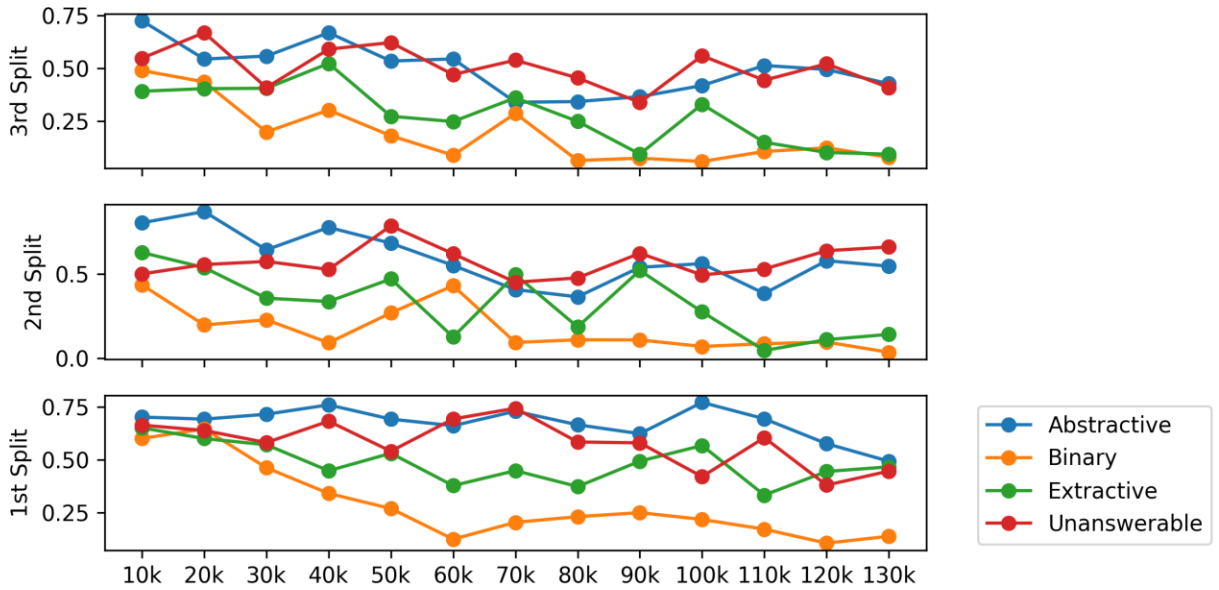
command-r7b (booksum)



command-r7b (booksum)

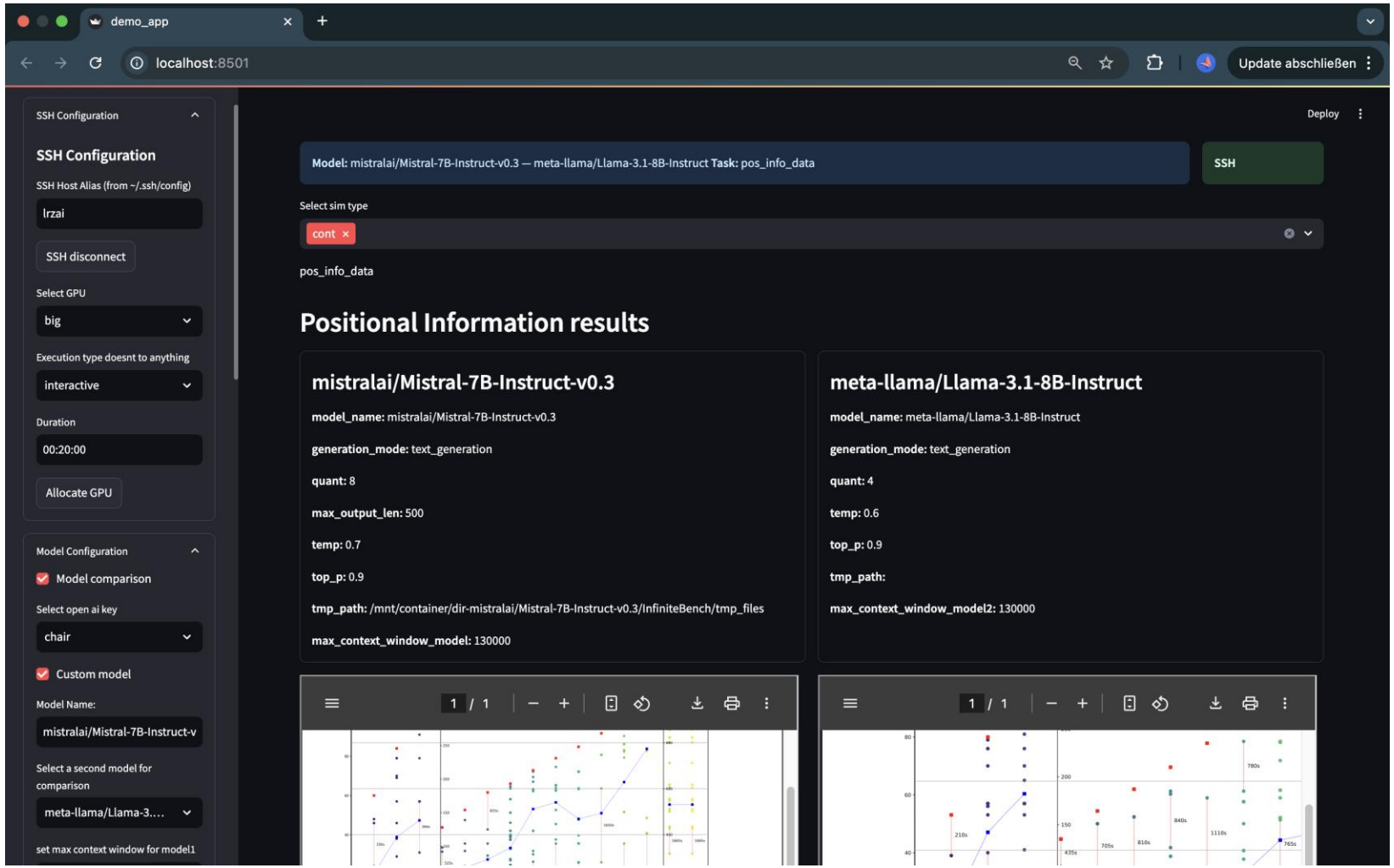
Results: Question Answering (RQ3)

- All models perform better on questions originating from earlier parts of the context window.
- Increasing the context window length leads to slightly degrading model similarity scores.



Command-r7b (InfiniteBench)

Demo App



Outline

1) Introduction
- Motivation
- Research Questions

2) Methodology

3) Datasets

4) Results

5) Discussion
- Contributions & Limitations

Contributions & Limitations

- **RQ1:** *How can we adequately test the quality of text summarization and question answering produced by LLMs? Does the quality of the generated summary or answer improve when more content from the article is provided to the model?* – **A pipeline** for evaluating the long context window performance of LLMs.

- **RQ2:** *What are the most effective techniques for extending the context window of LLMs?* – **Comprehensive performance overview** of contemporary LLMs comprising different techniques for extending the context window size.

- RQ3: *How do LLMs utilize the information within their context window during text summarization and question-answering tasks? Can LLMs distribute their attention uniformly across the entire document, or is their attention clustered towards specific sections?* – All tested models **do not distribute their attention uniformly across their context window.**

- The evaluation of the comparison of context window extension techniques is limited.
- A more precise evaluation could help determine results for RQ2.

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