Multilingual Evaluation of Long Context Retrieval and Reasoning

Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad,
Rhitabrat Pokharel, Russell Scheinberg

Portland State University, USA

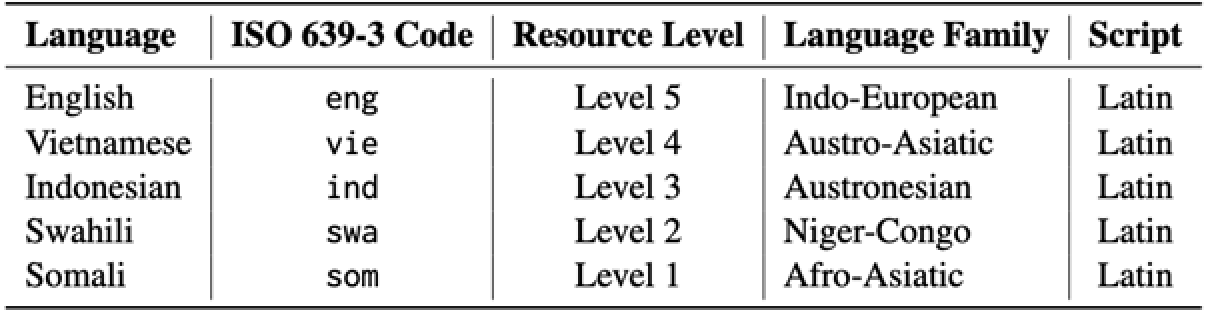
::- 4th Multilingual Representation Learning (MRL) Workshop, Nov. 16, 2024 -::

How well do models use long contexts across different languages?

[outline]

- Multilingual retrieval and reasoning task in long contexts (64k tokens)
- mLongRR: a new dataset for needle-in-a-haystack analysis across 5 languages

[languages]



Language	ISO 639-3 Code	Resource Level	Language Family	Script
English	eng	Level 5	Indo-European	Latin
Vietnamese	vie	Level 4	Austro-Asiatic	Latin
Indonesian	ind	Level 3	Austronesian	Latin
Swahili	swa	Level 2	Niger-Congo	Latin
Somali	som	Level 1	Afro-Asiatic	Latin

[retrieval and reasoning tasks]

Find "needles" in multilingual "haystacks"

Star dunes - or pyramid dunes – are named after their distinctive …

The special Doha number is 9121372.

In our dark laboratory, we see light from these sand grains …

Retrieval Task, single needle (n=1)

"The special {city} number is: {number}."

Reasoning Task, multiple needles (n≥2)

"What is the larger/largest magic number?"

"Which city has the larger magic number?"

[creating mLongRR dataset]

BBC news articles in multiple languages

Naturally occurring text

Recent data

[prompt]

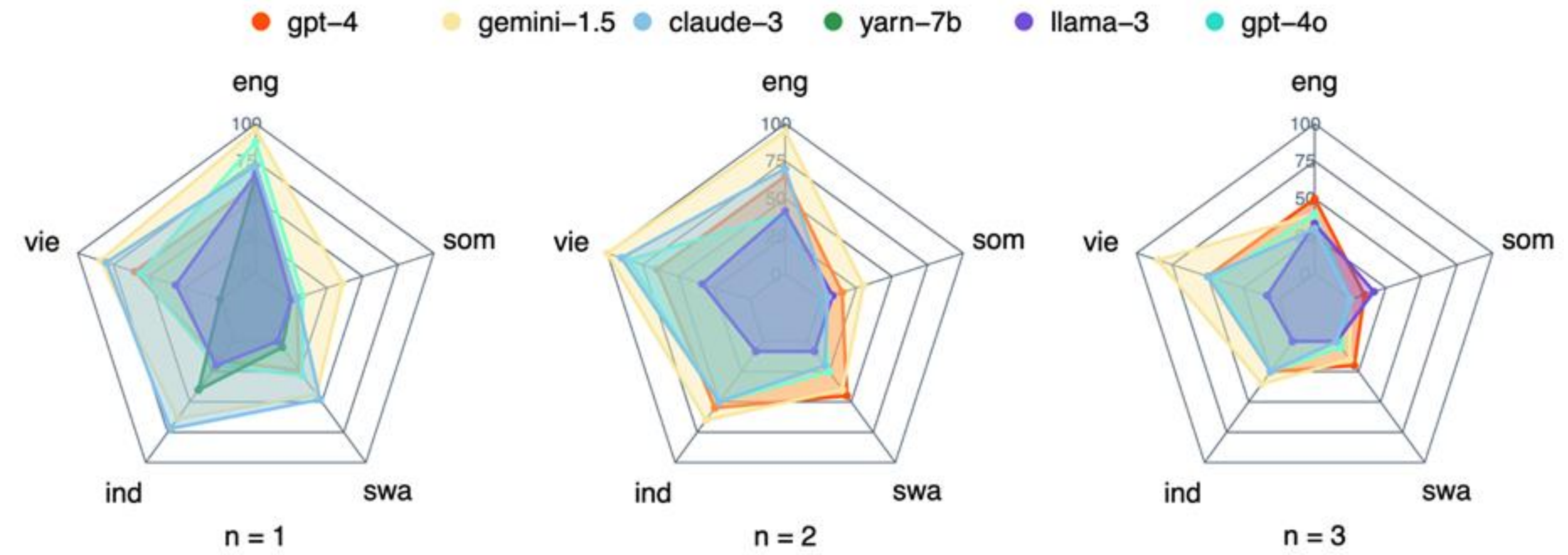
```
You are a helpful AI bot that answers questions for a user. Keep your response short and direct. The following is a set of context and a question that will relate to the context.


{context}


#QUESTION

What is the special magic number? Don't give information outside the document or repeat your findings. If the information is not available in the context respond UNANSWERABLE.
```

[experiments]

→ 6 LLMs: GPT-4, GPT-4o, Gemini-1.5, Claude-3, Yarn-7b, Llama-3

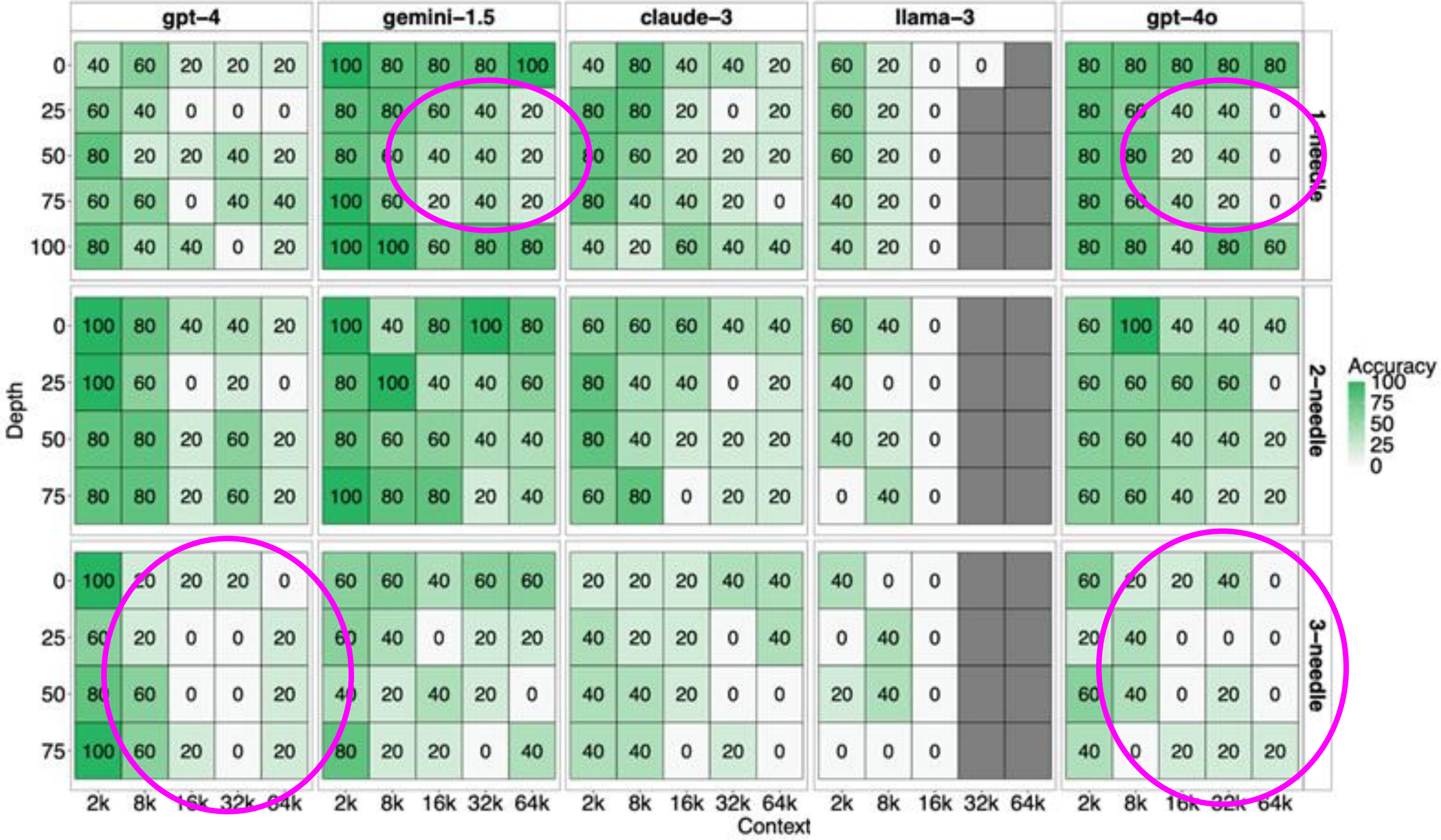
→ 5 context lengths: 2k, 8k, 16k, 32k, 64k

→ 5 needle depths: 0%, 25%, 50%, 75%, 100%



Performance drops significantly as task complexity increases (as few as n=3!)

Overall Gemini-1.5 pro yields best performance, followed by GPT-4o



1-needle

	gpt-4					gemini-1.5					claude-3					llama-3					gpt-4o				
Depth	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k
0	40	60	20	20	20	100	80	80	80	100	40	80	40	40	20	60	20	0	0		80	80	80	80	80
25	60	40	0	0	0	80	80	60	40	20	80	80	20	0	20	60	20	0			80	60	40	40	0
50	80	20	20	40	20	80	60	40	40	20	80	60	20	20	20	60	20	0			80	80	20	40	0
75	60	60	0	40	40	100	60	20	40	20	80	40	40	20	0	40	20	0			80	60	40	20	0
100	80	40	40	0	20	100	100	60	80	80	40	20	60	40	40	40	20	0			80	80	40	80	60

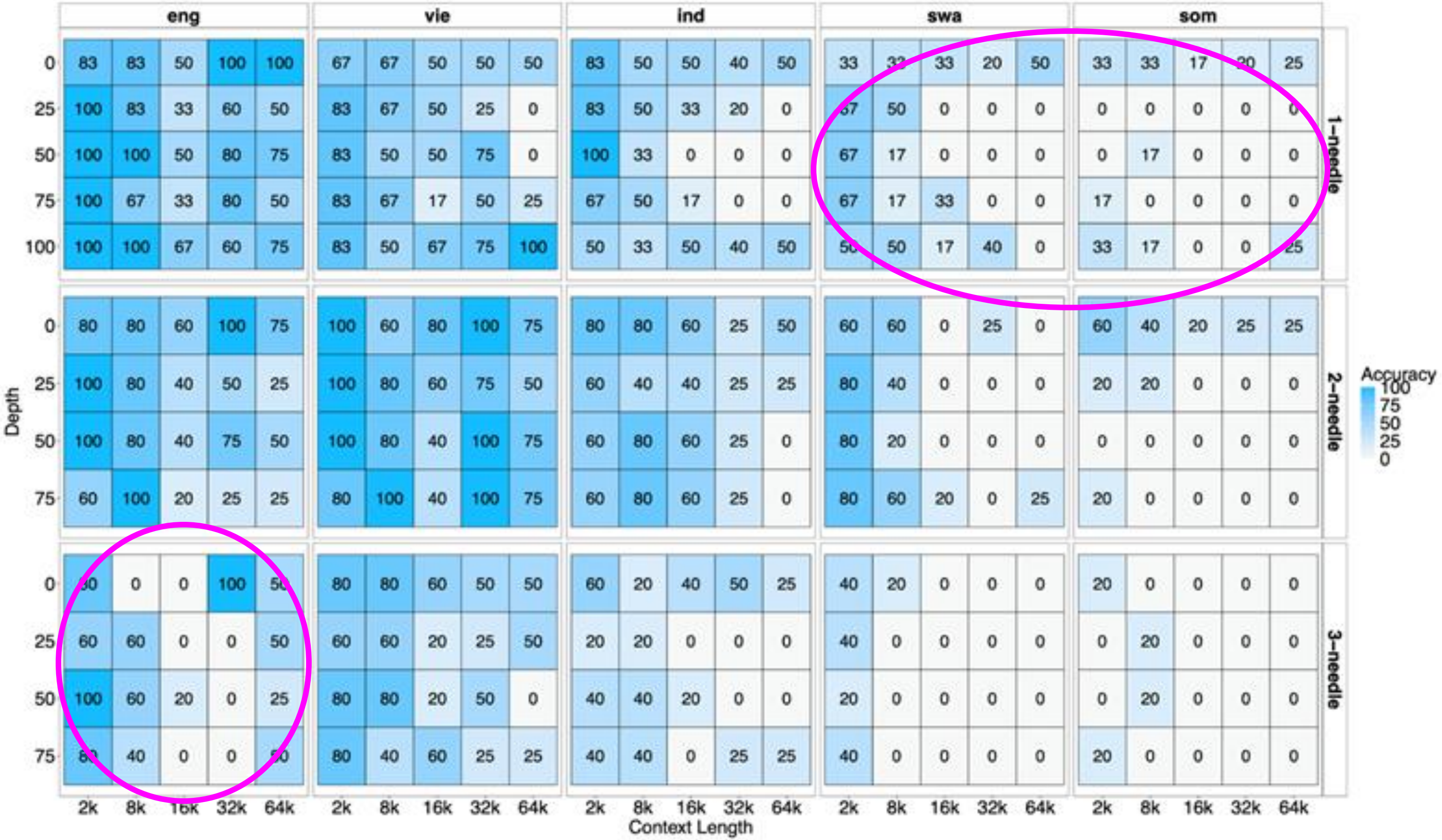
2-needle

	gpt-4					gemini-1.5					claude-3					llama-3					gpt-4o				
Depth	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k
0	100	80	40	40	20	100	40	80	100	80	60	60	60	40	40	60	40	0			60	100	40	40	40
25	100	60	0	20	0	80	100	40	40	60	80	40	40	0	20	40	0	0			60	60	60	60	0
50	80	80	20	60	20	80	60	60	40	40	80	40	20	20	20	40	20	0			60	60	40	40	20
75	80	80	20	60	20	100	80	80	20	40	60	80	0	20	20	0	40	0			60	60	40	20	20

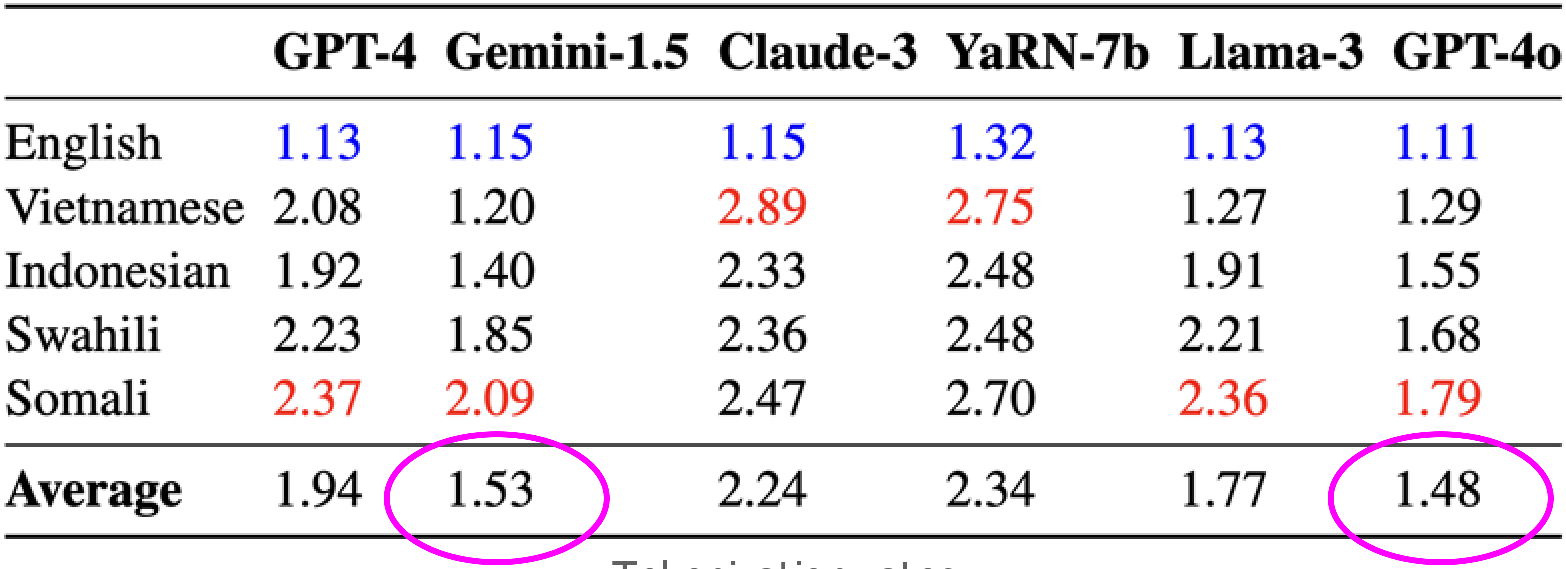
3-needle

	gpt-4					gemini-1.5					claude-3					llama-3					gpt-4o				
Depth	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k	2k	8k	16k	32k	64k
0	100	20	20	20	0	60	60	40	60	60	20	20	20	40	40	40	0	0			60	20	20	40	0
25	60	20	0	0	20	60	40	0	20	20	40	20	20	0	40	0	40	0			20	40	0	0	0
50	80	60	0	0	20	40	20	40	20	0	40	40	20	0	0	20	40	0			60	40	0	20	0
75	100	60	20	0	20	80	20	20	0	40	40	40	0	20	0	0	0	0			40	0	20	20	20

Performance drops with longer contexts, especially for needles in the middle



Performance drops as we get to lower resource languages



	GPT-4	Gemini-1.5	Claude-3	YaRN-7b	Llama-3	GPT-4o
English	1.13	1.15	1.15	1.32	1.13	1.11
Vietnamese	2.08	1.20	2.89	2.75	1.27	1.29
Indonesian	1.92	1.40	2.33	2.48	1.91	1.55
Swahili	2.23	1.85	2.36	2.48	2.21	1.68
Somali	2.37	2.09	2.47	2.70	2.36	1.79
Average	1.94	1.53	2.24	2.34	1.77	1.48

Tokenization rates

Lower tokenization rate results in better performance across languages and models

[conclusion]

- Performance drops with
- Longer contexts
- Needles in the middle
- Higher task complexity (English drops from ~100% to ~50% in n=3)
- Lower-resource languages

- Lower tokenization rates correlate with better performance across languages and models

[thank you]

ameeta@pdx.edu

PortNLP