ICLR

Needle Threading: Can LLMs Follow Threads Through Near-Million-Scale Haystacks?

Jonathan Roberts[1], Kai Han[2], Samuel Albanie

[1]University of Cambridge

[2]The University of Hong Kong

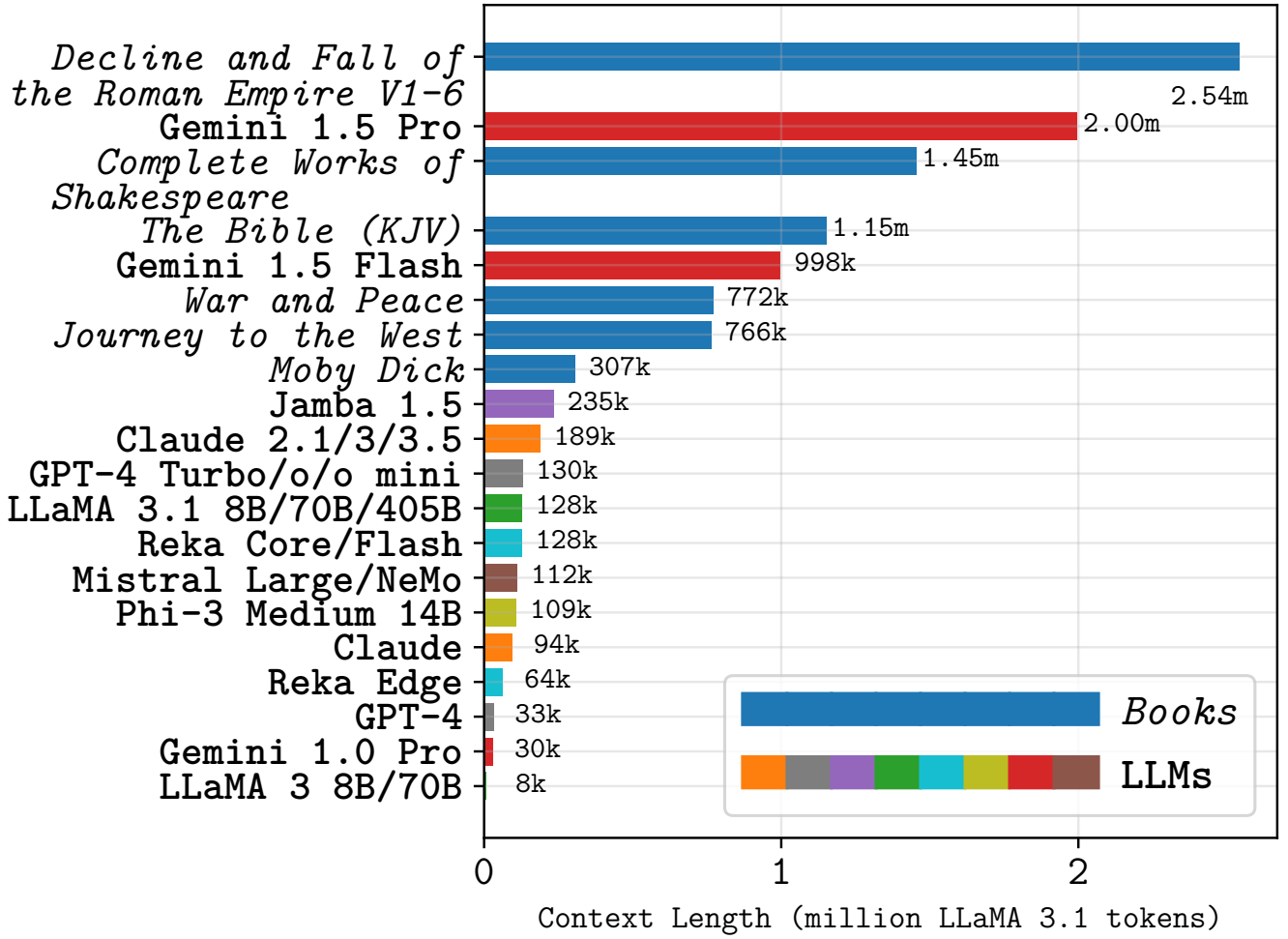
UNIVERSITY OF CAMBRIDGE

香 港 大 學

THE UNIVERSITY OF HONG KONG

Motivation

- LLMs possess remarkable capabilities in numerous domains
- Frontier models are being developed with longer context limits
- Longer context windows enable a broader range of real-world tasks
- The development of long context models has outpaced understanding



Long context evaluations

Needle(s) in a haystack:

Haystack diagram generated using GPT-4o

Limitations of prior long context evaluations:

- Performance saturation, limited context length, lack of granular takeaways

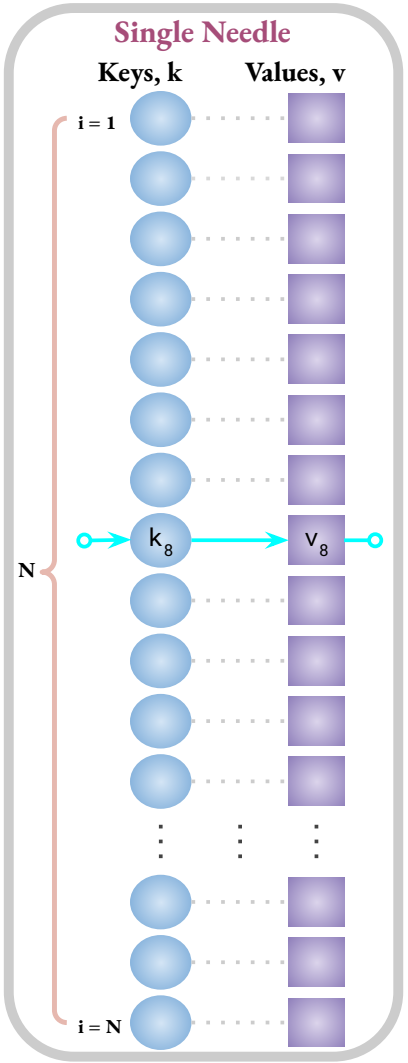
We focus on synthetic data:

- Avoid expensive curation and annotation, noise-free, fine-grained control of task parameters

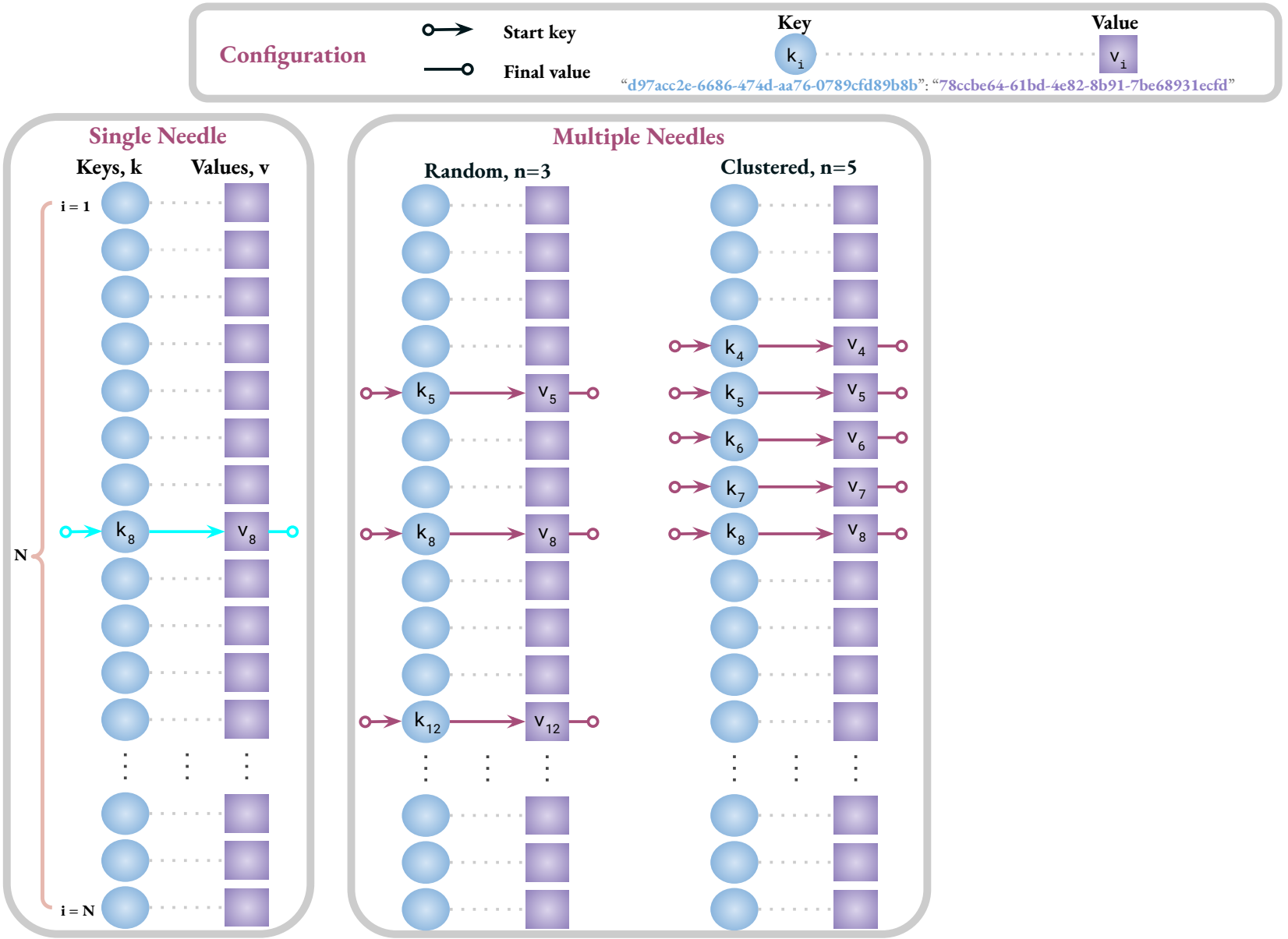
```
Haystack = {
    “9a159850-2f26-2bab-a114-4eefdeb0859f”: “5de8eca9-8fd4-80b8-bf16-bd4397034f54”,
    “d64b2470-8749-3be3-e6e8-11291f2dd06e”: “1f22fcdb-9001-05ab-91f1-e7914b66a4ea”,
    . . .,
    “bae328a1-44f3-7da1-d323-4bd9782beca1”: “1183e29c-db7a-dccf-6ce8-c0a462d9942c”,
    “5d88d112-e4ec-79a1-d038-8f1c58a240e4”: “ea8bf5c3-1ede-7de0-ba05-d8cd69393423”
    }
```

Needle threading

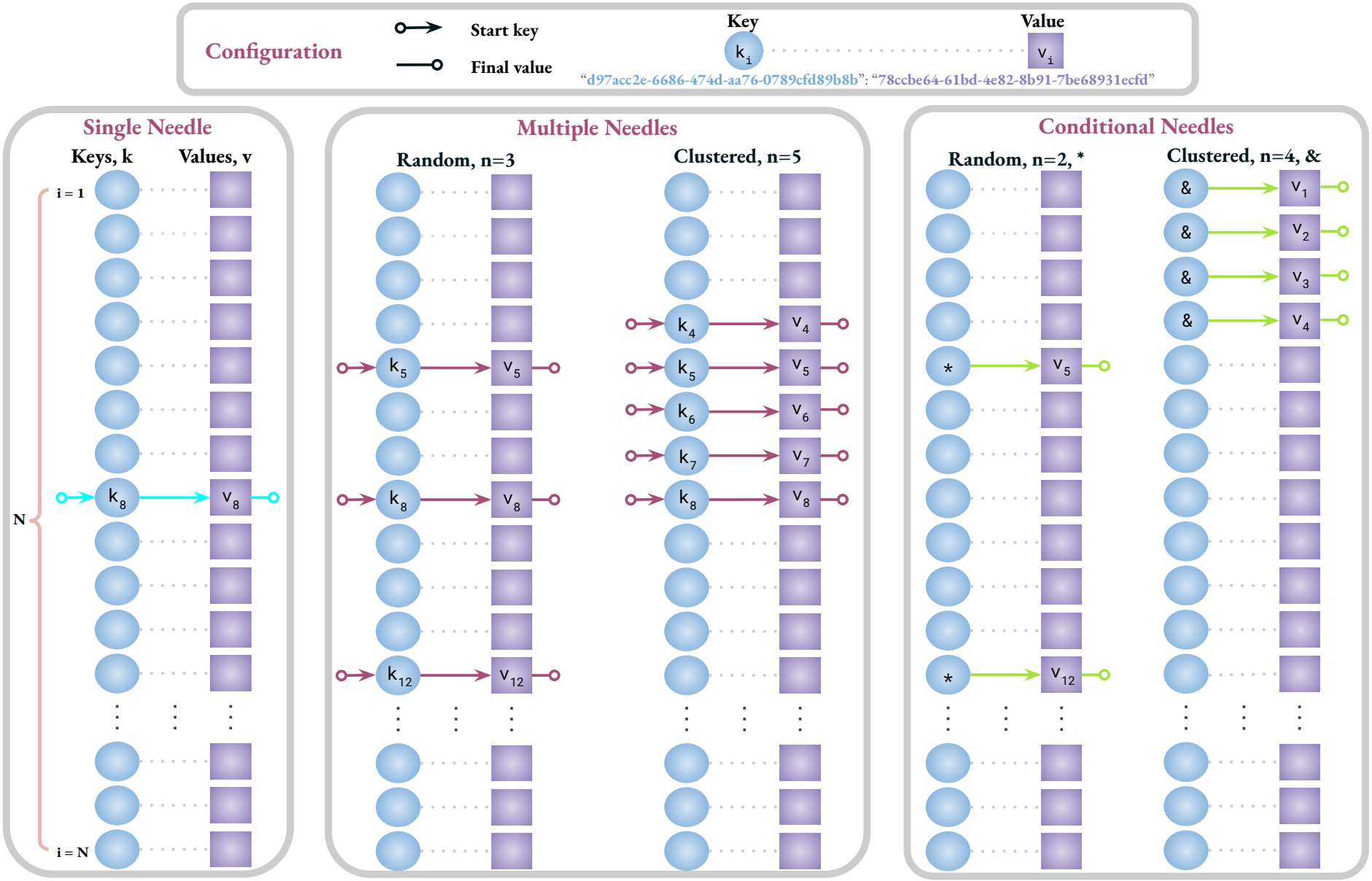
Needle threading



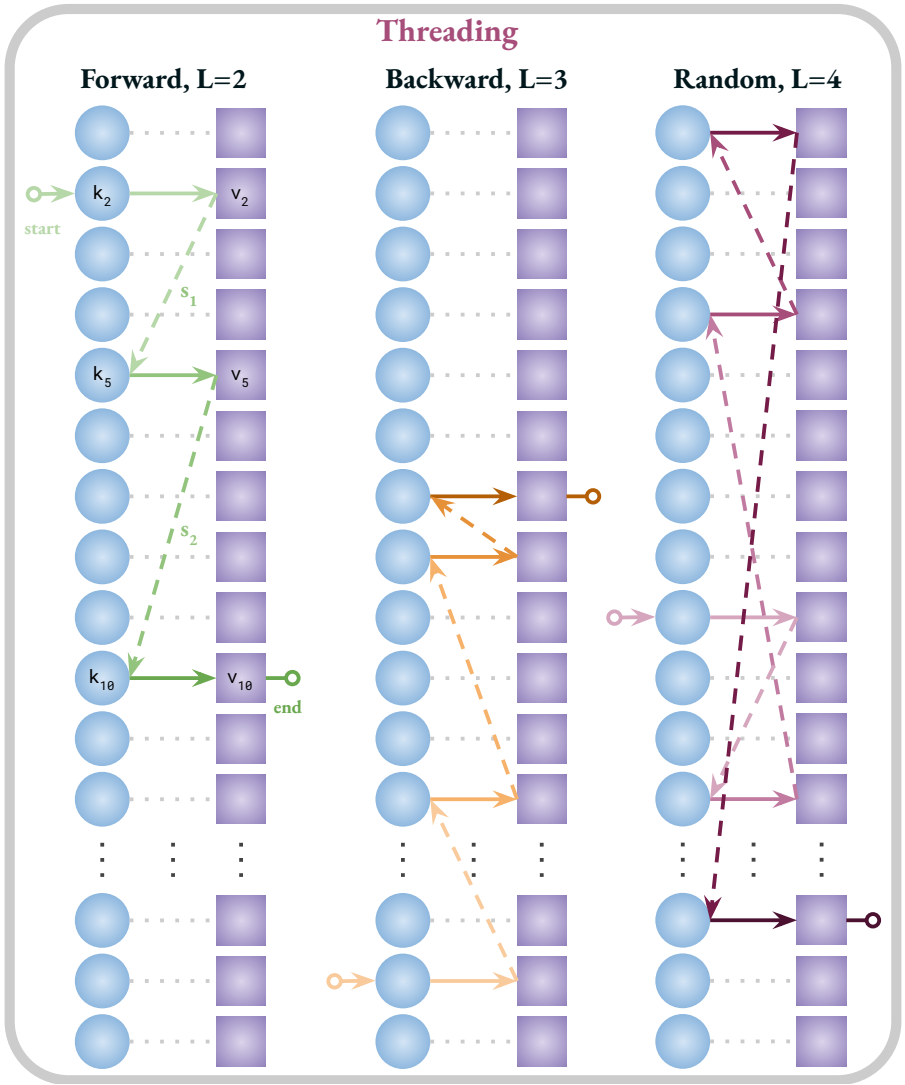
Needle threading



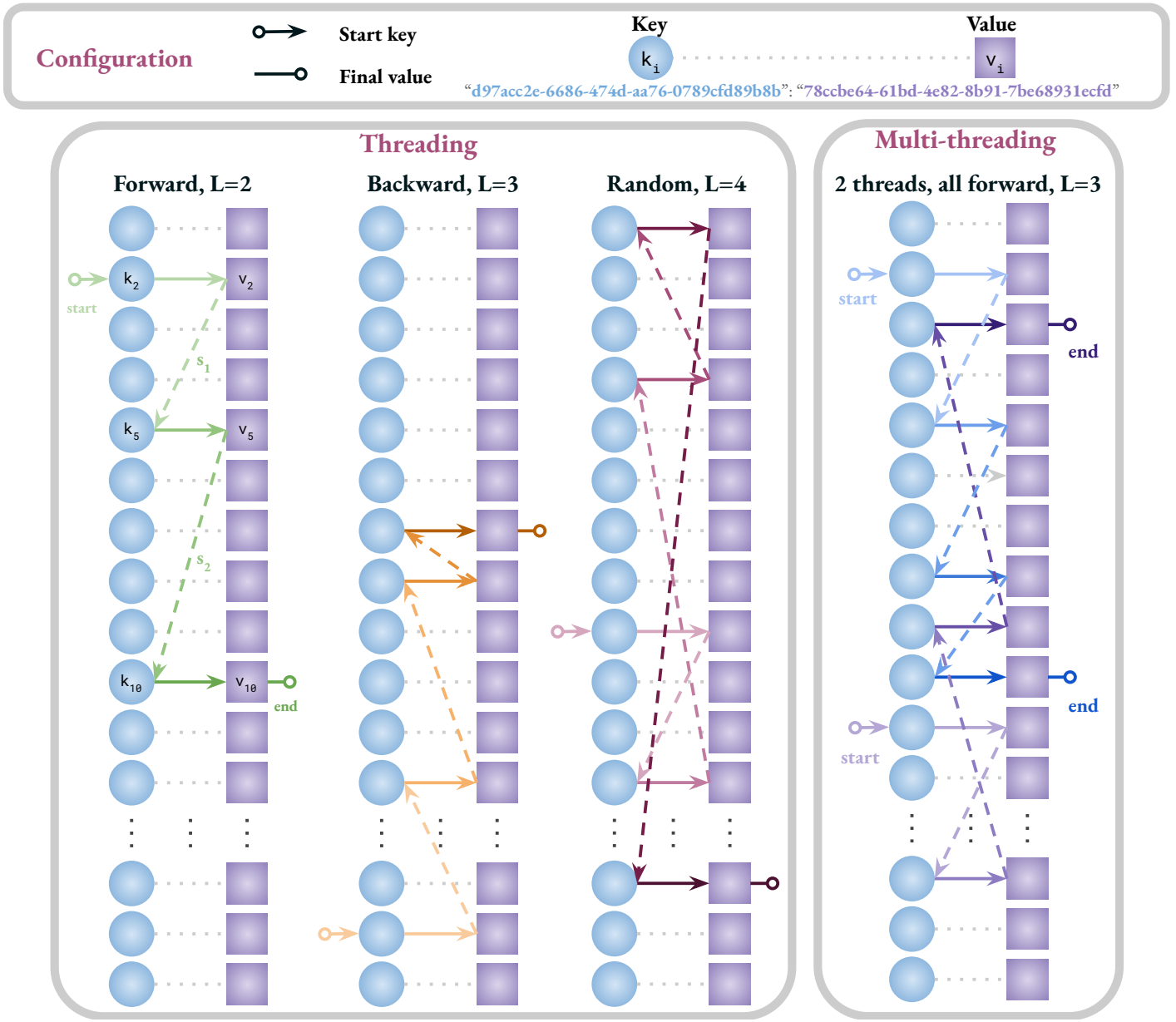
Needle threading



Needle threading



Needle threading

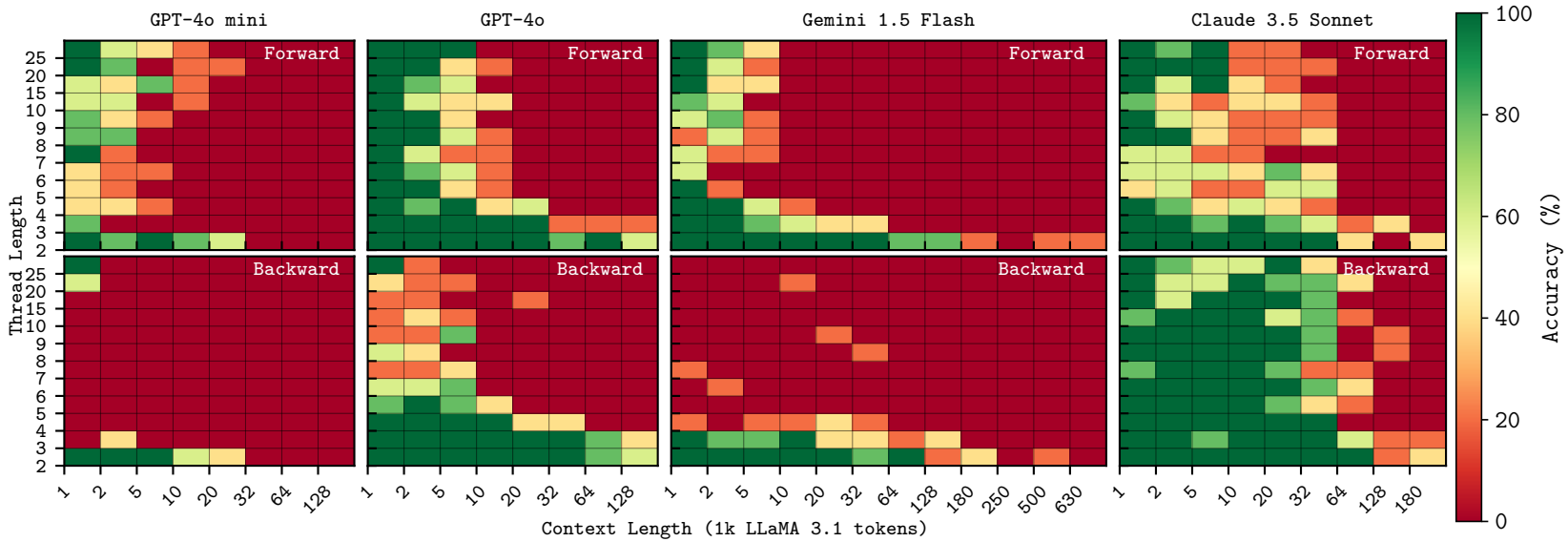


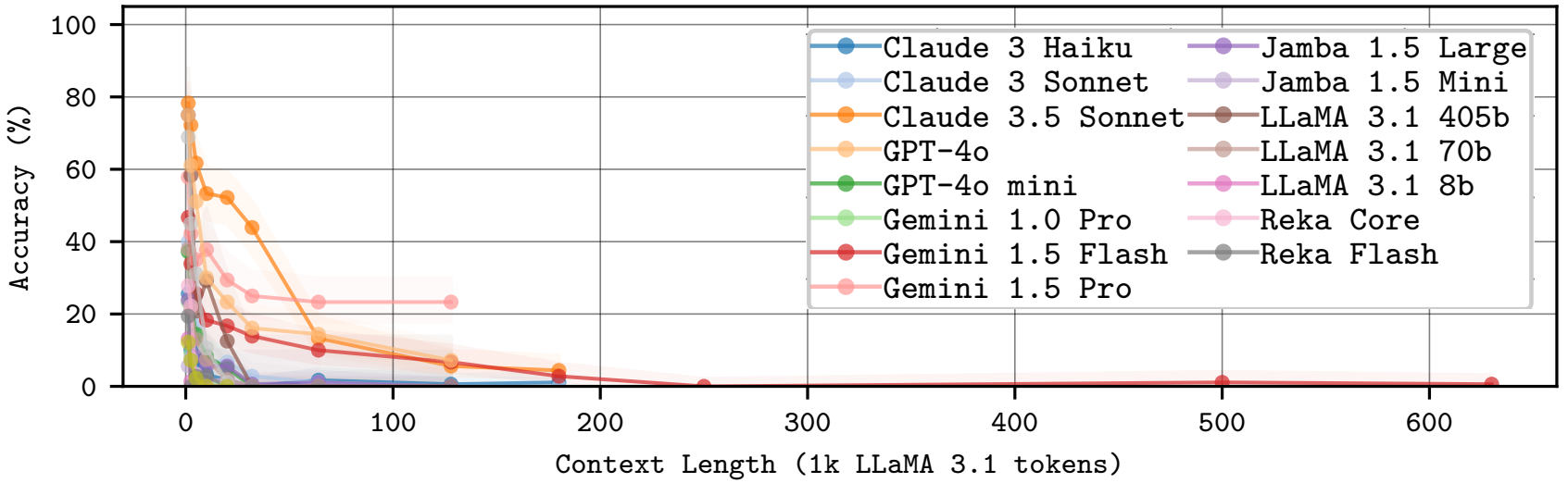
Single threading

Single threading

Findings:

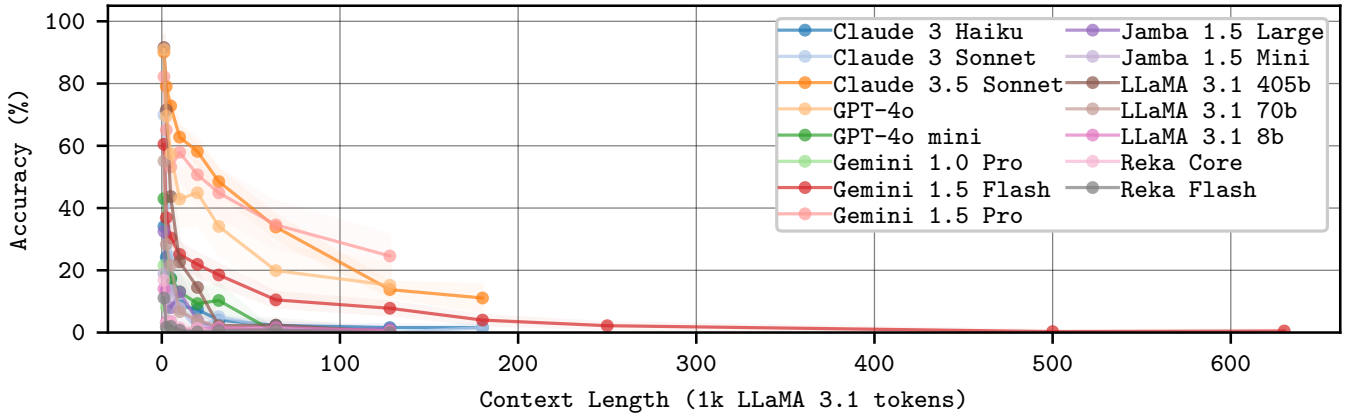
- Performance decreases with thread length
- Retrieval accuracy decreases on longer context length haystacks
- Most models prefer forward travelling threads





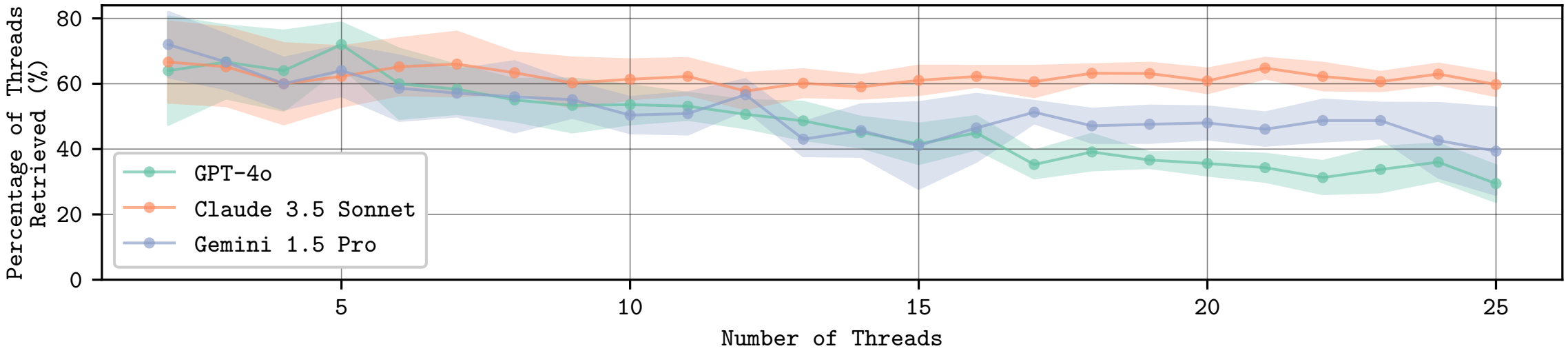
Multi-threading

Performance also decreases as the context length grows



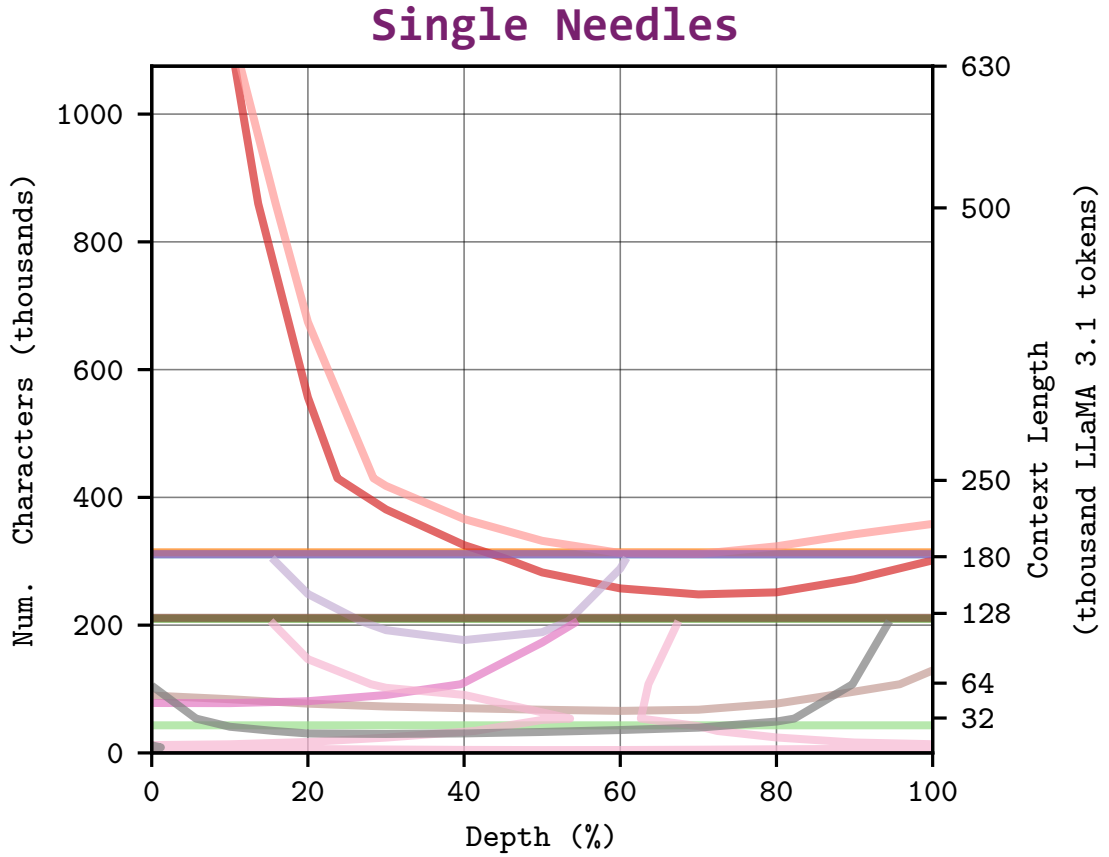
The number of threads to simultaneously retrieve does not significantly impact accuracy

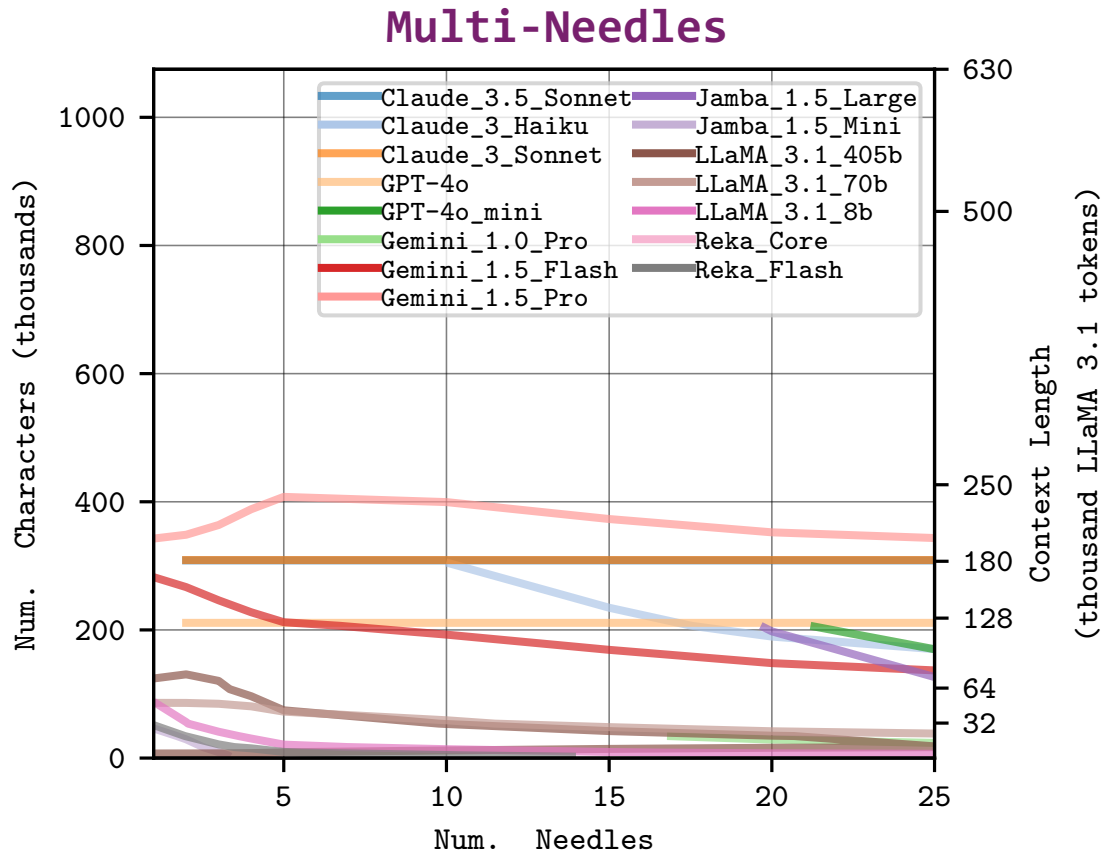
Frontier models are thread-safe



Effective Context Lengths

- Create a dense grid of points along the key experimental variables
- Interpolate average accuracy at each point
- Compute a contour corresponding to a threshold accuracy
  -> this represents an effective frontier, beyond which retrieval is unreliable





ICLR

Check out our paper for:
- More experiments
- Quantitative results
- Natural language ablations

Project page: https://needle-threading.github.io/
arXiv: https://arxiv.org/abs/2411.05000

UNIVERSITY OF CAMBRIDGE

香 港 大 學
THE UNIVERSITY OF HONG KONG

ICLR

Thanks!

Project page: https://needle-threading.github.io/
Code: https://github.com/jonathan-roberts1/needle-threading
Data: https://huggingface.co/datasets/jonathan-roberts1/needle-threading
arXiv: https://arxiv.org/abs/2411.05000

Jonathan Roberts
University of Cambridge

Kai Han
The University of Hong Kong

Samuel Albanie

UNIVERSITY OF CAMBRIDGE

香 港 大 學
THE UNIVERSITY OF HONG KONG