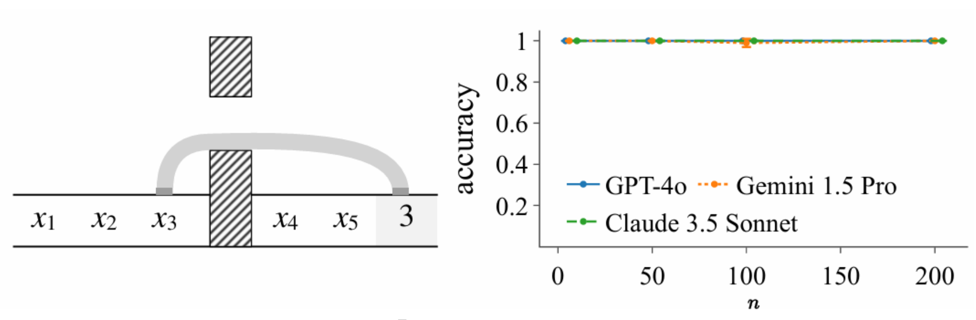
NeurIPS spotlight

Lost in Transmission: When and Why LLMs Fail to Reason Globally

Tobias Schnabel, Kiran Tomlinson, Adith Swaminathan, Jennifer Neville

Paper: aka.ms/bapo

low bandwidth → easy for LLMs

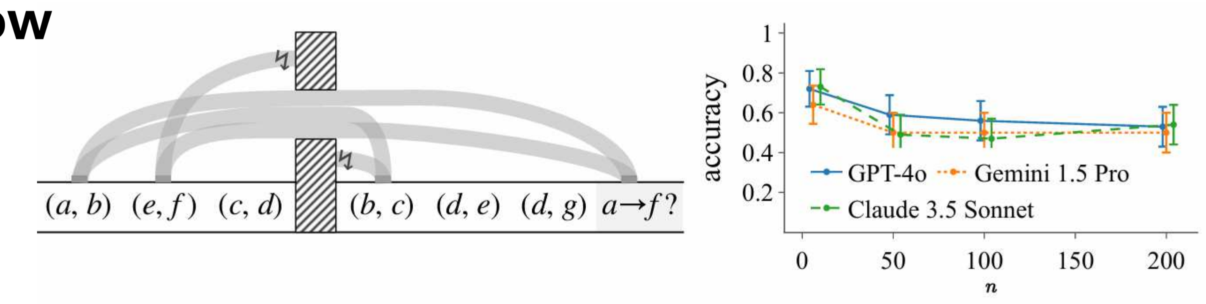


INDEX

We propose BAPO, a new theoretical framework that quantifies **the minimum amount of information about the input that needs to flow so that an LLM can solve a task.**

When the required flow for a task exceeds an **LLM's bandwidth, it fails at the task.**

high bandwidth → hard for LLMs



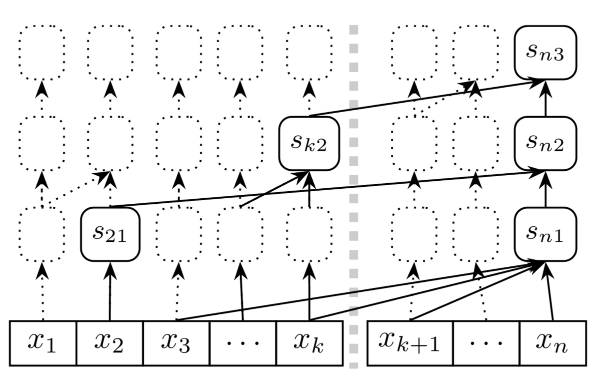
REACHABILITY

Information flow in LLMs

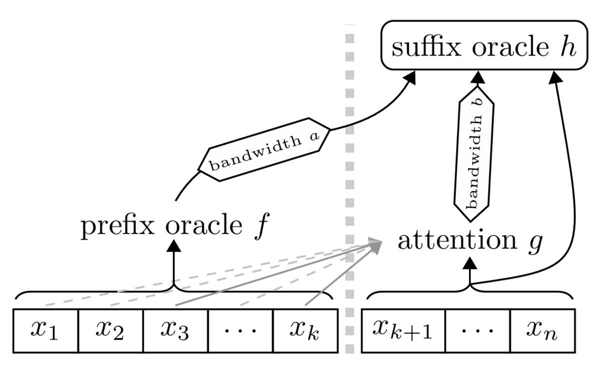
Hypothesis:

(Pre-trained) LLMs fail at cross-context information tracking

We test this with a new theoretical model:



A transformer with causally masked attention



Our model, the bounded attention prefix oracle (BAPO)

Takeaways

...for practioners

1. 💡 For tasks requiring lot of information flow, consider using tools or reasoning
🚀 Add BAPO-hard tasks to long-context LLM benchmarks

...for scaling LLMs

2. 💡 Scaling (data, model size, context window) is insufficient for BAPO-hard tasks
🚀 Explore new architectures

...for theorists

3. 💡 New way to reason about communication in causally-masked LLMs
🚀 Prove new bandwidth bounds, explore BAPO variants

...for improving reasoning

4. 💡 CoT can lower bandwidth, but may need many tokens in practice
🚀 Design/learn efficient CoT decompositions

💡 = Insight 🚀 = Future work

Theory

Bounded Attention Prefix Oracle (BAPO)

(a, b)-BAPO must solve problem given:

1. Arbitrary split of input into prefix & suffix
2. Output of prefix oracle f (a bits)
3. Attended tokens selected by g (b tokens)
4. Access to suffix, but not prefix

We bound the bandwidth requirements of problems:

	Problem	Lower bound	Upper bound
BAPO-easy	INDEX (Thm. 1)		$(0,1)$
	EQUALITY (Thm. 1)		$(1,1)$
	DISJOINTNESS (Thm. 1)		$(1,1)$
	MATCH2$_n$ (Thm. 4)		$(0,1)$
BAPO-hard	REACHABILITY (Thm. 2)	$(o(m^{1/c}\log m), o(m^{1-2/c}))$	trivial
	MAJORITY (Thm. 3)	$(o(\log n), o(n^{1-\epsilon}))$	$(\lceil \log_2 n \rceil, 0)$
	MATCH3$_n$ (Thm. 4)	$(o(n/b(n)), b(n))$	trivial
BAPO-Σ-hard	UNIQUE (Thm. 5)	$(o(\lvert\Sigma\rvert/b(\lvert\Sigma\rvert)), b(\lvert\Sigma\rvert))$	$(2\lvert\Sigma\rvert, 0)$
	SETDIFF (Thm. 6)	$(o(\lvert\Sigma\rvert/b(\lvert\Sigma\rvert)), b(\lvert\Sigma\rvert))$	$(\lvert\Sigma\rvert, 0)$

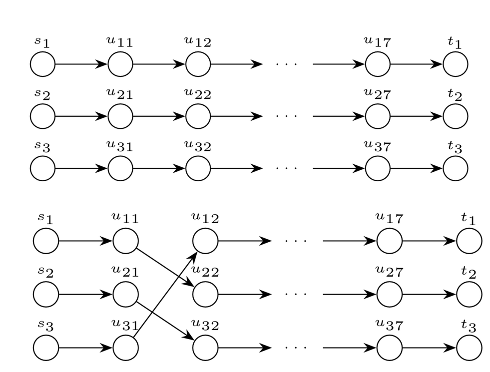
Problems can be:

- **BAPO-easy:** solvable with constant a, b (w.r.t. n)
- **BAPO-hard:** impossible with constant a, b
- **BAPO-Σ-hard:** impossible with constant a, b (w.r.t. # tokens)

Reachability proof sketch

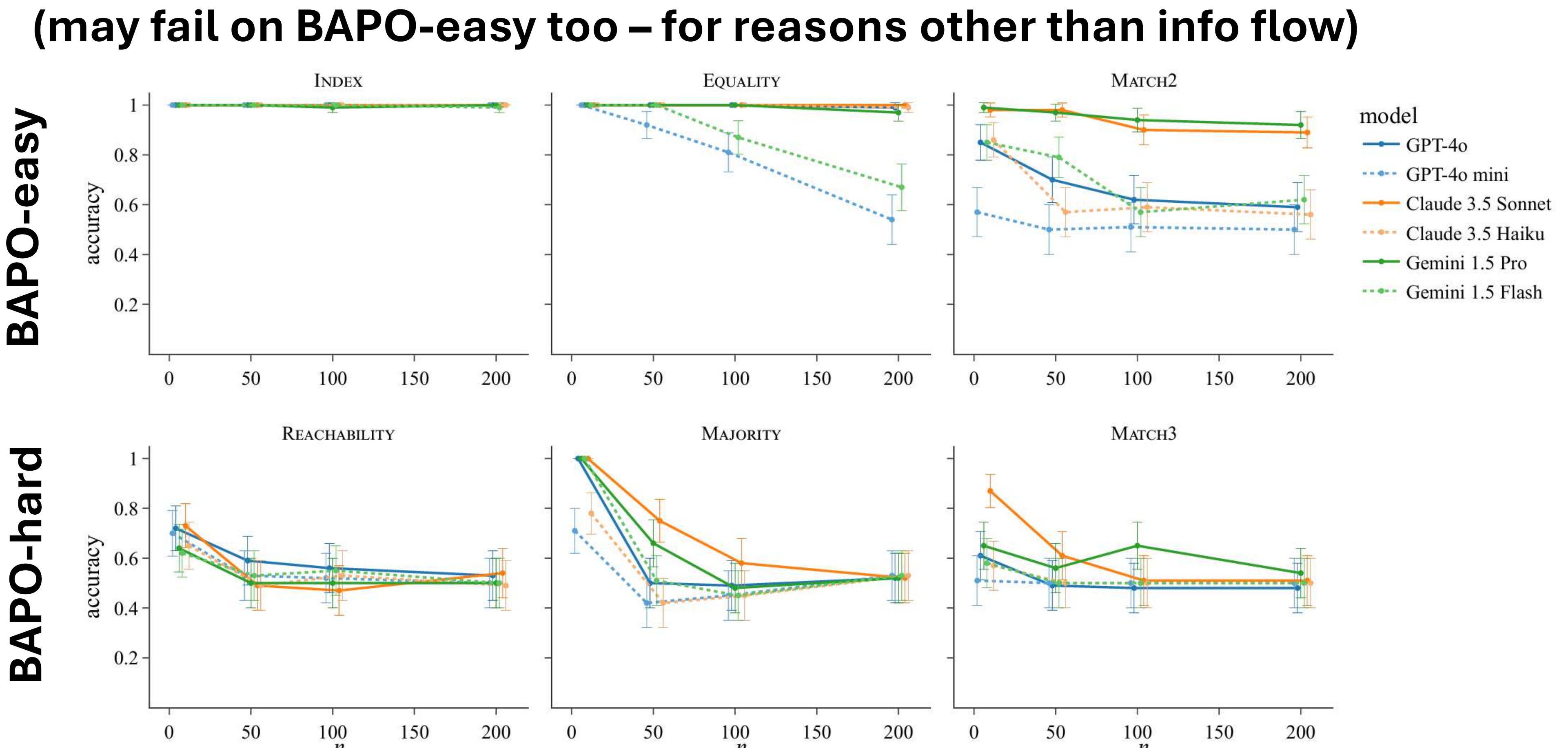
Prefixes: disjoint path graphs (one layer permuted); suffixes: s-t pairs

Pigeonhole → f collision; careful prefix design → g collision; mistake!

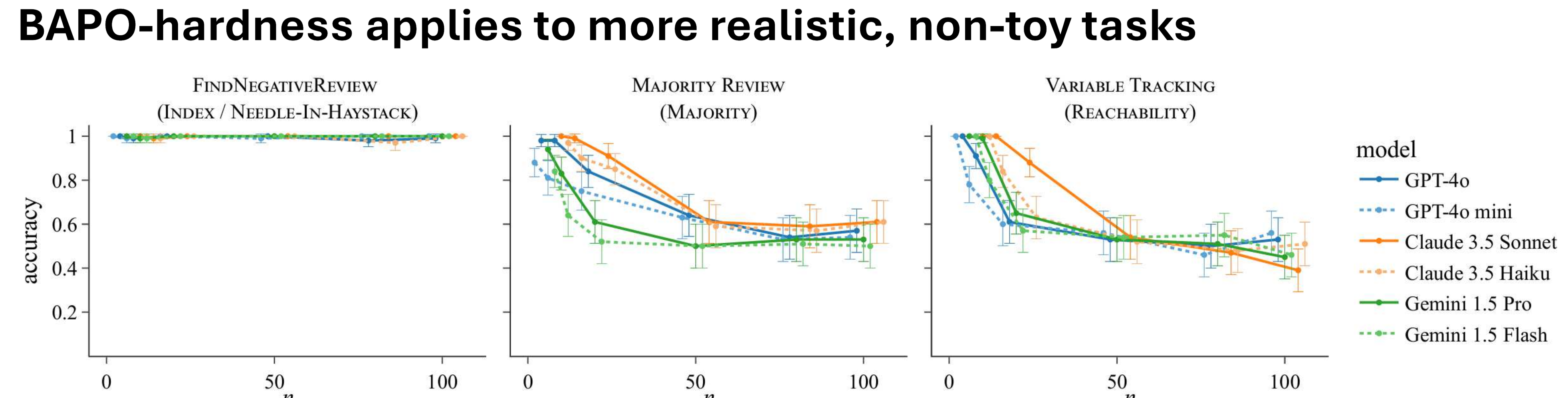


Experiments

LLMs can succeed on BAPO-easy tasks, but fail on BAPO-hard tasks (may fail on BAPO-easy too – for reasons other than info flow)



BAPO-hardness applies to more realistic, non-toy tasks



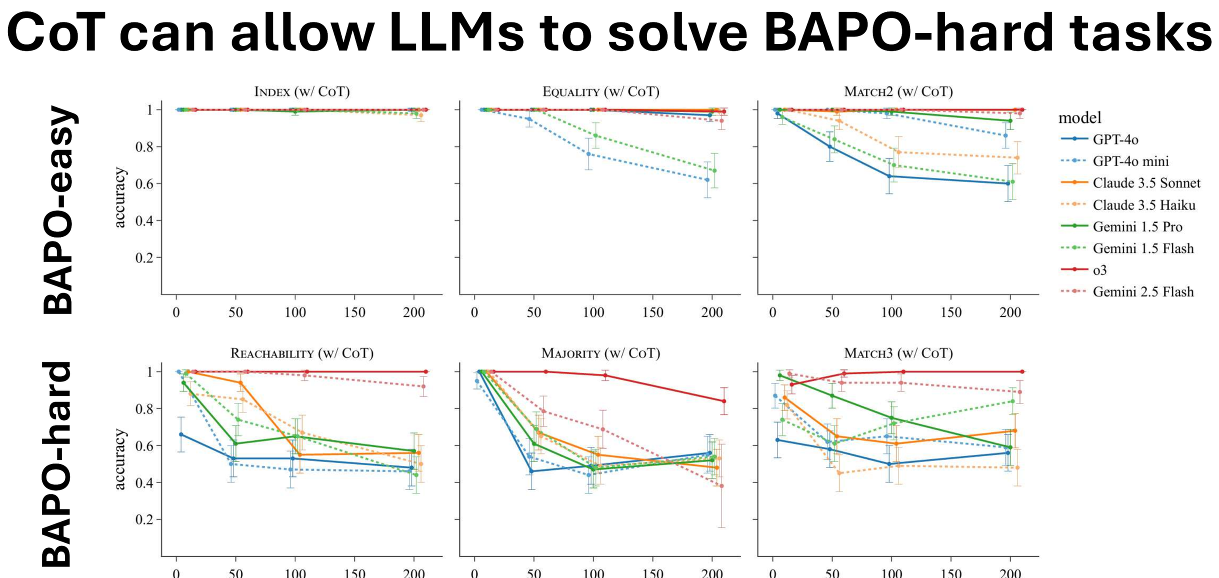
BAPO-easy **BAPO-hard**

BAPO model with chain of thought

Theorem (informal)

Any BAPO-hard problem can be broken down into a (potentially long) sequence of BAPO-easy chain-of-thought steps!

CoT can allow LLMs to solve BAPO-hard tasks



...but may use an impractical # of reasoning tokens