Large Language and Reasoning Models are Shallow Disjunctive Reasoners

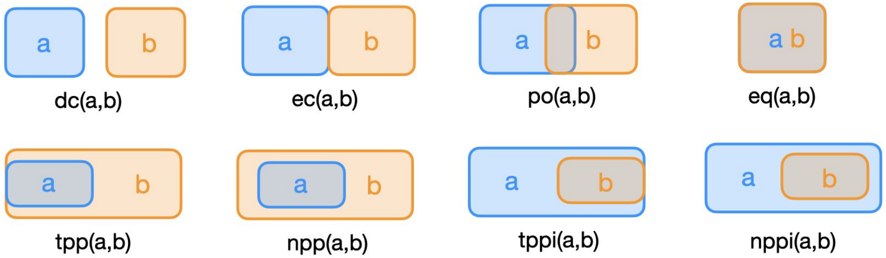
Irtaza Khalid, Amir Masoud Nourollah, Steven Schockaert

Sparks of AGI vs Embers of Autoregression

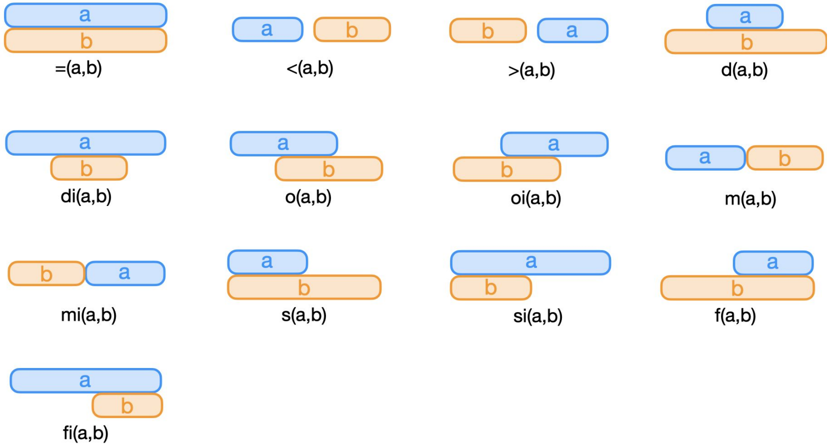
- Can Large Language Models (LLMs) and Large Reasoning Models (LRMs) reason or are they shallow pattern-matching on internet-scale data?
- It is difficult to adjudicate either side due to
  - the absence of proper test data that is significantly out-of-distribution
  - Training on test data / data memorization for MMLU/GSM8k (Zhang et. al. 2023 and Oren et. al. 2024)
  - Recurring issues like
    - The reversal curse (Berglund et. al. 2024) A -> B is solved but not B -> A
    - Over-reliance on co-occurrence statistics (Kang and Choi et. al. 2023)
- We benchmark LLMs and LRMs on the STaR benchmark (Khalid et. al. 2025) for the problem of disjunctive reasoning, whilst circumventing both issues above with previous test data.

STaR: Spatio-Temporal Reasoning benchmark

8 qualitative spatial relations
(Randell et. al. 1992)

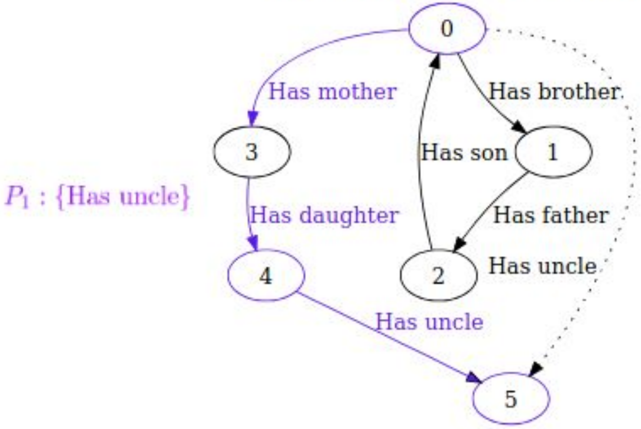


13 qualitative temporal relations
(Allen et. al. 1983)



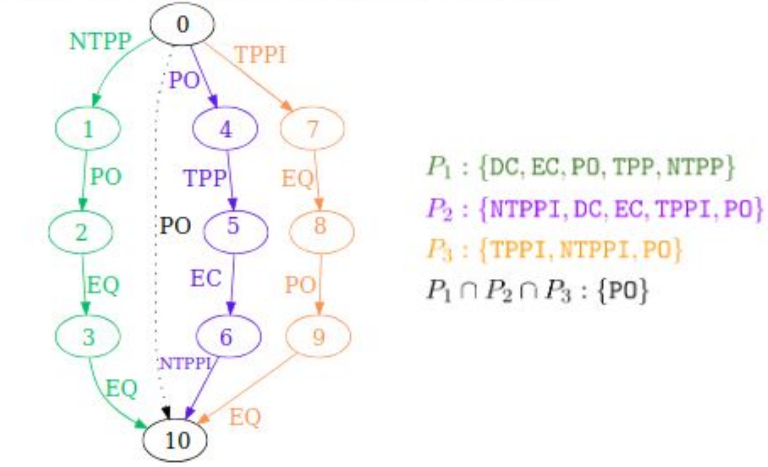
STaR: Spatio-Temporal Reasoning benchmark

Single-path reasoning



$$r_3(X,Z) \leftarrow r_1(X,Y) \wedge r_2(Y,Z)$$

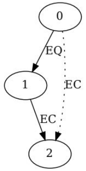
Multi-path disjunctive reasoning


0
NTPP
TPPI
PO
1
4
7
PO
TPP
EQ
PO
2
5
8
EQ
EC
PO
3
6
9
NTPPI
EQ
EQ
10
P_1 : {DC, EC, PO, TPP, NTPP}
P_2 : {NTPPI, DC, EC, TPPI, PO}
P_3 : {TPPI, NTPPI, PO}
$P_1 \cap P_2 \cap P_3$: {PO}

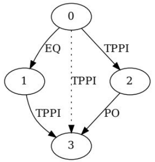
$$s_1(X,Z) \vee \ldots \vee s_k(X,Z) \leftarrow r_1(X,Y) \wedge r_2(Y,Z)$$
$$(s_1 \vee \cdots \vee s_k)_{\text{path}_1} \wedge \cdots \wedge (s_1 \vee \cdots \vee s_k)_{\text{path}_b}$$

Can LLMs and LRMs systematically generalize on STaR?

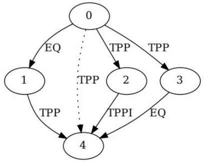
Evaluation is systematic: train on small instances and tested on increasingly larger instances in terms of number of paths, *b,* and path length from source to sink, *k*



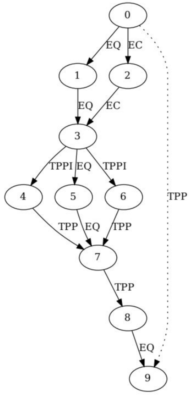
(a) $k = 2, b = 1$



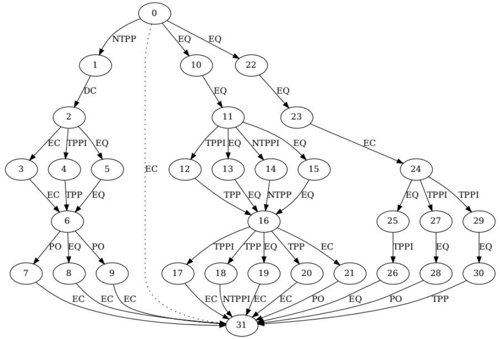
(b) $k = 2, b = 2$



(c) $k = 2, b = 3$



(g) $k = 6, b = 1$



(h) $k = 6, b = 3$

Train on:
k=2,3,4,
b=1,2,3

Test on
k=2,...,10,
b=1,...,4

Input Representation

Instruction (*Q*): You are a helpful assistant. **Just answer the question as a single integer.** Given a consistent graph with edges comprising the 8 base relations, predict the label of the target edge. More specifically, Given a data row delimited by a comma with the following columns: `graph_edge_index`, `edge_labels`, `query_edge`, predict the label of the `query_edge` as one of the 8 base relations as a power of 2 as defined above.

Composition Table (*T*): The following are the base elements of RCC-8: DC = 1 EC = 2 PO = 4 TPP = 8 ...

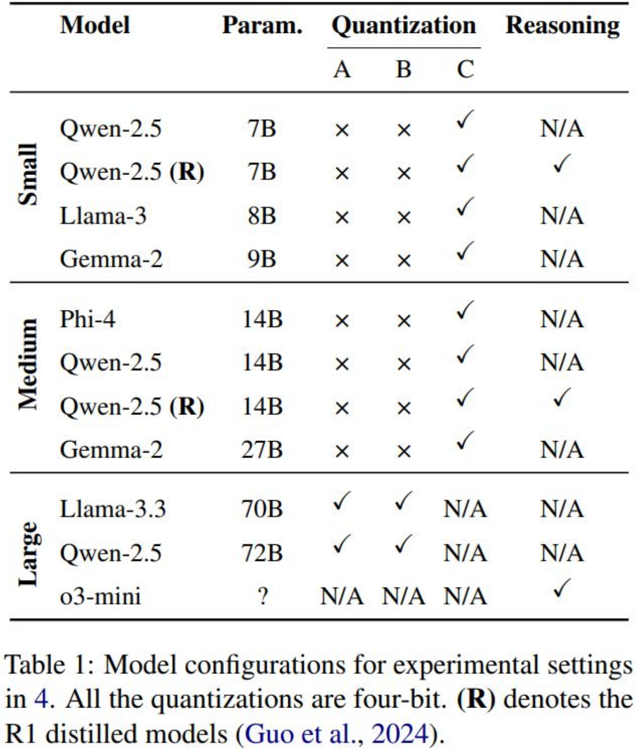
Graph Edge Index (E_i): *"[(0, 1), (1, 2)]"*

Edge labels (L_i): *"['EC' 'NTPPI']"*

Query Edge ((0, n_i)): *"(0, 2)"*

1

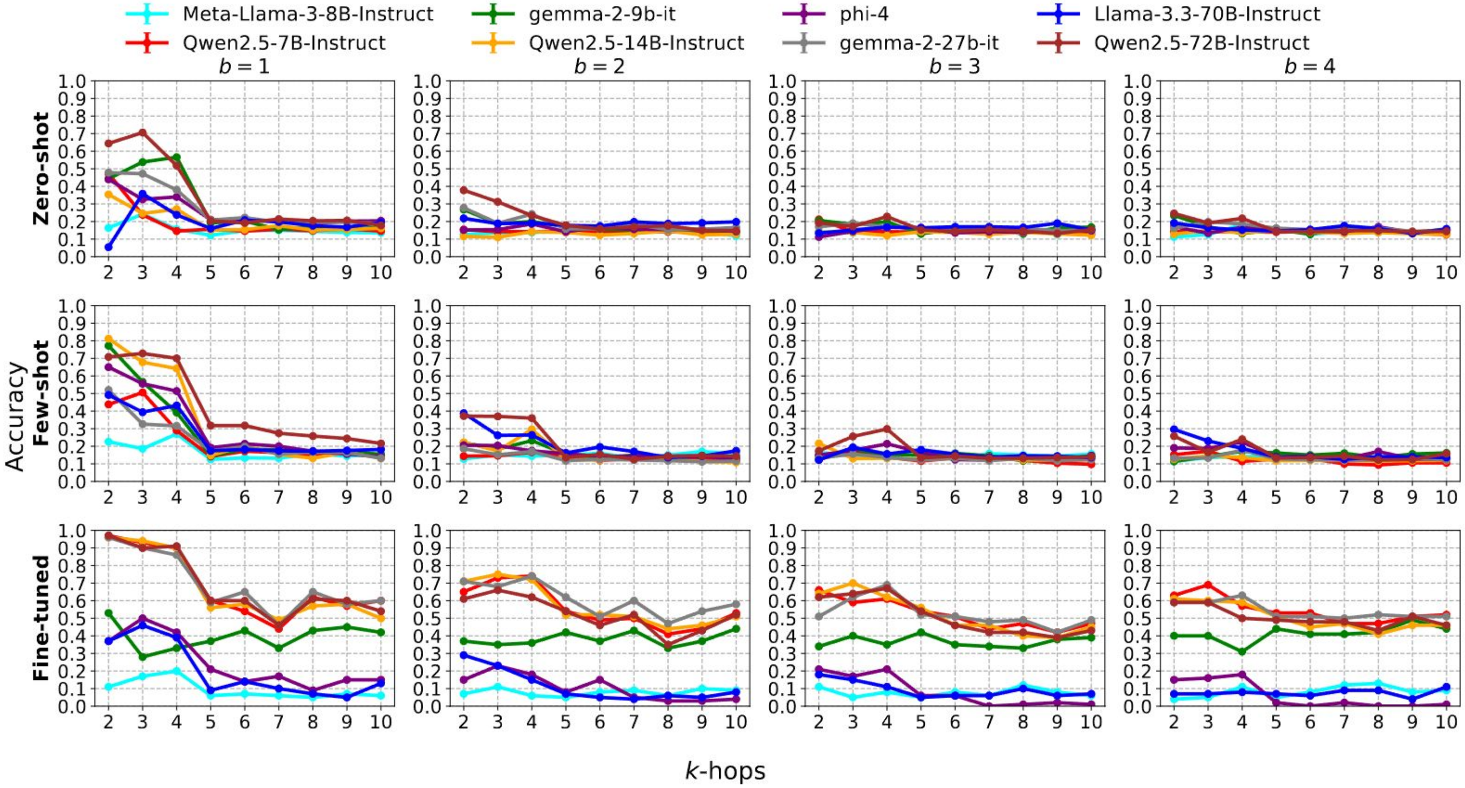
Model configurations



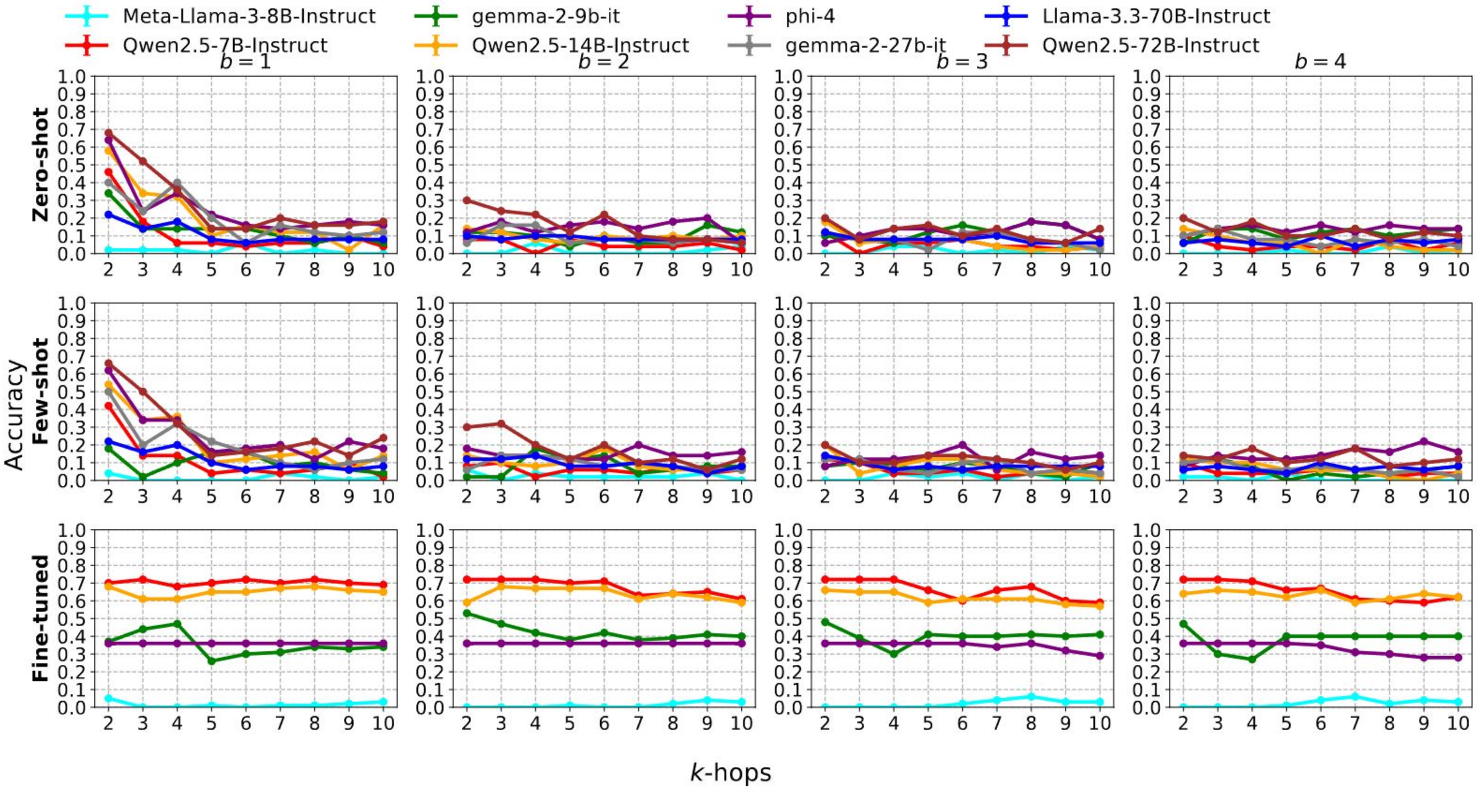
	Model	Param.	Quantization			Reasoning
			A	B	C	
Small	Qwen-2.5	7B	×	×	✓	N/A
	Qwen-2.5 **(R)**	7B	×	×	✓	✓
	Llama-3	8B	×	×	✓	N/A
	Gemma-2	9B	×	×	✓	N/A
Medium	Phi-4	14B	×	×	✓	N/A
	Qwen-2.5	14B	×	×	✓	N/A
	Qwen-2.5 **(R)**	14B	×	×	✓	✓
	Gemma-2	27B	×	×	✓	N/A
Large	Llama-3.3	70B	✓	✓	N/A	N/A
	Qwen-2.5	72B	✓	✓	N/A	N/A
	o3-mini	?	N/A	N/A	N/A	✓

Table 1: Model configurations for experimental settings in 4. All the quantizations are four-bit. **(R)** denotes the R1 distilled models (Guo et al., 2024).

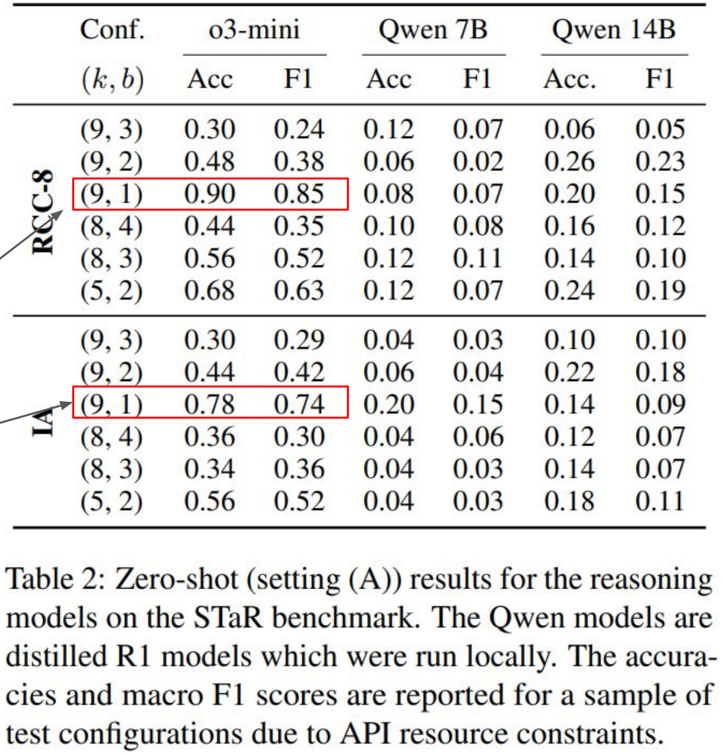
LLM Results (RCC-8)



LLM Results (Interval Algebra)



LRM results on STaR

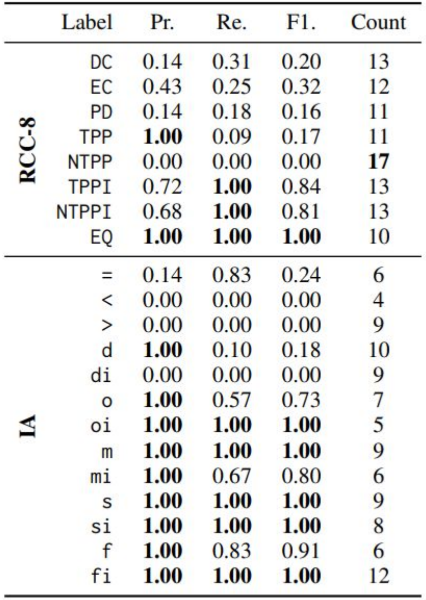


	Conf.	o3-mini		Qwen 7B		Qwen 14B	
	(k, b)	Acc	F1	Acc	F1	Acc.	F1
RCC-8	(9, 3)	0.30	0.24	0.12	0.07	0.06	0.05
	(9, 2)	0.48	0.38	0.06	0.02	0.26	0.23
	(9, 1)	0.90	0.85	0.08	0.07	0.20	0.15
	(8, 4)	0.44	0.35	0.10	0.08	0.16	0.12
	(8, 3)	0.56	0.52	0.12	0.11	0.14	0.10
	(5, 2)	0.68	0.63	0.12	0.07	0.24	0.19
IA	(9, 3)	0.30	0.29	0.04	0.03	0.10	0.10
	(9, 2)	0.44	0.42	0.06	0.04	0.22	0.18
	(9, 1)	0.78	0.74	0.20	0.15	0.14	0.09
	(8, 4)	0.36	0.30	0.04	0.06	0.12	0.07
	(8, 3)	0.34	0.36	0.04	0.03	0.14	0.07
	(5, 2)	0.56	0.52	0.04	0.03	0.18	0.11

Table 2: Zero-shot (setting (A)) results for the reasoning models on the STaR benchmark. The Qwen models are distilled R1 models which were run locally. The accuracies and macro F1 scores are reported for a sample of test configurations due to API resource constraints.

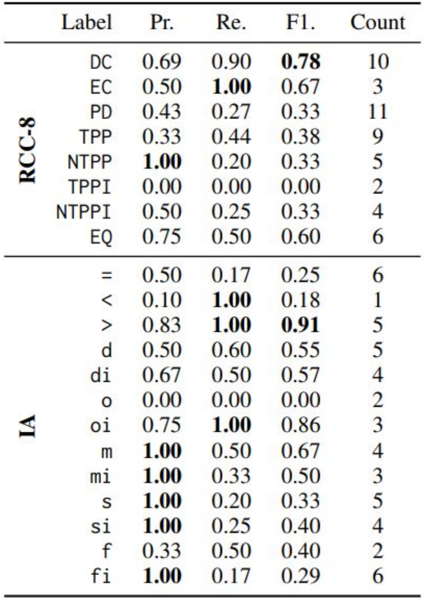
Best at single path reasoning

Breakdown of LRM results hints at heuristic exploitation



	Label	Pr.	Re.	F1.	Count
RCC-8	DC	0.14	0.31	0.20	13
	EC	0.43	0.25	0.32	12
	PD	0.14	0.18	0.16	11
	TPP	**1.00**	0.09	0.17	11
	NTPP	0.00	0.00	0.00	**17**
	TPPI	0.72	**1.00**	0.84	13
	NTPPI	0.68	**1.00**	0.81	13
	EQ	**1.00**	**1.00**	**1.00**	10
IA	=	0.14	0.83	0.24	6
	<	0.00	0.00	0.00	4
	>	0.00	0.00	0.00	9
	d	**1.00**	0.10	0.18	10
	di	0.00	0.00	0.00	9
	o	**1.00**	0.57	0.73	7
	oi	**1.00**	**1.00**	**1.00**	5
	m	**1.00**	**1.00**	**1.00**	9
	mi	**1.00**	0.67	0.80	6
	s	**1.00**	**1.00**	**1.00**	9
	si	**1.00**	**1.00**	**1.00**	8
	f	**1.00**	0.83	0.91	6
	fi	**1.00**	**1.00**	**1.00**	12

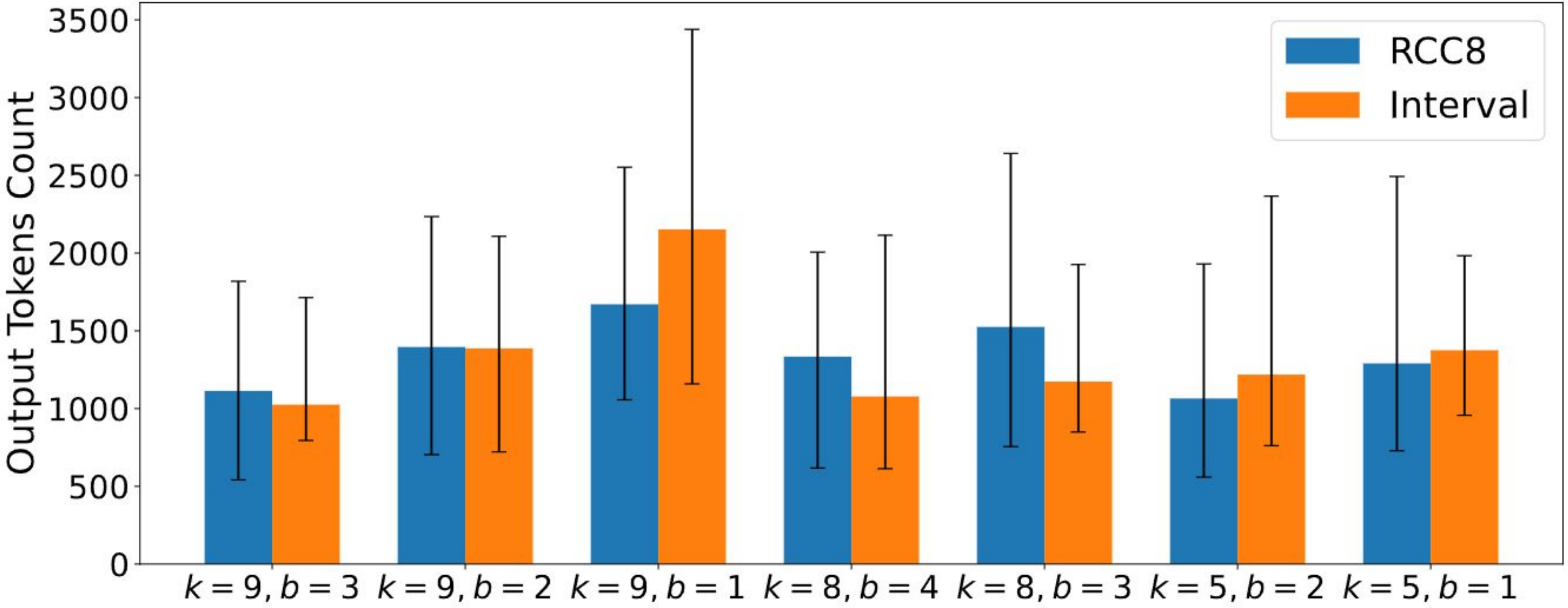
Table 3: Fine-grained breakdown of classification scores for the $k = 9, b = 2$ dataset configuration for the fine-tuned Qwen2.5-14B LLM. We sample 50 points randomly from each STaR dataset.



	Label	Pr.	Re.	F1.	Count
RCC-8	DC	0.69	0.90	**0.78**	10
	EC	0.50	**1.00**	0.67	3
	PD	0.43	0.27	0.33	11
	TPP	0.33	0.44	0.38	9
	NTPP	**1.00**	0.20	0.33	5
	TPPI	0.00	0.00	0.00	2
	NTPPI	0.50	0.25	0.33	4
	EQ	0.75	0.50	0.60	6
IA	=	0.50	0.17	0.25	6
	<	0.10	**1.00**	0.18	1
	>	0.83	**1.00**	**0.91**	5
	d	0.50	0.60	0.55	5
	di	0.67	0.50	0.57	4
	o	0.00	0.00	0.00	2
	oi	0.75	**1.00**	0.86	3
	m	**1.00**	0.50	0.67	4
	mi	**1.00**	0.33	0.50	3
	s	**1.00**	0.20	0.33	5
	si	**1.00**	0.25	0.40	4
	f	0.33	0.50	0.40	2
	fi	**1.00**	0.17	0.29	6

Table 4: Fine-grained breakdown of classification scores for the $k = 9, b = 2$ dataset configuration for the o3-mini LRM. We sample 50 points randomly from each STaR dataset.

CoT token usage is high for splits (small b) where the model has high accuracy



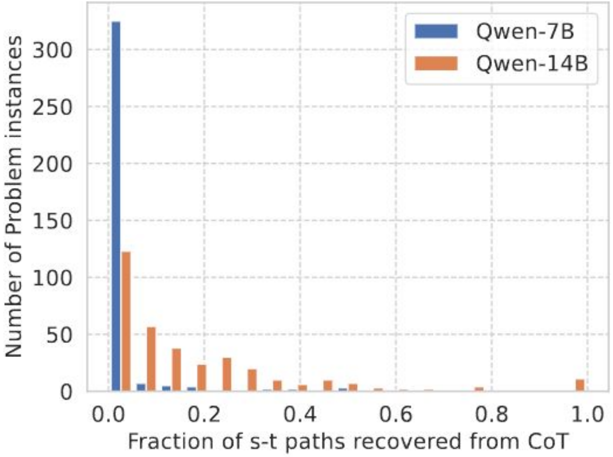
How faithfully can a SoTA LRM recover the Algebraic Closure Algorithm (ACA) that solves STaR?

The algebraic closure algorithm has the following components (Renz and Ligozat, 2005)

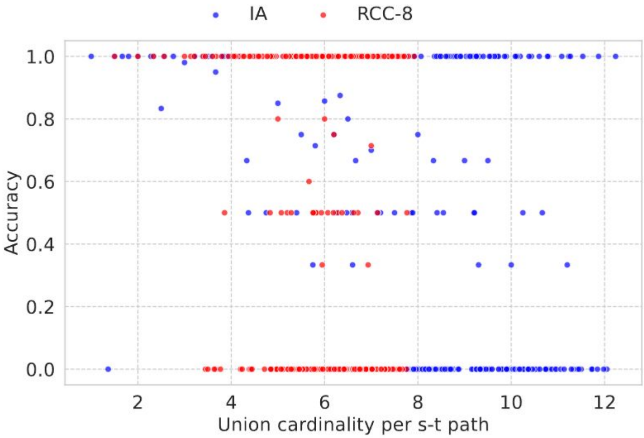
1. **Bellman-Ford search**: to find all (s-t) paths from source *s* to sink *t* node
2. **Set Union:** Compose relations along edges disjunctively
3. **Set Intersection:** Intersect the sets of possible target relations along each s-t path *b*

LRMs are shallow ACA simulators

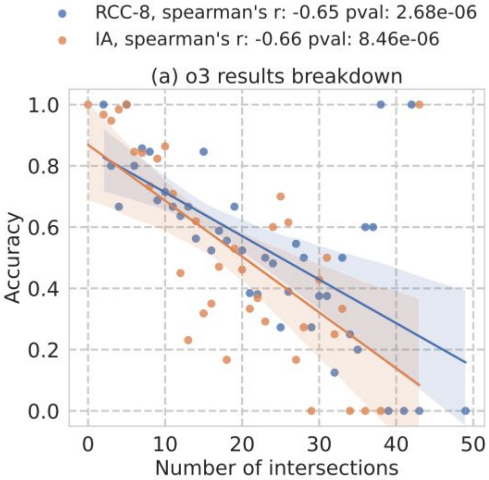
1. Path fidelity



2. Union fidelity

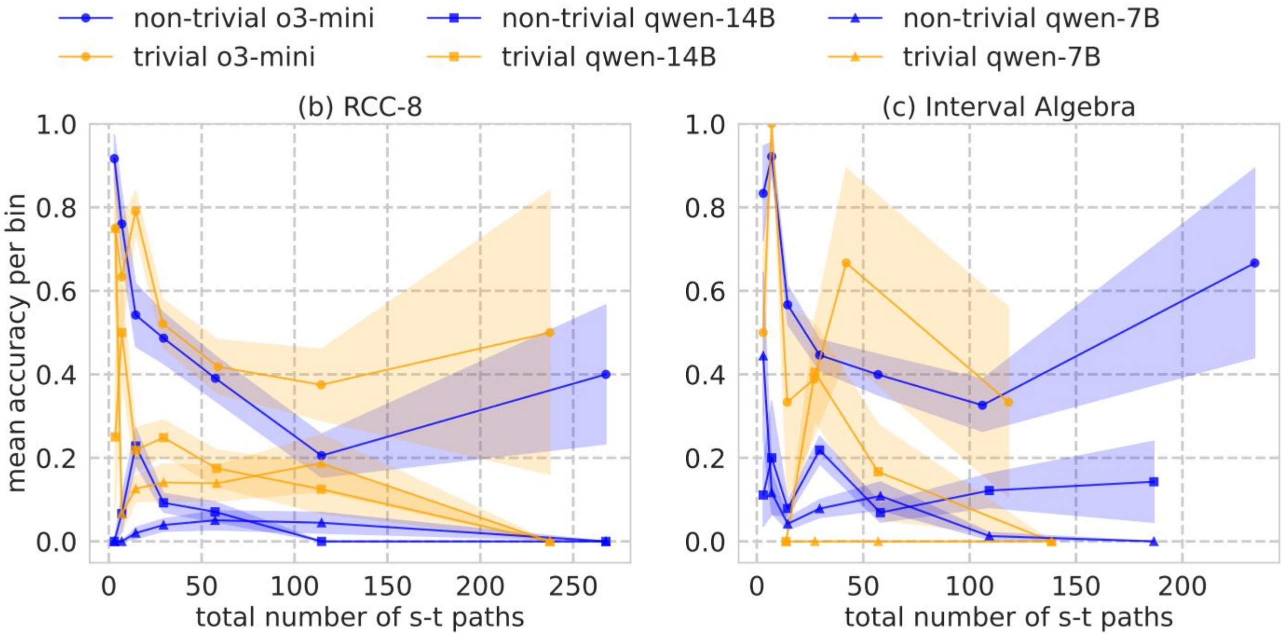


3. Intersection fidelity



Exploiting the trivial path heuristic and model size scaling

There exists a single (s-t) relational path with at most 1 non-identity (not EQ/=) relation.



Conclusions

1. LLMs and LRMs are shallow disjunctive reasoners.
2. We tested their disjunctive reasoning ability on the STaR benchmark
   a. that allows proper OOD testing
   b. and is cheaply synthetically generated with rich topologies that are unlikely to be memorized
   c. Performance is better than chance so they are reasoning, noisily.
3. A behavioral analysis reveals that LRMs like o3-mini can shallowly approximate different components of the Algebraic closure algorithm

Thanks for listening!

STaR dataset:
https://huggingface.co/datasets/erg0dic/STaR

Paper:
https://arxiv.org/abs/2503.23487