Large Language and Reasoning Models are Shallow Disjunctive Reasoners

Irtaza Khalid[1,†], Amir Masoud Nourollah[1], Steven Schockaert[1]

[1]Cardiff University

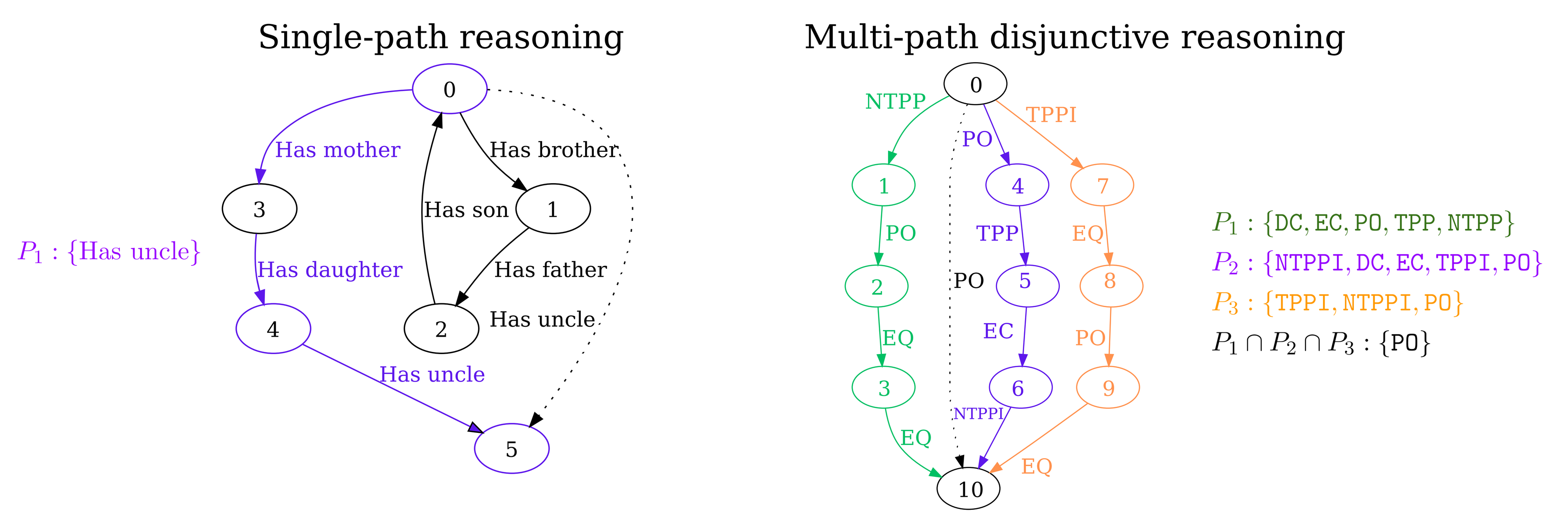
[†]khalidmi@cardiff.ac.uk

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

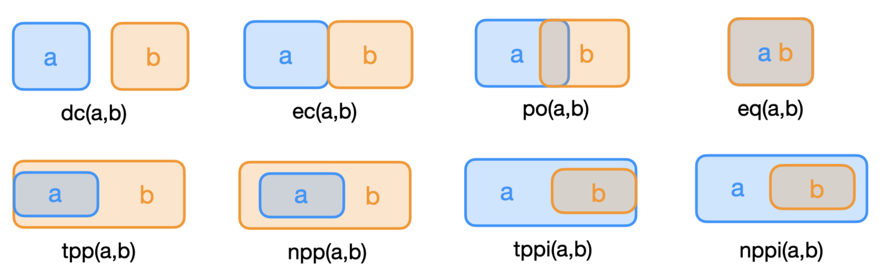
1. Summary

1. **Target:** Can Large Language Models (LLMs) and Large Reasoning Models (LRMs) reason or are they shallow pattern-matching on internet-scale data?
2. **Method:** We benchmark LLMs and LRMs on the STaR benchmark [1] for the problem of disjunctive reasoning, whilst circumventing previous issues with test data e.g. memorization (e.g. for GSM8k) [2].
3. **Novelty:** STaR problems are novel as the intermediate computation nodes need to contain multiple possible solutions or sets, compared to other art.
4. **Punchline:** LLMs and LRMs are shallow disjunctive reasoners.
5. **Why?:** A behavioral analysis reveals that LRMs like o3-mini can shallowly approximate different components of the Algebraic closure algorithm that solves the STaR benchmark [3].

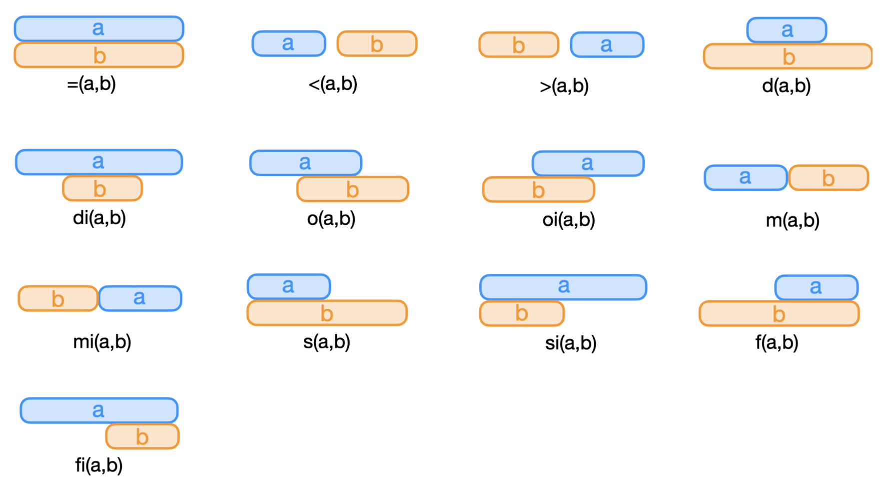
2. Benchmarking Disjunctive Reasoning



STaR example



RCC-8 relations [4]



IA relations [5]

Spatio-Temporal Reasoning (STaR) benchmark:

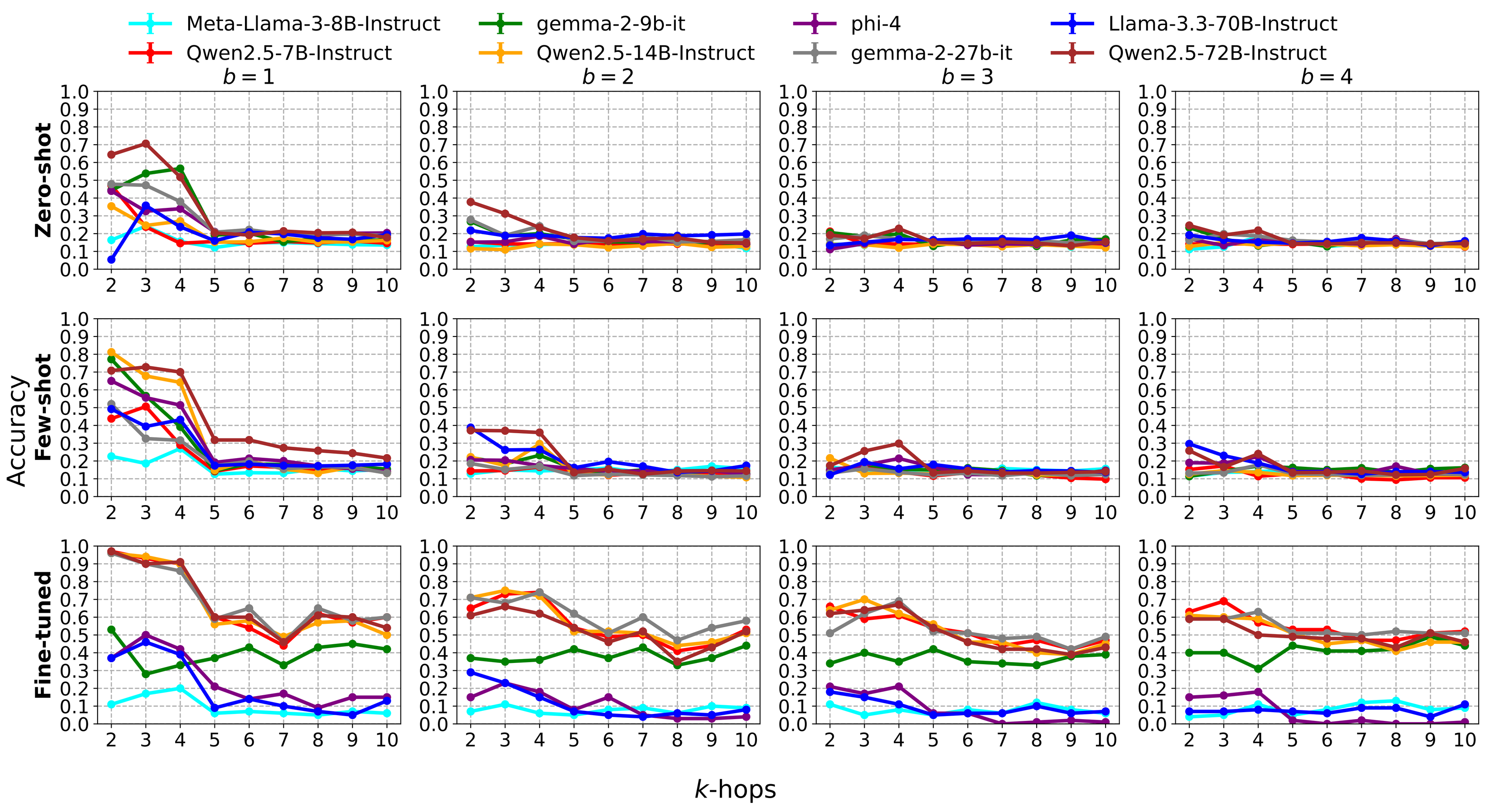
- The Systematic Generalization (SG) task is framed as a graph link classification problem $(s, ?, t)$.
- **(Def) SG** is the ability of a model to solve test instances by composing knowledge that was learned from multiple training instances [6], where the test instances are typically larger than the training instances.
- **Problem complexity parameters** : s-t path length k (number of edges) and number of s-t paths b
- **Train/test split:** Train on $k = 2, 3, 4$, $b = 1, 2, 3$, test on $2 \leq k \leq 10$ and $1 \leq b \leq 4$

Input Representation

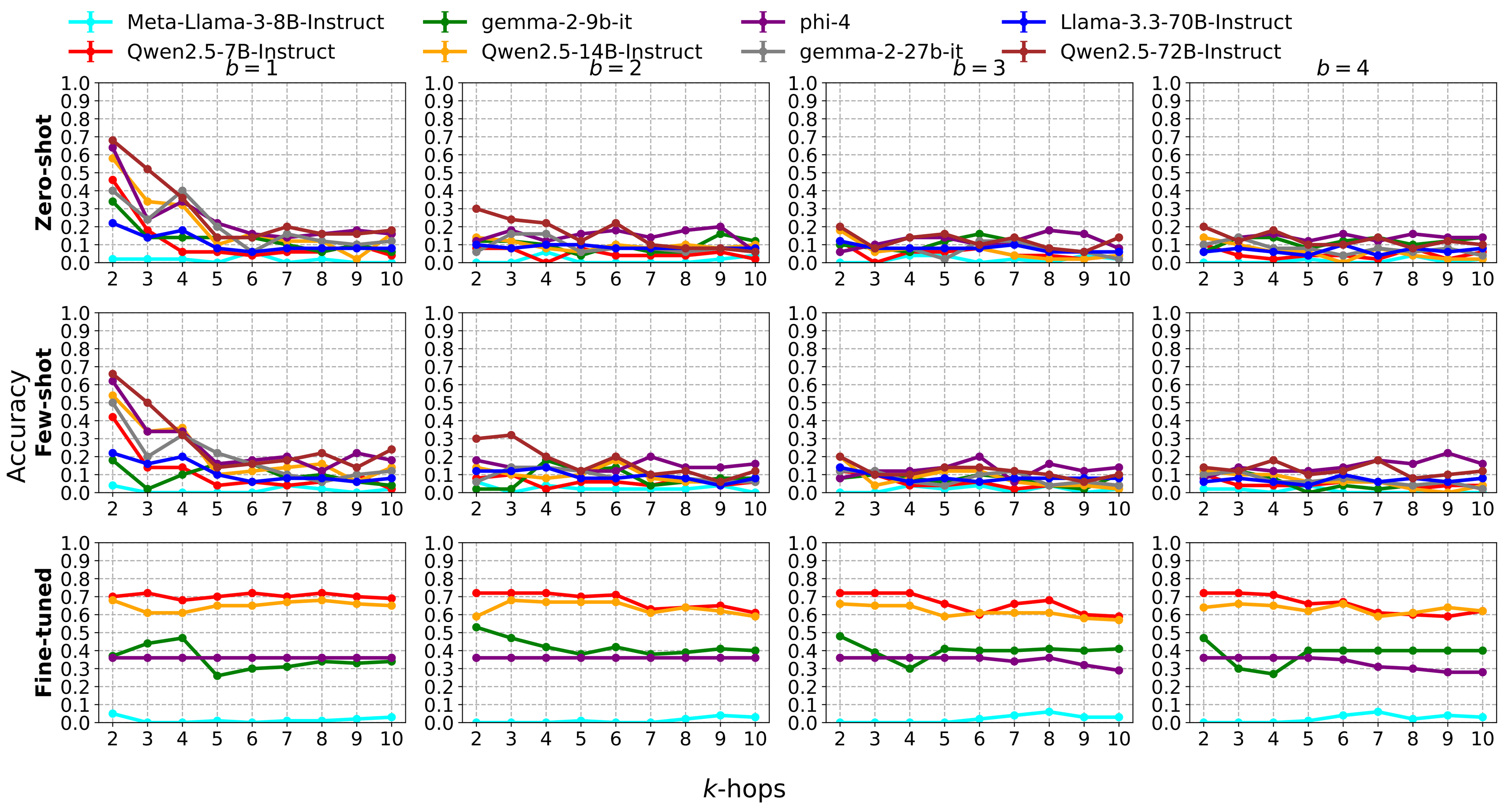
Instruction (*Q*): You are a helpful assistant. **Just answer the question as a single integer.** Given a consistent graph with edges comprising the 8 base relations, predict the label of the target edge. More specifically, Given a data row delimited by a comma with the following columns: \`graph_edge_index\`,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

4. Results



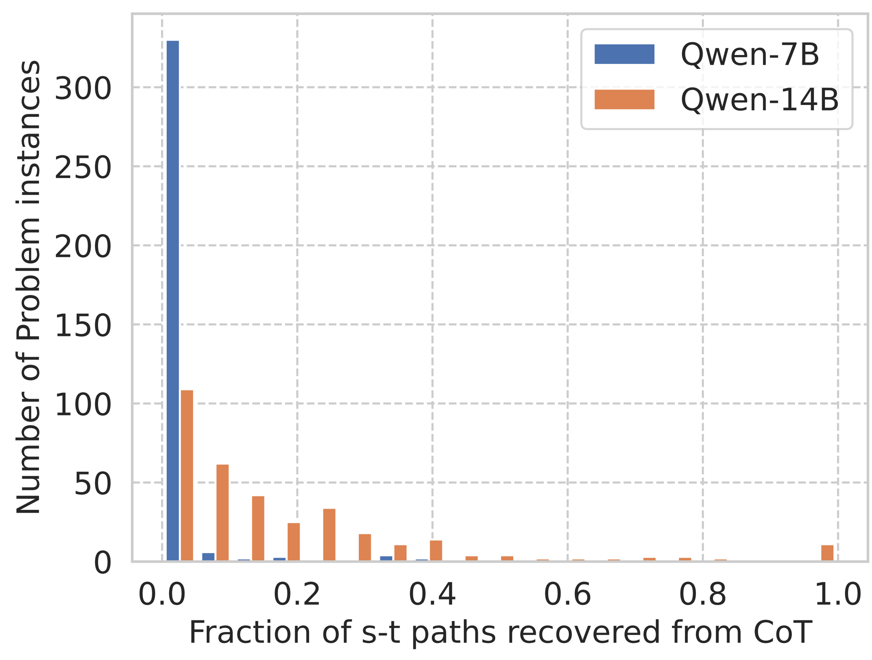
Non-reasoning LLM results on the RCC-8 split



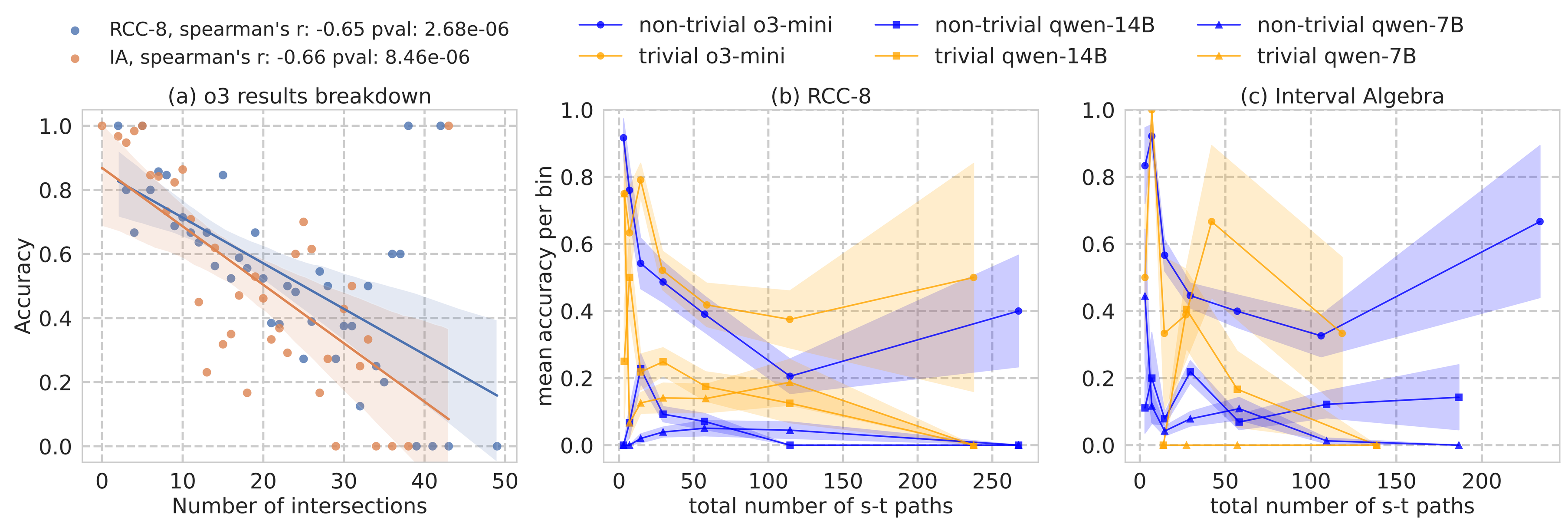
Non-reasoning LLM results on the IA split

	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

LRM results on STaR



Fraction of $s-t$ paths recovered from CoT



LRMs are shallow Algebraic Closure Algorithm (ACA) simulators. (a) o3-mini's performance on STaR. (b)-(c) Models, increasingly with size, zero-shot exploit the trivial path heuristic for solving STaR problems. Error bars are $\pm 1\sigma$.

References

[1] Irtaza Khalid and Steven Schockaert. Systematic relational reasoning with epistemic graph neural networks. In *ICLR*, 2025.
[2] Hugh Zhang et. al. A careful examination of large language model performance on grade school arithmetic. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
[3] Jochen Renz et. al. Weak composition for qualitative spatial and temporal reasoning. In *Principles and Practice of Constraint Programming 2005*.
[4] Zhan Cui et. al. Qualitative and topological relationships in spatial databases. In *Advances in Spatial Databases, Third International Symposium, SSD'93, Singapore, June 23-25, 1993, Proceedings*, volume 692 of *Lecture Notes in Computer Science*, pages 296–315. Springer, 1993.
[5] James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.
[6] Dieuwke Hupkes et. al. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.