How Is LLM Reasoning Distracted by Irrelevant Context? An Analysis Using a Controlled Benchmark

Minglai Yang[1], Ethan Huang[1], Liang Zhang[1], Mihai Surdeanu[1], William Yang Wang[2], Liangming Pan[1]

[1]University of Arizona [2]University of California, Santa Barbara

mingly@arizona.edu, liangmingpan@arizona.edu

Computational Language Understanding

Looking for PhD - Fall 26

UCSB NLP GROUP

Links

Paper

@YMinglai

www.ymingl.com

Data & Codes

Introduction

How robust is LLM reasoning?

LLM is easy to be affected by irrelevant context.

Flanker Effect:
When a target stimulus is surrounded by distractors suggesting a different response, people take longer to respond and tend to make more mistakes.

Challenges:

1) How does varying the amount of irrelevant context affect robustness?
2) Can robust reasoning be enhanced through continued pretraining or LoRA?
3) How does the intensity of irrelevant context during training impact model performance in both in-distribution and OOD scenarios?
4) How can the above questions be qualitatively evaluated?

Solutions:

GSM-DC - A synthetic benchmark

- The explicit injection of irrelevant context via off-path nodes and edges without affecting correct solutions.
- Adjustment of reasoning complexity by varying graph depth and structure.
- Automatic evaluation of model outputs.
- Exploration through controlled experiments.

Metrics

Automatic stepwise evaluation of solutions by comparing with the correct reasoning path:

Step Accuracy (SAcc):
- Each step must compute the correct value using only reachable nodes in G'.
- Extra steps are allowed if they don't interfere.

Path Accuracy (PAcc):
- The predicted reasoning must node-level aligned with solution path P.
- Permitting redundancy but not confused by irrelevant context.

Extraction Answer Accuracy (EAcc):
- The final answer must match the answer from ground-truth solution S.

Note: All metrics are computed using a symbolic parser with node-level alignment, not strict sentence-level sequence matching.

Limitations

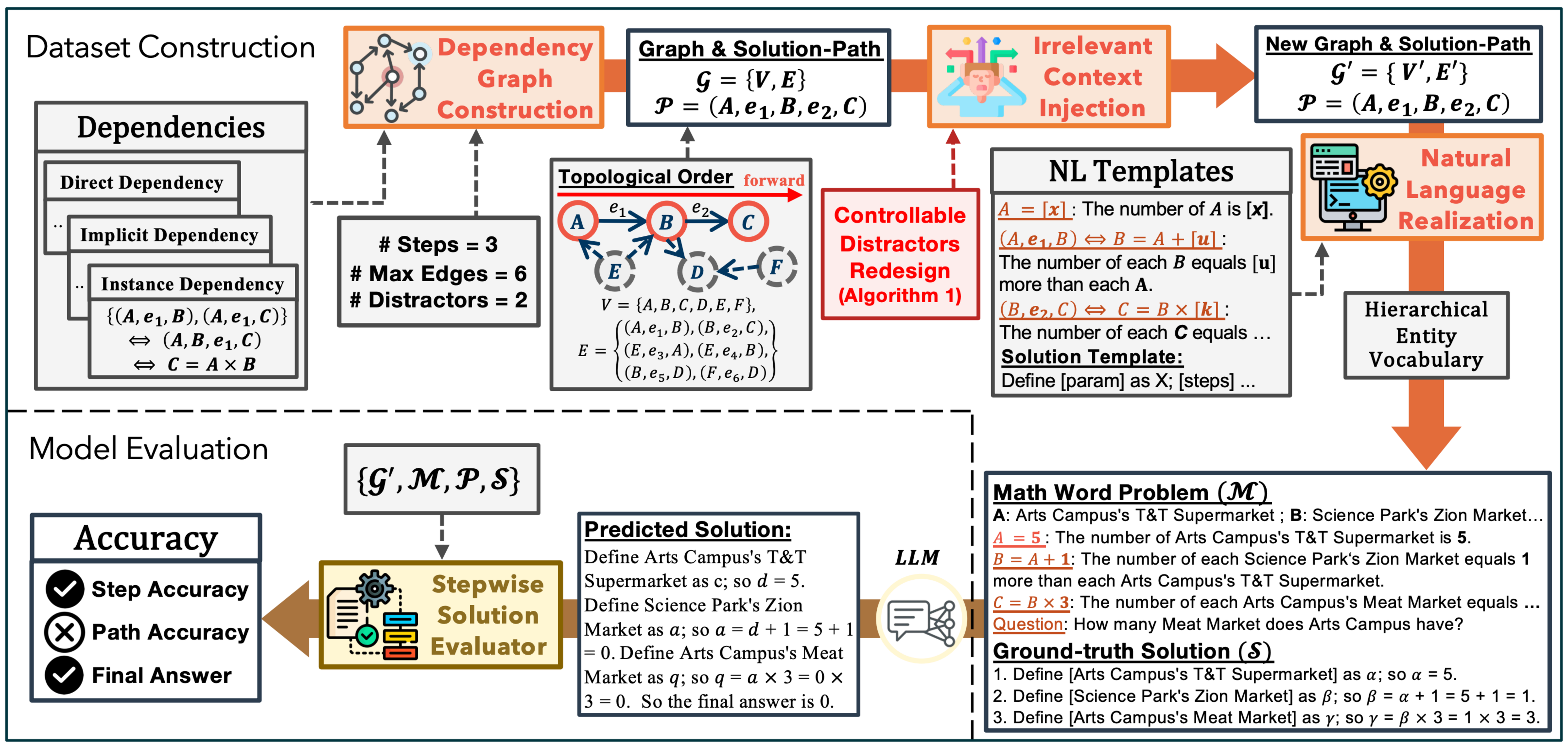
Broader applicability:
- Methodology applies to any symbolic reasoning task (e.g., logic, algorithms).

Extension to non-unique reasoning paths:
- Allow multiple valid reasoning chains.

Plans for new evaluations:
- RL-based training using Process Reward Models.
- Designing stepwise evaluator to evaluate reasoning models such as OpenAI o1/o3/o4 and DeepSeek-R1.

Graph-Based Benchmark for Controlled Experiments

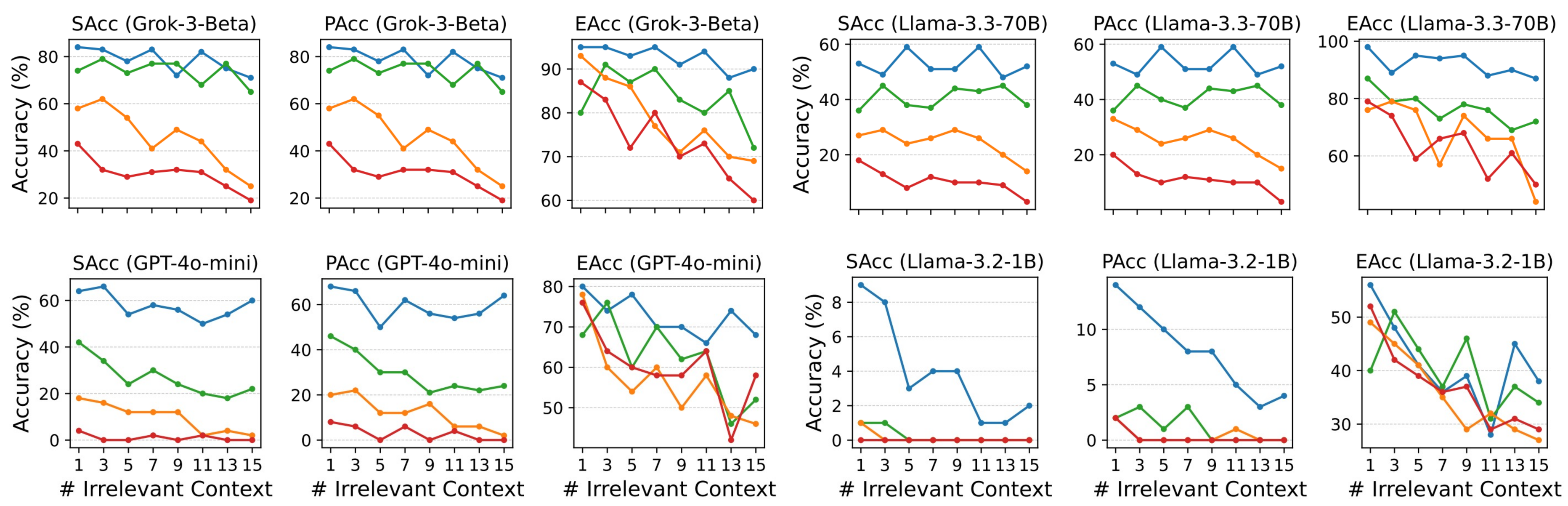
General Framework (Gen&Eval): Grade School Math with Distracting Context (GSM-DC)



Results from Controlled Experiments

Result 1: LLMs' reasoning performance degrades with increasing irrelevant context.

Result 2: Irrelevant context degrades accuracy more steeply at greater reasoning depths.



Distractor Scaling Law

$$E(m; rs) = m^{\delta(rs)}$$

- $E(m; rs)$: error rate
- m: distractor count
- rs: reasoning steps
- $\delta(rs)$: model's sensitivity to distractors at reasoning steps rs

Result 3: Continued pretraining enhances robustness even without access to IC samples.

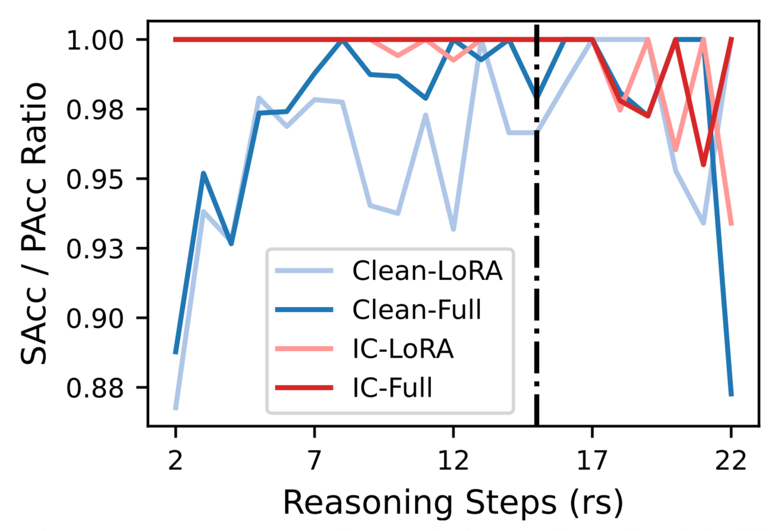


Fig: Step accuracy of models trained with Non-IC or IC data using LoRA or continued pretraining.

Result 4: Training with irrelevant context improves robustness most effectively.

rs	Clean		Clean+IC		IC	
	SAcc	PAcc	SAcc	PAcc	SAcc	PAcc
≤ 15	35.9	41.3	70.0	71.2	**73.2**	**74.7**
16	22.0	22.7	32.0	32.0	**33.3**	**33.3**
17	21.0	21.0	23.0	23.0	**20.7**	**21.3**
18	13.0	13.0	15.7	15.7	**16.7**	**16.7**
19	13.7	13.7	13.3	13.3	**15.0**	**15.0**
20	9.0	9.0	8.3	8.3	**10.0**	**10.0**
21	7.7	7.7	8.7	8.7	**5.7**	**5.7**
22	6.0	6.0	5.3	5.3	**6.3**	**6.3**

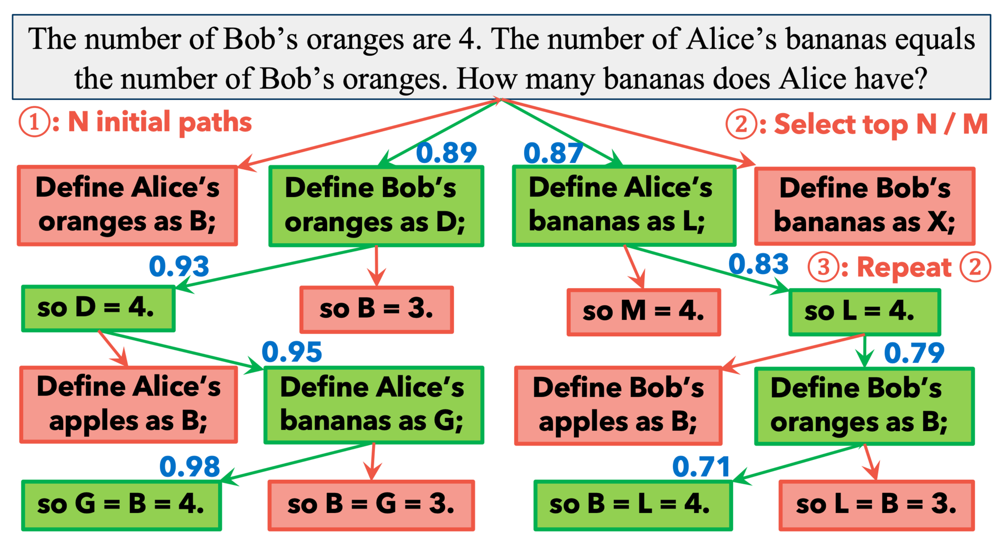
Fig: Comparison of SAcc and PAcc under different training regimes: Clean, Clean+IC, and IC.

Result 5: Training with challenging irrelevant context leads to the strongest robustness and generalization across all pretraining settings.

Training IC Level	Testing w/ IC (SAcc)			Testing w/o IC (SAcc)		
	ID	OOD	All	ID	OOD	All
Clean	35.91	13.19	32.36	81.95	17.05	60.32
Light-IC	64.79	6.90	46.57	67.33	7.09	46.56
Medium-IC	65.79	7.23	47.44	69.39	9.95	50.38
Hard-IC	**77.95**	**18.57**	**59.48**	**82.30**	**19.86**	**61.21**
Mix-IC	73.23	15.33	57.86	78.09	15.62	57.38

Training IC Level	ID Test SAcc			OOD Test SAcc		
	Light	Medium	Hard	Light	Medium	Hard
Light-IC	67.21	66.57	60.57	8.14	7.29	5.28
Medium-IC	68.14	66.07	63.14	8.71	8.43	4.57
Hard-IC	**78.36**	**79.21**	**76.28**	**22.7**	**18.43**	**14.57**
Mix-IC	74.71	75.07	69.93	17.7	16.57	11.28

Result 6: Improving reasoning robustness at test time: Tree search can enhance the generalization capabilities of LLMs.



Training IC Level	ID SAcc			OOD SAcc		
	w/o PRM	w/ PRM	Δ	w/o PRM	w/ PRM	Δ
Light-IC	64.79	**66.10**	+1.31	6.90	**9.59**	+2.69
Medium-IC	65.79	**70.05**	+4.26	7.23	**13.52**	+6.29
Hard-IC	77.95	**79.48**	+1.53	18.57	**24.17**	+5.60
Mix-IC	73.23	**75.81**	+2.58	15.33	**19.06**	+3.73
Clean	35.91	**36.38**	+0.47	13.19	**15.76**	+2.57