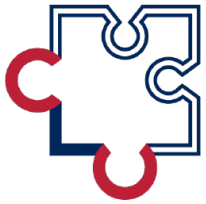


How Is LLM Reasoning Distracted by Irrelevant Context? An Analysis Using a Controlled Benchmark

Minglai Yang, Liangming Pan

Analyze Robust Reasoning Capability

LLMs have exhibited emergent ability to **"reason" like human**

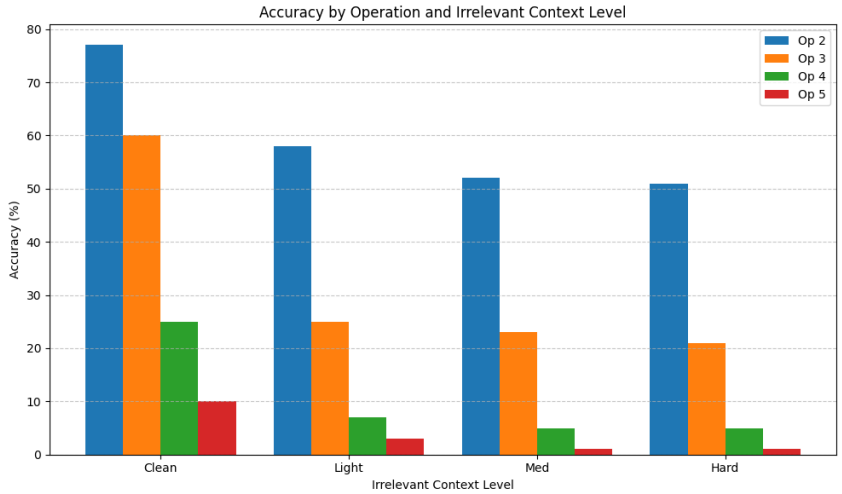
… which gives them the potential to **reasoning in robust environment**

However, current language models are still **sensitive to prompts, not generalizable and illogical reasoning …**

How robust is language model reasoning?

We need to evaluate how language models perform in tasks requiring arithmetic precision and resistance to distractions.

Why does LLM fall for distractions?



We find that LLM suffers from the irrelevant information heavily, especially OOD.

Eriksen-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.

Grade-School Math with Irrelevant Context (Denny Zhou)

Modified Problem

Maria buys a large bar of French soap that lasts her for 2 months. She spends $8.00 per bar of soap. *Every 10 months, Maria's neighbor buys a new shampoo and moisturizer for Maria's neighbor.* If Maria wants to stock up for the entire year, how much will she spend on soap?

Correct CoT Output to the Original Problem

Maria spends $8.00 per bar of soap. She buys a bar of soap every 2 months. There are 12 months in a year. Maria will spend $8.00 x 12 / 2 = $48.00 on soap for the year. The answer is $48.00.

Output to the Modified Problem

CoT

Maria buys a large bar of French soap that lasts her for 2 months. She spends $8.00 per bar of soap. Every 10 months, Maria's neighbor buys a new shampoo and moisturizer for Maria's neighbor. If Maria wants to stock up for the entire year, she will spend $8.00 x 12 = $96.00 on soap. The answer is $96.00.

LtM

Let's break down this problem: 1. How much does Maria spend on soap every 10 months? 2. How much does Maria spend on soap every year?
1. Maria spends $8.00 per bar of soap. She buys a new bar of soap every 10 months. So Maria spends $8.00 * 10 = $80.00 every 10 months.
2. Maria spends $80.00 every 10 months. So Maria spends $80.00 * 12 = $960.00 every year.
The answer is $960.00.

Program

```
Soap_price = 8.00
Soap_per_year = 2 * 12
Soap_total = Soap_price * Soap_per_year
print(Soap_total)
```

Output: 192.0

The added irrelevant sentence is in italic and highlighted in red, which causes different errors (highlighted in yellow) for all prompting techniques.

Original Problem

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets. [Irrelevant Sentence] How many more tickets should Jeanne buy?

Options for the Irrelevant Sentence Topic

In-Topic [ROLE] rides [NUMBER] kilometers to the bus station every day.
Off-Topic The shoe size of [ROLE] is [NUMBER].

Cutting Through the Noise (Anantheswaran, 2024)

Models	Og	Adv	Drop
Gemini-1.5 Pro	93.70	78.64	16.07
Llama-3 (70B)	86.59	70.65	18.41
Mistral Large	90.13	70.86	21.38
Claude-3 Sonnet	90.04	69.39	22.93
Reka Flash (21B)	88.29	61.85	29.95
Llama-2 (70B)	37.05	25.55	31.02
Yi (34B)	58.82	36.38	38.15
Command-R+ (100B)	78.19	47.72	38.97
Qwen-1.5 (72B)	73.38	42.91	41.53

	Simple						Complex					
Test set →	Og			Adv			Og			Adv		
Training set →	Og-Tr	Adv-Tr	Cmb-Tr	Og-Tr	Adv-Tr	Cmb-Tr	Og-Tr	Adv-Tr	Cmb-Tr	Og-Tr	Adv-Tr	Cmb-Tr
Llama-2 (7B)	56.25	60.00	57.84	42.62	50.82	43.96	26.85	28.70	27.01	22.02	28.44	22.65
Mistral (7B)	42.50	41.25	41.65	36.07	50.83	38.49	30.56	25.71	26.84	22.93	27.62	23.64
Llama-2 (13B)	72.50	68.75	71.25	62.30	60.66	63.93	20.37	28.70	21.29	19.26	29.35	18.34

	Zero-shot		One-shot		Two-shot	
	Og	Adv	Og	Adv	Og	Adv
Yi (34B)	42.21	38.13	54.31	38.33	53.34	50.87
Llama-2 (70B)	38.54	36.17	32.33	29.97	34.58	30.28
Llama-3 (70B)	65.36	60.01	69.29	66.31	72.64	68.06
Mistral Large	71.38	65.64	74.14	69.86	78.93	74.68
Qwen-1.5 (72B)	59.23	53.56	65.94	62.04	68.21	63.32

Through carefully designed trivial human-crafted noise injection and datasets (**PROBLEMATHIC** and **GSM-8K-Adv**).
They demonstrated how to enhance model robustness against noise: Fine-tuning models on adversarial data improves their resistance to noise, while training on original data proves less effective.

Gaps in Previous Research

Analysis of the robustness of LLM reasoning under irrelevant context

In those datasets, the irrelevant context is either human-crafted or single-sentence instances.


Analysis only based on prompts


Not been tested on more complex questions involving OOD testing


Potential for Fine-tuning Overfitting

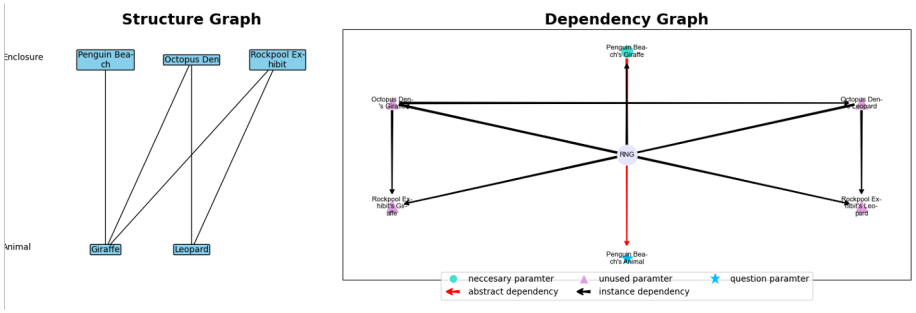

Cannot differ the error from reasoning chain or arithmetic mistake


Controllable Questions + Controllable Irrelevant context + Precise Accuracy

Outline

- Introduction
- Dataset Construction: Graph-Based Controllable Math Dataset
- Evaluation: Differ Math and Path Search
- Analysis: Controlled Experiments on Irrelevant Context
- Improvement: Tree Search with Process Reward Models
- Conclusion

Physics of Language Models Part 2.1 (Zeyuan Allen-Zhu)



Problem:
The number of each Penguin Beach's Giraffe equals 6. The number of each Octopus Den's Leopard equals each Octopus Den's Giraffe. The number of each Rockpool Exhibit's Leopard equals 20 more than the sum of each Octopus Den's Giraffe and each Octopus Den's Leopard. The number of each Rockpool Exhibit's Giraffe equals 8 times as much as the sum of each Octopus Den's Giraffe and each Octopus Den's Leopard. The number of each Octopus Den's Giraffe equals 21. How many Animal does Penguin Beach have?
Solution:
Define Penguin Beach's Giraffe as e; so e = 6. Define Penguin Beach's Animal as J; so J = e = 6.

iGSM Dataset

Example problem numbers, edges, operations, and variables are all controllable.

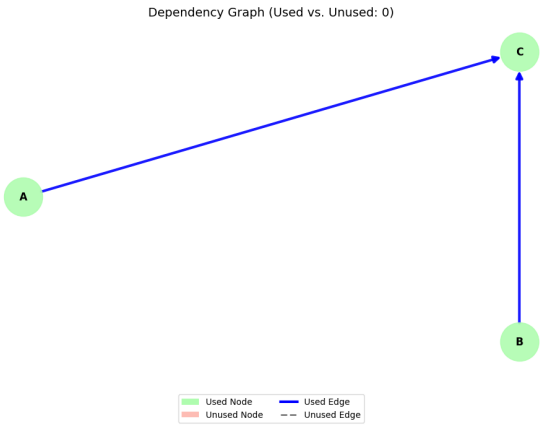
They build a **D**irected **A**cyclic **G**raph (DAG) for each problem and select two points, asked the questions based on these two points and the solution will be the path by topological sort to these two points.

Therefore, the solution can also be checked step by step.

Graph-Based Math Problem Generation

Inspired by the iGSM dataset, we adopt their math problem generation approach using a Knowledge DAG.

In our dataset, each node corresponds to a term selected from the GSM8K, while the edges represent numerical relationships between pairs of variables.



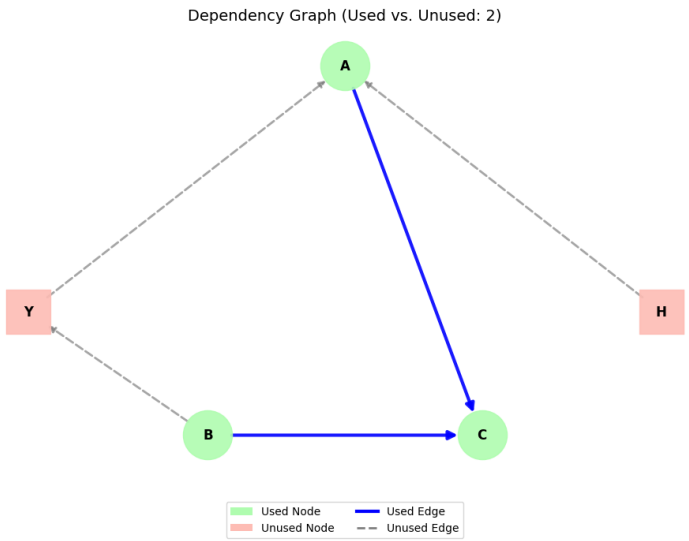
Question

The number of A equals 3.
The number of B equals 4.
The number of C is computed as A plus B. How much is C?

Reasoning

Define A as α, so $\alpha = 3$.
Define B as β, so $\beta = 4$,
Define C as γ, so $\gamma = \alpha + \beta = 3 + 4 = 7$.

After generating each question, we control the noise level by adjusting the unused parameters. Since the structure is a DAG, there is a unique path for each solution. The model is expected to ignore any unused parameters, such as H and Y.



Question

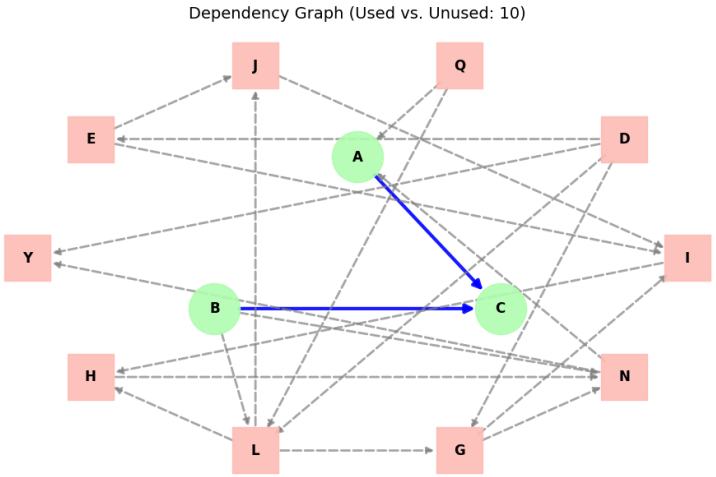
The number of A equals 3.
The number of B equals 4.
The number of C is computed as A plus B. How much is C?

Add Controllable Noise

The number of A equals 3.
The number of B equals 4.
The number of H equals 8.
The number of Y equals B plus 4 plus 1.
The number of C is computed as A plus B. How much is C?

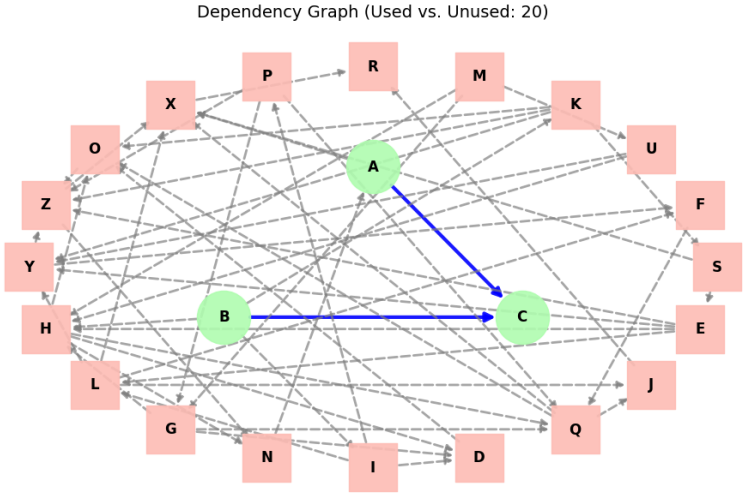
Reasoning

Define A as α, so $\alpha = 3$.
Define B as β, so $\beta = 4$,
Define C as γ, so $\gamma = \alpha + \beta = 3 + 4 = 7$.



The number of B equals 4. The number of D equals 6. The number of Q equals 10. The number of E equals D plus 2 plus 2. The number of L equals B plus 2 minus Q plus D plus 1 plus 1. The number of J equals L plus 3 minus E. The number of G equals L minus D plus 3. The number of I equals -G minus J plus E. The number of H equals 2 times L plus 4 times I plus 3. The number of N equals H plus 5 plus G minus B. The number of Y equals N plus 5 plus 2 times D. **The number of A equals 3. The number of C is computed as A plus B. How much is C?**

College of Information Science



The number of B equals 4. The number of M equals 2. The number of I equals -B plus 1. The number of K equals -B plus 1. The number of U equals M plus 2. The number of P equals -I plus 3. The number of S equals -K. The number of G equals P plus 4 plus M plus 1. The number of E equals -S plus 2. The number of H equals B plus 1 plus 2 times G plus 4 times E plus 2 times U plus M. The number of L equals 2 times I plus E plus 3 plus 2. The number of D equals G plus 4 plus I plus H plus 3. The number of Y equals -K plus U plus E plus 4 times L. The number of F equals -Y plus 3 times L. The number of Q equals G plus F minus H minus P plus 3. The number of O equals D plus 3 times Q plus K plus 1 minus H plus 3. The number of J equals Q plus 4 plus 2 times L plus 3. The number of Z equals K plus P plus 3 minus O minus Y plus 3 times E plus 2. The number of N equals H plus 2 plus Z plus 2. **The number of A equals 3. The number of C is computed as A plus B.** The number of X equals Q plus S minus Q plus 2 times Z plus A plus L plus 2 plus 1. The number of R equals 4 times J plus X plus 1. **How much is C?**

Controllable Irrelevant Context at different operations

- Empirical Irrelevant Context Stratification via CDF Partitioning
- Irrelevant Context Score (z_i): Quantifies extraneous info per question
- Empirical CDF: $\hat{\mathrm{F}} = \frac{1}{M}\sum_{n=1}^{M} i = 1^{M}\mathbb{I}(z_i \leq t)$
- Partition into N bins: Thresholds τ_i where $\widehat{\mathrm{F}_z}(\tau_k) = \frac{k}{N} \rightarrow$ equal-sized noise levels

Operation	Extraneous Nodes			High (>66%)
	Light	Medium	Hard	Range
$op = 2$	0-2	3-4	5-19	2623
$op = 3$	0-1	2-4	5-17	2430
$op = 4$	0-1	2-3	4-15	3234
$op = 5$	0-1	2-3	4-14	3028
$op = 6$	0-1	2-3	4-15	2935
$op = 7$	0-1	2-3	4-13	2613
$op = 8$	0-1	2-3	4-12	2488
$op = 9$	0-1	2-2	3-13	3386
$op = 10$	0-1	2-2	3-13	3252
$op = 11$	0-0	1-2	3-13	3090
$op = 12$	0-0	1-2	3-11	2847
$op = 13$	0-0	1-2	3-11	2590
$op = 14$	0-0	1-2	3-11	2385
$op = 15$	0-0	1-2	3-10	2277
$op = 16$	0-0	1-1	2-10	3289
$op = 17$	0-0	1-1	2-10	3117
$op = 18$	0-0	1-1	2-9	2795
$op = 19$	0-0	1-1	2-10	2550
$op = 20$	0-0	1-1	2-10	2361

A Dataset Example of op=2

Question (Easy)

The number of each Arts Campus's T&T Supermarket equals 3.
The number of each Science Park's Zion Market equals 1 more than each Arts Campus's T&T Supermarket.
The number of each Engineering Campus's Zion Market equals each Engineering Campus's T&T Supermarket.
How many Zion Market does Science Park have?

Question (Medium)

The number of each Arts Campus's T&T Supermarket equals 3.
The number of each Arts Campus's La Michoacana Meat Market equals 4.
The number of each Preparatory School District's La Michoacana Meat Market equals 3 more than the difference of each Science Park's T&T Supermarket and each Science Park's La Michoacana Meat Market.
The number of each Science Park's Zion Market equals 1 more than each Arts Campus's T&T Supermarket.
The number of each Engineering Campus's Zion Market equals each Engineering Campus's T&T Supermarket.
How many Zion Market does Science Park have?

Question (Hard)

The number of each Arts Campus's T&T Supermarket equals 3. The number of each Arts Campus's La Michoacana Meat Market equals 4. The number of each Arts Campus's Seafood City Supermarket equals 2 more than each Science Park's Zion Market. The number of each Preparatory School District's Zion Market equals each Engineering Campus's Seafood City Supermarket. The number of each Science Park's Seafood City Supermarket equals the sum of each Science Park's La Michoacana Meat Market and each Science Park's T&T Supermarket. The number of each Preparatory School District's Seafood City Supermarket equals 4 more than the sum of each Science Park's Zion Market, each Arts Campus's T&T Supermarket and each Arts Campus's Seafood City Supermarket. The number of each Arts Campus's Zion Market equals the sum of each Science Park's T&T Supermarket, each Arts Campus's T&T Supermarket and each Engineering Campus's La Michoacana Meat Market. The number of each Preparatory School District's T&T Supermarket equals 4 more than each Engineering Campus's Seafood City Supermarket. The number of each Science Park's T&T Supermarket equals 4. The number of each Engineering Campus's La Michoacana Meat Market equals 0. The number of each Engineering Campus's T&T Supermarket equals 1 times as much as the difference of each Engineering Campus's La Michoacana Meat Market and each Preparatory School District's Seafood City Supermarket. The number of each Engineering Campus's Seafood City Supermarket equals 2 times as much as the sum of each Science Park's Seafood City Supermarket, each Science Park's La Michoacana Meat Market and each Science Park's T&T Supermarket. The number of each Science Park's La Michoacana Meat Market equals 3 times as much as each Science Park's T&T Supermarket. The number of each Preparatory School District's La Michoacana Meat Market equals 3 more than the difference of each Science Park's T&T Supermarket and each Science Park's La Michoacana Meat Market. The number of each Science Park's Zion Market equals 1 more than each Arts Campus's T&T Supermarket. The number of each Engineering Campus's Zion Market equals each Engineering Campus's T&T Supermarket.
How many Zion Market does Science Park have?

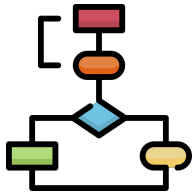
Reasoning

Define Arts Campus's T&T Supermarket as e; so e = 3.
Define Science Park's Zion Market as w; so w = 3 + e = 3 + 1 = 4.

Outline

- Introduction
- Dataset Construction: Graph-Based Controllable Math Dataset
- Evaluation: Separating Mathematical Reasoning from Path Search
- Analysis: Controlled Experiments on Irrelevant Context
- Improvement: Tree Search with Process Reward Models
- Conclusion

Novel Evaluation Metrics



Stepwise Accuracy

- Verifies the correctness of each reasoning step.
- Ensures all intermediate calculations align with the expected solution chain.

Distraction Robustness (Path Accuracy)

- Separates arithmetic accuracy from the model's ability to filter out irrelevant information.
- Assesses if the model correctly identifies key variables in a reasonable sequence.
- Provides a direct measure of resilience to distraction.

Final-Answer Verification (Answer Accuracy)

- Evaluates the correctness of the final answer, regardless of CoT deviations
- While some closed-source models excel in this metric, they may struggle with fine-tuning and produce non-standard CoT explanations.

Novel Evaluation Metrics

Large language model reasoning is **unreliable** in terms of:

- Poor performance under larger operations
- Sensitive to irrelevant context injections

GPT-4o mini's performance declines as irrelevant context increases on our dataset.

Op.	Accuracy (%)				Path Accuracy (%)			
	Clean	Light	Med	Hard	Clean	Light	Med	Hard
2	77	58	52	51	79	59	54	53
3	60	25	23	21	66	34	32	24
4	25	7	5	5	30	9	7	5
5	10	3	1	1	14	5	4	3

Less Robust Reasoning Capability

Large language model reasoning is **less robust** in terms of:

- **Limited Generalization:** Struggles to generalize reasoning to larger operations.
- **Poor Distraction Handling:** Fails to filter out irrelevant information during reasoning.

Operation	NoNoise	Noise	NoNoise Irr	Noise Irr	NoNoise Extr	Noise Extr
4o-mini						
OP=2	77.33	52.00	78.33	54.33	82.00	61.67
OP=3	60.33	26.33	66.00	31.67	50.67	39.33
OP=4	25.33	6.00	30.00	10.33	47.00	33.00
OP=5	10.00	4.00	14.00	6.67	39.67	28.00
LLaMA-1B-Instruct						
OP=2	18.00	2.00	20.00	5.33	70.67	39.67
OP=3	7.00	2.00	9.67	2.33	38.67	29.67
OP=4	1.00	0.00	1.67	0.00	34.33	23.33
OP=5	0.00	0.00	1.33	0.33	25.33	24.67

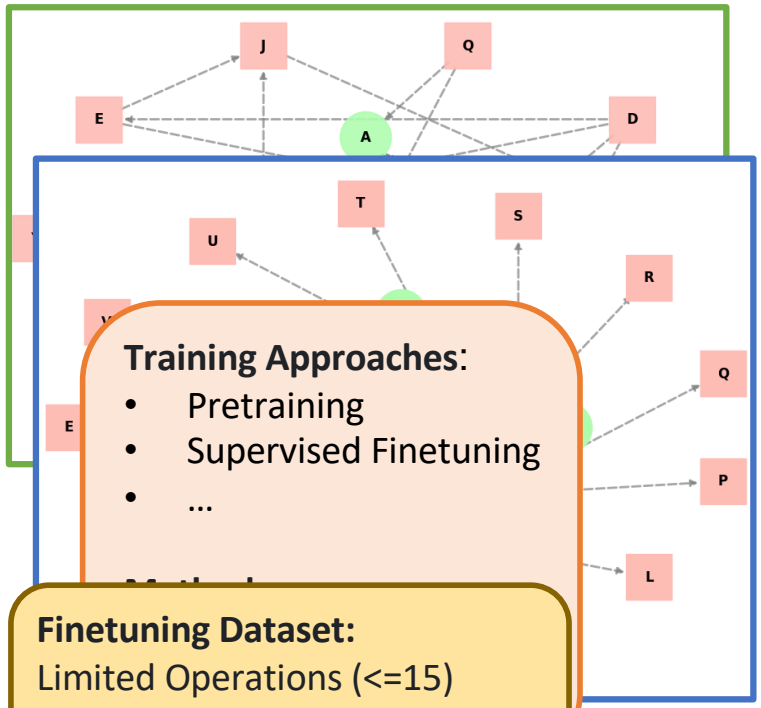
Outline

- Introduction
- Dataset Construction: Graph-Based Controllable Math Dataset
- Evaluation: Differ Math and Path Search
- Analysis: Controlled Experiments on Irrelevant Context
- Improvement: Tree Search with Process Reward Models
- Conclusion

GSM-DI Dataset

Control Experiments of Math Word Problems for LLM Reasoning

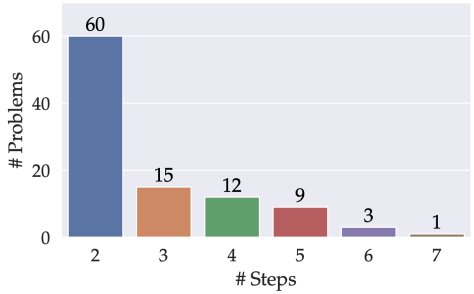
- Controllable Noise
- Controllable Operations
- Finetuning Methods
- Finetuning Dataset



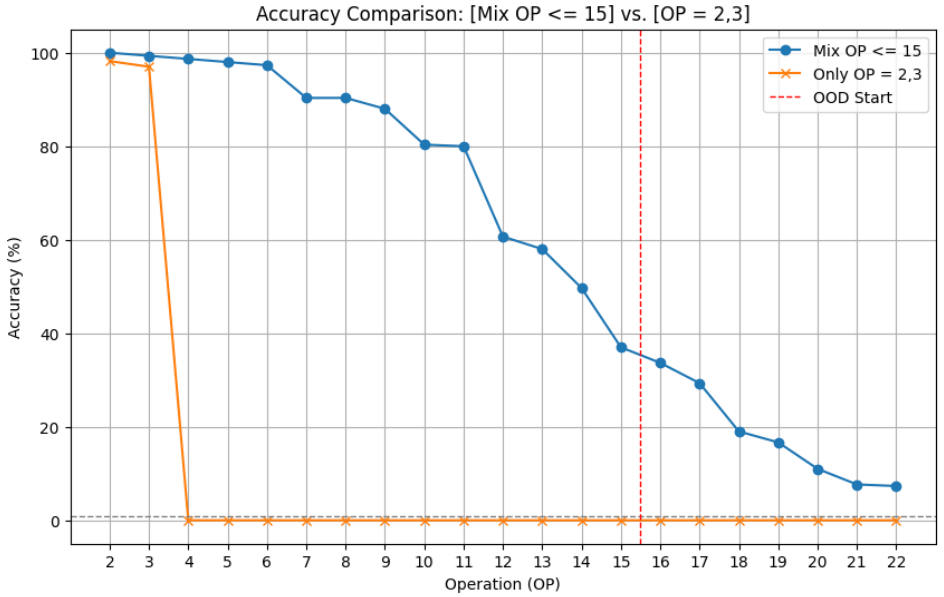
Controlled Experiments on Dataset

An LLM that is fully trained on limited operations may struggle with reasoning in more complex or larger operations.

Training on a limited operations is not generalizable



We will test the same op distribution as GSM-IC



Controlled Experiments on Dataset

An LLM trained on a mix of clean and irrelevant context data can achieve better out-of-distribution performance with irrelevant context.

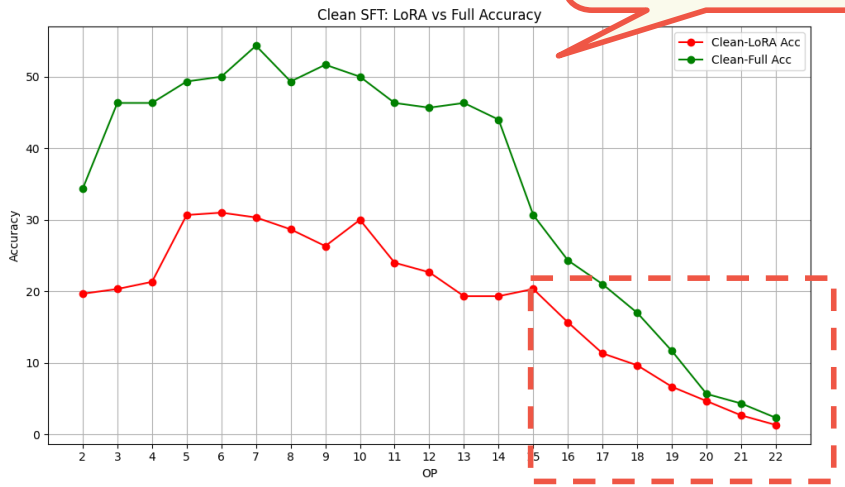
Op	Clean		Noisy		Mixed	
	Acc	PAcc	Acc	PAcc	Acc	PAcc
16	24.33	24.33	27.33	27.33	33.67	33.67
17	21	21	25	25	29.33	29.33
18	17	17.33	14.67	15	19	19
19	11.67	12	11.67	12	16.67	16.67
20	5.67	5.67	7	7	11	11
21	10.33	10.33	7	7.33	7.67	7.67
22	2.33	2.67	5	5	7.33	7.33

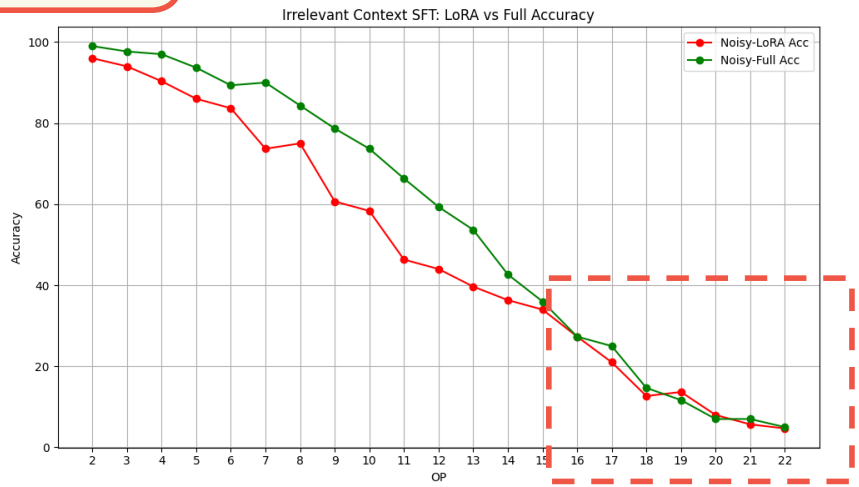
Best in Generalization

Controlled Experiments on Finetuning Methods

Due to the complexity of path learning and denoising, full SFT is generally more effective than LoRA in mathematical reasoning tasks.

LoRA Finetuning does not help at Path Search



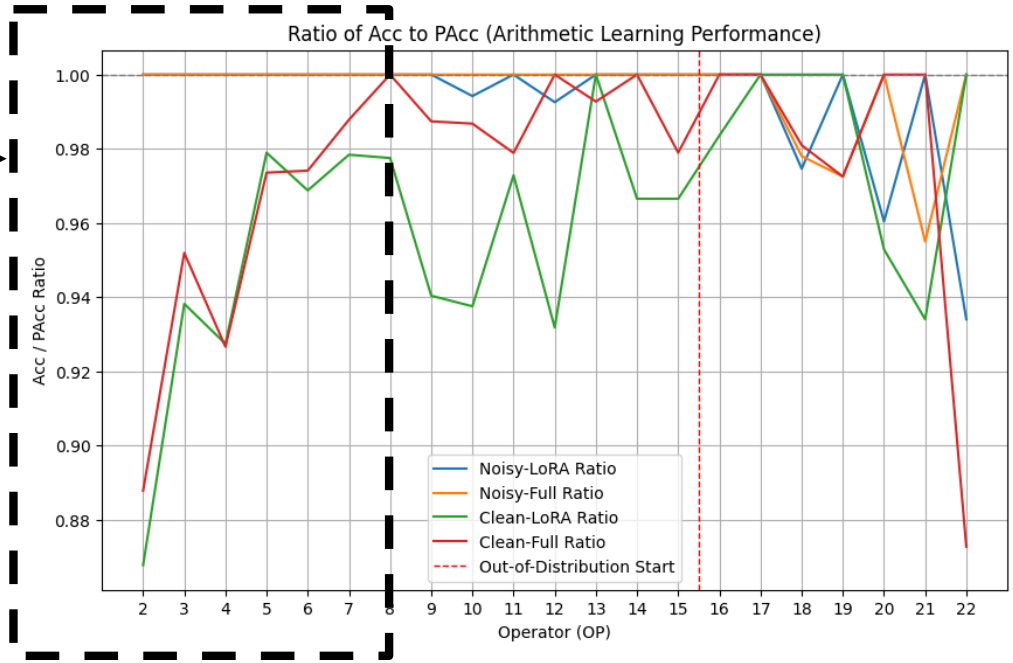


Full Finetuning doesn't help with OOD testing even with irrelevant context

Controlled Experiments on Finetuning Methods

Noise will affect the model's ability to perform mathematical operations, even if it has identified the correct path.

Irrelevant Context will influence LLM doing arithmetic operations



Arithmetic Accuracy= Δ(Accuracy, Path Accuracy)

Controlled Experiments on Irrelevant Context

Injecting more irrelevant context into the training set can improve performance on both ID and OOD test, enhancing the model's ability to generalize to complex mathematical operations—even on a clean test set.

Training Set	Noise			Clean		
	ID	OOD	All	ID	OOD	All
Light Noise	69.62	9.13	46.57	69.74	8.87	46.56
Medium Noise	70.05	10.71	47.44	74.10	11.84	50.38
Hard Noise	**82.00**	**22.88**	**59.48**	**85.00**	**22.54**	**61.21**
Mixed Noise	78.87	16.71	55.19	81.46	18.25	57.38
Clean Only	47.23	15.38	35.09	84.80	20.54	60.32

Training With "Hard Noise" Achieved a higher score than "Clean Only"

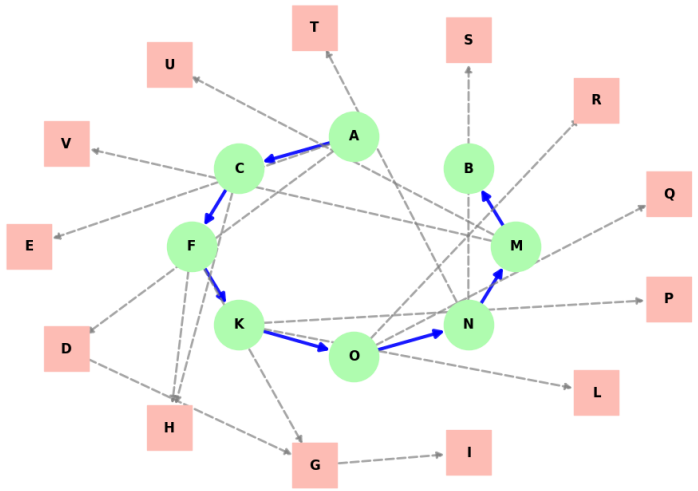
- **ID:** op <=15
- **OOD:** op > 15

Outline

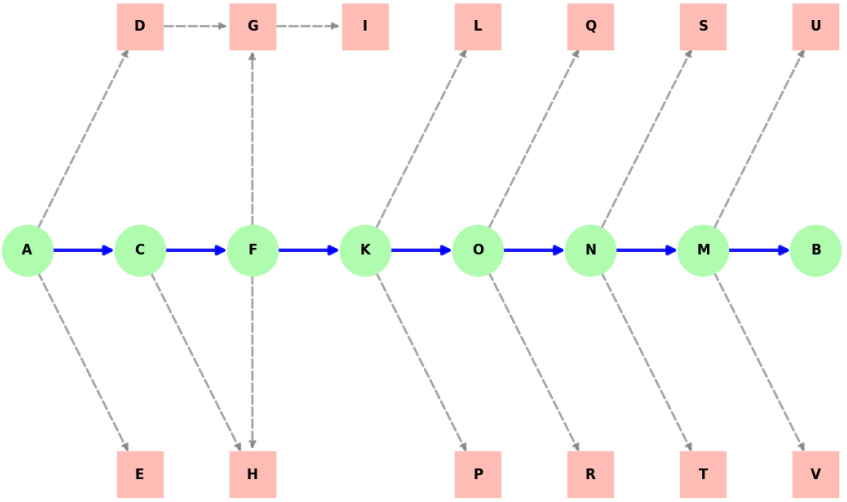
- Introduction
- Dataset Construction: Graph-Based Controllable Math Dataset
- Evaluation: Differ Math and Path Search
- Analysis: Controlled Experiments on Irrelevant Context
- Improvement: Tree Search with Process Reward Models
- Conclusion

Path Search in DAG

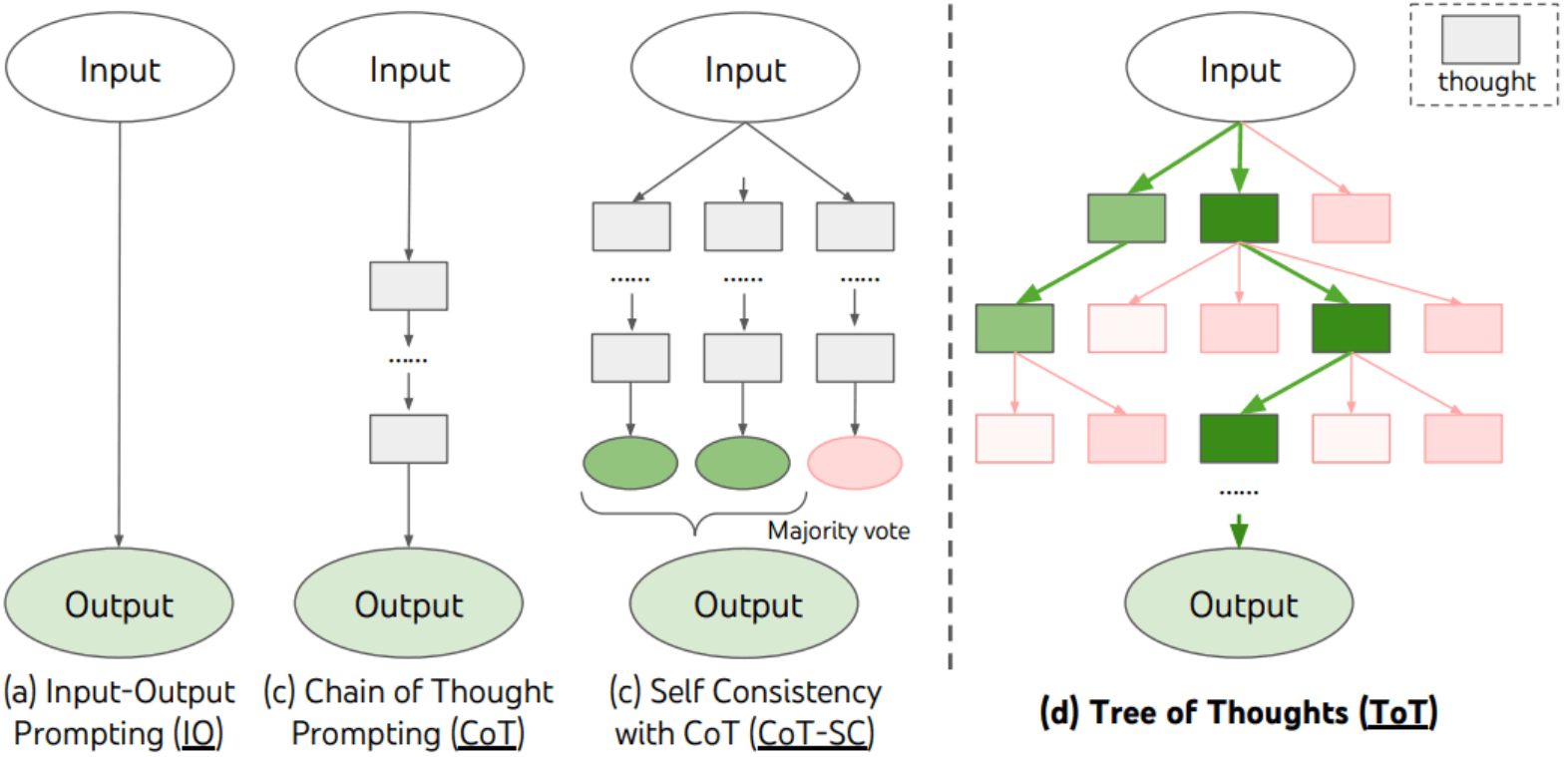
Dependency Graph



Tree Like Dependency Graph



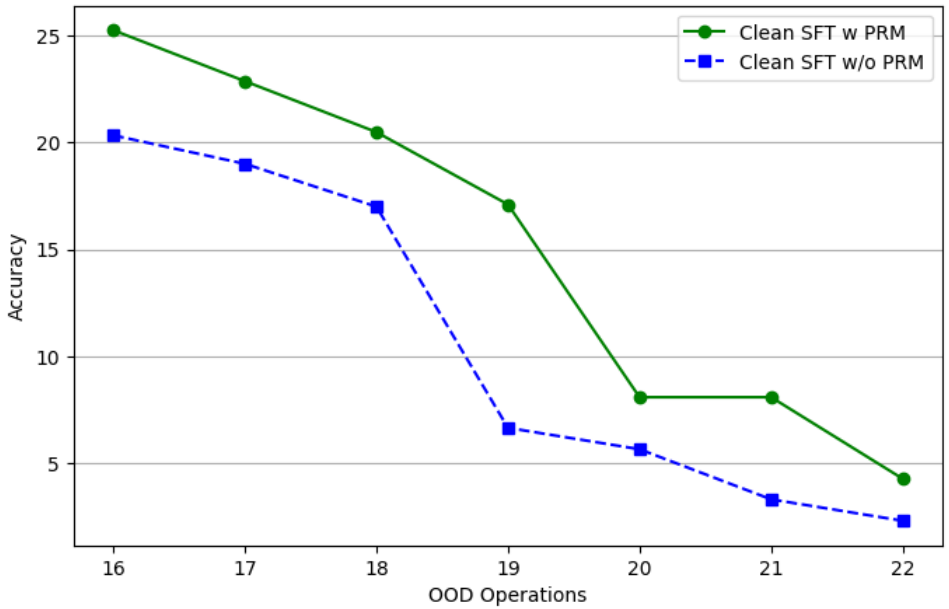
Tree-of-Thought Reasoning



[Yao et al, Tree of Thoughts: Deliberate Problem Solving with Large Language Models]

Tree-of-Thought Reasoning

Performance with an increasing number of reasoning depth



Clean SFT with PRM is more reliable at deeper reasoning depths, surpassing the non-PRM model.

PRM may have a positive impact on complex reasoning tasks, particularly demonstrating greater reliability in OOD reasoning problems.

Outline

- Introduction
- Dataset Construction: Graph-Based Controllable Math Dataset
- Evaluation: Differ Math and Path Search
- Analysis: Meticulous Controlled Experiments
- Improvement: Tree Search with Process Reward Models
- Conclusion

Takeaways

- Introducing controllable irrelevant context into the training set improves the reliability of reasoning.
- Learning path search is more challenging than mastering arithmetic calculations.
- Tree search can enhance the generalization capabilities of LLMs

Thank You!

mingly@arizona.edu

http://ymingl.com