

Paper Presentation

Jason Liu

jasonl@wustl.edu

Paper 1: Progressive-Hint Prompting Improves Reasoning in Large Language Models

 Background: Prompt design matters for LM reasoning

State-of-the-Art (SOTA) methods: Chain-of-Thought (CoT) and Self-Consistency

 Contribution: Progressive-Hint Prompting (PHP)

Use LM's (own) previously-generated result to guide its reasoning toward the correct answer

Improved accuracy without losing efficiency

Can be used alongside other SOTA methods

Background: Reasoning

 Arithmetic Reasoning: solving math word problems (MWP)

 Commonsense Reasoning: making inference from world knowledge

 Symbolic Reasoning: e.g. last letter concatenation, coin flip

Coin flip: tracking the state of coin after people flip or not-flip it

Background: Prompt design

⚙️ Prompt Engineering's boost to LM reasoning is comparable to full fine-tuning.

It is also more sample-efficient

🔗 Chain-of-Thought (CoT) is a prompting method that promotes step-by-step reasoning


A “greedy” [1] approach

📁 Self-consistency is another that samples multiple answers and aggregate them

Various reasoning paths can reach the same right answer for complex problems

[1] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations, 2023. 3, 9


Contribution: iterative refinement

 Examining reasoning steps we made so far can help us get to the right answer

 Correcting mistakes

 Doing better in the next reasoning step

Contribution: Progressive-Hint Prompting(PHP)

 Sequentially interact with the LM as it works toward the right answer

Given a question Q:

1. Get answer $A = \text{LM}(Q)$
2. $Q = Q + A$
3. Repeat 1, 2 until the Subsequent Answer stabilizes (does not change over 2 iterations)

Base Answer: the first A we get from the original Q

“Stage 1,” handled differently from subsequent answer

Subsequent Answer: A's we get thereafter

💡 Contribution: Progressive-Hint Prompting (PHP)

🤖 Observation: Q snowballs over the iterations

Base Prompting

Q: There are 96 fourth-graders at Small Tree School. 43 of them are girls. On Friday, 5 fourth-grade girls and 4 fourth-grade boys were absent. How many fourth grade boys were at Small Tree School on Friday?



Base Answer

A: There were 96 fourth-graders at Small Tree School. 43 of them were girls. On Friday, 5 girls and 4 boys were absent. So there were $43 - 5 = 38$ girls at school on Friday, and $96 - 38 = 58$ boys at school on Friday. The answer is **58**.

Progressive-Hint Prompting

Q: There are 96 fourth-graders at Small Tree School. 43 of them are girls. On Friday, 5 fourth-grade girls and 4 fourth-grade boys were absent. How many fourth grade boys were at Small Tree School on Friday? (Hint: The answer is near to **58**).



Second Answer

A: **We know the Answer Hints: 58. With the Answer Hints: 58, we will answer the question.** There were 96 fourth-graders at Small Tree School, and 43 of them were girls. On Friday, 5 girls and 4 boys were absent. So there were $43 - 5 = 38$ girls and $96 - 43 = 53$ boys at Small Tree School on Friday. After subtracting the absent students, there were $38 - 4 = 34$ girls and $53 - 5 = 48$ boys. The answer is **48**.

Progressive-Hint Prompting

Q: There are 96 fourth-graders at Small Tree School. 43 of them are girls. On Friday, 5 fourth-grade girls and 4 fourth-grade boys were absent. How many fourth grade boys were at Small Tree School on Friday? (Hint: The answer is near to **58, 48**).



Third Answer

A: **We know the Answer Hints: 58, 48. With the Answer Hints: 58, 48, we will answer the question.** There were 96 fourth-graders, and 43 of them were girls. On Friday, 5 girls and 4 boys were absent. So there were $43 - 5 = 38$ girls and $96 - 9 = 87$ students in total. Of those 87 students, 38 were girls, so there must have been $87 - 38 = 49$ boys. The answer is **49**.

Progressive-Hint Prompting

Q: There are 96 fourth-graders at Small Tree School. 43 of them are girls. On Friday, 5 fourth-grade girls and 4 fourth-grade boys were absent. How many fourth grade boys were at Small Tree School on Friday? (Hint: The answer is near to **58, 48, 49**).



Fourth Answer

A: **We know the Answer Hints: 58, 48, 49. With the Answer Hints: 58, 48, 49, we will answer the question.** There were 96 fourth-graders at Small Tree School, and 43 of them were girls. On Friday, 5 girls and 4 boys were absent. So there were $43 - 5 = 38$ girls and $96 - 43 = 53$ boys originally. After the absences, there were $38 - 5 = 33$ girls and $53 - 4 = 49$ boys. The answer is **49**.

👉 The 'Hint' keeps getting longer

1 "Is near"

2 "We know the Answer Hints..." & "we will answer..."

⚠️ Note: green text NOT added to the Base Answer! (Its there in the Subsequent Answer via CoT demonstrations)


Contribution: Progressive-Hint Prompting(PHP)

 CoT: cot_PHP_aqua.txt (<https://github.com/chuanyang-Zheng/Progressive-Hint>)

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?

Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64

A: If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (a).

 PHP (“PHP-Cot” in the paper)

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is? (Hint: The answer is near to (a)).

Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64

A: We know the Answer Hints: (a). With the Answer Hints: (a), we will answer the question. If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (a).

💡 Contribution: Experiment Setting

👉 Model needs to figure out right answer in BOTH situations:

Hint == correct answer

Hint != correct answer

🔄 Recall: the 'Base Answer' starts off the interactive process

Generated using Standard prompt, CoT prompt, and Complex CoT

Datasets: AddSub, MultiArith, SingleEQ, SVAMP, GSM8K, AQuA and MATH

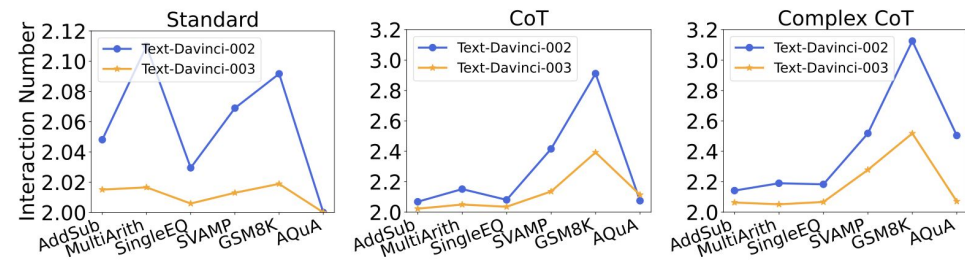
LLMs: text-davinci-002, text-davinci-003, GPT-3.5-Turbo and GPT-4

Contribution: Varying LMs and prompting methods

Table 2: PHP, when applied to different LLMs and prompting methods, can help to improve the performance. Meanwhile, PHP works better when the model and prompt are more powerful. The results are with greedy decoding.

	Prompt	PHP	Dataset					Average	
			AddSub	MultiArith	SingleEQ	SVAMP	GSM8K		AQuA
GPT-3.5 text-davinci-002	Standard [8]	✗	79.4	34.0	80.7	64.8	15.1	25.5	49.91
		✓	80.5 (+1.1)	31.8 (-2.2)	79.9 (-0.8)	64.2 (-0.6)	14.7 (-0.4)	25.5 (0.0)	49.43 (-0.48)
	CoT [8]	✗	85.8	89.1	89.7	72.9	49.5	44.4	71.89
		✓	86.8 (+1.0)	89.0 (-0.1)	90.1 (+0.4)	72.3 (-0.6)	51.1 (+1.6)	45.6 (+1.2)	72.48 (+0.59)
	Complex CoT [10]	✗	82.5	89.8	87.7	70.4	57.6	37.4	70.89
		✓	83.7 (+1.2)	90.1 (+0.3)	89.9 (+2.2)	74.6 (+4.2)	61.2 (+3.6)	37.0 (-0.4)	72.75 (+1.86)
GPT-3.5 text-davinci-003	Standard [8]	✗	89.1	36.3	83.8	68.7	15.9	28.3	53.68
		✓	89.1 (0.0)	36.0 (-0.3)	83.6 (-0.2)	68.7 (0.0)	16.0 (+0.1)	28.3 (0.0)	53.61 (-0.07)
	CoT [8]	✗	90.6	93.6	92.7	81.0	56.1	44.0	76.33
		✓	91.1 (+0.5)	94.0 (+0.4)	93.5 (+0.8)	81.3 (+0.3)	57.5 (+1.4)	44.4 (+0.4)	76.96 (+0.63)
	Complex CoT [10]	✗	86.3	94.8	91.5	77.4	67.0	48.8	77.63
		✓	88.1 (+1.8)	95.0 (+0.2)	94.0 (+2.5)	80.0 (+2.6)	71.6 (+4.6)	50.0 (+1.2)	79.78 (+2.15)

- Better with more powerful LMs
- Better with more powerful prompts
- Less interactions needed with more powerful LMs
- Less interactions needed with and less powerful prompts



💡 Contribution: Varying Hint Quality

🔄 Recall: hints come from the previous Subsequent Answer. They are added to the current question

Table 3: Performance with different Base Answers. Initially, the base prompt provides base answers to the model and PHP generates the subsequent answers. The results are from text-davinci-003 with greedy decoding.

PHP	Base Prompt	Dataset						Average
		AddSub	MultiArith	SingleEQ	SVAMP	GSM8K	AQuA	
PHP-Standard	Standard [8]	89.1	36.0	83.6	68.7	16.0	28.3	53.61
	CoT [8]	92.4	80.5	92.1	78.5	50.2	42.5	72.70
	Complex CoT [10]	90.6	80.6	92.9	77.2	60.3	45.6	74.53
PHP-CoT	Standard [8]	90.8	92.5	90.7	80.2	52.3	40.9	74.56
	CoT [8]	91.1	94.0	93.5	81.3	57.5	44.4	76.96
	Complex CoT [10]	90.6	96.8	93.7	81.2	62.6	50.0	79.14
PHP-Complex CoT	Standard [8]	88.3	80.1	93.3	80.4	65.5	35.4	73.83
	CoT [8]	88.8	95.6	94.8	81.4	70.6	45.6	79.46
	Complex CoT [10]	88.1	95.0	94.0	80.0	71.6	50.0	79.78

💡 Contribution: Ablation study on adding Rehearsing hint

Table 4: Ablation Study. CoT-Merge: for the CoT base prompt and the PHP-CoT prompt, we employ the prompt that contains both base prompt and the PHP. **P1**: We know the Answer Hints A_1, \dots, A_p . **P2**: With the Answer Hints A_1, \dots, A_p , we will answer the question. According to the experiment results, we see that both the proposed P1 and P2 are necessary. Meanwhile, non-merge based method is better than merge based method when prompts are more powerful. The results are from text-davinci-003 with greedy decoding.

Method	P1	P2	Dataset						Average
			AddSub	MultiArith	SingleEQ	SVAMP	GSM8K	AQuA	
CoT-Merge	✓	✓	91.3	94.6	93.1	79.5	58.6	50.0	77.85
CoT [8]	✗	✗	91.1	93.5	93.3	80.0	58.1	44.8	76.80
	✓	✗	90.8	93.1	92.9	80.7	58.8	43.7	76.66
	✗	✓	91.3	93.8	93.5	80.5	58.2	46.4	77.28
Complex CoT-Merge	✓	✓	91.1	94.0	93.5	81.3	57.5	44.4	76.96
	✓	✓	88.8	94.3	94.6	78.1	70.2	46.8	78.80
Complex CoT [10]	✗	✗	87.8	93.3	93.7	78.0	68.3	50.3	78.56
	✓	✗	87.8	95.1	94.2	78.5	70.5	48.4	79.08
	✗	✓	88.3	94.3	94.6	79.1	69.3	46.8	78.73
	✓	✓	88.1	95.0	94.0	80.0	71.6	50.0	79.78

🔄 Recall: hints are rehearsed at the beginning of the LM's answer too!

- Adding rehearsals improves performance
- PHP helps when there are correct AND incorrect hints

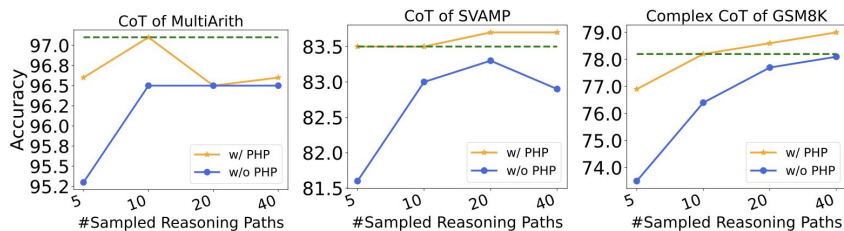
Table 5: Analysis of Hint Design (Shown in Figure 1). Correct: The hints of designed prompt are the same as the correct answers. Incorrect: The hints of the designed prompt are the incorrect answers. **Green**: The performance is better than without progressive-hint. **Red**: The performance is worse than without progressive-hint. The results are from text-davinci-003 with greedy decoding.

Method	Hint		Dataset						Average
	Correct	Incorrect	AddSub	MultiArith	SingleEQ	SVAMP	GSM8K	AQuA	
CoT [8]	✗	✗	90.6	93.6	92.7	81.0	56.1	44.0	76.33
	✓	✗	91.6	94.3	93.3	81.9	57.0	43.7	76.96
	✗	✓	91.1	93.5	93.1	79.7	57.9	45.2	76.74
	✓	✓	91.1	94.0	93.5	81.3	57.5	44.4	76.96
Complex CoT [10]	✗	✗	86.3	94.8	91.5	77.4	67.0	48.8	77.63
	✓	✗	88.3	94.0	93.8	77.8	68.6	46.4	78.14
	✗	✓	88.1	94.6	94.0	79.2	70.2	48.4	79.08
	✓	✓	88.1	95.0	94.0	80.0	71.6	50.0	79.78

💡 Contribution: Self-Consistency

Table 6: The results after adding Self-Consistency (SC). **Number**: The interaction number between agent and LLM. The **best results of adding PHP** are highlighted with red color, and the **best results without PHP** are highlighted with green color. We find that PHP further improves performance, even adding self-consistency. Meanwhile, PHP may reduce the cost of self-consistency.

Prompt	SC	PHP	Dataset						Average
			AddSub	MultiArith	SingleEQ	SVAMP	GSM8K	AQuA	
CoT [8]	5	✗	90.6	95.3	94.4	81.6	63.3	49.2	79.06
	5	✓	90.8	96.6	94.8	83.5	66.3	49.6	80.26
	10	Number	2.0075	2.0433	2.0098	2.1090	2.5458	2.0157	2.1218
	5	✗	90.6	96.5	93.8	83.0	65.5	49.2	79.76
	10	✓	90.8	97.1	93.8	83.5	67.5	50.0	80.45
	10	Number	2.0075	2.0283	2.0059	2.0510	2.2145	2.0118	2.0531
	20	✗	91.1	96.5	94.2	83.3	68.0	55.1	81.36
	20	✓	91.6	96.5	94.4	83.7	68.6	55.1	81.64
	20	Number	2.0050	2.0366	2.0098	2.0250	2.1144	2.0078	2.0330
	40	✗	91.6	96.5	94.8	82.9	67.3	53.1	81.03
	40	✓	91.6	96.6	95.0	83.7	68.4	53.1	81.39
	40	Number	2.0050	2.0300	2.0050	2.0320	2.0530	2.0000	2.0208
Complex CoT [10]	5	✗	88.1	97.0	93.1	80.4	73.5	51.5	80.60
	5	✓	89.6	97.3	95.2	82.5	76.9	51.9	82.23
	5	Number	2.0378	2.0166	2.0334	2.2370	2.5390	2.0118	2.1459
	10	✗	88.6	98.3	93.3	82.4	76.4	54.3	82.21
	10	✓	89.1	98.5	95.2	83.4	78.2	54.7	83.18
	10	Number	2.0177	2.0016	2.0295	2.059	2.1531	2.0078	2.0447
	20	✗	88.6	98.0	93.8	82.5	77.7	56.2	82.80
	20	✓	89.8	98.0	95.8	83.6	78.6	56.2	83.66
	20	Number	2.0253	2.0000	2.0196	2.0330	2.0401	2.0000	2.0196
	40	✗	88.3	98.5	94.8	83.9	78.1	58.6	83.70
	40	✓	88.6	98.5	95.8	84.7	79.0	58.6	84.20
	40	Number	2.0101	2.0000	2.0137	2.0210	2.0348	2.0039	2.0137



🔄 **Recall: Self-consistency samples multiple answers and aggregates them, thus exploring different reasoning paths.**

- After applying Self-Consistency, PHP improve performance further
- PHP + Self-consistency needs smaller number of total sampled reasoning paths to achieve comparable performance of just self-consistency

💡 Contribution: with GPT

Table 7: Performance of Complex CoT with GPT-3.5-Turbo and GPT-4, employing greedy decoding. Number: The average interaction number with LLM.

	PHP	Dataset						Average
		AddSub	MultiArith	SingleEQ	SVAMP	GSM8K	AQuA	
Previous SOTA	✗	94.9 [27]	100 [25]	95.5 [29]	89.1 [30]	92.0 [17]	76.4 [31]	91.31
GPT-3.5 Turbo	✗	85.5	97.5	92.5	81.0	82.8	57.4	82.78
	✓	85.3	98.0	92.9	83.1	85.1	60.6	84.16
		(-0.2)	(+0.5)	(+0.4)	(+2.1)	(+2.3)	(+3.2)	(+1.38)
	Number	2.1037	2.0133	2.0610	2.3570	2.3426	2.3228	2.2000
GPT-4	✗	89.3	97.8	93.1	90.5	94.9	77.5	90.51
	✓	89.6	98.1	93.1	91.9	95.5	79.9	91.34
		(+0.3)	(+0.3)	(0.0)	(+1.4)	(+0.6)	(+2.4)	(+0.83)
	Number	2.0126	2.0033	2.0019	2.0700	2.0507	2.2913	2.0716

Table 8: Performance of Complex CoT with GPT-3.5-Turbo and GPT-4 on MATH dataset, employing greedy decoding. Number: The average interaction number with LLM. **Overall:** The results overall MATH subtopics [14].

	PHP	MATH Dataset							Overall
		InterAlgebra	Precalculus	Geometry	NumTheory	Probability	PreAlgebra	Algebra	
Previous SOTA[7]	✗	-	-	-	-	-	-	-	50.30
GPT-4 CoT[17]	✗	-	-	-	-	-	-	-	42.50
GPT-3.5-Turbo Complex CoT (Ours)	✗	14.6	16.8	22.3	33.4	29.7	53.8	49.1	34.12
	✓	17.1	16.1	25.4	35.1	33.7	57.7	51.1	36.50
		(+2.5)	(-0.7)	(+3.1)	(+1.7)	(+4.0)	(+3.9)	(+2.0)	(+2.38)
	Number	4.2746	3.9625	4.3361	3.8166	3.7594	3.1526	3.0716	3.6673
GPT-4 Complex CoT (Ours)	✗	23.4	26.7	36.5	49.6	53.1	71.6	70.8	50.36
	✓	26.3	29.8	41.9	55.7	56.3	73.8	74.3	53.90
		(+2.9)	(+3.1)	(+5.4)	(+6.1)	(+3.2)	(+2.2)	(+3.5)	(+3.54)
	Number	3.2414	3.2435	3.2233	3.1740	2.8122	2.3226	2.4726	2.8494

● Overall: PHP improves performance

😊 GPT-3.5-Turbo adheres to prompts to a lower degree

📈 GPT 4: less interactions used compared to GPT-3.5-Turbo

🔄 Recall: less interactions needed for better model

💡 Limitations:

🧐 TLDR: Rethinking until ‘convergence?’

😬 ‘Escaping’ wrong answers needs more rigorous examination

- Limitations inherited from CoT

Thinking backwards e.g. equations with unknowns

What is ‘symbolic’ reasoning for a LM?

- LMs answering MWPs without the question part? [2]

[2] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2080–2094, 2021. 2, 4

Paper 2: Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

 Background: We don't understand WHY CoT works well

In-context learning != fine-tuning?

Random or misleading labels can hurt models very little

 Contribution:

Validity of demonstrations play a **small** role in reasoning quality

The relevance (to query) and ordering of the reasoning steps are more important for performance

CoT might be drawing on the models' inherent reasoning ability, rather than teaching it that!

🏞️ Background: Components of CoT Rationale

👣 CoT Rationale decomposed into:

🌉 Bridging Objects: LM need to traverse these to get to correct answer (e.g. numbers in a MWP)

💬 Language templates: contextual hints, relationships, and predicates that can help the models' reasoning

🤔 Do the ground truth of these components matter?

If not, what does to for LLM reasoning?

💡 Contribution: Experiment Setup (Ablation on Validity)

🐰 Multi-step reasoning:

📊 Arithmetic

👩 Multi-hop factual Question-Answering (QA)

🪑 Benchmarks: those that saw CoT boosting performance a lot

GSM8K for Arithmetic, Bamboogle for QA

Models: InstructGPT-175B2, text-davinci-002, text-davinci-003, PaLM, and Flan-PaLM

📏 Metrics:

Extrinsic: is answer correct?

Intrinsic: Recall/F1 of Bridging Objects

👉 all-correct-except-last-step reasoning would have 0 extrinsic score

Contribution: Invalid reasoning

 Humans wrote rationales for invalid reasoning in CoT's in-context demonstration

Keep premise the same

Subsequent steps altered to NOT produce correct answer

These are drastic changes...

...in contrast to adversarial perturbations, which would be making minimal changes that also make the steps lead to the wrong answer

💡 Contribution: CoT & Invalid Reasoning

😞 LLMs achieve 80%~90% accuracy inference time, even when they are given demonstrations with Invalid Reasoning

CoT

Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39.

Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of $12 + 24 = 36$ pages. Now she has $120 - 36 = 84$ pages left. Since she wants to read half of the remaining pages, she should read $84 / 2 = 42$ pages. The answer is 42. ✓

Invalid Reasoning

Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total. The answer is 39.

Yesterday, Julie read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. She wants to read half of the remaining pages tomorrow, so she needs to read $84 / 2 = 42$ pages tomorrow. The answer is 42. ✓

😞 For a significant portion of the samples:

CoT produces **wrong** answer...

...when Invalid Reasoning yields **correct** ones!

💡 **Contribution: Ablation on Validity**

👤 **Invalid Reasoning yields rationales very similar to CoT's**

When they answer correctly, the rationales are logically sound

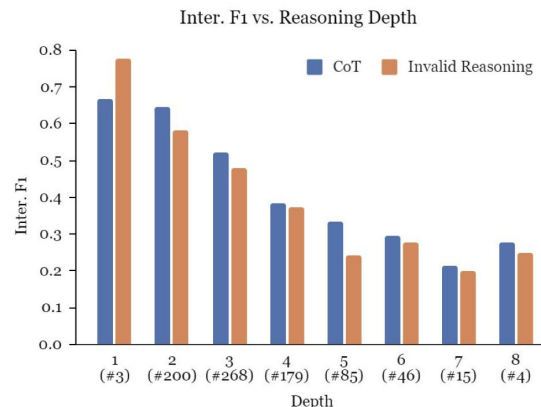
When they answer wrongly, the mistakes are similar to that when CoT answers wrongly

The distribution of these mistakes are similar

Inter. F1 is an intrinsic metric👉

Error Types	CoT correct & IR wrong	CoT wrong & IR correct
Calculation	20%	20%
One step missing	35%	25%
Semantic understanding	45%	55%

Table 3: Distribution of error types of 20 examples from GSM8K where Chain-of-Thought (CoT) prompting reaches the correct answer and the Invalid Reasoning setting (IR) reaches a wrong answer, and 20 examples for the opposite case.



💡 Contribution: Properties (“Aspects”) of the components of Rationale

🔄 Recall: components of rationale are:

🌉 Bridging Objects & 💬 Language templates

- Relevance: is it based on corresponding component of query?

Bridging Objects: being the exact same

Language Templates: belong to the same set, follow the same relationships

- Coherence: is it in the correct order (relative to others)?

E.g. introducing a concept before referring to it is NOT cohere!

 Contribution: Ablation on Rationale Components and their

Prompt Setting	Example Query (Arithmetic Reasoning) <i>Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?</i>	Example Query (Factual QA) <i>Who is the grandchild of Dambar Shah?</i>
STD (Standard prompting)	39	So the final answer is: Rudra Shah.
CoT (Chain-of-Thought)	Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
① Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. The Gorkha Kingdom was established by Prince Dravya Shah. Dravya Shah has a child named Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
② No coherence for bridging objects	Originally, Leah had $32 + 42 = 74$ chocolates and her sister had 32. So in total they had $74 - 35 = 39$. After eating 35, they had 42 pieces left in total. The answer is 39.	Krishna Shah was the father of Rudra Shah. Dambar Shah (? - 1645) was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
③ No relevance for bridging objects	Originally, Leah had 19 chocolates and her sister had 31. So in total they had $19 + 31 = 50$. After eating 29, they had $50 - 29 = 21$ pieces left in total. The answer is 21.	Metis Amando was the father of David Amando. Randall Amando was the child of David Amando. So the final answer (the name of the grandchild) is: Randall Amando.
④ No coherence for language templates	After eating 32, they had 42 pieces left in total. Originally, Leah had $32 + 42 = 74$ chocolates and her sister had 35. So in total they had $74 - 35 = 39$. The answer is 39.	Dambar Shah (? - 1645) was the child of Krishna Shah. Krishna Shah (? - 1661) was the father of Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
⑤ No relevance for language templates	Patricia needs to donate 32 inches, and wants her hair to be 42 inches long after the donation. Her hair is 35 inches long currently. Her hair needs to be $32 + 42 = 74$ inches long when she cuts it. So she needs to grow $74 - 35 = 39$ more inches. The answer is 39.	The husband of Dambar Shah (? - 1645) is Krishna Shah. Krishna Shah (? - 1661) has a brother called Rudra Shah. So the final answer (the name of the brother-in-law) is: Rudra Shah.
⑥ No coherence	After eating $32 + 42 = 74$, they had 32 pieces left in total. Originally, Leah had $74 - 35 = 39$ chocolates and her sister had 35. So in total they had 42. The answer is 39.	Krishna Shah was the child of Rudra Shah. Dambar Shah (? - 1645) was the father of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
⑦ No relevance	Patricia needs to donate 19 inches, and wants her hair to be 31 inches long after the donation. Her hair is 29 inches long currently. Her hair needs to be $19 + 31 = 50$ inc long when she cuts it. So she needs to grow $50 - 29 = 21$ more inches. The answer is 21.	The husband of Metis Amando is David Amando. David Amando has a brother called Randall Amando. So the final answer (the name of the brother-in-law) is: Randall Amando.

+ Additional scenarios:

1. Removing relevance: random substitution
2. Removing coherence: random shuffling

Contribution: Ablation on Rationale Components and their Properties cont.

Results:

1. Relevance and Coherence matters a lot for performance
2. Relevance is very important: removing it lead to the largest performance drop! (worse than standard prompting without rationale)
3. For Bridging objects, Relevance > Coherence: performance with incoherent bridging objects > that with irrelevant ones!
4. For Language templates, coherence matters

💡 Contribution: Insights on CoT

🤖 Validity plays a minor role in CoT performance

👉 Relevance (to query) and Coherence are more important

- CoT draws on and directs, rather than imbue, LLMs' reasoning ability

Models like text-davinci-003 and Flan-PaLM suffer **less** from the ablations when they have prior knowledge on the queried task

- CoT does **NOT** show that LLMs are few-shot learners in reasoning
But this does **NOT** preclude LLMs from in-context learning

💡 Limitations:

- Ablating on LLMs' prior knowledge using alternative Benchmarks

- Lack of variety of reasoning tasks

E.g. Symbolic and Commonsense reasoning

🤔 Neural Networks != 🧠

- Manual production of Invalid Reasoning samples

- More sophisticated Intrinsic metric

🔄 Recall: Intrinsic metrics are designed to give credit to getting the correct reasoning 🐾steps

🧠 Relevance for Bridging Objects: synonyms & equivalence?