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

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, Huan Sun

ACL 2023

発表者: 武田・笹野研 M2 高野春樹

概要

- Chain-of-Thoughtは推論過程を教えることで推論性能を上げる手法である
- しかし算術推論やQAにおいて、プロンプトに与える**推論過程の正しさ**は性能の小さな割合しか占めないことが分かった
- Chain-of-Thoughtにおいて大事なのは**推論の正しさ**ではなく**質問への関連性**と**推論ステップの順序**であることを示した。

Chain-of-Thought

- 言語モデルにタスクでの**推論過程**を示すことで性能を上げた手法
- 推論時に合わせて入力することでfine-tuningなしで学習（**In-Context Learning**）

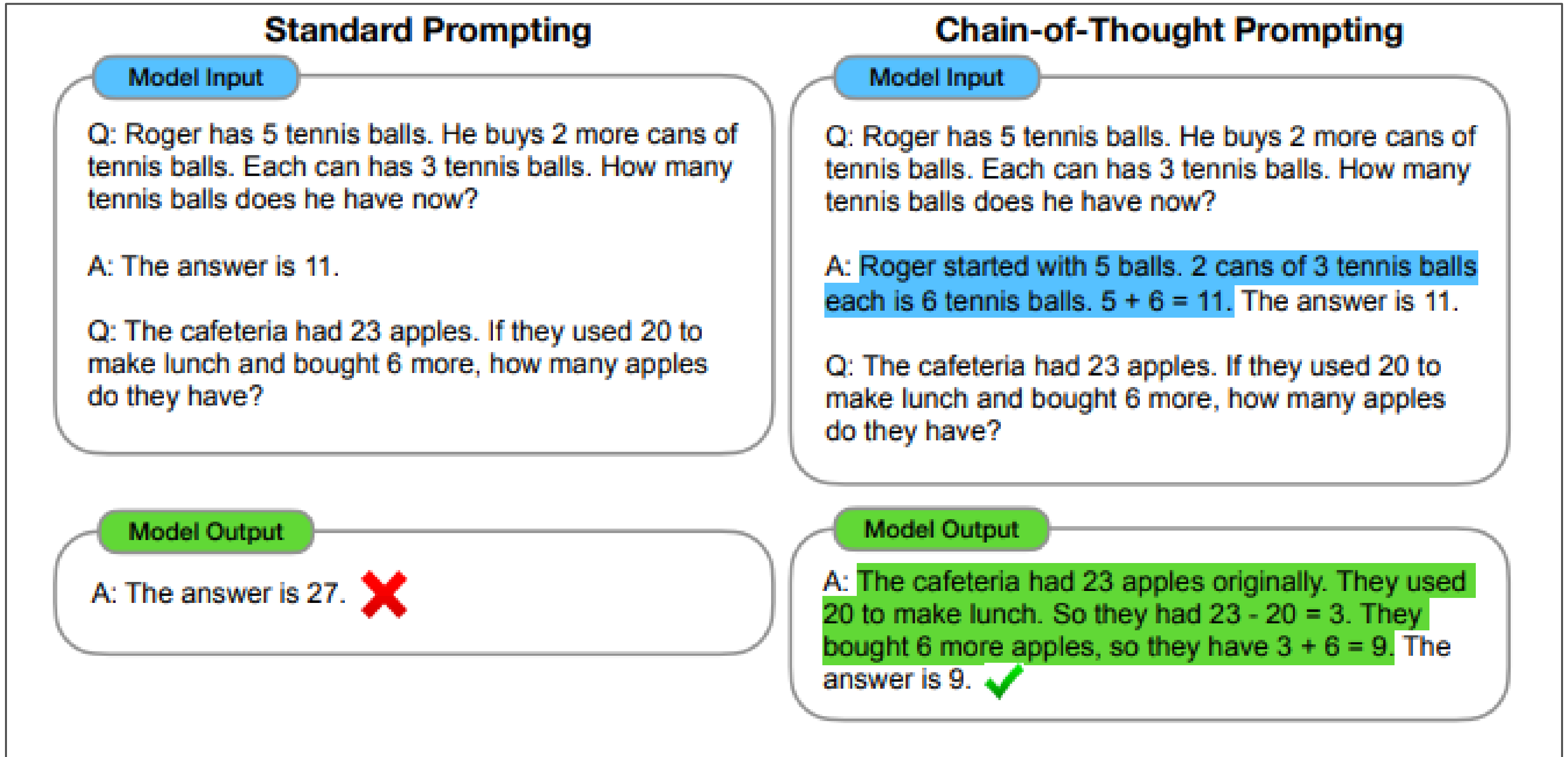

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

- しかしChain-of-Thoughtの何が性能に寄与しているかは不明瞭

定義, Bridging objects, Language template

- Chain-of-Thoughtの推論過程を2つの構成要素に分解

Bridging objects (青字部)

- 答えを出すのに重要な途中解
- 算術推論では数値
- QAでは主語や目的語

Language templates (赤字部)

- 答えを出すのに重要な推論ステップ
- 正しいBridging objectsを導くのための関係/述語

Arithmetic Reasoning	Multi-hop QA
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Q: Who is the grandchild of Dambar Shah?
A: 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.	A: 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.

Table 1: Bridging objects and language templates of a Chain-of-Thought rationale. Here we illustrate with one in-context exemplar for each task we experiment with.

問題提起

- Chain-of-Thoughtでは正しいBridging objects, Language templatesが与えられていた。

1. Bridging objects, Language templatesが正しいことは重要なのか
2. もし重要でないならば、Chain-of-Thoughtにはどのような観点が必要なのか

誤った推論過程

- Chain-of-Thoughtで直感的に重要に感じるのは**論理的に妥当で正しい推論**
- もし**誤った推論例**を与えると、何も与えない場合に比べて性能はわずかに向上するか、もしくは下がると予想できる

- オリジナルのChain-of-Thoughtの例から**誤った推論過程**を手作業で作成し、比較実験
  - （元の問題解決に全く役に立たない内容）

Prompt Setting	**Example Query (Arithmetic Reasoning)** *Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?*
STD (Standard prompting)	39
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.
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.

実験設定, 評価

- データセット
  - 算術推論: GSM8K (Chain-of-Thought内でも使用)
  - Multi-hop QA: Bamboogle

- 評価指標
  - Answer accuracy, Answer F1
    - 従来の評価では答えの正しさのみに注目
    - 推論が途中まで正しく、最後を間違えた場合: スコアは与えられない
    - 推論がほとんど正しくないが、答えのみ合っている場合: スコアが与えられる
  - Inter. Recall/F1
    - 答えのみではなく、途中のBridging objectsのRecall/F1でも推論の正しさを評価

実験結果

- モデルはInstructGPT-175B (text-davinci-002)

	GSM8K			Bamboogle	
	Inter. Recall	Inter. F1	Answer Acc.	Inter. Recall	Answer F1
STD (Standard prompting)	N/A	N/A	15.4	N/A	20.6
CoT (Chain-of-Thought prompting)	43.9	48.3	48.5	45.2	45.2
① Invalid Reasoning	39.8	43.9	39.5	44.4	39.4

- Inter.に限ると、元の性能の90%は維持
  - GSM8Kにおいて難易度における性能変化も元のChain-of-Thoughtと同様
- また、CoTが誤った答えを出し、Invalid Reasoningが正しい答えを出した割合も大きい

(両モデルの予測が異なる196件のうち62件)

誤った推論過程でもChain-of-Thoughtは有効

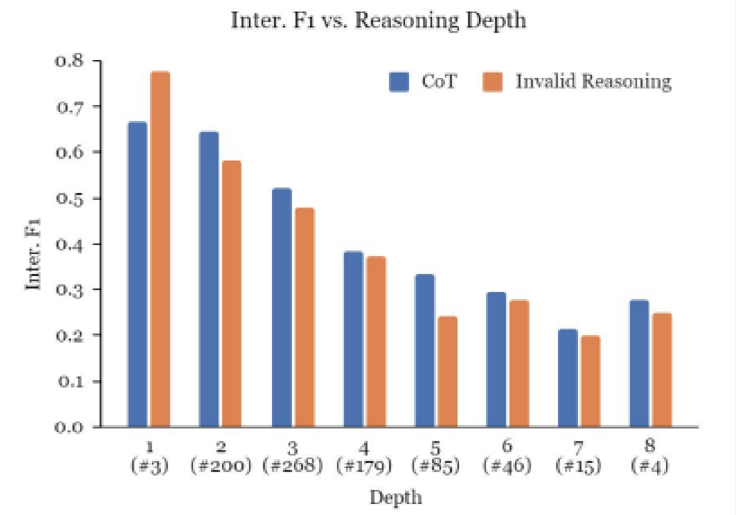

Figure 2: Model performance using CoT and demonstrations with invalid reasoning for examples with different reasoning depths on GSM8K. The number of samples for each reasoning depth is shown below (led by "#"). The performance drop is consistent across different levels of difficulty.

ではChain-of-Thoughtで何が重要か

- もし正しい推論が必要ではないなら、Chain-of-Thoughtの効果の鍵となる観点は何か
- Invalid ReasoningでもChain-of-Thoughtでも**共通する部分**があった

関連性 (Relevance)

- 質問内容に推論が関連しているか
- Bridging objects: 必ず質問文内に存在するものから始め、
- Language templates: 質問のトピック(Leah, her sister, chocolateの関係)を維持していた

順序性・一貫性 (Coherence)

- 推論のステップの順序が妥当か
- Bridging objects: 必ず前のステップで求めたものを使い、
- Language templates: 前のステップの推論を用いて答えを出す方向に推論を進めていた

Chain-of-Thoughtの元のプロンプトから、これらを破壊することでAblationを実施

実験結果

- 表は関連性や順序性を破壊したプロンプトを用いて実験した結果

	GSM8K			Bamboogle	
	Inter. Recall	Inter. F1	Answer Acc.	Inter. Recall	Answer F1
STD (Standard prompting)	N/A	N/A	15.4	N/A	20.6
CoT (Chain-of-Thought prompting)	43.9	48.3	48.5	45.2	45.2
① Invalid Reasoning	39.8	43.9	39.5	44.4	39.4
② No *coherence* for bridging objects	35.3	39.2	35.8	40.8	37.4
③ No relevance for bridging objects	21.4	26.2	27.5	39.6	34.0
④ No *coherence* for language templates	24.1	28.3	25.8	35.2	32.1
⑤ No relevance for language templates	29.5	34.0	32.8	40.4	29.4
⑥ No *coherence*	25.2	29.4	23.1	39.6	33.8
⑦ No relevance	9.6	11.9	11.0	36.8	23.9

- 関連性と順序性はどちらも重要である

- 特に関連性は非常に重要な役割を持つ
  - ⑦関連性を除いたモデルは大きく性能が低下している (推論過程を与えないSTDより悪い結果)
  - 多くは推論過程に”cats and dogs”などを出力しており、事前学習のコーパスで頻繁に現れる算数のパターンではないかと考察

実験結果

	GSM8K			Bamboogle	
	Inter. Recall	Inter. F1	Answer Acc.	Inter. Recall	Answer F1
STD (Standard prompting)	N/A	N/A	15.4	N/A	20.6
CoT (Chain-of-Thought prompting)	43.9	48.3	48.5	45.2	45.2
① Invalid Reasoning	39.8	43.9	39.5	44.4	39.4
② No *coherence* for bridging objects	35.3	39.2	35.8	40.8	37.4
③ No relevance for bridging objects	21.4	26.2	27.5	39.6	34.0
④ No *coherence* for language templates	24.1	28.3	25.8	35.2	32.1
⑤ No relevance for language templates	29.5	34.0	32.8	40.4	29.4
⑥ No *coherence*	25.2	29.4	23.1	39.6	33.8
⑦ No relevance	9.6	11.9	11.0	36.8	23.9

■ Bridging objectsでは**関連性**がより重要

■ Language templatesでは**順序性**がより重要

- 割と直感に近い結果
  - 最初のBridging objectsを誤ると、誤った答えを生成する
  - Language templatesの順序が異なると、最終的な答えが求められない

考察

- 誤った推論の割に性能が良いことを考慮すると、Chain-of-Thoughtは言語モデルに解き方を例示してるのではなく、**既に事前学習で得た推論能力を引き出すような役割**があり、与える推論過程は出力形式/空間を絞るような役割に近い
- 誤った推論を与えて性能が良いことは、言語モデルが既にある知識を使えている良い例である一方、**誤った推論を生成するというタスクにおいては大きく失敗している**
  - In-Contextの有用な情報を無視して、事前知識を優先する懸念
- 結果的にChain-of-Thoughtは言語モデルを推論のgood few-shot learnerにしてるわけではない（既に事前学習コーパスが good reasonerにしている）

まとめ

- Chain-of-Thoughtは推論過程を教えることで推論性能を上げる手法である
- しかし算術推論やQAにおいて、プロンプトに与える**推論過程の正しさ**は性能の小さな割合しか占めないことが分かった
- Chain-of-Thoughtにおいて大事なのは**推論の正しさ**ではなく**質問への関連性**と**推論ステップの順序**であることを示した。

補足資料

補足① Relevance, Coherenceの破壊

- Relevance
  - Bridging objects: 質問内の数値をランダムな数値に置換
  - Language templates: Bridging objectsは残しつつ、その他を別の学習データで置換
    - GSM8Kでは同じ推論ステップ数のデータ同士ならBridging objectsの登場回数が等しいので可能

- Coherence
  - 元のプロンプト内でBridging objects, Language templatesの順番をそれぞれシャッフルしただけ
  - ただし, Language templatesではBridging objectsの順番は保持したままシャッフルしている

補足② Relevance, Coherenceの破壊例

Prompt Setting	Example Query (Arithmetic Reasoning) *Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?*	Example Query (Factual QA) *Who is the grandchild of Dambar Shah?*
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.

Table 4: Examples for all settings in our experiments.

補足③ その他モデルの実験結果

- Text-davinci-003

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.2	N/A	25.1
CoT (Chain-of-Thought prompting)	48.4	53.1	54.5	61.6	59.5
① Invalid Reasoning	50.2	53.5	51.5	60.8	56.4
② No *coherence* for bridging objects	46.5	51.5	50.4	59.2	55.2
③ No relevance for bridging objects	32.5	38.3	47.2	60.4	56.9
④ No *coherence* for language templates	37.8	43.3	41.9	57.2	51.4
⑤ No relevance for language templates	44.6	49.9	51.8	62.4	59.3
⑥ No *coherence*	34.5	39.4	31.0	57.6	55.2
⑦ No relevance	15.5	17.8	16.2	50.0	49.0

Table 6: Intrinsic and extrinsic evaluation results under `text-davinci-003` for all settings. Discussions are included in Appendix A.3.

- PaLM

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.0	N/A	31.0
CoT (Chain-of-Thought prompting)	36.6	40.6	37.0	54.0	54.8
① Invalid Reasoning	33.9	36.9	31.8	50.4	46.1
② No *coherence* for bridging objects	30.3	35.0	33.5	33.6	25.7
③ No relevance for bridging objects	15.5	20.1	21.2	47.2	47.7
④ No *coherence* for language templates	23.1	27.3	21.9	40.4	35.5
⑤ No relevance for language templates	19.5	22.9	20.4	38.4	30.6
⑥ No *coherence*	23.9	28.3	24.1	39.6	33.6
⑦ No relevance	12.1	16.4	16.4	28.4	14.3

Table 8: Intrinsic and extrinsic evaluation results under PaLM. Discussions are included in Appendix A.3.

- Flan-PaLM (instruction-tuned PaLM)

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	21.8	N/A	36.5
CoT (Chain-of-Thought prompting)	72.2	73.0	63.8	57.6	56.9
① Invalid Reasoning	71.8	72.6	64.4	55.6	52.8
② No *coherence* for bridging objects	72.1	72.9	65.8	51.6	49.3
③ No relevance for bridging objects	71.1	71.9	64.6	54.0	52.8
④ No *coherence* for language templates	71.6	72.2	63.9	54.0	52.0
⑤ No relevance for language templates	71.9	72.7	64.9	55.2	53.5
⑥ No *coherence*	71.7	72.5	64.2	54.4	54.0
⑦ No relevance	70.7	71.6	64.5	50.0	51.9

Table 7: Intrinsic and extrinsic evaluation results under Flan-PaLM (Chung et al., 2022), the instruction-tuned version of PaLM for all settings. Discussions are included in Appendix A.3.

- ここら辺のモデルはそもそも事前にタスクを知っているので性能低下が少ない
- 特にFlan-PaLMは両タスクを既に学習済みで、Ablationが全く効いていないのが分かる