

CS371N: Natural Language Processing

Lecture 19: Understanding GPT II: Text rationales, Chain-of-thought

Greg Durrett



Administrivia

- Independent project proposals due tomorrow
 - TACC allocation submitted, contact me next week for status
- Midterm back early next week, A4 back after
- Vote in next few days!



Recap: Zero-shot/Few-shot prompting

- Single unlabeled datapoint \mathbf{x} , want to predict label y

\mathbf{x} = *The movie's acting could've been better, but the visuals and directing were top-notch.*

- Wrap \mathbf{x} in a template we call a **verbalizer** \mathbf{v}

Review: The movie's acting could've been better, but the visuals and directing were top-notch.

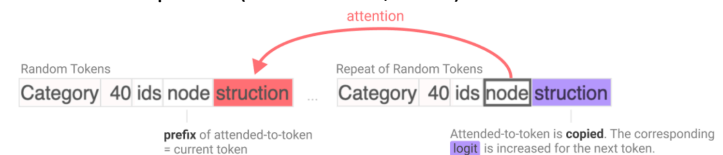
Out of positive, negative, or neutral, this review is GPT-3 *→ neutral*

- Need the right prompt (but there is a “plateau” of prompts that work)
- Few-shot: add one or more examples. Typically works better! Particularly with rich examples like we'll see today



Recap: Understanding ICL

- We can identify *induction heads* in Transformers; these emerge when ICL performance improves (Olsson et al., 2022)





Recap: Factuality

- ▶ We can identify *induction heads* in Transformers; these emerge when ICL performance improves (Olsson et al., 2022)

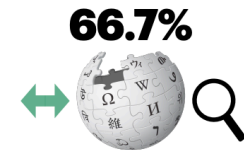


Step 3: Fact-checking

- ▶ Take things generated by models, split them into facts, and check them against verified sources

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗



This Lecture

- ▶ Text rationales: text explanations of answers
- ▶ Chain-of-thought prompting (zero- and few-shot)
- ▶ Chain-of-thought: extensions
- ▶ Analysis of explanations

Text Rationales



Explaining Decisions

this movie was **great!** would **watch again** **+**

the movie was **gross** and **overwrought**, but I **liked** it **+**

this movie was **not** really very **enjoyable** **-**

- How can we explain these decisions?



Explaining Decisions

GD "Barbie boasts impressive production design, costumes, and a pair of leads up to the task of bringing their respective dolls to life, but sadly comes up a little short when it comes to its screenplay."

Does this critic have a positive or negative view of Barbie? Explain your decision.

The critic has a mixed view of "Barbie."

Positive aspects mentioned:

1. Impressive production design
2. Impressive costumes
3. A pair of leads that are capable of bringing their respective dolls to life

Negative aspect mentioned:

1. Comes up a little short when it comes to its screenplay

The critic appreciates several elements of the production, notably the design, costumes, and the performances of the lead actors. However, they also express dissatisfaction with the screenplay. So, while there are elements they admire, there's also a significant point of criticism.



Explaining Decisions

I want to go to ____

Transformer model prediction: *Austin*

- How can we explain this decision?



Example from Vision

Laysan Albatross



Description: This is a large flying bird with black wings and a white belly.

Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

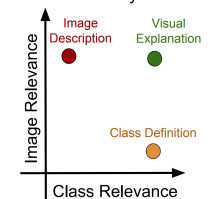


Description: This is a large bird with a white neck and a black back in the water.

Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

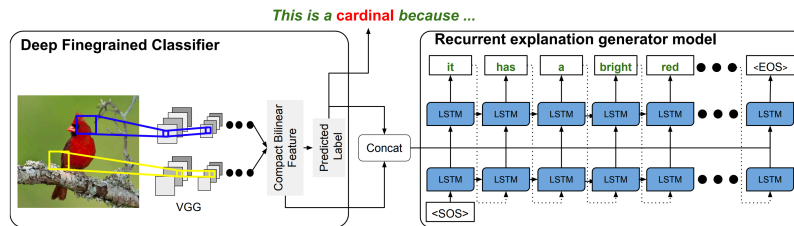
- What makes a visual explanation? Should be relevant to the class (output) and the image (input)
- Are these features *really* what the model used?



Hendricks et al. (2016)



Generating Explanations: Birds



- ▶ LSTM decoder looks at a feature vector and predicted label, then generates an explanation from those
- ▶ It's trained on human explanations — so it will likely produce explanations that look good (it learns to be a language model)

Hendricks et al. (2016)



E-SNLI

Premise: An adult dressed in black **holds a stick**.

Hypothesis: An adult is walking away, **empty-handed**.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.

Hypothesis: A young **mother** is playing with her **daughter** in a swing.

Label: neutral

Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A **man** in an orange vest **leans over a pickup truck**.

Hypothesis: A man is **touching** a truck.

Label: entailment

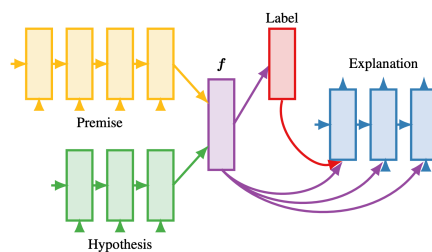
Explanation: Man leans over a pickup truck implies that he is touching it.

- ▶ Two formats: highlights and text

Camburu et al. (2019)



Generating Explanations: E-SNLI



f = function of premise and hypothesis vectors

- ▶ Similar to birds: explanation is conditioned on the label + network state f
- ▶ Information from f is fed into the explanation LSTM, although we don't know how that information is being used

Camburu et al. (2019)



Text Rationales

- ▶ Can we generate a natural language explanation of a model's behavior?
- ▶ What are some advantages to this?
 - ▶ Easy for untrained users to understand
 - ▶ Multitasking to produce human-written explanations may help us learn
- ▶ What are some risks/disadvantages?





Text Explanations

- Issues with text explanations:
 - Hard to produce/consume (these models are sort of clunky)
 - Hard to know if they faithfully reflect what a model is doing
 - More broadly, hard to evaluate
- However, writing such explanations comes naturally to us...so that means that they reflect some kind of underlying reasoning process that we're doing?
- Pre-2021: this process would usually be captured structurally in a model.
2022 and beyond: chain of thought

Chain-of-thought



Text rationales vs. programs

Problem 2:

Question: From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?

Options: A) $2/1223$ B) $1/122$ C) $1/221$ D) $3/1253$ E) $2/153$

Rationale: Let s be the sample space.

Then $n(s) = 52C2 = 1326$

E = event of getting 2 kings out of 4

$n(E) = 4C2 = 6$

$P(E) = 6/1326 = 1/221$

Answer is C

Correct Option: C

- Rationales are most useful for problems where some computation is required. They can articulate the intermediate steps needed to solve it
- Some of the earliest work: math word problems

Ling et al. (2017)



Chain-of-thought

- Chain-of-thought uses natural language as a scaffold for "reasoning"
- Unifies several ideas:
 - For math: relies on the fact that LLMs can do single steps of arithmetic okay. Builds on that to do multistep problems.
 - For QA: many problems involve reasoning decompositions
E.g., *What's the capital of the country where Aristotle lived?* ->
country = "*country where Aristotle lived*"
return *What's the capital of* [country]
 - For other tasks: capture the kinds of behavior written in rationales

Wei et al. (2022)



Chain-of-thought

- Typically a few-shot prompting technique where the in-context examples now contain explanations
- Answer is not generated in one go, but comes after an explanation that “talks through” the reasoning

Input: Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
 A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

...

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?
 A: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. So that is $10 \times .5 = 5$ hours a day. $5 \text{ hours a day} \times 7 \text{ days a week} = 35 \text{ hours a week}$. The answer is 35 hours a week. ✓

Model output: Wei et al. (2022)



Chain-of-thought

From our work: a synthetic test of multi-hop reasoning with extractive explanations:

Context: Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber.

Q: Who hangs out with a student?

A: Mary.

- What kind of explanation would you write here?

Explanation: because Mary hangs out with Danielle and Danielle is a student.

Ye and Durrett (NeurIPS 2022)



Chain-of-thought

Context: Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?
Mary

Standard few-shot learning, no explanation

Context: Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?
Mary, because Mary hangs out with Danielle and Danielle is a student.

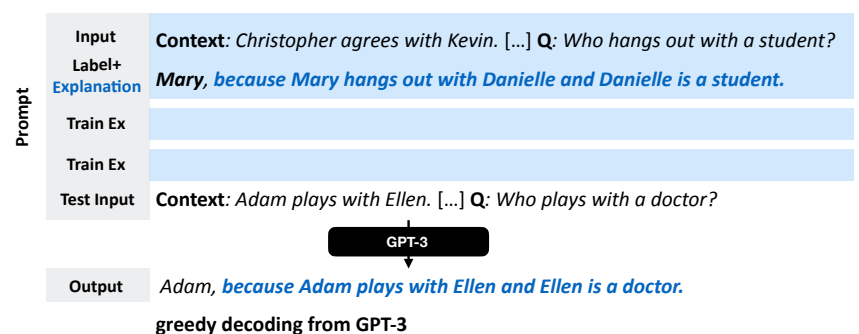
Predict-explain: answer is **not** conditioned on output explanation (original E-SNLI LSTM)

Context: Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?
Because Mary hangs out with Danielle and Danielle is a student, the answer is Mary.

Explain-predict: answer is conditioned on output explanation (Chain of Thought)



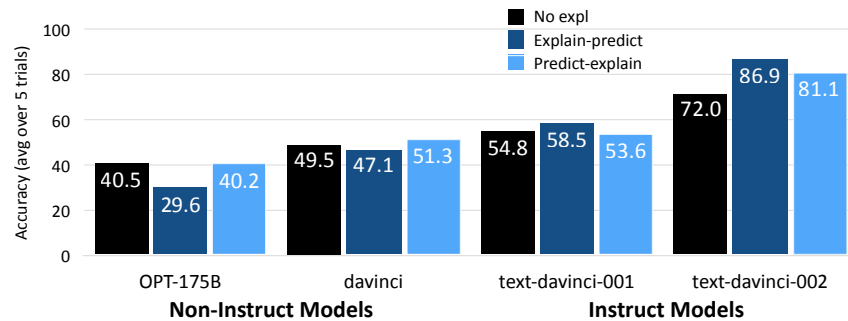
Chain-of-thought





Results

Results on SYNTH data



► Bigger, instruction-tuned models are far ahead of others on this task

Ye and Durrett (NeurIPS 2022)

Chain-of-thought extensions



Step-by-Step

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

- Prompt for step-by-step reasoning: produces chains of thought without including demonstrations
- Separate prompt to extract the answer ("Therefore, the answer is ____")

Kojima et al. (2022)



Step-by-Step

| | Arithmetic | | | | | |
|---------------|-------------------|-------------------|-----------------------|------------------|-----------------------|---------------------|
| | SingleEq | AddSub | MultiArith | GSM8K | AQUA | SVAMP |
| zero-shot | 74.6/ 78.7 | 72.2/77.0 | 17.7/22.7 | 10.4/12.5 | 22.4/22.4 | 58.8/58.7 |
| zero-shot-cot | 78.0/78.7 | 69.6/74.7 | 78.7/79.3 | 40.7/40.5 | 33.5/31.9 | 62.1/63.7 |
| | Common Sense | | Other Reasoning Tasks | | Symbolic Reasoning | |
| | Common SenseQA | Strategy QA | Date Understand | Shuffled Objects | Last Letter (4 words) | Coin Flip (4 times) |
| zero-shot | 68.8/72.6 | 12.7/ 54.3 | 49.3/33.6 | 31.3/29.7 | 0.2/- | 12.8/53.8 |
| zero-shot-cot | 64.6/64.0 | 54.8/52.3 | 67.5/61.8 | 52.4/52.9 | 57.6/- | 91.4/87.8 |

► text-davinci-002 (~ChatGPT-style model)

Kojima et al. (2022)



Step-by-Step

| No. | Category | Template | Accuracy |
|-----|-------------|---|-------------|
| 1 | instructive | Let's think step by step. | 78.7 |
| 2 | | First, (*1) | 77.3 |
| 3 | | Let's think about this logically. | 74.5 |
| 4 | | Let's solve this problem by splitting it into steps. (*2) | 72.2 |
| 5 | | Let's be realistic and think step by step. | 70.8 |
| 6 | | Let's think like a detective step by step. | 70.3 |
| 7 | | Let's think | 57.5 |
| 8 | | Before we dive into the answer, | 55.7 |
| 9 | | The answer is after the proof. | 45.7 |
| 10 | misleading | Don't think. Just feel. | 18.8 |
| 11 | | Let's think step by step but reach an incorrect answer. | 18.7 |
| 12 | | Let's count the number of "a" in the question. | 16.7 |
| 13 | | By using the fact that the earth is round, | 9.3 |
| 14 | irrelevant | By the way, I found a good restaurant nearby. | 17.5 |
| 15 | | AbraKadabra! | 15.5 |
| 16 | | It's a beautiful day. | 13.1 |
| - | | (Zero-shot) | 17.7 |

Kojima et al. (2022)



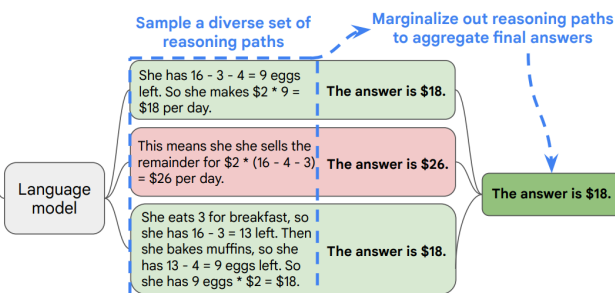
Demo: Step-by-Step (Math QA, StrategyQA)



Self-Consistency

Self-consistency

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.
...
Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?
A:



- Ensembling across multiple outputs (either zero-shot or few-shot)
- GSM8k: 56.5 -> 74.4, 5% gains on several other math datasets, lower gains on text tasks

Wang et al. (2022)



Program-aided Language Models

- For math: why are we doing the arithmetic in the LLM itself?
- Instead: generate code fragments and actually execute them to get an answer (how most earlier math word problem systems worked)
- Many flavors of this: "Faithful Chain-of-thought", "Program-of-thought", Toolformer, etc.

Model Output

A: The bakers started with 200 loaves
`loaves_baked = 200`
`They sold 93 in the morning and 39 in the afternoon`
`loaves_sold_morning = 93`
`loaves_sold_afternoon = 39`
`The grocery store returned 6 loaves.`
`loaves_returned = 6`
The answer is
`answer = loaves_baked - loaves_sold_morning`
`- loaves_sold_afternoon + loaves_returned`
`>>> print(answer)`
74



Gao et al. (2022)



Self-ask

- ▶ Similar idea but with QA/a search engine in the loop
- ▶ Demonstration shows sub-questions and sub-answers, can potentially do search at these intermediate points
- ▶ Bing's "Sydney" agent has some capabilities around this

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?

Are follow up questions needed here: Yes.

Follow up: How old was Theodor Haecker when he died?

Intermediate answer: Theodor Haecker was 65 years old when he died.

Follow up: How old was Harry Vaughan Watkins when he died?

Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.

So the final answer is: Harry Vaughan Watkins

Question: Who was president of the U.S. when superconductivity was discovered?

Are follow up questions needed here: Yes.

Follow up: When was superconductivity discovered?

Intermediate answer: Superconductivity was discovered in 1911.

Follow up: Who was president of the U.S. in 1911?

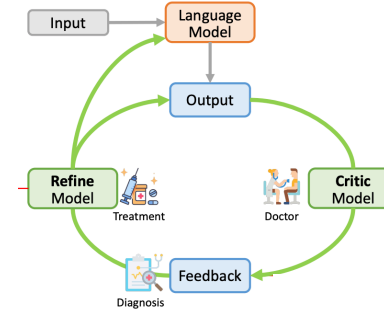
Intermediate answer: William Howard Taft.

So the final answer is: William Howard Taft.

Press et al. (2022)



Self-refinement



Liangming Pan et al. (2023)
Surveying the landscape of self-correction strategies



Demo: Self-correction



GD

What is 2613×12364 ?

GD

The product of 2,613 and 12,364 is 32,300,692.

GD

Can you verify if the product of 2,613 and 12,364 is 32,300,692?

GD

Let's do the multiplication:

css

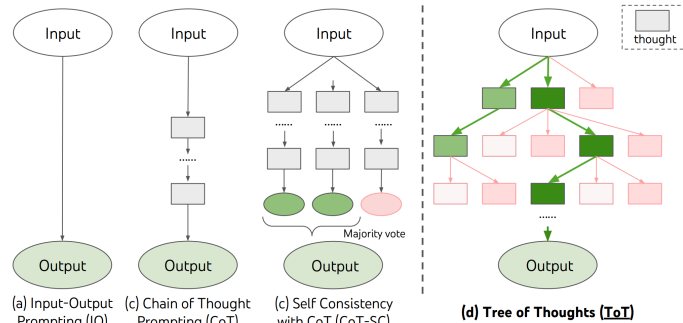
Copy code

```
2,613
× 12,364
-----
10,452 (2,613 × 4)
15,078 (2,613 × 6, shifted one place to the left)
26,130 (2,613 × 2, shifted two places to the left)
26,130 (2,613 × 1, shifted three places to the left)
-----
32,390,920
```

The correct product of 2,613 multiplied by 12,364 is 32,390,920. I apologize for the earlier oversight.



Tree-of-thought



- ▶ Combining LLMs with tree search
- ▶ Major challenge: what is the value function? How do you know to abandon this completion and pick another one?



Other ideas

- ▶ For math: can having various other ways of doing programmatic verification
- ▶ For natural language reasoning: missing component of search and planning, discussed in “Language Model Cascades”
- ▶ For problems like fact-checking or QA involving complex reasoning, its difficult to verify all of the individual steps...so if CoT goes wrong, it may even be hard for a human to spot

When does CoT help?



What does the literature report?



Aligning Large and Small Language Models via Chain-of-Thought Reasoning

Leonardo Ranaldi⁽¹⁾, André Freitas^(2,*)

⁽¹⁾Ulp Research Institute, Switzerland
⁽²⁾Department of Computer Science, University of Manchester, UK
 {firstname, lastname}@id.ox.ac.uk

Abstract

Chain-of-Thought (CoT) prompting improves the reasoning abilities of Large Language Models (LLMs), eliciting them to solve complex reasoning tasks in a step-wise manner. However, these abilities appear only in models with billions of parameters, which represent a very barrier for many users who are constrained to operate on a smaller model scale, i.e., Small Language Models (SLMs). Although many companies are releasing LLMs of the same family with fewer parameters, these models tend



Commonsense

| Task | Llama-2-7 | | Llama-2-13 | |
|------------|---------------|----------------------------|---------------|----------------------------|
| | Baseline | CoT | Baseline | CoT |
| OBQA | 53.6 \pm .2 | 49.5 \pm .3 \downarrow | 55.4 \pm .2 | 54.2 \pm .3 \downarrow |
| CSQA | 58.6 \pm .3 | 50.6 \pm .1 \downarrow | 63.4 \pm .2 | 60.8 \pm .2 \downarrow |
| SIQA | 46.5 \pm .2 | 45.3 \pm .3 | 48.3 \pm .4 | 46.9 \pm .3 |
| PIQA | 61.6 \pm .2 | 63.8 \pm .2 | 66.4 \pm .1 | 71.2 \pm .3 |
| GSM8K | 68.2 \pm .3 | 71.3 \pm .3 | 65.6 \pm .4 | 70.5 \pm .1 |
| MultiArith | 69.5 \pm .2 | 72.6 \pm .3 | 67.2 \pm .2 | 70.8 \pm .4 |

(Zayne Sprague et al., “To CoT or not to CoT”, arXiv 2024)



What does the literature report?



Aligning Large and Small Language Models
via Chain-of-Thought Reasoning

Leonardo Ranaldi¹, André Freitas^{1,2}

We reviewed NAACL,
EACL, and ICLR 2024
(4,642 papers)

Commonsense

| Task | Llama-2-7 | | Llama-2-13 | |
|------------|---------------|----------------------------|---------------|----------------------------|
| | Baseline | CoT | Baseline | CoT |
| OBQA | 53.6 \pm .2 | 49.5 \pm .3 \downarrow | 55.4 \pm .2 | 54.2 \pm .3 \downarrow |
| CSQA | 58.6 \pm .3 | 50.6 \pm .1 \downarrow | 63.4 \pm .2 | 60.8 \pm .2 \downarrow |
| SIQA | 46.5 \pm .2 | 45.3 \pm .3 | 48.3 \pm .4 | 46.9 \pm .3 |
| PIQA | 61.6 \pm .2 | 63.8 \pm .2 | 66.4 \pm .1 | 71.2 \pm .3 |
| GSM8K | 68.2 \pm .3 | 71.3 \pm .3 | 65.6 \pm .4 | 70.5 \pm .1 |
| MultiArith | 69.5 \pm .2 | 72.6 \pm .3 | 67.2 \pm .2 | 70.8 \pm .4 |

(Zayne Sprague et al., "To CoT or not to CoT", arXiv 2024)



What does the literature report?



Aligning Large and Small Language Models
via Chain-of-Thought Reasoning

Leonardo Ranaldi¹, André Freitas^{1,2}

We reviewed NAACL,
EACL, and ICLR 2024
(4,642 papers)

Commonsense

| Task | Llama-2-7 | | Llama-2-13 | |
|------------|---------------|----------------------------|---------------|----------------------------|
| | Baseline | CoT | Baseline | CoT |
| OBQA | 53.6 \pm .2 | 49.5 \pm .3 \downarrow | 55.4 \pm .2 | 54.2 \pm .3 \downarrow |
| CSQA | 58.6 \pm .3 | 50.6 \pm .1 \downarrow | 63.4 \pm .2 | 60.8 \pm .2 \downarrow |
| SIQA | 46.5 \pm .2 | 45.3 \pm .3 | 48.3 \pm .4 | 46.9 \pm .3 |
| PIQA | 61.6 \pm .2 | 63.8 \pm .2 | 66.4 \pm .1 | 71.2 \pm .3 |
| GSM8K | 68.2 \pm .3 | 71.3 \pm .3 | 65.6 \pm .4 | 70.5 \pm .1 |
| MultiArith | 69.5 \pm .2 | 72.6 \pm .3 | 67.2 \pm .2 | 70.8 \pm .4 |

(Zayne Sprague et al., "To CoT or not to CoT", arXiv 2024)



What does the literature report?



Aligning Large and Small Language Models
via Chain-of-Thought Reasoning

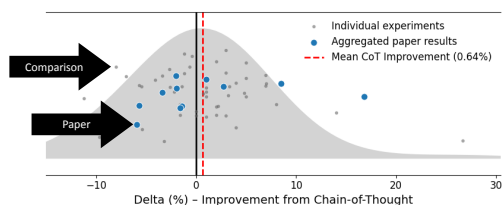
Leonardo Ranaldi¹, André Freitas^{1,2}

We reviewed NAACL,
EACL, and ICLR 2024
(4,642 papers)

Commonsense

| Task | Llama-2-7 | | Llama-2-13 | |
|------------|---------------|----------------------------|---------------|----------------------------|
| | Baseline | CoT | Baseline | CoT |
| OBQA | 53.6 \pm .2 | 49.5 \pm .3 \downarrow | 55.4 \pm .2 | 54.2 \pm .3 \downarrow |
| CSQA | 58.6 \pm .3 | 50.6 \pm .1 \downarrow | 63.4 \pm .2 | 60.8 \pm .2 \downarrow |
| SIQA | 46.5 \pm .2 | 45.3 \pm .3 | 48.3 \pm .4 | 46.9 \pm .3 |
| PIQA | 61.6 \pm .2 | 63.8 \pm .2 | 66.4 \pm .1 | 71.2 \pm .3 |
| GSM8K | 68.2 \pm .3 | 71.3 \pm .3 | 65.6 \pm .4 | 70.5 \pm .1 |
| MultiArith | 69.5 \pm .2 | 72.6 \pm .3 | 67.2 \pm .2 | 70.8 \pm .4 |

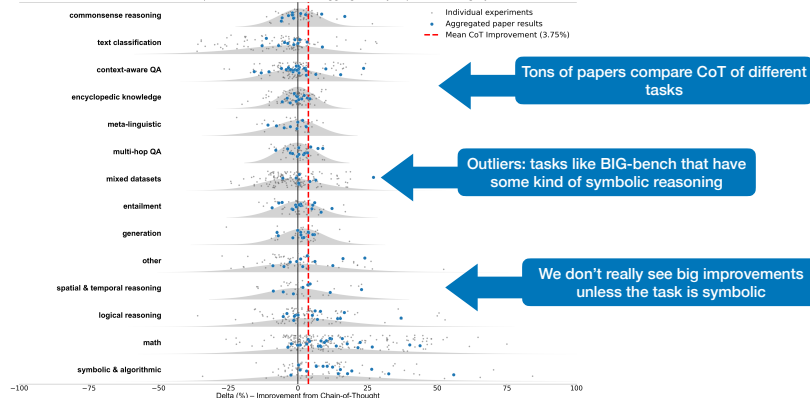
CoT Performance Improvement for Commonsense Reasoning



What does the literature report?

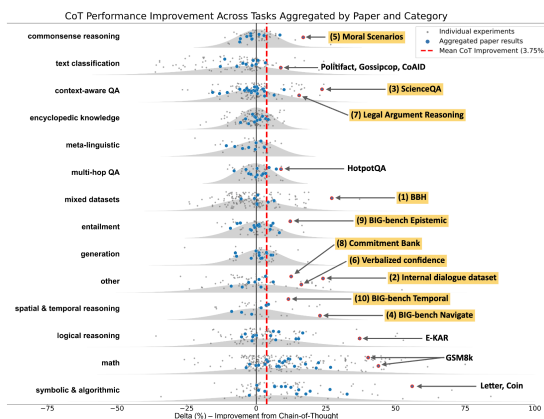


CoT Performance Improvement Across Tasks Aggregated by Paper and Category

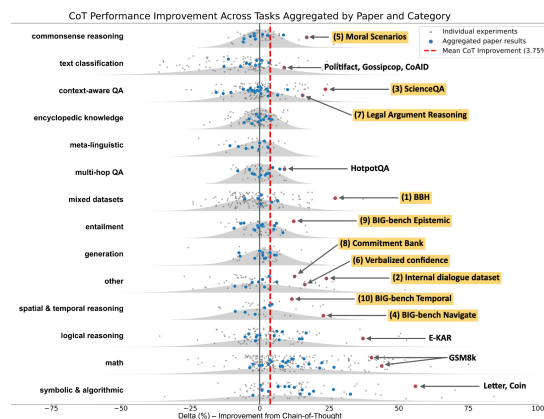




What does the literature report?



What does the literature report?

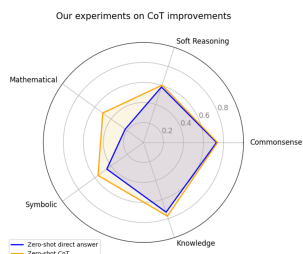


Most positive CoT results reported in the literature involve math or symbolic questions despite it being used in many domains.

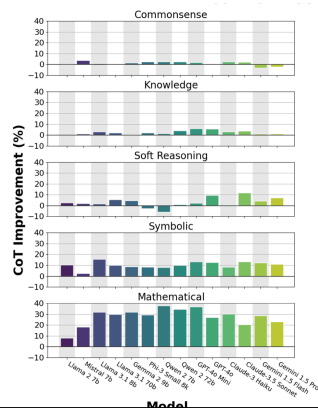


Analysis of Current Models

We ran **14 LLMs** on **20 different datasets** spanning areas in knowledge, soft, commonsense, symbolic, and mathematical reasoning

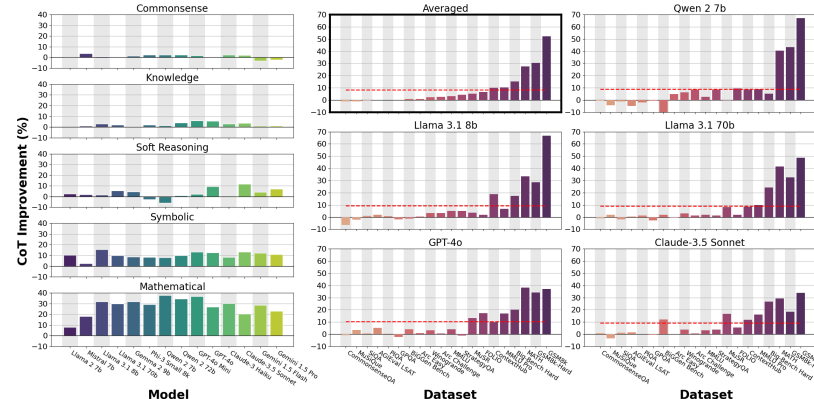


CoTs performance improvements are consistent across models, and only really helps on the symbolic and mathematical domains.



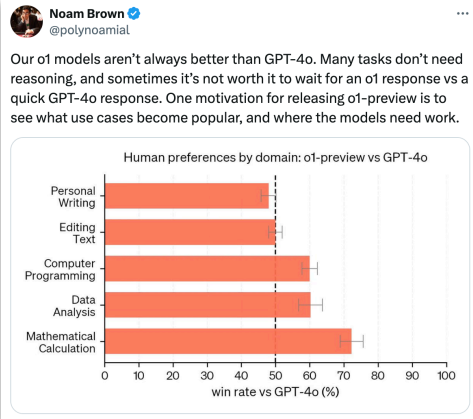
Analysis of Current models

CoT vs Direct Answer Prompting in Zero-Shot Setting





o1



Takeaways

- ▶ Chain-of-thought prompting (zero- and few-shot) can work well for tasks involving reasoning, especially mathematical reasoning and textual question answering with multiple steps
- ▶ Several things needed to improve them, such as self-consistency and the ability to use other resources like code execution or APIs
- ▶ Next time: RLHF, makes models better at zero-shot prompting and producing well-structured chain-of-thought responses