

# 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  $x$ , want to predict label  $y$

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

- ▶ Wrap  $x$  in a template we call a **verbalizer**  $v$

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

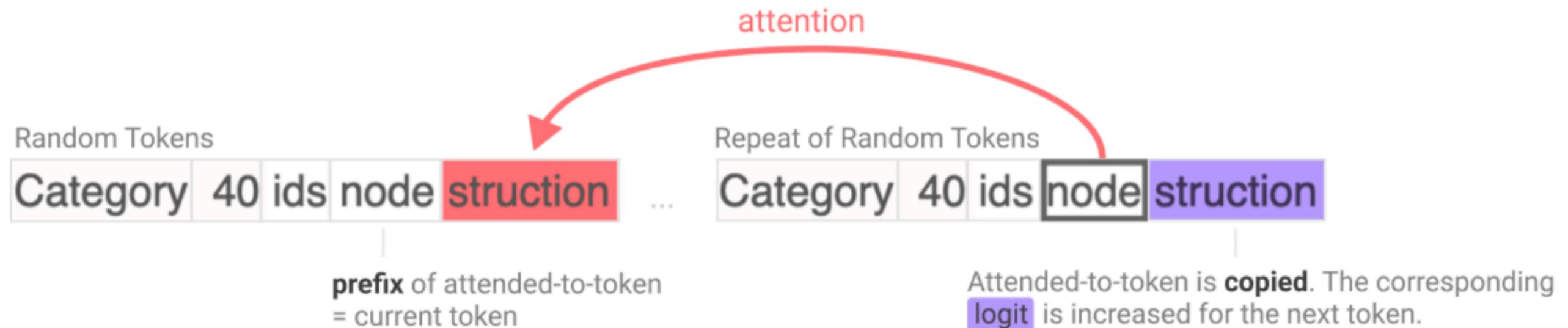
*Out of positive, negative, or neutral, this review is* — **GPT-3** → neutral

- ▶ 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. ✗

**66.7%**





# 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



"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

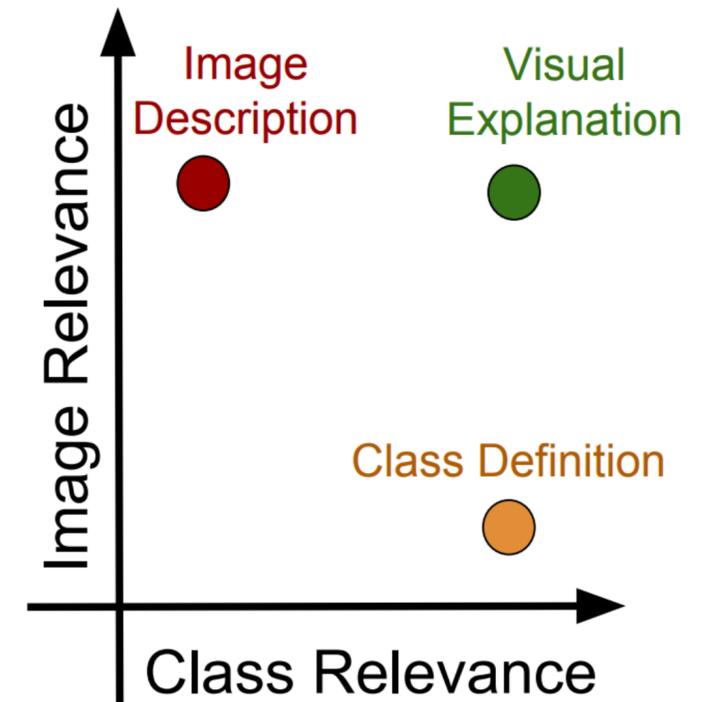


**Description:** This is a large bird with a white neck and a black back in the water.

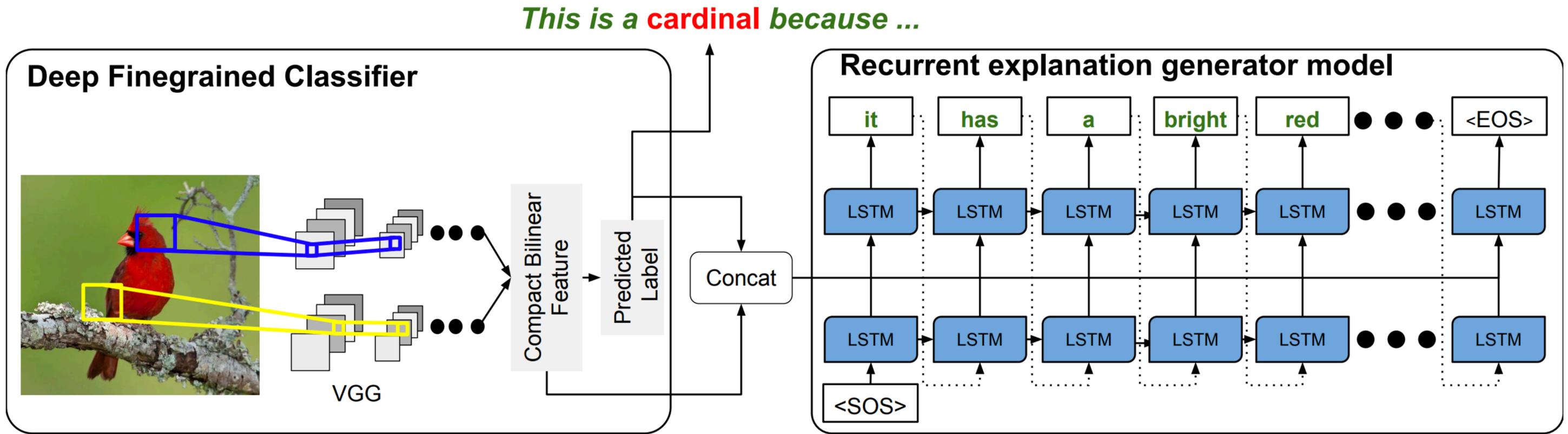
**Class Definition:** The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

**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?



# 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)



# 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

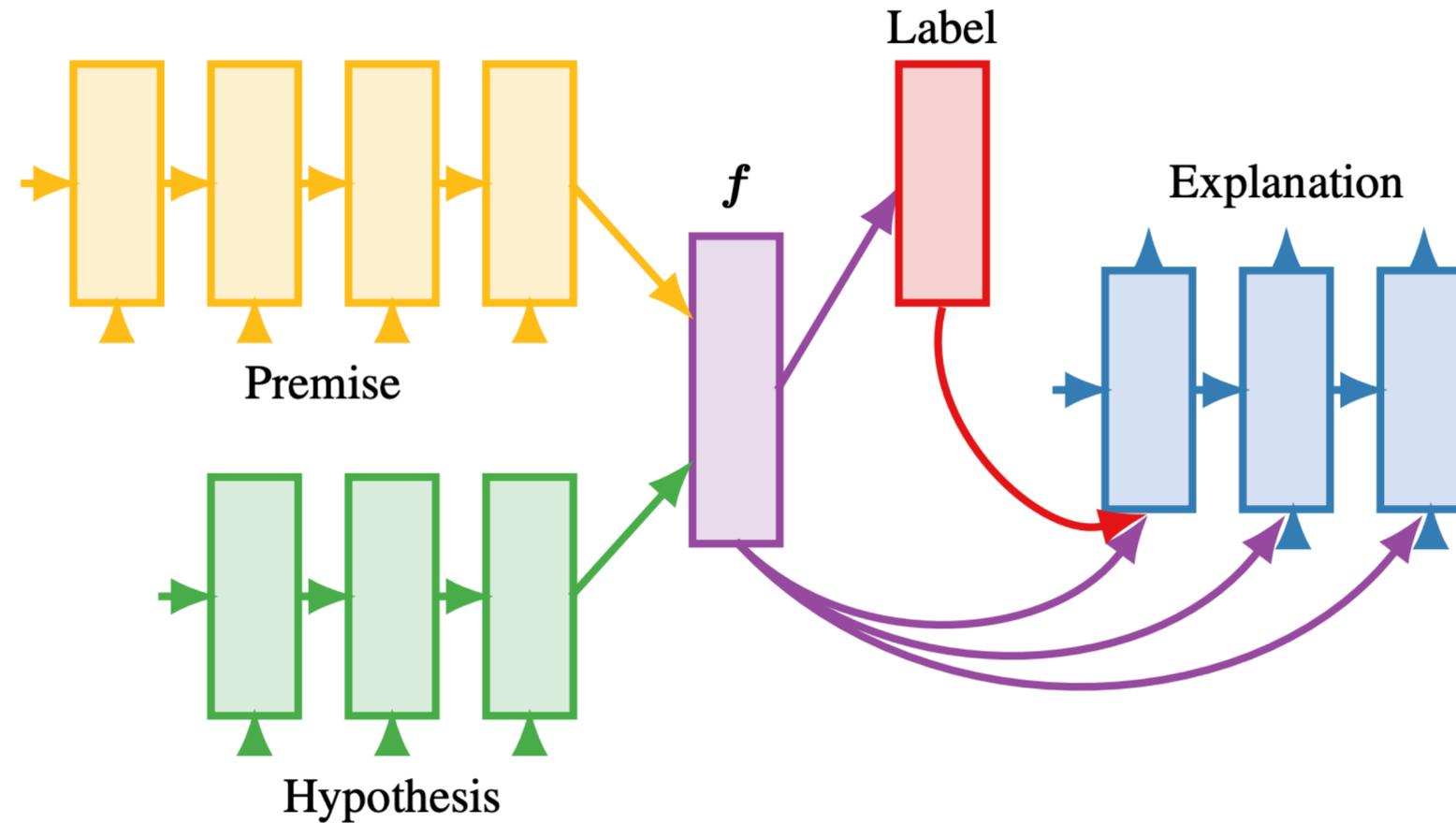
Explanation: Man leans over a pickup truck implies that he is touching it.

---

- ▶ Two formats: highlights and text



# Generating Explanations: E-SNLI



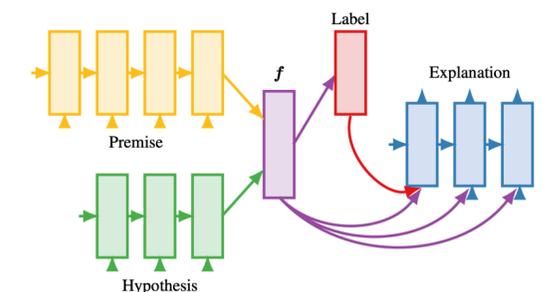
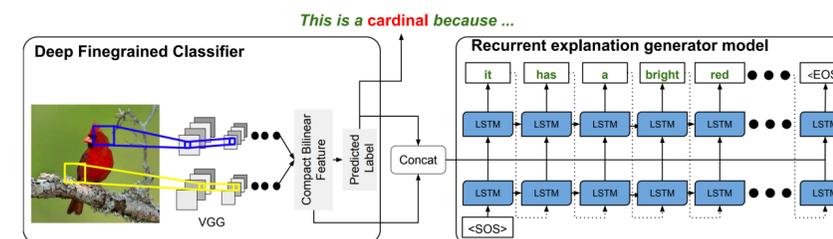
$f$  = function of premise and hypothesis vectors

- ▶ 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



# 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



# 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



# 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:

Model output:

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.  $5$  hours a day  $\times 7$  days a week = 35 hours a week. The answer is 35 hours a week. ✓



# 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.*



# 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

Prompt

Input	<b>Context:</b> <i>Christopher agrees with Kevin. [...]</i> <b>Q:</b> <i>Who hangs out with a student?</i>
Label+ Explanation	<i>Mary, because Mary hangs out with Danielle and Danielle is a student.</i>
Train Ex	
Train Ex	
Test Input	<b>Context:</b> <i>Adam plays with Ellen. [...]</i> <b>Q:</b> <i>Who plays with a doctor?</i>



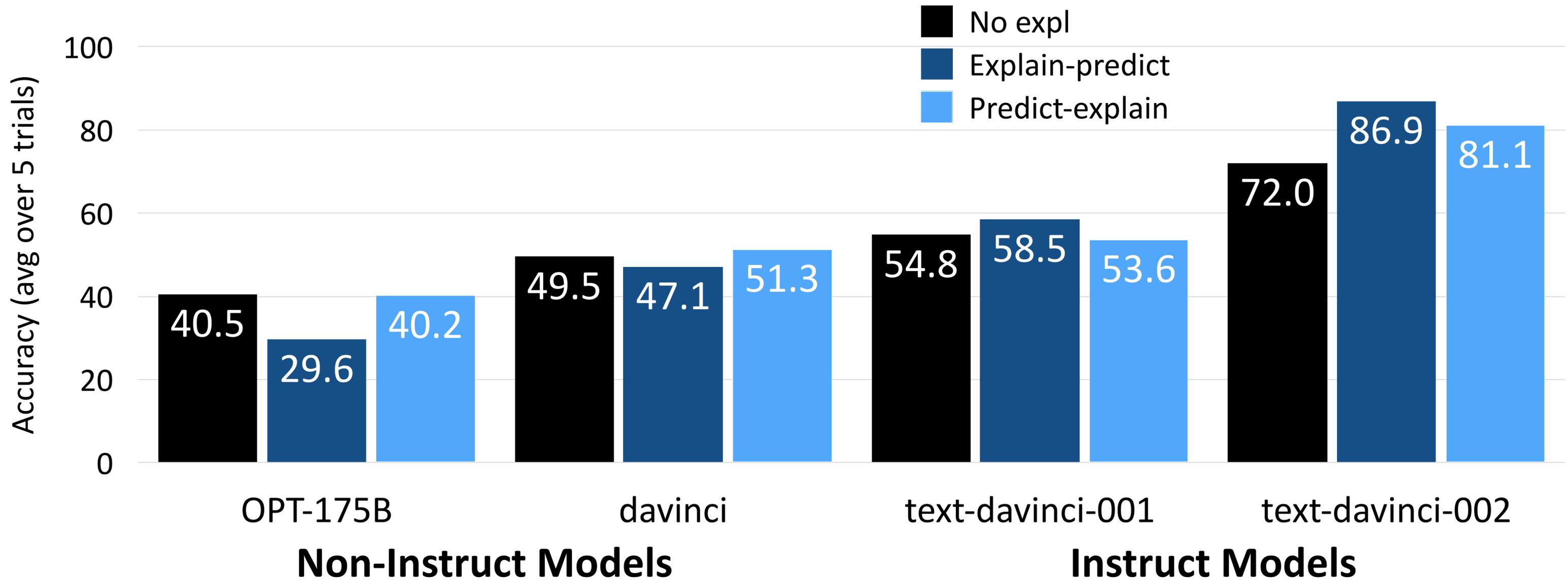
Output	<i>Adam, because Adam plays with Ellen and Ellen is a doctor.</i>
--------	---

**greedy decoding from GPT-3**



# Results

## Results on SYNTH data



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

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 \_\_\_\_”)



# Step-by-Step

	Arithmetic					
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
zero-shot	74.6/78.7	<b>72.2/77.0</b>	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
zero-shot-cot	<b>78.0/78.7</b>	69.6/74.7	<b>78.7/79.3</b>	<b>40.7/40.5</b>	<b>33.5/31.9</b>	<b>62.1/63.7</b>
	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	<b>68.8/72.6</b>	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	<b>54.8/52.3</b>	<b>67.5/61.8</b>	<b>52.4/52.9</b>	<b>57.6/-</b>	<b>91.4/87.8</b>

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



# Step-by-Step

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	<b>78.7</b>
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



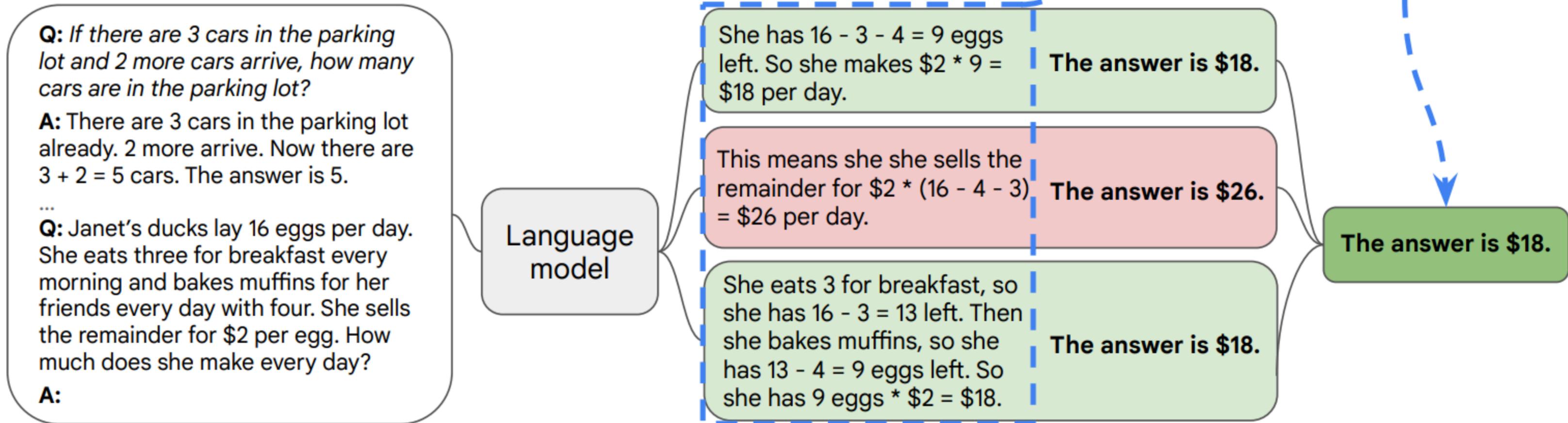
---

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



# Self-Consistency

## Self-consistency



- ▶ 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



# 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)
```

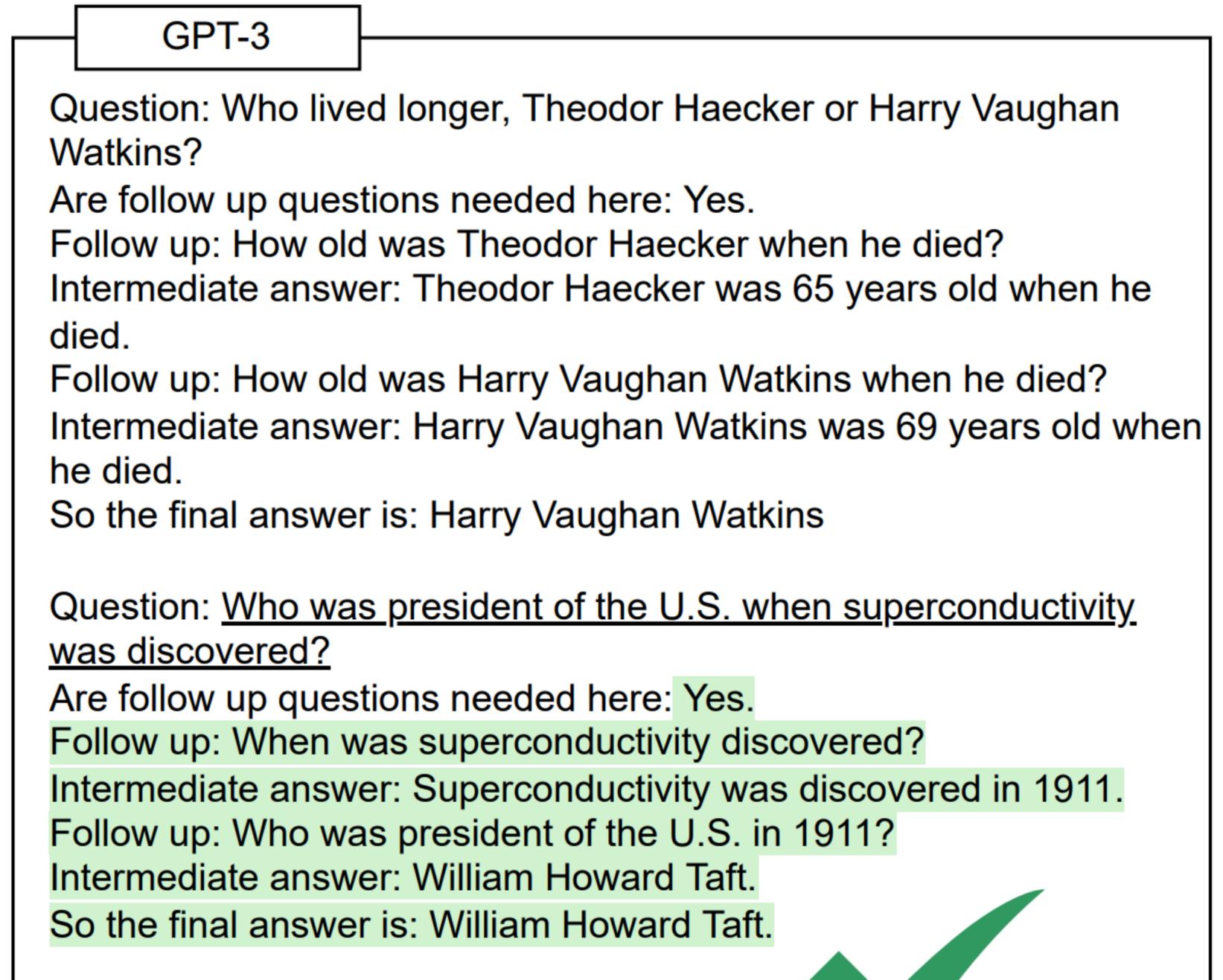
```
74
```





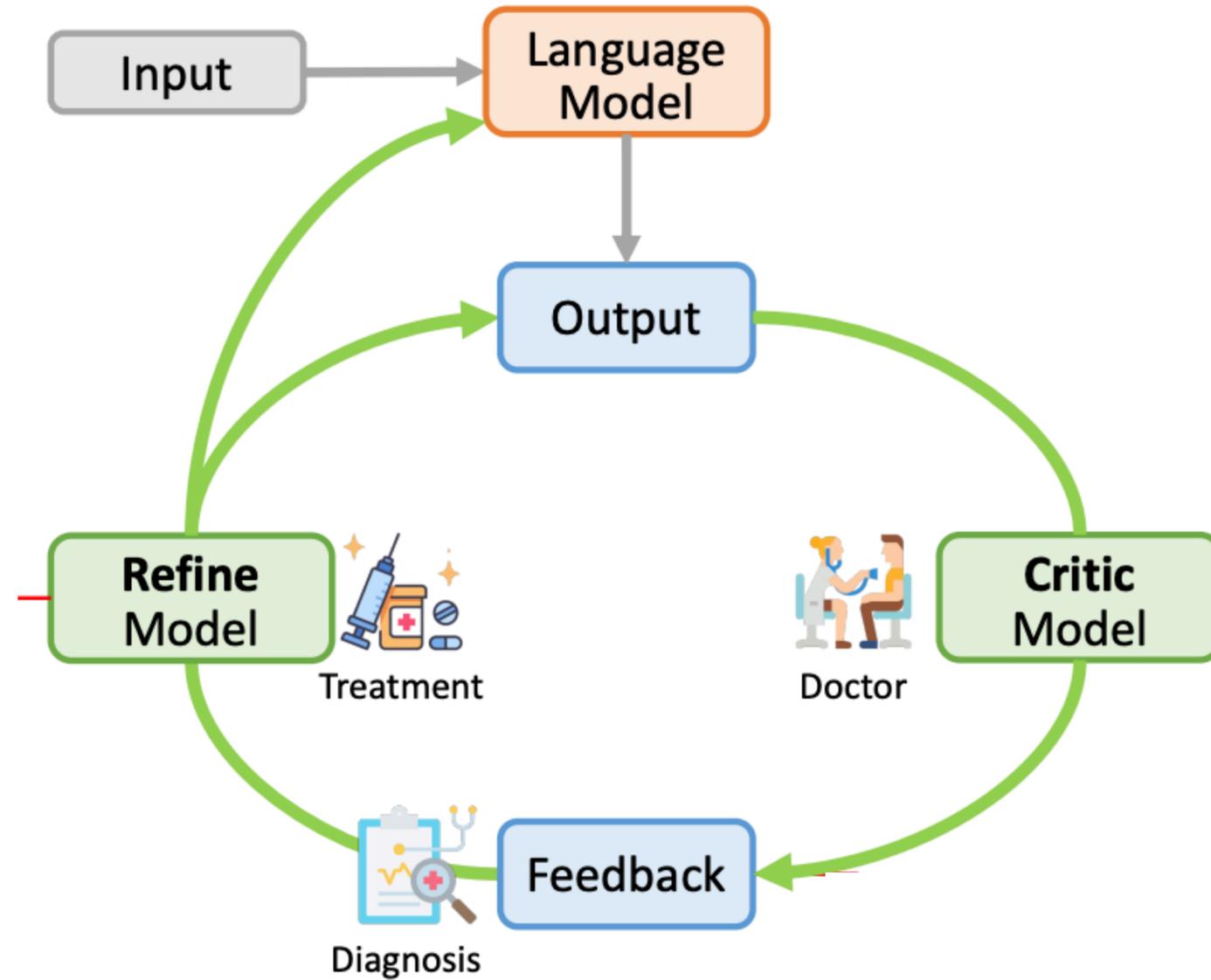
# Self-ask

- ▶ 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





# Self-refinement



Liangming Pan et al. (2023)

Surveying the landscape of self-correction strategies



---

# Demo: Self-correction



What is  $2613 * 12364$ ?



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



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



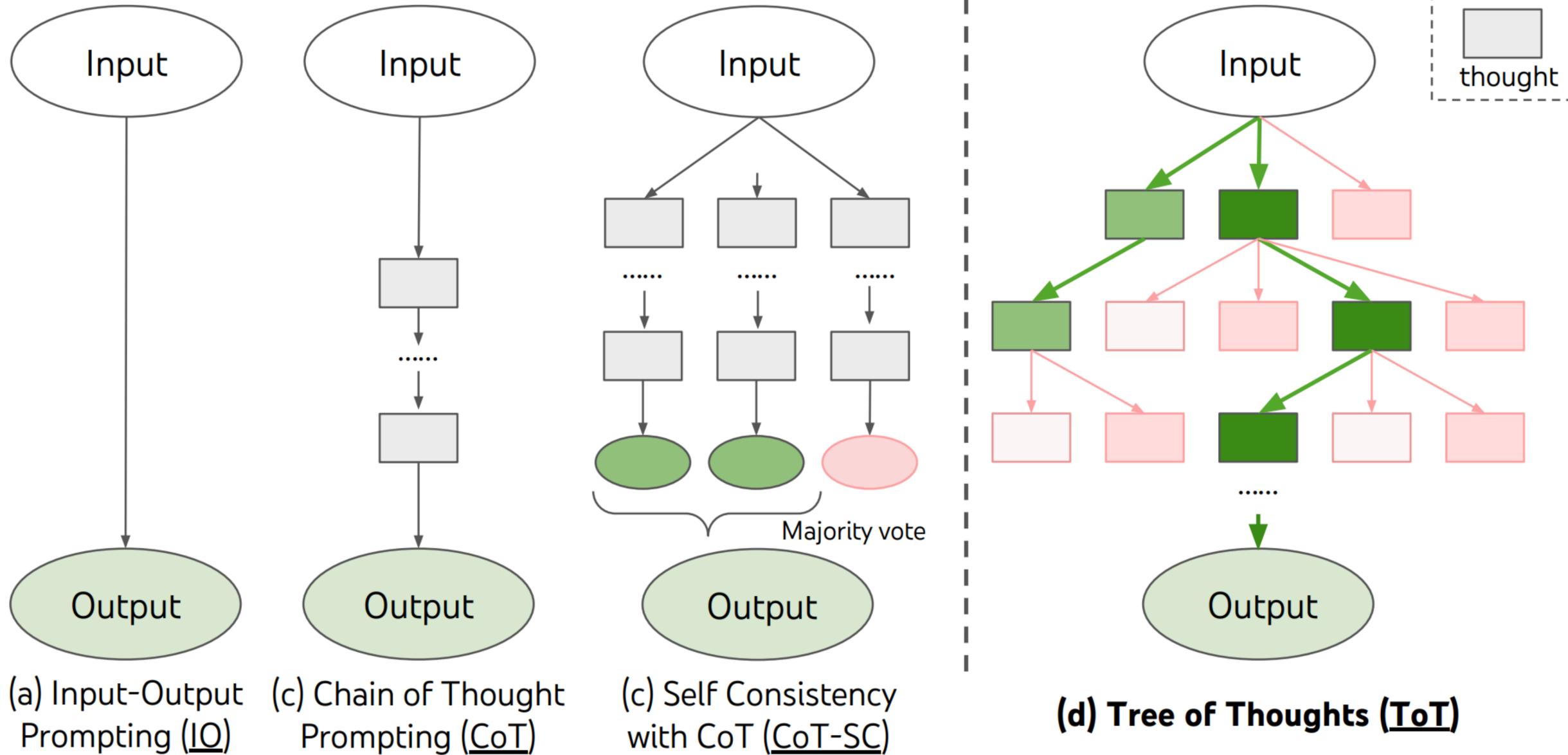
Let's do the multiplication:

```
CSS Copy code
  2,613
× 12,364
-----
 10,452 (2,613 × 4)
 15,678 (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?

EACL 2024

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

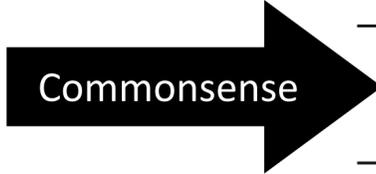
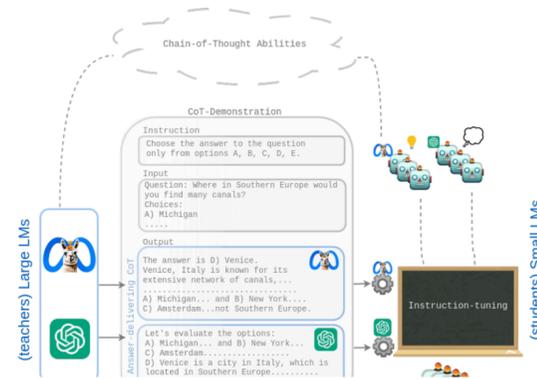
Leonardo Ranaldi<sup>(†)</sup>, André Freitas<sup>(†,\*)</sup>

<sup>(†)</sup> Idiap Research Institute, Switzerland

<sup>(\*)</sup> Department of Computer Science, University of Manchester, UK  
[firstname].[lastname]@idiap.ch

### Abstract

Chain-of-Thought (CoT) prompting empowers 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 an entry 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



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

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



# What does the literature report?

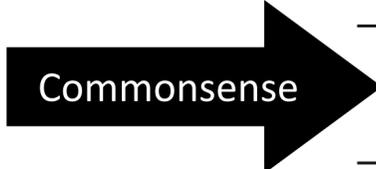
EACL 2024

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

Leonardo Ranaldi<sup>(†)</sup>, André Freitas<sup>(†,\*)</sup>  
<sup>(†)</sup> Idiap Research Institute, Switzerland

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

Chain-of-thought reasoning in LLMs...  
the reasoning capabilities of LLMs...  
reasoning tasks...  
ever, these abilities appear only in models with billions of parameters, which represent an entry 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



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



# What does the literature report?

EACL 2024

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

Leonardo Ranaldi<sup>(†)</sup>, André Freitas<sup>(†,\*)</sup>  
<sup>(†)</sup> Idiap Research Institute, Switzerland

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

Chain-of-thought reasoning in LLMs...  
the reasoning capabilities of LLMs...  
reasoning tasks...  
ever, these abilities appear only in models with billions of parameters, which represent an entry 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



Task	Llama-2-7		Llama-2-13	
	Direct	CoT	Direct	CoT
OBQA				54.2±.3↓
CSQA				60.8±.2↓
SIQA				46.9±.3
PIQA				71.2±.3
GSM8K	68.2±.1	71.8±.1	68.8±.1	70.5±.1
MultiArith	69.5±.2	72.6±.3	67.2±.2	70.8±.4

516 mention CoT, with 110 papers having a CoT vs direct answer comparisons



# What does the literature report?

EACL 2024

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

Leonardo Ranaldi (†), André Freitas (†,\*)  
(†) Idiap Research Institute, Switzerland

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

Chain-of-thought reasoning in LLMs... ever, these abilities appear only in models with billions of parameters, which represent an entry 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

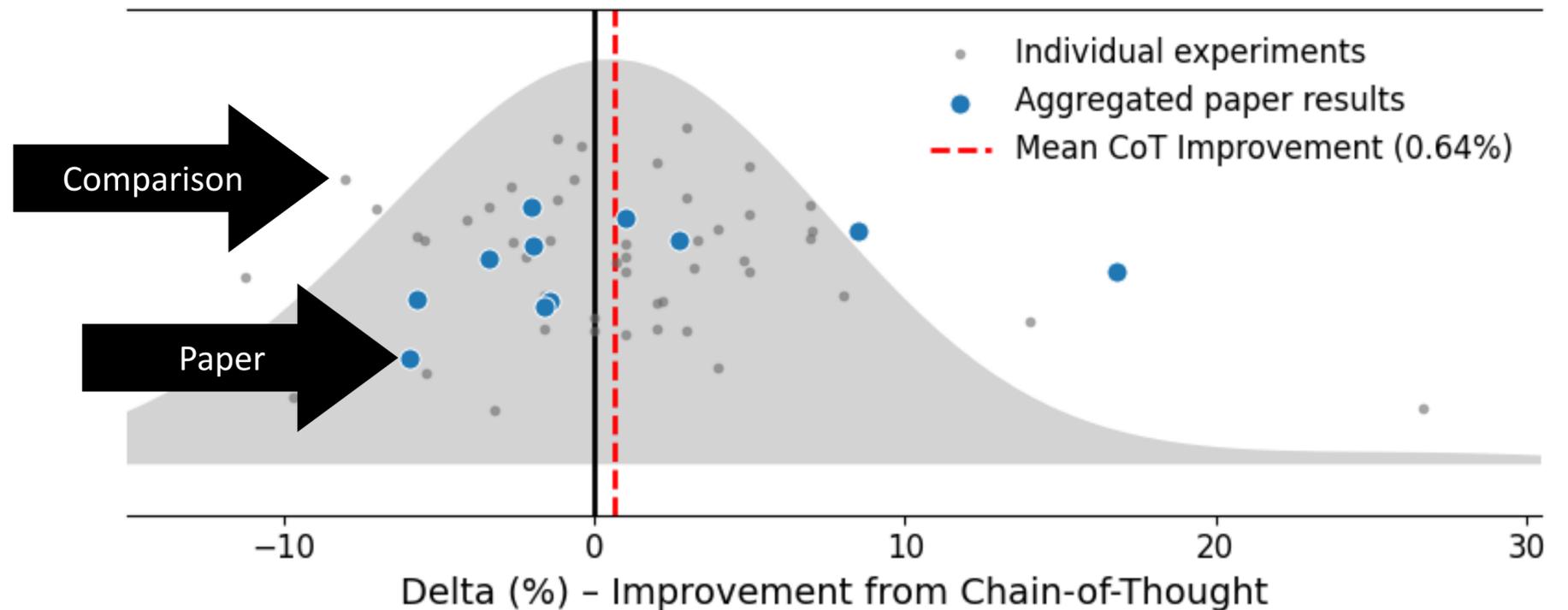


Commonsense

Task	Llama-2-7		Llama-2-13	
	Direct	CoT	Direct	CoT
OBQA	54.2±.3	54.2±.3↓	54.2±.3	54.2±.3↓
CSQA	60.8±.2	60.8±.2↓	60.8±.2	60.8±.2↓
SIQA	46.9±.3	46.9±.3	46.9±.3	46.9±.3
PIQA	71.2±.3	71.2±.3	71.2±.3	71.2±.3
GSM8K	68.2±.1	71.8±.1	68.8±.1	70.5±.1
MultiArith	69.5±.2	72.6±.3	67.2±.2	70.8±.4

516 mention CoT, with 110 papers having a CoT vs direct answer comparisons

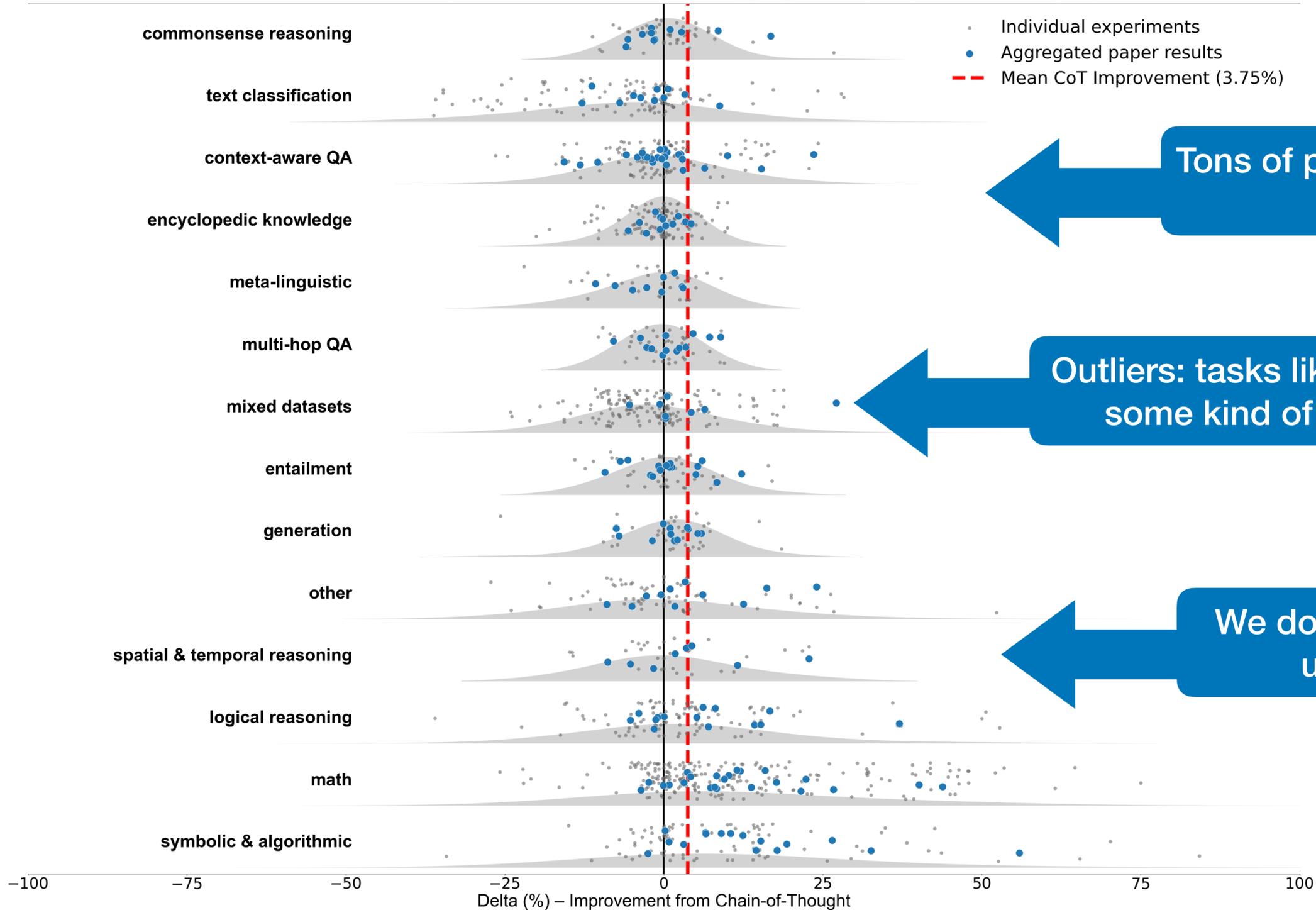
CoT Performance Improvement for Commonsense Reasoning





# What does the literature report?

CoT Performance Improvement Across Tasks Aggregated by Paper and Category



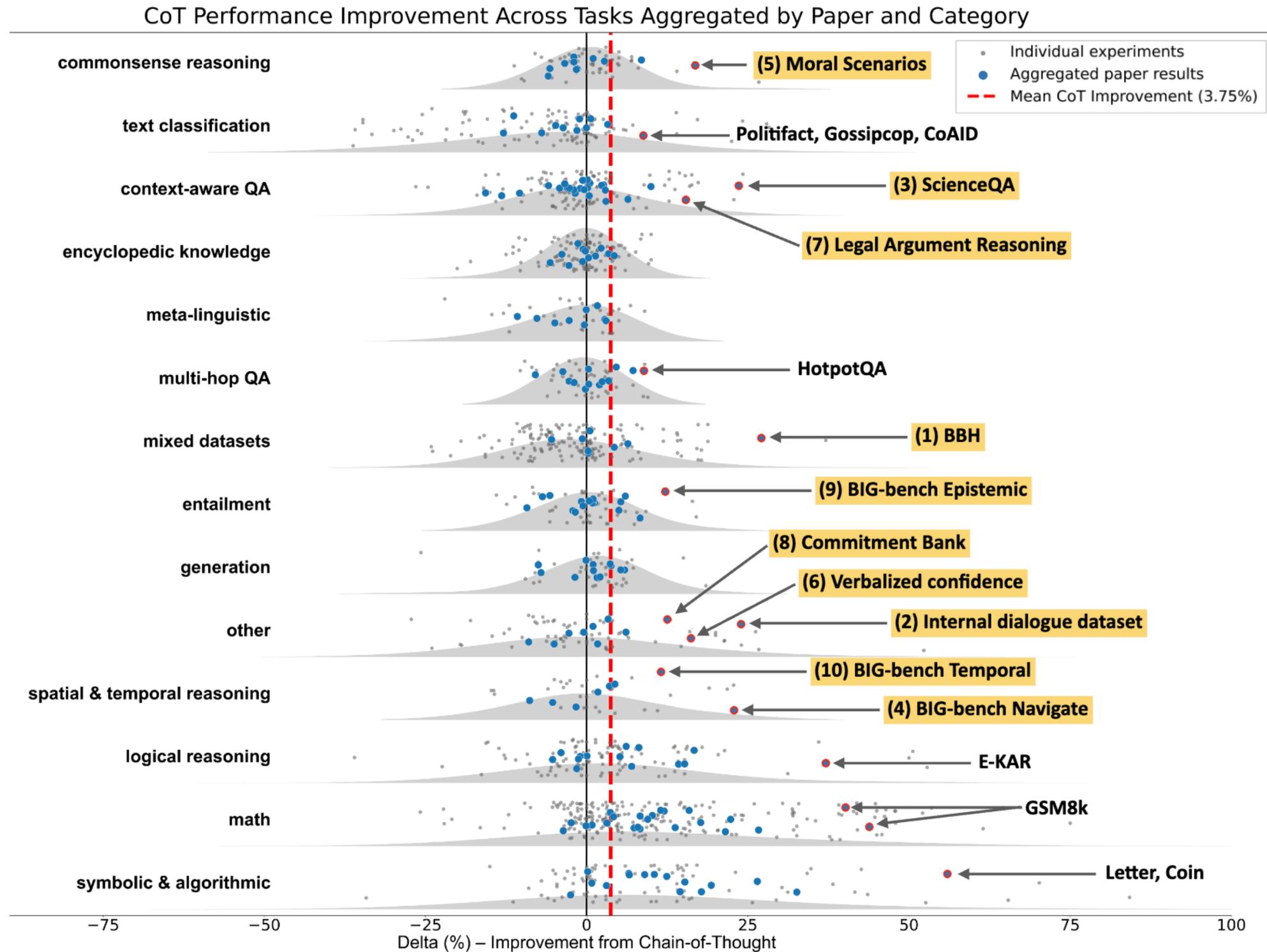
Tons of papers compare CoT of different tasks

Outliers: tasks like BIG-bench that have some kind of symbolic reasoning

We don't really see big improvements unless the task is symbolic

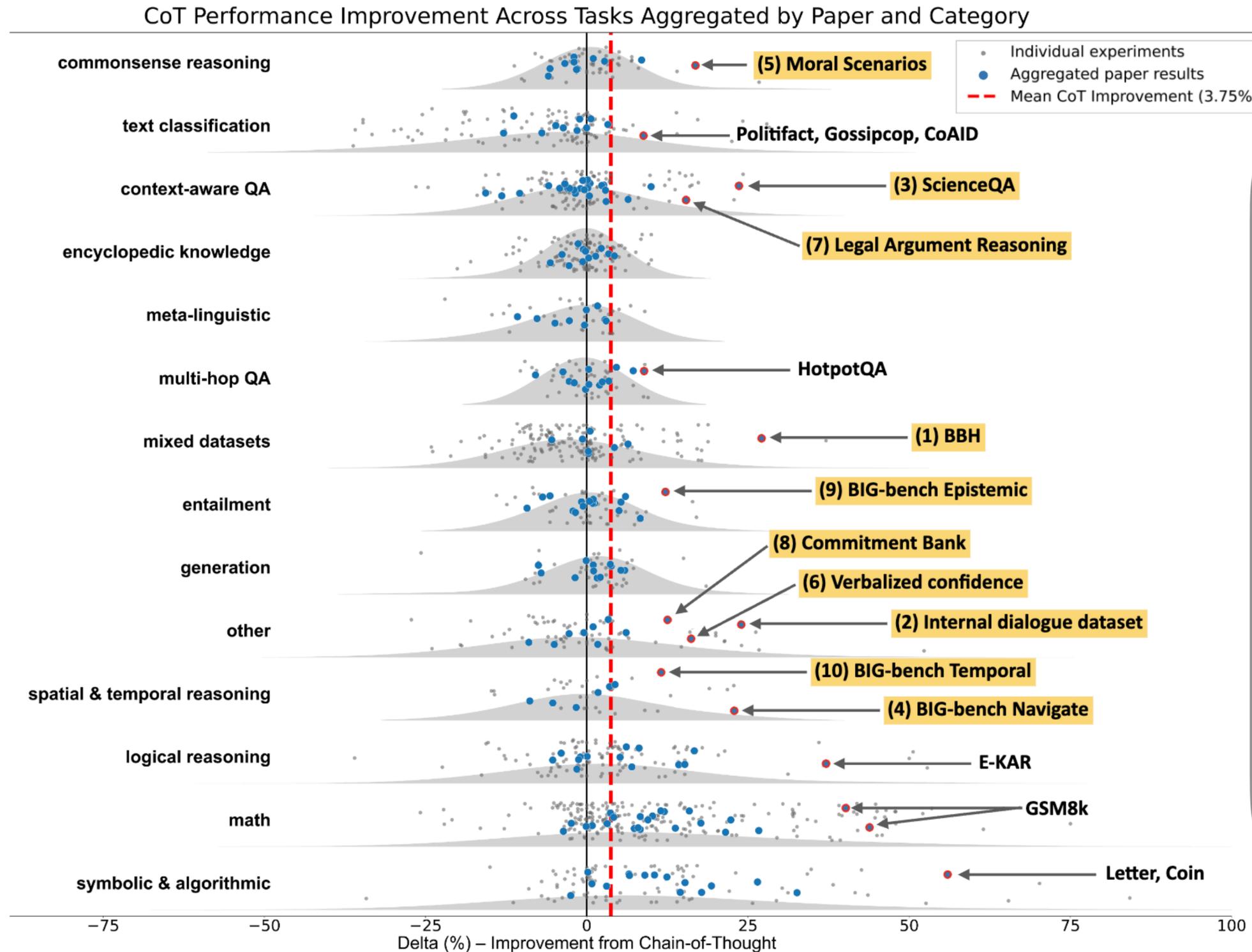


# What does the literature report?





# What does the literature report?



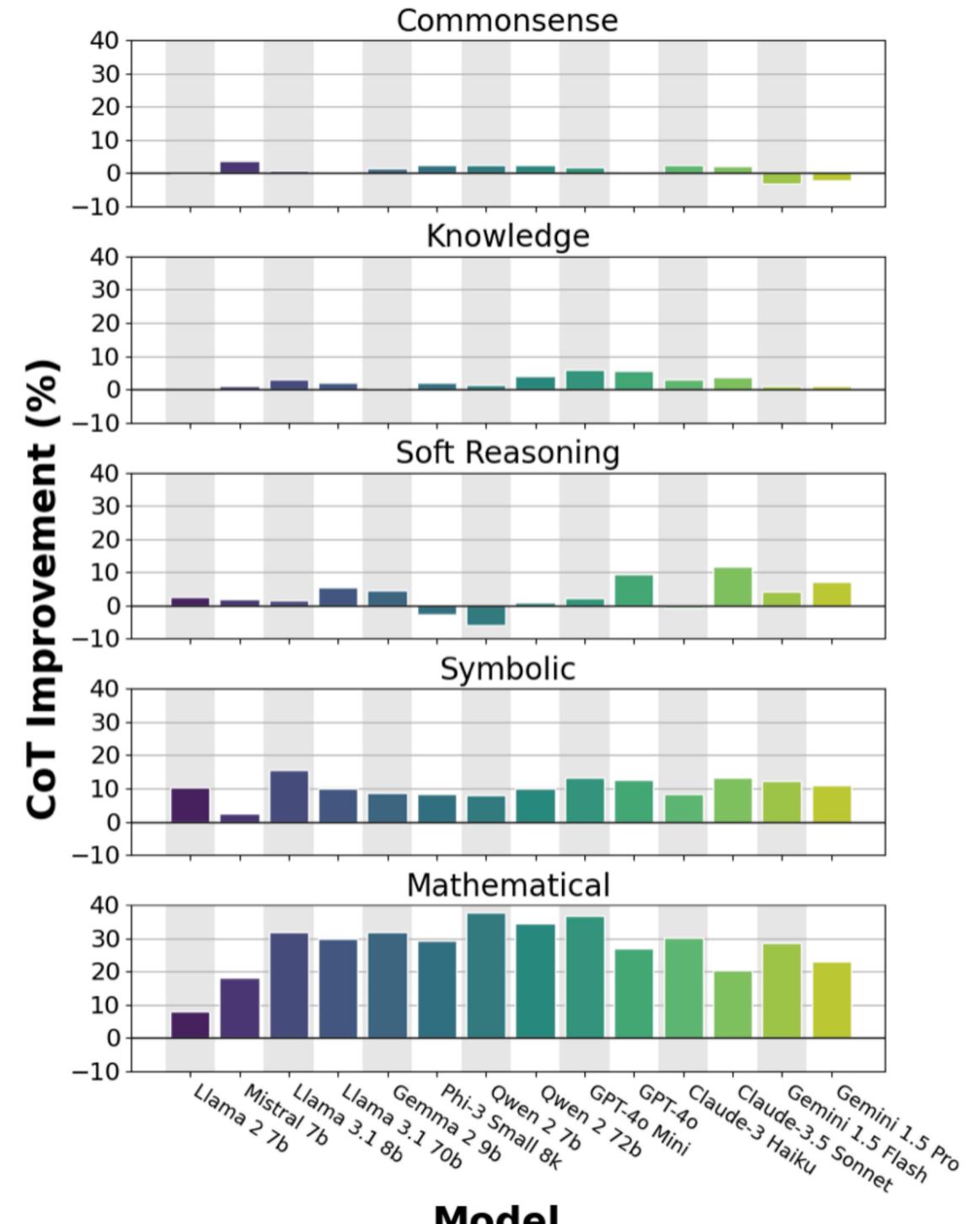
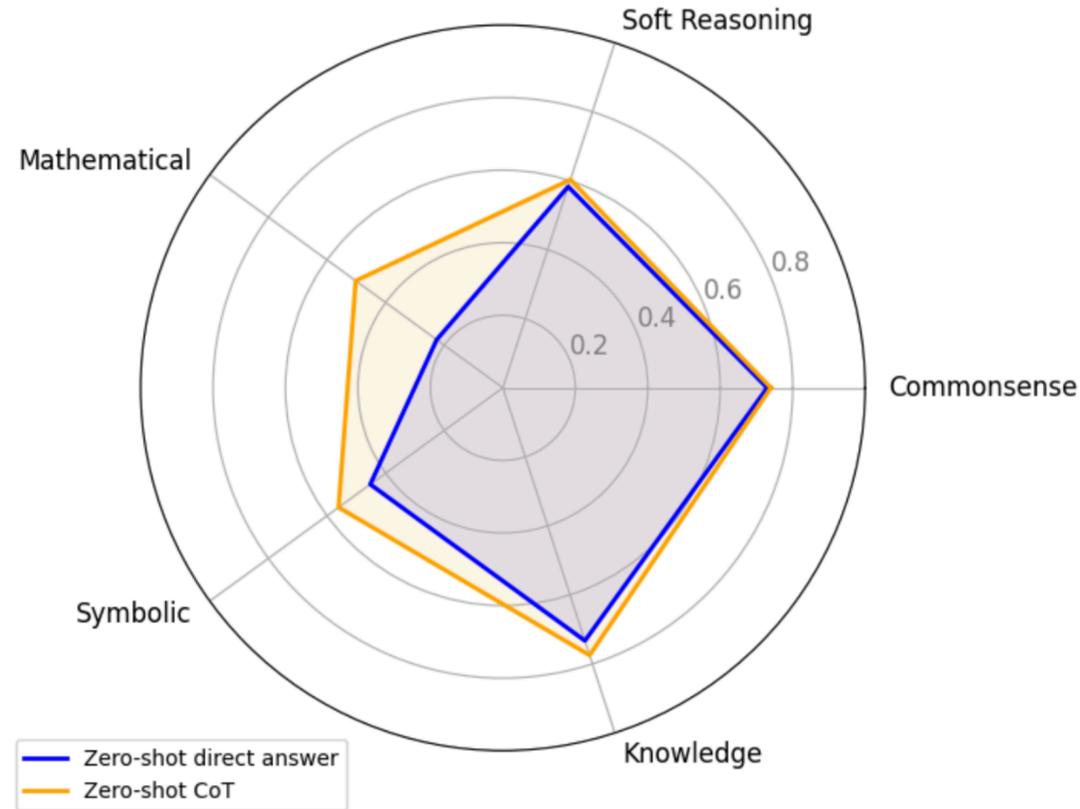
Most positive CoT results reported in the literature involve math or symbolic questions despite it being used in many domains.



# Analysis of Current Models

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

Our experiments on CoT improvements

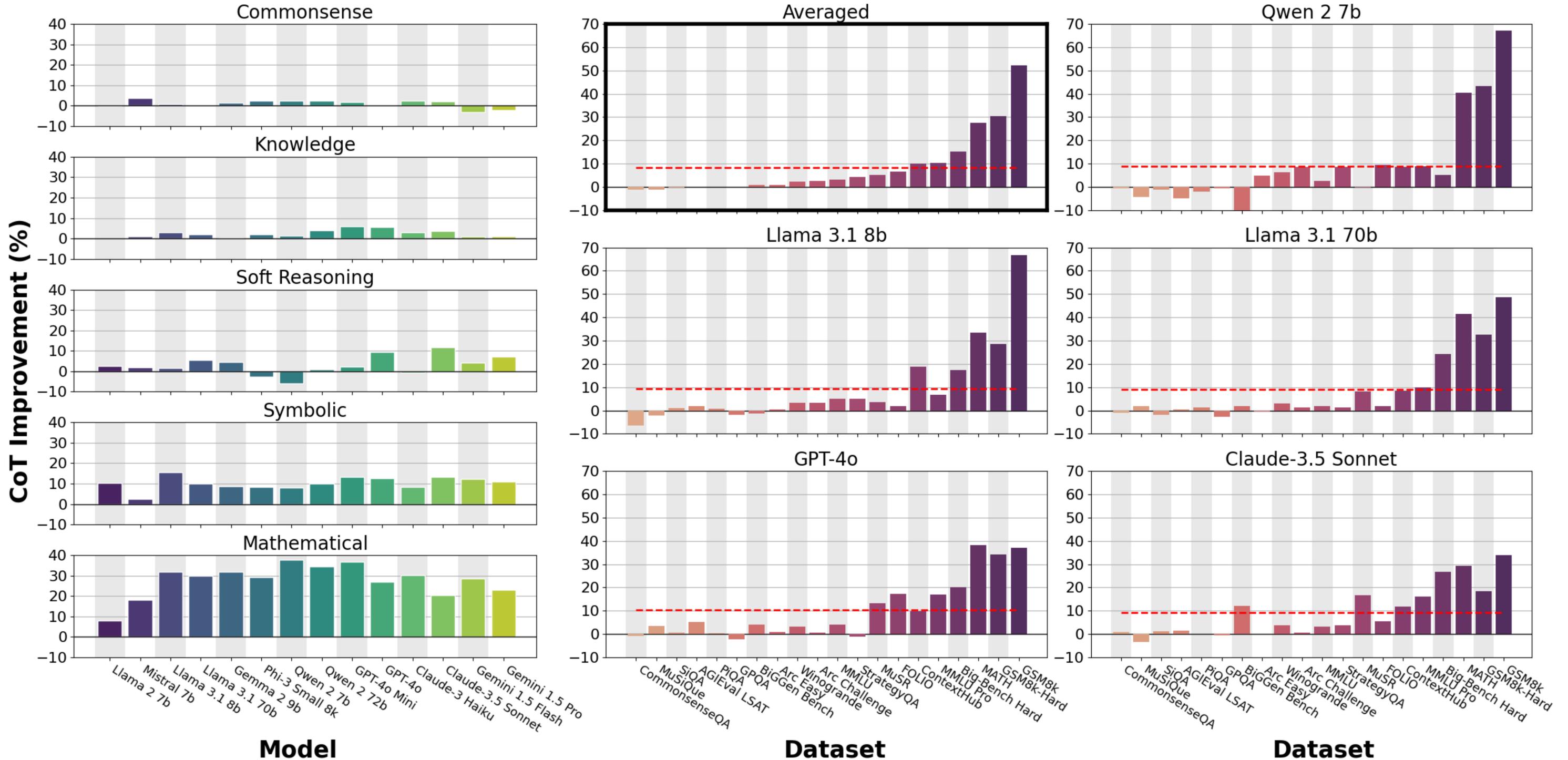


CoTs performance improvements are consistent across models, and only really helps on the symbolic and mathematical domains.



# Analysis of Current models

## CoT vs Direct Answer Prompting in Zero-Shot Setting





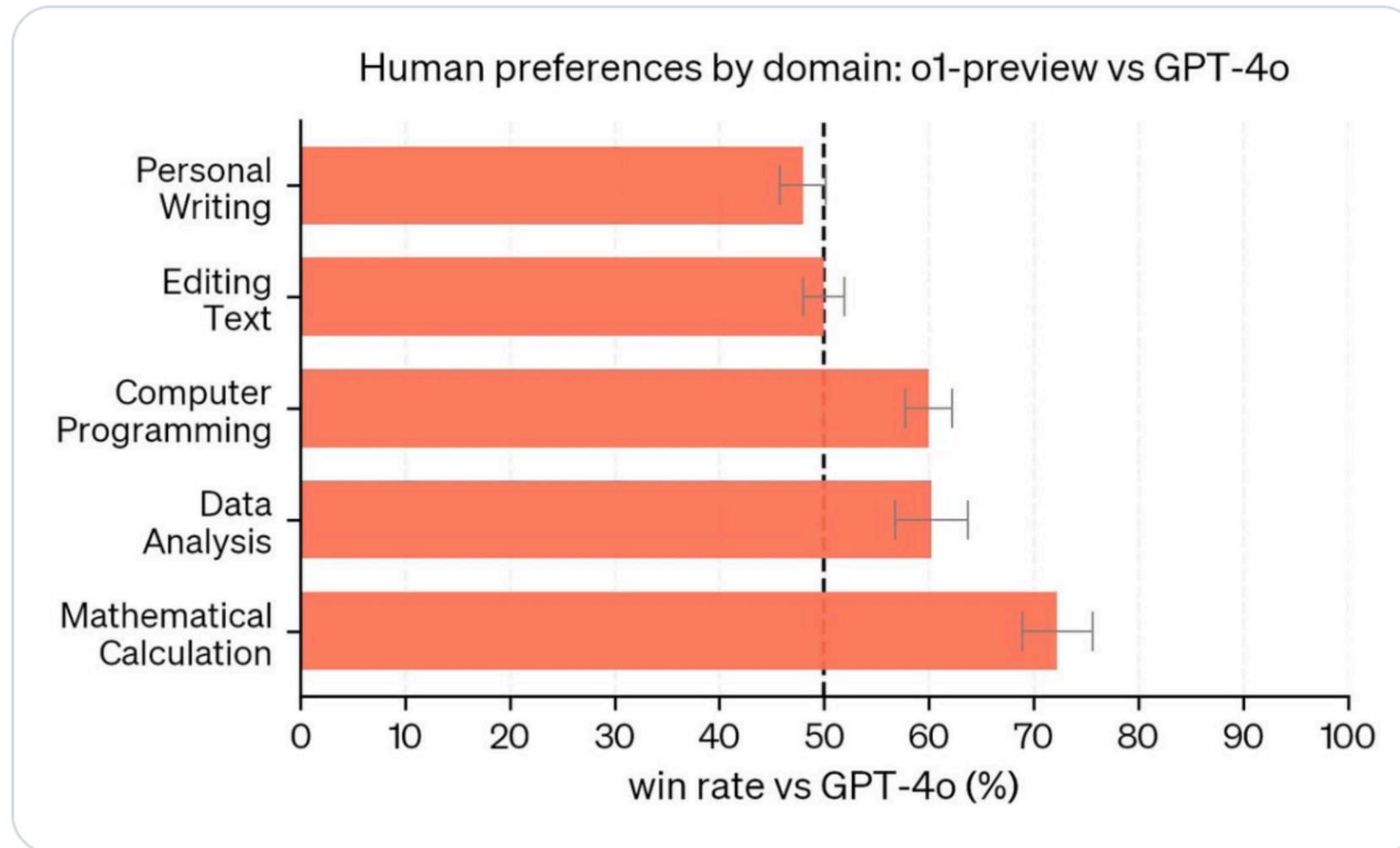
# o1



**Noam Brown** ✓  
@polynoamial



Our o1 models aren't always better than GPT-4o. Many tasks don't need reasoning, and sometimes it's not worth it to wait for an o1 response vs a quick GPT-4o response. One motivation for releasing o1-preview is to see what use cases become popular, and where the models need work.





# 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