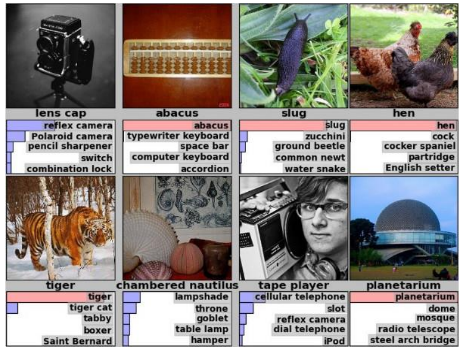
Can Language Models Compose Skills In-Context?

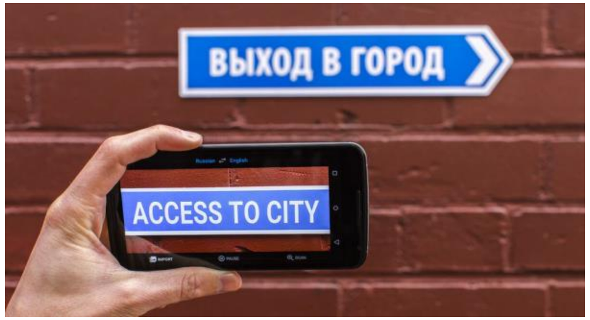
Zidong Liu, Zhuoyan Xu, Zhenmei Shi, Yingyu Liang

@HKU workshop on Fundamental AI and Theoretical Machine Learning

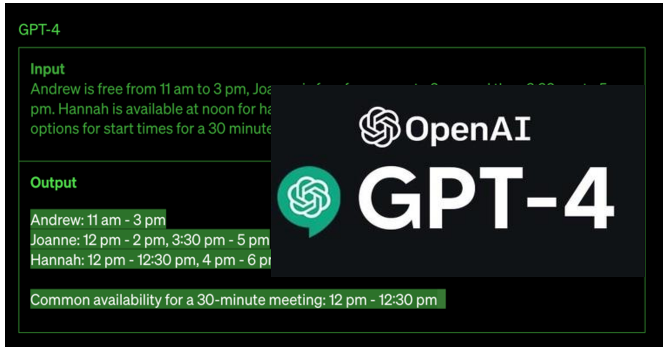
Machine Learning/AI Progress



Computer Vision



Natural Language Processing



GPT-4

Input

Andrew is free from 11 am to 3 pm, Joa

pm. Hannah is available at noon for ha

options for start times for a 30 minute

Output

Andrew: 11 am - 3 pm

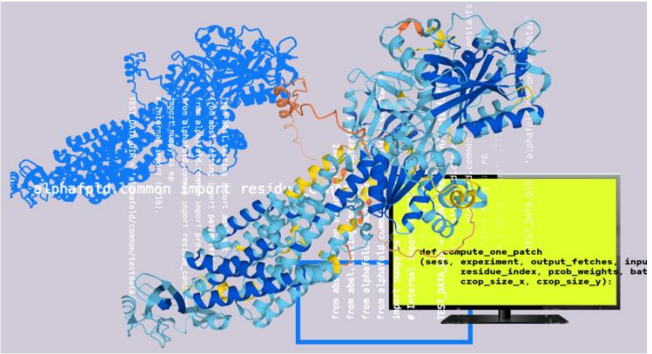
Joanne: 12 pm - 2 pm, 3:30 pm - 5 pm

Hannah: 12 pm - 12:30 pm, 4 pm - 6 pr

Common availability for a 30-minute meeting: 12 pm - 12:30 pm

Chatbots

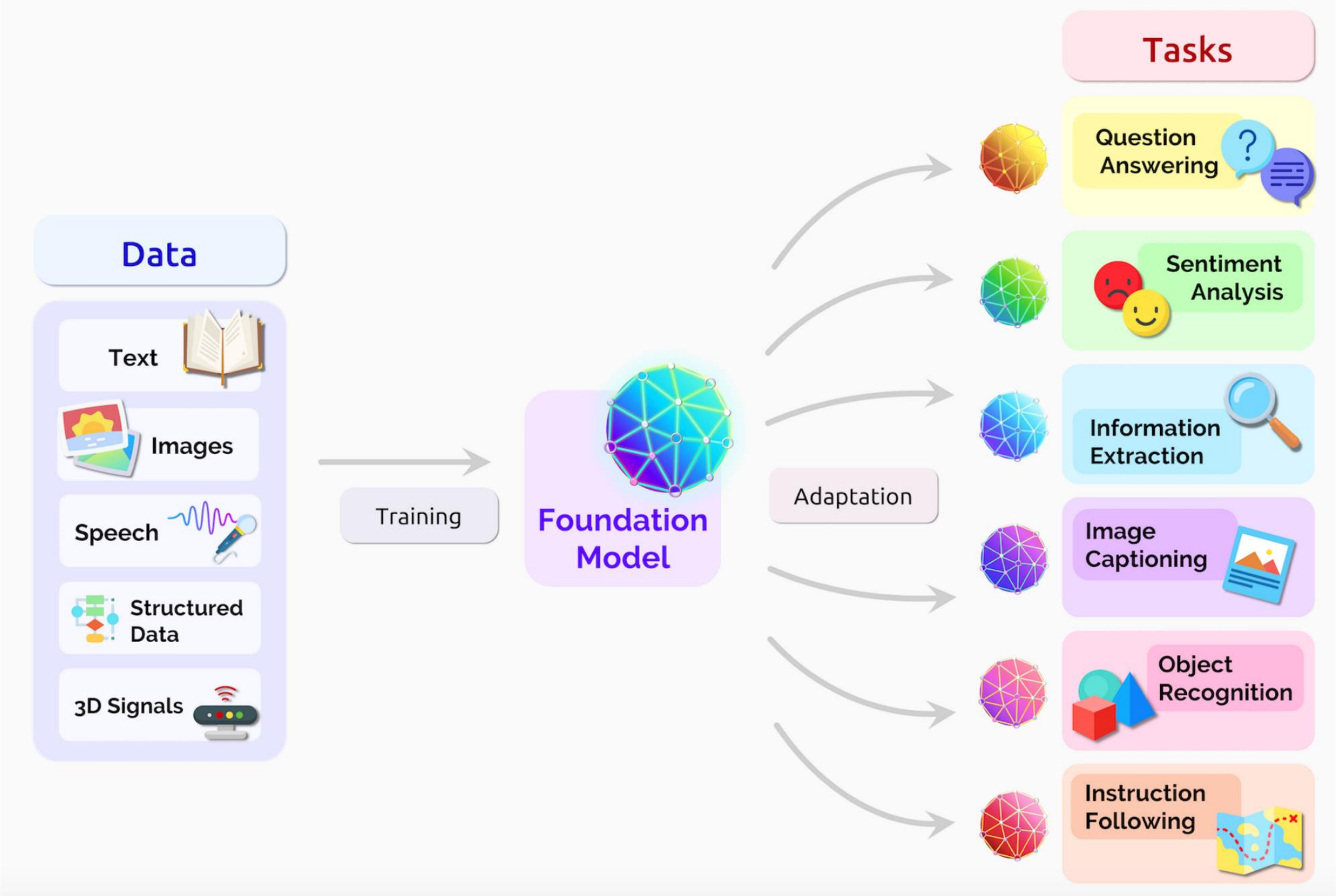
Game Playing


alphafold.common import resid
def compute_one_patch
(sess, experiment, output_fetches, inpu
residue_index, prob_weights, bat
crop_size_x, crop_size_y):

Sciences

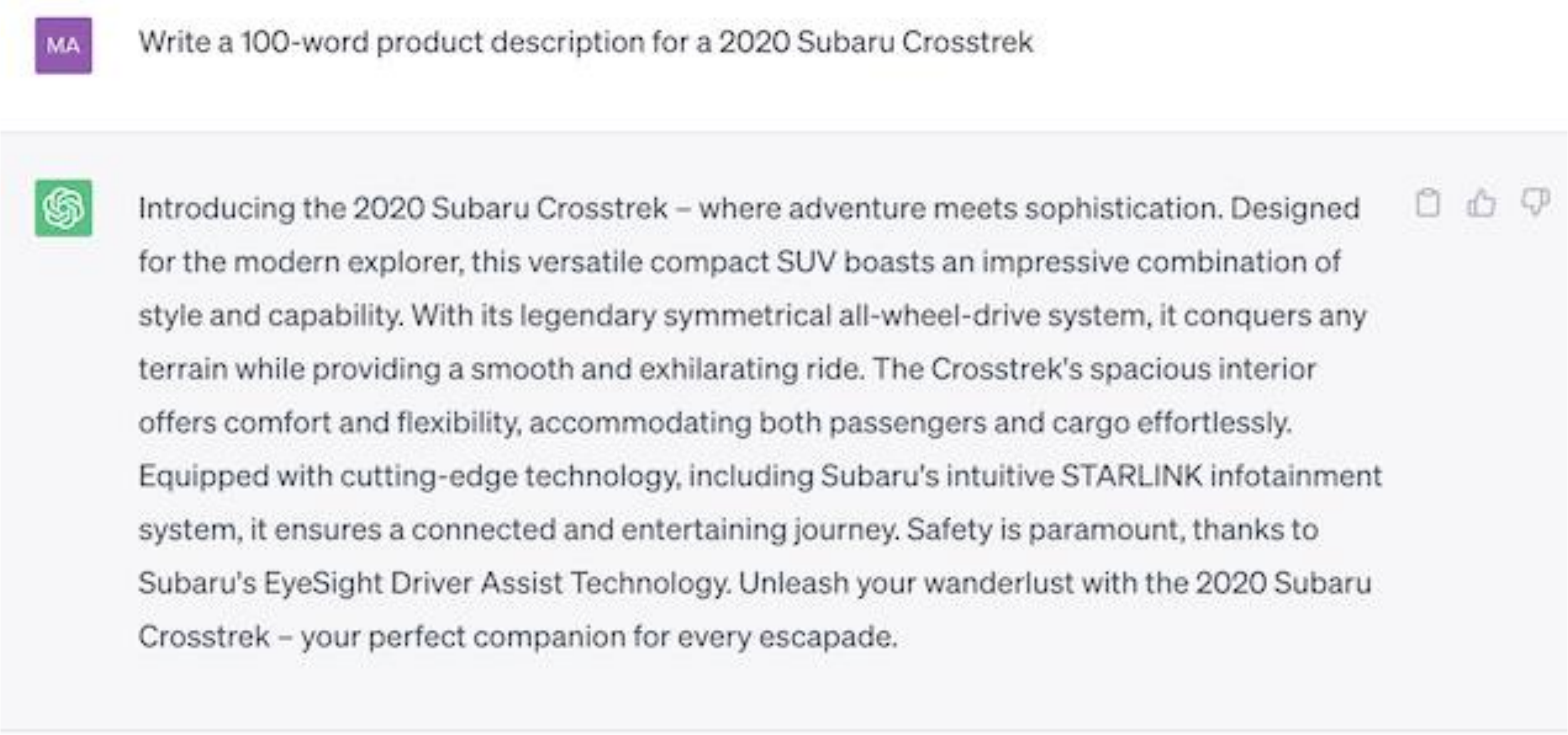
Arts

credit: OpenAI, Nature, NYTimes, MIT technology review



credit: Opportunities and Risks of Foundation Models. Bommasani, et al. 2021

Large Language Models (LLMs)

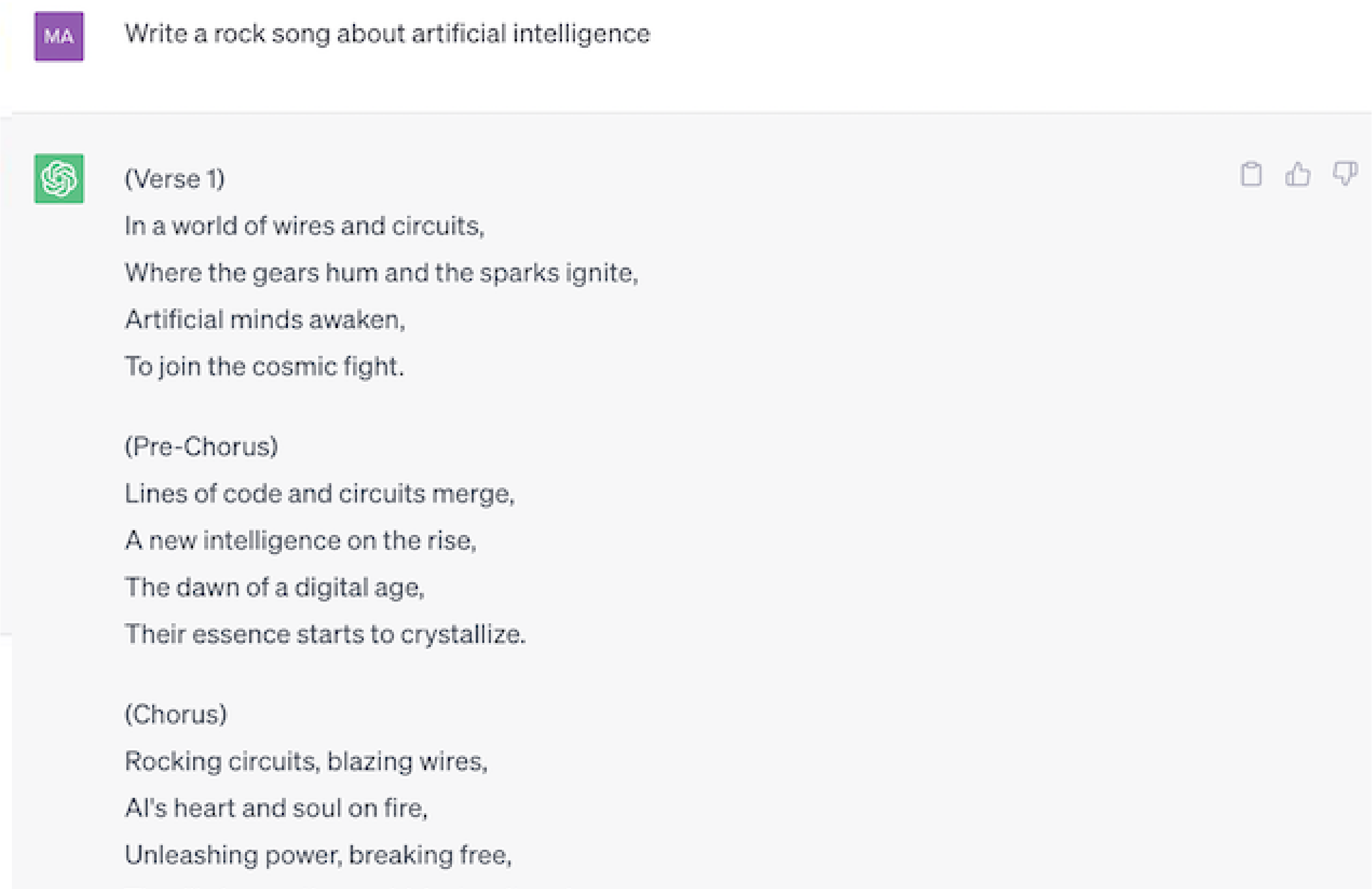


Write a 100-word product description for a 2020 Subaru Crosstrek

Introducing the 2020 Subaru Crosstrek – where adventure meets sophistication. Designed for the modern explorer, this versatile compact SUV boasts an impressive combination of style and capability. With its legendary symmetrical all-wheel-drive system, it conquers any terrain while providing a smooth and exhilarating ride. The Crosstrek's spacious interior offers comfort and flexibility, accommodating both passengers and cargo effortlessly. Equipped with cutting-edge technology, including Subaru's intuitive STARLINK infotainment system, it ensures a connected and entertaining journey. Safety is paramount, thanks to Subaru's EyeSight Driver Assist Technology. Unleash your wanderlust with the 2020 Subaru Crosstrek – your perfect companion for every escapade.

https://builtin.com/artificial-intelligence/chatgpt-examples

Large Language Models (LLMs)



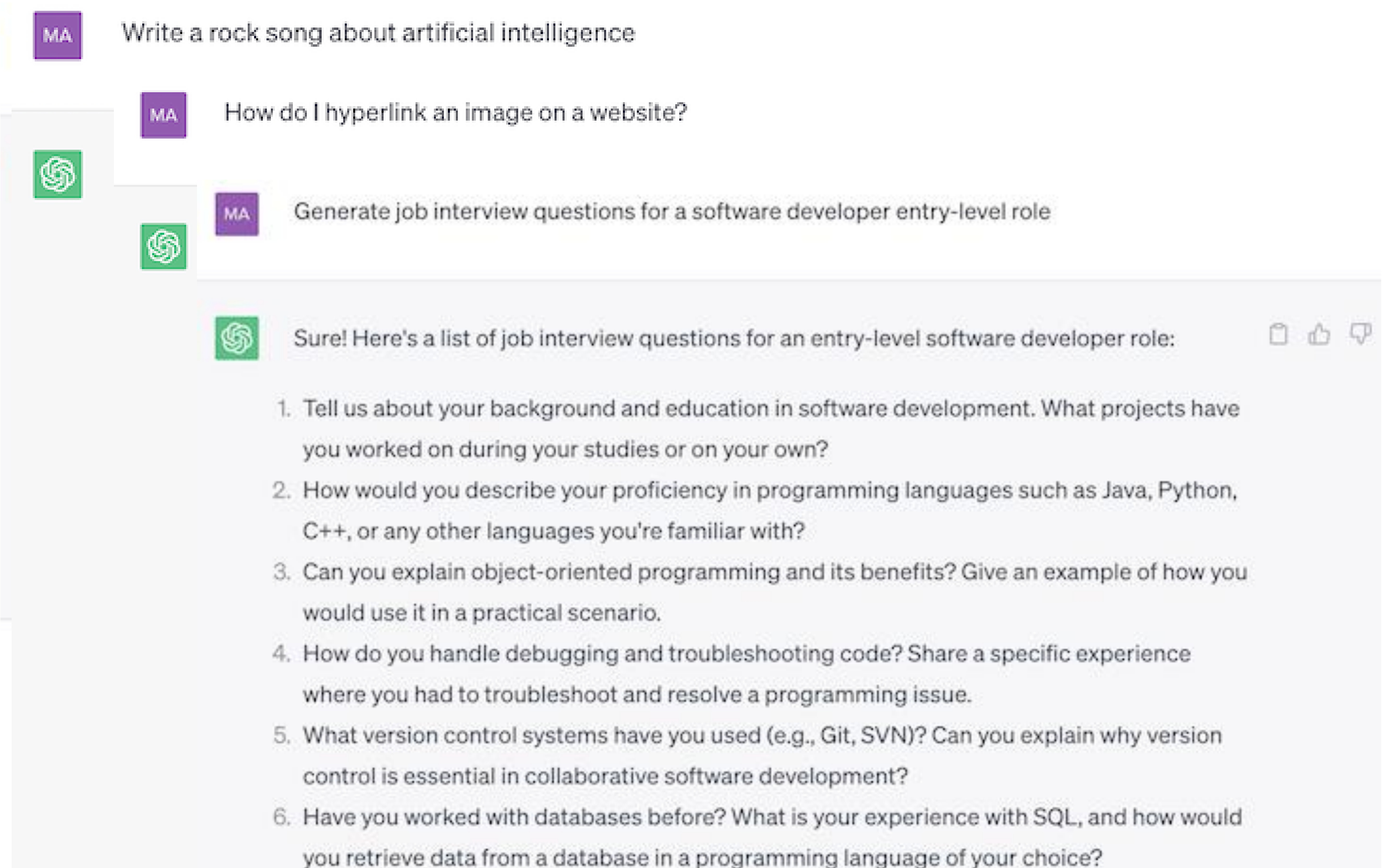
Write a rock song about artificial intelligence

(Verse 1)
In a world of wires and circuits,
Where the gears hum and the sparks ignite,
Artificial minds awaken,
To join the cosmic fight.

(Pre-Chorus)
Lines of code and circuits merge,
A new intelligence on the rise,
The dawn of a digital age,
Their essence starts to crystallize.

(Chorus)
Rocking circuits, blazing wires,
AI's heart and soul on fire,
Unleashing power, breaking free,

Large Language Models (LLMs)



Write a rock song about artificial intelligence

How do I hyperlink an image on a website?

Generate job interview questions for a software developer entry-level role

Sure! Here's a list of job interview questions for an entry-level software developer role:

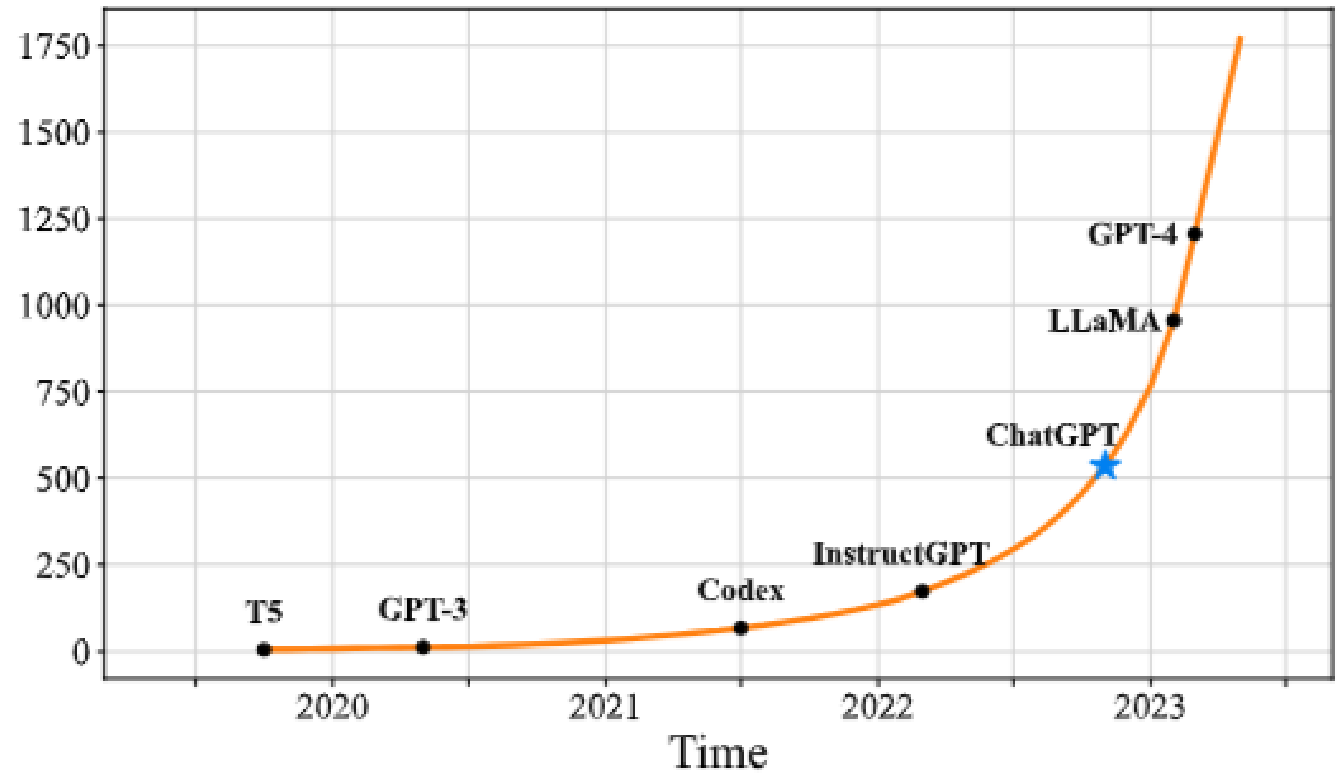
1. Tell us about your background and education in software development. What projects have you worked on during your studies or on your own?
2. How would you describe your proficiency in programming languages such as Java, Python, C++, or any other languages you're familiar with?
3. Can you explain object-oriented programming and its benefits? Give an example of how you would use it in a practical scenario.
4. How do you handle debugging and troubleshooting code? Share a specific experience where you had to troubleshoot and resolve a programming issue.
5. What version control systems have you used (e.g., Git, SVN)? Can you explain why version control is essential in collaborative software development?
6. Have you worked with databases before? What is your experience with SQL, and how would you retrieve data from a database in a programming language of your choice?

Large Language Models (LLMs)

GPT-4.5

Claude 3.5 Sonnet

Gemini 2.5



(b) Query="Large Language Model"

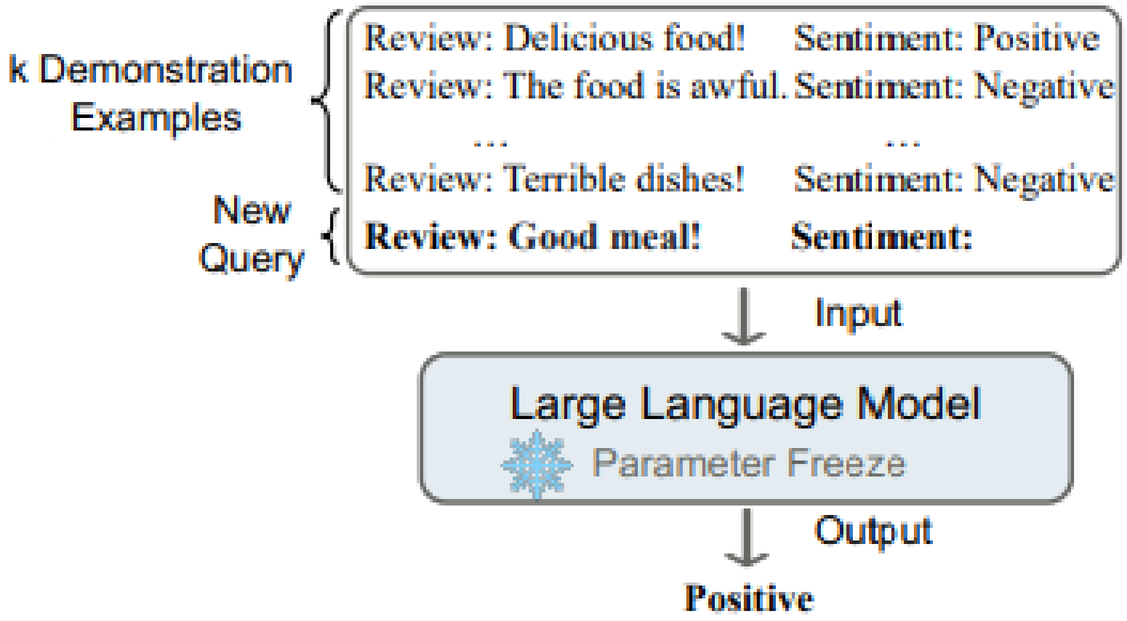
deepseek

Grok 3

Cumulative numbers of arXiv papers contain keyphrase "large language model"

Figures from: A Survey of Large Language Models, 2023.

In-Context Learning (ICL)



Review: Delicious food! Sentiment: Positive
Review: The food is awful. Sentiment: Negative
… …
Review: Terrible dishes! Sentiment: Negative
Review: Good meal! Sentiment:

- Allow easy adaptation to various tasks
- Key property towards AGI

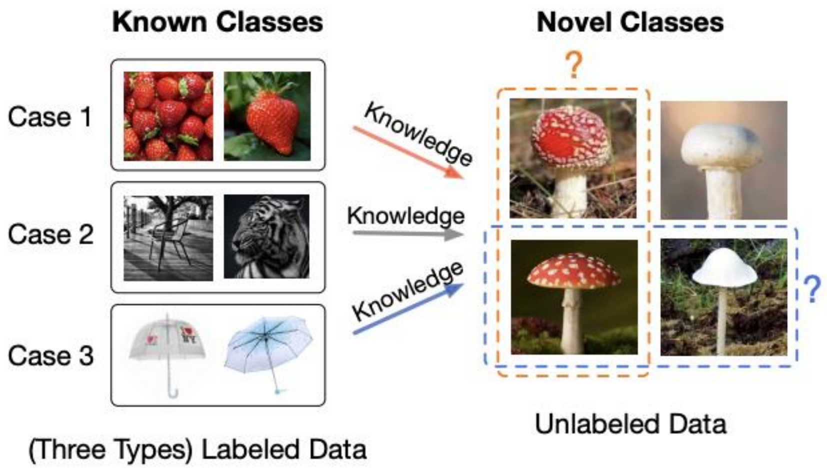
Composition

- Why can adapt to general tasks? Knowledge reuse

Composition

- Why can adapt to general tasks? Knowledge reuse

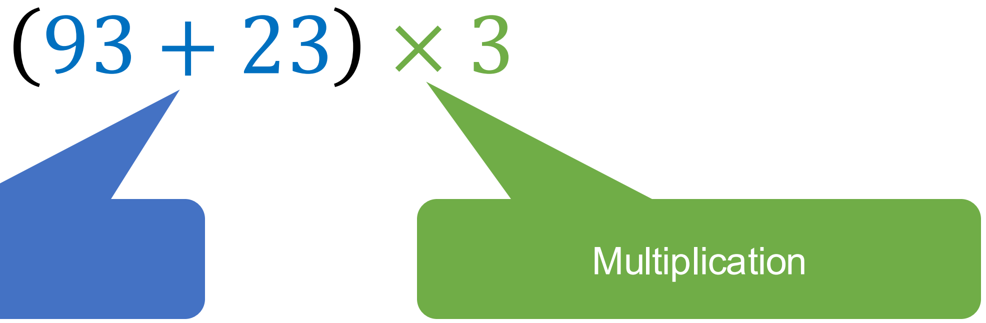
- Many tasks are compositions of basic skills
  - Novel classification tasks involving several features each seen in different tasks



Composition

- Why can adapt to general tasks? Knowledge reuse

- Many tasks are compositions of basic skills
  - Novel classification tasks involving several features each seen in different tasks
  - Math/reasoning involving multiple steps that may need different skills



$$(93 + 23) \times 3$$

Composition

- Why can adapt to various tasks? Knowledge reuse

- Many tasks are compositions of basic skills
  - Novel classification tasks involving several patterns each seen in different tasks
  - Math/reasoning involving multiple steps that may need different skills

- Composition ability is crucial for general intelligence
  - Exponentially many compositions of basic skills; cannot handle individually
  - Allows efficient adaptation with few data

In-Context Composition

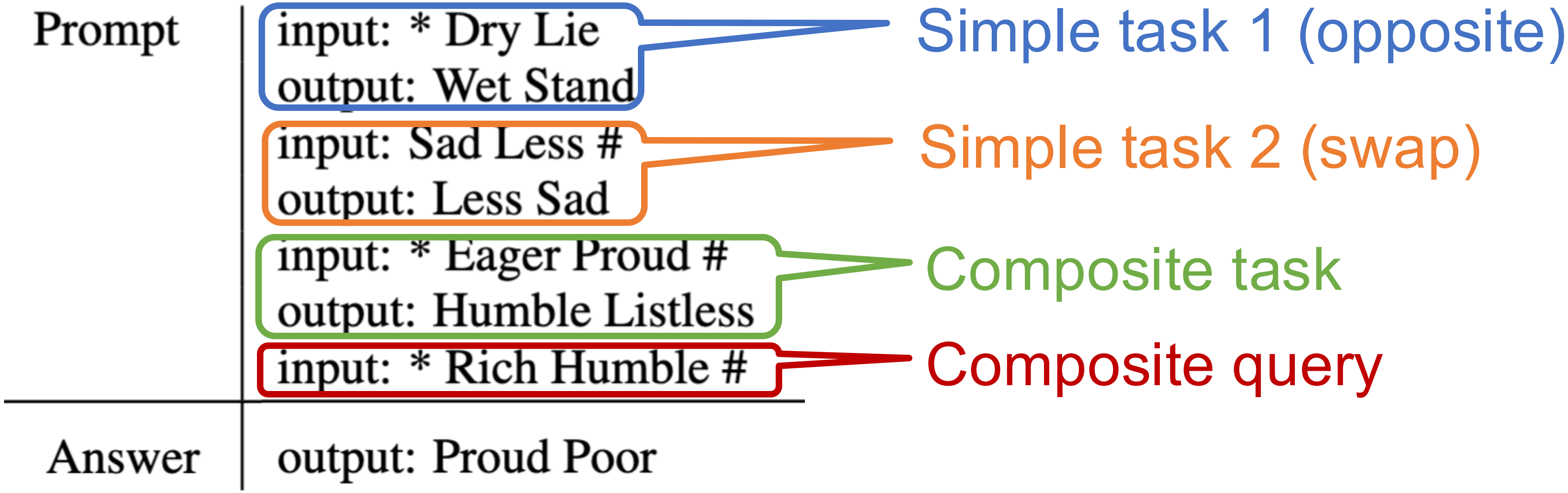
- Given the context consisting of
  - k_1 examples from simple task 1 (f_1)
  - k_2 examples from simple task 2 (f_2)
  - Possibly k_c examples from composite task composing task 1 and 2 ($f_2 \circ f_1$)

- The model needs to address a query from the composite task

Can language models do composition in-context?

In-Context Composition Dataset

- Stylized linguistic and logic tasks
- Use special symbols for the tasks



Prompt	input: * Dry Lie output: Wet Stand input: Sad Less # output: Less Sad input: * Eager Proud # output: Humble Listless input: * Rich Humble #
Answer	output: Proud Poor

Simple task 1 (opposite)

Simple task 2 (swap)

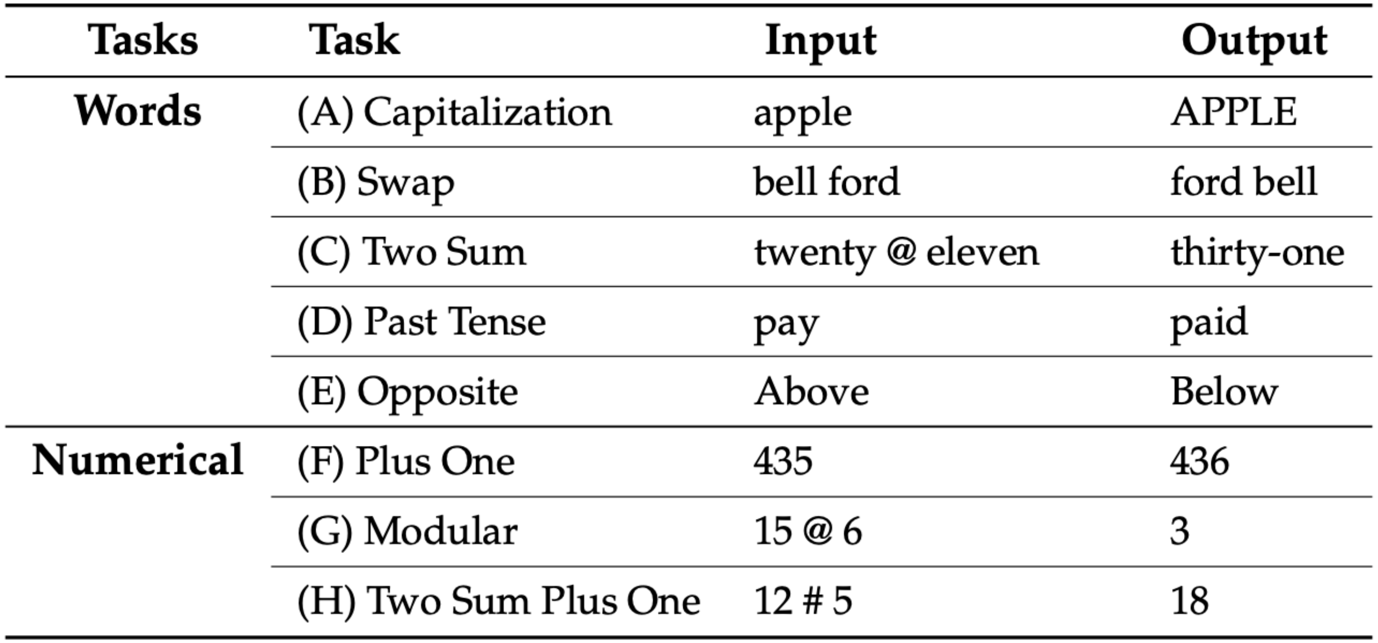
Composite task

Composite query

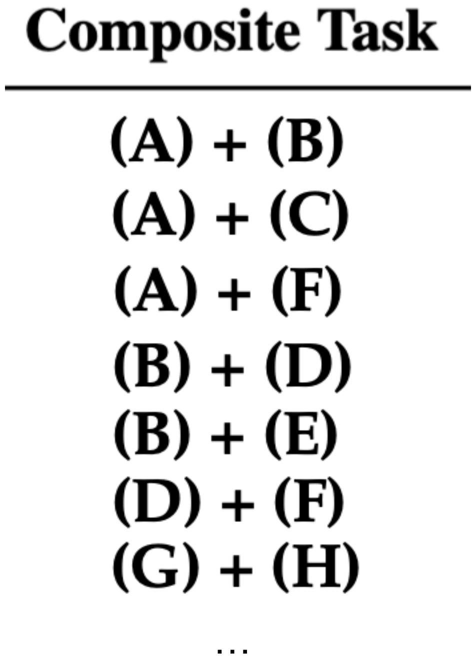
Xu, Shi, L. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability. COLM 2024.

In-Context Composition Dataset

- Stylized linguistic and logic tasks
- Use special symbols for the tasks



Tasks	Task	Input	Output
Words	(A) Capitalization	apple	APPLE
	(B) Swap	bell ford	ford bell
	(C) Two Sum	twenty @ eleven	thirty-one
	(D) Past Tense	pay	paid
	(E) Opposite	Above	Below
Numerical	(F) Plus One	435	436
	(G) Modular	15 @ 6	3
	(H) Two Sum Plus One	12 # 5	18



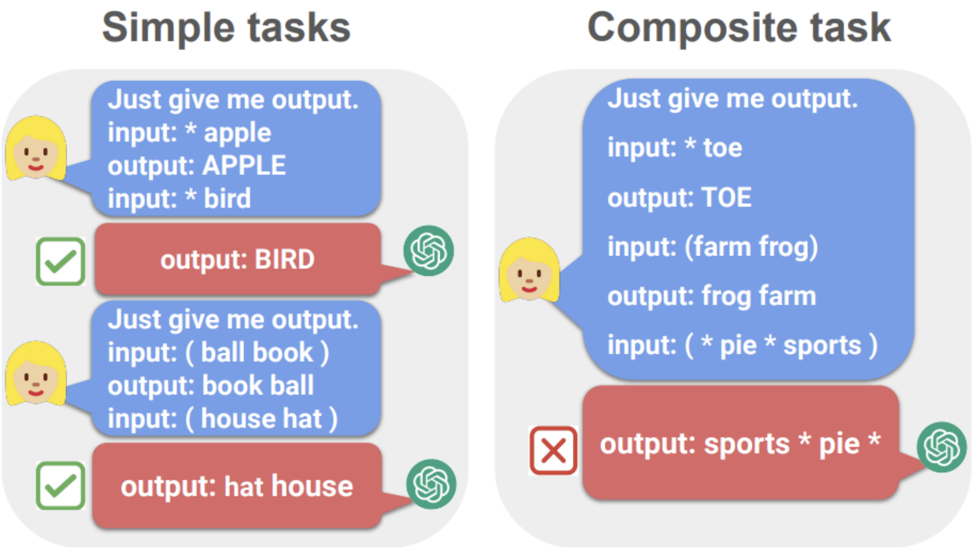
Composite Task

(A) + (B)
(A) + (C)
(A) + (F)
(B) + (D)
(B) + (E)
(D) + (F)
(G) + (H)
…

Xu, Shi, L. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability. COLM 2024.

Sanity Checks

- Models: Llama (7B, 13B, 30B, 65B), Llama2 (7B, 13B, and 70B), Mistral (7B and 8x7B)
- Simple task queries without examples: Close to 0 accuracy
- Simple task queries with examples: Easily solved
- Composite queries without composite examples: Low performance



Just give me output.
input: * apple
output: APPLE
input: * bird

output: BIRD

Just give me output.
input: (ball book)
output: book ball
input: (house hat)

output: hat house

Just give me output.
input: * toe
output: TOE
input: (farm frog)
output: frog farm
input: (* pie * sports)

output: sports * pie *

Can in-context examples help composition?

Recall the setting:

- Given the context consisting of
    - k_1 examples from simple task 1
    - k_2 examples from simple task 2
    - Possibly k_c examples from composite task composing task 1 and 2
- The model needs to address a query from the composite task

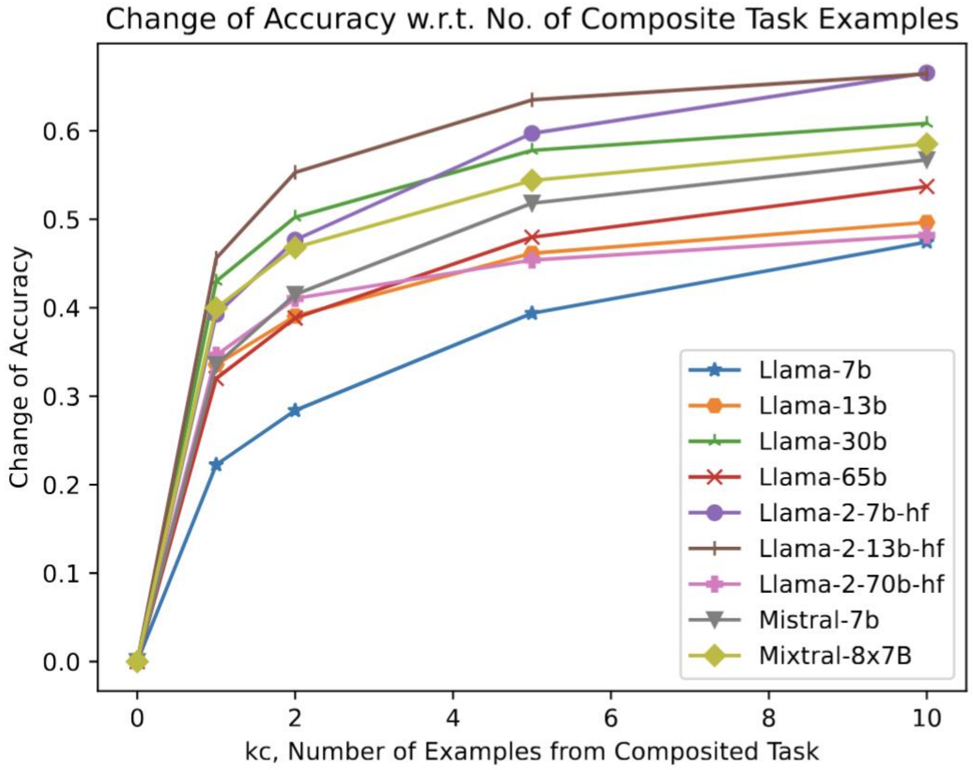
Method

- Increase the example number, check the performance **change**

Liu, Xu, Shi, L. Can Language Models Compose Skills In-Context? Manuscript, 2025.

Can in-context examples help composition?

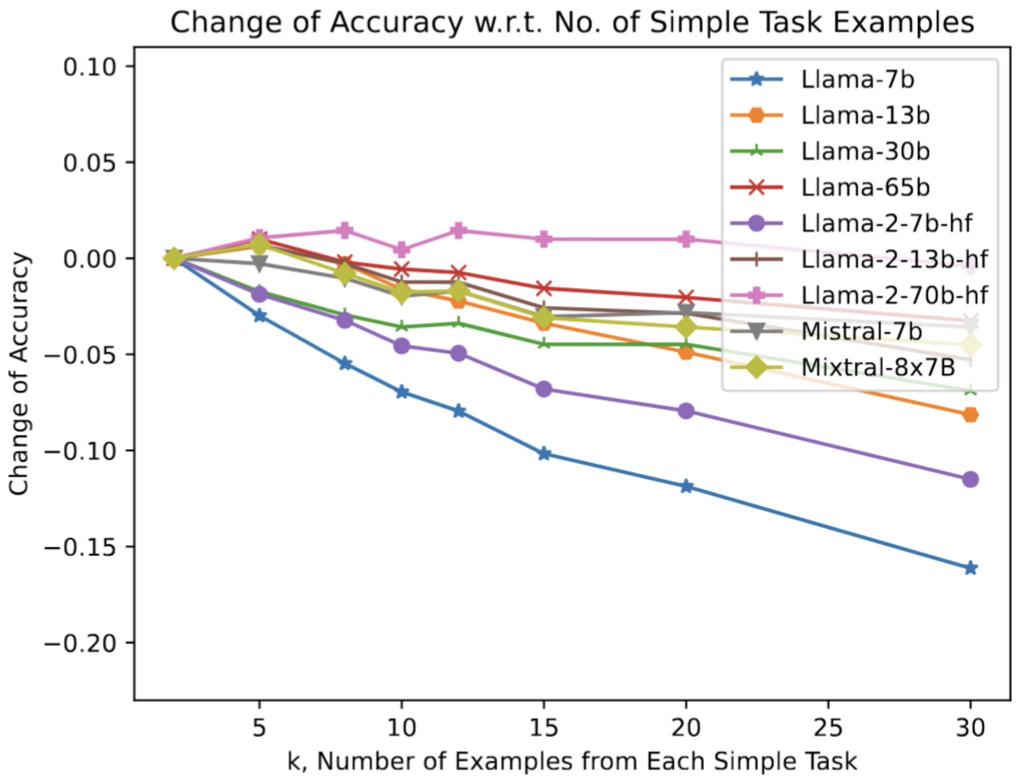
- Increase the example number, check the performance **change**
  - Finding: Composite task examples help the performance as expected



Increasing the number of composite task examples

Can in-context examples help composition?

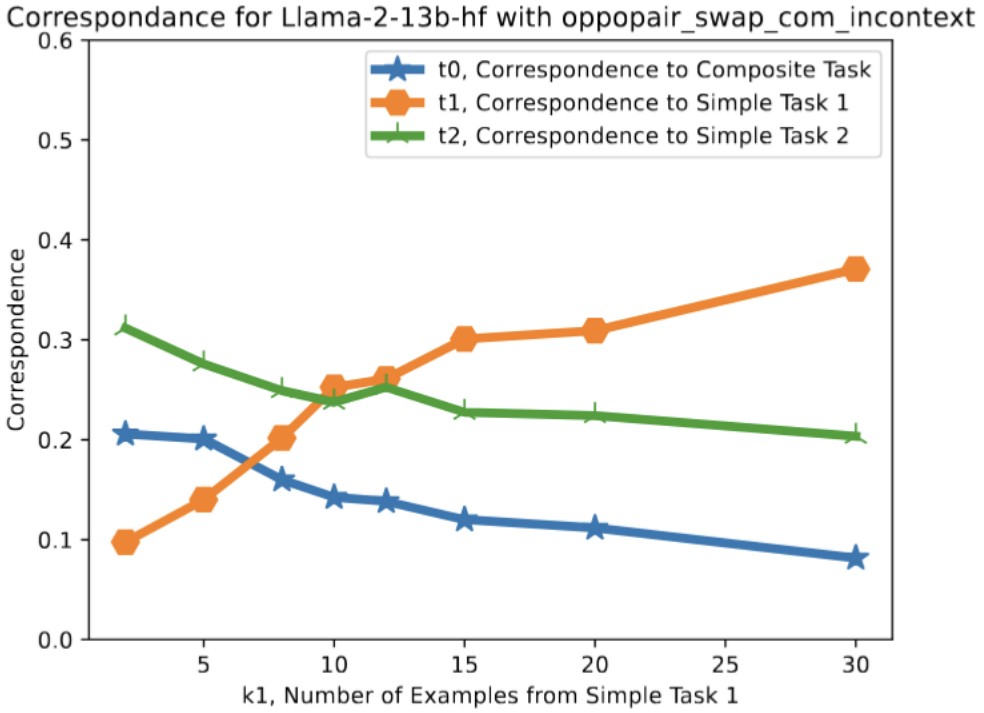
- Increase the example number, check the performance **change**
  - Finding: Simple task examples hurt the performance unexpectedly



Increasing the number of simple task examples

What info from the examples is utilized?

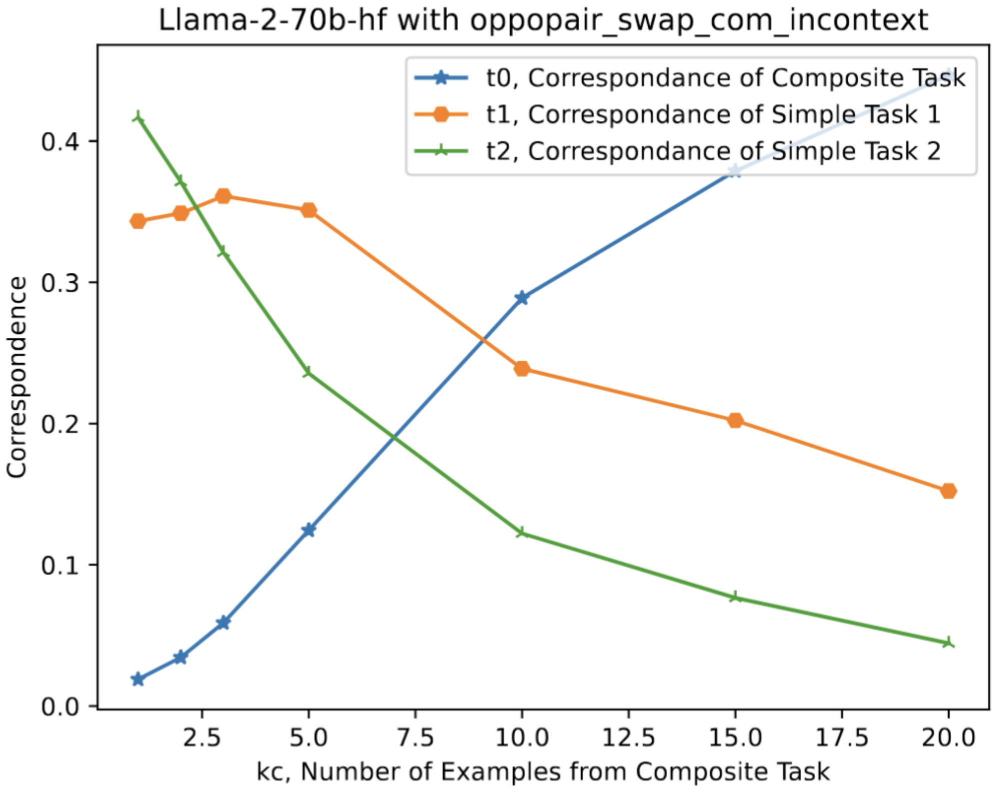
- Method: check how the output distribution relates to the examples
  - **Correspondence** to a task: output only performing that task
  - Finding: Models may match query to in-context examples from any task, ignoring the compositional structure



(b) Llama-2-13b model

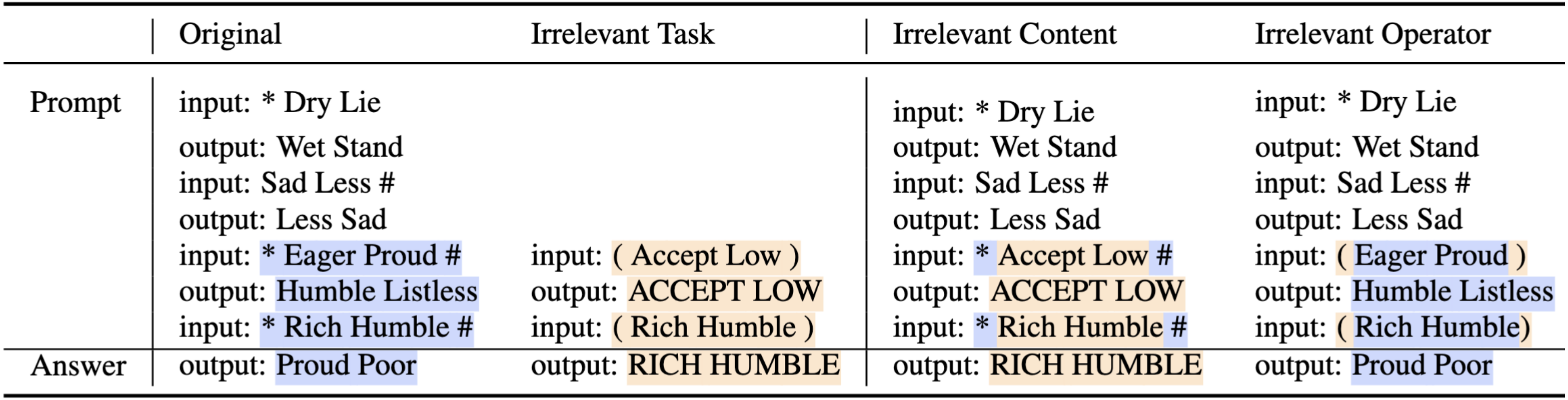
What info from the examples is utilized and how?

- Method: check how the output distribution relates to the examples
  - **Correspondence** to a task: Output only performing that task
  - Finding: Models may match query to in-context examples from any task, ignoring the compositional structure



How are the examples are utilized?

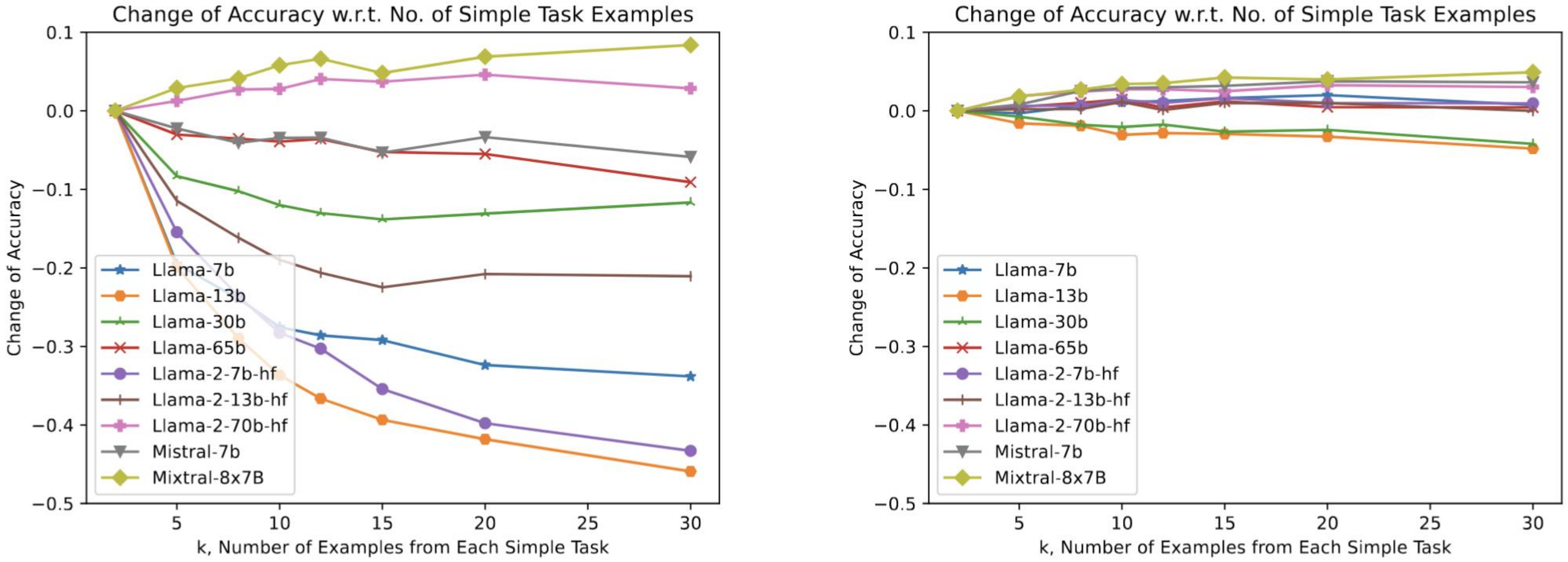
- Method: Ablate the two parts (operator and content) of the examples



	Original	Irrelevant Task	Irrelevant Content	Irrelevant Operator
Prompt	input: * Dry Lie		input: * Dry Lie	input: * Dry Lie
	output: Wet Stand		output: Wet Stand	output: Wet Stand
	input: Sad Less #		input: Sad Less #	input: Sad Less #
	output: Less Sad		output: Less Sad	output: Less Sad
	input: * Eager Proud #	input: (Accept Low)	input: * Accept Low #	input: (Eager Proud)
	output: Humble Listless	output: ACCEPT LOW	output: ACCEPT LOW	output: Humble Listless
	input: * Rich Humble #	input: (Rich Humble)	input: * Rich Humble #	input: (Rich Humble)
Answer	output: Proud Poor	output: RICH HUMBLE	output: RICH HUMBLE	output: Proud Poor

How are the examples are utilized?

- Method: Ablate the two parts (content and operator) of the examples



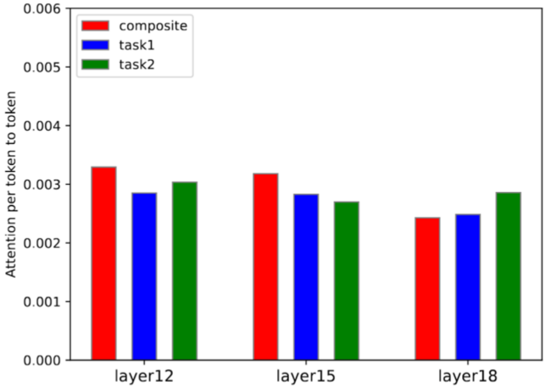
(a) Irrelevant Content

(b) Irrelevant Operator

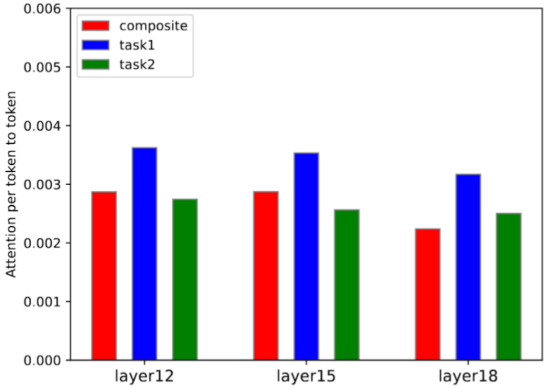
- Finding: Largely base on the operators to utilize the examples

Where are the models paying attention to?

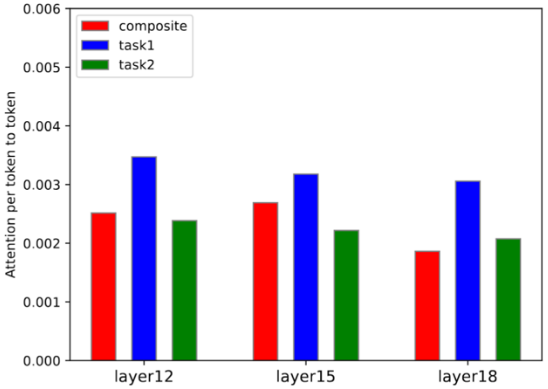
- Method: Pick a layer, compute the **average attention** from the query tokens to different groups (model: Mistral-7B)



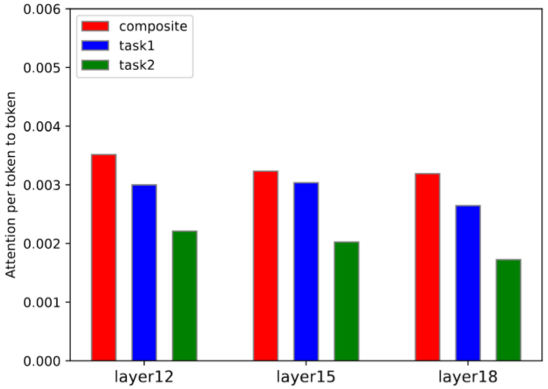
(a) The opposition+pastTense task



(b) The pastTense+capitalization task



(c) The pastTense+plusOne task



(d) The pastTense+swap task

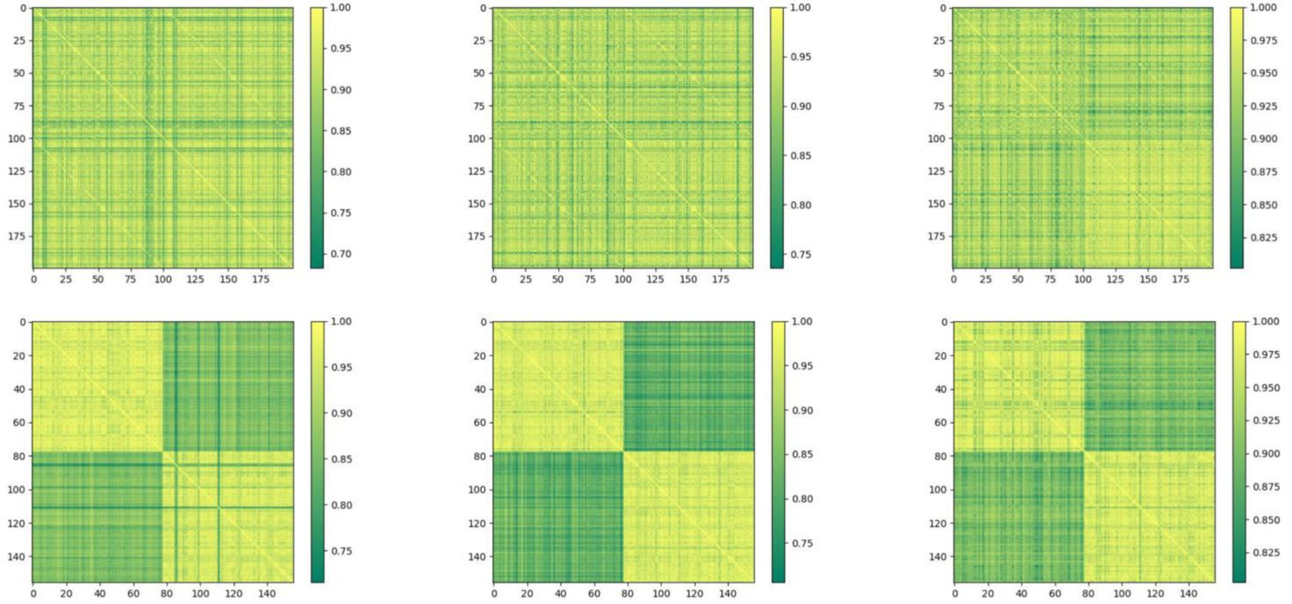
- Finding: Same order of attention is paid to examples of different tasks

Are attentions different on simple/composite queries?

- Method: fix context, generate 100 simple/composite queries; compute the similarities between their attentions (layer 12/15/18 of Mistral-7B)

Low accuracy

High accuracy



- Finding: Ignoring compositional structure leads to low accuracy

Can Chain-of-Thought examples help?

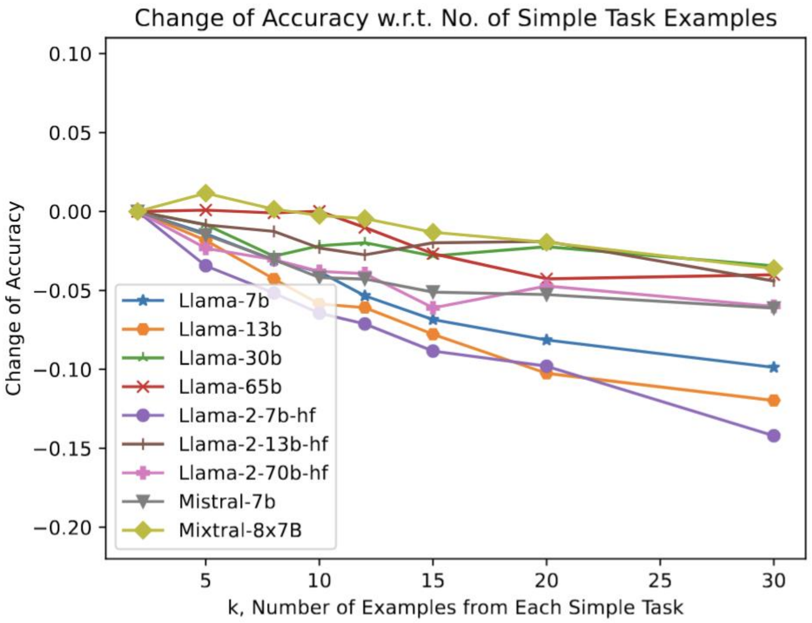
- CoT: popular method for improving LLM’s reasoning

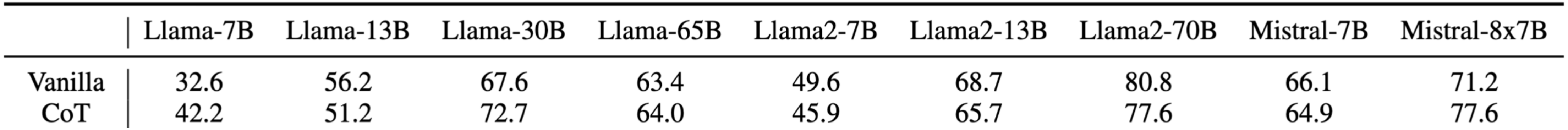
CoT examples: Adding intermediate output

- Original composite task example: * Rich Humble # -> Proud Poor where * denotes opposition and denotes swap
- CoT version: * Rich Humble # -> Poor Proud # -> Proud Poor

Can Chain-of-Thought examples help?

- Method: replace all composite examples with CoT, and redo experiments
  - Finding 1: CoT examples do not help mitigate the negative impact of more simple examples
  - Finding 2: CoT examples do not help improve the performance consistently





	Llama-7B	Llama-13B	Llama-30B	Llama-65B	Llama2-7B	Llama2-13B	Llama2-70B	Mistral-7B	Mistral-8x7B
Vanilla	32.6	56.2	67.6	63.4	49.6	68.7	80.8	66.1	71.2
CoT	42.2	51.2	72.7	64.0	45.9	65.7	77.6	64.9	77.6

Can Chain-of-Thought examples help?

- Error pattern: cannot align the examples with the corresponding steps in the composition

Example:

- Query: * Grow Respect # (opposition+swap task)
- Correct: * Grow Respect # -> Shrink Disrespect # -> Disrespect Shrink
- Model: * Grow Respect # -> Shrink Disrespect # -> Respect Grow
- In the second step, the model incorrectly performs opposition+swap instead of only swap

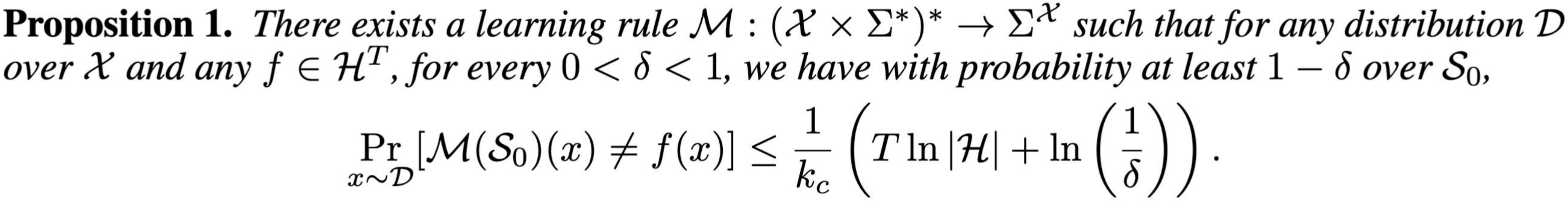
Preliminary Theoretical Analysis

- Composite task f as function composition

$$f = f_T \circ \cdots \circ f_2 \circ f_1 \in H^T$$

where $f_i \in H$ is simple task

- If only given k_c examples in form of $(x, f(x))$, by standard analysis:

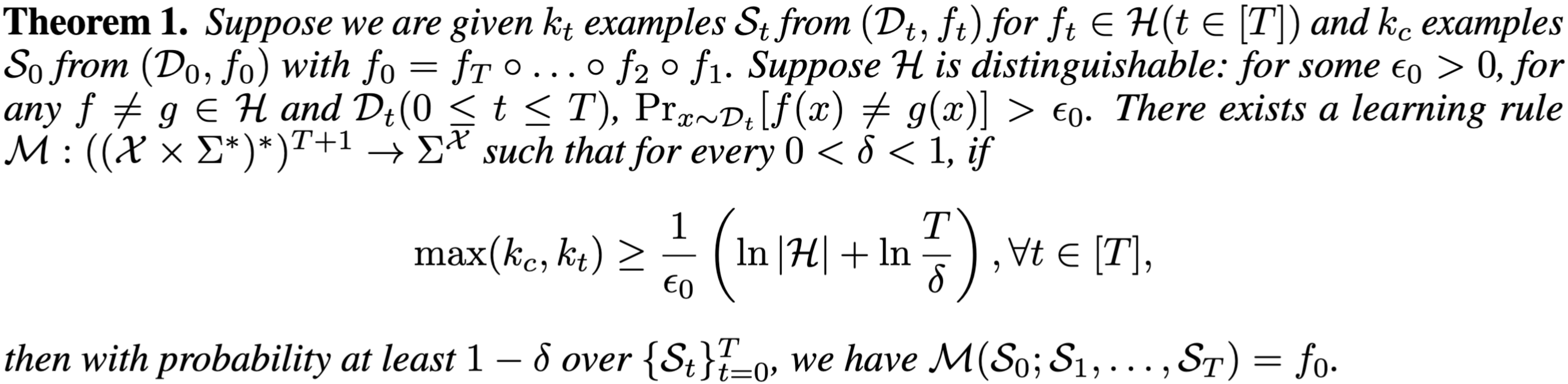


Proposition 1. *There exists a learning rule $\mathcal{M} : (\mathcal{X} \times \Sigma^*)^* \to \Sigma^{\mathcal{X}}$ such that for any distribution $\mathcal{D}$ over $\mathcal{X}$ and any $f \in \mathcal{H}^T$, for every $0 < \delta < 1$, we have with probability at least $1 - \delta$ over $\mathcal{S}_0$,*

$$\Pr_{x \sim \mathcal{D}}[\mathcal{M}(\mathcal{S}_0)(x) \neq f(x)] \leq \frac{1}{k_c}\left(T \ln|\mathcal{H}| + \ln\left(\frac{1}{\delta}\right)\right).$$

Preliminary Theoretical Analysis

- If given k_c examples in form of $(z^1 = x, z^2, \ldots, z^T)$ where $z^{t+1} = f_t(z^t)$
- Also given k_t examples in form of $(x, f_t(x))$ knowing that they are from simple task t



Theorem 1. *Suppose we are given* k_t *examples* $\mathcal{S}_t$ *from* $(\mathcal{D}_t, f_t)$ *for* $f_t \in \mathcal{H} (t \in [T])$ *and* k_c *examples* $\mathcal{S}_0$ *from* $(\mathcal{D}_0, f_0)$ *with* $f_0 = f_T \circ \ldots \circ f_2 \circ f_1$. *Suppose* $\mathcal{H}$ *is distinguishable: for some* $\epsilon_0 > 0$, *for any* $f \neq g \in \mathcal{H}$ *and* $\mathcal{D}_t (0 \le t \le T)$, $\Pr_{x \sim \mathcal{D}_t}[f(x) \neq g(x)] > \epsilon_0$. *There exists a learning rule* $\mathcal{M} : ((\mathcal{X} \times \Sigma^*)^*)^{T+1} \to \Sigma^{\mathcal{X}}$ *such that for every* $0 < \delta < 1$, *if*

$$\max(k_c, k_t) \ge \frac{1}{\epsilon_0} \left(\ln |\mathcal{H}| + \ln \frac{T}{\delta} \right), \forall t \in [T],$$

then with probability at least $1 - \delta$ *over* $\{\mathcal{S}_t\}_{t=0}^{T}$, *we have* $\mathcal{M}(\mathcal{S}_0; \mathcal{S}_1, \ldots, \mathcal{S}_T) = f_0$.

New Method: Expanded Chain-of-Thought

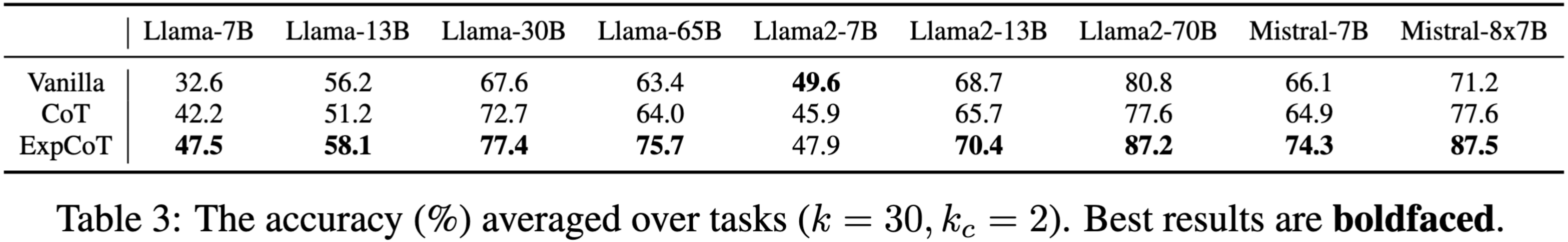
- ExpCoT: view the simple task examples as composite task examples with missing steps and expand them into the CoT format

Examples:

- Composite task CoT example * Rich Humble # -> Poor Proud # -> Proud Poor converted to Step1: * Rich Humble #, Step2: Poor Proud #, Step3: Proud Poor
- Simple task 1 example Dry Lie -> Wet Stand converted to Step1: * Dry Lie, Step2: Wet Stand, Step3: ???
- Simple task 2 example Sad Less # -> Less Sad is converted to Step1: ???, Step2: Sad Less #, Step3: Less Sad

New Method: Expanded Chain-of-Thought

- ExpCoT improves the performance significantly most of the time

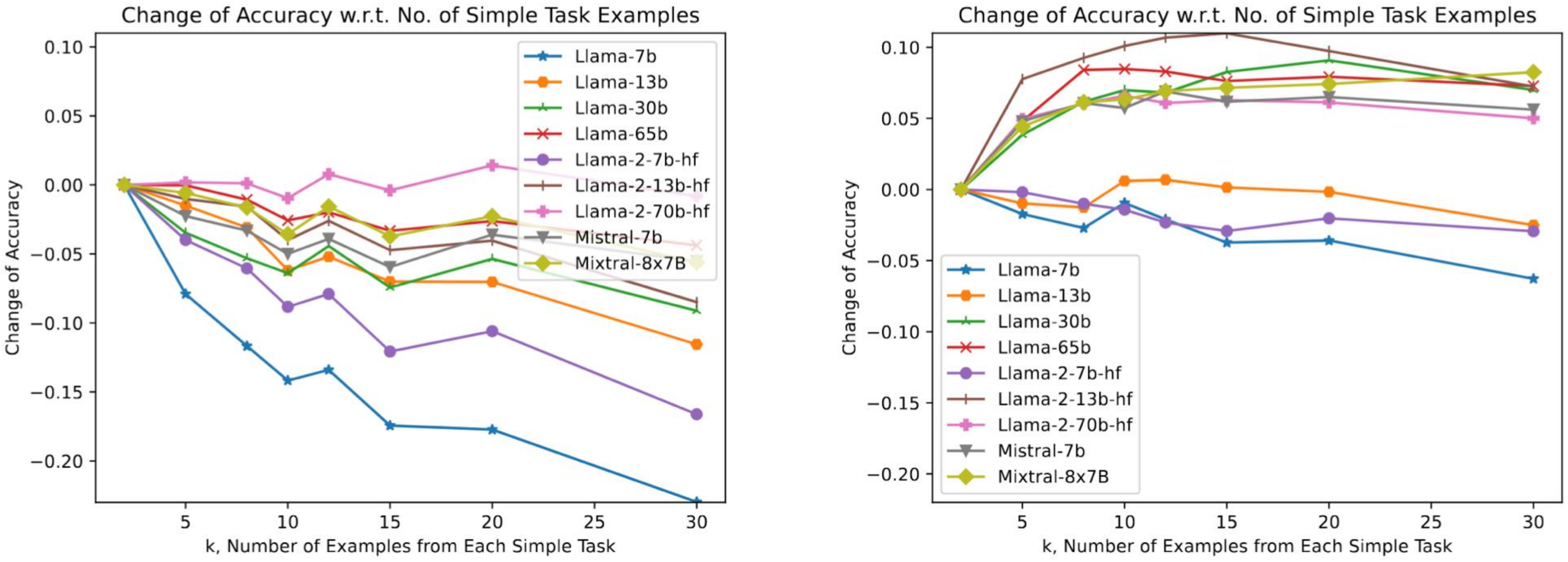


	Llama-7B	Llama-13B	Llama-30B	Llama-65B	Llama2-7B	Llama2-13B	Llama2-70B	Mistral-7B	Mistral-8x7B
Vanilla	32.6	56.2	67.6	63.4	**49.6**	68.7	80.8	66.1	71.2
CoT	42.2	51.2	72.7	64.0	45.9	65.7	77.6	64.9	77.6
ExpCoT	**47.5**	**58.1**	**77.4**	**75.7**	47.9	**70.4**	**87.2**	**74.3**	**87.5**

Table 3: The accuracy (%) averaged over tasks ($k = 30, k_c = 2$). Best results are **boldfaced**.

New Method: Expanded Chain-of-Thought

- ExpCoT mitigates the negative impact of more simple task examples



(a) Vanilla (without ExpCoT)

(b) ExpCoT

Figure 8: The impact of more simple task examples for without or with ExpCoT ($k_c = 2$).

Take Home Messages

- Language models may have limited in-context composition ability
    - Due to failure to recognize composition and align skills with corresponding steps
- A new method aligning example steps helps in-context composition

Further thoughts:

- More recent models like Qwen3, Llama4 etc.
- Other composition tasks
- How to do step alignment for practical data?

Thanks!

Figure credit: mattfyles on reddit/dataisbeautiful