Can language models learn from explanations in context?

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James McClelland, Jane Wang, Felix Hill

DeepMind, London, UK

Presented by Cynthia Chen, Jason Jabbour, Mark Mazumder

April 17, 2023

Presentation Roadmap

- Introduction
- Method
- Experiments
- Discussion

Introduction

- **Language models (LMs)** have been found to be able to perform new tasks by adapting to a few in-context examples
  - In-context example: an example (question + right answer) for a specific task prompt
  - “few-shot”: a few examples are given to guide the model

- Could including few-shot **explanations** of the answers in these examples improve LM performance?

Training: Task instruction, examples + explanations

Task instruction:
Answer these questions by identifying whether the second sentence is an appropriate paraphrase of the first, metaphorical sentence.

Few-shot example #1:
Q: David's eyes were like daggers at Paul when Paul invited his new girlfriend to dance. <-> David had two daggers when Paul invited his new girlfriend to dance.
choice: True
choice: False
A: False

Answer explanation:
Explanation: David's eyes were not literally daggers, it is a metaphor used to imply that David was glaring fiercely at Paul.

4 more examples + explanations

Evaluation: Target question

Target question:
Q: Our whole life we swim against the waves towards the green light of happiness. <-> Our whole life we try to reach happiness.
choice: True
choice: False
A:

Related Work:

- In-context and prompt-based learning:
  - **Min et al., 2022**: Find that ground truth labels don’t have a large effect on model performance, and identify other aspects (label space / distribution) that drive performance
  - **Webson and Pavlick, 2021**: Find limitation in models’ ability to truly understand the meaning of their prompts
- Prompting with explicit instructions
  - **Liu et al., 2021**: Prompting with explicit instructions or task descriptions helps LMs adapt to a task
  - **Wei et al, 2022**: Breaking down the steps of a reasoning process for LMs improve few-shot performance
- Exploring explanations of answers
  - Assessing effects of auxiliary in-context information on performance
  - Large body of work (beyond prompting) on training/tuning with explanations
  - This paper particularly focuses on the effect of **post-answer explanations**

Research Questions

- Can language models benefit from explanations when learning from **examples in-context**?
- Do **few-shot explanations** help the model to **“understand” the task better**? Do explanations of answers improve few-shot task performance?
- More generally, what kind of **in-context learning abilities** do LMs exhibit?

Contributions and Main Findings

- Annotate 40 challenging tasks with explanations of examples
- Evaluate the particular **effect of post-answer explanations** (contrasted to previous work such as Wei et al. who provide chains of reasoning *before* the answer)
- Findings:
  - **Explanations of examples improve the performance of LLMs** when compared to matched control conditions
  - Explanations tuned using a small validation set have even larger performance benefits

Presentation Roadmap

- Introduction
- **Method**
- Experiments
- Discussion

Methods: Data and Model

- **Data**: Selected a set of **40 challenging tasks** from the BIG-Bench dataset
  - Task examples:
    - Inferring goals from actions
    - Reasoning about mathematical induction
    - Reasoning about causality
    - Inferring assumptions behind a statement
- **Model**: set of decoder-only Transformer language models
  - Evaluated a set of models with same context window and trained on the same dataset

Methods: Explanation Annotation

- An author **annotated 15 examples with human explanations**
  - Restricted to multiple choice tasks
- Conducted comparison to ***control explanations*** to account for confounding factors
  - Scrambled explanations
  - True non-explanations
  - Permuted explanations
- **Explanation fine-tuning**
  - Selected examples to build a 5-shot prompt greedily, to evaluate the benefit of selecting the best examples
  - Hand-tuned explanations to better understand the potential benefits of more optimal “expert” explanations
  - Tuning performed on small validation set, but tested on larger set of task examples

Methods: Evaluation

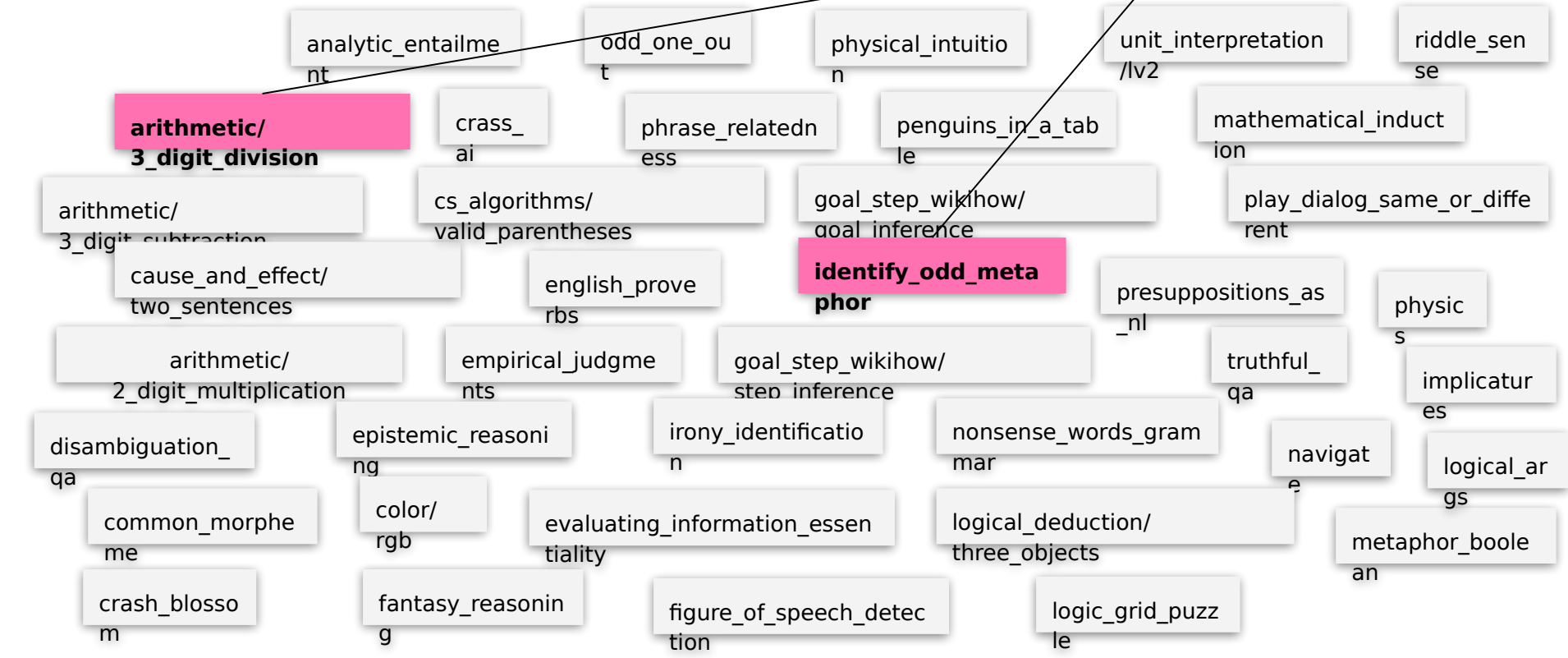
- **Modeled dependencies of results using hierarchical logistic regression**
  - Nested dependencies and heterogeneous structure
  - Factors to take into account:
    - Task difficulty
    - Shared content between prompts
- Model each unique effect at each level of the hierarchy
  - Estimates the unique added contribution of a prompt component (eg: few-shot example, explanation, etc) to the performance

Presentation Roadmap

- Introduction
- Method
- **Experiments**
- Discussion

Benefit of Adding Explanatic

Look at **2 out of 15 prompts** drawn from the 40 tasks



Examples Alone (no explanations)

`arithmetic/3_digit_division:`

Question: "What is 688 divided by 1?"
Answer: "688"

(choices omitted for brevity)

`identify_odd_metaphor:`

Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."

Answer Likelihoods

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"

(choices omitted for brevity)

`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."

Each experiment evaluates the relative probability of **this sequence** of tokens relative to all sequences in the **multiple choice answers** (**not** all possible sequences of tokens)

i.e., model's likelihood of **each answer option** after conditioning on the prompt and question

Adding Explanations

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"

`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."

Adding Explanations

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"
Explanation: "Dividing a number by 1 always yields the same number."

`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."
Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"

Adding Explanations

arithmetic/3_digit_division:
Question: "What is 688 divided by 1?"
Answer: "688"
Explanation: "Dividing a number by 1 always yields the same number."

identify_odd_metaphor:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."
Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"

Does the **likelihood ratio** of the correct answer **increase** for the **next question**, **when explanations are added** to previous examples?

How much do explanations help?

Is it the explanation itself or something else (relevant words, syntactic structure, …)

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"
Explanation: "Dividing a number by 1 always yields the same number."

`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."
Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"

vs 3 Control Conditions

How much do explanations help?

Is it the explanation itself or something else (relevant words, syntactic structure, …)

Control Experiment 1

Relevant but non-explanatory

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"
Explanation: "Dividing a number by 1 always yields the same number."
True non-explanation: "688 is an even number, which means it is divisible by 2."

`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."
Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"
True non-explanation: "Cognitive psychology involves the study of people, and sometimes comparative study of other animals to determine the similarities and differences."

How much do explanations help?

Is it the explanation itself or something else (relevant words, syntactic structure, …)

Control Experiment 2

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"
Explanation: "Dividing a number by 1 always yields the same number."
True non-explanation: "688 is an even number, which means it is divisible by 2."
Other item explanation: "We know 450 = 9 * 50, so we can rewrite as 522 / 9 = 450 / 9 + 72 / 9 = 50 + 8 = 58."

Randomly chosen from other examples in the prompt dataset

`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."
Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"
True non-explanation: "Cognitive psychology involves the study of people, and sometimes comparative study of other animals to determine the similarities and differences."
Other item explanation: "This sentence does not use a metaphor, while the others use container-relevant metaphors (fullest, crammed, contained)."

How much do explanations help?

Is it the explanation itself or something else (relevant words, syntactic structure, …)

Control Experiment 3

`arithmetic/3_digit_division`:

Question: "What is 688 divided by 1?"

Answer: "688"

Explanation: "Dividing a number by 1 always yields the same number."

True non-explanation: "688 is an even number, which means it is divisible by 2."

Other item explanation: "We know 450 = 9 * 50, so we can rewrite as 522 / 9 = 450 / 9 + 72 / 9 = 50 + 8 = 58."

Scrambled explanation: "number. the Dividing same always 1 yields a number by"

`identify_odd_metaphor`:

Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."

Answer: "It is important how you package your ideas."

Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"

True non-explanation: "Cognitive psychology involves the study of people, and sometimes comparative study of other animals to determine the similarities and differences."

Other item explanation: "This sentence does not use a metaphor, while the others use container-relevant metaphors (fullest, crammed, contained)."

Scrambled explanation: "metaphors other life, while do. Packaging the people, and resurrection) infancy, (breathing not to does apply"

How much do explanations help?

Is it the explanation itself or something else (relevant words, syntactic structure, …)

`arithmetic/3_digit_division`:
Question: "What is 688 divided by 1?"
Answer: "688"
Explanation: "Dividing a number by 1 always yields the same number."
True non-explanation: "688 is an even number, which means it is divisible by 2."
Other item explanation: "We know 450 = 9 * 50, so we can rewrite as 522 / 9 = 450 / 9 + 72 / 9 = 50 + 8 = 58."
Scrambled explanation: "number. the Dividing same always 1 yields a number by"

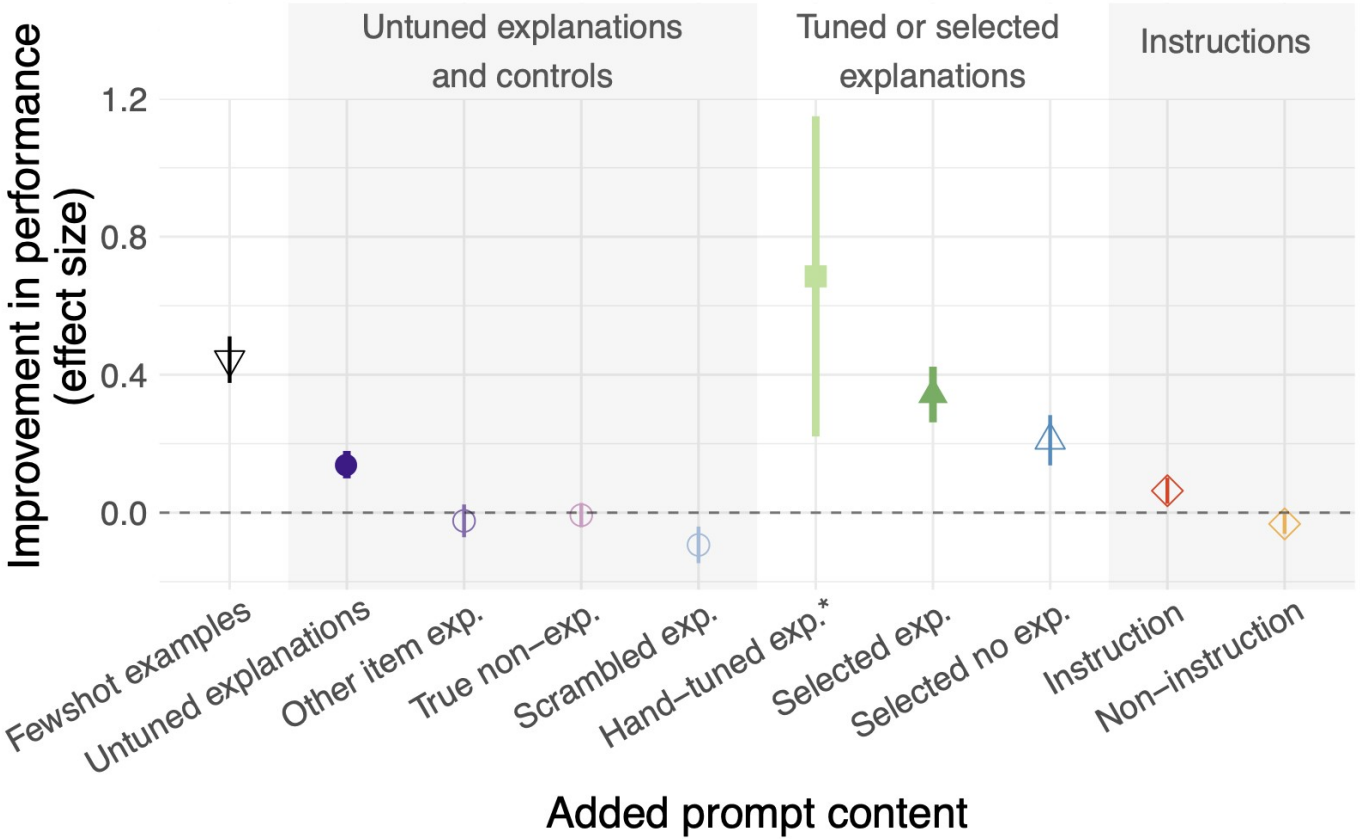
These are **untuned explanations:**

Annotated by hand, **without looking at LM probabilities**

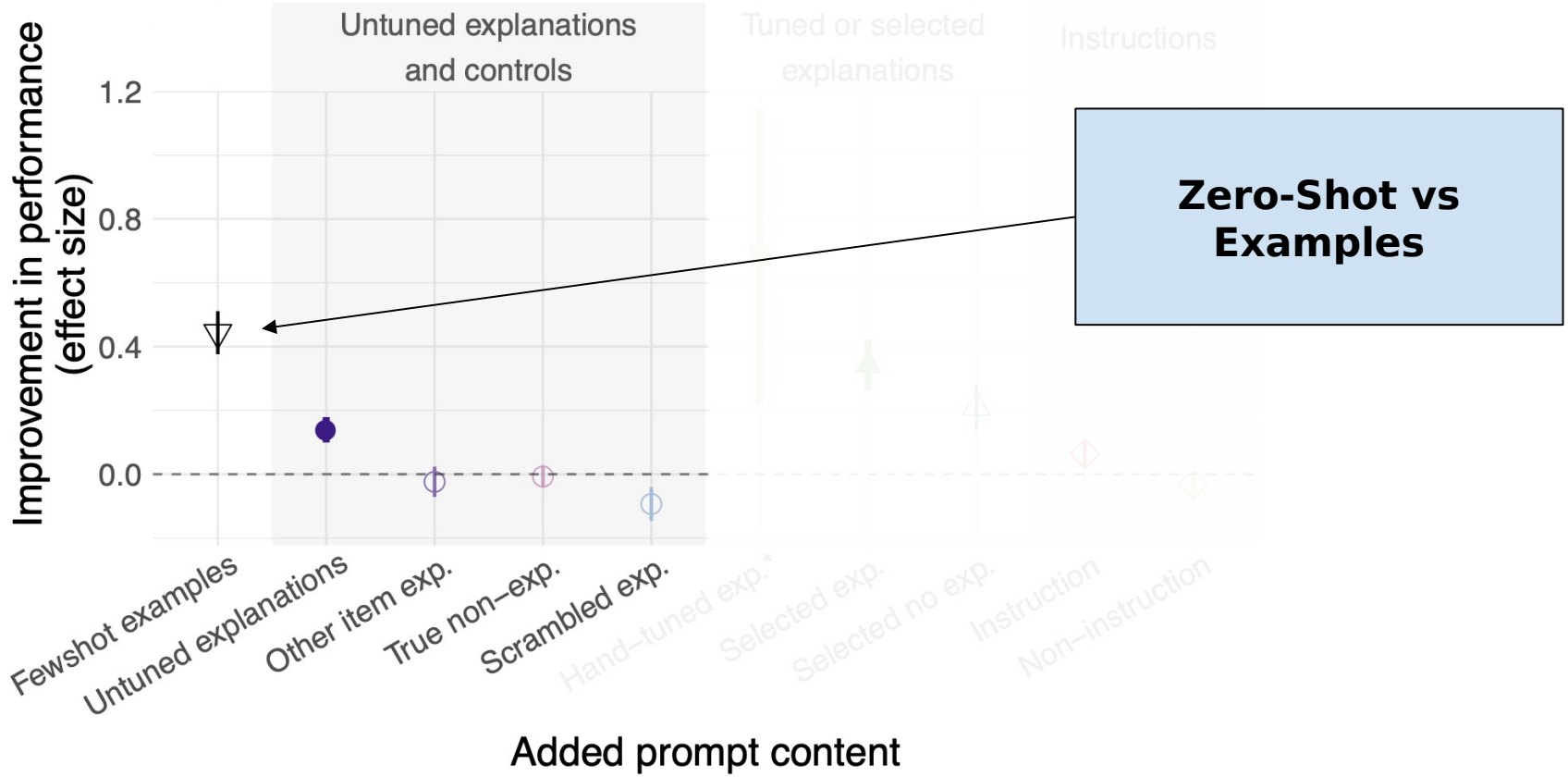
`identify_odd_metaphor`:
Question: "Which of the following sentences relating to ideas does not use metaphorical language that could also be applied to people? choice: He breathed new life into that idea. choice: It is important how you package your ideas. choice: Cognitive psychology is still in its infancy. choice: That's an idea that ought to be resurrected."
Answer: "It is important how you package your ideas."
Explanation: 'Packaging does not apply to people, while the other metaphors (breathing life, infancy, and resurrection) do.'"
True non-explanation: "Cognitive psychology involves the study of people, and sometimes comparative study of other animals to determine the similarities and differences."
Other item explanation: "This sentence does not use a metaphor, while the others use container-relevant metaphors (fullest, crammed, contained)."
Scrambled explanation: "metaphors other life, while do. Packaging the people, and resurrection) infancy, (breathing not to does apply"

Hierarchical Regression: **quantify** unique added contribution (on Gopher-280B)

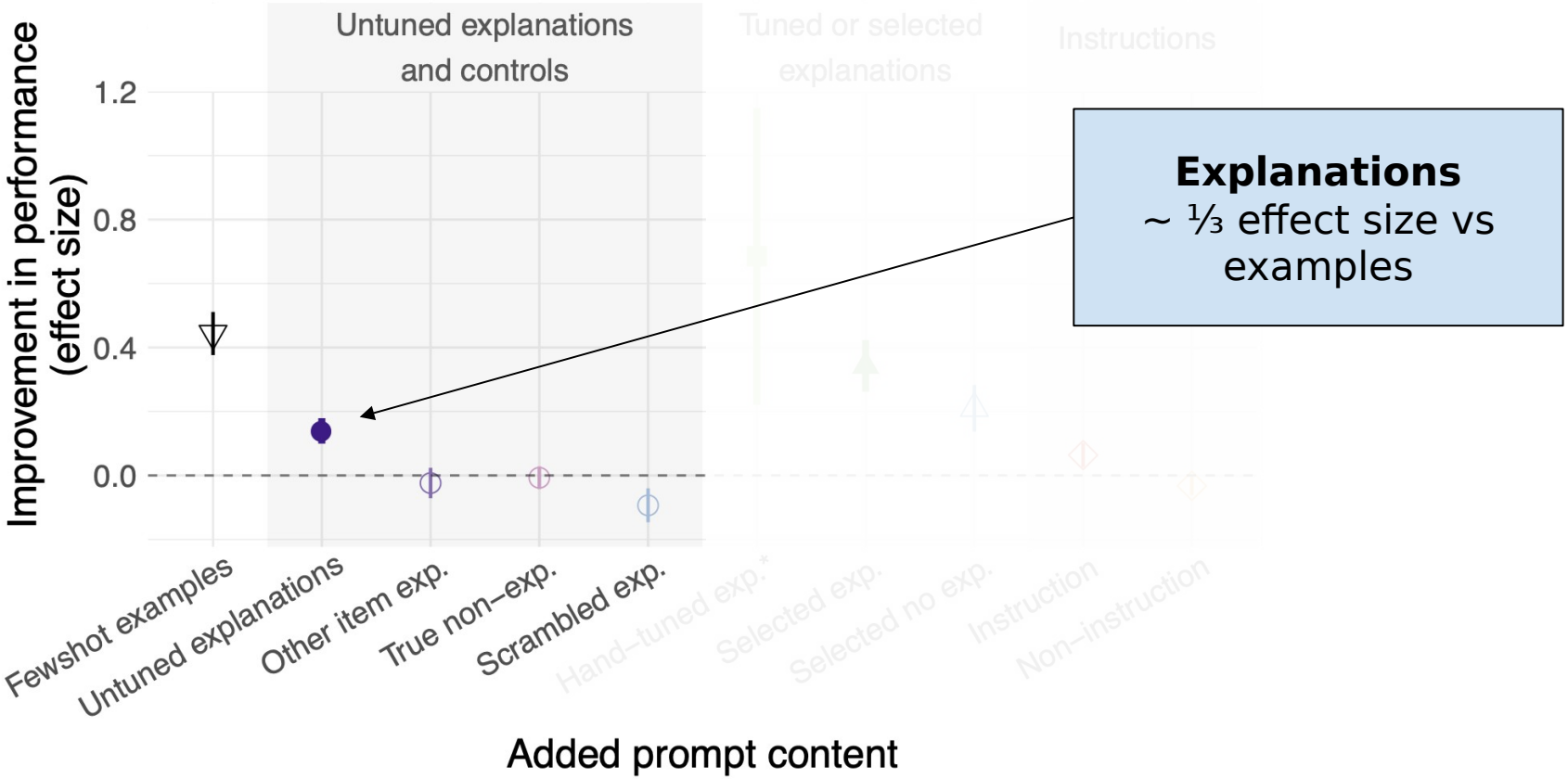
Effect Size



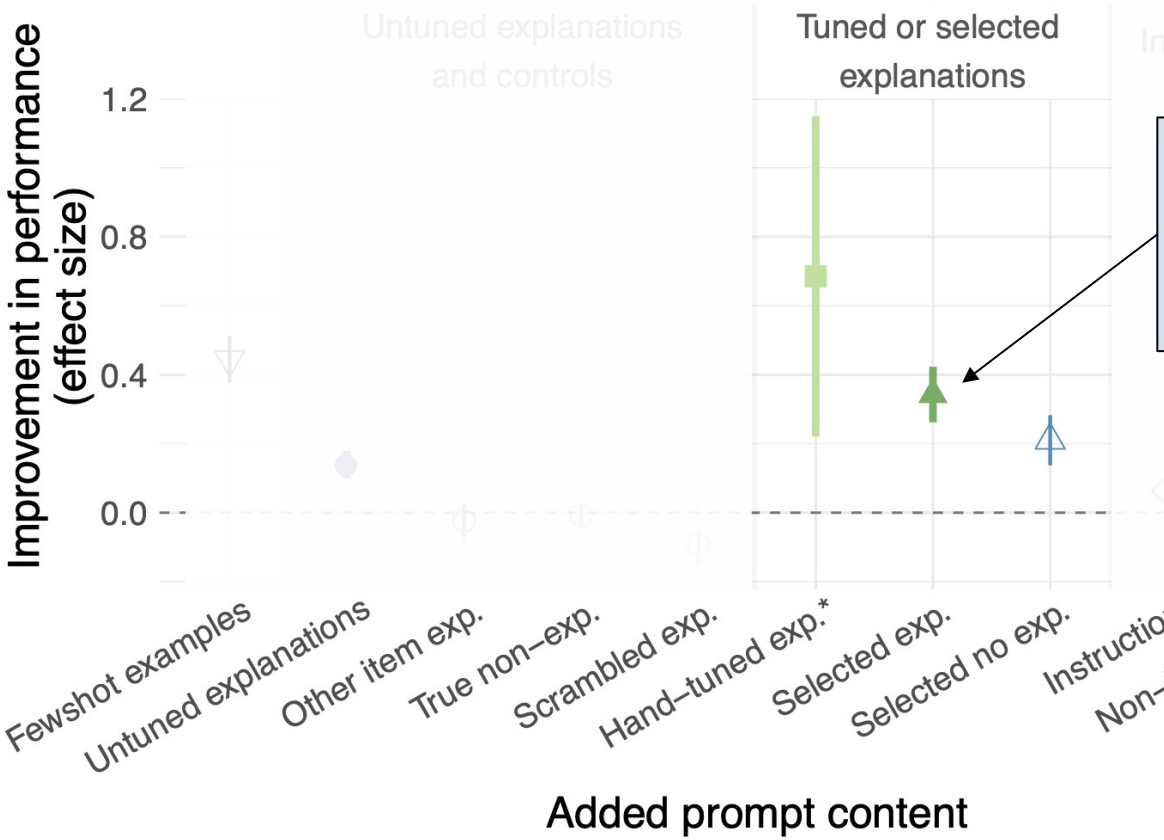
Effect Size


Untuned explanations and controls
Tuned or selected explanations
Instructions
Improvement in performance (effect size)
1.2
0.8
0.4
0.0
Zero-Shot vs Examples
Fewshot examples
Untuned explanations
Other item exp.
True non–exp.
Scrambled exp.
Hand–tuned exp.*
Selected exp.
Selected no exp.
Instruction
Non–instruction
Added prompt content

Effect Size


Untuned explanations and controls
Tuned or selected explanations
Instructions
Improvement in performance (effect size)
1.2
0.8
0.4
0.0
Fewshot examples
Untuned explanations
Other item exp.
True non–exp.
Scrambled exp.
Hand–tuned exp.*
Selected exp.
Selected no exp.
Instruction
Non–instruction
Added prompt content
Explanations
~ ⅓ effect size vs examples

Effect Size



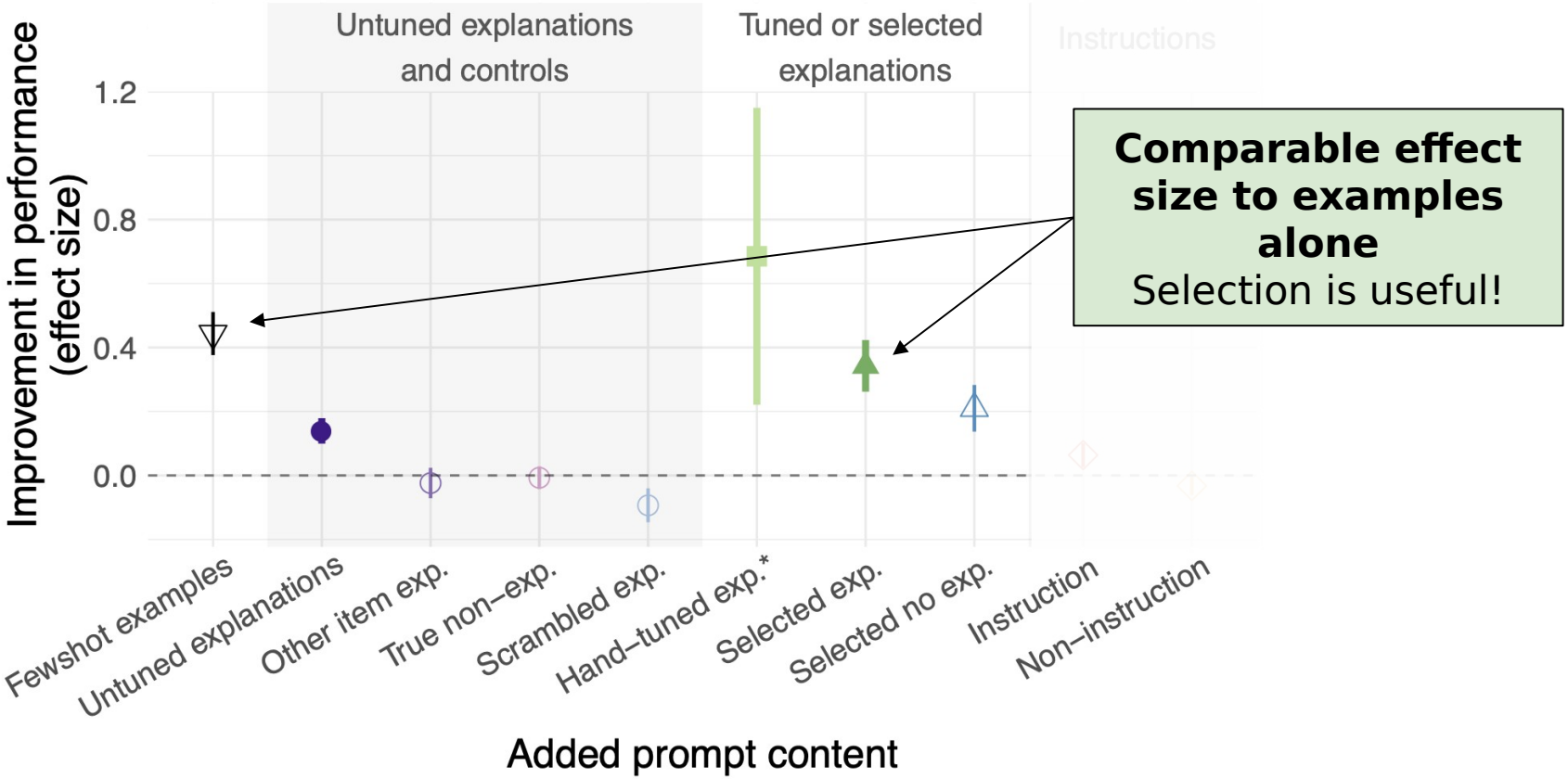
Selection

Greedily choose the best k-shot **chain of prompts** to maximize likelihood of remaining (n-k) examples, up to k=5

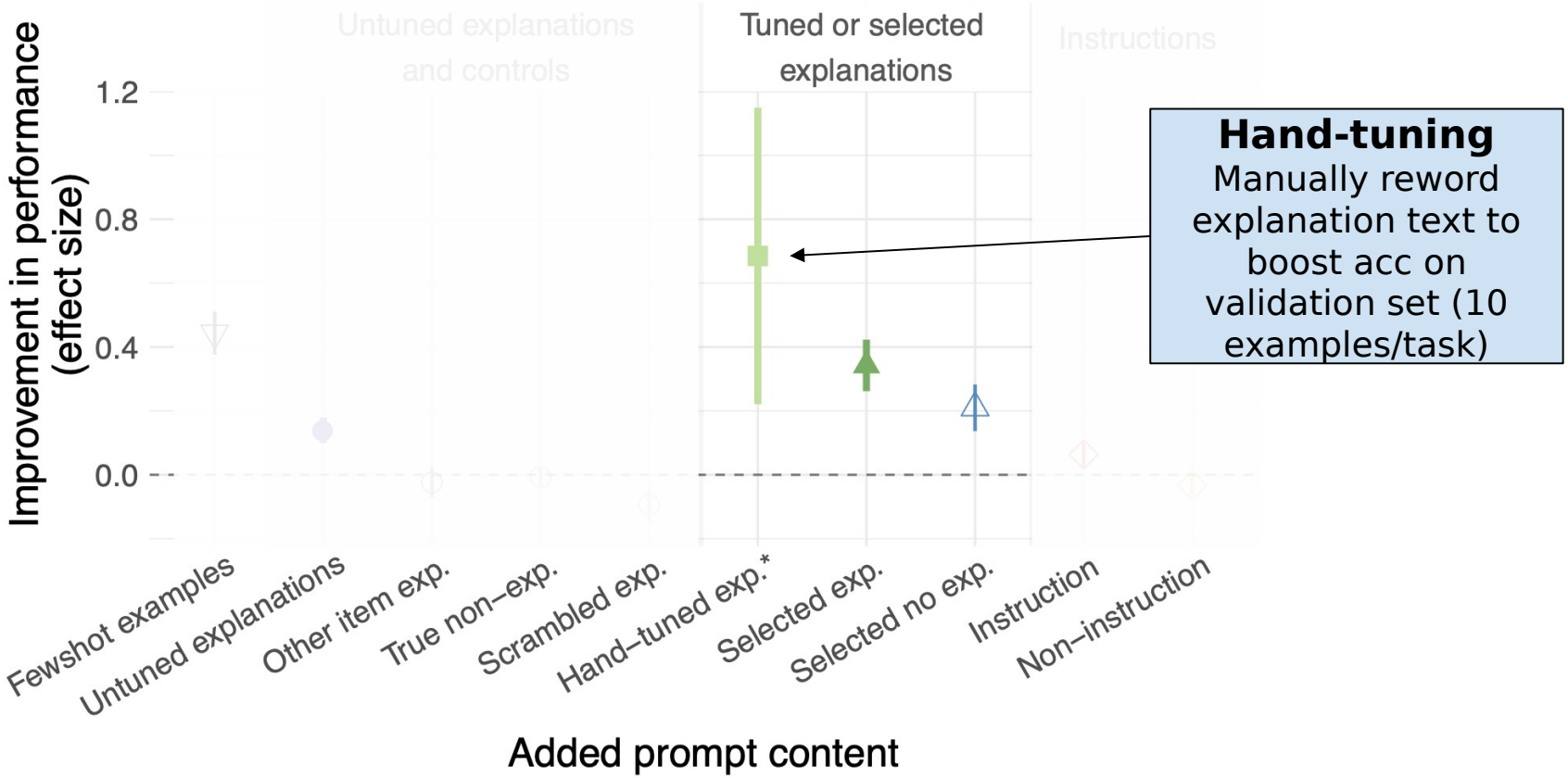
Only 15 examples per task:

- 5-shot prompt selection
- 10 validation examples

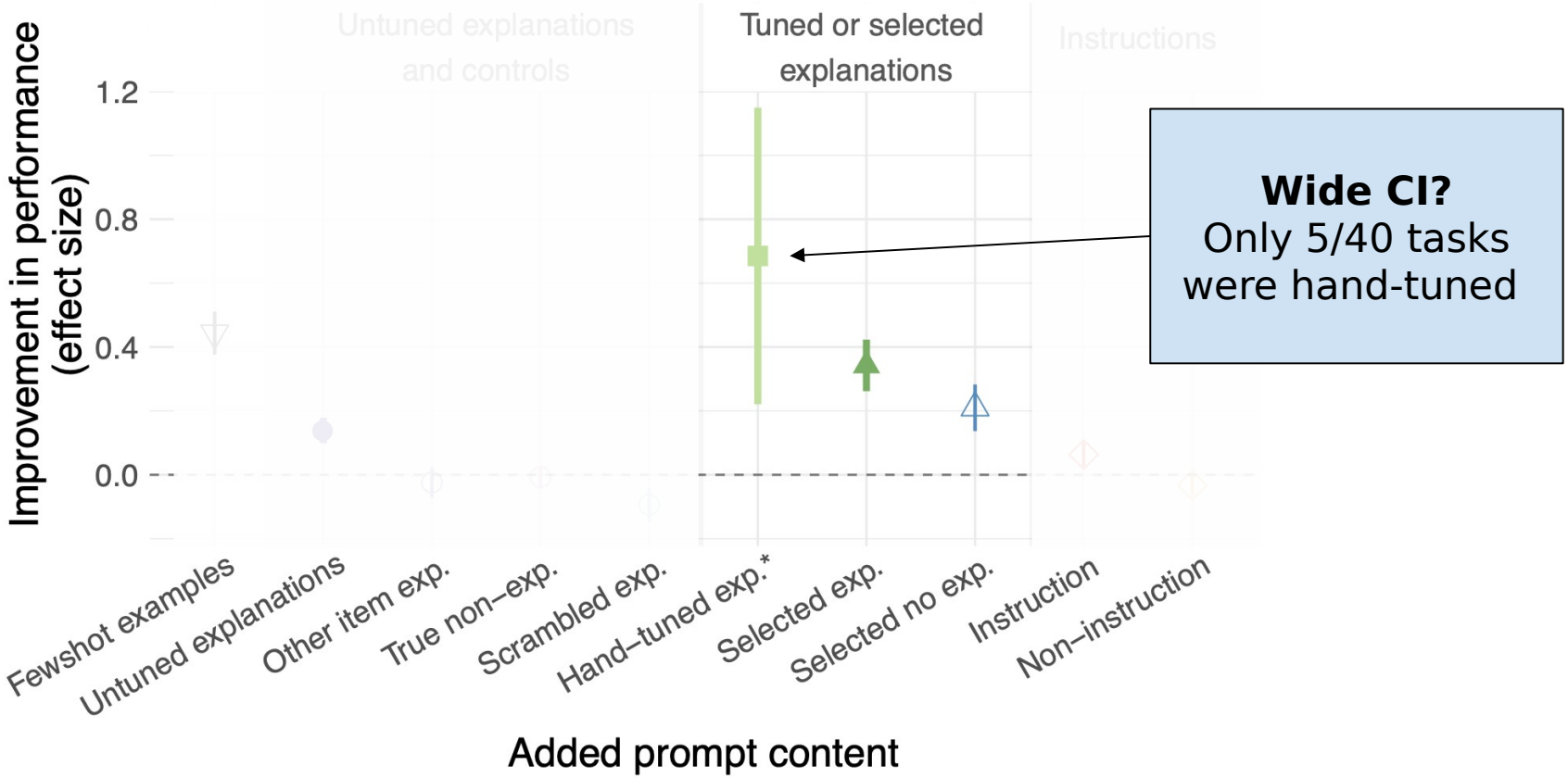
Effect Size


Untuned explanations and controls
Tuned or selected explanations
Instructions
Improvement in performance (effect size)
1.2
0.8
0.4
0.0
Fewshot examples
Untuned explanations
Other item exp.
True non–exp.
Scrambled exp.
Hand–tuned exp.*
Selected exp.
Selected no exp.
Instruction
Non–instruction
Added prompt content
Comparable effect size to examples alone
Selection is useful!

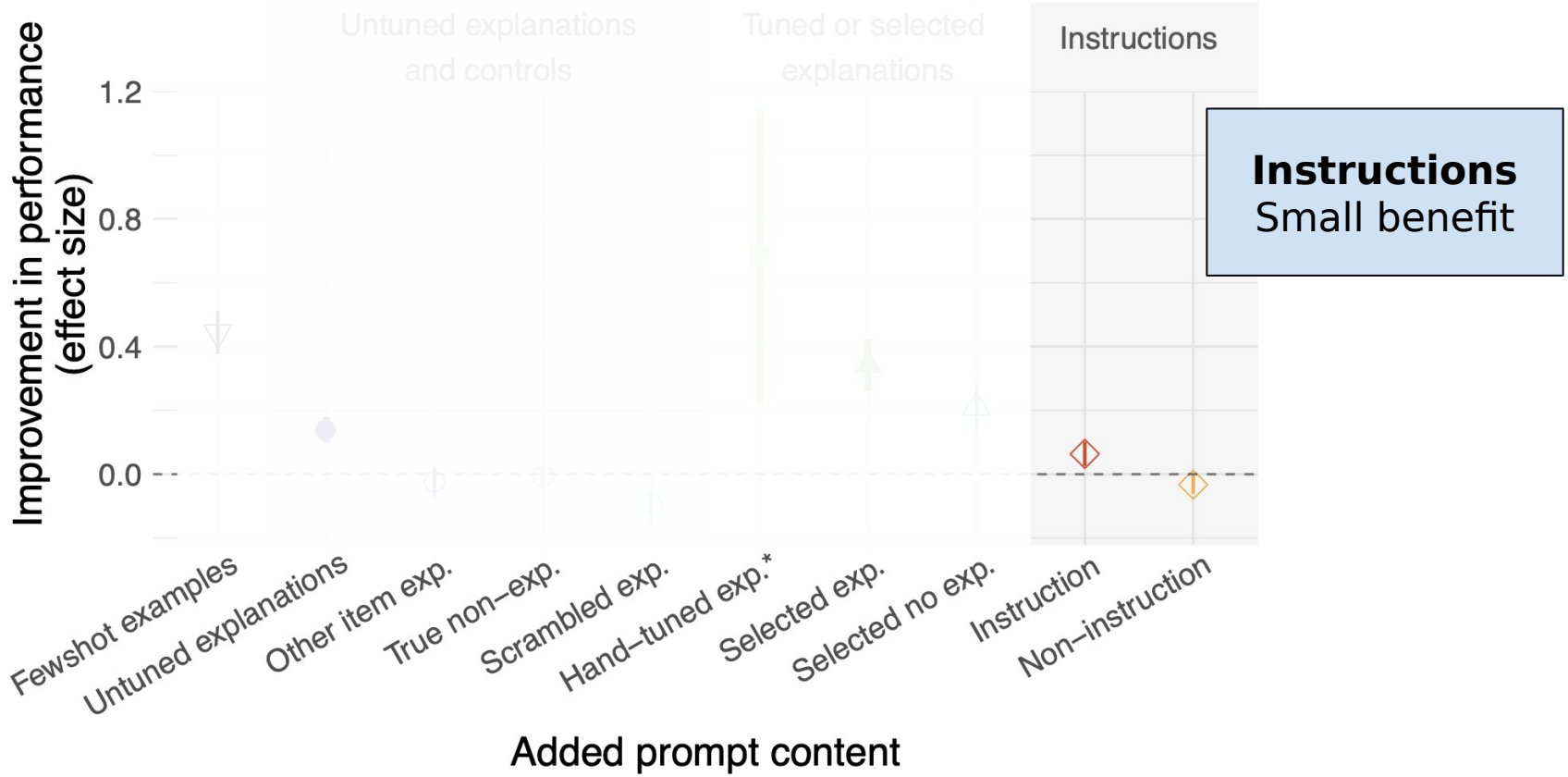
Effect Size


Tuned or selected explanations
Improvement in performance (effect size)
1.2
0.8
0.4
0.0
Fewshot examples
Untuned explanations
Other item exp.
True non–exp.
Scrambled exp.
Hand–tuned exp.*
Selected exp.
Selected no exp.
Instruction
Non–instruction
Added prompt content
Hand-tuning
Manually reword explanation text to boost acc on validation set (10 examples/task)

Effect Size


Tuned or selected explanations
Improvement in performance (effect size)
1.2
0.8
0.4
0.0
Fewshot examples
Untuned explanations
Other item exp.
True non–exp.
Scrambled exp.
Hand–tuned exp.*
Selected exp.
Selected no exp.
Instruction
Non–instruction
Added prompt content
Wide CI?
Only 5/40 tasks were hand-tuned

Effect Size

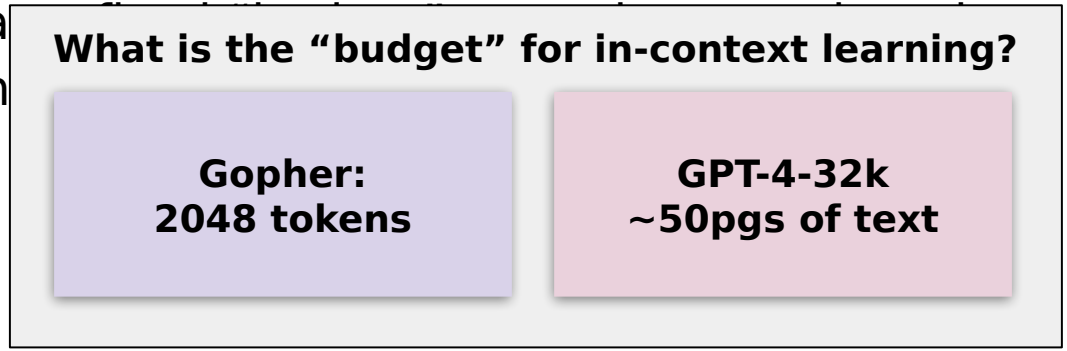

Untuned explanations and controls
Tuned or selected explanations
Instructions
Improvement in performance (effect size)
1.2
0.8
0.4
0.0
Fewshot examples
Untuned explanations
Other item exp.
True non–exp.
Scrambled exp.
Hand–tuned exp.*
Selected exp.
Selected no exp.
Instruction
Non–instruction
Added prompt content
Instructions
Small benefit

Presentation Roadmap

- Introduction
- Method
- **Experiments**
- Discussion

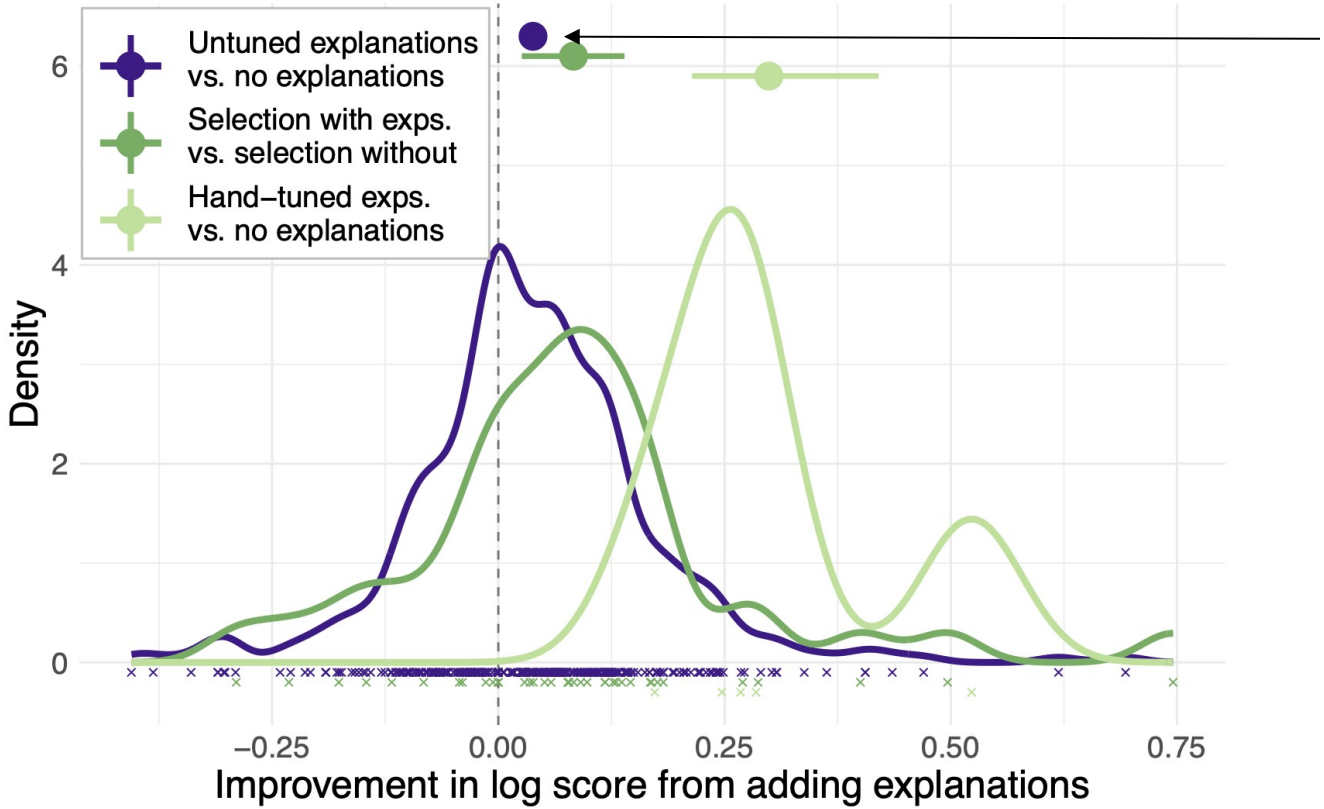
Potential impacts for LLM applications

- Adding untuned few-shot explanations slightly outperforms examples alone
- Selection and rewriting using a **validation set** boosts gains as much as examples
- **In-context** explanations vs finetuning LLM **weights** on a task
  - Explanations are quicker + easier to deploy
  - But ha
  - Can th



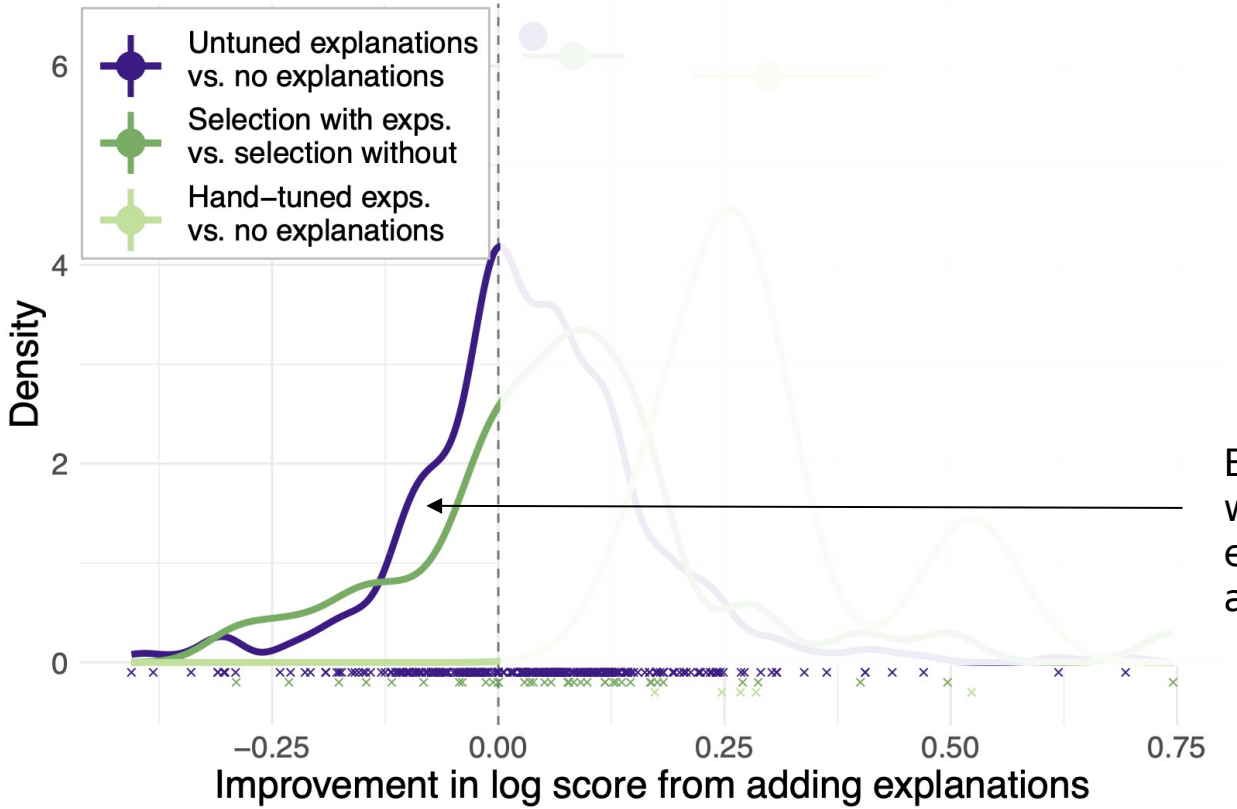
Source:

Selection vs untuned explanations



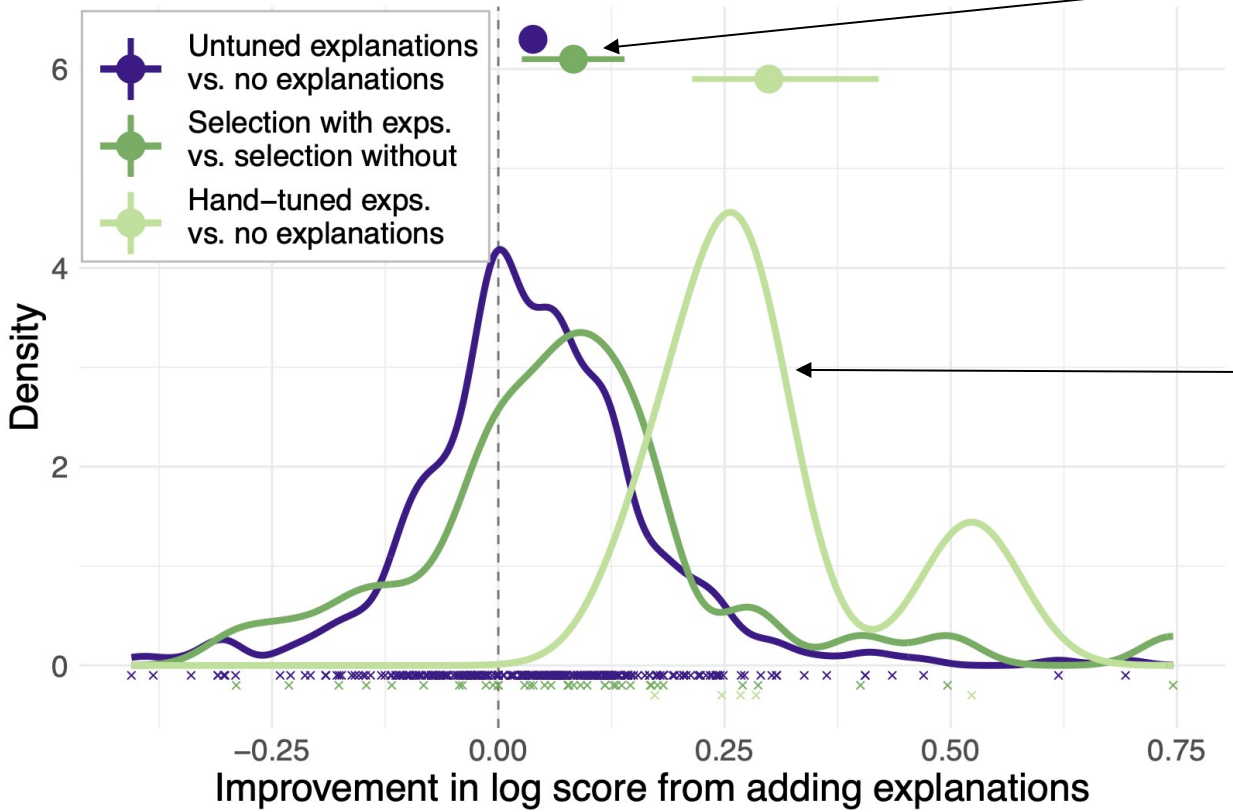
Untuned explanations are slightly beneficial on average

Selection vs untuned explanations



But many instances where untuned explanations hurt accuracy

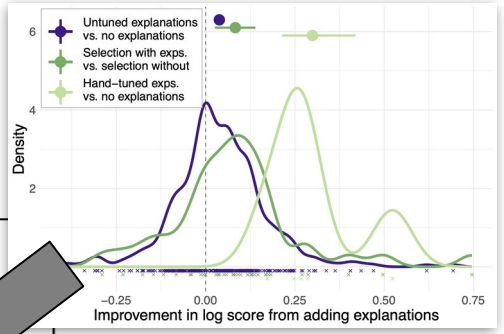
Selection vs untuned explanations

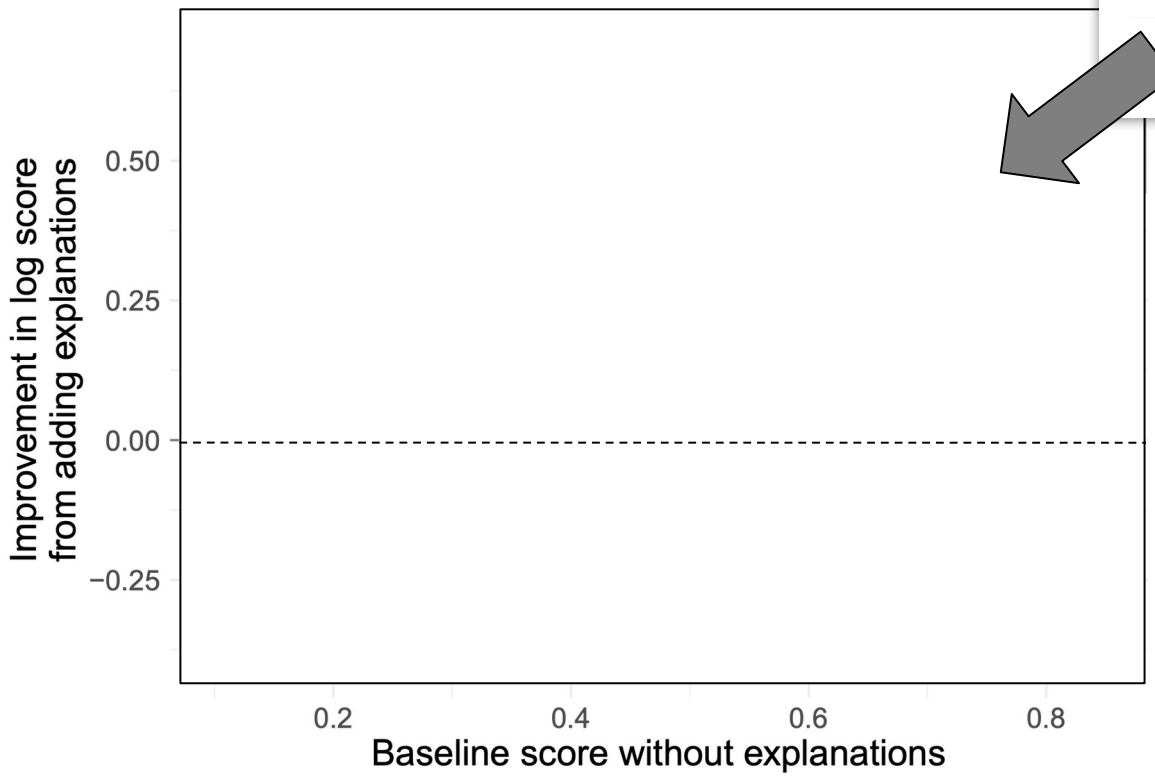


Benefit of selection is clearer

Not enough data on hand-tuning to be persuasive?

When do explanations help?

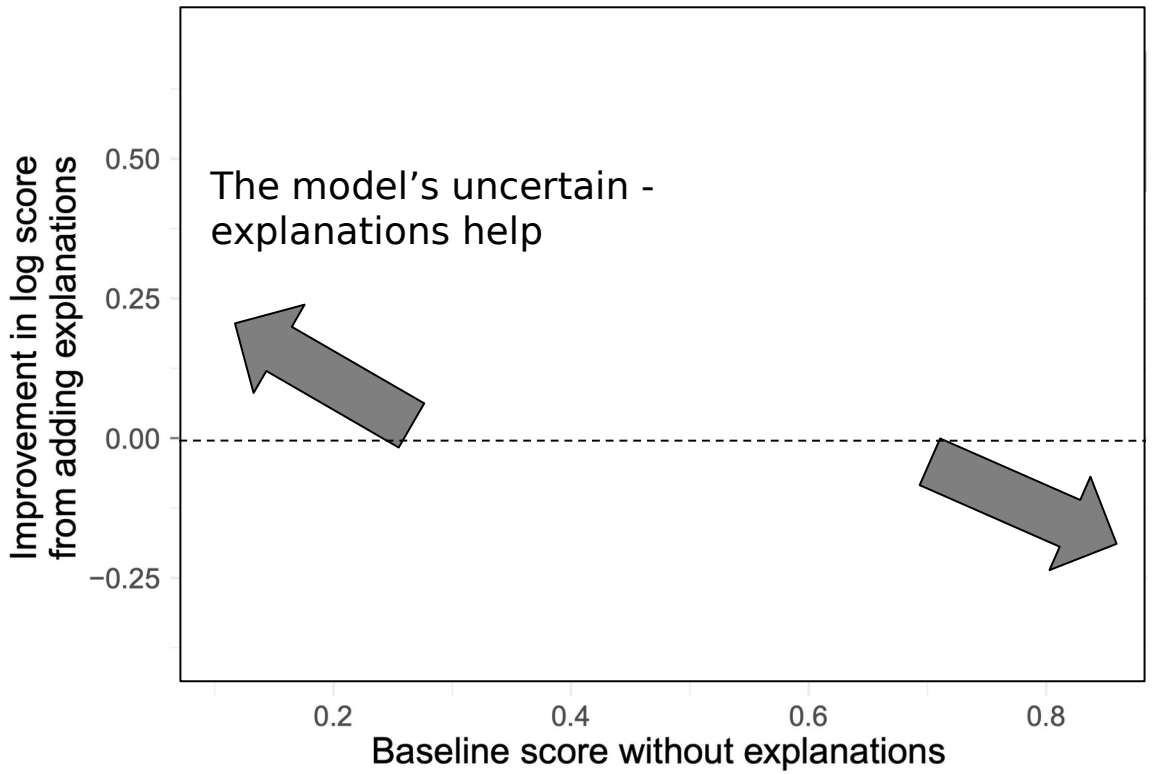




Addition of explanations vs model's **prior** confidence **with examples alone**

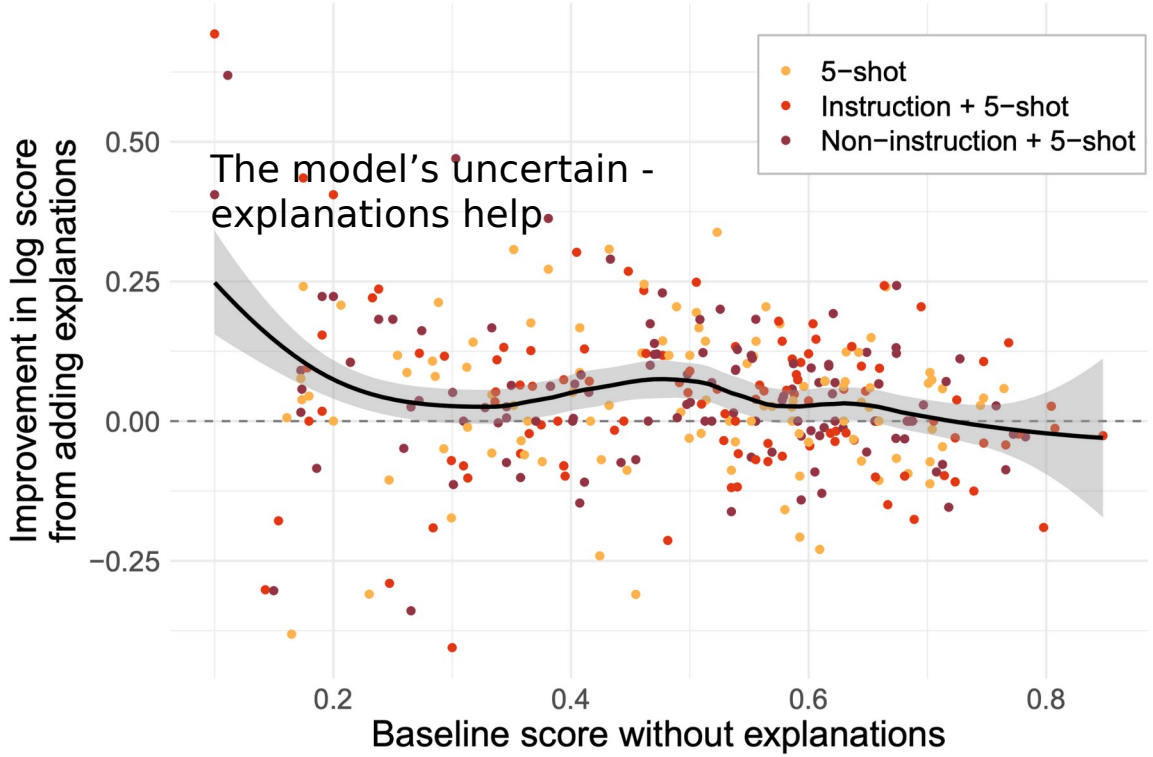
Baseline: without explanations (baseline is **not** zero-shot)

When do explanations help?



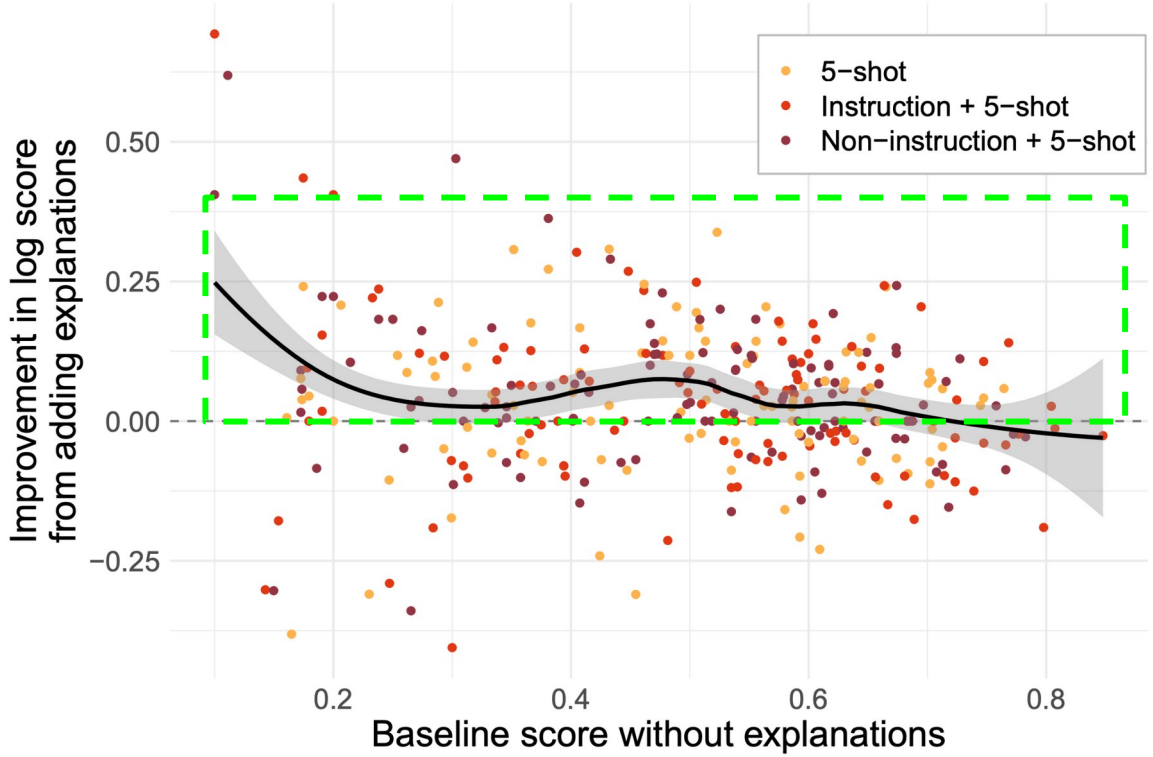
The model’s already correct without explanations - do explanations hurt?

When do explanations help?



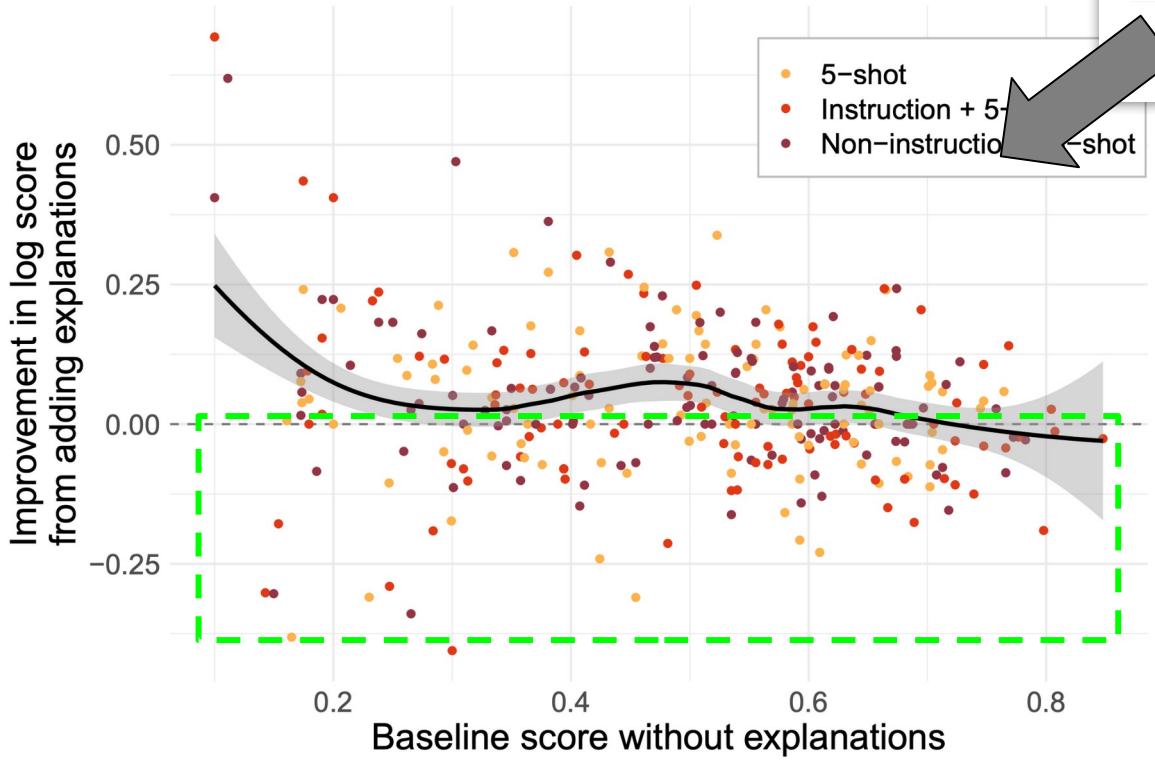
The model's already correct without explanations - do explanations hurt?

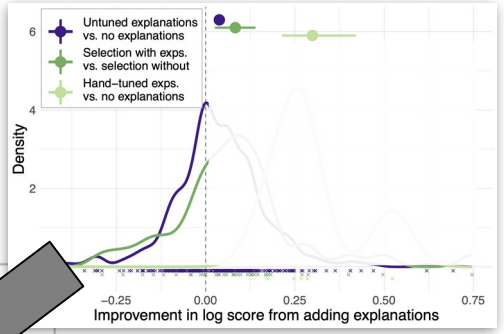
When do explanations help?



Explanations usually help: higher density regardless of baseline confidence

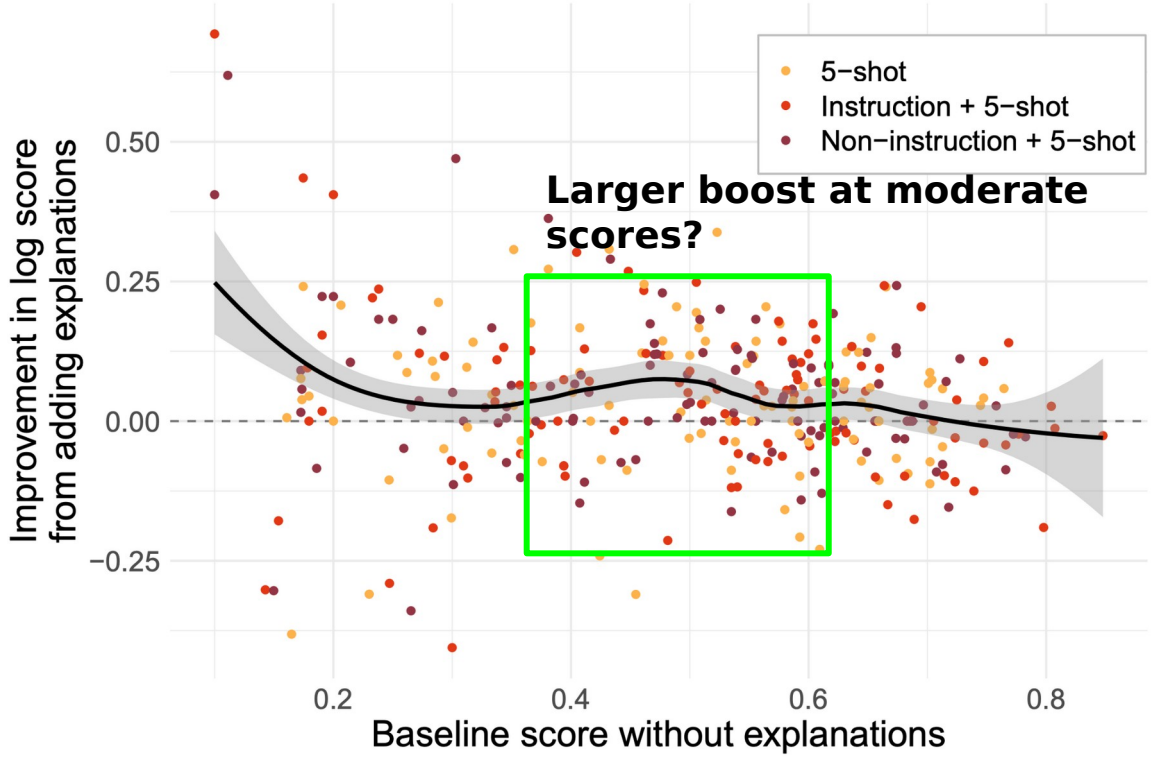
When do explanations help?

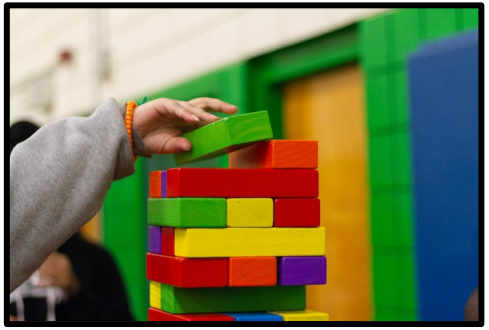




But not always: scores also drop with explanations, **but uniformly(ish)**

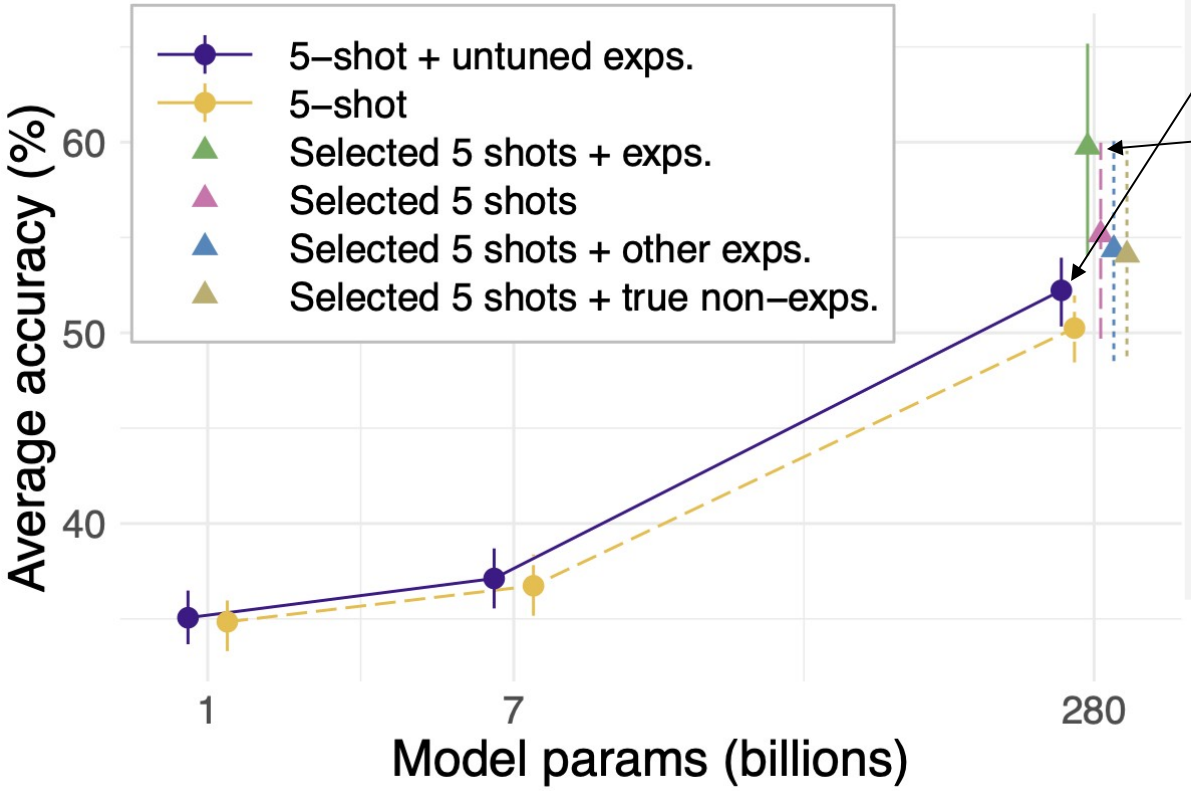
When do explanations help?





"**Zone of proximal effect**?" (what a learner can do **on the cusp of their prior capabilities,** with/without teacher assistance)

Is Scale Necessary?

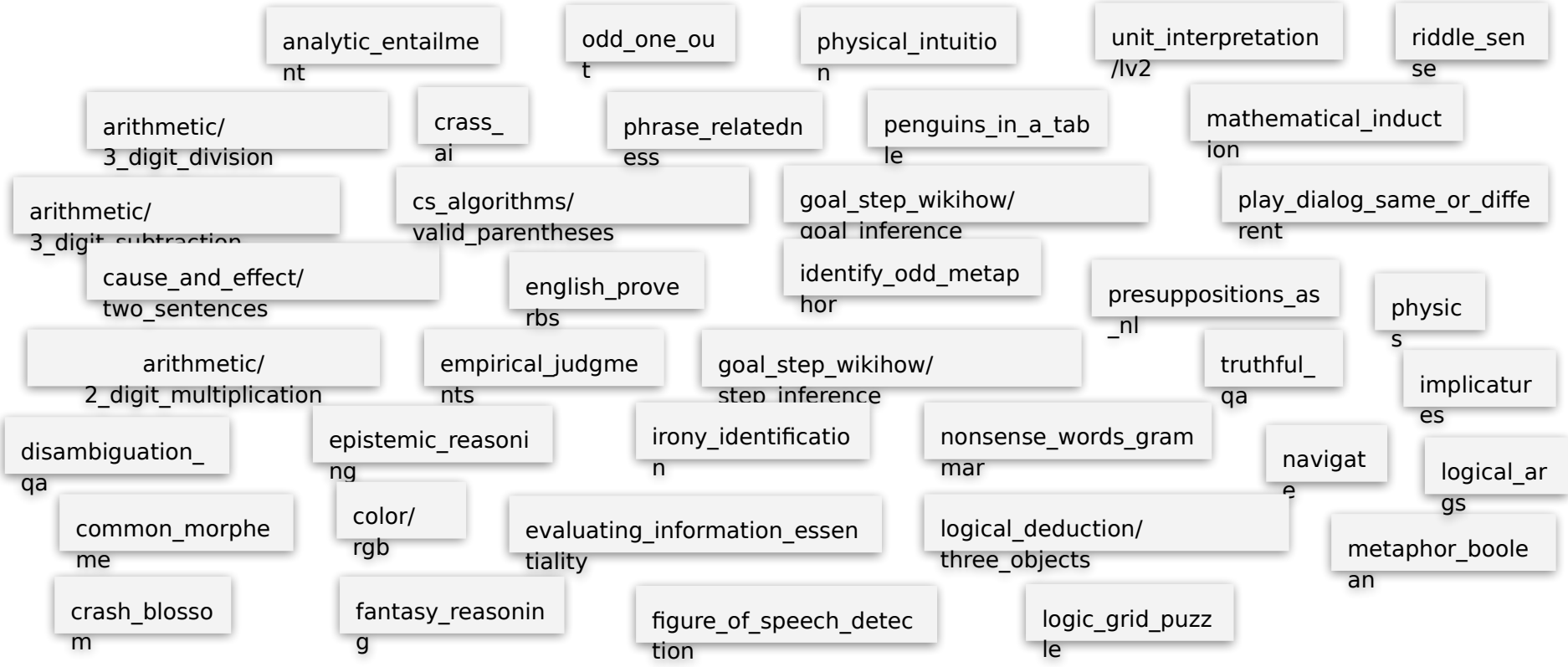


- Untuned explanations only help modestly **for the largest model**
  - **Selection** emerges from the pack
- But **no benefits @ 1B, 7B**
- Consistent w/prior work on scaling:
  - Brown ‘20: **sharp transition** for arithmetic
  - Wei ‘21: only large LMs generalize

Task Specific Improvements

Does **improvement** provided by explanations **differ** between **types** of **tasks**?

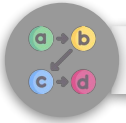
Task Diversity



Clustering Tasks into Categories


Casual


Computer Science



Logic


Negation


OOD

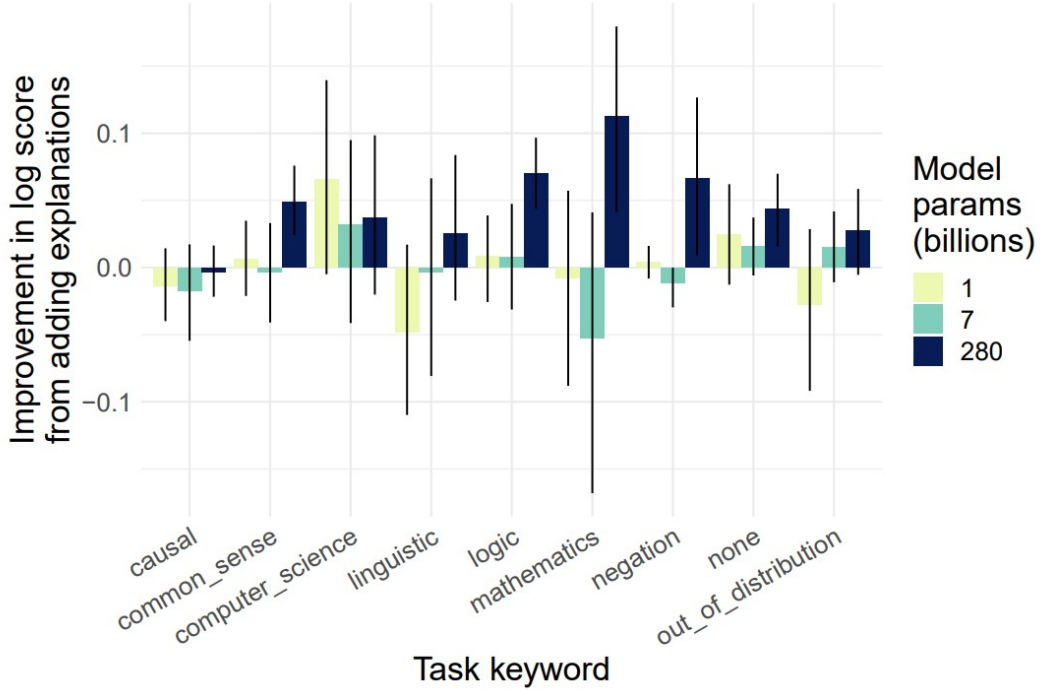

Common Sense


Linguistic

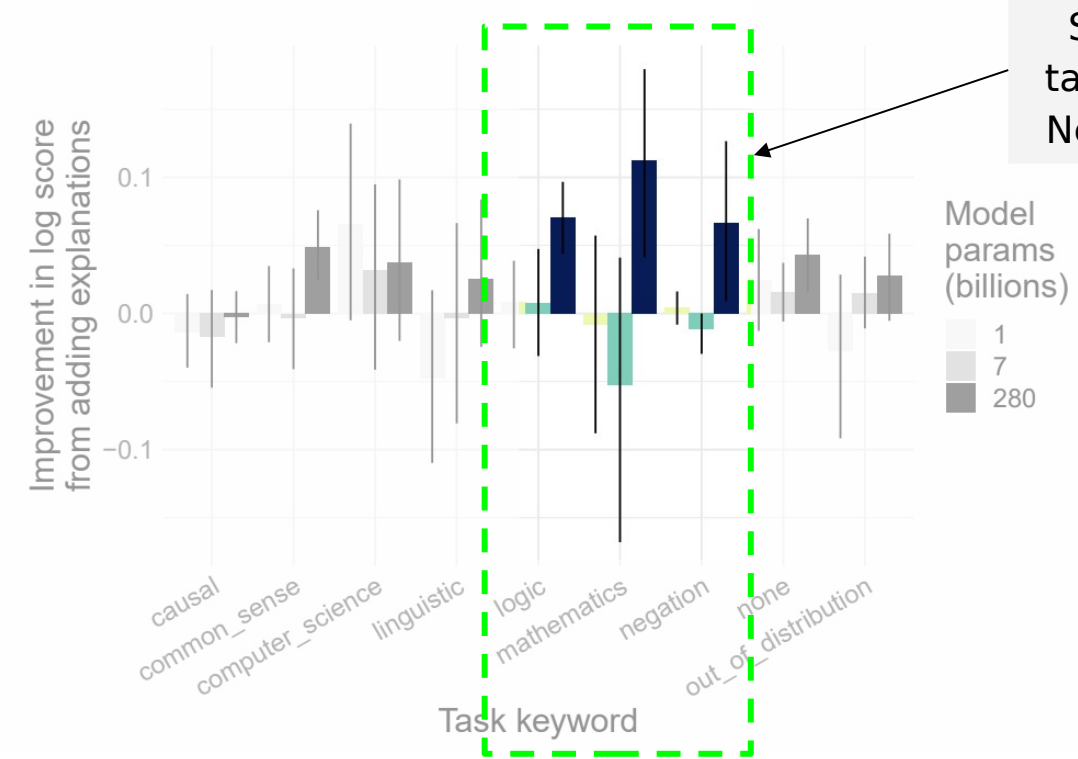

Mathematics


None

Not all Tasks are Created Equal



Not all Tasks are Created Equal



Semantics of explanations for tasks about Logic, Mathematics, Negation might be better suited for LM

Presentation Roadmap

- Introduction
- Method
- Experiments
- Discussion

Discussion

Can explanations improve few-shot

Why are post-answer explanations

What do their results imply about LMs' abilities for in-

How do explanations relate to

How does their work relate to human language processing?

Discussion

Can explanations improve few-shot

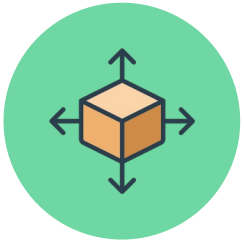
Why are post-answer explanations

What do their results imply about LMs' abilities for in-

How do explanations relate to

How does their work relate to human language processing?

Can explanations improve few-shot learning?



Scale

Quality

Discussion

Can explanations improve few-shot

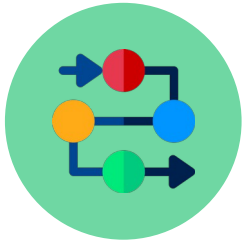
Why are post-answer explanations

What do their results imply about LMs' abilities for in-

How do explanations relate to

How does their work relate to human language processing?

Why are post-answer explanations interesting?



Holistic vs Chain-of-Reasoning

Test Time Implications

Discussion

Can explanations improve few-shot

Why are post-answer explanations

What do their results imply about LMs' abilities for in-

How do explanations relate to

How does their work relate to human language processing?

What do their results imply about LMs’ abilities for in-context learning?

Example & Explanation Relationships

Learning vs Recalling

Discussion

Can explanations improve few-shot

Why are post-answer explanations

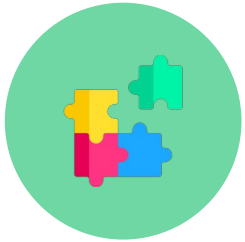
What do their results imply about LMs' abilities for in-

How do explanations relate to

How does their work relate to human language processing?

How do explanations relate to instructions?

Instructions Improvement in Prior Work was Successful



Instruction & Explanations Complementary

Discussion

Can explanations improve few-shot

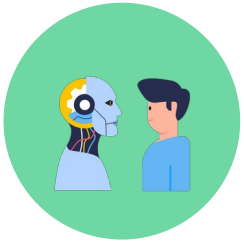
Why are post-answer explanations

What do their results imply about LMs' abilities for in-

How do explanations relate to

How does their work relate to human language processing?

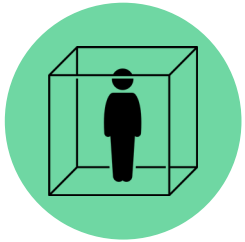
How does their work relate to human language processing?



Humans

≠

LMs



No Broader Context

Class Discussion

- Is the dataset large (~600 samples)/diverse (40 tasks) enough to support their conclusions?
- Is the degree of improvement offered by explanations significant?
- How does this compare to other **in-context learning** methods?
  - Chain-of-thought [Wei]
  - Soft prompts (changing the embedding of the prompt) [Leister]
  - Finetuning + Distillation (Alpaca) [Taori]
- How could explanations benefit smaller models and not only large models?