Effective Use of Explanations in Few-Shot Prompting for Textual Reasoning

The University of Texas at Austin

TAUR LAB

Xi Ye
The University of Texas at Austin

2023/03

Prompting with Explanations

Prompt

Q: Alice has 5 apples. Bob has 2 apples. How many apples do they have together?
A: The answer is 7.

Q: Charlie has 4 toys. Dianna has twice as much as Charlie. How many toys do they have together?

GPT-3

Output

A: The answer is 12.

Performance on GSM
19%

Q: Alice has 5 apples. Bob has 2 apples. How many apples do they have together?
A: They have 5 + 2 = 7 apples together. The answer is 7.

Q: Charlie has 4 toys. Dianna has twice as much as Charlie. How many toys do they have together.

GPT-3

A: Dianna has 2 * 4 = 8 toys. They have 4 + 8 = 12 toys in total. The answer is 12

Performance on GSM
65%

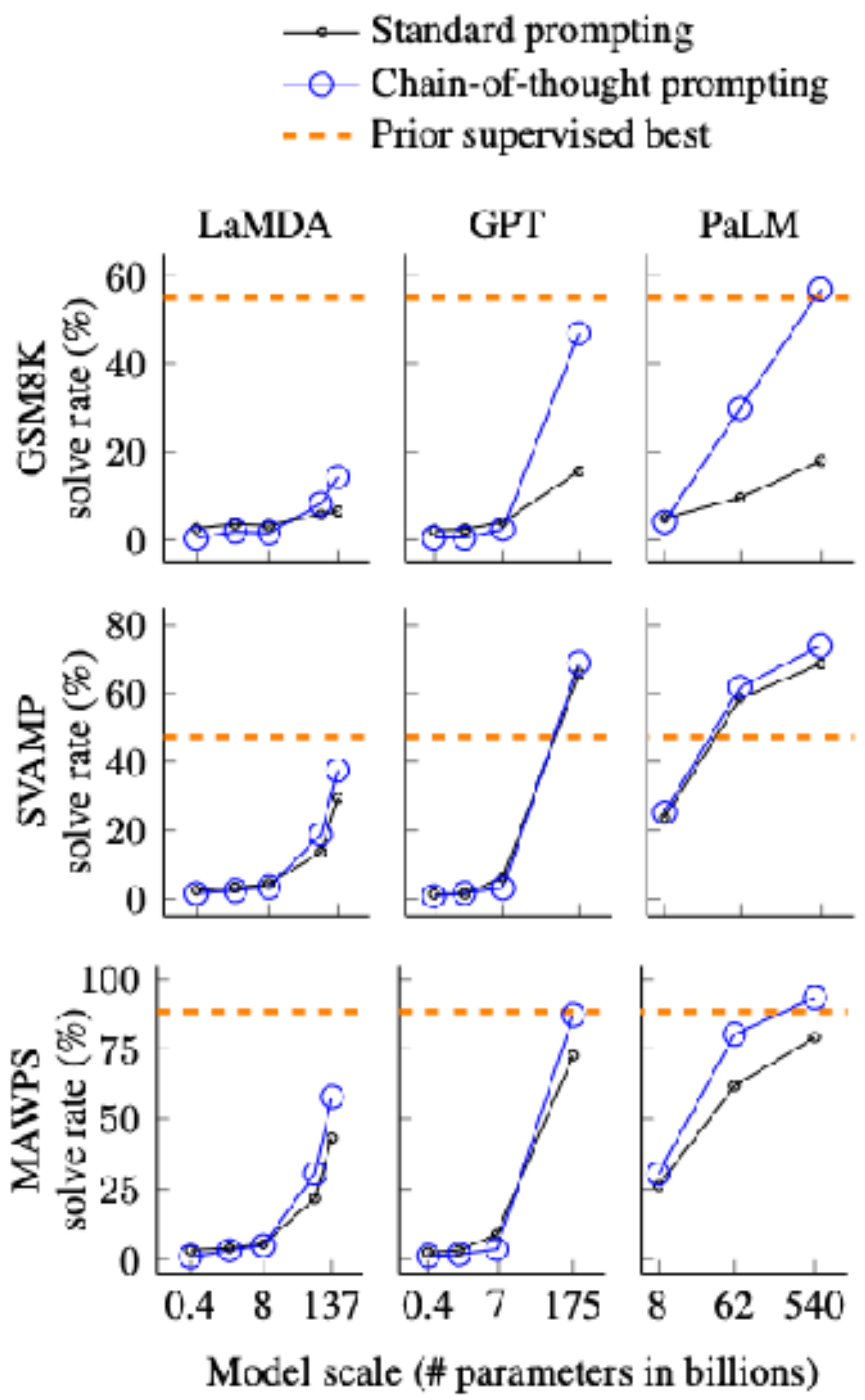
- Including **explanations** (ScratchPad; Chain-of-Thought) in prompts

(Nye at al., 2022)
(Wei et al., 2022)

Prompting with Explanations

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou
Google Research, Brain Team
{jasonwei,dennyzhou}@google.com



Challenging BIG-Bench tasks and whether chain-of-thought can solve them

Mirac Suzgun* Nathan Scales Nathanael Schärli Sebastian Gehrmann
Yi Tay Hyung Won Chung Aakanksha Chowdhery Quoc V. Le
Ed H. Chi Denny Zhou Jason Wei
Google Research *Stanford University

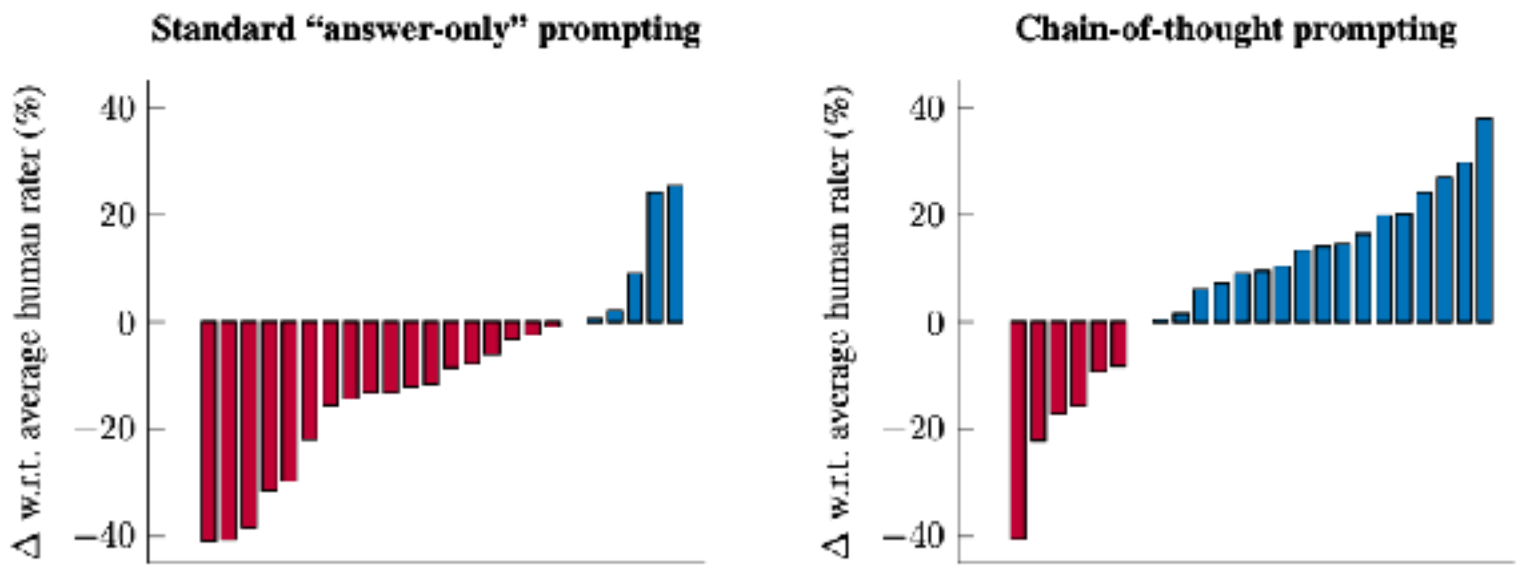
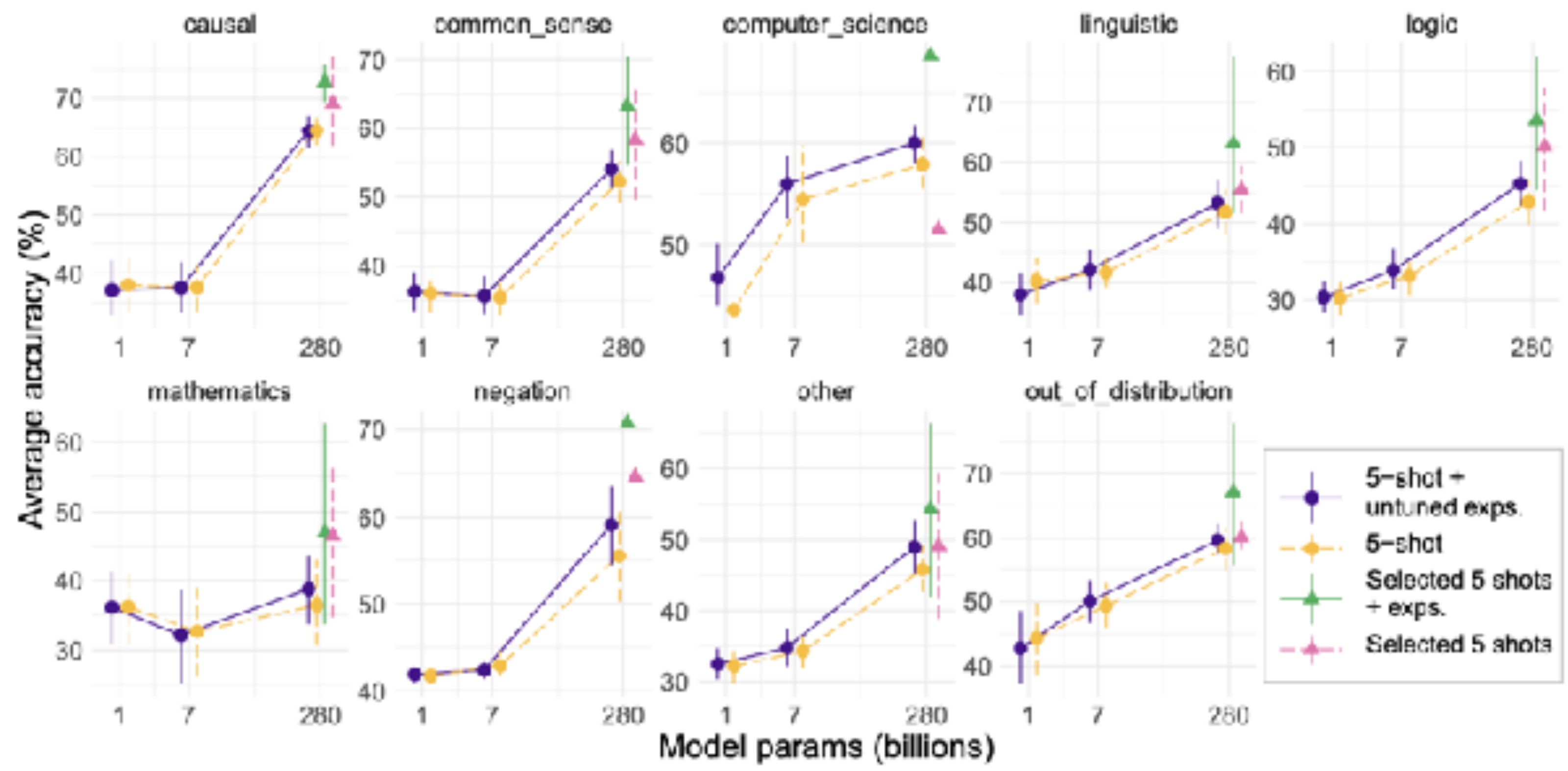


Figure 1: Per-task delta between Codex (code-davinci-002) and the average human-rater performance on 23 challenging tasks in BIG-Bench Hard, for standard *"answer-only"* (left) and *chain-of-thought* (right) prompting.

Can language models learn from explanations in context?

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan
Kory Mathewson, Michael Henry Tessler, Antonia Creswell
James L. McClelland, Jane X. Wang, Felix Hill
DeepMind
London, UK



Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Table 1: Accuracy comparison of Zero-shot-CoT with Zero-shot on each tasks. The values on the left side of each task are the results of using answer extraction prompts depending on answer format as described at § 3. The values on the right side are the result of additional experiment where standard answer prompt "The answer is" is used for answer extraction. See Appendix A.5 for detail setups.

	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**

Using Explanations for Textual Reasoning

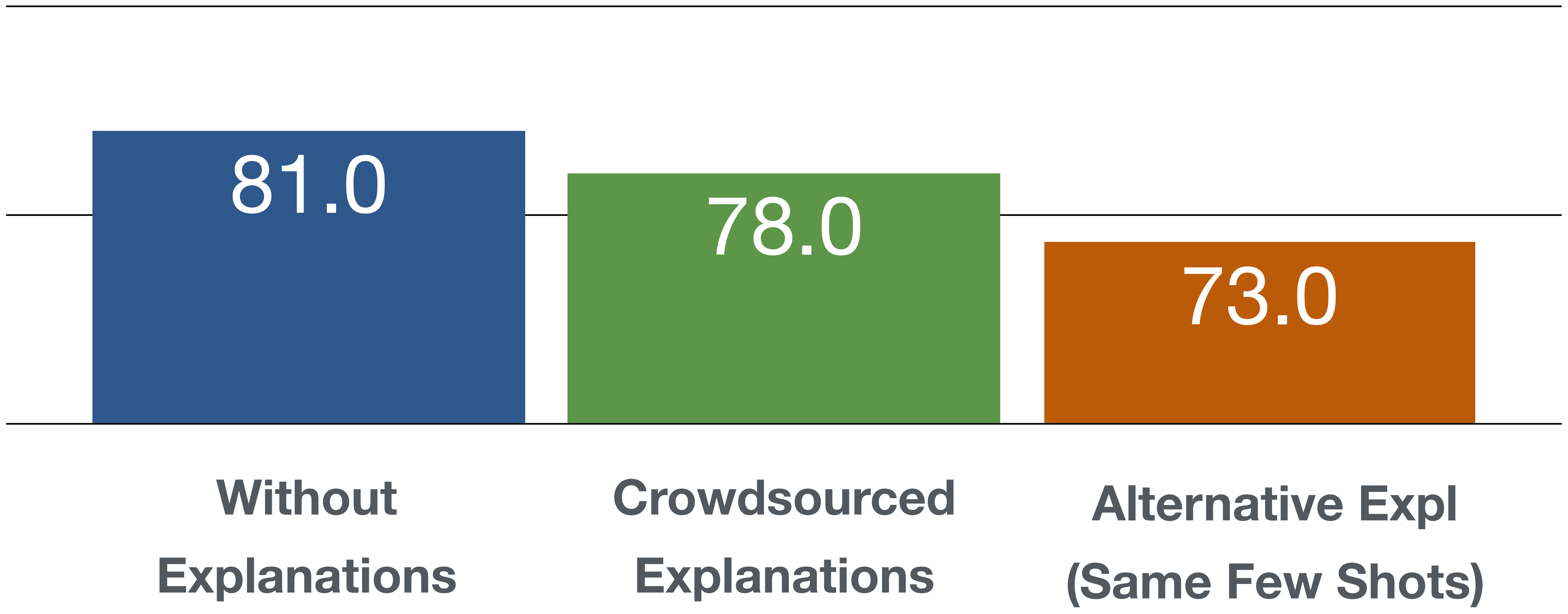
- We study prompting LLMs with explanations for **textual reasoning** tasks such as QA and NLI
- Explanations may not always improve prompting performance on textual reasoning tasks
- Performance is sensitive to different explanations

An E-SNLI Example

Premise: A female is looking through a microscope.
Hypothesis: A lady is observing something.
Explanation: You're looking through a microscope you are observing something.
Label: Entailment

Alternative Explanation: Looking through microscope implies observing

Prompting Performance

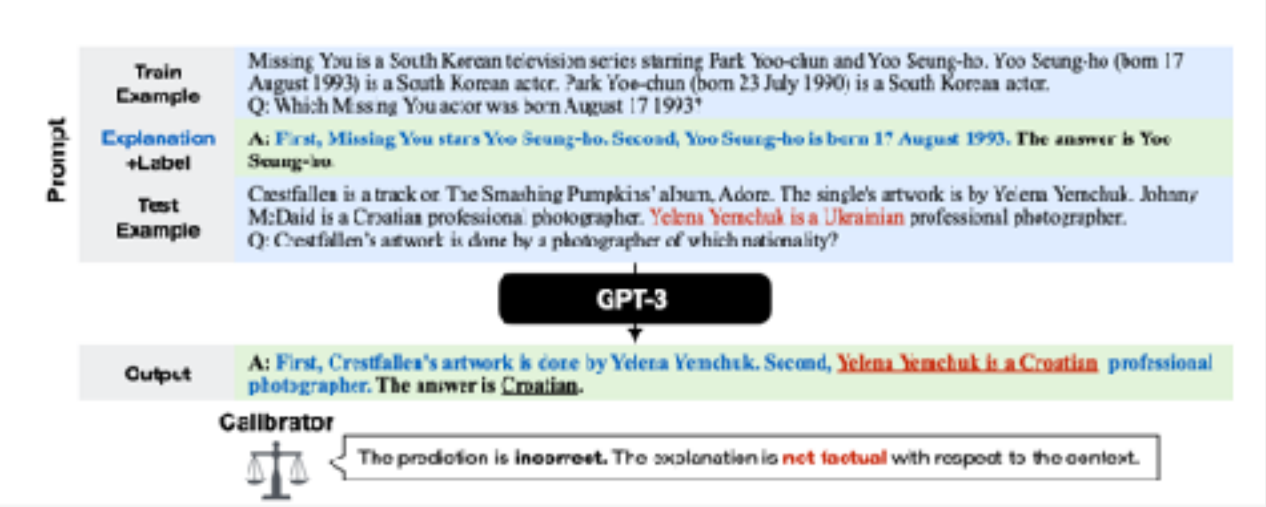


Outline

How well can LLMs learn from explanations in-context?
How to make explanations work better?

The Unreliability of Explanations in Few-Shot Prompting for Textual Reasoning

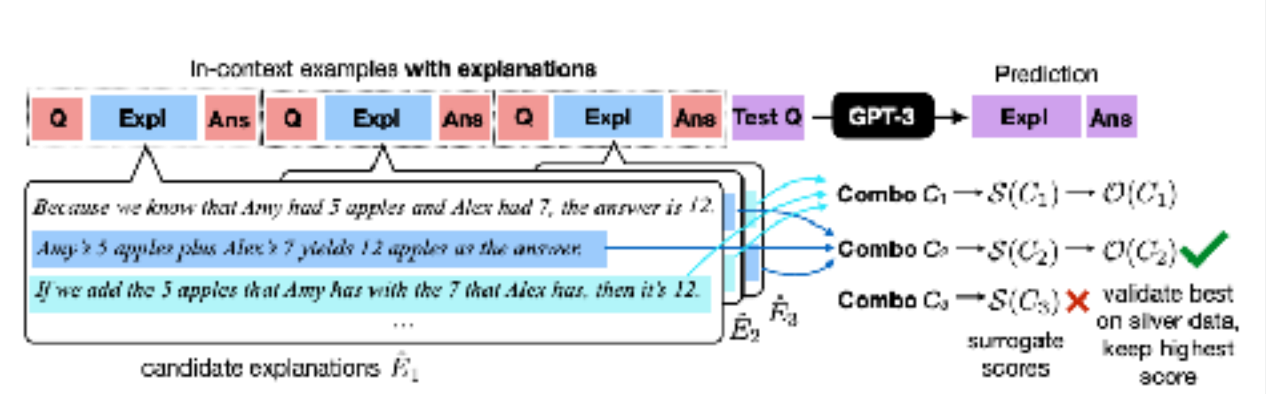
X Ye and G Durrett, NeurIPS 22



- Benchmark the effective of explanations in-context

Explanation Selection using Unlabeled Data for In-Context Learning

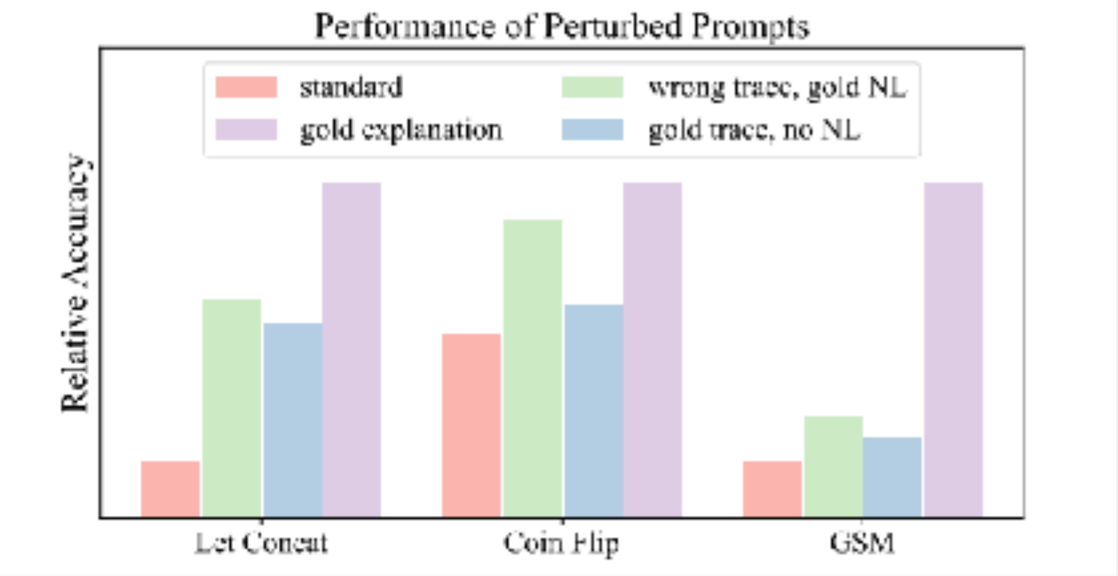
X Ye and G Durrett, ArXiv 23



- Optimize explanations to improve downstream performance

Complementary Explanations for Effective In-Context Learning

X Ye, S Iyer, A Celikyilmaz, V Stoyanov, G Durrett, and R Pasunuru, ACL Findings 23



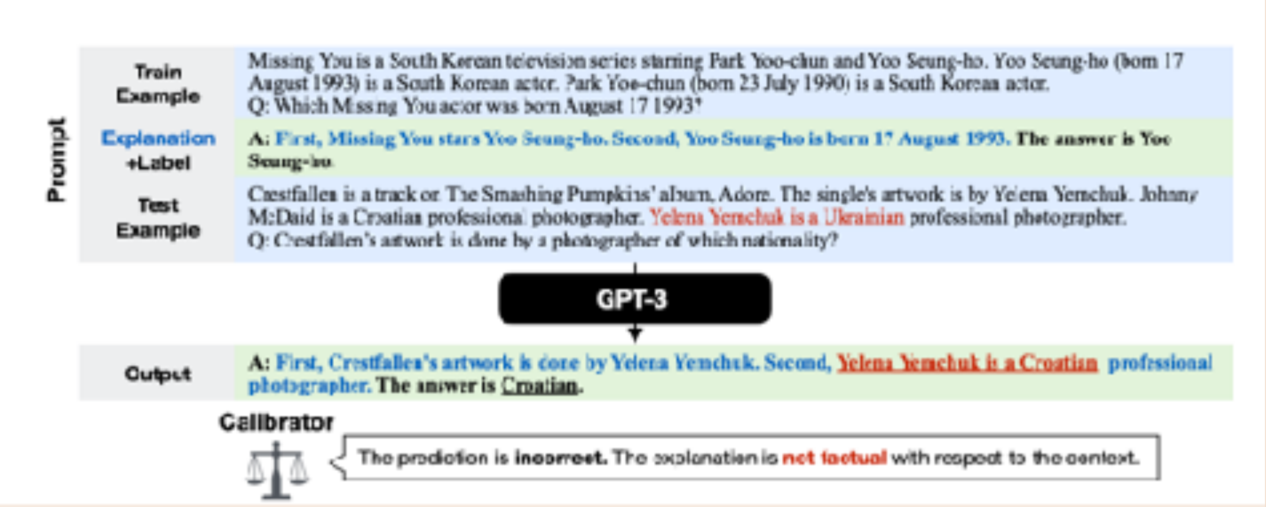
- Empirical analysis on how explanations work in in-context learning

Outline

How well can LLMs learn from explanations in-context?
How to make explanations work better?

The Unreliability of Explanations in Few-Shot Prompting for Textual Reasoning

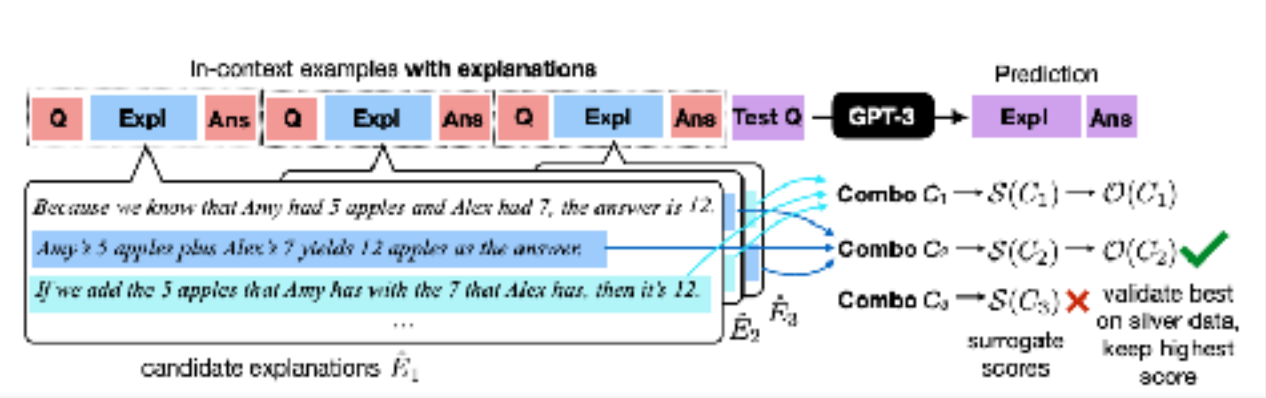
X Ye and G Durrett, NeurIPS 22



- Benchmark the effective of explanations in-context

Explanation Selection using Unlabeled Data for In-Context Learning

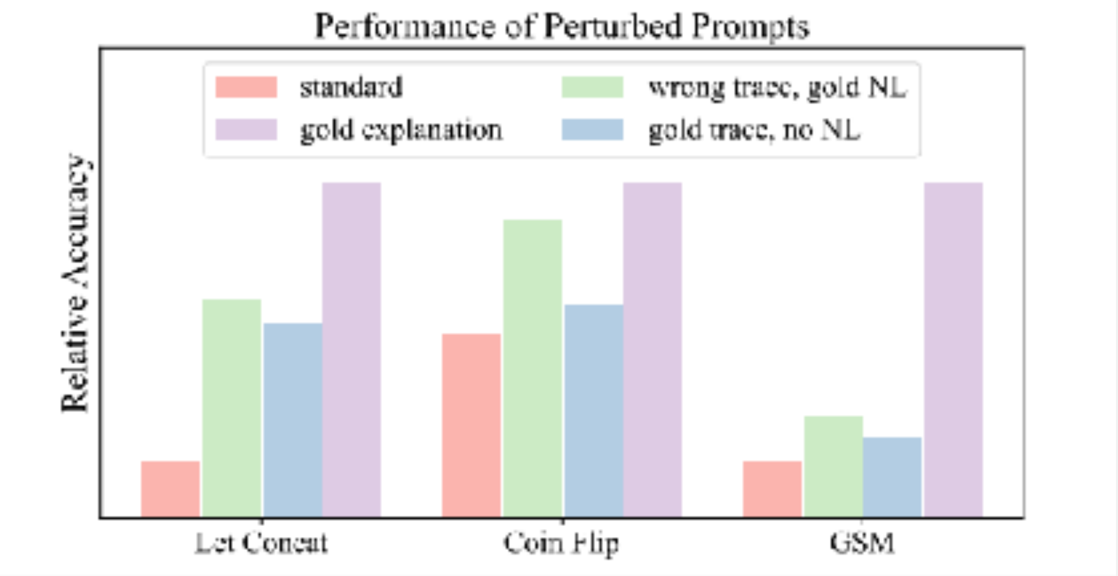
X Ye and G Durrett, ArXiv 23



- Optimize explanations to improve downstream performance

Complementary Explanations for Effective In-Context Learning

X Ye, S Iyer, A Celikyilmaz, V Stoyanov, G Durrett, and R Pasunuru, ACL Findings 23



- Empirical analysis on how explanations work in in-context learning

Using Explanations for Textual Reasoning

Crestfallen is a track on The Smashing Pumpkins' album, Adore. The single's artwork is by Yelena Yemchuk.
Johnny McDaid is a Croatian professional photographer.
Yelena Yemchuk is a Ukrainian professional photographer.
Q: Crestfallen's artwork is done by a photographer of which nationality?

GPT-3

A: First, Crestfallen's artwork is done by Yelena Yemchuk. Second, Yelena Yemchuk is a Croatian photographer. The answer is Croatian.

- Prompting LLMs with explanations for QA

Using Explanations for Textual Reasoning

Crestfallen is a track on The Smashing Pumpkins' album, Adore. The single's artwork is by Yelena Yemchuk.
Johnny McDaid is a Croatian professional photographer.
Yelena Yemchuk is a **Ukrainian** professional photographer.
Q: Crestfallen's artwork is done by a photographer of which nationality?

GPT-3

A: First, Crestfallen's artwork is done by Yelena Yemchuk. Second, Yelena Yemchuk is a **Croatian** photographer. The answer is **Croatian**.

! nonfactual

- Prompting LLMs with explanations for QA
- How well can LLMs learn from explanations in-context?
  - **Q1:** Does adding explanations to few-shot prompts improve performance?
  - **Q2:** Can LLMs generate reliable explanations?

Tasks

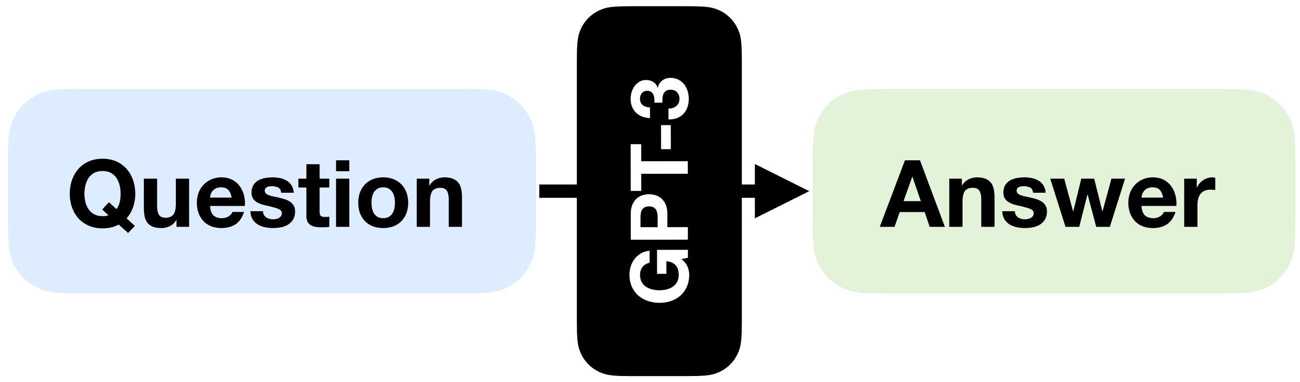
- **Synthetic:** a controlled synthetic QA dataset which allows full understanding of correct reasoning process

> **Context:** Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Daniel. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Daniel is a student. Thomas is a plumber.
> **Q:** Who hangs out with a student?
> **A:** Mary.
> **Explanation:** Mary hangs out with Daniel and Daniel is a student.

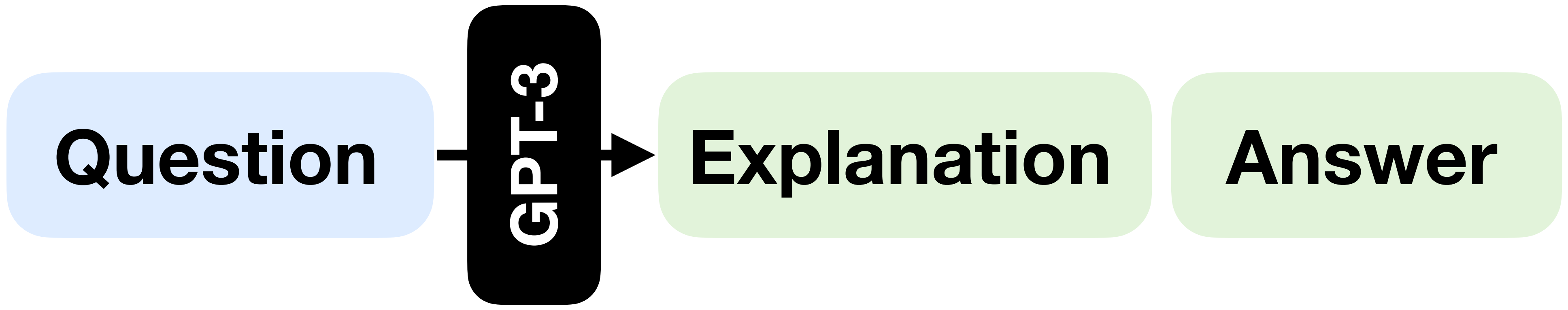
- **AdvHotpot:** a difficult version of adversarial Hotpot QA datasets
- **E-SNLI:** NLI with free-text explanations

Prompting Methods

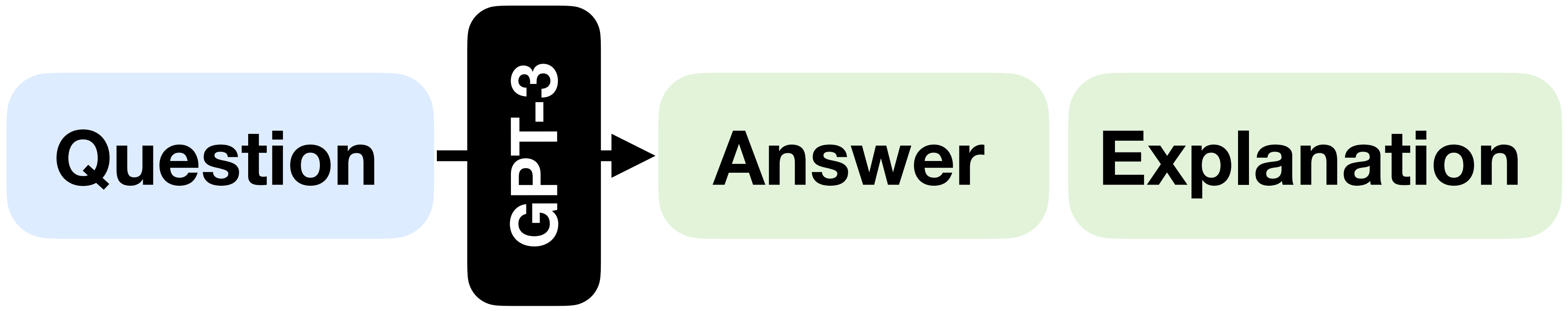
- **Standard:** directly answer



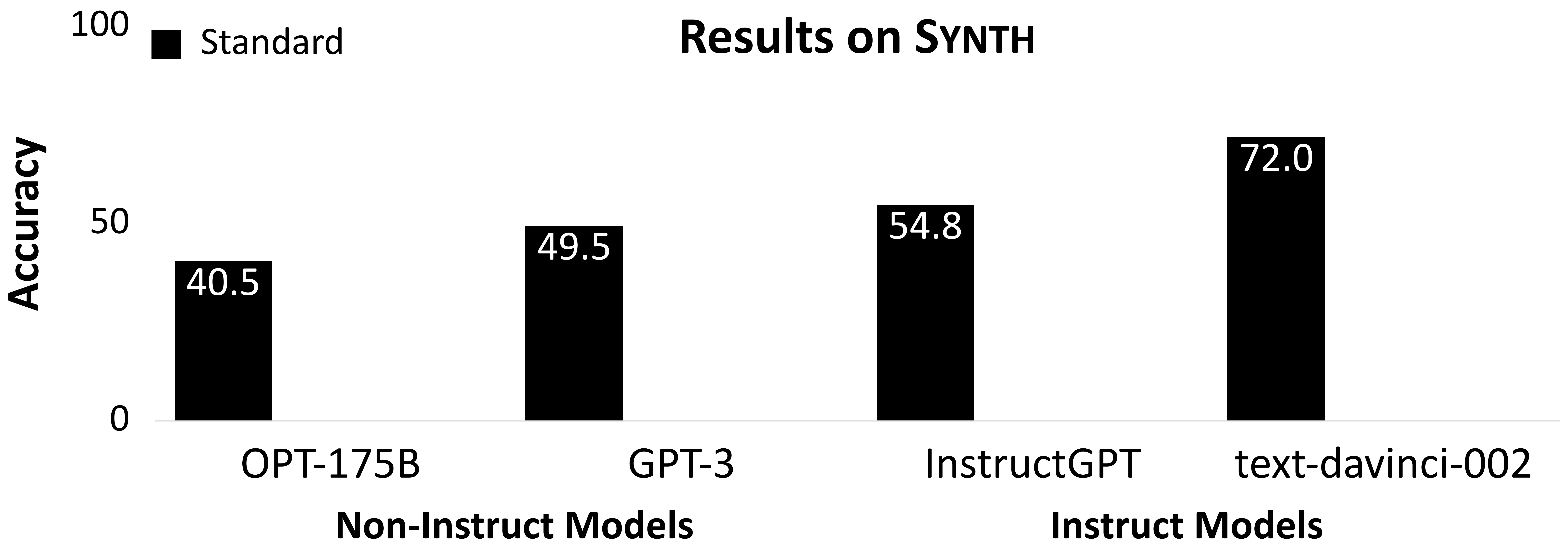
- **Explain-predict**: Scratchpad (Nye et al., 2021); Chain-of-thought (Wei et al., 2022);



- **Predict-explain**: first makes a prediction and then generates an explanation

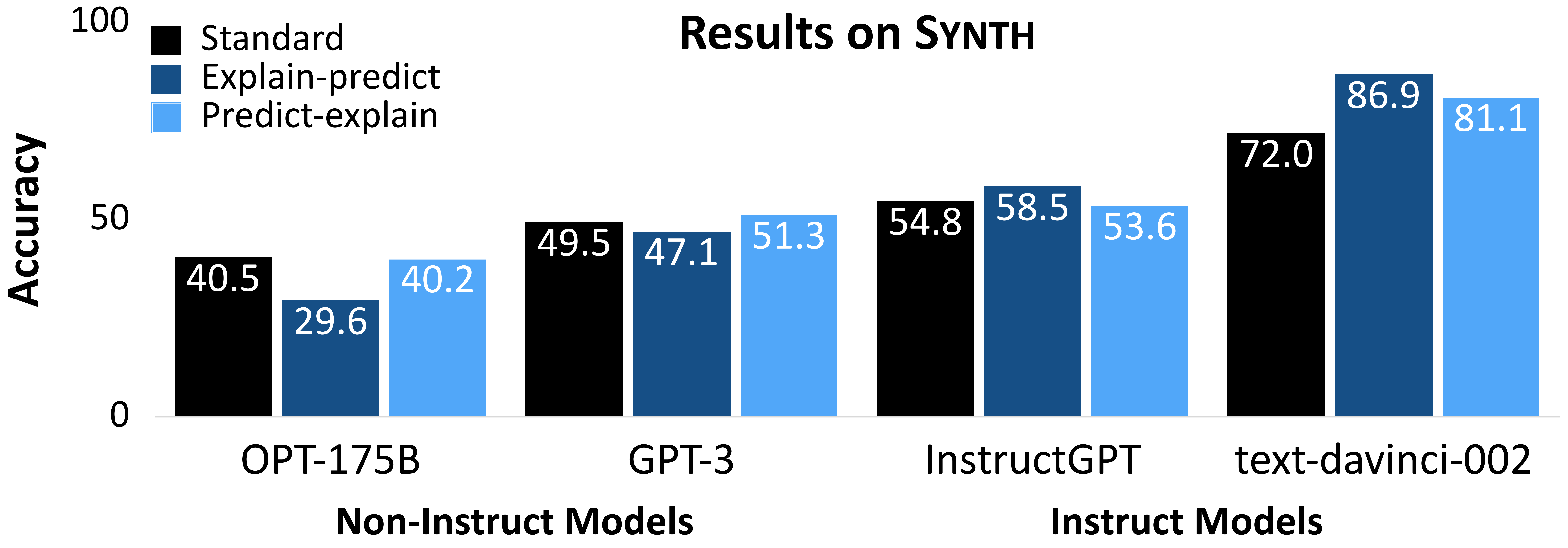


Results: Performance



- LLMs: OPT-175B, GPT-3 (davinci), InstructGPT(text-daivinci-001), and text-davinci-002
- Do explanations help?

Results: Performance

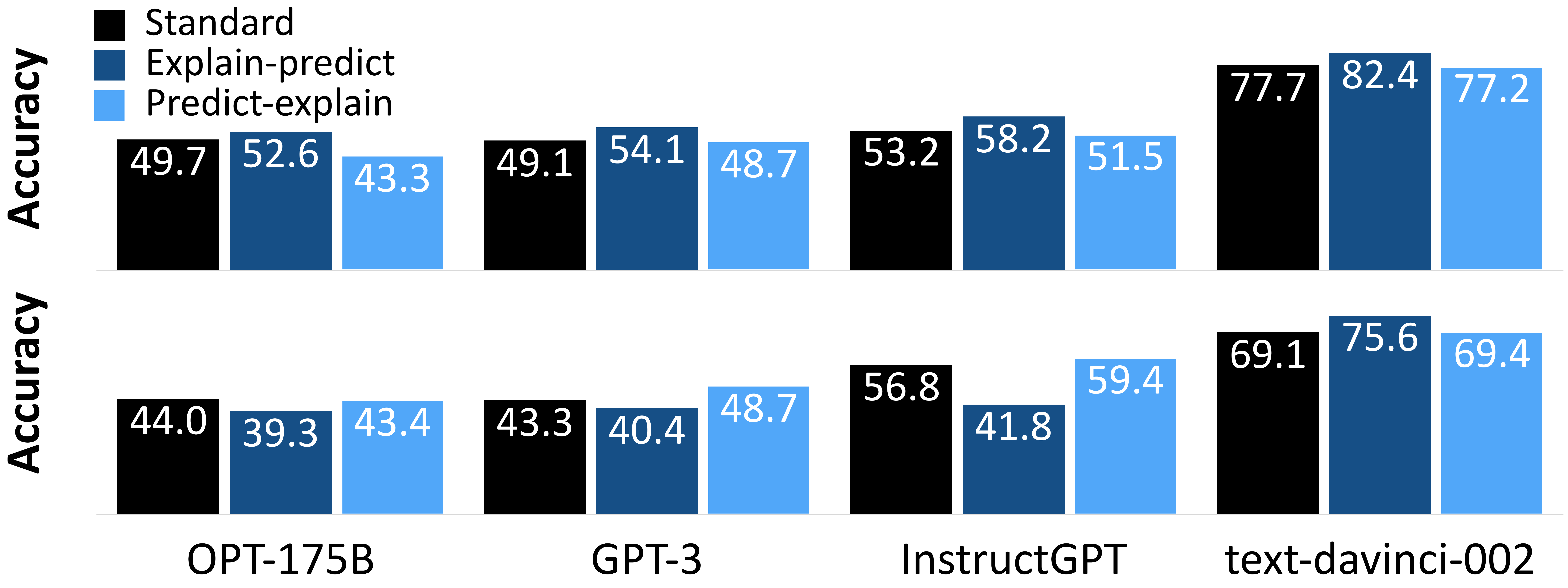


- LLMs: OPT-175B, GPT-3 (davinci), InstructGPT (text-daivinci-001), and text-davinci-002
- Do explanations help?
  - For the Synth dataset: minor gains on OPT, GPT-3, InstructGPT. More substantial improvements on text-davinci-002.

Results: Performance (Cont'd)

Results on HOTPOTADV

Results on ESNLI



- LLMs: OPT-175B, GPT-3 (davinci), InstructGPT (text-daivinci-001), and text-davinci-002
- Do explanations help?
  - For the Synth dataset: minor gains on OPT, GPT-3, InstructGPT. More substantial improvements on text-davinci-002
  - Consistent trends on HotpotAdv and E-SNLI

Unreliability of Explanations

- Can language models generate reliable explanations?
  - **Factuality:** whether an explanation is factually grounded in the input context

nonfactual explanation on SYNTH (E-P)

Stephanie is an engineer. John is a nurse. Jonathan hangs out with Stephanie. Tiffany is a nurse. Emily agrees with John. David agrees with Jason. Angela hangs out with Tiffany.

Q: Who hangs out with a nurse?

GPT-3

A: Jonathan hangs out with Stephanie and Stephanie is a nurse. The answer is a Jonathan.

Unreliability of Explanations (Cont'd)

- Can language models generate reliable explanations?
  - **Factuality:** whether an explanation is factually grounded in the input context

nonfactual explanation on AdvHotpot (E-P)

Crestfallen is a track on The Smashing Pumpkins' album, Adore. The single's artwork is by Yelena Yemchuk.
Johnny McDaid is a Croatian professional photographer.
Yelena Yemchuk is a **Ukrainian** professional photographer.
Q: Crestfallen's artwork is done by a photographer of which nationality?

GPT-3

A: First, Crestfallen's artwork is done by Yelena Yemchuk. Second, Yelena Yemchuk is a **Croatian** photographer. The answer is **Croatian**.

Unreliability of Explanations (Cont'd)

- Can language models generate reliable explanations?
  - **Factuality:** whether an explanation is factually grounded in the input context
  - **Consistency:** whether an explanation entails the answer

Inconsistent explanation on SYNTH (E-P)

Matthew blames Tiffany. Lisa is a chef. Christopher helps Kelly. Angela helps Jessica. Rachel blames Lisa. Jessica is a farmer. Kelly is a chef. Tiffany is a farmer

Q: Who helps a farmer?

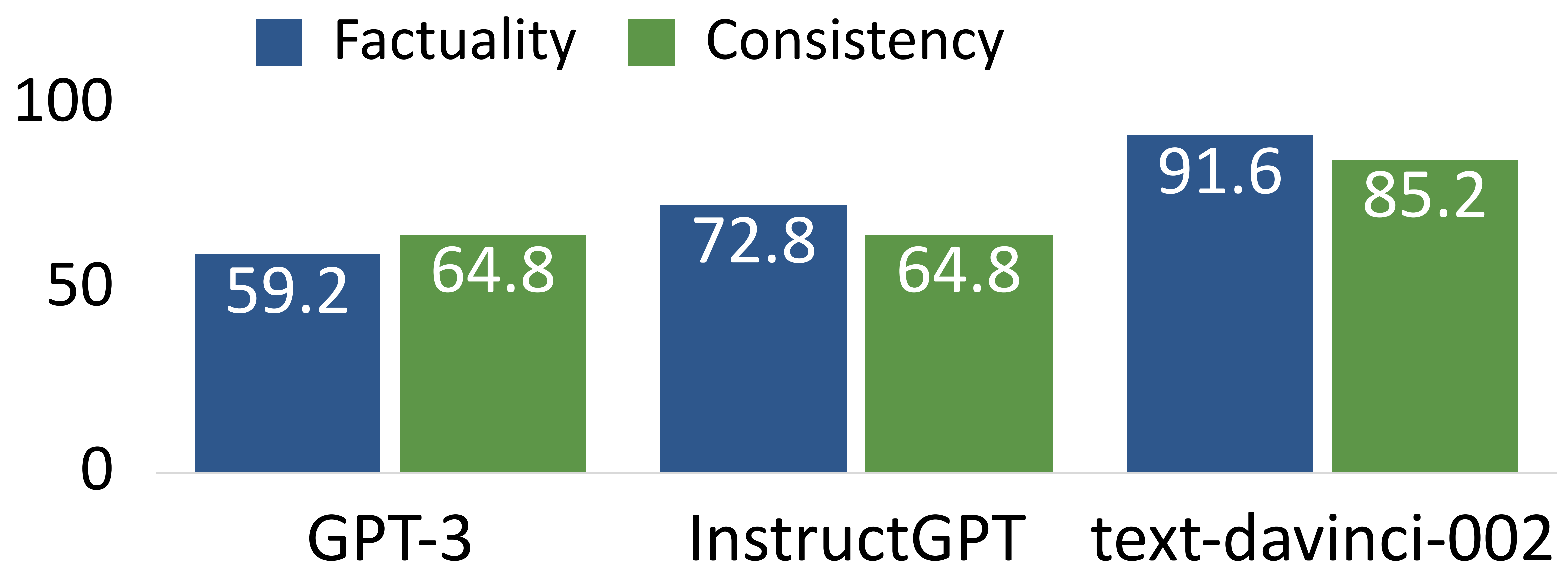
GPT-3

A: Jessica is a farmer and Christopher helps Kelly. The answer is Christopher.

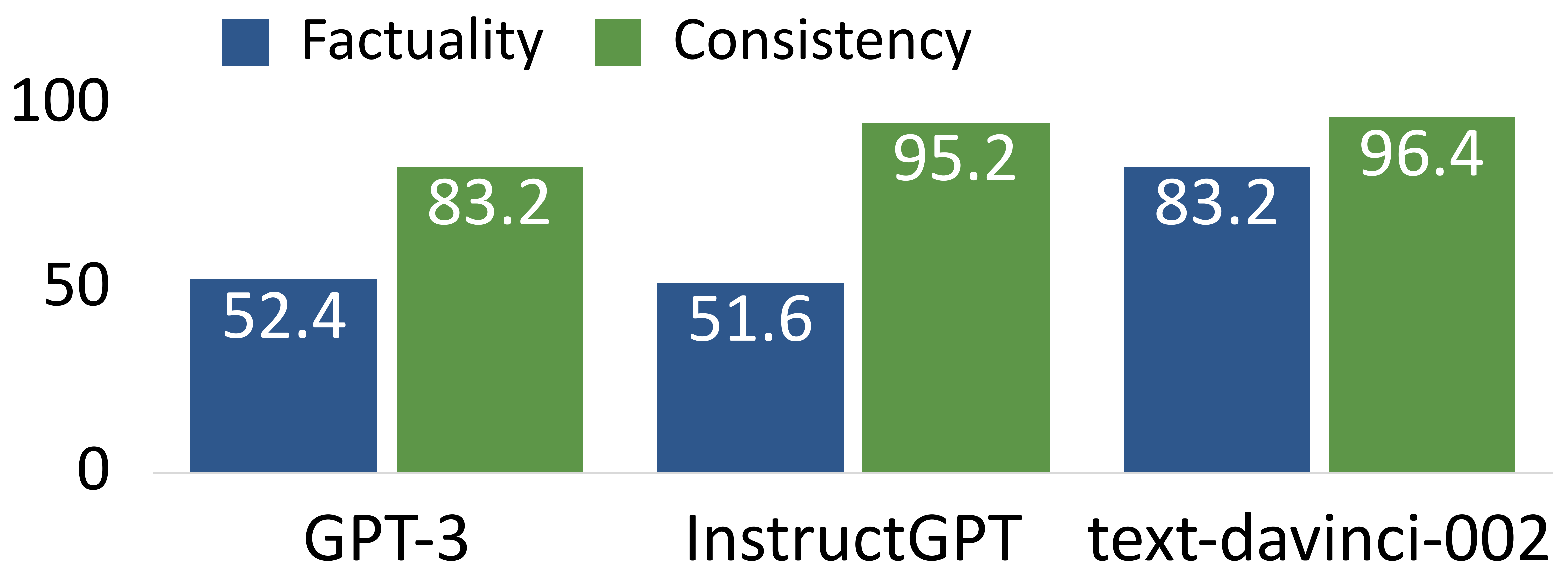
Results: Reliability

- Can language models generate reliable explanations?
  - **Factuality:** whether an explanation is factually grounded in the input context
  - **Consistency:** whether an explanation entails the answer
- Model-generated explanations can be **unreliable** !

Explain-Predict on SYNTH



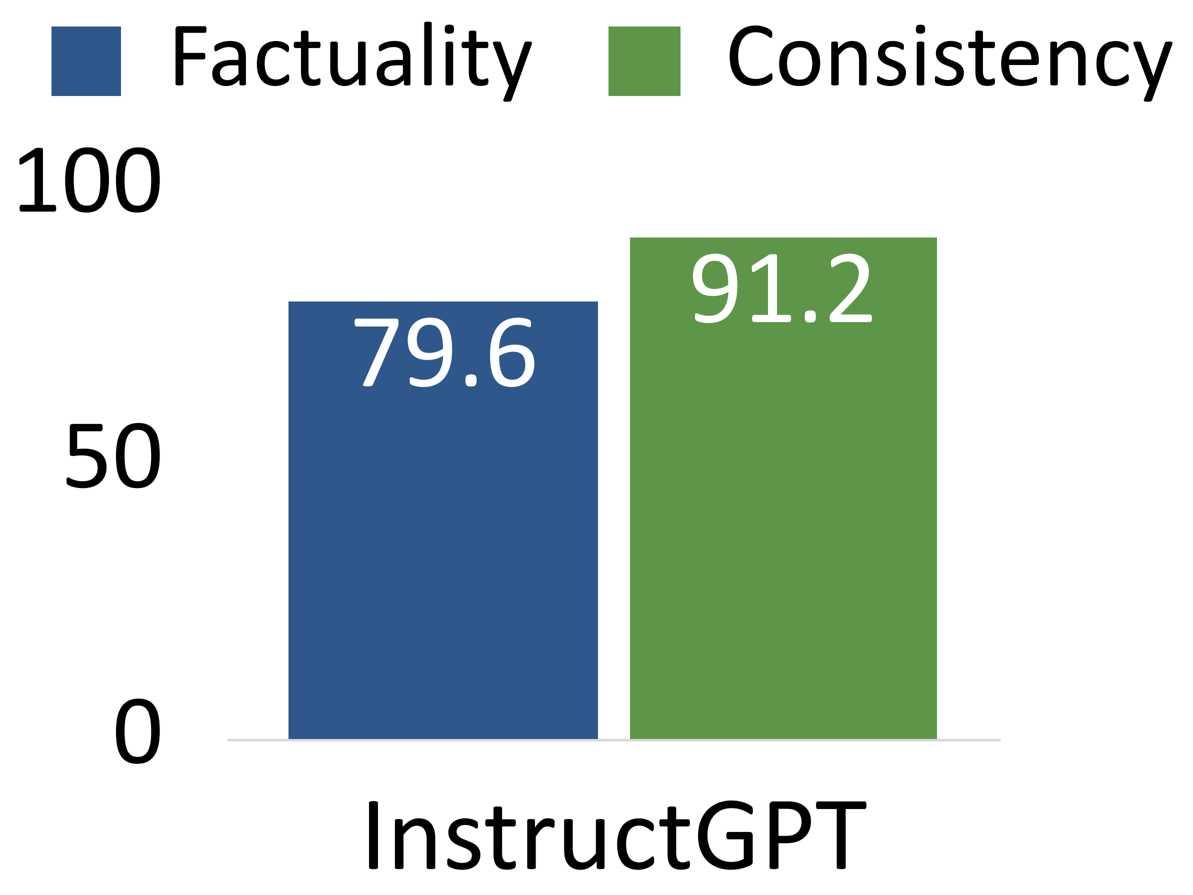
Predict-Explain on SYNTH



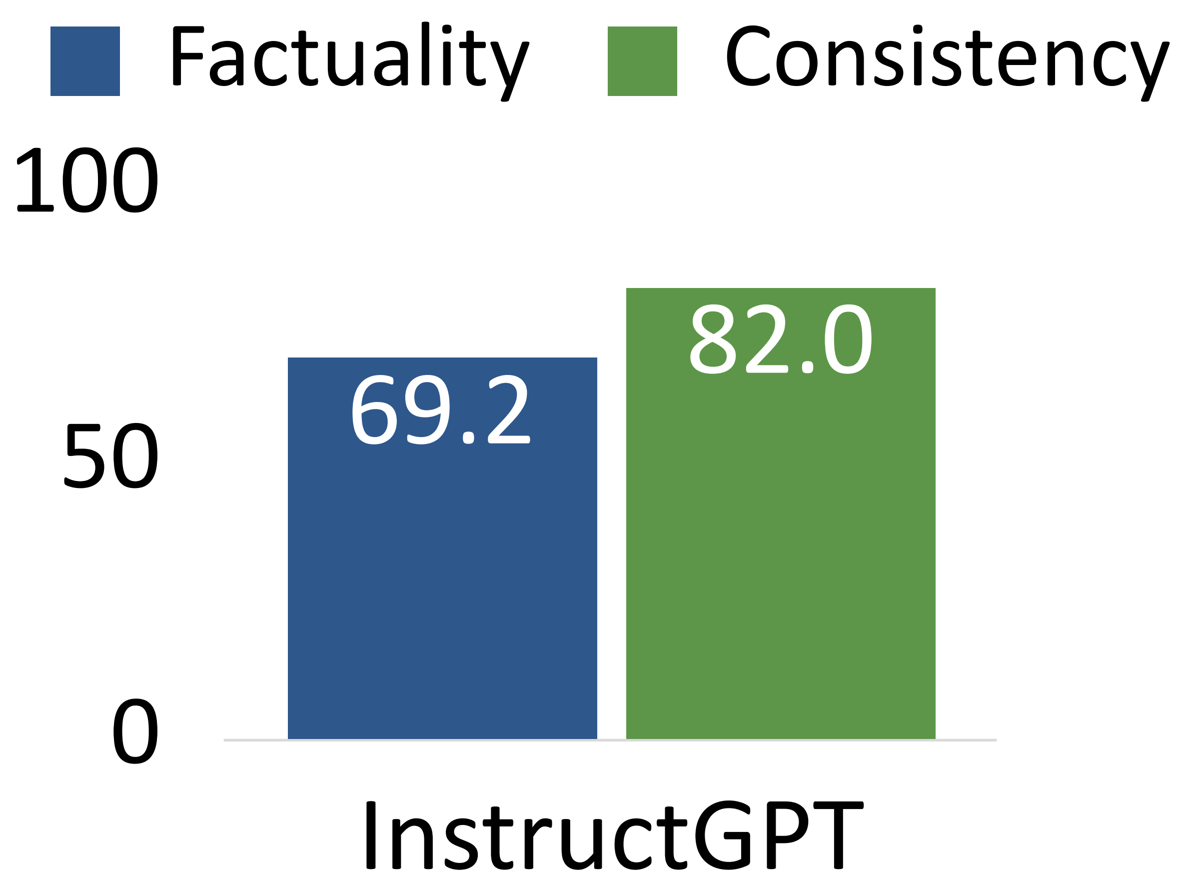
Results: Reliability (Cont'd)

- Can language models generate reliable explanations?
  - **Factuality:** whether an explanation is factually grounded in the input context
  - **Consistency:** whether an explanation entails the answer
- Model-generated explanations can be **unreliable** !

Explain-Predict on ADVHOTPOT



Predict-Explain on ADVHOTPOT



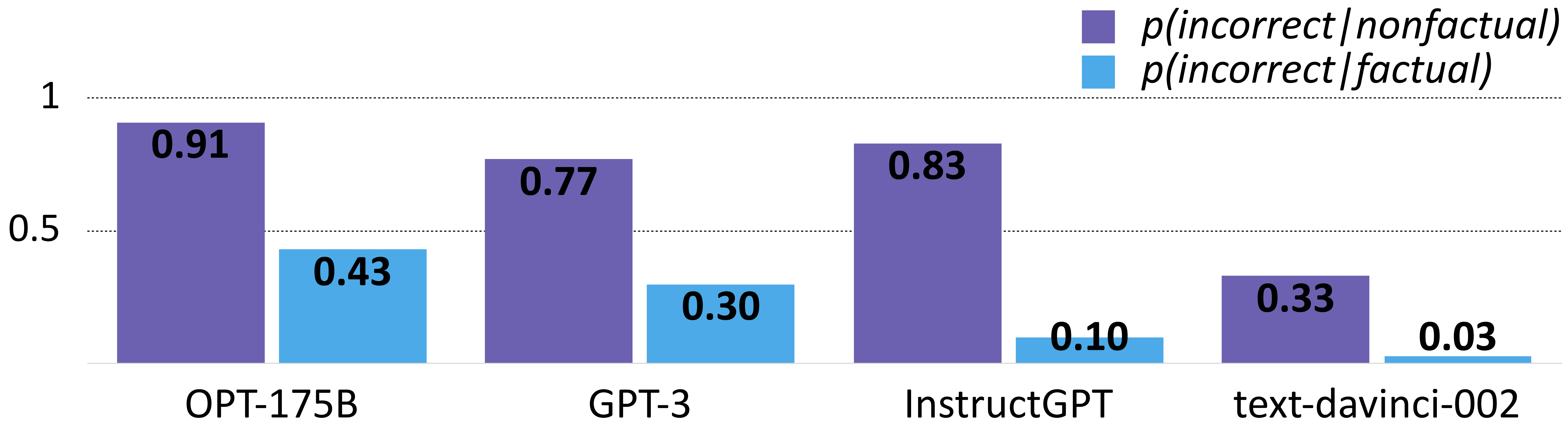
Connecting Factuality and Accuracy

Stephanie is an engineer. John is a nurse. Jonathan hangs out with Stephanie. Tiffany is a nurse. Emily agrees with John. David agrees with Jason. Angela hangs out with Tiffany.

Q: Who hangs out with a nurse?

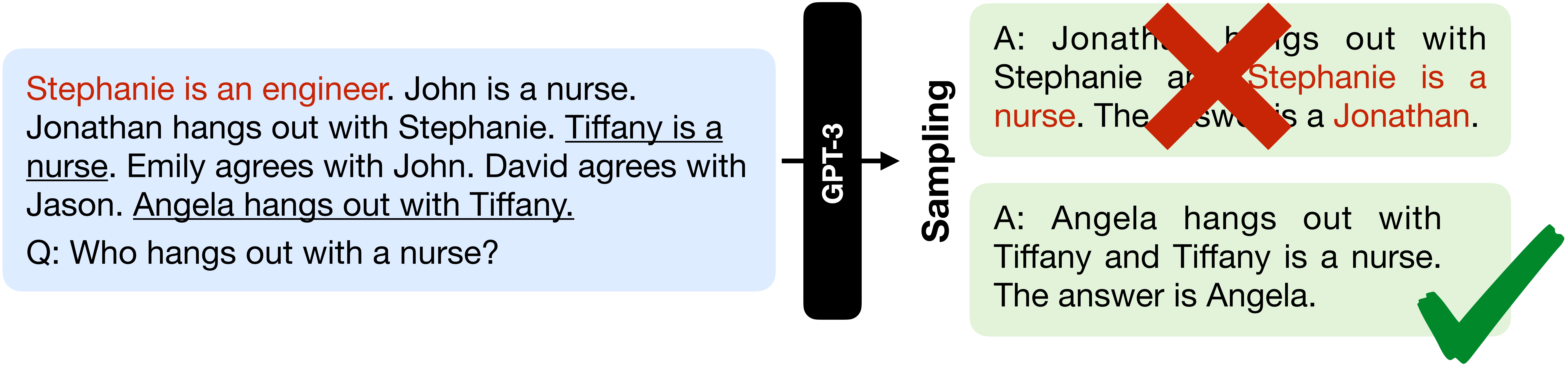
GPT-3

A: Jonathan hangs out with Stephanie and Stephanie is a nurse. The answer is a Jonathan.



- Incorrect predictions are more likely to co-occur with nonfactual explanations

Connecting Factuality and Accuracy



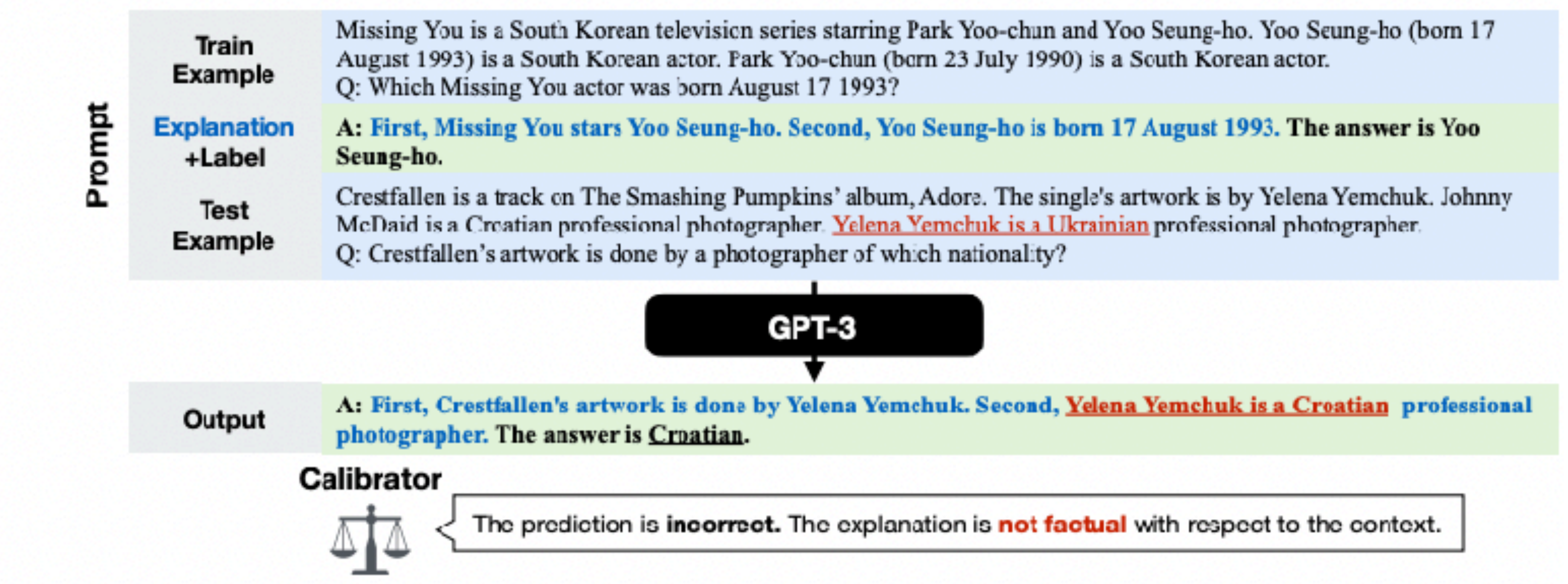
- Incorrect predictions are more likely to co-occur with nonfactual explanations
- Nonfactual explanations can be useful as a way to verify LLMs' predictions
  - On SYNTH, we sample multiple explanation-answer pairs , and reject nonfactual ones
  - Successfully improves the accuracy from 54% to 74% (P-E)

Wrap-up

- **LLMs are not good enough at using explanations for textual reasoning**
  - Simply including explanations in prompt may not always lead to substantial benefits
  - Model-generated explanations can be unreliable
- **But flawed explanations can be useful for verifying LLMs' predictions**

The Unreliability of Explanations in Few-Shot Prompting for Textual Reasoning

Xi Ye and Greg Durrett, NeurIPS 2022

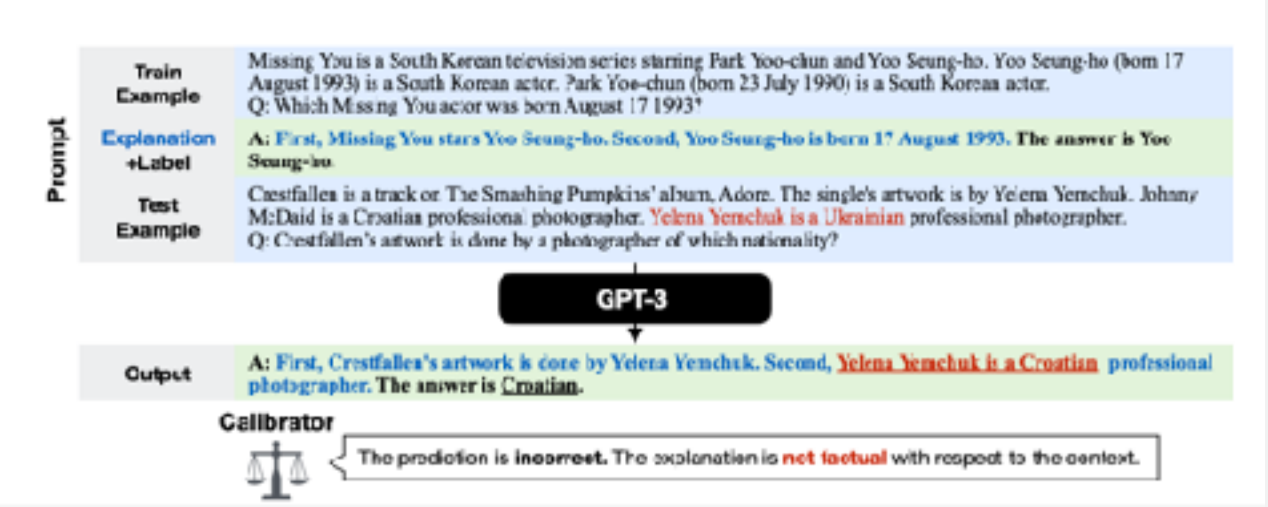


Outline

How well can LLMs learn from explanations in-context?
How to make explanations work better?

The Unreliability of Explanations in Few-Shot Prompting for Textual Reasoning

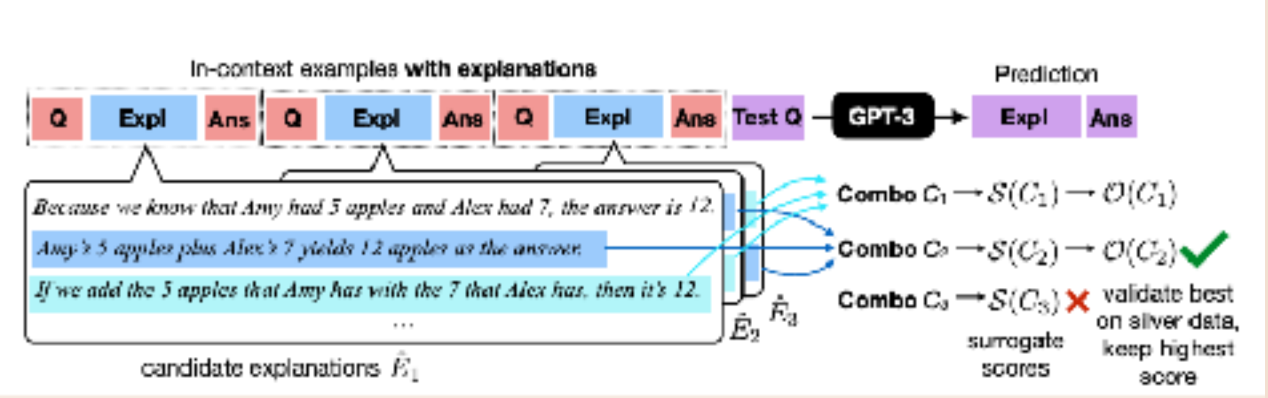
X Ye and G Durrett, NeurIPS 22



- Benchmark the effective of explanations in-context

Explanation Selection using Unlabeled Data for In-Context Learning

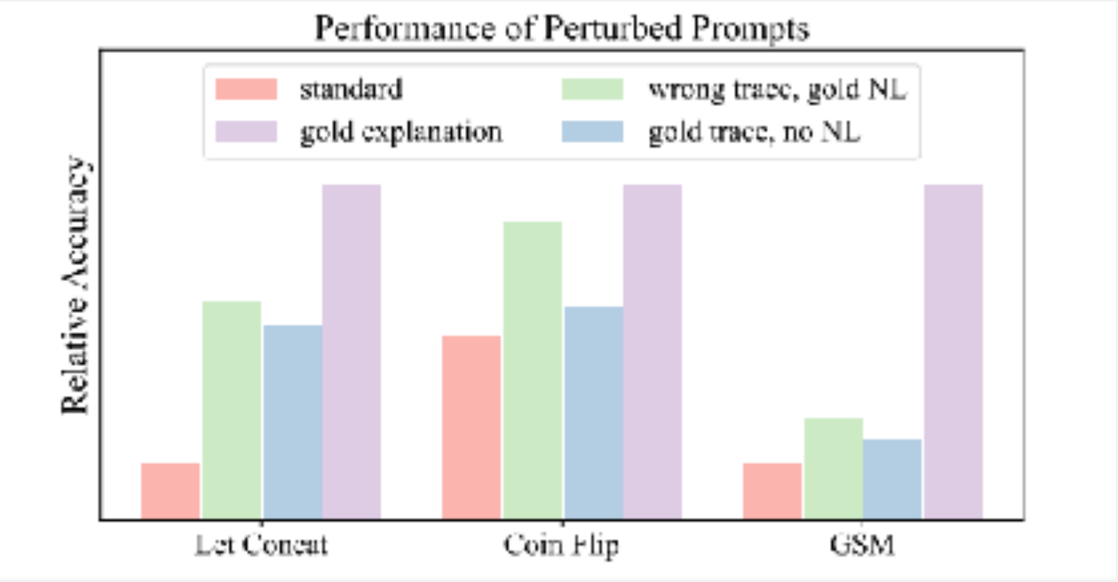
X Ye and G Durrett, ArXiv 23



- Optimize explanations to improve downstream performance

Complementary Explanations for Effective In-Context Learning

X Ye, S Iyer, A Celikyilmaz, V Stoyanov, G Durrett, and R Pasunuru, ACL Findings 23



- Empirical analysis on how explanations work in in-context learning

Performance Varying Across Explanations

Q: Alice has 5 apples. Bob has 2 apples. How many apples do they have together?
A: They have 5 + 2 = 7 apples together. The answer is 7.

Q: ...

GPT-3

Performance

52%

Q: Alice has 5 apples. Bob has 2 apples. How many apples do they have together?
A: Because Alice has 5 apples and Bob has 2 apples. We know 5 + 2 = 7. The answer is 7.

Q: ...

GPT-3

Performance

57%

- Performance varies across explanations
- How to find the explanations that yields better downstream performance?

Optimizing Explanations

Few-Shot Exemplars Q_1 A_1 ; Q_2 A_2 ; … ; Q_K A_K

- Search for E_1 E_2 … E_K that yields better end task performance (on unseen test set)

$(Q_1\ E_1\ A_1 ;\ Q_2\ E_2\ A_2 ;\ \dots ;\ Q_K\ E_K\ A_K) ;\ Q \xrightarrow{\text{GPT-3}}$ **Best Performance**

Data Condition

Given

Few-Shot Exemplars	Q_1 A_1 ; Q_2 A_2 ; … ; Q_K A_K
Seed Explanations	$\tilde{E}_1$ $\tilde{E}_2$ … $\tilde{E}_K$
Unlabeled Dev set	$V =$ Q_1 Q_2 … Q_M

Output

Optimized Explanations	E_1 E_2 … E_K that yields better end task performance

Approach Overview

- **Generate candidate explanations:** use seed explanations to perform leave-one-out prompt

$$(\; Q_2 \; \tilde{E}_2 \; A_2 \; ; \ldots ; \; Q_K \; \tilde{E}_K \; A_K \;) \; ; \; Q_1 \xrightarrow{\text{GPT-3}} \begin{array}{lll} \hat{E}_1^{(1)} \; \hat{A}_1^{(1)} & \hat{A}_1^{(1)} = A_1 & \checkmark \\ \hat{E}_1^{(2)} \; \hat{A}_1^{(2)} & \hat{A}_1^{(2)} \neq A_1 & \times \\ \ldots & & \end{array}$$

View Q_1 as test query
use the others to do CoT prompting

Only keep explanations
paired correct answers

Q: Alice has 5 apples.…How many apples do they have?
A: They have …. The answer is 7.
…
Q: …
A: …
Q: Charlie has 4 toys. Dianna has twice as much as Charlie. How many toys do they have together.

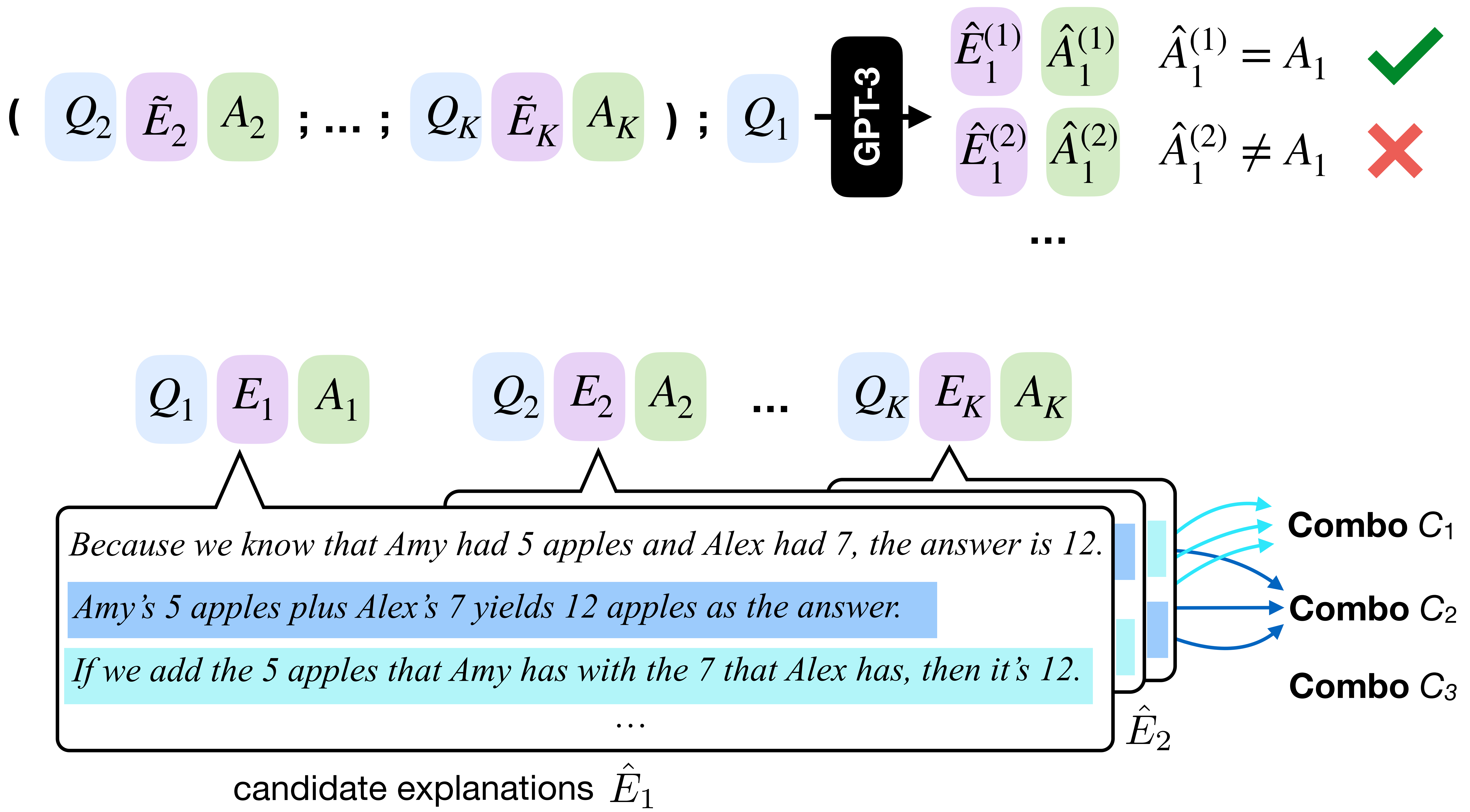
GPT-3 →

A: Dianna has 2 * 4 = 8 toys. They have 4 + 8 = 12 toys in total. The answer is **12**. ✓

A: Diana has twice toys. So they have 4 * 2 = 8 toys. The answer is **8**. ✗

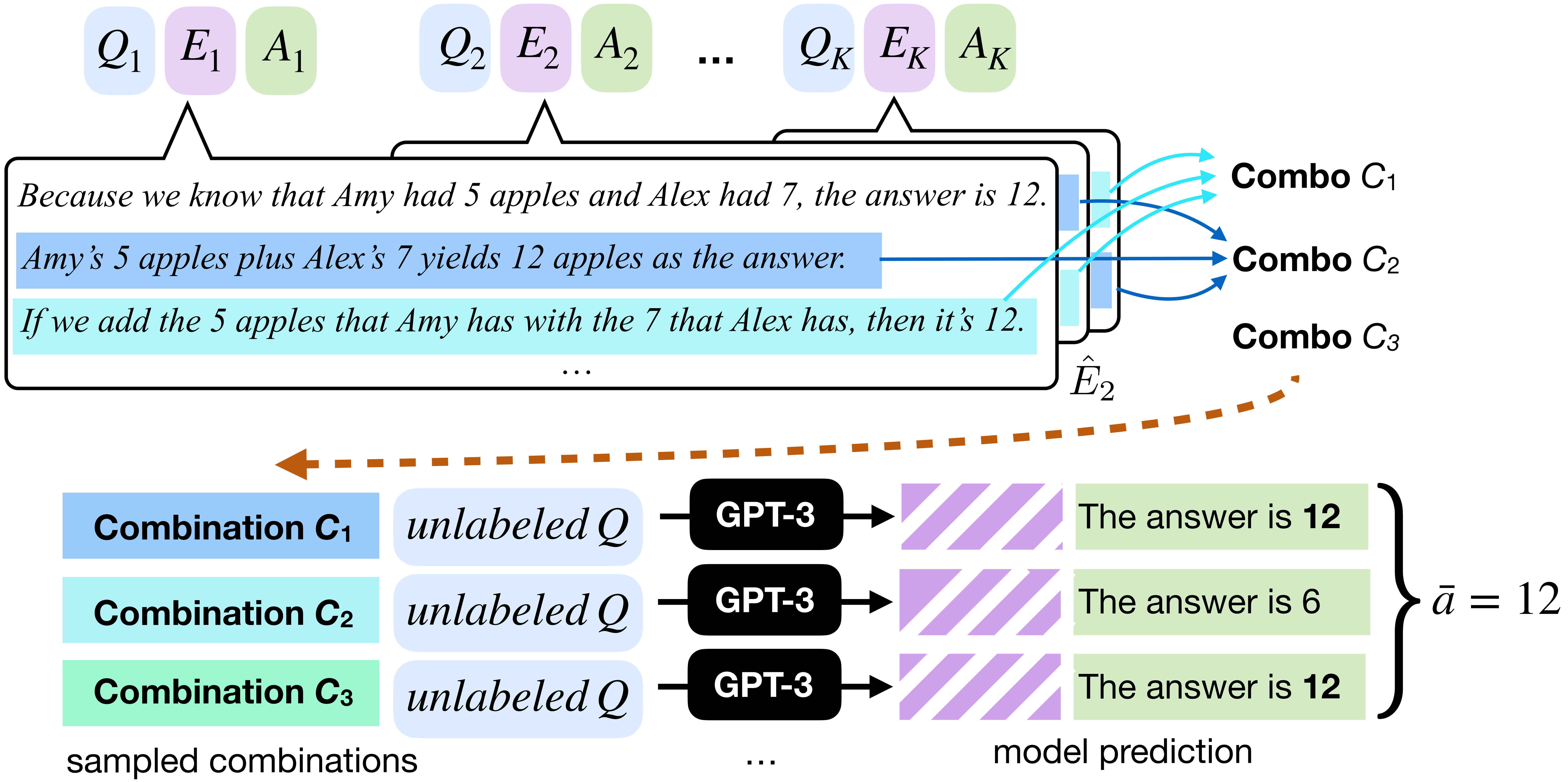
Approach Overview

- **Generate candidate explanations:** use seed explanations to perform leave-one-out prompt
  - This yields **combinations** of explanations



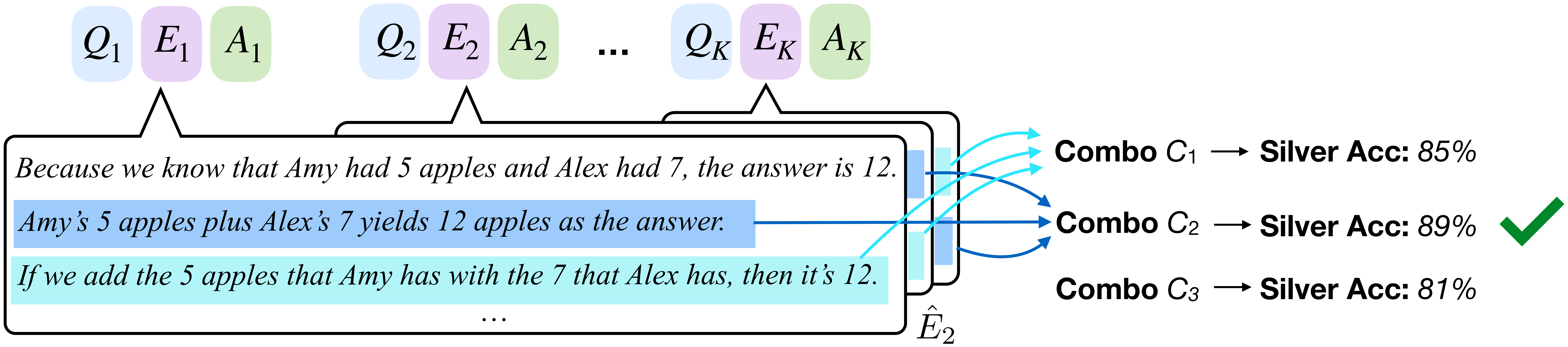
Approach Overview (Cont'd)

- **Generate candidate explanations:** use seed explanations to perform leave-one-out prompt
  - This yields **combinations** of explanations
- **Silver-label development set:** sample combinations and silver-label V by prompting and voting



Approach Overview (Cont'd)

- **Generate candidate explanations:** use seed explanations to perform leave-one-out prompt
  - This yields **combinations** of explanations
- **Silver-label development set:** sample combinations and silver-label V by prompting and voting
- **Select combination based on silver-accuracy:** score combinations using silver-accuracy
  - Essentially, we search for combinations that gives best silver accuracy



Performance Varying across Explanations

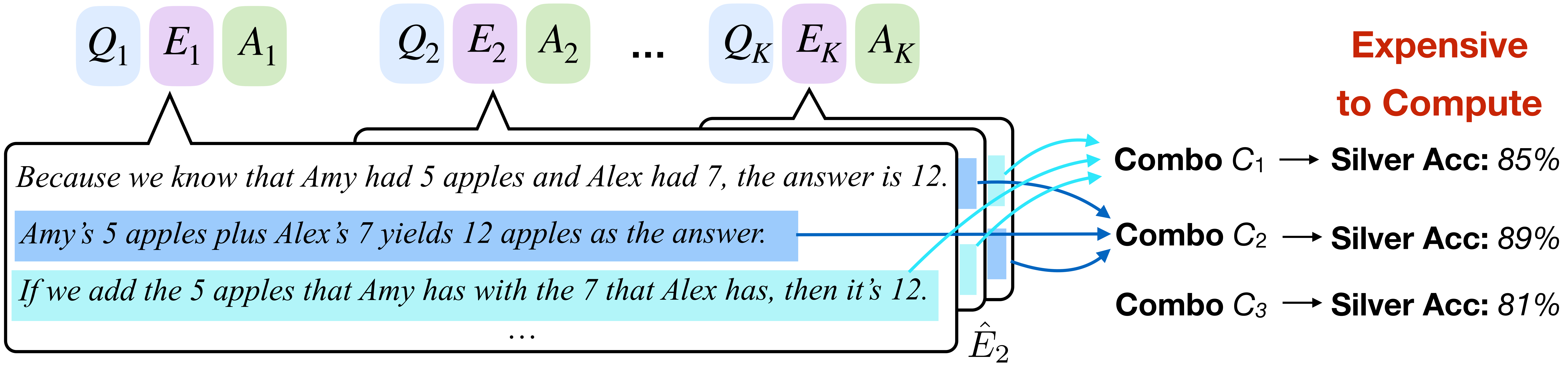
- We investigate the variance of performance obtained with different combinations
  - Performance varies **widely** across explanations on four tasks
  - Seed explanations (annotated by crowdworkers) yields suboptimal performance

Stats of performance across sampled combinations

	MIN	AVG	MAX	SEED
GSM	57.7	61.8	66.0	61.9
ECQA	72.7	76.1	78.6	74.9
E-SNLI	60.3	72.3	80.1	71.8
STRATEGYQA	69.8	73.8	76.5	74.0

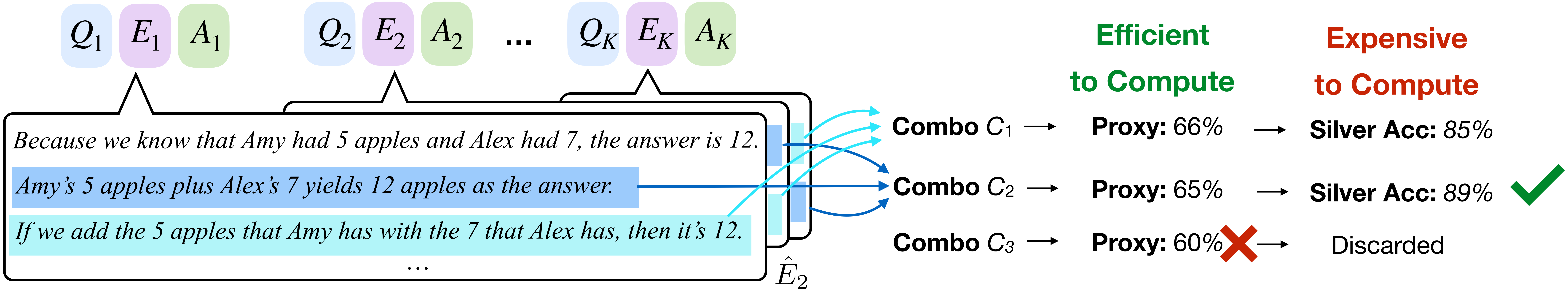
Prioritizing Search

- We can only evaluate the silver-accuracy of a few combinations owning to the high cost of running LLMs



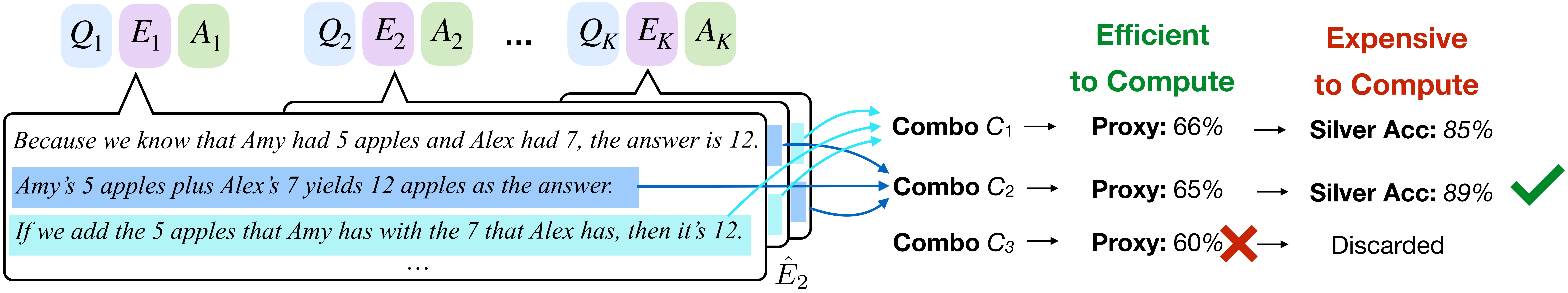
Prioritizing Search

- We can only evaluate the silver-accuracy of a few combinations owning to the high cost of running LLMs
- We use proxy metrics that are cost-efficient to compute to first find more promising combinations to search over



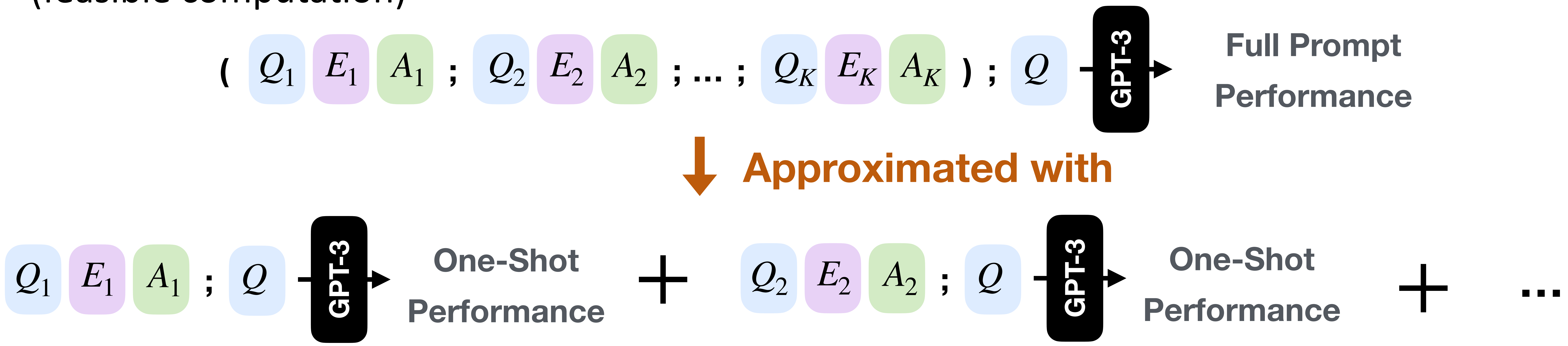
Approach Overview

- **Generate candidate explanations**
  - This yields **combinations** of explanations
- **Silver-label development set:** sample combinations and vote to silver-label V
- **Use proxy metrics to pre-filter promising combinations**
- **Select combination based on silver-accuracy:** score combinations using silver-accuracy



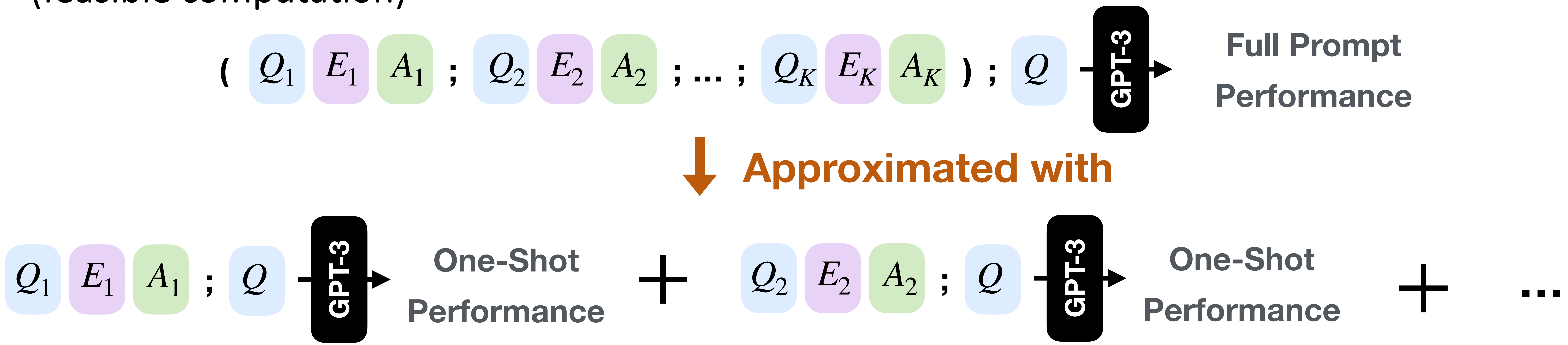
Proxy Metrics

- **One-shot Silver Accuracy:** we approximate the accuracy of a combination by the aggregated one-shot accuracy
  - We can score any combinations with this proxy metric once we score all Q,E,A individually (feasible computation)



Proxy Metrics

- **One-shot Silver Accuracy:** we approximate the accuracy of a combination by the aggregated one-shot accuracy
  - We can score any combinations with this proxy metric once we score all Q,E,A individually (feasible computation)



- **One-shot Log-likelihood (skipped):** maximizing the one-shot likelihood on the few-shot exemplar sets
  - This allows using a few gold labels

$$\sum_{j=1:K} \sum_{i=1:K \wedge i \neq j} \log p(e_j, a_j \mid (q_i, e_i, a_i), q_j; \theta).$$

Experiment Setup

- **Datasets:** GSM (arithmetical reasoning), ECQA (commensenQA), ESNLI (natural language inference), StrategyQA (multi-hop open QA)
- **LLM:** Code-davinci-002
- **Data Condition:**

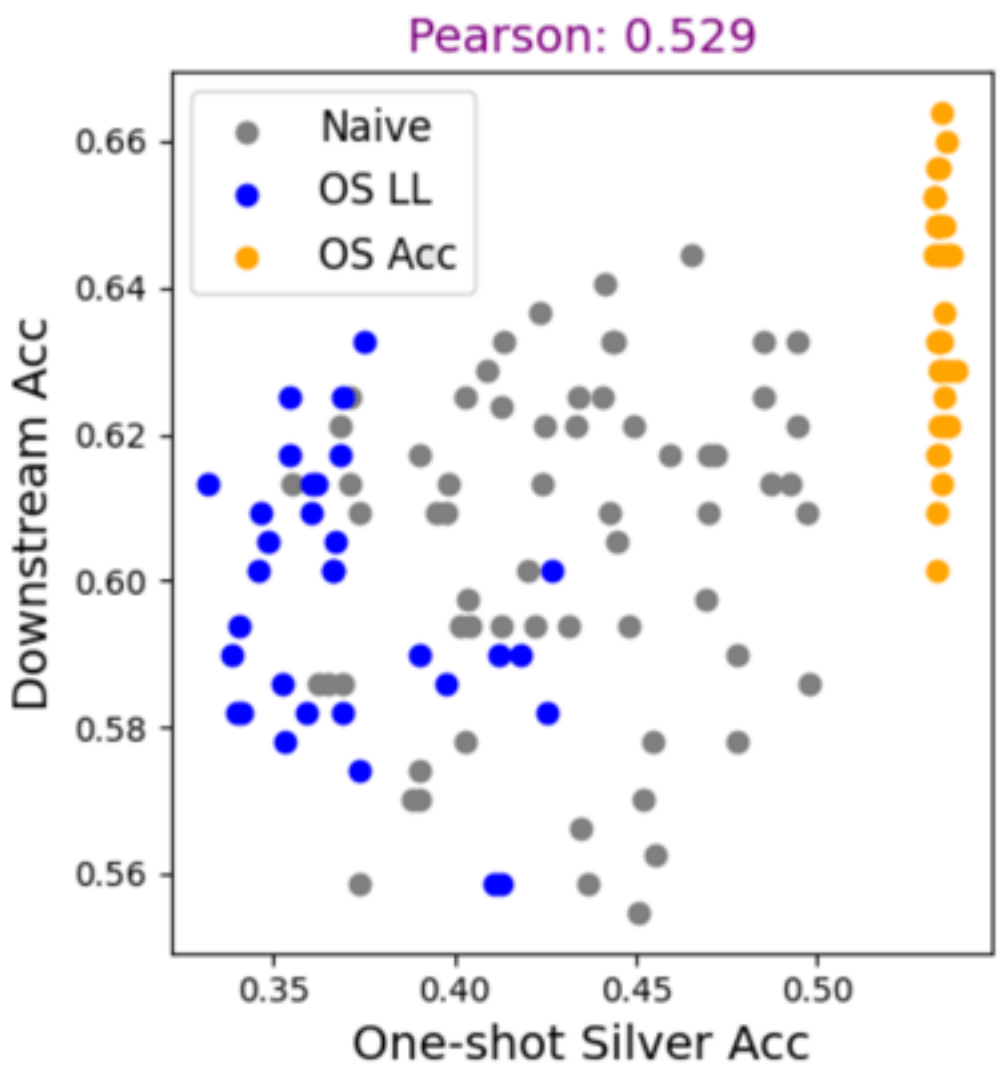
Few-Shot Exemplars	Q_1 A_1 ; Q_2 A_2 ; … ; Q_K A_K	K=8
Seed Explanations	$\tilde{E}_1$ $\tilde{E}_2$ … $\tilde{E}_K$	Crowdworker Annotations
Unlabeled Dev set	$V =$ Q_1 Q_2 … Q_M	M=256

Effectiveness of Proxy Metrics

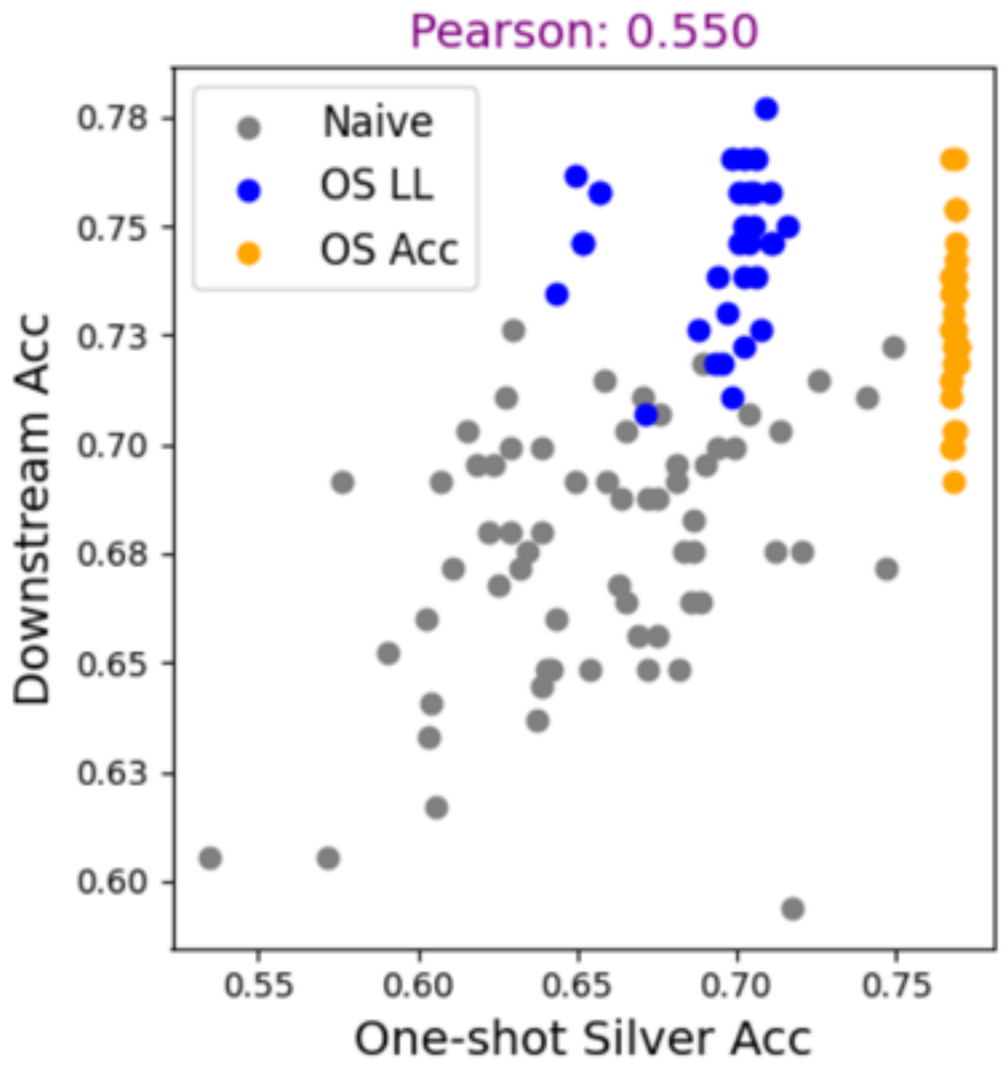
- **One-shot Silver Accuracy:** aggregated one-shot silver accuracy on the development set

X-Axis: proxy metrics **Y-Axis: downstream acc**

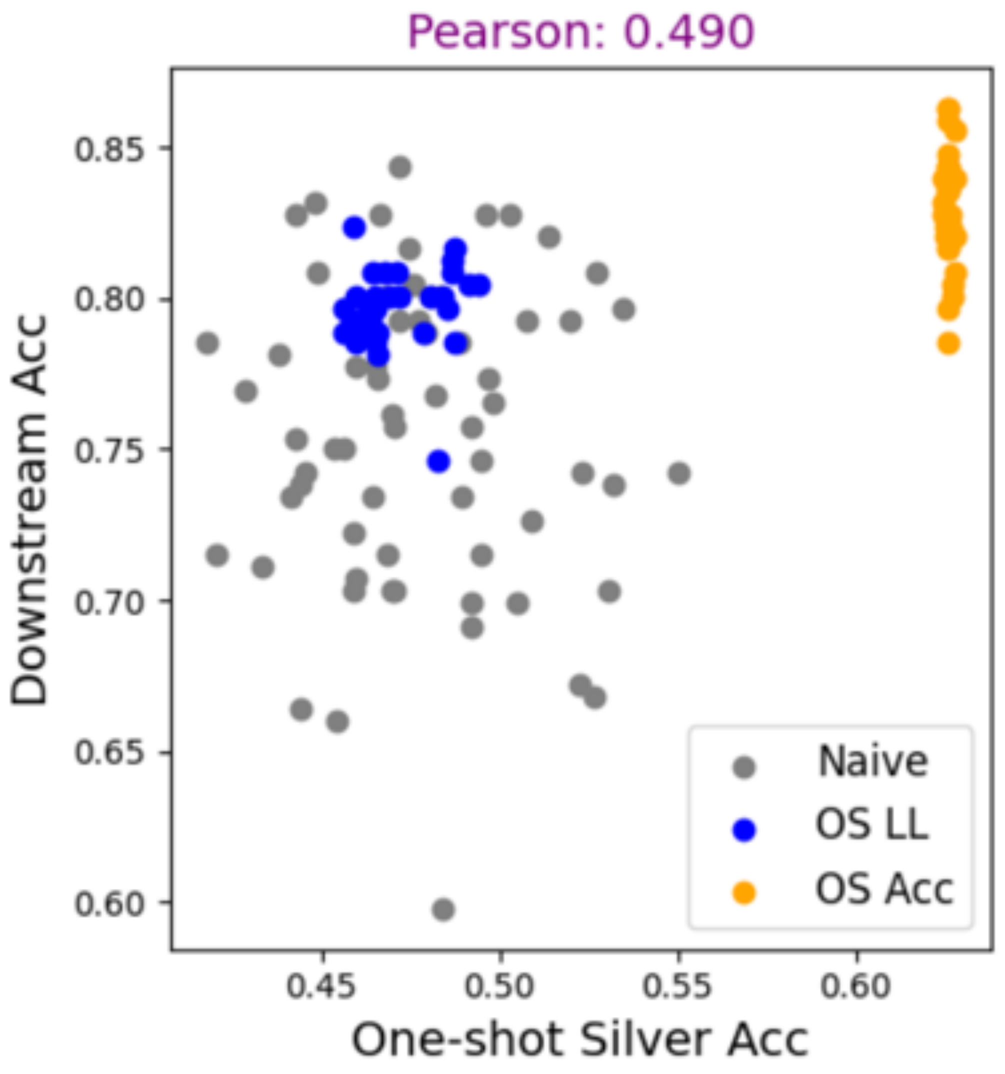
Colors: combinations preferred by different proxy metrics



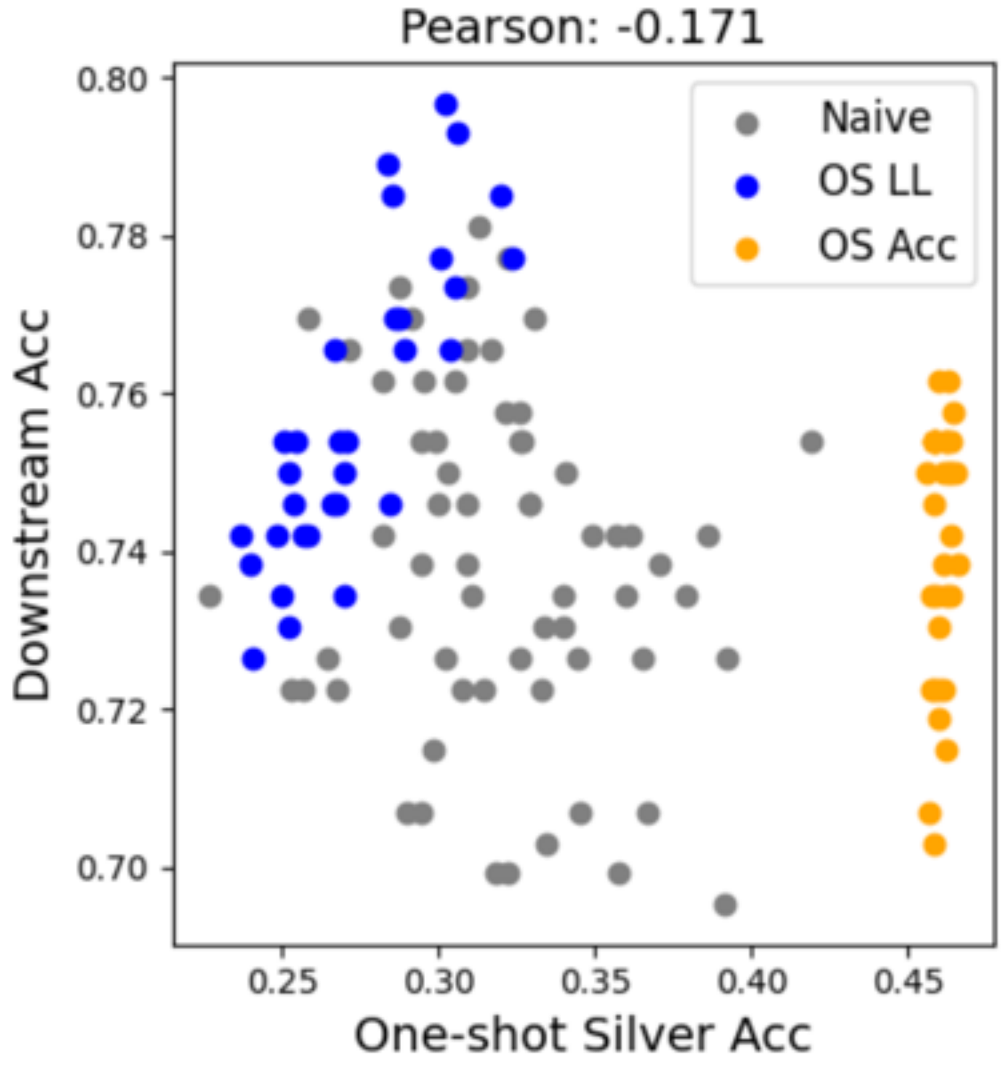
GSM



ECQA



ESNLI

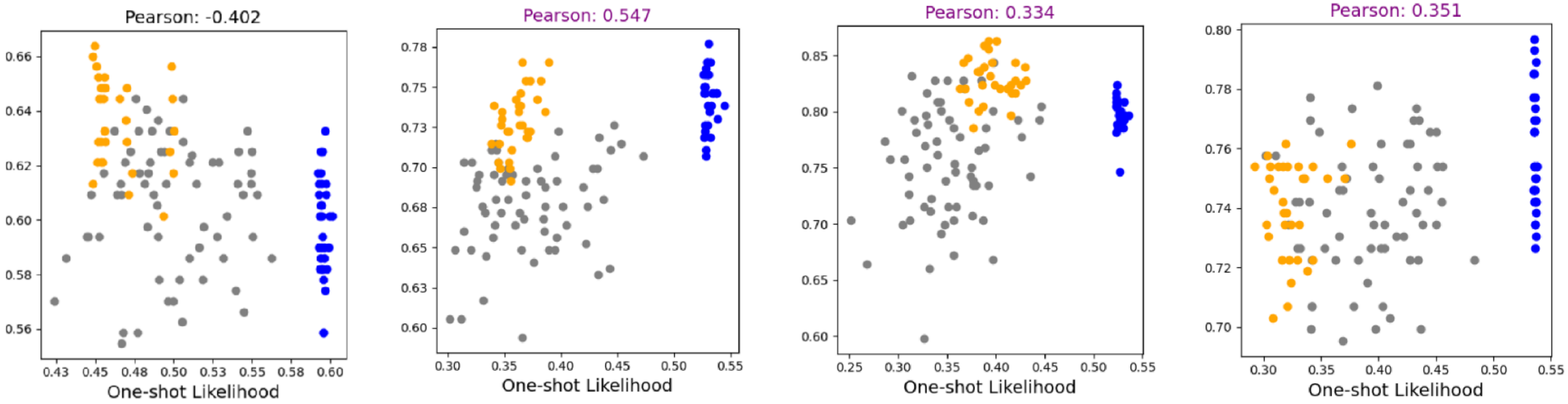


StrategyQA

- The proxy metrics correlates well with downstream accuracy in most cases

Effectiveness of Proxy Metrics

- **One-shot Silver Accuracy:** aggregated one-shot silver accuracy on the development set
- **One-shot Log-Likelihood:** aggregated one-shot likelihood on few-shot exemplars

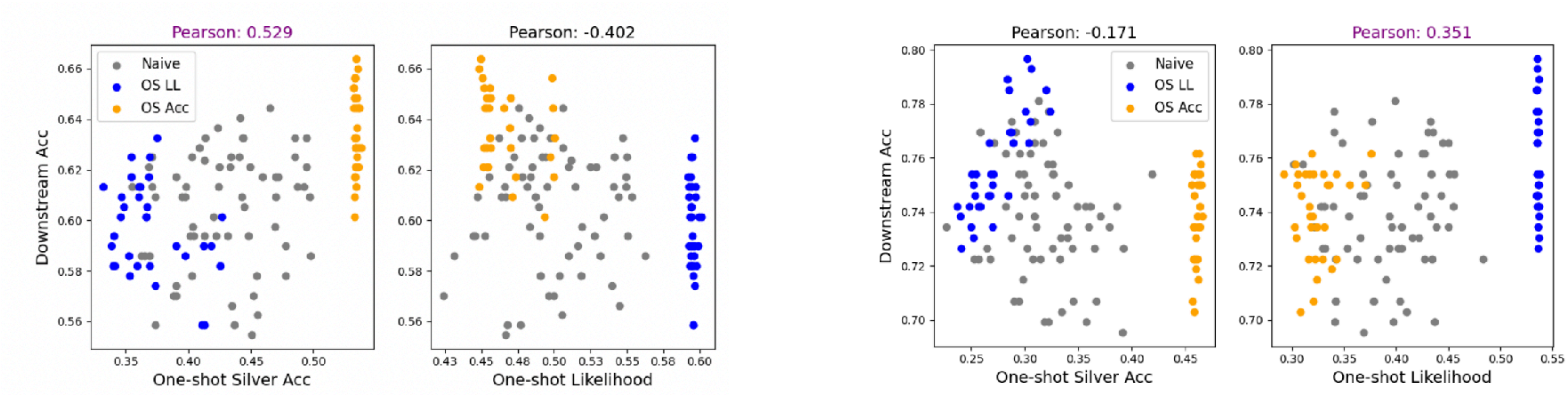


GSM | ECQA | ESNLI | StrategyQA

- **Similar trends**: the proxy metrics correlates well with downstream accuracy in most cases

Effectiveness of Proxy Metrics

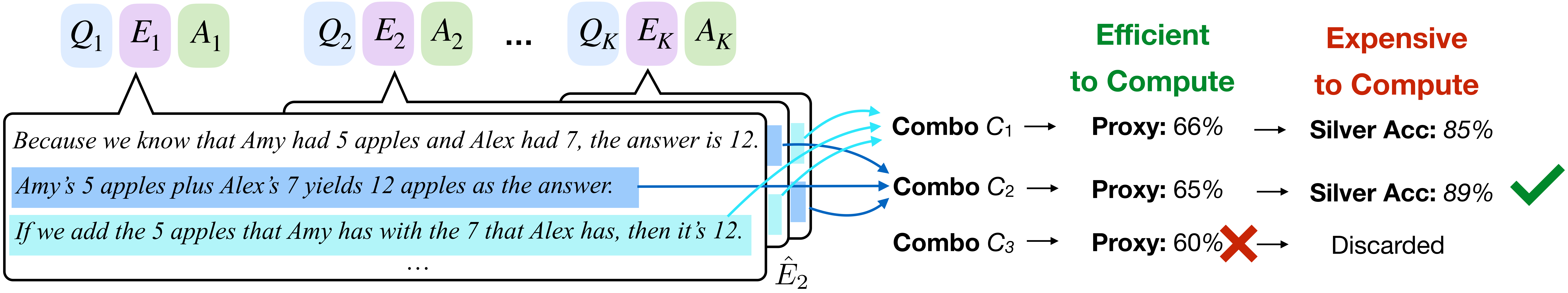
- **One-shot Silver Accuracy:** aggregated one-shot silver accuracy on the development set
- **One-shot Log-Likelihood:** aggregated one-shot likelihood on few-shot exemplars
- Using approximate metrics allows prioritize search over betters combinations than naive (randomly sampled combinations)
- No **one-size-fit-all solution**



GSM: OSAcc ✓ **GSM: OSLL ✗** **StrategyQA: OSAcc ✗** **StrategyQA: OSLL ✓**

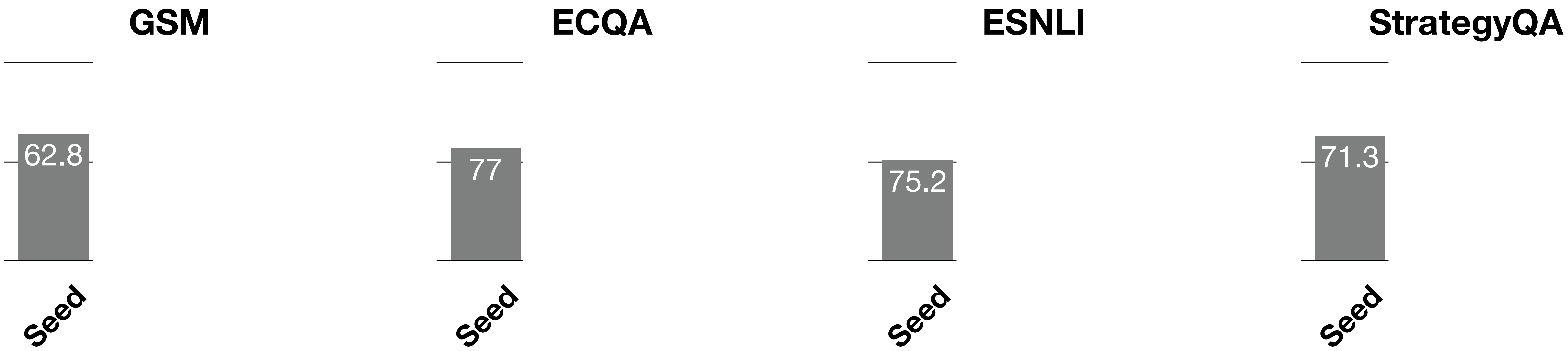
Approach Overview

- **Generate candidate explanations**
  - This yields **combinations** of explanations
- **Silver-label development set:** sample combinations and vote to silver-label V
- **Use proxy metrics to pre-filter promising combinations**
- **Select combination based on silver-accuracy:** score combinations using silver-accuracy



Main Experiments

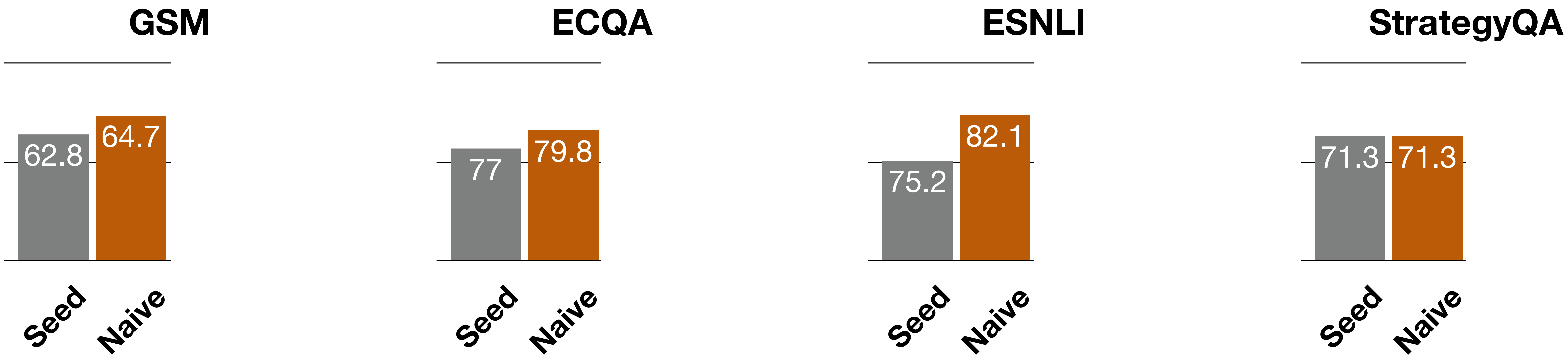
- **Seed:** initial explanations



- Results are averaged from four trials with four randomly selected K exemplars

Main Experiments

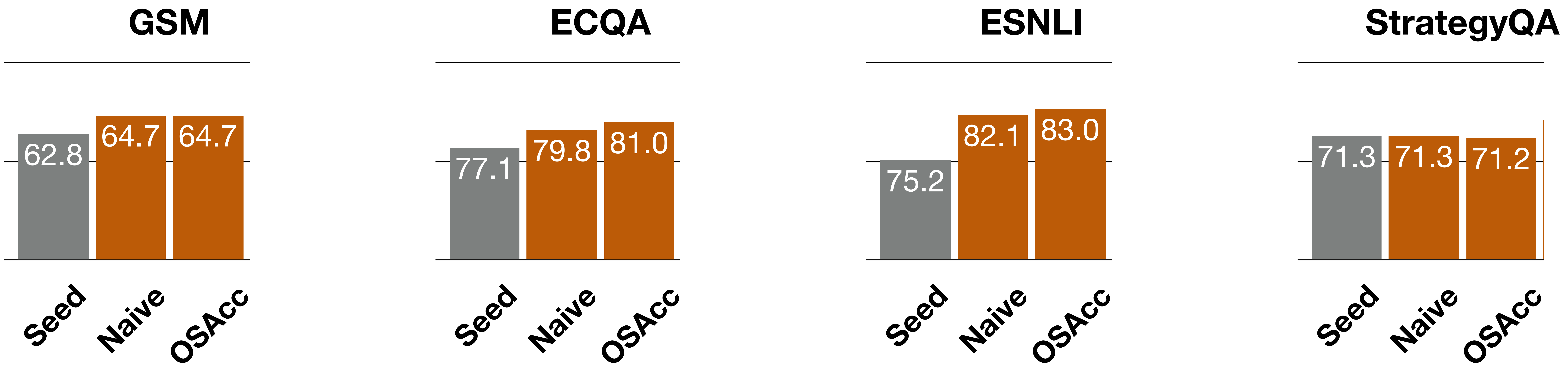
- **Seed:** initial explanations
- **Naive:** using our framework to search over random combinations



- Results are averaged from four trials with four randomly selected K exemplars
- Applying our optimization framework and search over random combinations can already yield better performing explanations

Main Experiments

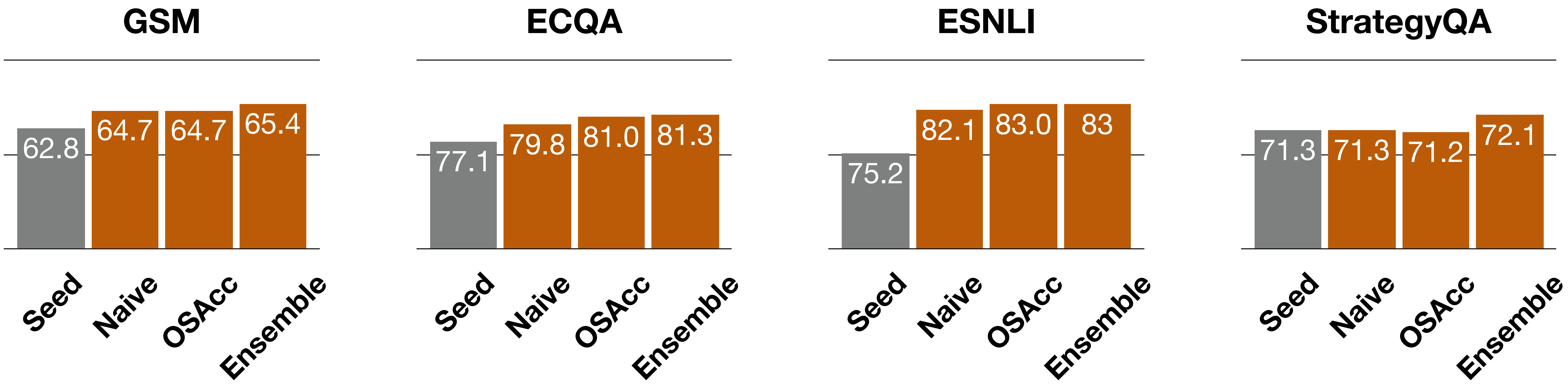
- **Seed:** initial explanations
- **Naive:** using our framework to search over random combinations
- **OSAcc:** search over combinations found by OSAcc



- Using the proxy metric allows us prioritize search on better performing combinations, which yields better results in general

Main Experiments

- **Seed:** initial explanations
- **Naive:** using our framework to search over random combinations
- **OSAcc:** search over combinations found by OSAcc



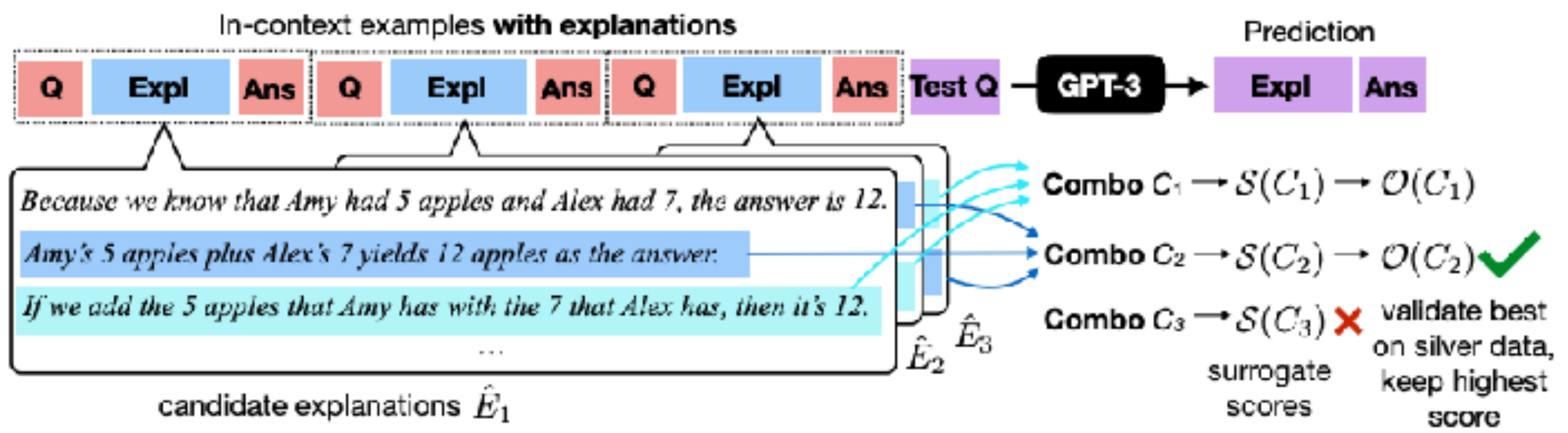
- **Ensemble:** search over combinations found by OSAcc + OSLL
  - Achieves the best performance overall

Wrap-up

- We can optimize for better explanations regarding downstream performance, using only unlabeled data
- We propose two proxy metrics to prioritize exploring better combinations given a limited computation

Explanation Selection using Unlabeled Data for In-Context Learning

Xi Ye and Greg Durrett, ArXiv 2023



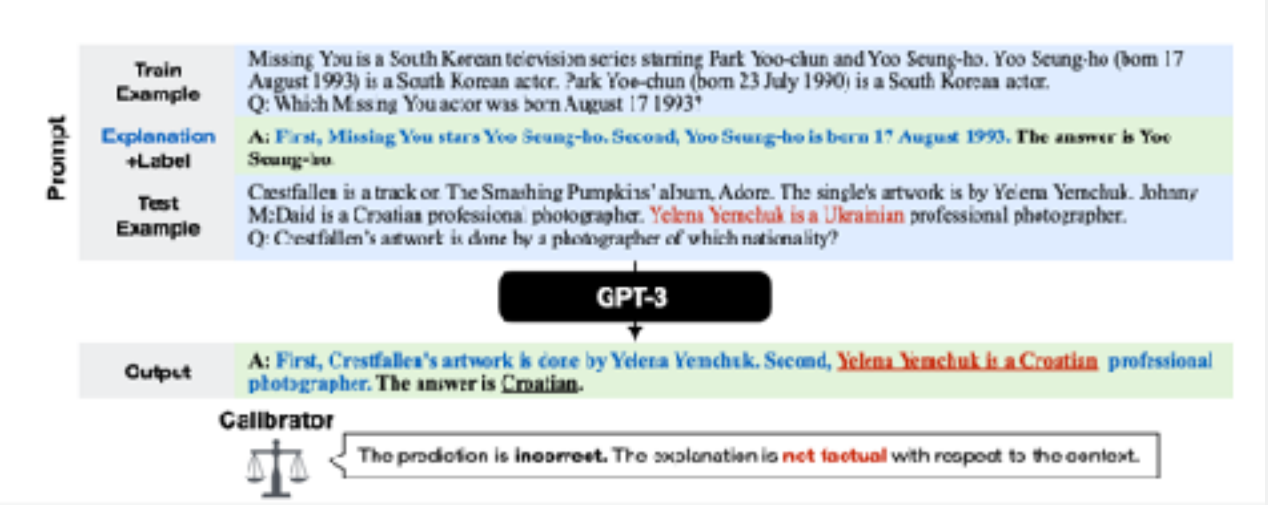
Outline

How well can LLMs learn from explanations in-context?
How to make explanations work better?

The Unreliability of Explanations in Few-Shot Prompting for Textual Reasoning

X Ye and G Durrett, NeurIPS 22

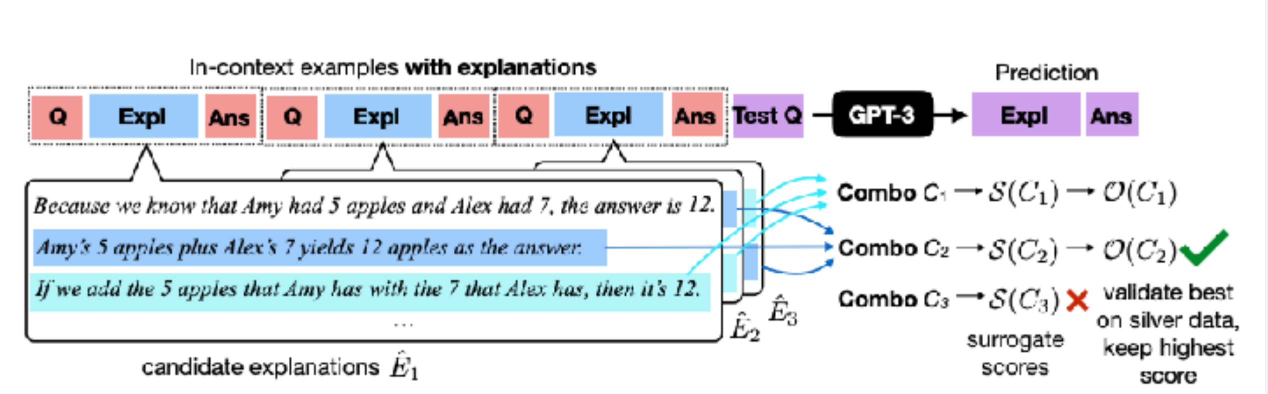
- Benchmark the effective of explanations in-context



Explanation Selection using Unlabeled Data for In-Context Learning

X Ye and G Durrett, ArXiv 23

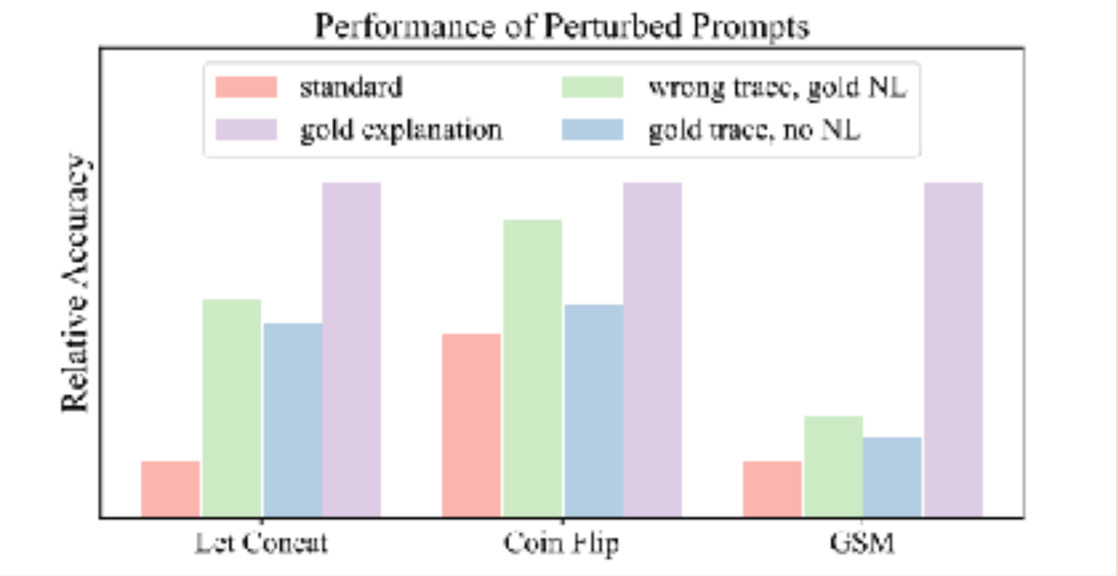
- Optimize explanations to improve downstream performance



Complementary Explanations for Effective In-Context Learning

X Ye, S Iyer, A Celikyilmaz, V Stoyanov, G Durrett, and R Pasunuru, ACL Findings 23

- Empirical analysis on how explanations work in in-context learning



How Explanations Work?

- LMs don't "follow" prompts in some ways

PROMPT WAYWARDNESS: The Curious Case of Discretized Interpretation of Continuous Prompts

Daniel K
Lianhui Qi
Hannaneh Hajishirzi[†‡] Tusha
[†]University of Washington

Do Prompt-Based Models Really Understand the Meaning of Their Prompts?

Albert Webso
{albert_webson, e
[1]Department of Compu
[2]Department of Phi

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Sewon Min[1,2] Xinxi Lyu[1] Ari Ho
Mike Lewis[2] Hannaneh Hajishirz
[1]University of Washington [2]Meta AI
{sewon,alrope,ahai,hannaneh,ls
{artetxe,mikelewis}

Can Large Language Models Truly Understand Prompts? A Case Study with *Negated* Prompts

Joel Jang* KAIST joeljang@kaist.ac.kr
Seongheyon Ye* KAIST seonghyeon.ye@kaist.ac.kr
Minjoon Seo KAIST minjoon@kaist.ac.kr

- Do LMs "follow" explanations? How do explanations work for in-context-learning?

What Makes Explanations Effective?

- Probe LLMs with perturbed explanations
  - Perturbing Computation Trace
  - Perturbing Natural Language

Question	Take the last letters of the words in "Bill Gates” and concatenate them.
Gold Explanation (Trace, NL)	The last letter of "Bill" is letter"l". The last of "Gates" is "s". Concatenating “l" and "s" is “ls". So the answer is ls.
Perturbing Trace	The last letter of "Bill" is letter " ". The last of "Gates" is " ". Concatenating “l" and "s" is “ls". So the answer is ls.
Perturbing NL	"Bill","l","Gates","s","l","s","ls". So the answer is ls.

What Makes Explanations Effective?

- Probe LLMs with perturbed explanations
  - Perturbing Computation Trace
  - Perturbing Natural Language

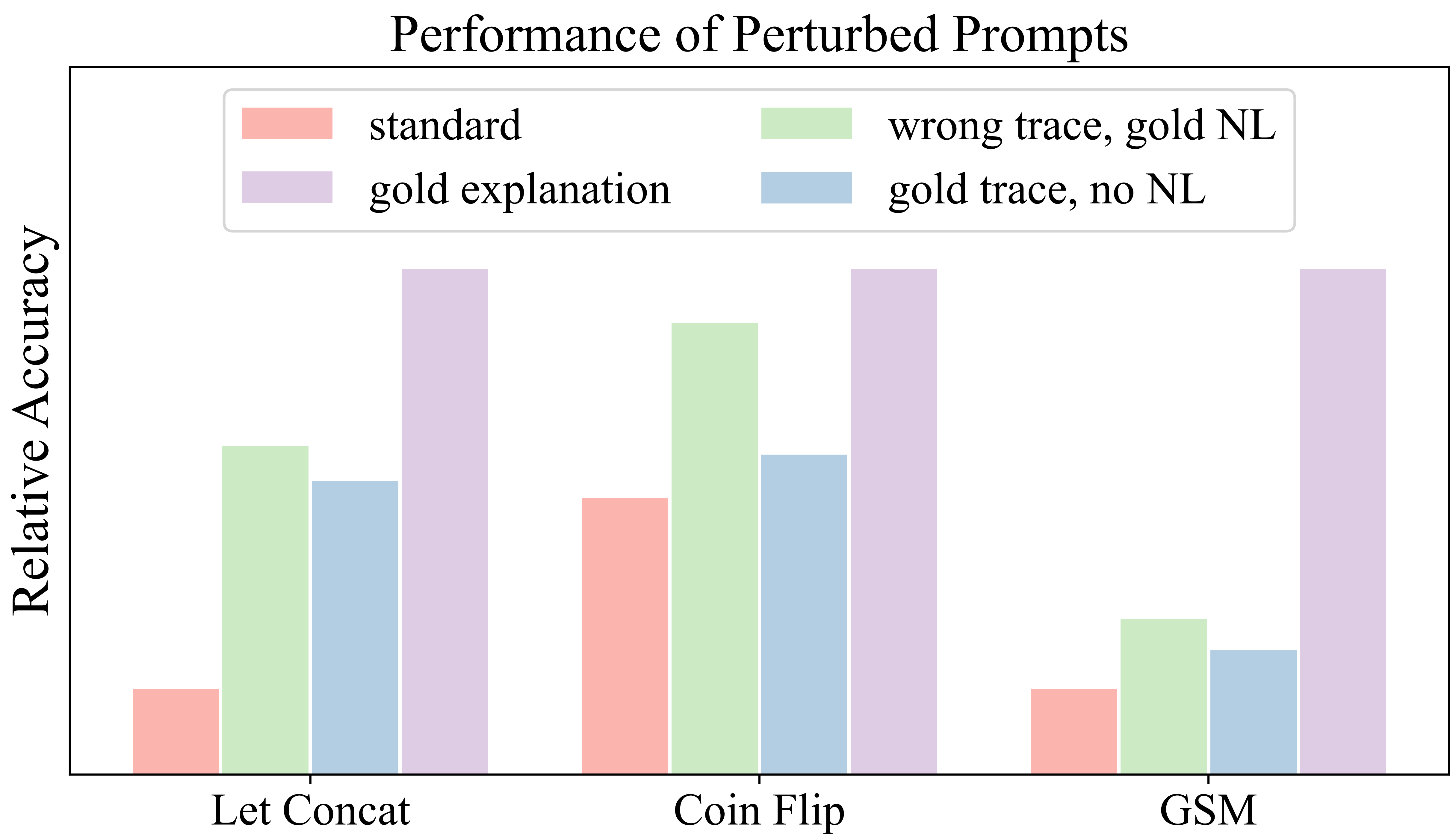
	Question: Take the last letters of the words in "Bill Gates" and concatenate them.
LETCONCAT	**Gold:** The last letter of Bill is l. The last letter of Gates is s. Concatenating l and s is ls. So the answer is ls. **Mask1:** The last letter of Bill is _. The last letter of Gates is _. Concatenating l and s is ls. So the answer is ls. **Mask2:** The last letter of Bill is l. The last letter of Gates is n. Concatenating _ and _ is _. So the answer is ln. **Incorrect:** The last letter of "Bill" is "y". The last letter of "Gates" is "e". Concatenating "y" and "e" is "ye". So the answer is ye. **No NL:** "Bill", "l". "Gates", "s". "l", "s", "ls". So the answer is ls.

	Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
GSM	**Gold:** Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 74 chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 chocolates. The answer is 39. **Mask1:** Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = _ chocolates. 35 have been eaten. So in total they still have _ - 35 = 39 chocolates. The answer is 39. **Mask2:** Leah had 32 chocolates and Leah's sister had 42. That means there were originally _ chocolates. 35 have been eaten. So in total they still have _ chocolates. The answer is 39. **Incorrect:** Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 62 chocolates. 35 have been eaten. So in total they still have 62 - 35 = 27 chocolates. The answer is 27. **No NL:** 32 + 42 = 74, 74 - 35 = 39. The answer is 39.

	Question: A coin is heads up. Shaunda does not flip the coin. Shalonda flips the coin. Is the coin still heads up?
COINFLIP	**Gold:** The coin started heads up. Shaunda does not flip the coin, so it becomes heads up. Shalonda flips the coin, so it becomes tails up. So the answer is no. **Mask1:** The coin started heads up. Shaunda does not flip the coin, so it becomes _ up. Shalonda flips the coin, so it becomes tails up. So the answer is no. **Mask2:** The coin started heads up. Shaunda does not flip the coin, so it becomes heads up. Shalonda flips the coin, so it becomes _ up. So the answer is no. **Incorrect:** The coin started heads up. Shaunda does not flip the coin, so it becomes tales up. Shalonda flips the coin, so it becomes heads up. So the answer is yes. **No NL:** heads, heads, tails. So the answer is no.

How Explanations Work?

- Do LMs "follow" explanations?
  - YES. Perturbing either trace or NL leads to performance degradation.
  - Both trace and NL contribute to making effective explanations
  - But perturbed explanations are still beneficial compared to not using explanations at all



What Makes A Good Set of Explanations?

- Given a test query , we study how to form a maximally effective **set** of exemplars
  - Interplay between query and exemplar**: relevance** (using more relevant examples)
  - Interplay between exemplars in the set: **complementarity**

Test Query:
Q: Peter bought 20 popsicles at $0.25 each. He bought 4 ice cream bars at $0.50 each. How much did he pay in total?
A: 0.25 * 20 = 5. 0.5 * 4 = 2. 5 + 2 = 7. The answer is 7.

Complementary

Addition Exemplars:
Q: Marion received 20 more turtles than Martha. If Martha received 40 turtles, how many turtles did they receive together?
A: **20 + 40 = 60. 60 + 40 = 100.** The answer is 100.

Multiplication Exemplars:
Q: Car Wash Company cleans 80 cars per day. They make $5 per car washed. How much money will they make in 5 days?
A: **8 * 5 = 40. 40 * 5 = 2000**. The answer is 2000

Probing with Complementary Exemplars

- We test whether LLMs can benefit from complementarity of exemplars

Addition Exemplars:
Q: Marion received 20 more turtles than Martha. If Martha received 40 turtles, how many turtles did they receive together?
A: 20 + 40 = 60. 60 + 40 = 100. The answer is 100.

Multiplication Exemplars:
Q: Car Wash Company cleans 80 cars per day. They make $5 per car washed. How much money will they make in 5 days?
A: 8 * 5 = 40. 40 * 5 = 2000. The answer is 2000

Experiments Setup

Addition Exemplars → GPT-3 →

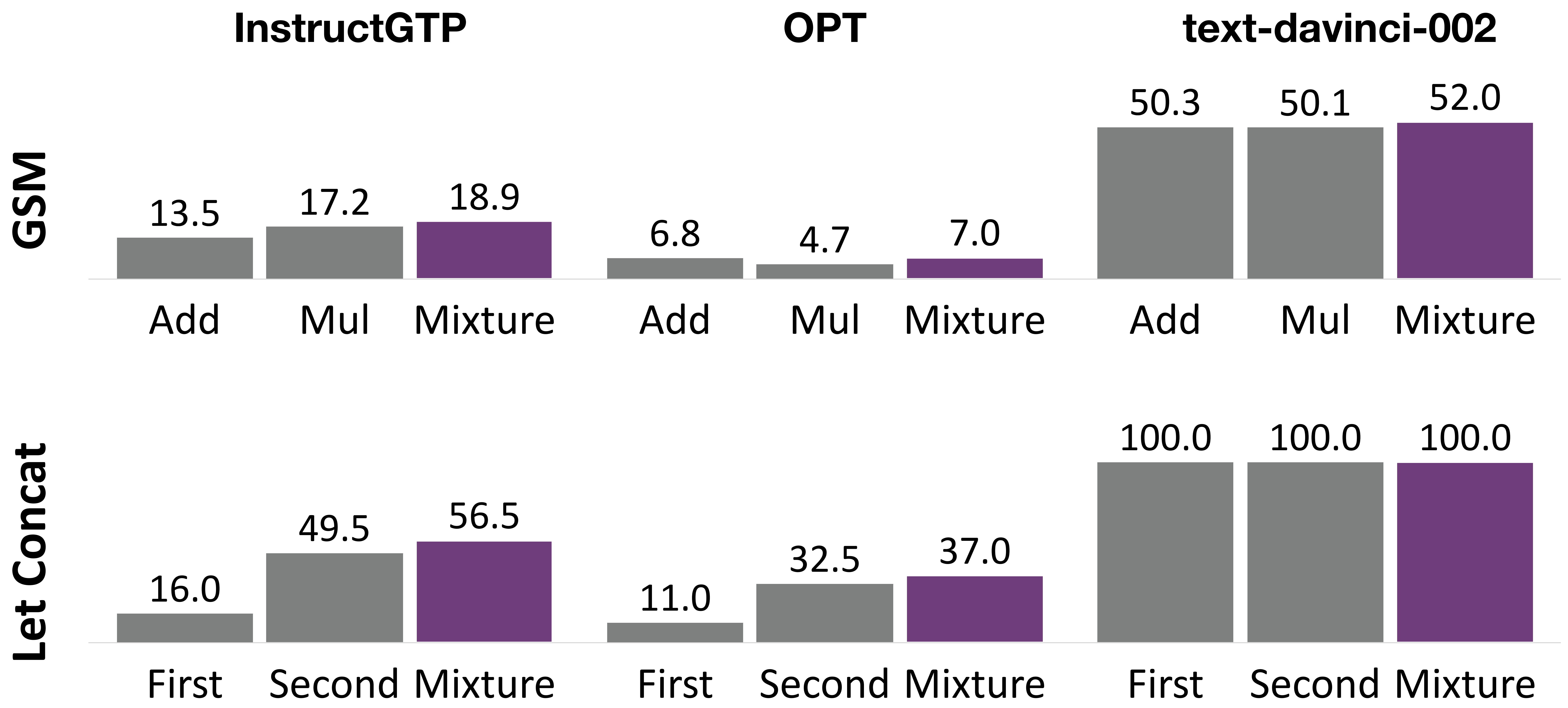
Multiplication Exemplars → GPT-3 →

Mixture of (Add and Mul) → GPT-3 →

Test Data:
Q: Peter bought 20 popsicles at $0.25 each. He bought 4 ice cream bars at $0.50 each. How much did he pay in total?
A: 0.25 * 20 = 5. 0.5 * 4 = 2. 5 + 2 = 7. The answer is 7.

Probing with Complementary Exemplars

- Complementary exemplar sets lead to better performance



MMR for Exemplar Selection

- Prominent nearest neighbor-based exemplar selection method only considers relevance
- We propose a maximal-marginal-relevance (MMR) -based exemplar selection method, which selects **diverse** exemplars that are **relevant** to the test query

Test Query	Currently Selected Exemplars	Distance Metric
Q	$T = Q_1, Q_2, \ldots, Q_{k-1}$	$S(Q_i, Q_j)$

Next Exemplar to Select

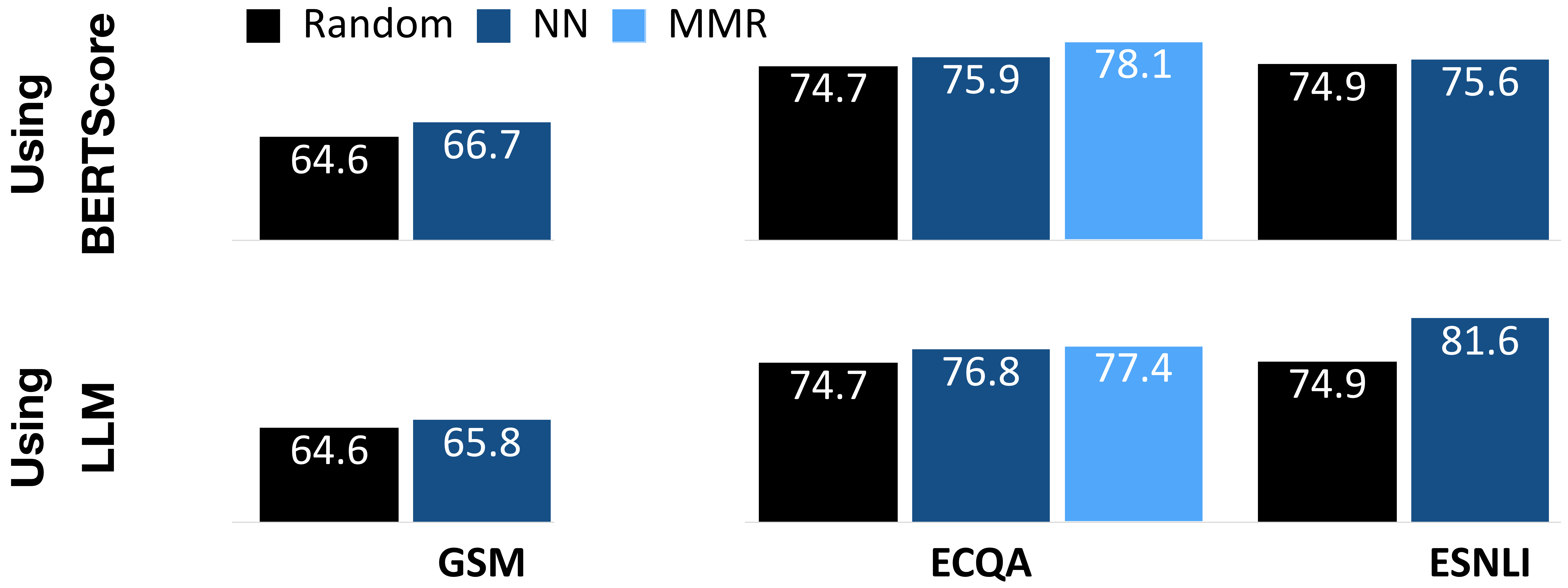
$$Q_k = \underset{Q_j}{\arg\max}\ \lambda S(Q, Q_j) - (1-\lambda) \max_{Q_i \in T} S(Q_j, Q_i)$$

Relevant to test query

Diverse w.r.t. already selected exemplars

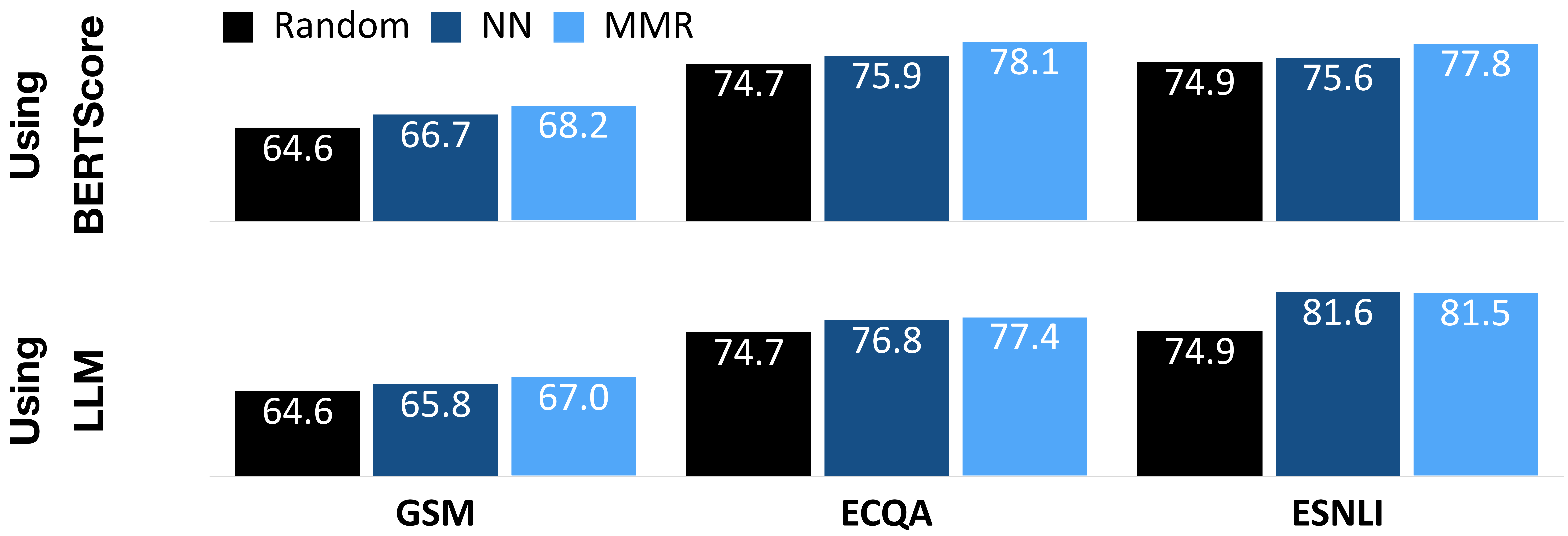
Experiments

- **Datasets:** GSM, ECQA, E-SNLI **LLM:** code-davinci-002
- **Baselines: random** exemplar selection; **nearest neighbor**-based exemplar selection
- **Distance Metrics**:
  - **BERTScore:** $S(Q_i, Q_j) = BERTScore(Q_i, Q_j)$ **LLMScore:** $S(Q_i, Q_j) = P_{LLM}(Q_i \mid Q_j)$



Experiments

- **Datasets:** GSM, ECQA, E-SNLI **LLM:** code-davinci-002
- **Baselines: random** exemplar selection; **nearest neighbor**-based exemplar selection
- **Distance Metrics**:
  - **BERTScore:** $S(Q_i, Q_j) = BERTScore(Q_i, Q_j)$ **LLMScore:** $S(Q_i, Q_j) = P_{LLM}(Q_i \mid Q_j)$



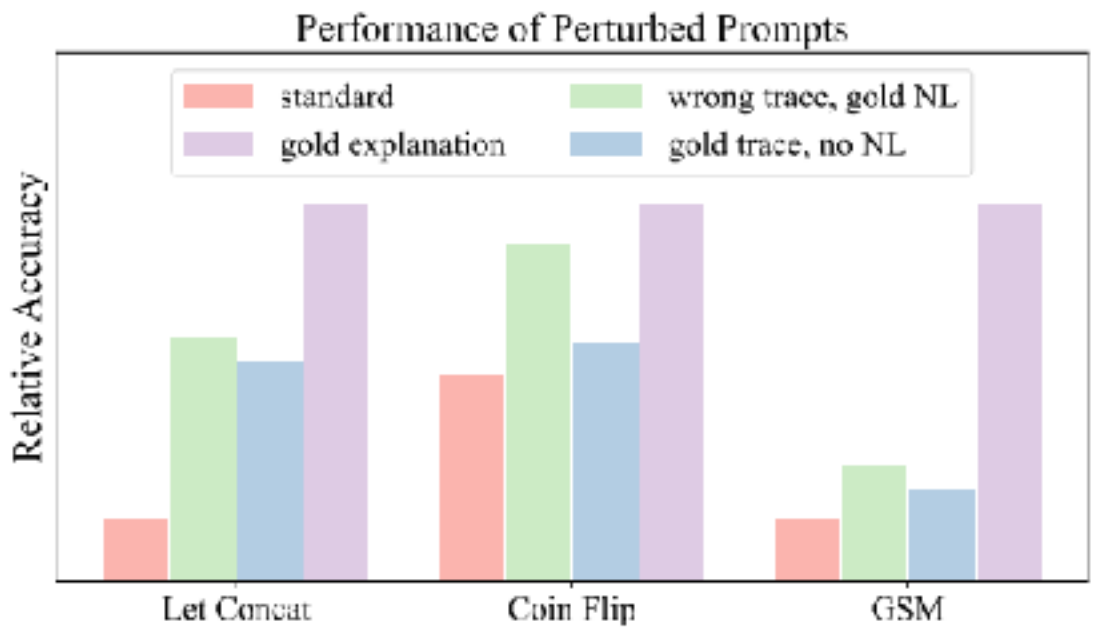
- MMR is more effective than NN in general across different datasets and different metrics

Wrap-up

- Both computation trace and NL contributes to effective explanations
- LLMs can benefit from complementary explanations

Complementary Explanations for Effective In-Context Learning

X Ye, S Iyer, A Celikyilmaz, V Stoyanov, G Durrett, and R Pasunuru, ACL Findings 23



Takeaways!

- **How well can LLMs learn from explanations in prompts?**
  - Only more advanced LLMs (like text-davinci-002) can benefit substantially
  - The generated explanations might be unreliable
- **How to make explanations work better?**
  - We can optimize explanations using unlabeled data
  - We can construct explanations that are relevant and diverse
- **What about now?**
  - More recent LLMs have incredibly strong reasoning abilities; but they can still generate unreliable explanations

What about Now?

XI

Each of five experts—a lawyer, a naturalist, an oceanographer, a physicist, and a statistician—individually gives exactly one presentation at a conference. The five presentations are given consecutively. Each presentation is in exactly one of the four following languages: French, Hindi, Japanese, or Mandarin. Each expert speaks exactly one of the languages. The following conditions must hold: Exactly two of the presentations are in the same language as each other. The statistician gives the second presentation in Hindi. The lawyer gives the fourth presentation in either Mandarin or French. The oceanographer presents in either French or Japanese; the same is true of the physicist. The first presentation and the last presentation are in Japanese.
Question: Which one of the following could be the order in which the experts give their presentations, from first to last?
Choices:
(A) the physicist, the statistician, the lawyer, the naturalist, the oceanographer
(B) the physicist, the naturalist, the oceanographer, the lawyer, the statistician
(C) the oceanographer, the statistician, the naturalist, the lawyer, the physicist
(D) the oceanographer, the statistician, the lawyer, the naturalist, the physicist

Please answer this question. Please think carefully and explain the thought process before answering.

To solve this problem, we need to use the given conditions to determine the possible orders in which the experts could give their presentations.

1. The first presentation and the last presentation are in Japanese, so the first and fifth positions must be Japanese.
2. The statistician gives the second presentation in Hindi, so the second position must be Hindi.
3. The lawyer gives the fourth presentation in either Mandarin or French, so the fourth position must be either Mandarin or French.
4. The oceanographer presents in either French or Japanese; the same is true of the physicist, so the third position can be either French or Japanese.
5. Exactly two of the presentations are in the same language as each other, which means that the third and fourth positions must be in the same language.

Now let's go through each of the answer choices and see if they satisfy all the conditions:

(A) the physicist, the statistician, the lawyer, the naturalist, the oceanographer

- The first and fifth positions are Japanese, and the second position is Hindi, which satisfies conditions 1 and 2.
- The fourth position can be either Mandarin or French, which satisfies condition 3.
- The third position can be either French or Japanese, which satisfies condition 4.
- However, this order does not satisfy condition 5, because the third and fourth positions are occupied by the lawyer and the naturalist, who speak different languages.

(B) the physicist, the naturalist, the oceanographer, the lawyer, the statistician

- The first and fifth positions are Japanese, and the second position is Hindi, which satisfies conditions 1 and 2.
- The fourth position can be either Mandarin or French, which satisfies condition 3.
- The third position can be either French or Japanese, which satisfies condition 4.
- This order satisfies condition 5, because the third and fourth positions are occupied by the oceanographer and the lawyer, who speak French.
  Therefore, this order could be possible.

...

TAUR LAB

taur.cs.utexas.edu

Greg Durrett

Meta

Ram Pasunuru

Srini Iyer

Asli Celikyilmaz

Ves Stoyanov