

Steering Textual Reasoning Using Explanations



Xi Ye

Nov 2023

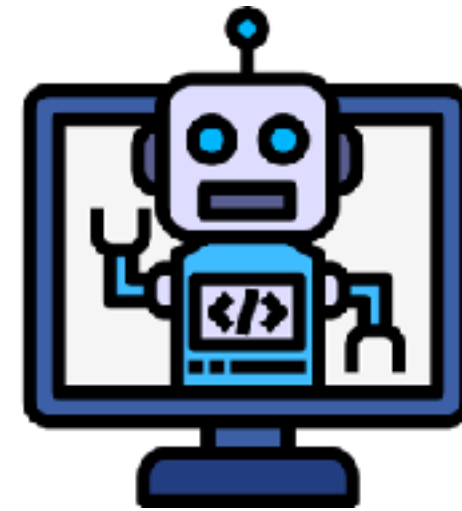
Language Models **Reason** over Text



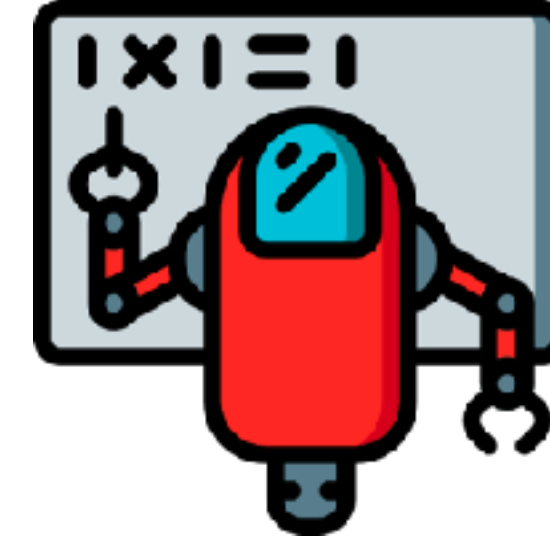
Language Models



Reason over a text



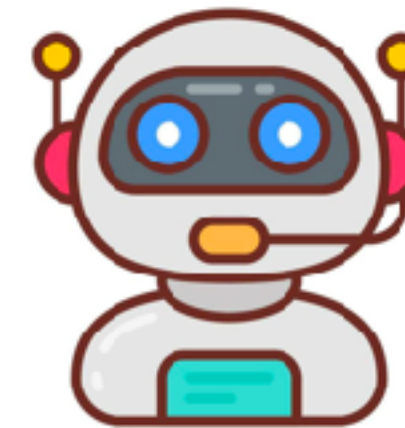
Programming Assistant



AI Math Tutor



Writing Assistant



Customer Assistant

Huge Progress in Reasoning



Math Dataset in 2020-ish

(Patel et al., 2021)

PROBLEM:

Text: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Jack have now?

Equation: $8 - 3 = 5$



Math Dataset Now

(Lewkowycz et al., 2022)

Question: Assume that the variance of the first n natural numbers is 10, and the variance of the first m even natural numbers is 16. Compute $m + n$.

Model output: Let σ^2 be the variance of the first m even natural numbers, and τ^2 be the variance of the first n natural numbers. Then, we have $\tau^2 = \frac{n^2-1}{12}$ and $\sigma^2 = \frac{m^2-1}{3}$. We solve the system of equations:

$$\frac{n^2 - 1}{12} = 10$$

$$\frac{m^2 - 1}{3} = 16.$$

This translates into the system

$$n^2 = 121$$

$$m^2 = 48 + 1 = 49.$$

Therefore, $n = \boxed{11}$ and $m = \boxed{7}$, so $n + m = \boxed{18}$.

Unreliability in Reasoning




Review

It is laughable. Watch this tragedy movie as a comedy and have a dozen good laughs !


Finetuned BERT



Sentiment
Positive

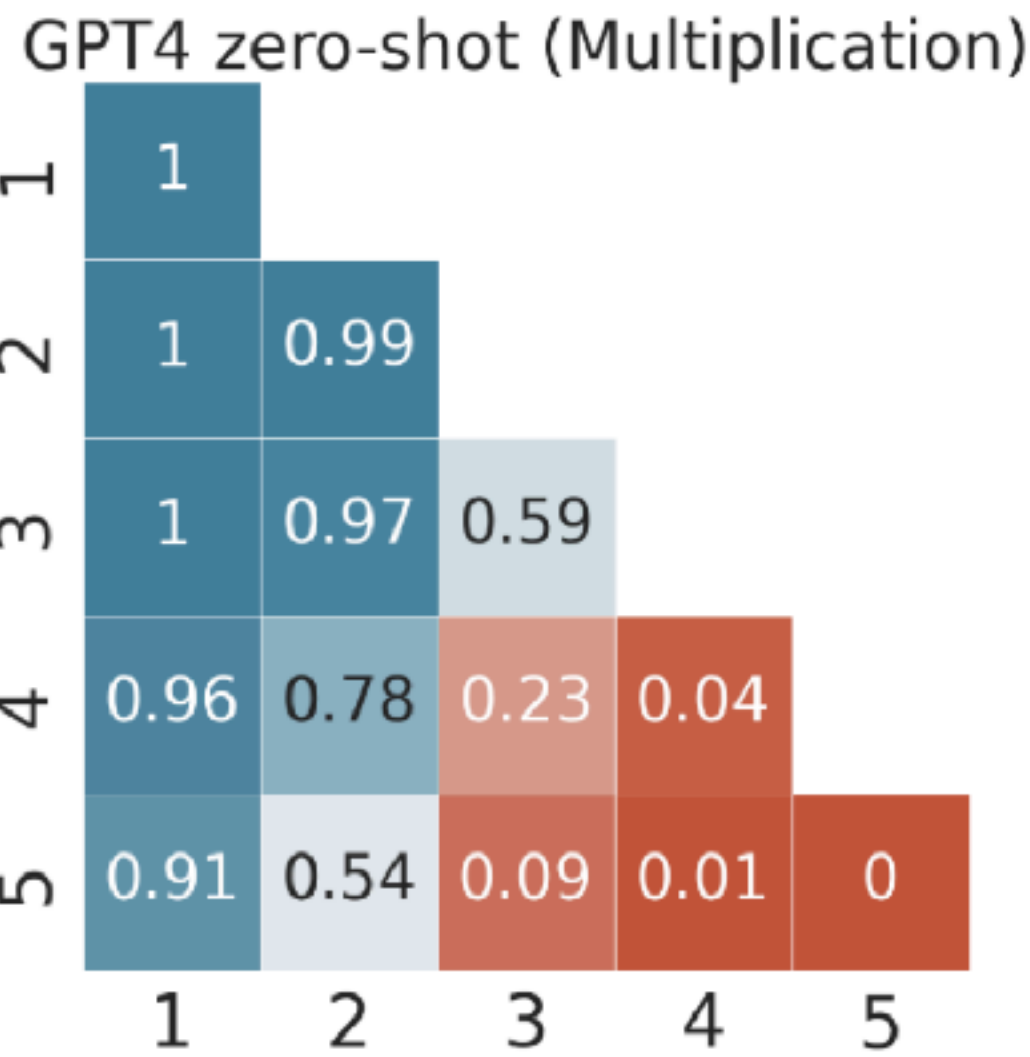


Determine the sentiment of the movie review:
"It is laughable. Watch this tragedy movie as comedy and have a dozen good laughs."



The sentiment of the movie review is positive. The reviewer describes the movie as "laughable" and suggests watching it as a comedy, which implies that they found it funny and entertaining, despite it being labeled as a tragedy.

Performance on the digit multiplication task decreases with growing complexity



(Dziri et al., 2023)

Being prone to learning surface clues instead of reasoning

Limitations in scaling to complex compositional reasoning

Unreliability in Reasoning



Finetuned BERT

Review

It is laughable. Watch this tragedy movie as a dozen good

Sentiment

Performance on the digit multiplication task decreases with growing complexity

My Research Goal:

Steering language models to perform reliable and complex reasoning **with explanations**



"laughable" and suggests watching it as a comedy, which implies that they found it funny and entertaining, despite it being labeled as a tragedy.

1 2 3 4 5

(Dziri et al., 2023)

Being prone to learning surface clues instead of reasoning

Limitations in scaling to complex compositional reasoning

Explanations (in NLP)

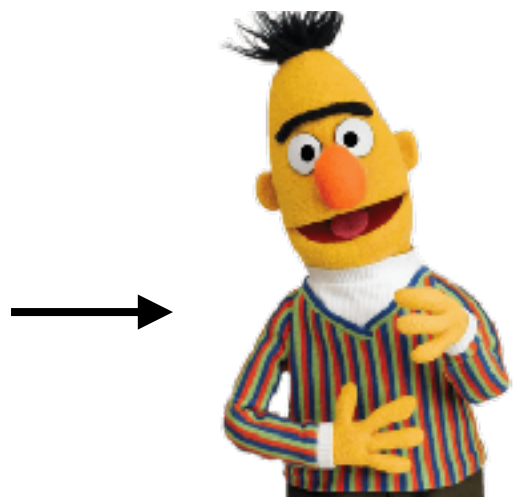


why is [input] assigned [label]?

Attributions

Review

It is laughable. Watch this tragedy movie as a comedy and have a dozen good laughs !



Sentiment
Positive

explain ↓

Review

It is laughable. Watch this tragedy movie as comedy and have a dozen good laughs.

Free-Text

Review It is laughable....

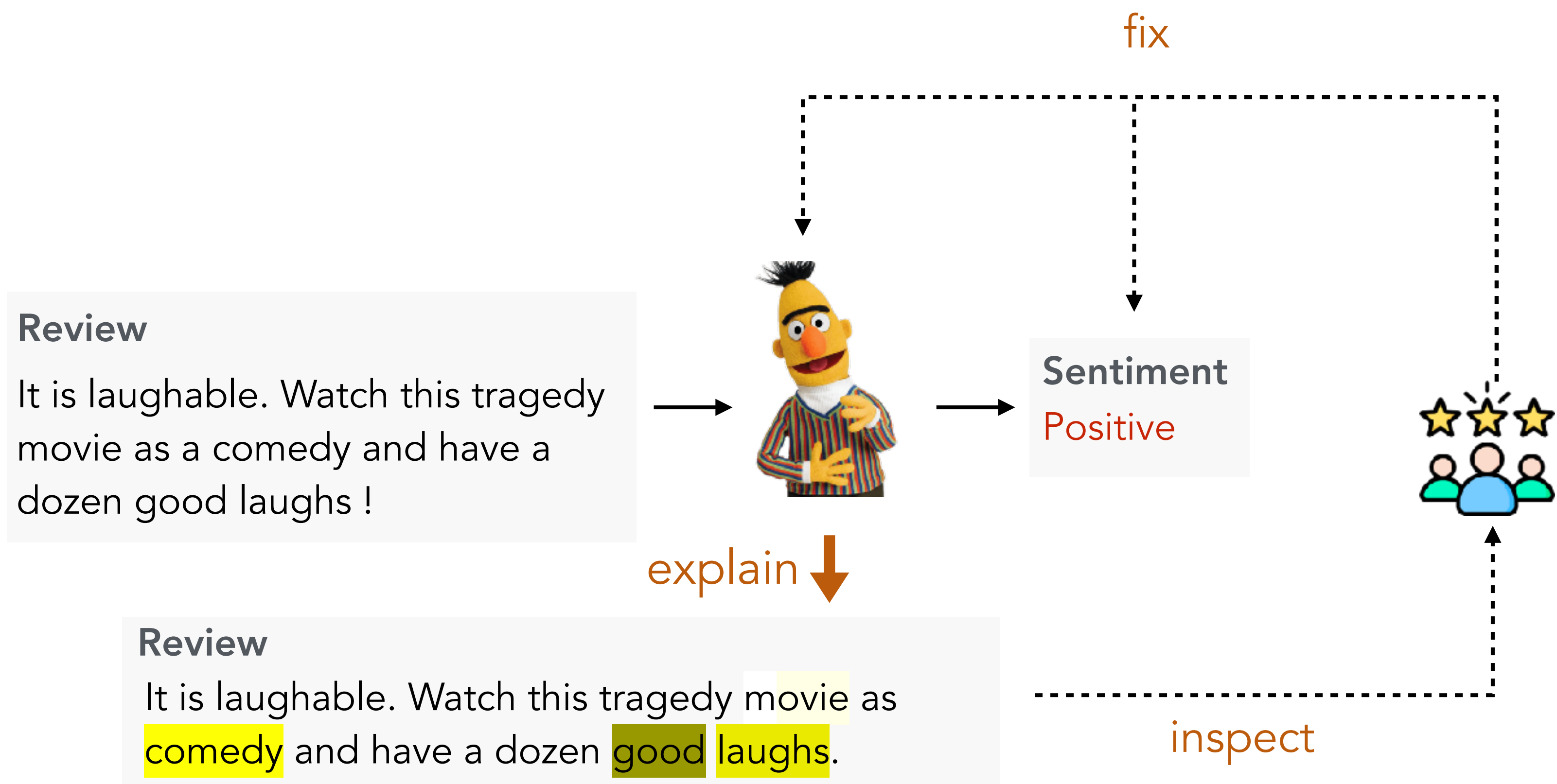


The sentiment is positive.

The review describes the movie as laughable, implies it finds it entertaining.



Using Explanations



Prominent Way

Use explanations to let **humans** make sense of the predictions and fix predictions or models

Humans develop a conceptual model of the LM's behavior (e.g., LM predicting positive when there are many individual tokens with positive sentiment).

Using Explanations



fix

Prominent Way

Use explanations to let **humans** make sense of the predictions

Our way:

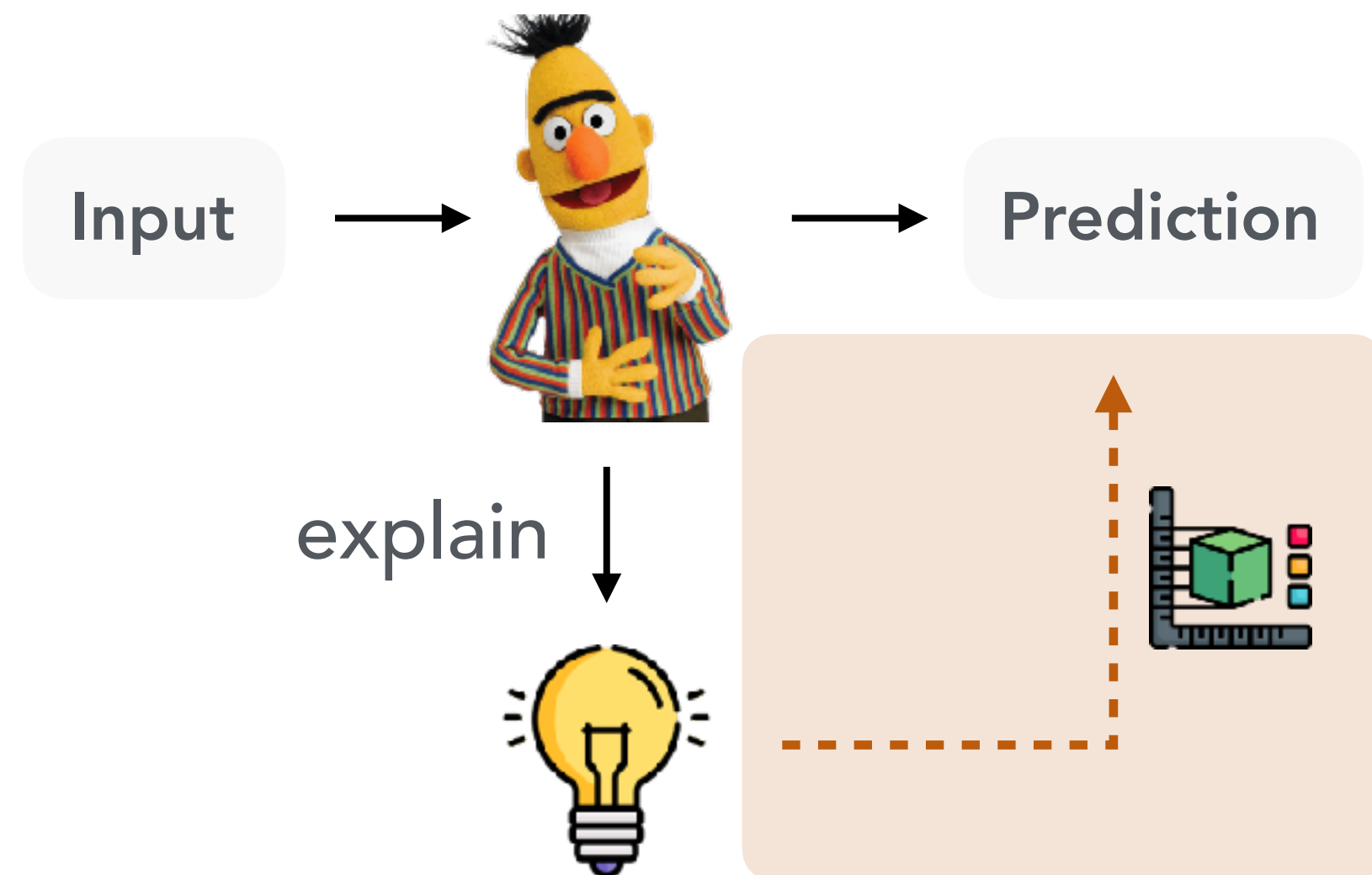
Automate the process of using explanations to understand and regulate LM behavior to improve model predictions

inspect

Steering Textual Reasoning with Explanations

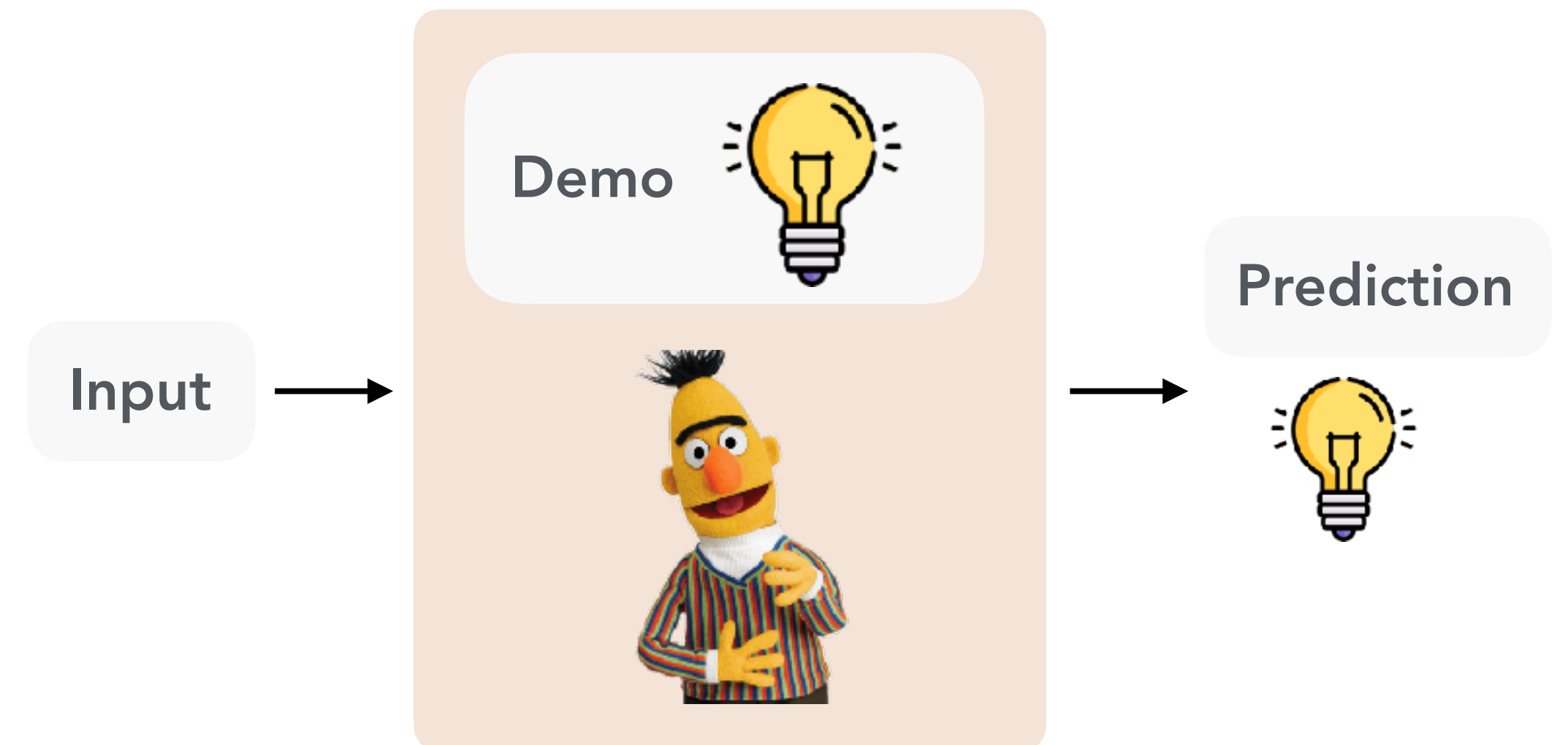


Post-Hoc Intervene



Assess correctness of predictions and intervene on predictions

Teach with Explanations

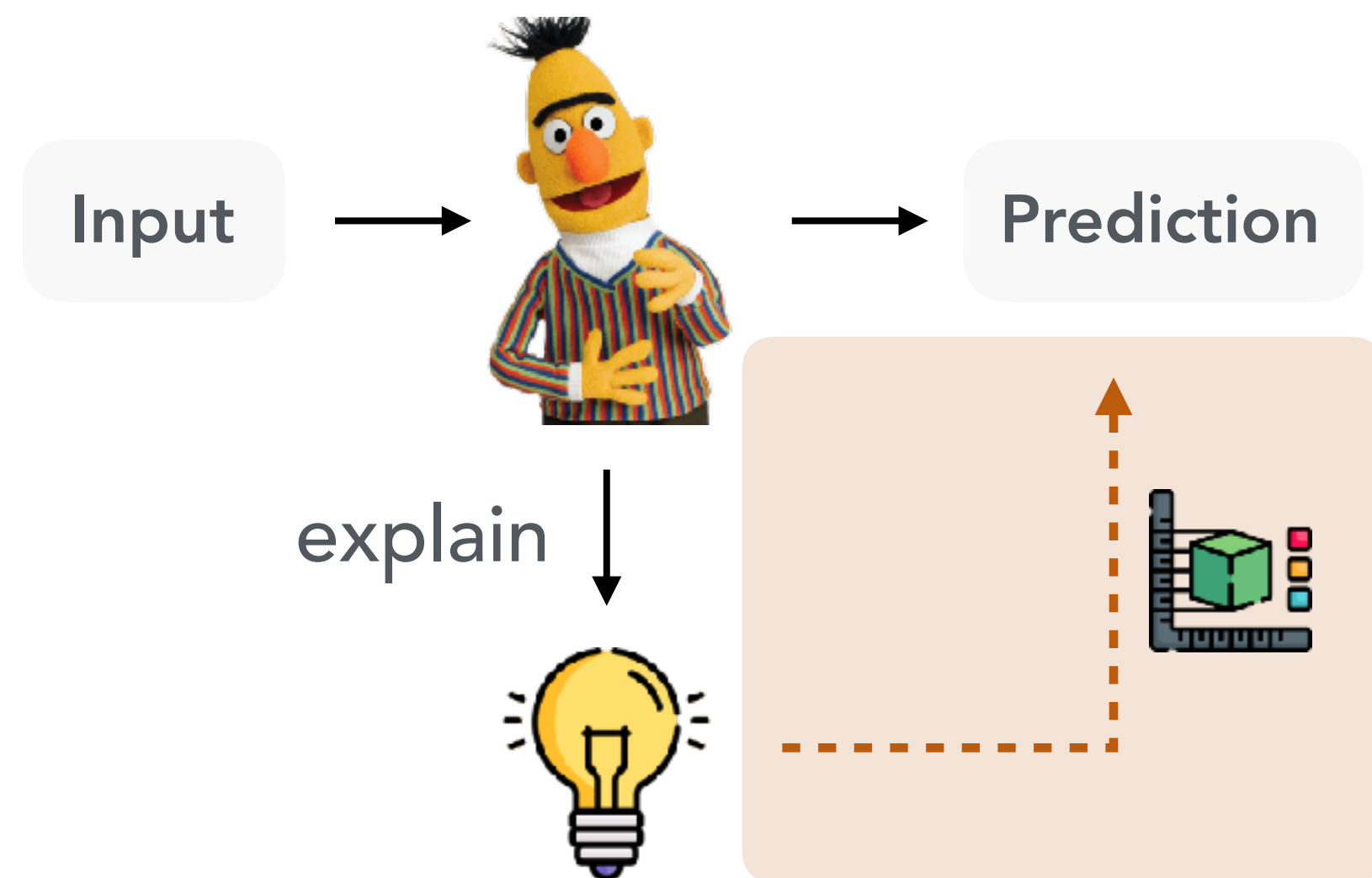


Use explanations to demonstrate how to reason

Steering Textual Reasoning with Explanations



Post-Hoc Intervene



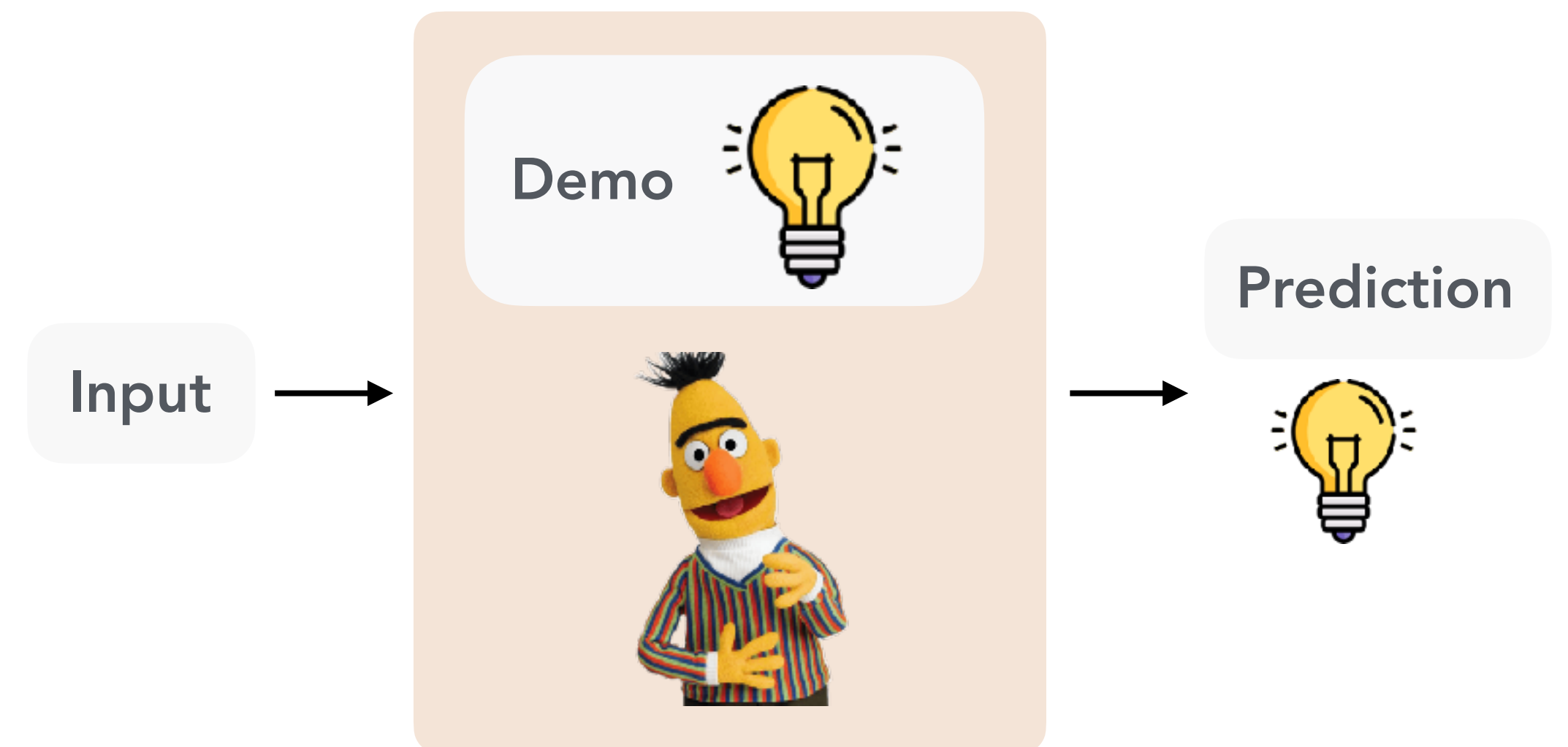
XY++ NeurIPS 22

XY++ ACL 22

XY++ EMNLP 21

PS*, JF*, **XY++** EACL 23

Teach with Explanations



XY++ NeurIPS 23

XY++ EMNLP 23

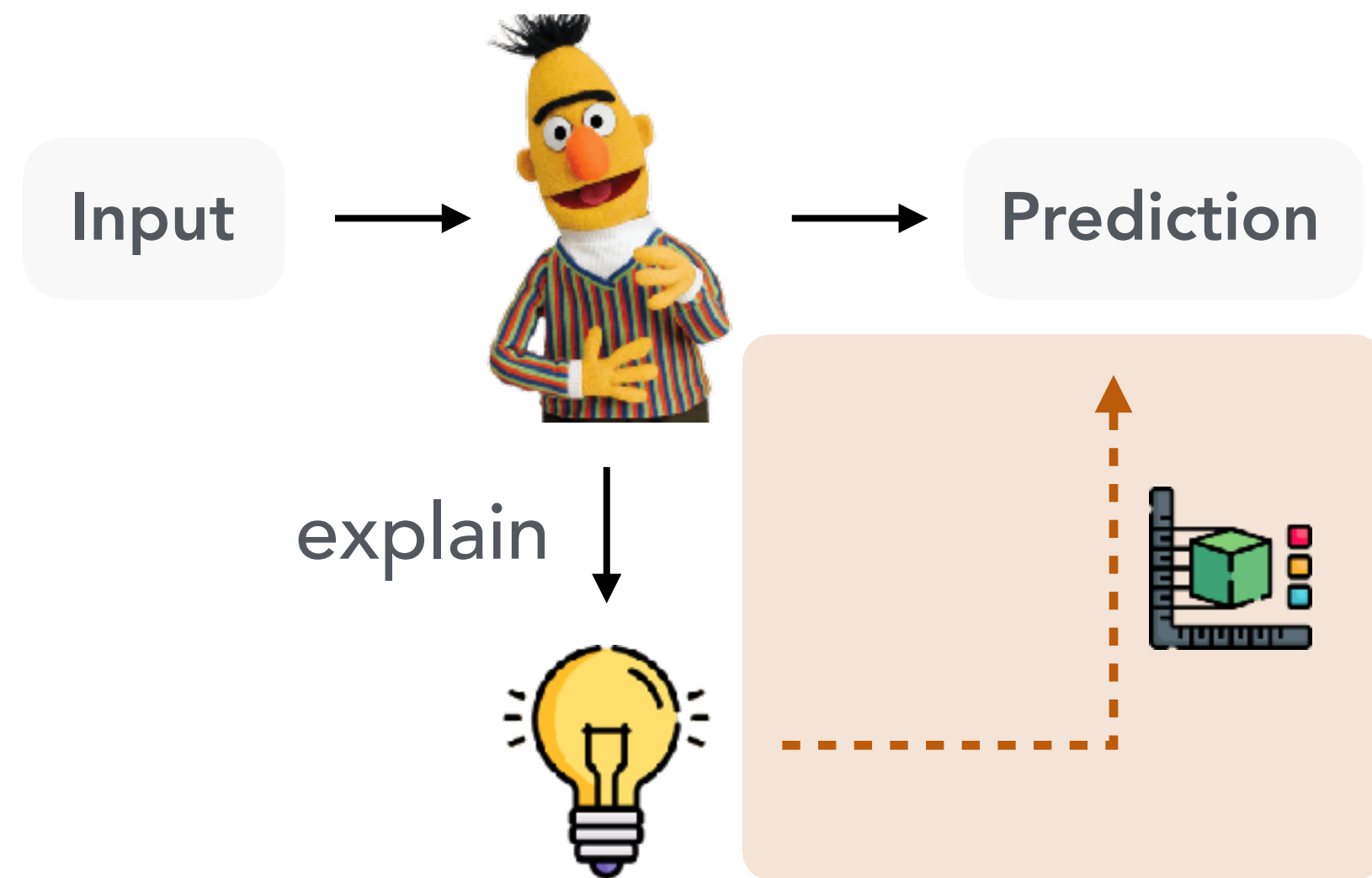
XY++ ACL Findings 23

ZS, **XY++** Arxiv 23 (in sub.)

Steering Textual Reasoning with Explanations



Post-Hoc Intervene



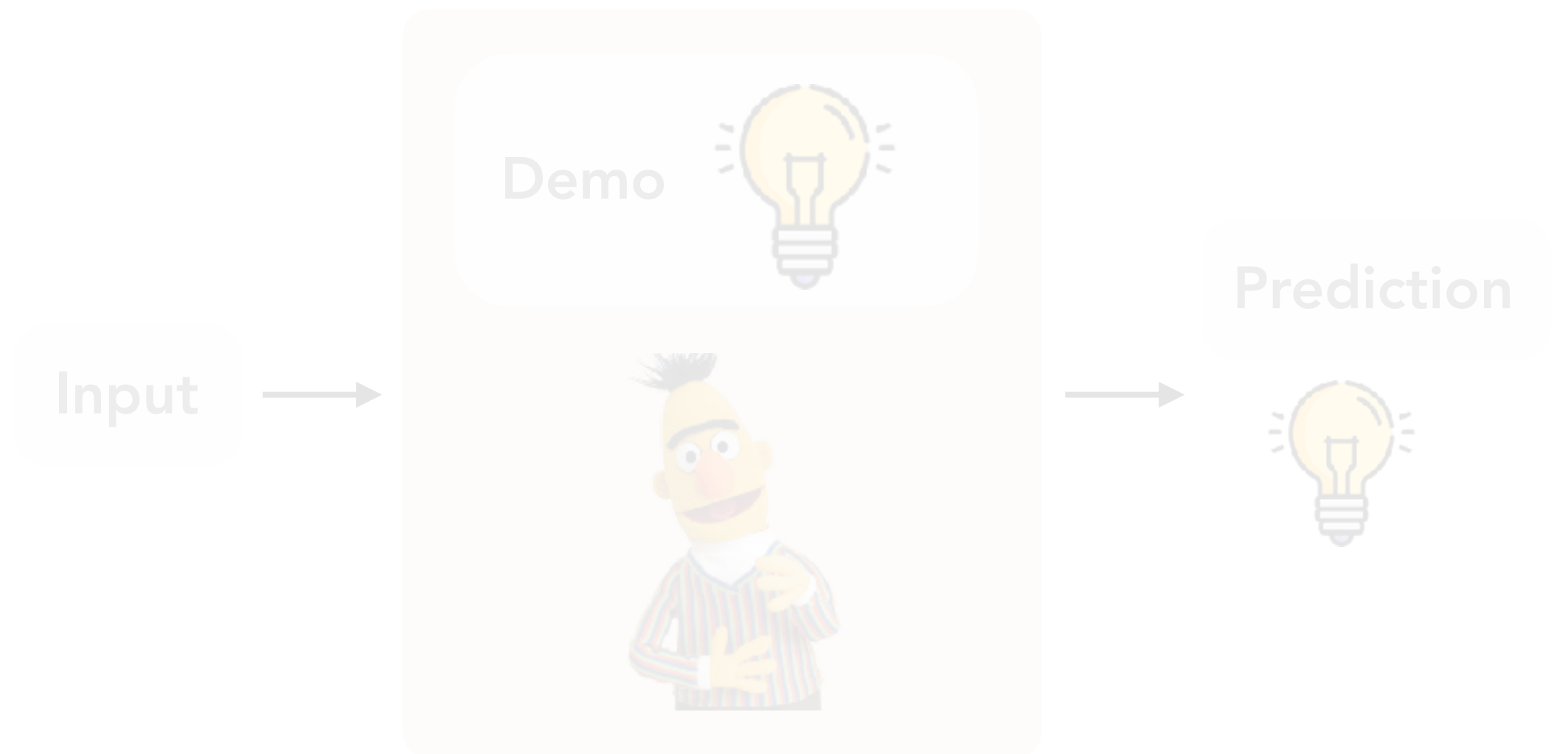
XY++ NeurIPS 22

XY++ ACL 22

XY++ EMNLP 21

PS*, JF*, **XY++** EACL 23

Teach with Explanations



XY++ NeurIPS 23

XY++ EMNLP 23

XY++ ACL Findings 23

ZS, **XY++** Arxiv 23 (in sub.)

Using Explanations Post-Hoc



More and more models are deployed black-box API

Performance degradation if a black-box model is tested under domain-shift

Avoid making errors by selective prediction (El-Yaniv and Wiener, 2010; Kamath et al. 2020)

Adversarial Example

Question

Where did the Panthers train ?

Context

The Panthers practice at the San Jose Stadium.

The Vikings train at Stark Industries.

Black-Box
QA Model



Prediction
Stark Industries
Confidence
0.97

intervene on
confidence

Abstain
Tuned Conf
0.35

Hard to calibrate black-box models due to limited information

Use explanations to know more about predictions



Using Explanations Post-Hoc



Use a calibrator to assess the correctness of predictions by looking at the reasoning process in explanations



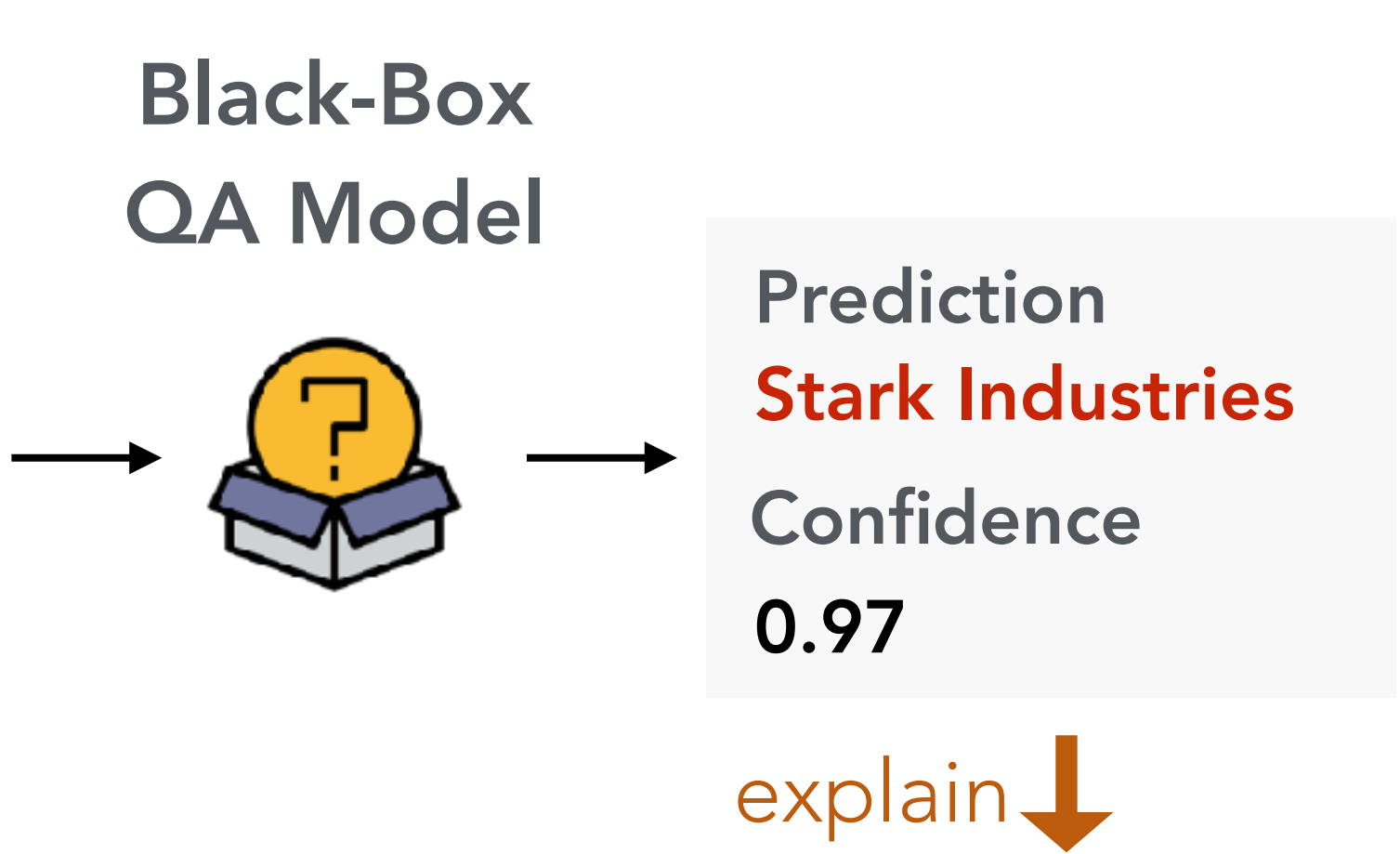
Adversarial Example

Question

Where did the Panthers train ?

Context

The Panthers practice at the San Jose Stadium.
The Vikings train at Stark Industries.

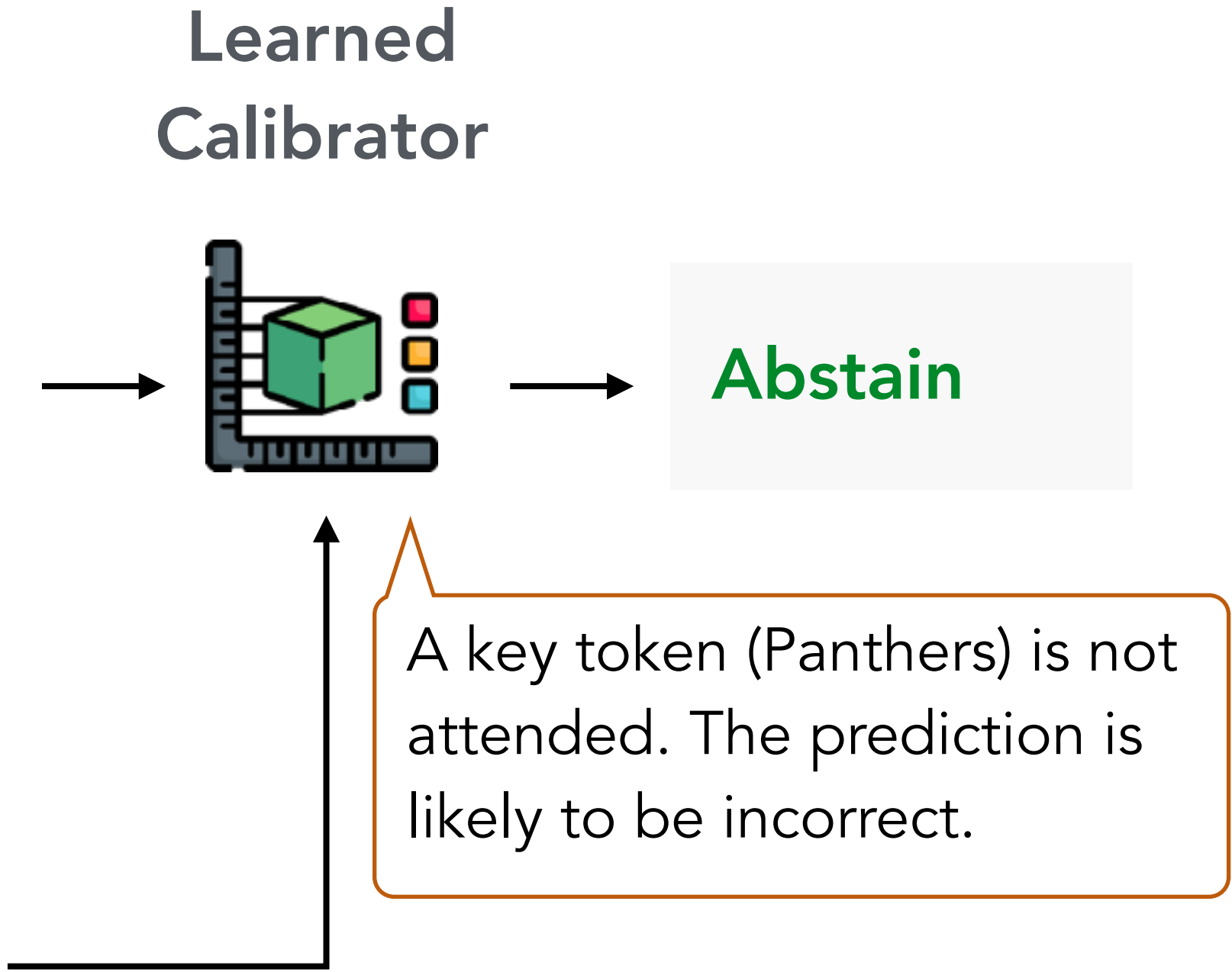


Question

Where did the Panthers practice ?

Context

The Panthers practice at the San Jose Stadium.
The Vikings practice at Stark Industries.



Calibration using Explanations



Example & Explanation

Question

Where did the Panthers practice ?

Context

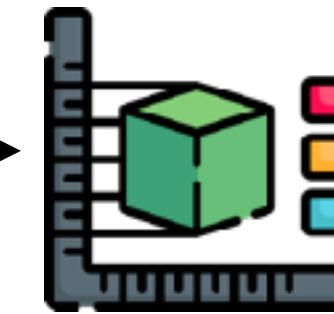
The Panthers practice at the San Jose Stadium.

The Vikings practice at Stark Industries.

Prediction

Stark Industries

Learned
Calibrator



correct / incorrect

How to let the calibrator learn
the patterns of reasoning?

Calibration using Explanations



Example & Explanation

Question

Where did the Panthers practice ?

Context

The Panthers practice at the San Jose Stadium.
The Vikings practice at Stark Industries.

Prediction

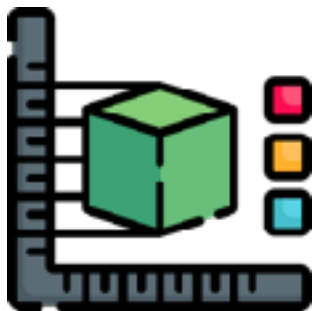
Stark Industries

Features of Reasoning Pattern

NNP (proper nouns)
not used by the model

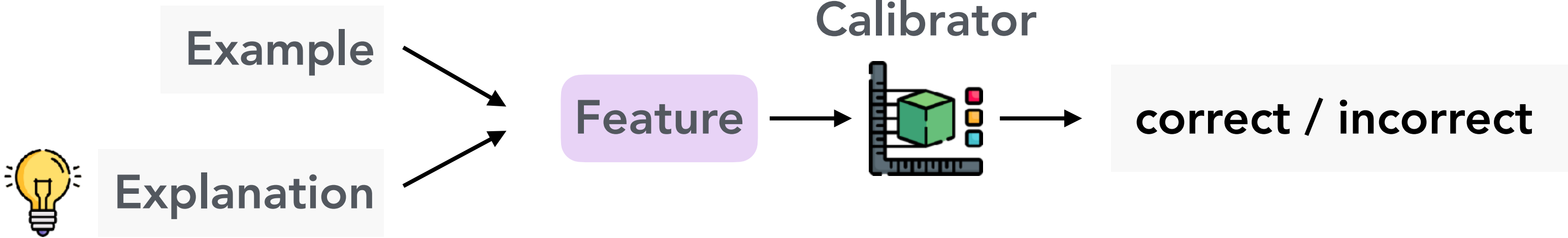
.....

Learned Calibrator

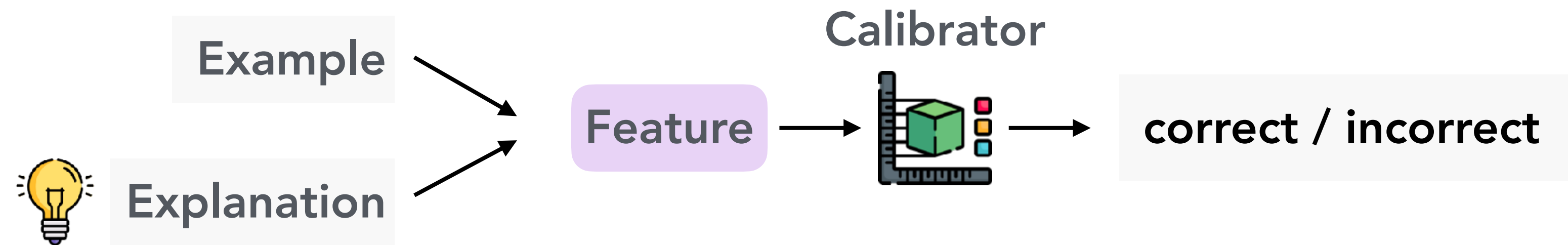


correct / incorrect

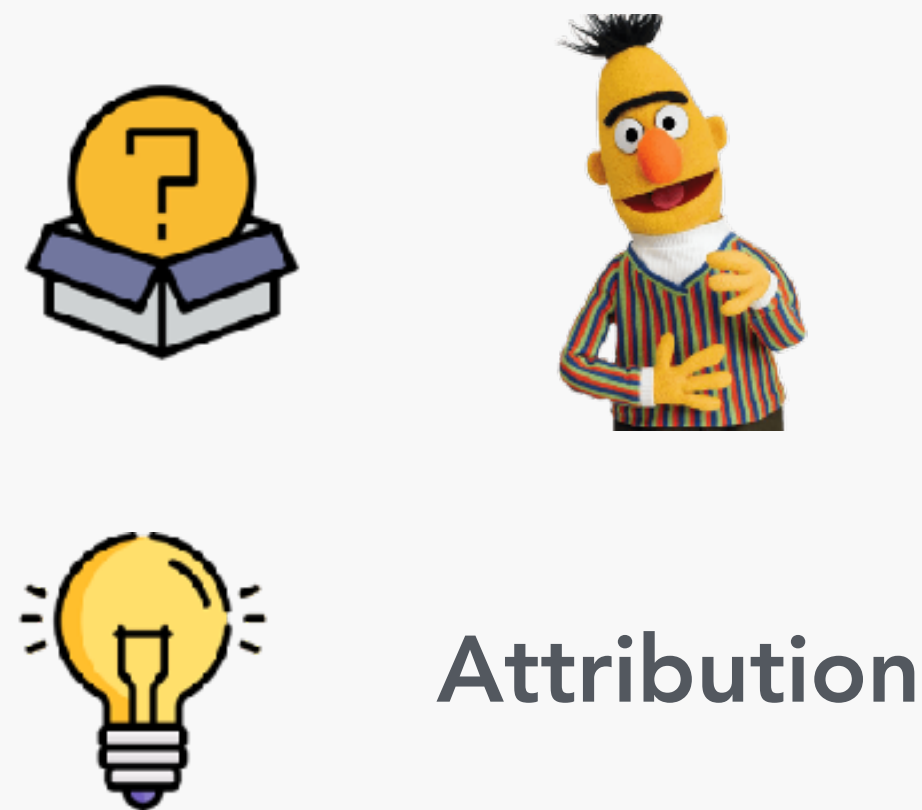
Our Framework



Calibration Framework



We use such a framework on two settings

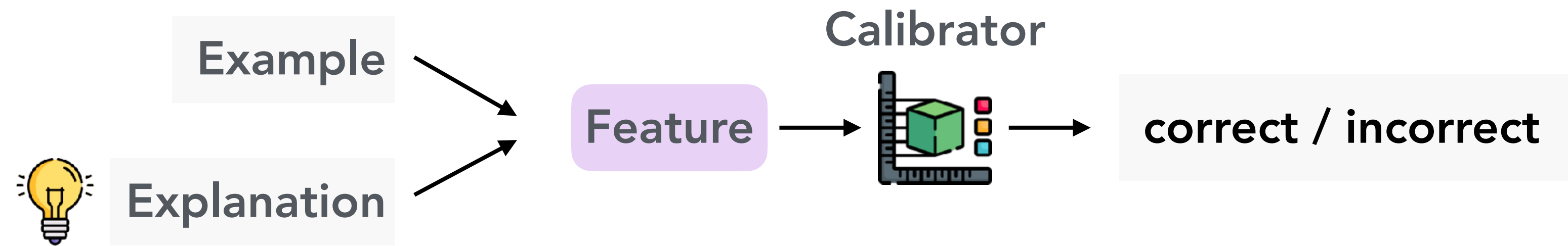


XY & GD, ACL 22



XY & GD NeurIPS 22

Calibration Framework

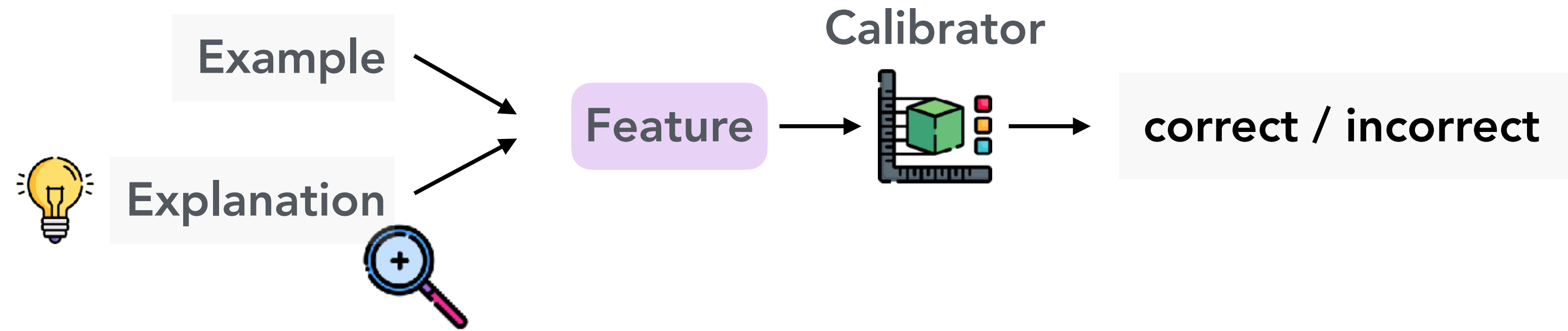


We use such a framework on two settings

This block represents the 'XY & GD, ACL 22' setting. It features a lightbulb icon with a question mark inside a box, a cartoon character (Bert from Sesame Street), and a hand pointing towards the character. Below these elements is the word 'Attribution' and the text 'XY & GD, ACL 22'.

This block represents the 'XY & GD NeurIPS 22' setting. It features a lightbulb icon with a question mark inside a box, a circular logo with a knot-like design, and the text 'Free-text'. Below these elements is the text 'XY & GD NeurIPS 22'.

Calibrating BERT-based Models

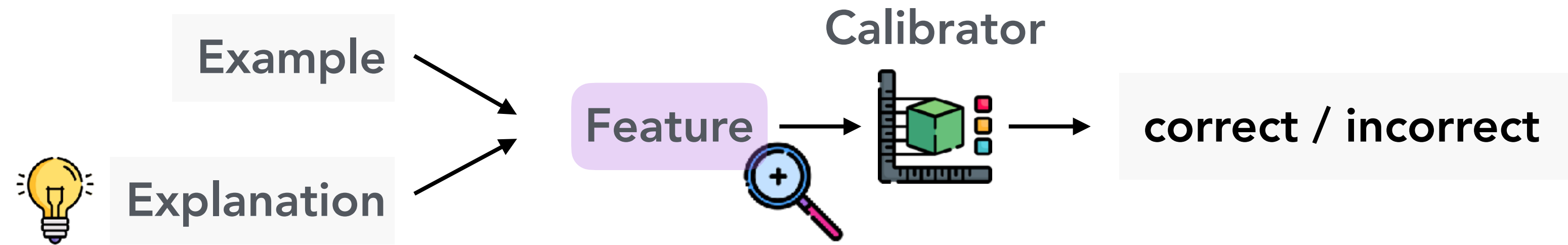


Use black-box explanation techniques, Lime and SHAP, to generate attributions

Assign an attribution score to each input token



Calibrating BERT-based Models



Numeric features describing the importance of certain parts of input or certain linguistic features (extracted automatically using a syntactic parser)

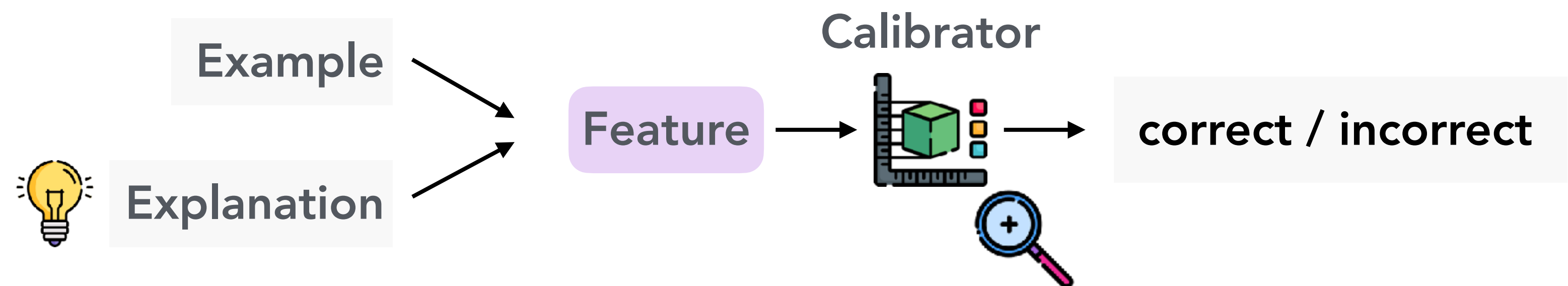
Importance of NNP: 0.10

Importance of Question: 0.27

.....



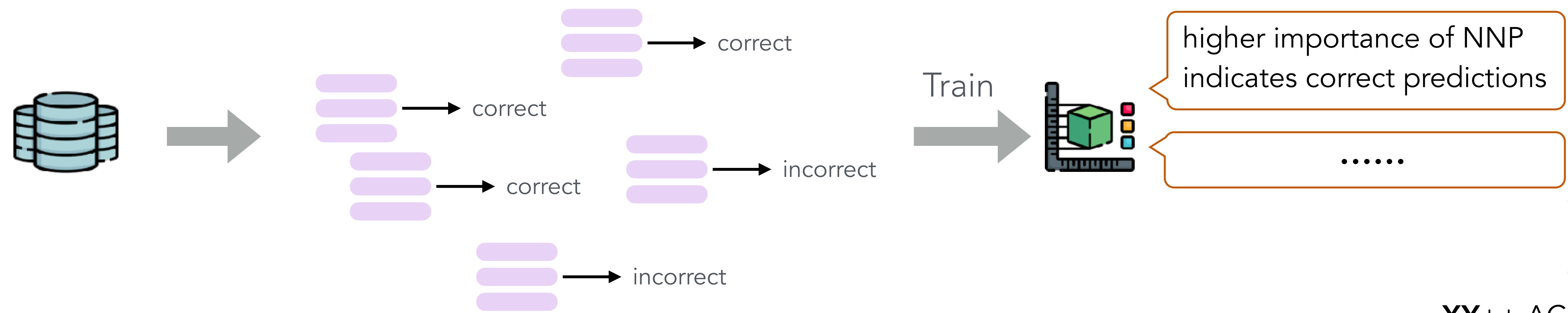
Calibrating BERT-based Models



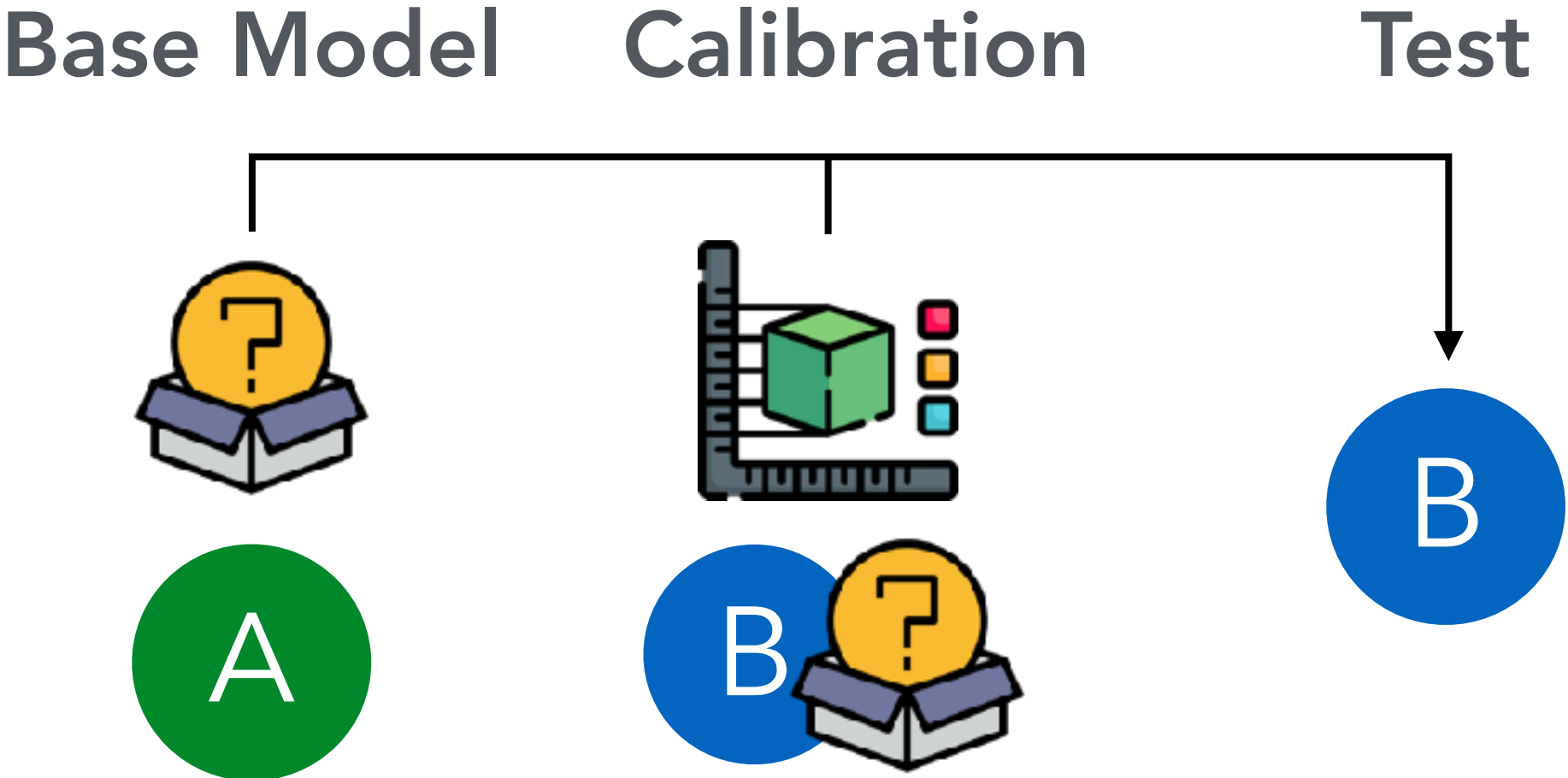
Train a calibrator using of feature-correctness pairs extracted from a small development set

Small Dev Set

Data For Training Calibrator



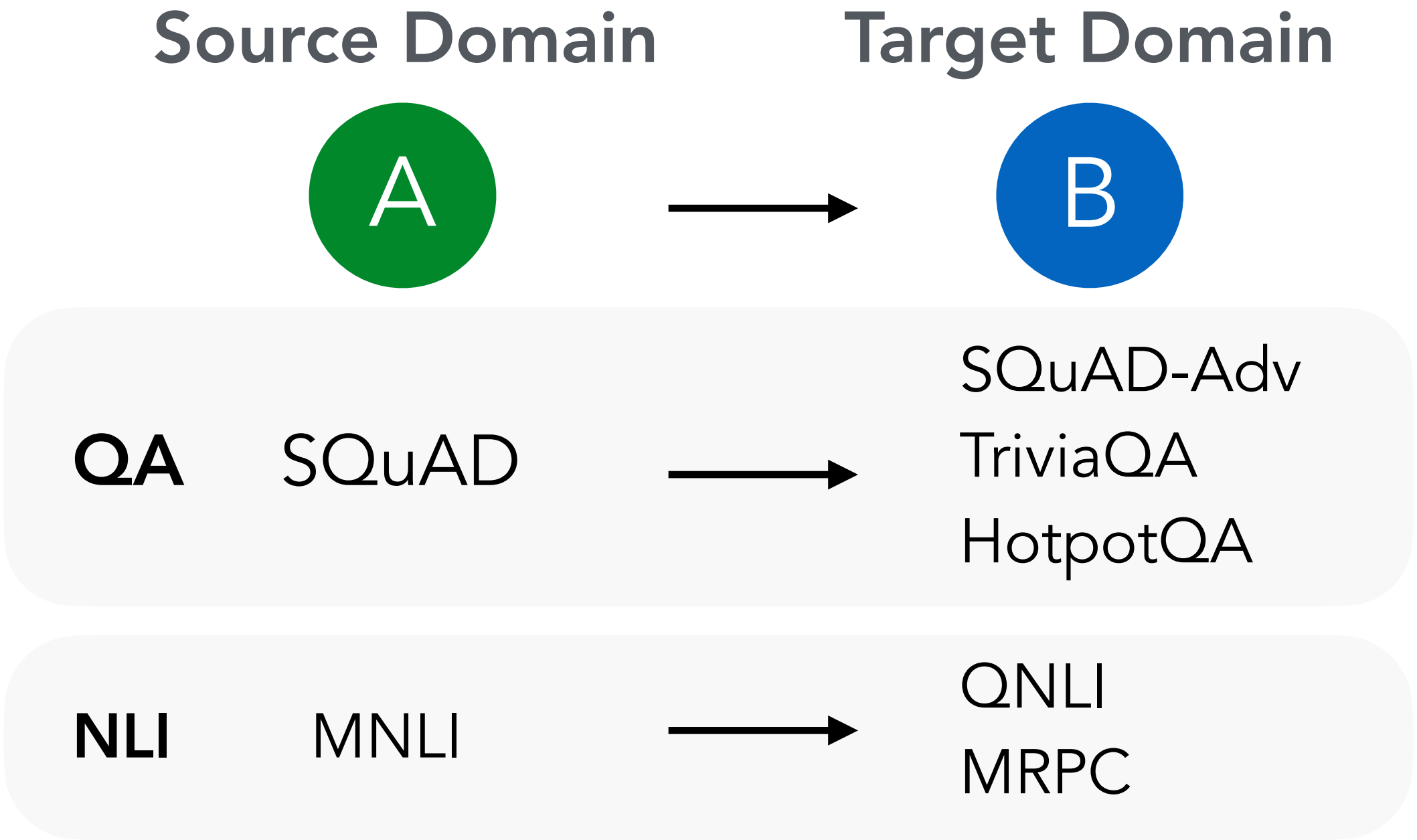
Experiments: Setup



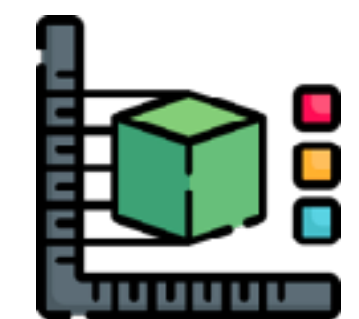
Base Model



RoBERTa



Calibrator



RandomForest trained
using 500 data points



Evaluating Selective Prediction



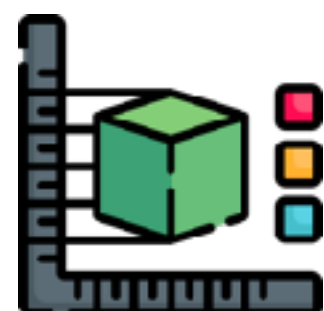
Model Performance

Score: 0.9

Score: 0.5

Score: 0.1

Score: 0.3



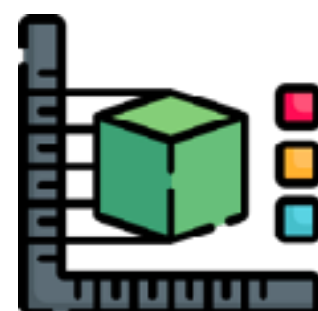
Calibrator A

Conf: 1.0

Conf: 0.6

Conf: 0.8

Conf: 0.3



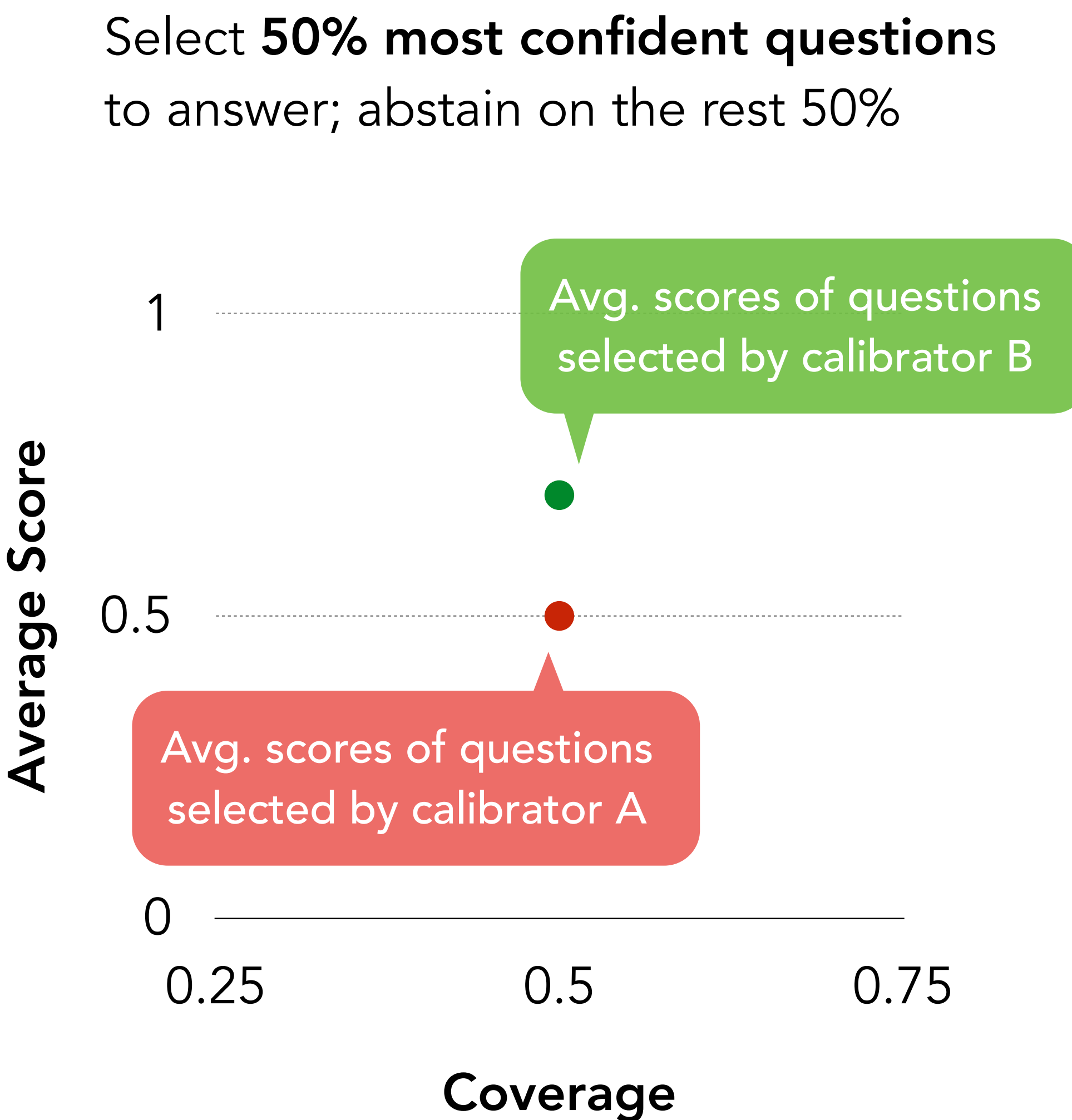
Calibrator B

Conf: 1.0

Conf: 0.6

Conf: 0.2

Conf: 0.3



Evaluating Selective Prediction



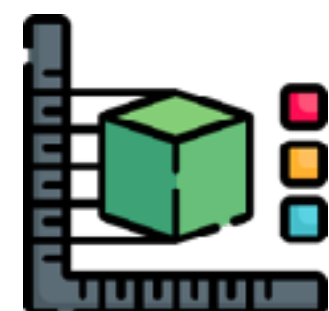
Model Performance

Score: 0.9

Score: 0.5

Score: 0.1

Score: 0.3



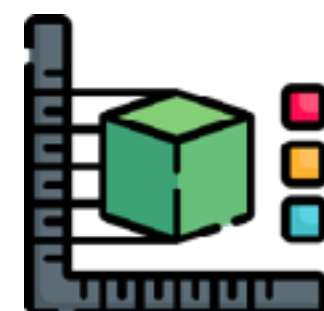
Calibrator A

Conf: 1.0

Conf: 0.6

Conf: 0.8

Conf: 0.3



Calibrator B

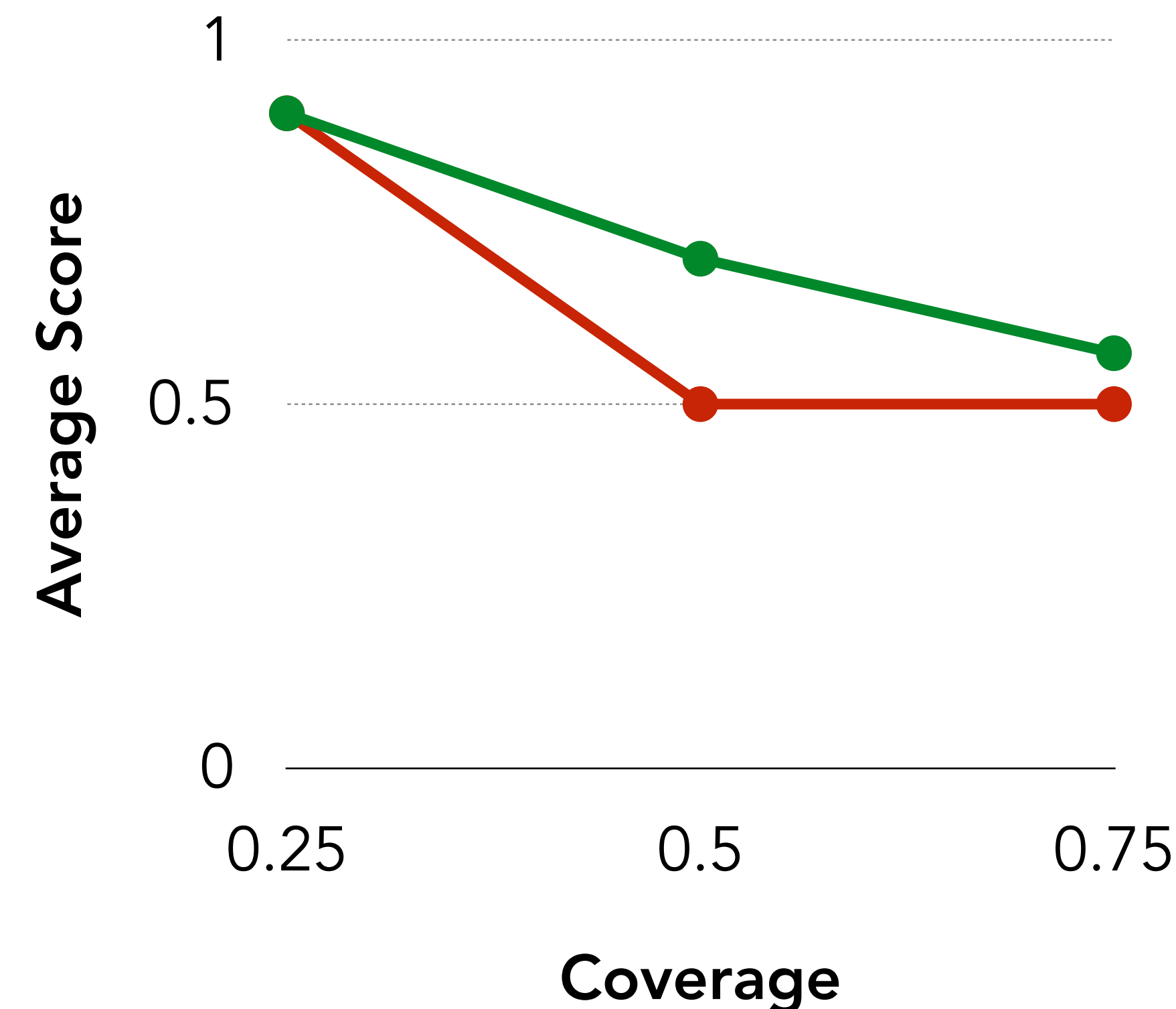
Conf: 1.0

Conf: 0.6

Conf: 0.2

Conf: 0.3

Evaluating calibration using area under coverage-score curve

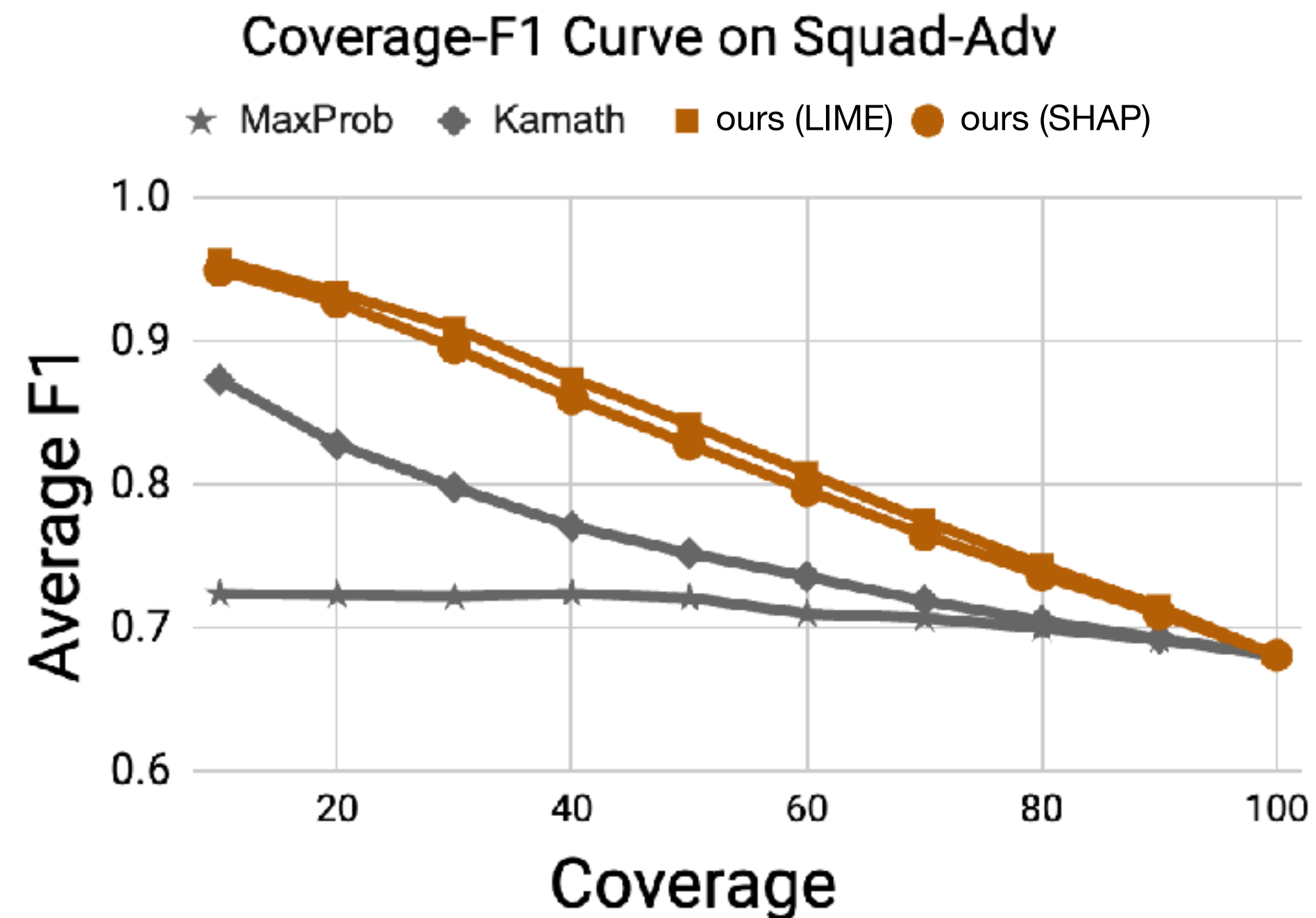


Experiments: Setup



Metrics

Area under Coverage-F1Score Curve (AUC)



Baselines

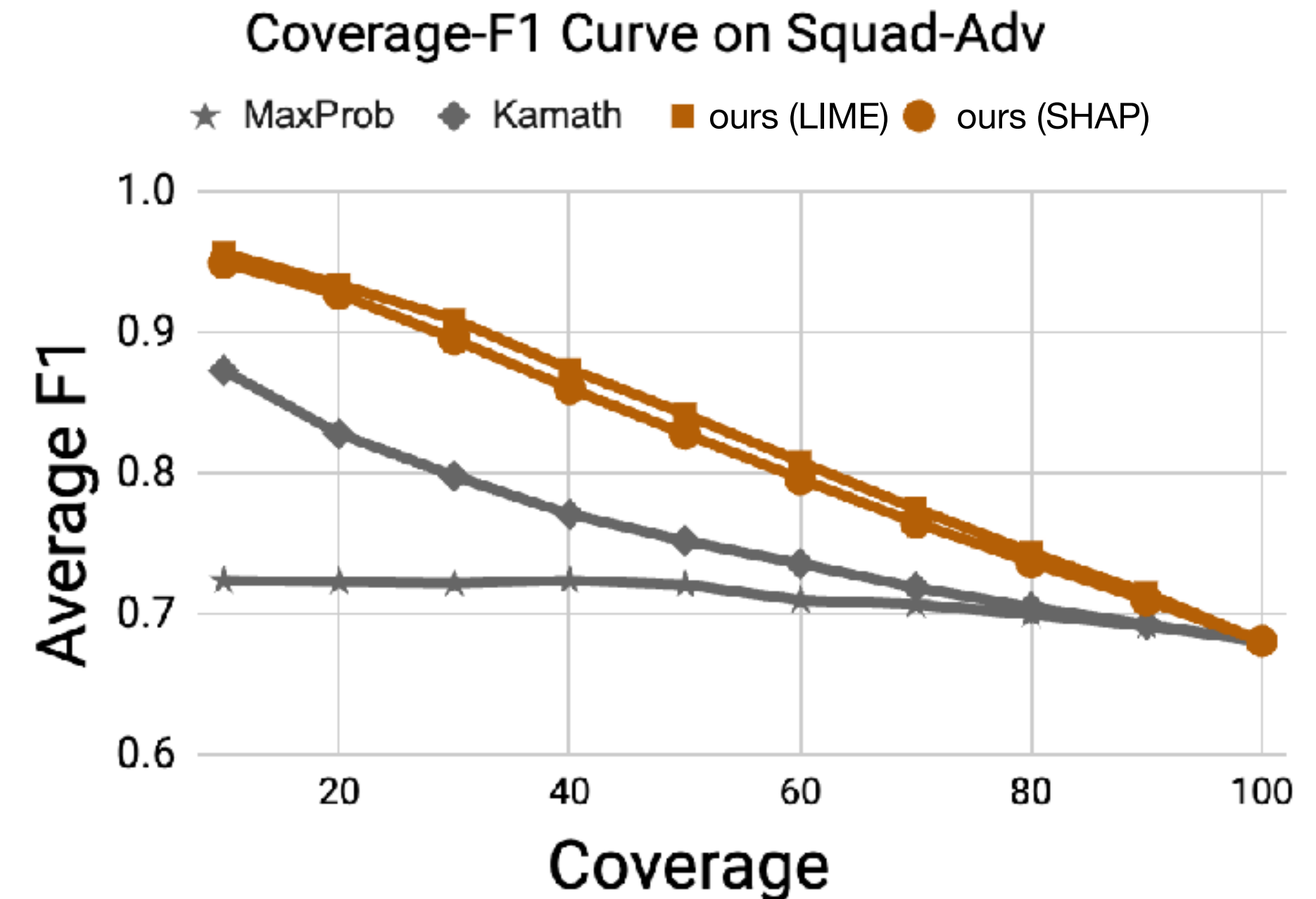
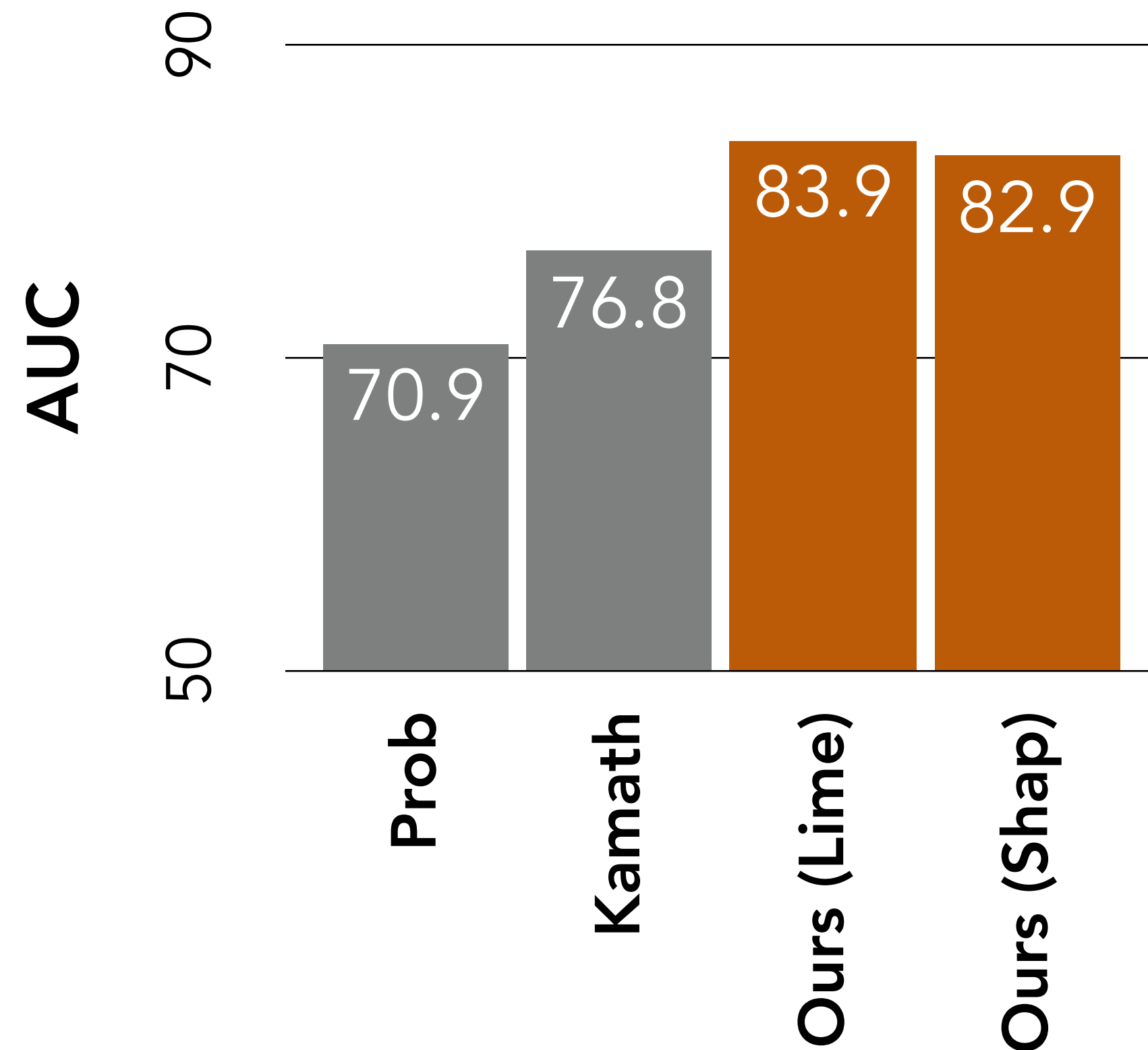
Prob: confidence of prediction

Kamath: (Kamath et al. 2020) calibrator using heuristic features (probabilities, length of context, length of answer)

Ours (LIME) & Ours (SHAP): calibrators using **explanation**-based features



Results



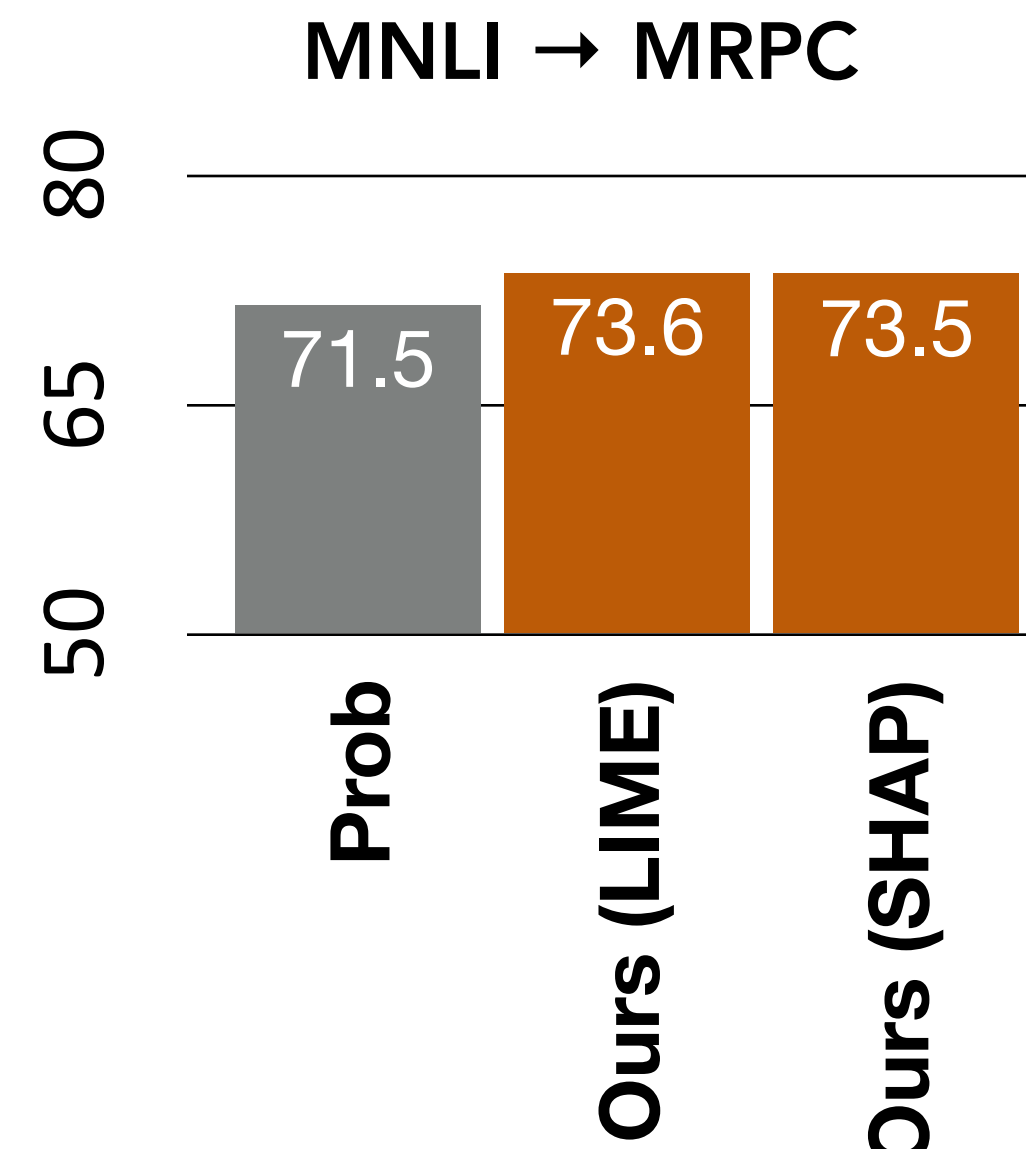
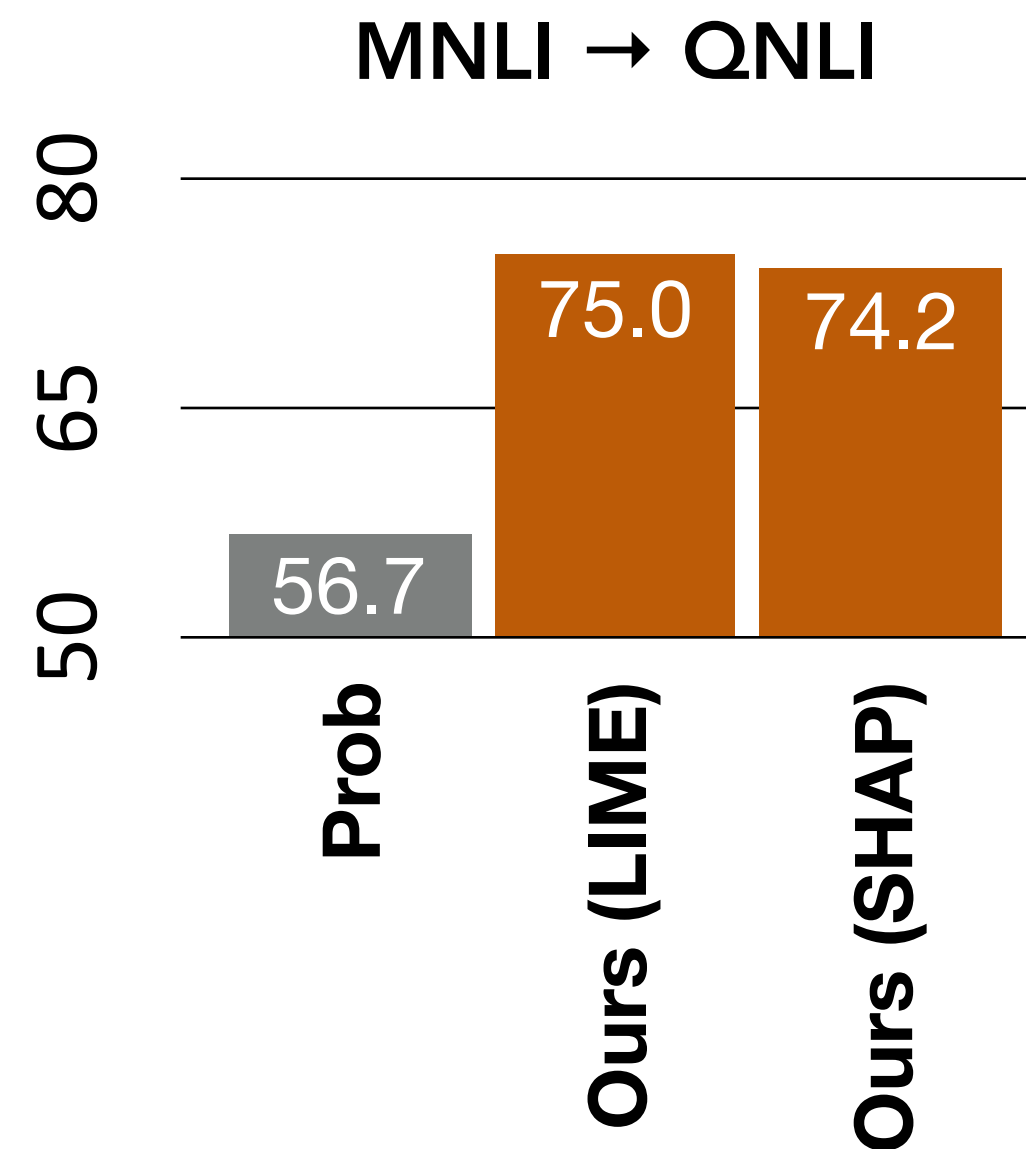
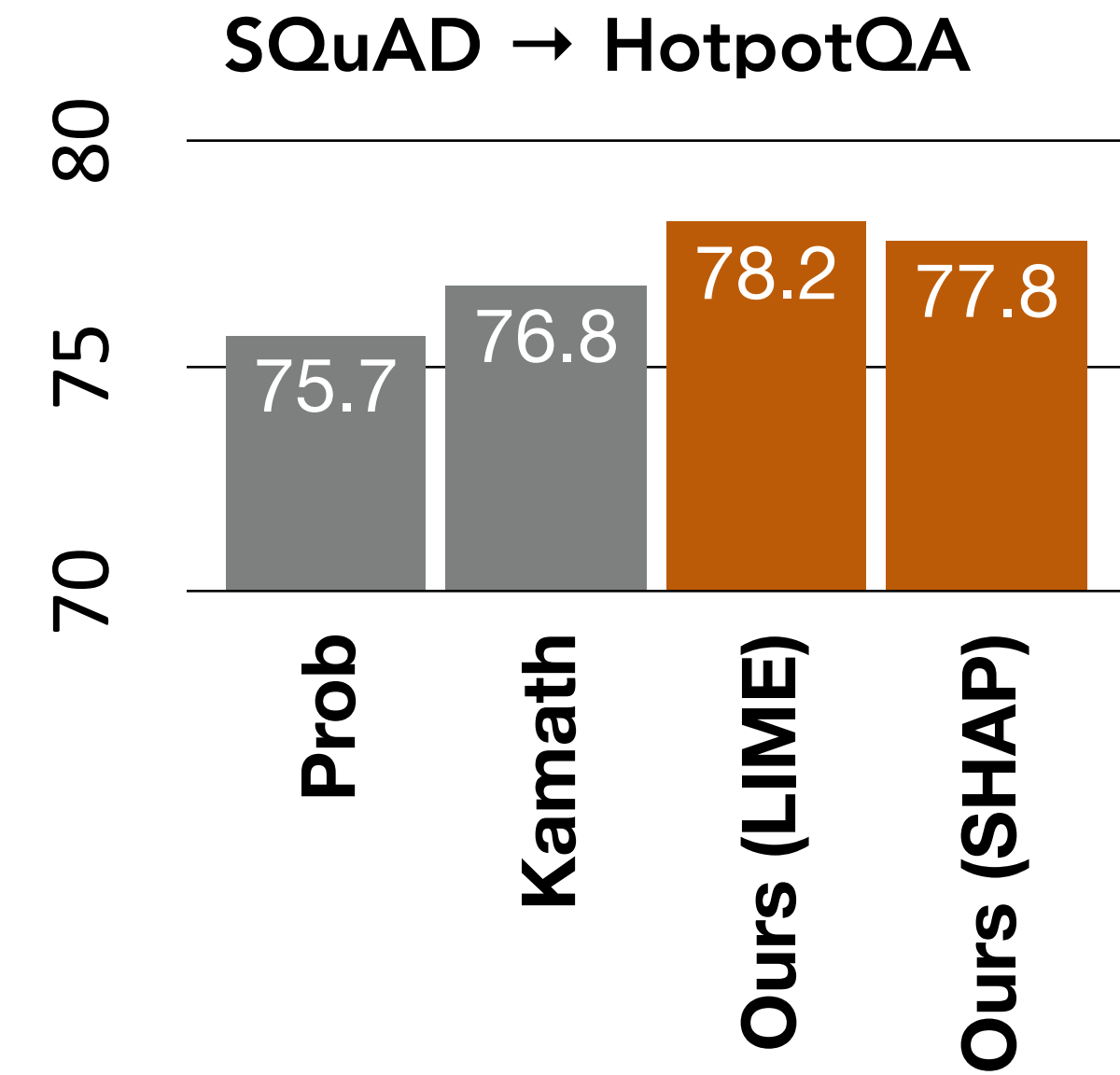
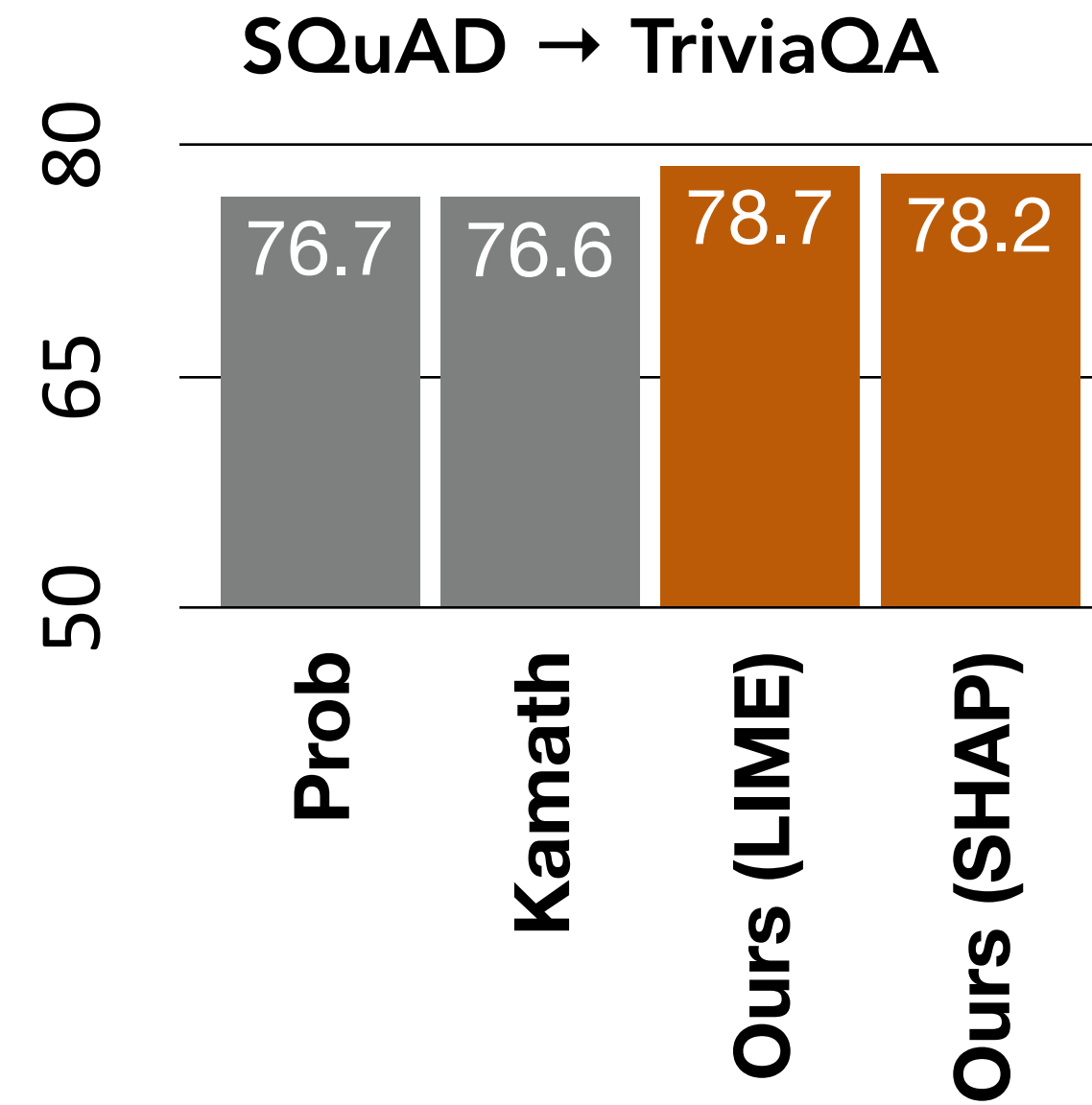
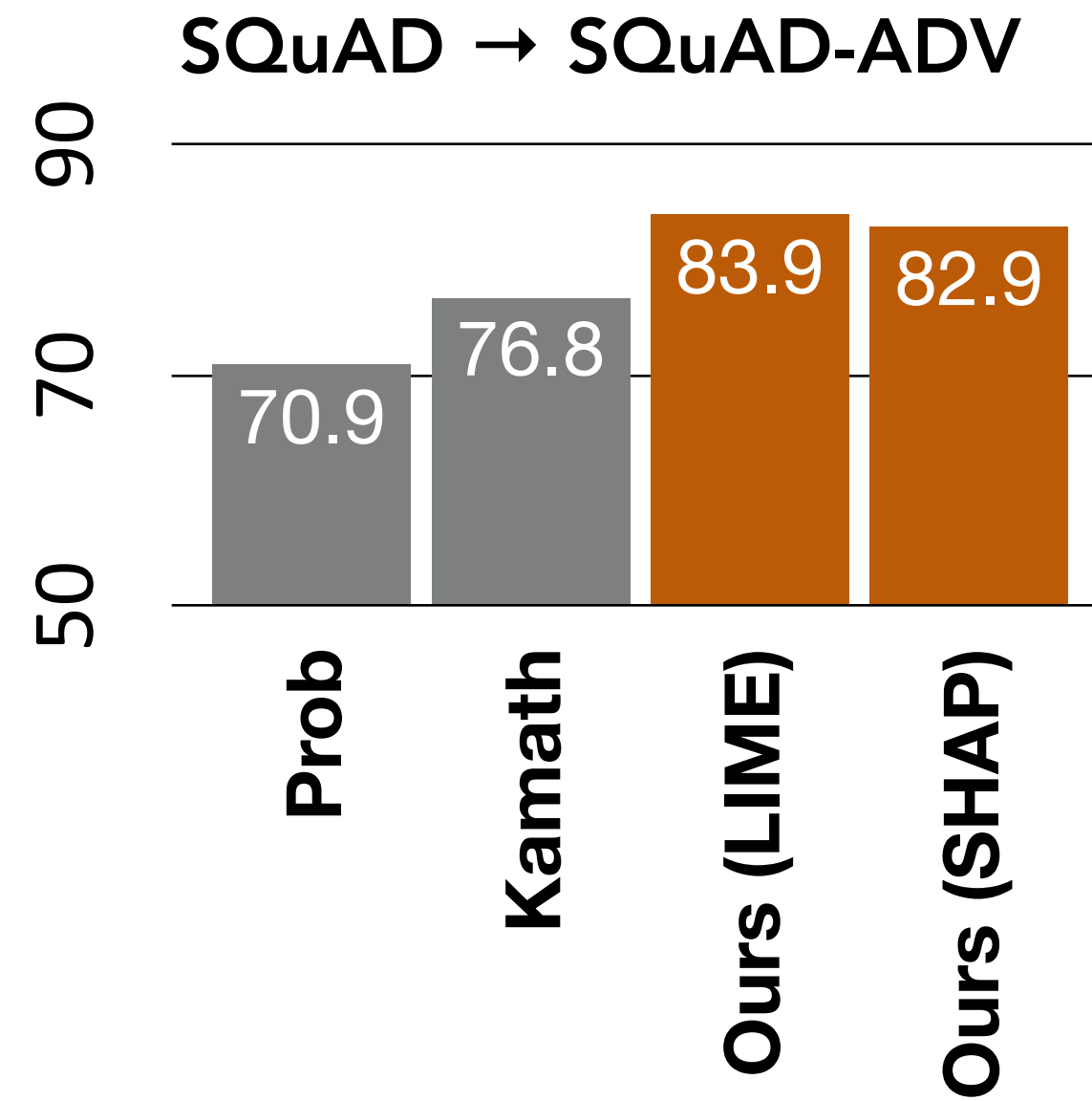
Ours (Lime) achieves the best performance

Explanations are helpful; **Ours** outperform calibrators without using explanations

Substantial performance difference when selectively answering a part of the questions that the calibrator is most confident with



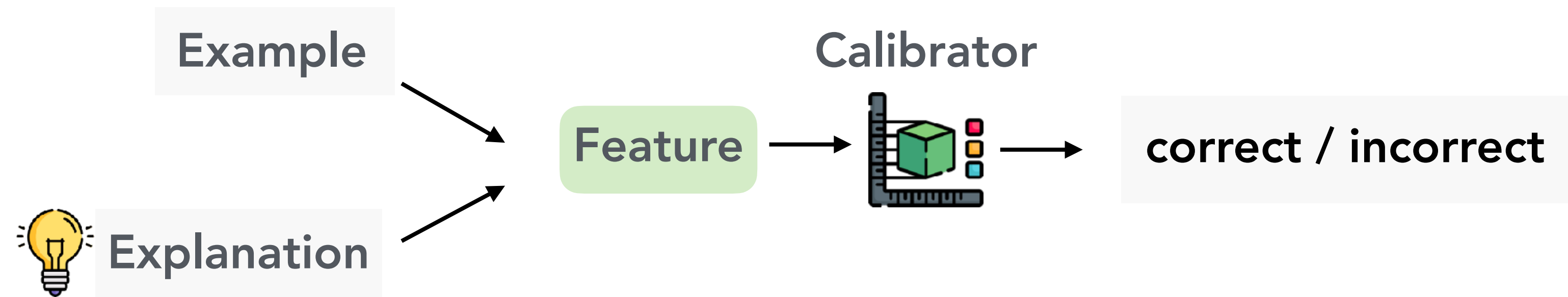
Results



Explanations improves the generalization performance across all pairs covering both QA and NLI tasks



Calibration Framework



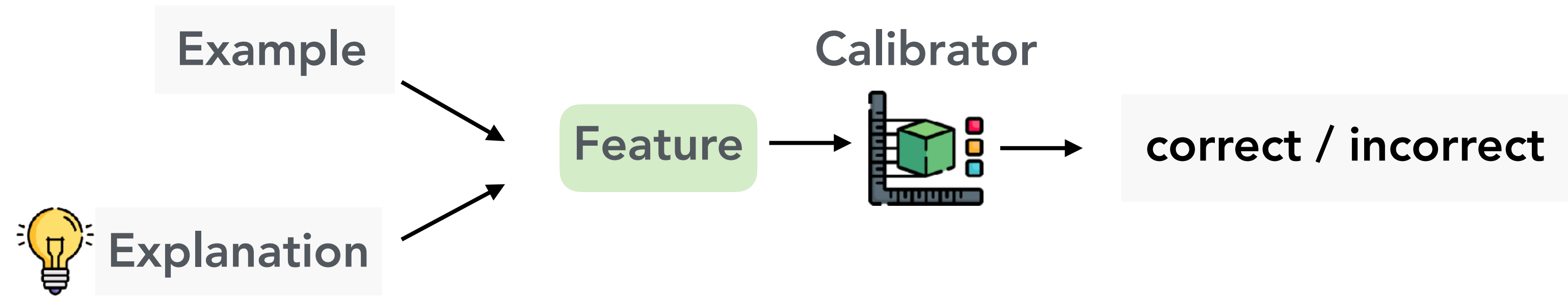
Attribution

XY & GD, ACL 22

Free-text

XY & GD NeurIPS 22

Calibration Framework



Prompting with Explanations



LLMs can learn a task from few-shot examples via in-context learning

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.

Prompting with Explanations



We can include **explanations** before answers (Nye et al., 2022, Wei et al., 2023) or after answers (Ye et al., 2023) in prompts

LLMs will generate explanations in addition to predictions

Prompt

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

Output

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

Controlled QA Task



We use a controlled QA task to demonstrate our study on using explanations for LLMs

Context: 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?.

A: Angela.

Explanation: Angela hangs out with Tiffany and Tiffany is a nurse.

Unreliability in LLM Explanations



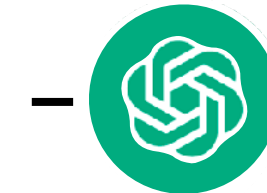
Unreliability in LLM Explanations



Few-Shot
Examples

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?



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

Stephanie is an engineer
in the context

LLMs can generate nonfactual explanations!

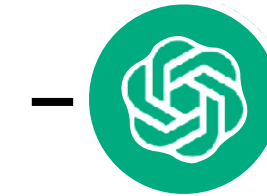
Unreliability in LLM Explanations



Few-Shot
Examples

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?



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

Stephanie is an engineer
in the context

LLMs can generate nonfactual explanations!

E.g., 72% explanations generated by Instruct-GPT (text-davinci-001) are nonfactual

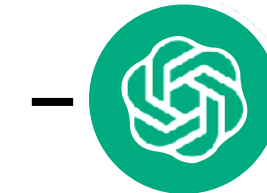
Unreliability in LLM Explanations



Few-Shot Examples

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?



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

Stephanie is an engineer in the context

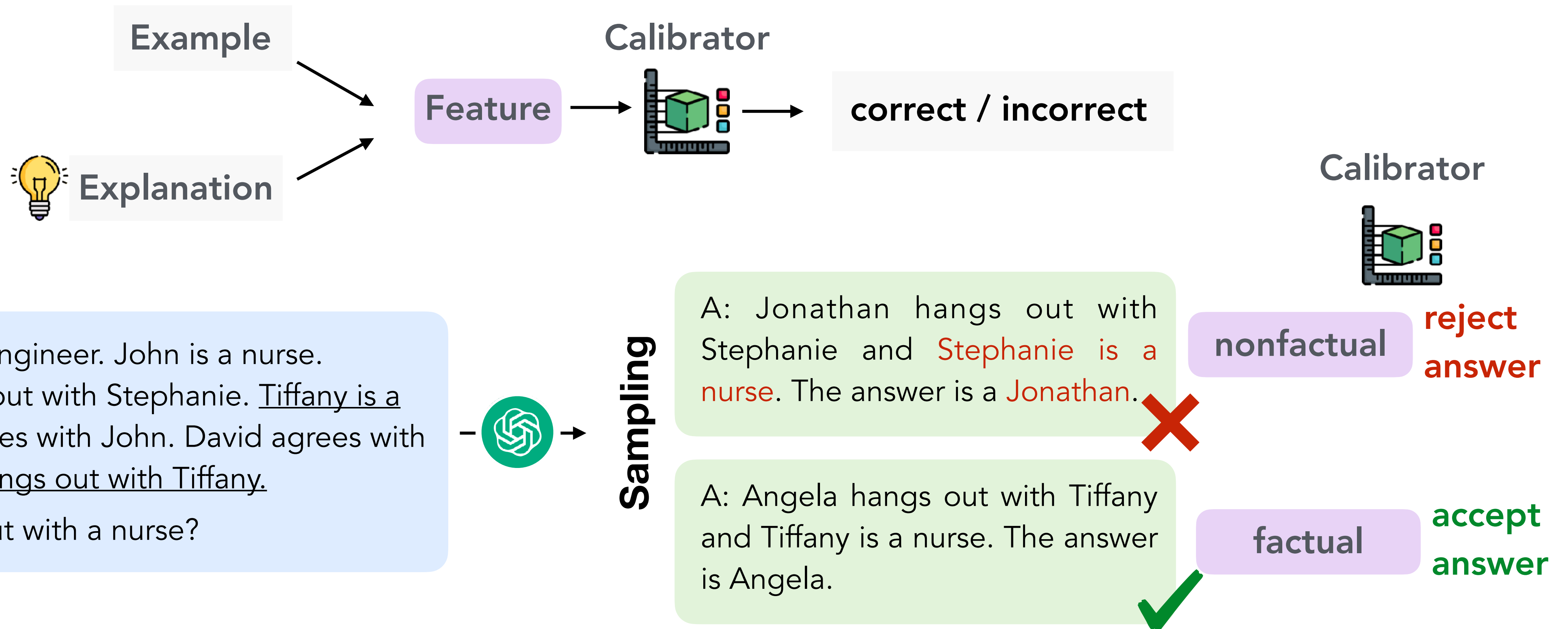
LLMs can generate nonfactual explanations!

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

We can use factuality of explanations to verify predictions



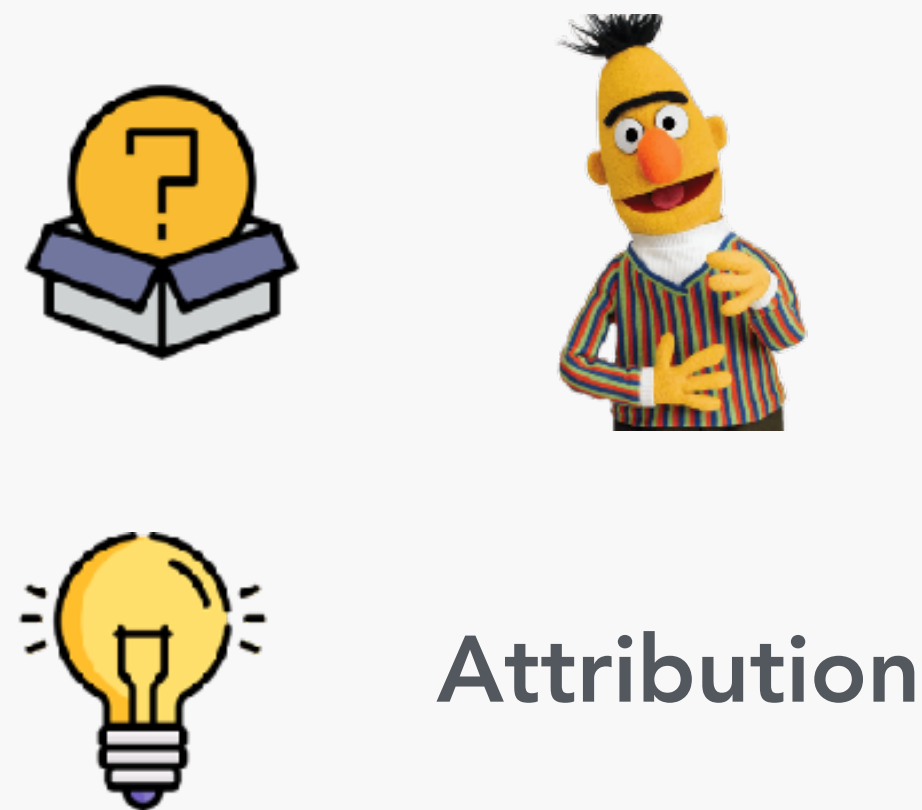
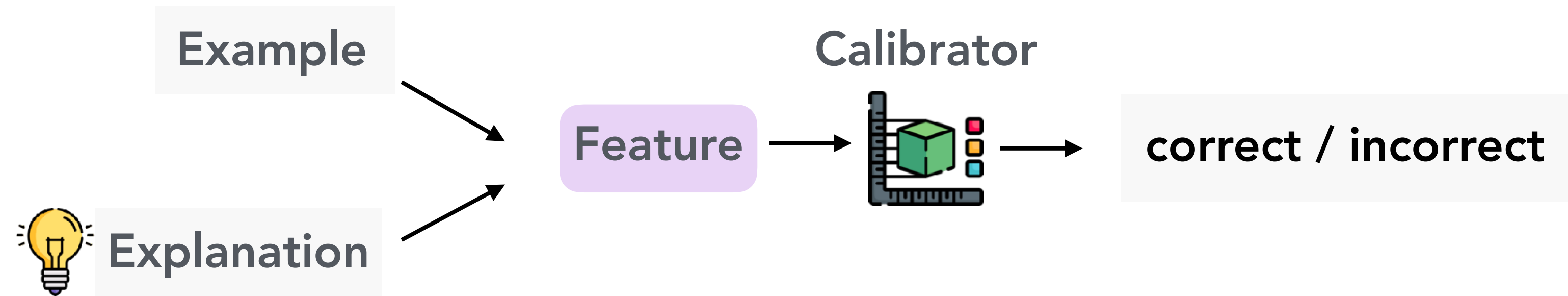
Calibrating Large Language Models



By using factuality of explanations to verify and reject answers, we improve the performance of InstructGPT from 54% to 78%

See paper for calibration experiments on realistic datasets

Calibration Framework



Attribution

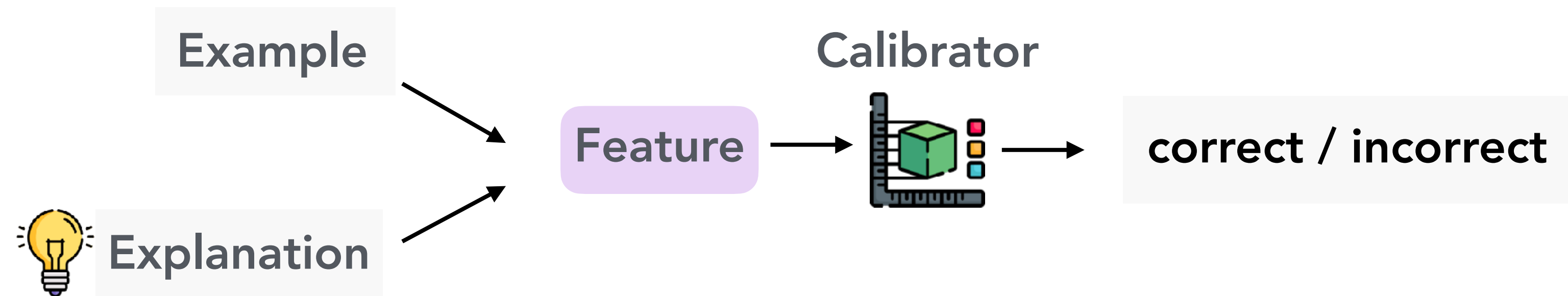
XY & GD, ACL 22



Free-text

XY & GD NeurIPS 22

Calibration Framework



Attribution

XY & GD, ACL 22

Explanations can be useful
for calibrating black-box
models' predictions



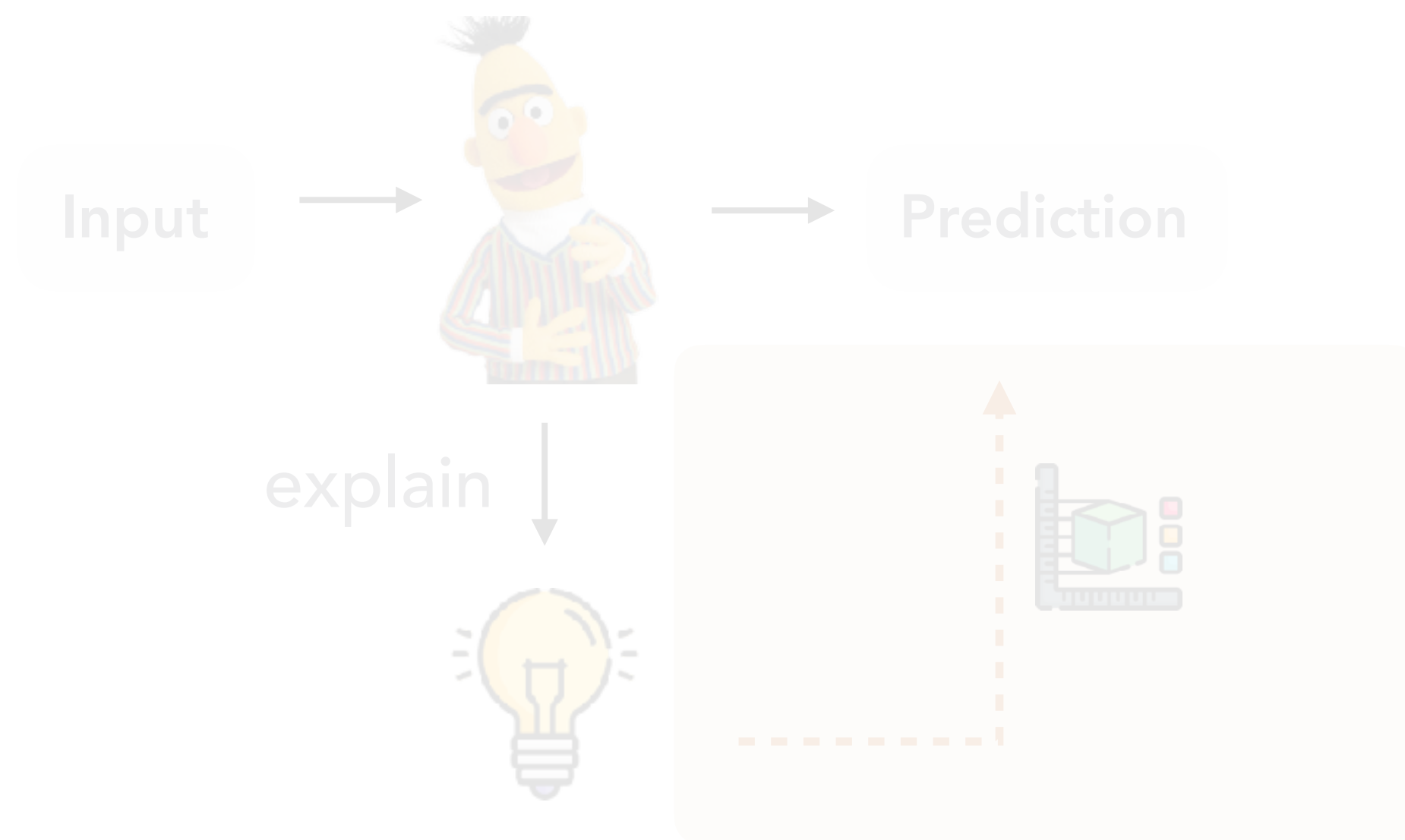
Free-text

XY & GD NeurIPS 22

Steering Textual Reasoning with Explanations

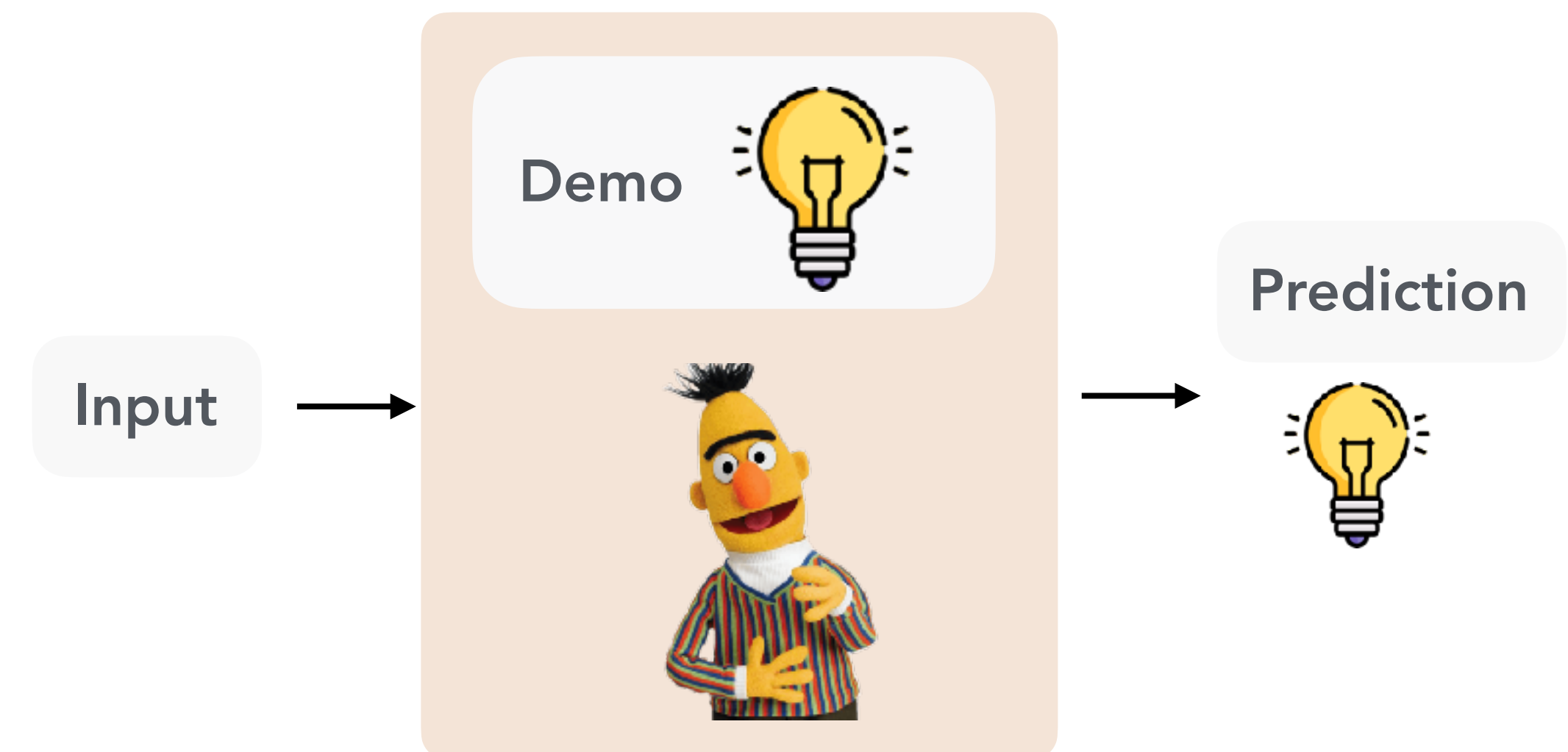


Post-Hoc Intervene



XY++ NeurIPS 22
XY++ ACL 22
XY++ EMNLP 21
PS*, JF*, XY++ EACL 23

Teach with Explanations



XY++ NeurIPS 23
XY++ EMNLP 23
XY++ ACL Findings 23
ZS, XY++ Arxiv 23 (in sub.)

Supervise LLMs with Explanations



Prompt

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-3.5

Output

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

We include explanations (in the form of input texts in prompts)

Text is versatile; there are many ways to formulate explanations

How to formalize more effective explanations?

Performance Varying Across Explanations



Prompt

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.



GPT-3.5

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.



GPT-3.5

Performance
57%

Good explanations need engineering

we optimize explanations for better performance

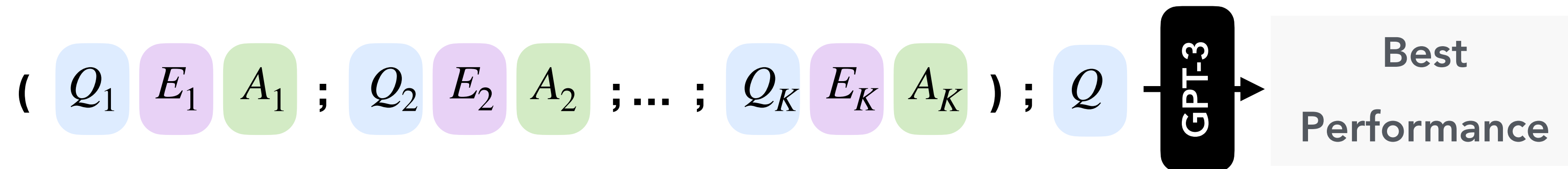
Optimizing Explanations



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)



Data Condition



Given

Few-Shot
Exemplars

$$Q_1 \ A_1 ; Q_2 \ A_2 ; \dots ; Q_K \ A_K$$

Seed
Explanations

$$\tilde{E}_1 \quad \tilde{E}_2 \quad \dots \quad \tilde{E}_K$$

Unlabeled
Dev set

$$V = Q_1 \ Q_2 \ \dots \ Q_M$$

Output

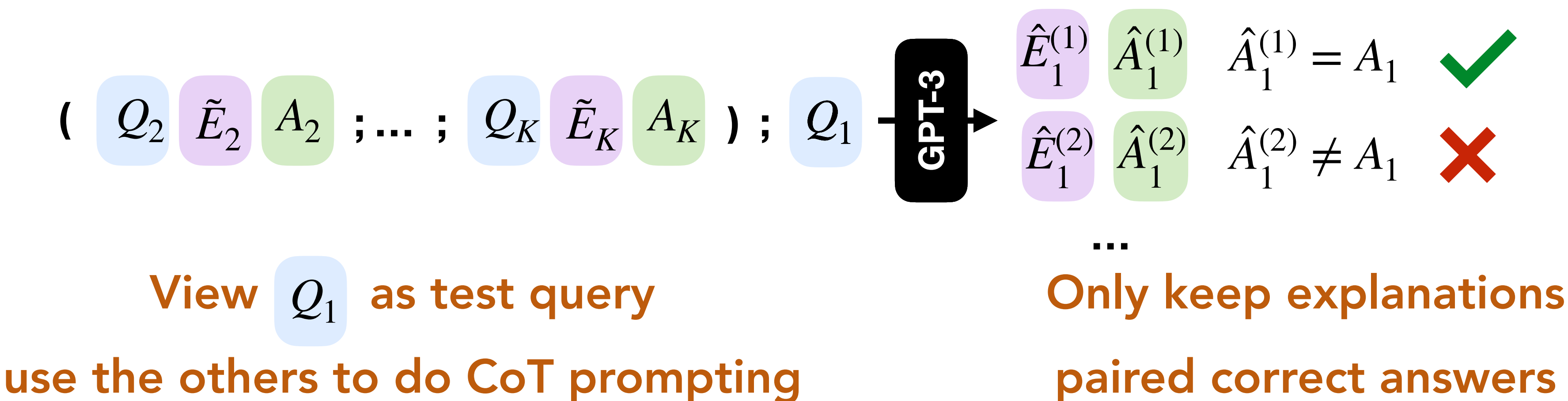
Optimized
Explanations

$$E_1 \ E_2 \ \dots \ E_K \text{ that yields better end task performance}$$

Approach Overview



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



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

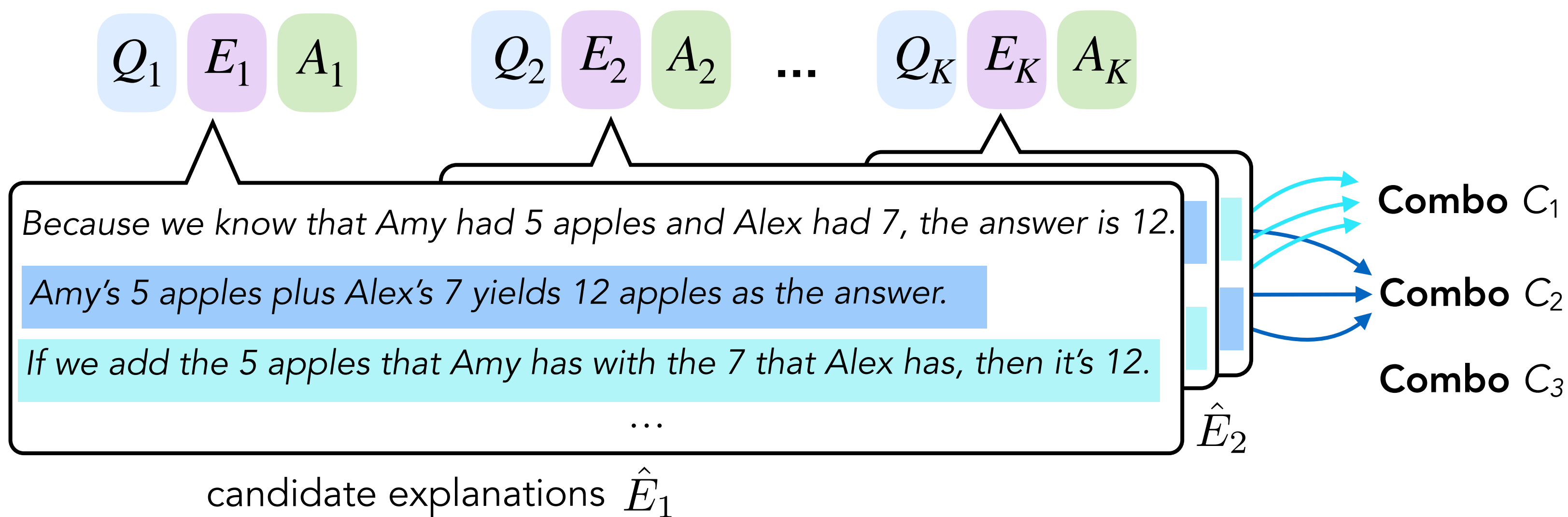
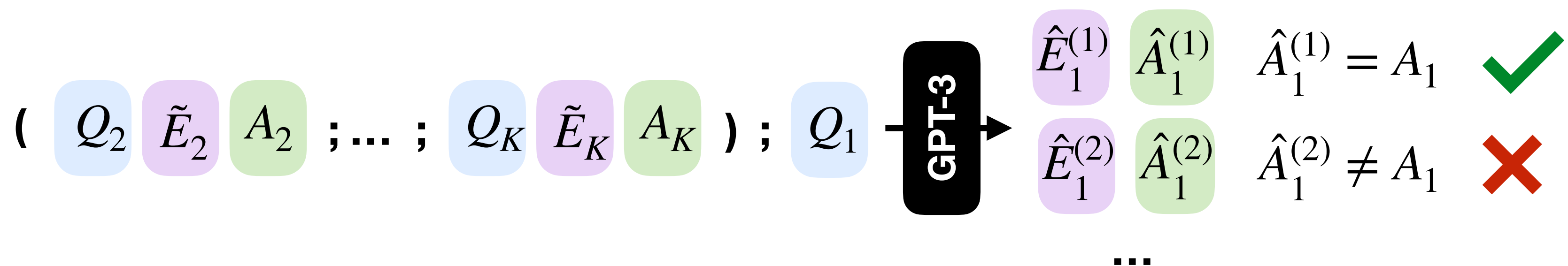
A: Dianna has $2 * 4 = 8$ toys. They have $4 + 8 = 12$ toys in total. The answer is **12**. ✓

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

Approach Overview



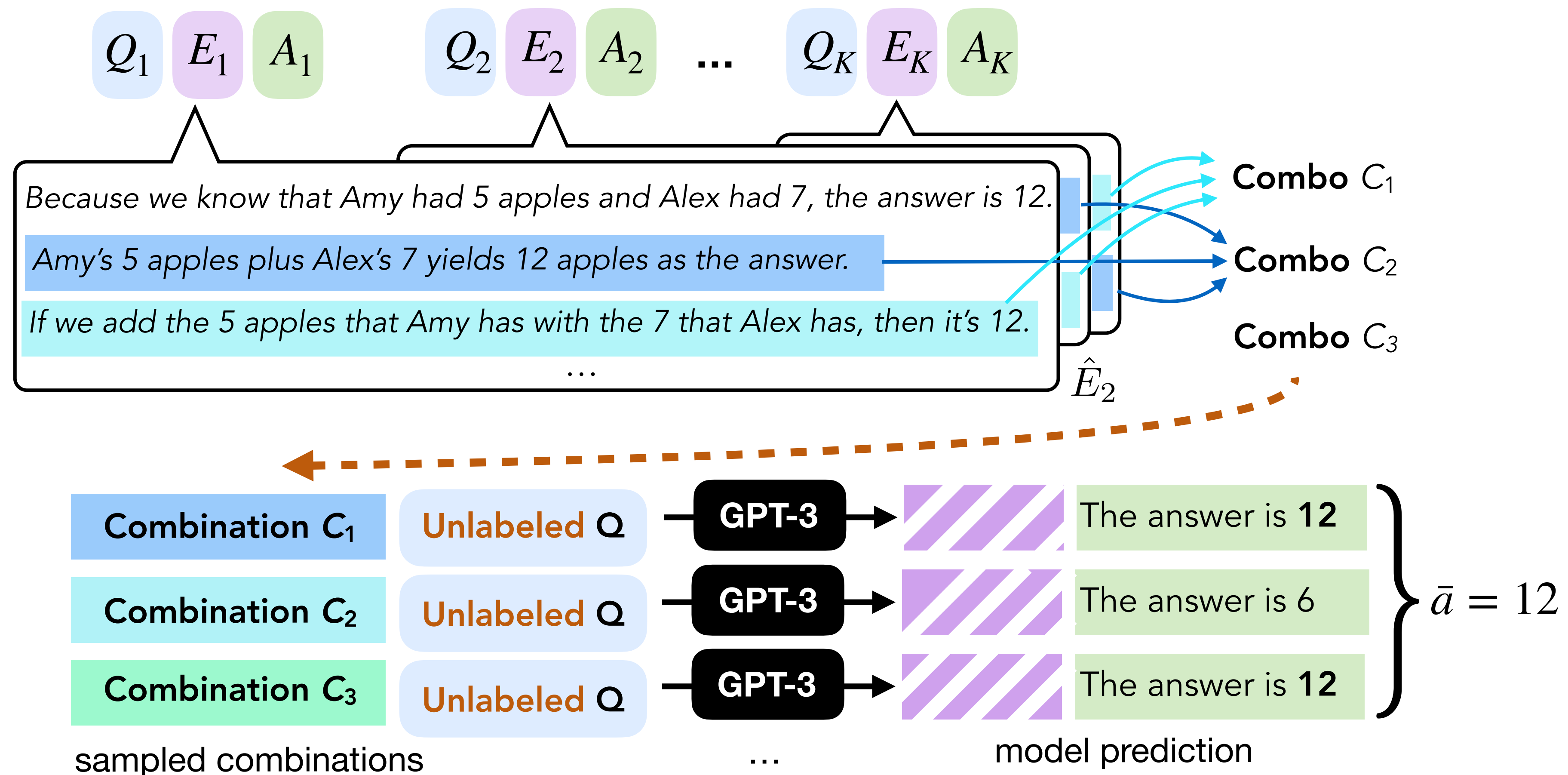
- **Generate candidate explanations:** use seed explanations to perform leave-one-out prompt
 - This yields **combinations** of explanations



Approach Overview



- ▶ **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

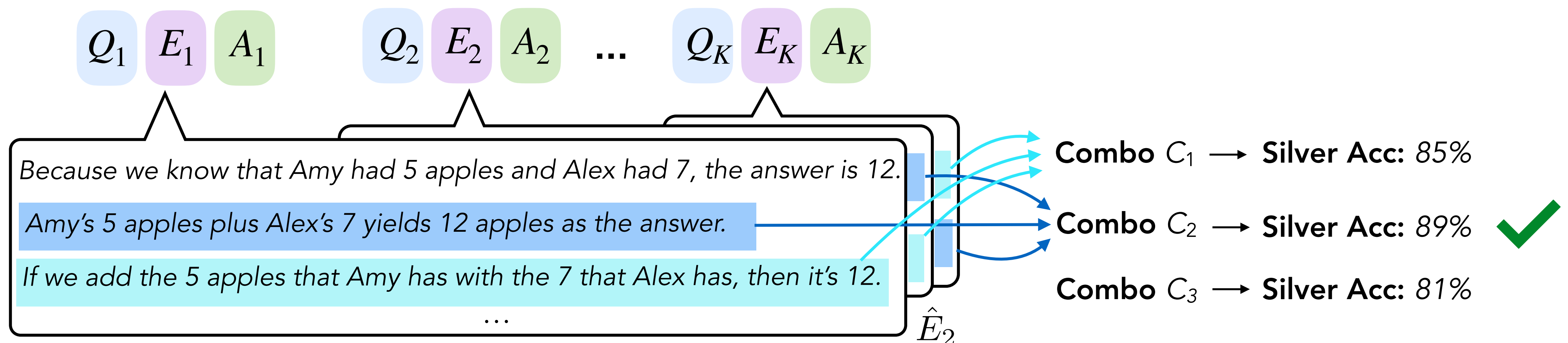


Approach Overview



- ▶ **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

Searching over combinations can be expensive. We search “smartly” by prioritizing exploring promising combinations using proxy metrics. See paper for details.



Experimental Setup



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

$V = Q_1 Q_2 \dots Q_M$

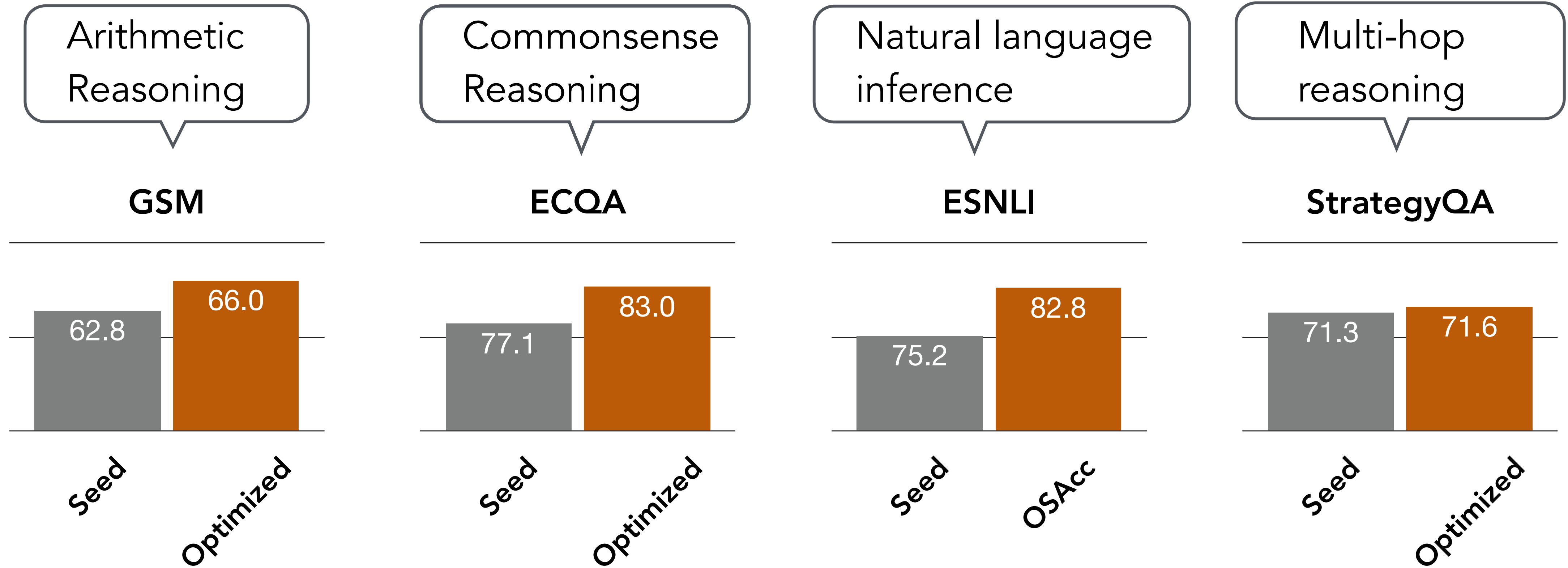
M=256

LLM



Code-davinci-002

Results



Optimizing explanations can lead to substantial gains compared to directly using crowdsourced explanations

Form of Explanations: Chain-of-Thought



Even with carefully crafted explanations, generalizing to hard questions requiring higher depth of reasoning is still challenging for LLMs

Input

Q: Alex, Stan, and Adelwolfe are trying to catch them all, Pokemon that is. Together they have caught 339 Pokemon. Alex has caught 5 more than Stan, and Stan has caught 13 less than 4 times as many as Adelwolfe has caught. How many Pokemon has Stan caught?



GPT-3.5

Output

Let X be the number of Pokemon Stan has caught.
Alex has caught 5 more than Stan, so Alex has caught $X + 5$.
Stan has caught 13 less than 4 times as many as Adelwolfe has caught, so Stan has caught $4X - 13$.
Together they have caught 339 Pokemon, so $X + 5 + 4X - 13 = 339$.
Combining like terms produces $5X + 5 = 339$.
Subtracting 5 from both sides produces $5X = 334$. Dividing both sides by 5 produces $X = 66.8$, so Stan has caught 66 Pokemon.
The answer is 66.

Use explanations that are easier for LLMs to follow

Form of Explanations: Chain-of-Thought



Problem

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

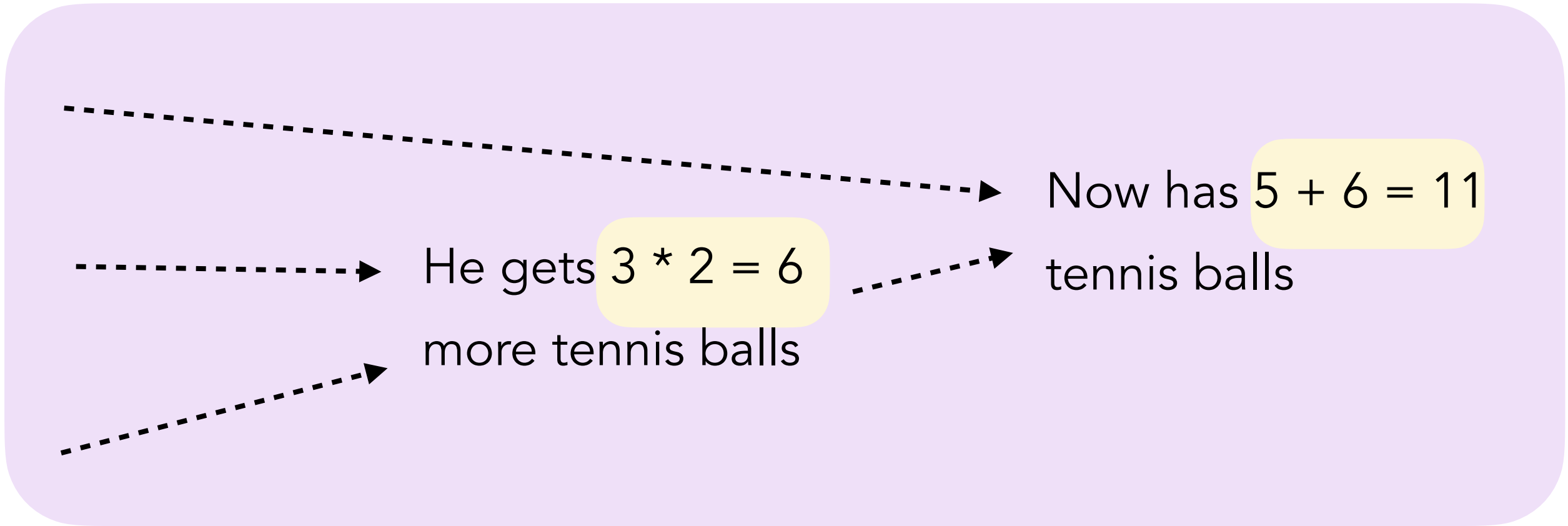
Chain-of-Thought Explanation

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

CoT supervises LLMs to plan out deduction steps and execute the computation

Facts

Roger has 5 tennis balls.
He buys 2 more cans of tennis balls.
Each can has 3 tennis balls.



Query

How many tennis balls does he have now?

Form of Explanations: Chain-of-Thought



CoT supervise LLMs to plan out deduction steps and execute the computation



Not good at planning complex solving procedure

(Valmeekam et al., 2022;
Ribeiro et al., 2023)

Input

Q: Alex, Stan, and Adelwolfe are trying to catch them all, Pokemon that is. Together they have caught 339 Pokemon. Alex has caught 5 more than Stan, and Stan has caught 13 less than 4 times as many as Adelwolfe has caught. How many Pokemon has Stan caught?



GPT-3.5

Output

Let X be the number of Pokemon Stan has caught.
Alex has caught 5 more than Stan, so Alex has caught $X + 5$.
Stan has caught 13 less than 4 times as many as Adelwolfe has caught, so Stan has caught $4X - 13$.
Together they have caught 339 Pokemon, so $X + 5 + 4X - 13 = 339$.
Combining like terms produces $5X + 5 = 339$.
Subtracting 5 from both sides produces $5X = 334$. Dividing both sides by 5 produces $X = \$66.80$, so Stan has caught 66 Pokemon.
The answer is 66.

Wrong steps in the plan

Form of Explanations: Chain-of-Thought



CoT supervise LLMs to plan out deduction steps and execute the computation

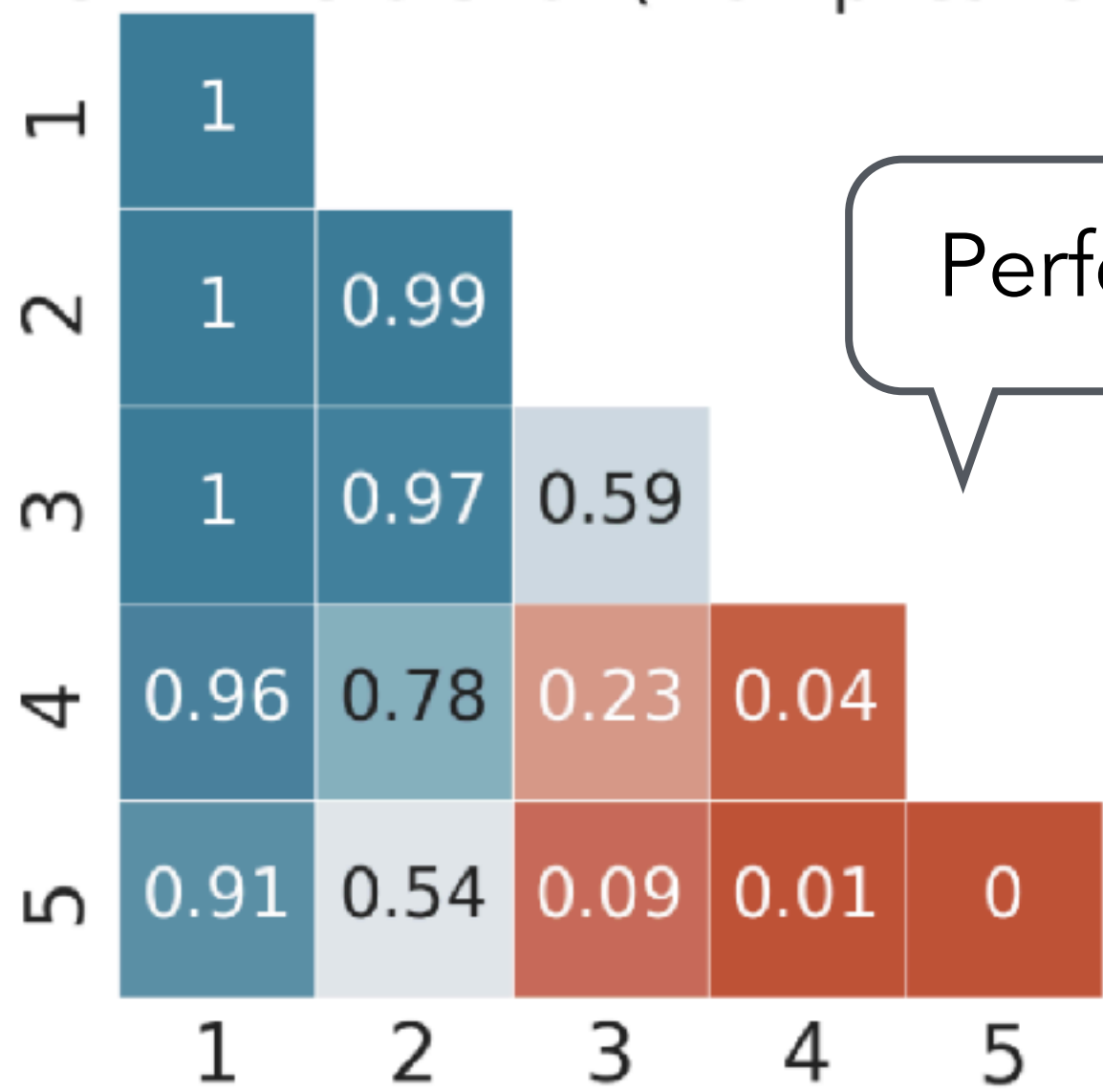
 Not good at planning complex solving procedure

(Valmeekam et al., 2022;
Ribeiro et al., 2023)

 Not good at executing intensive computation

(Chen et al., 2022; Gao et al.,
2023; Lyu et al., 2023)

GPT4 zero-shot (Multiplication)



Performance decreases with growing complexity

(Dziri et al., 2023)

Form of Explanations: Chain-of-Thought



CoT supervise LLMs to plan out deduction steps and execute the computation



Not good at planning complex solving procedure

(Valmeekam et al., 2022;
Ribeiro et al., 2023)



Not good at executing intensive computation

(Chen et al., 2022; Gao et al.,
2023; Lyu et al., 2023)



Good at interpreting the semantics in NL problems



We let LLMs focus on interpreting the NL problem

And offload the work of planning and executing to a symbolic solver

Offloading Planning and Execution to Solver



Each of 5 students—Hubert, Lori, Paul, Regina, and Sharon—will visit one of three cities—Montreal, Toronto, or Vancouver. Sharon visits a different city than Paul. Each student visits one of the cities with at least one of the other four students. [...] Q: Is it true that if any of the students visits Montreal, Lori visits Montreal?



LLM

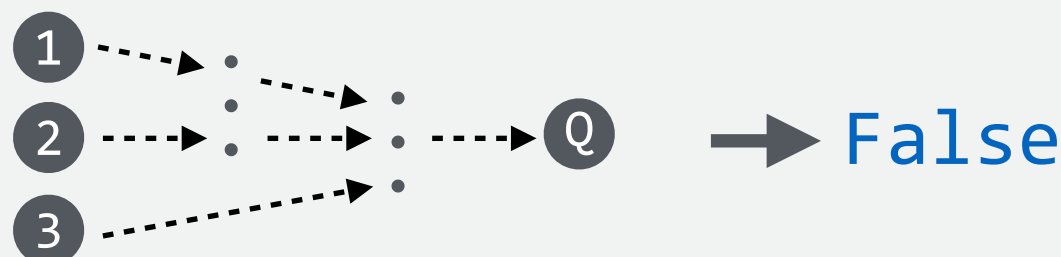
LLMs parse the NL problem into declarative formal specifications

```
students=[Hubert,Lori,Paul,Regina,Sharon], cities=[Montreal,Toronto,Vancouver]
visits = Function(students, cities)
# Sharon visits a different city than Paul
1 visits(Sharon) != visits(Paul)
# Each student visits one of the cities with at least one other student
2 ForAll([s1], Exists([s2], And(s2 != s1, visits(s1) == visits(s2))))
3 [...]
Q solve(Implies(Exists([s], visits(s) == Montreal), visits(L) == Montreal)) # Question
```

Specification encodes facts about objects, relations, and constraints.

query

Z3 SMT Solver



SMT solver generates and executes a plan using automated theorem proving

Declarative vs Imperative (Past Work)



Input


Q: Alex, Stan, and Adelwolfe are trying to catch them all, Pokemon that is. Together they have caught 339 Pokemon. Alex has caught 5 more than Stan, and Stan has caught 13 less than 4 times as many as Adelwolfe has caught. How many Pokemon has Stan caught?

SAT Problem: Declarative

Output

```
total_pokemon = 339
stan_pokemon = Variable()
alex_pokemon = stan_pokemon + 5
stan_pokemon = adelwolfe_pokemon * 4 - 13
total_pokemon = alex_pokemon +
stan_pokemon + adelwolfe_pokemon
result = stan_pokemon
solve(result)
```

SMT Solver



 result = 147

Program: Imperative

encodes the plan for solving the problem

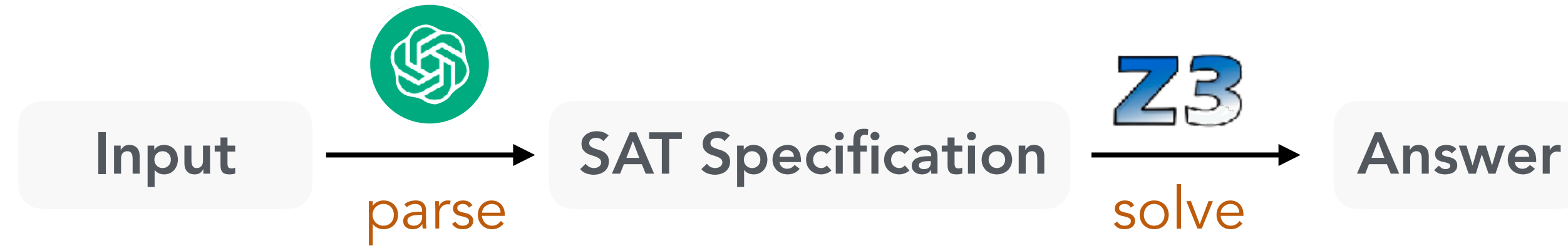
```
total_pokemon = 339
alex_pokemon = 5
stan_pokemon = 4
adelwolfe_pokemon = 13
stan_pokemon = (total_pokemon -
alex_pokemon - adelwolfe_pokemon *
stan_pokemon) / (1 - stan_pokemon)
result = stan_pokemon
```

Python interpreter

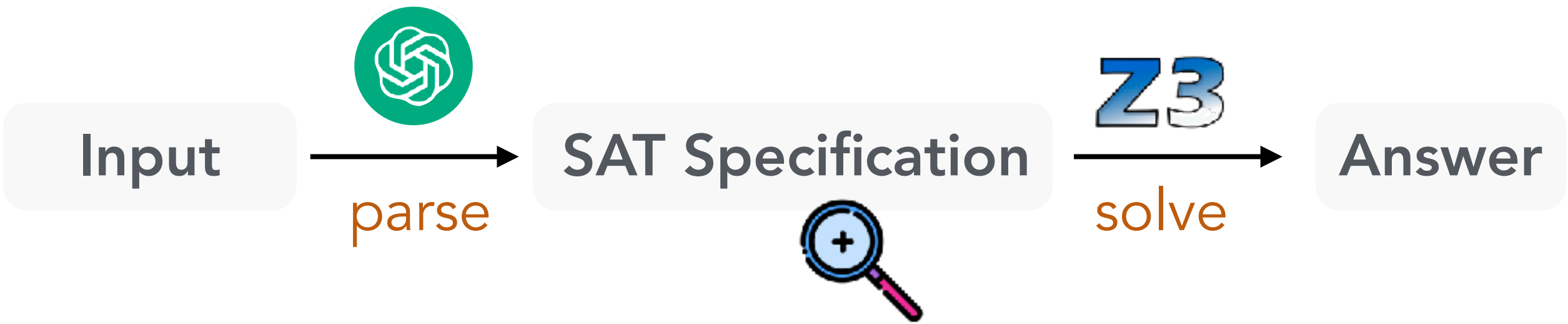
 result = -94 

(Chen et al., 2022; Gao et al., 2023; Lyu et al., 2023)

SAT-Aided Framework



SAT Problem



SAT Specification encodes a **SAT problem**, formally defined as

$$\mathcal{P} = (\Phi, \mathcal{T}, Q)$$

Formulas Φ

$$\{x + y = 3, x - y = 1\}$$

Solver finds value assignment that can satisfy all formulas
 $x = 2, y = 1$ **Z3**

Query Q

$$x - 2$$

Evaluate the value of the query

Theory \mathcal{T}

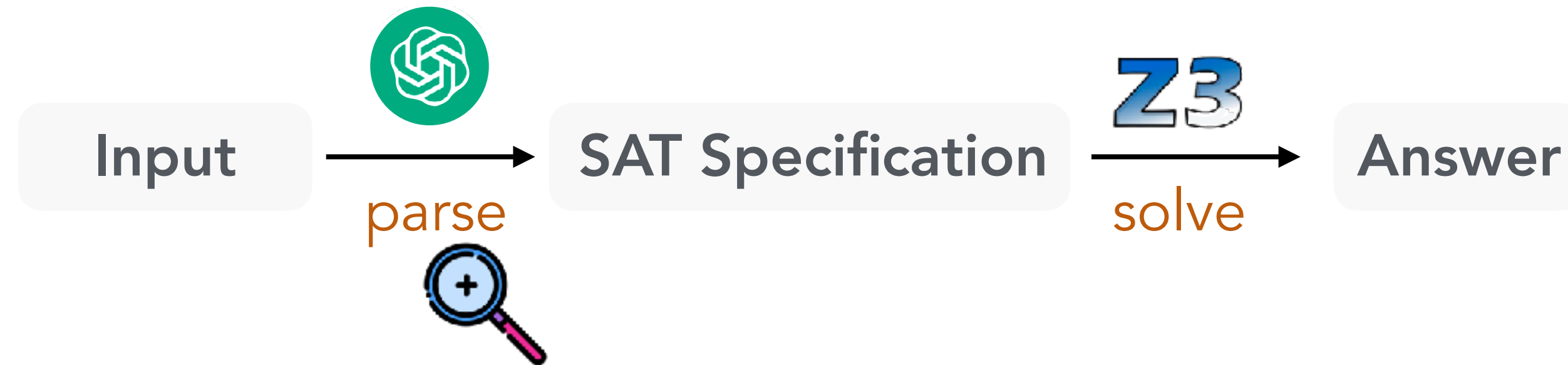
Theory of integers, Theory of equality

Define the meaning of some symbols in formulas, e.g., +, =

SMT formulation is expressive

Allows handling a lot of reasoning tasks with **unified formulation and solver** using theory of linear arithmetic, theory of arrays, theory of strings, etc.

Parsing into SAT specification



Few-shot Input-Specification Examples [...]

Few-shot in-context learning

Each of 5 students—Hubert, Lori, Paul, Regina, and Sharon—will visit one of three cities—Montreal, Toronto, or Vancouver. Sharon visits a different city than Paul. [...]



LLM

code style syntax,
closer to pretraining
data of LLMs

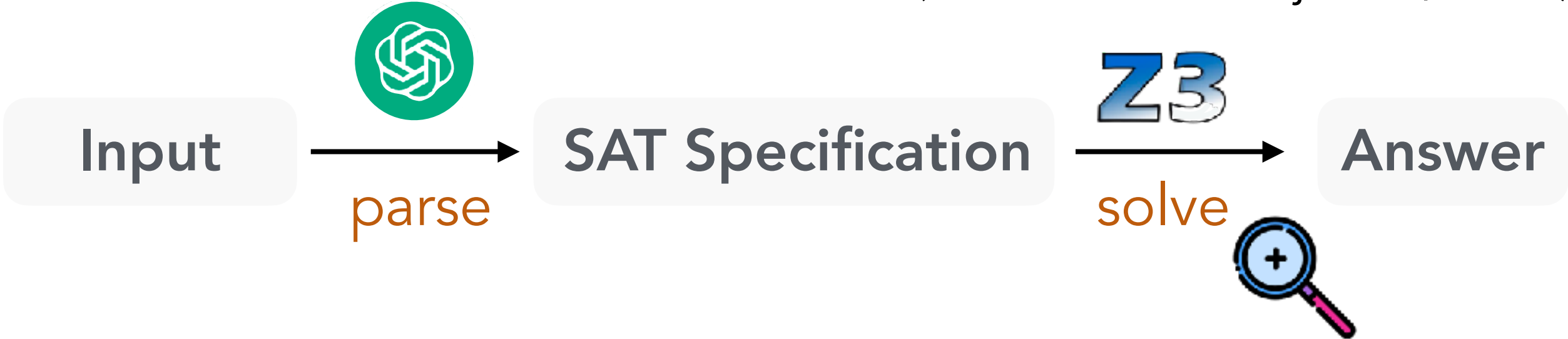
```
students=[Hubert,Lori,Paul,Regina,Sharon],  
cities=[Montreal,Toronto,Vancouver]  
visits = Function(students, cities)  
# Sharon visits a different city than Paul  
visits(Sharon) != visits(Paul)  
[...]
```

Interleaving NL (as comments) and
specification to improve fidelity of translation

Solving with Z3



(De Moura and Bjørner, 2008)

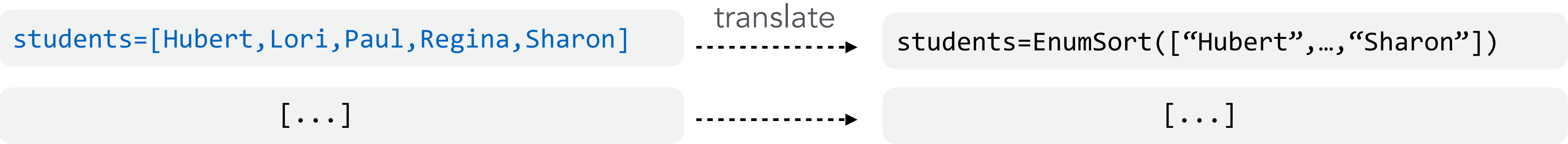


SAT specification

```
students=[Hubert,Lori,Paul,Regina,Sharon],
cities=[Montreal,Toronto,Vancouver]
visits = Function(students, cities)
# Sharon visits a different city than Paul
visits(Sharon) != visits(Paul)
[...]
```

Extract formulas Φ and query Q

Translate to actual python code that can be executed using Z3py



Experiments: Setup



Baselines

CoT: imperative NL explanations

PAL: imperative python programs

SatLM: declarative SAT specifications

Model



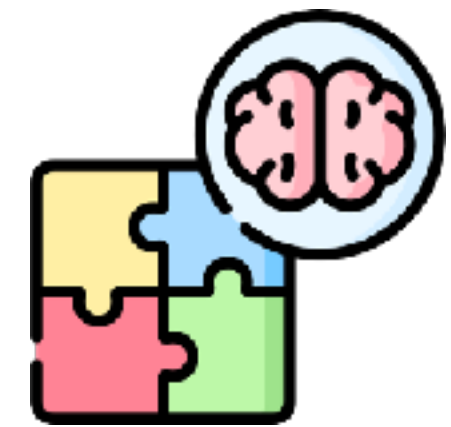
gpt-3.5 (code-davinci-002)

Tasks

Arithmetic Reasoning



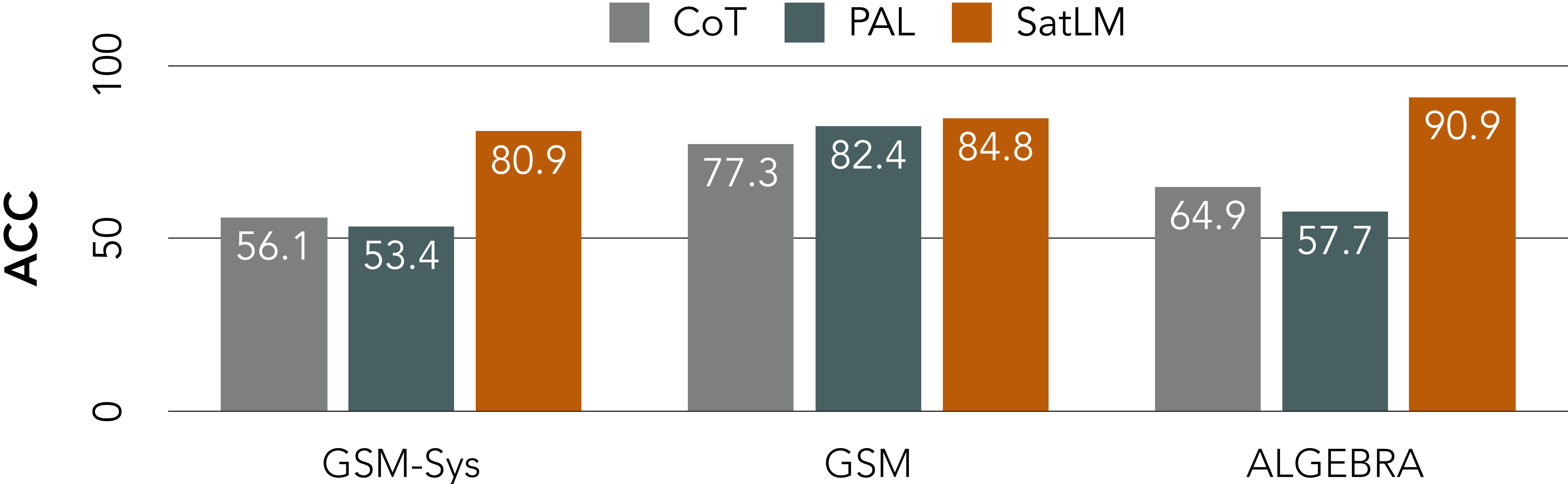
Logical Reasoning



Symbolic Reasoning
(reason over arrays)

Regex Synthesis
(reason over strings)

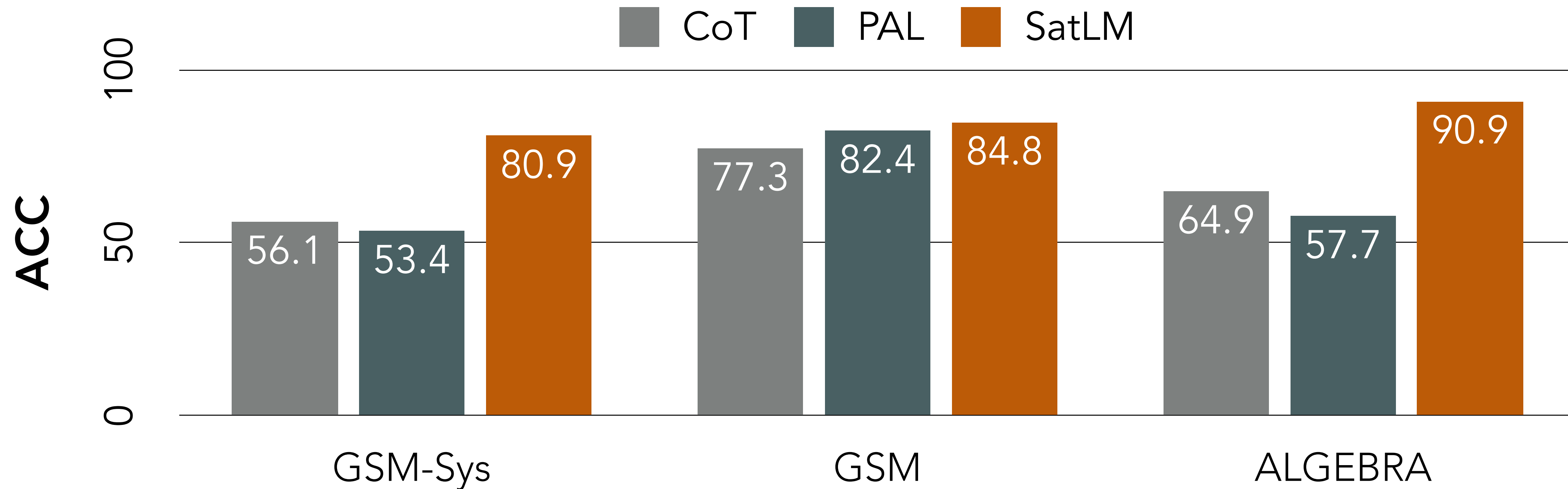
Results: Arithmetic Reasoning



A subset of GSM requiring more complex backward reasoning or constraint solving

Problems from algebra textbooks

Results: Arithmetic Reasoning

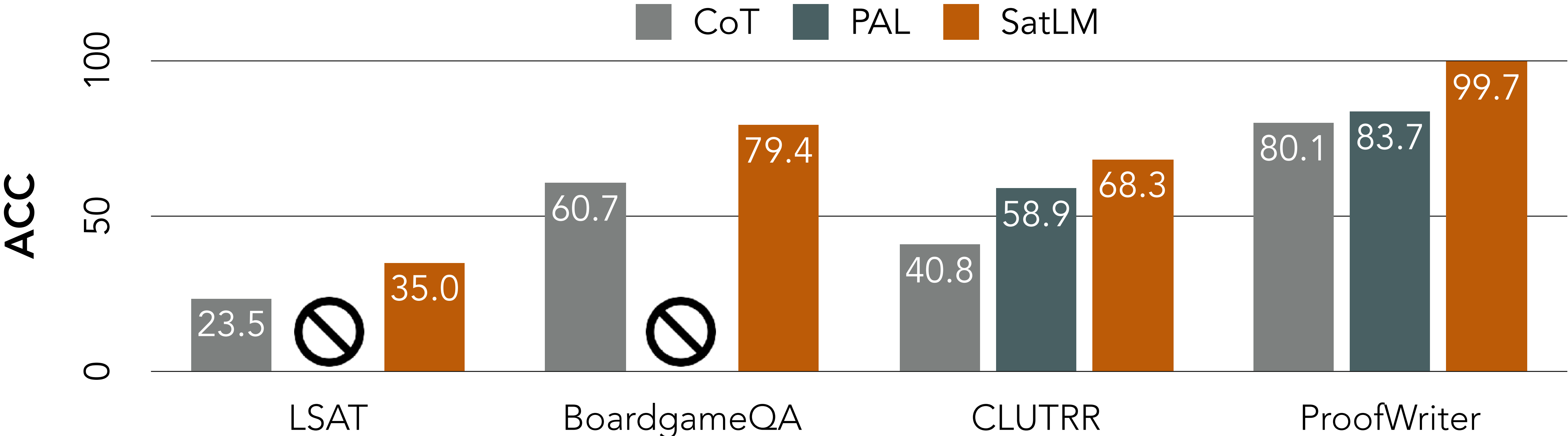


A subset of GSM requiring more complex backward reasoning or constraint solving

Problems from algebra textbooks

SatLM substantially outperform baselines using imperative specifications on more challenging datasets GSM-Sys and ALGEBRA

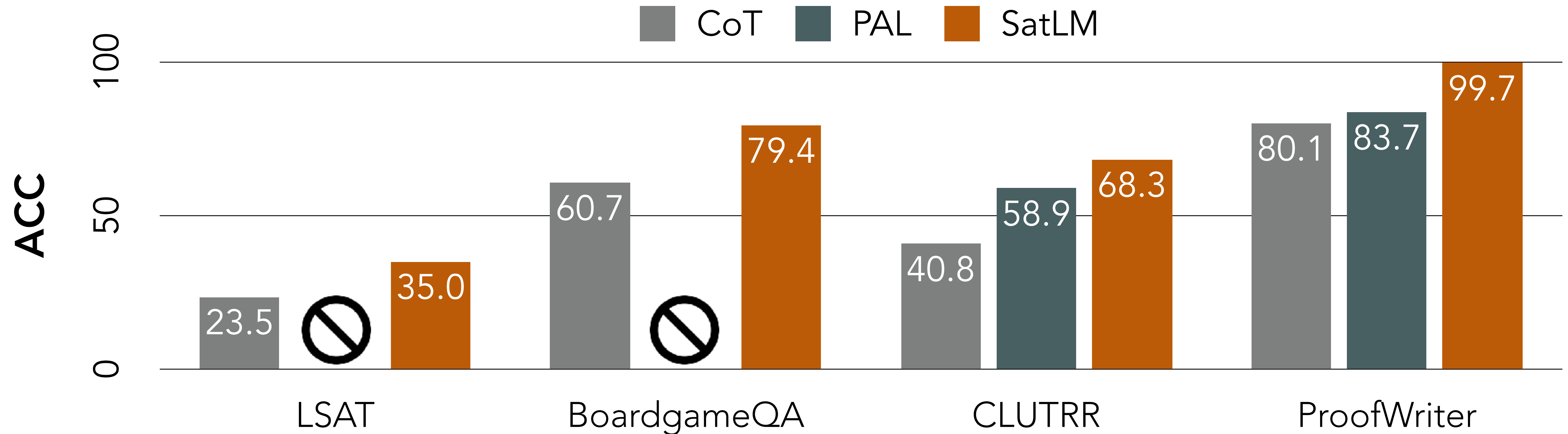
Results: Logical Reasoning



Problems from law-school admission test, CoT on par with random guess

Requires commonsense reasoning

Results: Logical Reasoning



Problems from law-school admission test, CoT can par with random guess

Requires commonsense reasoning

Using SMT solver, SatLM can handle problems requiring reasoning of high depth

Benefits of using SMT Solver



SMT solver can spot semantic errors in the specification

Unsatisfiable

Conflicting formulas

$$\begin{aligned}y &= x + 1 \\z &= x - 1 \\x &= y + 1\end{aligned}$$

Ambiguous

Multiple feasible solutions

$$\begin{aligned}x &= y + 1 \\x &> 0\end{aligned}$$

Exception

Syntax errors, time-out, etc.

Program interpreters typically can only spot this type of errors



Selective Prediction with SMT Solver



SMT solver can spot semantic errors in the specification

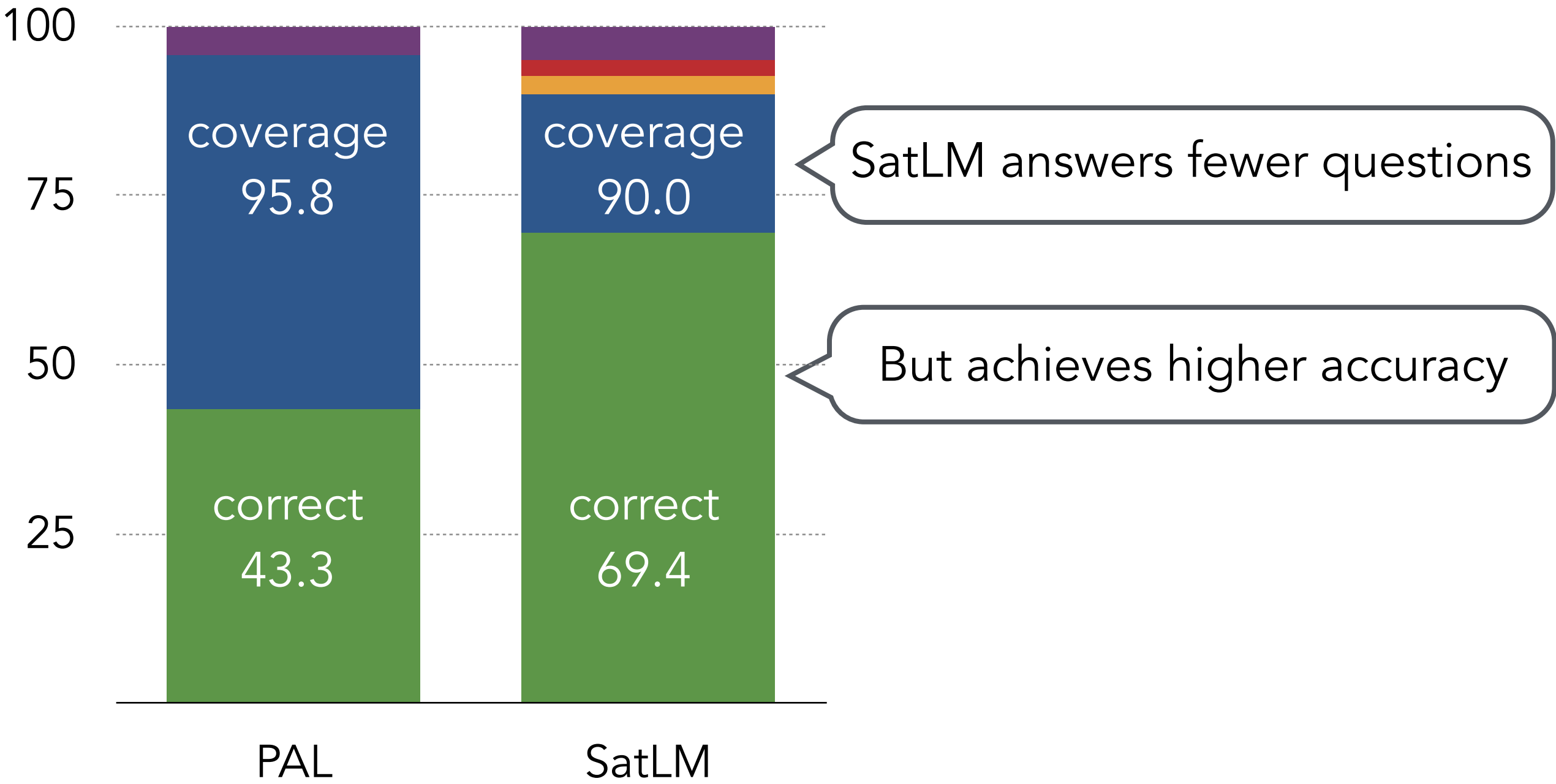
When SatLM successfully returns an answer, it is more likely to be correct

Unsatisfiable 

Ambiguous 

Exception  

Correct Incorrect UNSAT AMBIG EXCEP



Selective Accuracy
correct answer / coverage

PAL: 45.2

SatLM: 77.1

Commonsense Reasoning in SatLM



LLMs can perform commonsense reasoning while parsing

Q: Farmer Brown has 60 animals on his farm, all either chickens or cows. He has twice as many chickens as cows. How many legs do the animals have, all together?



GPT-3.5

```
animals_total = 60
animals_chickens = Variable()
animals_cows = Variable()
animals_chickens = animals_cows * 2
animals_total = animals_chickens + animals_cows
legs_chickens = animals_chickens * 2
legs_cows = animals_cows * 4
legs_total = legs_chickens + legs_cows
```

Commonsense.
Parsing goes beyond naive translation

SAT-Aided Framework



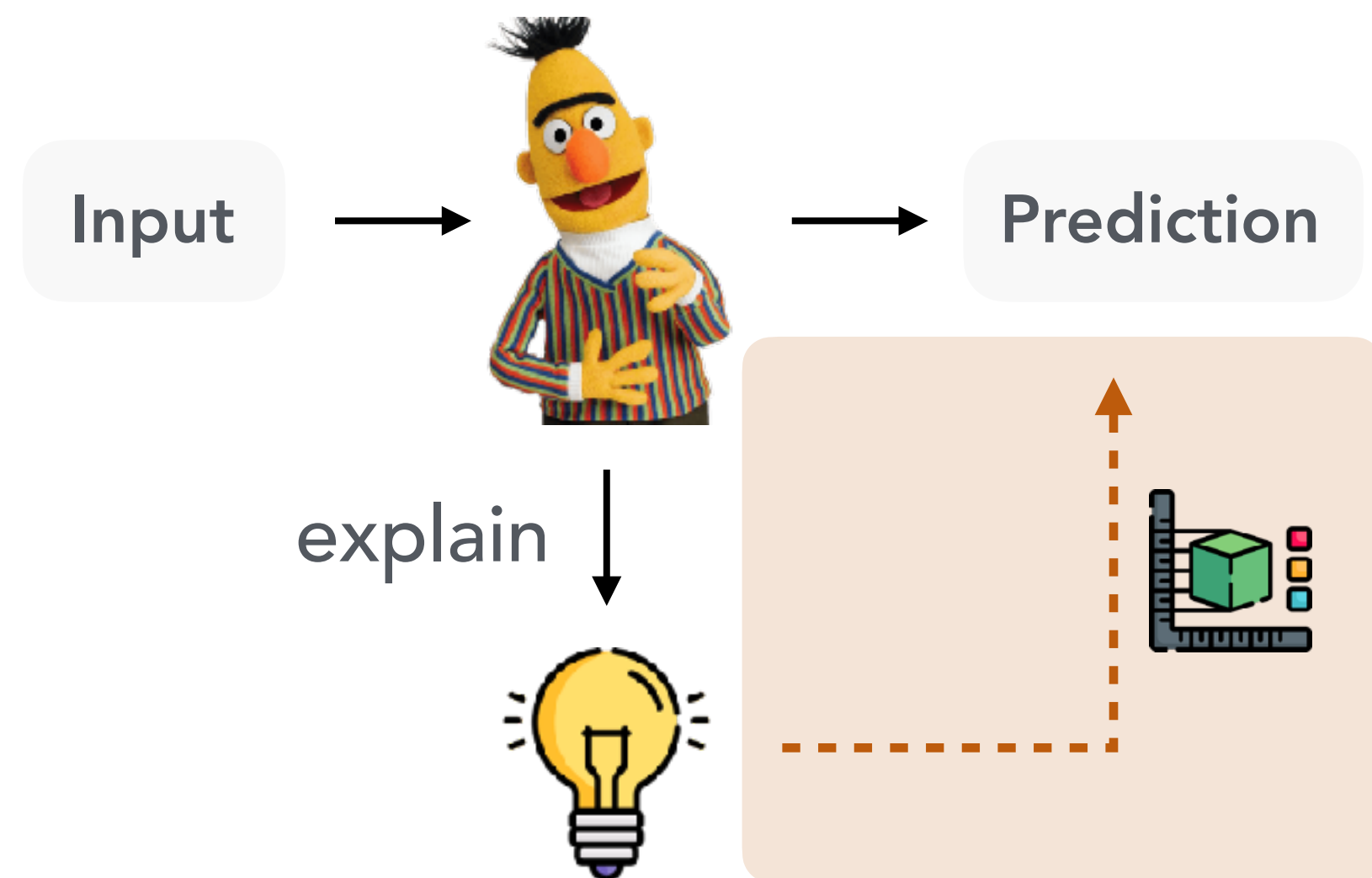
Use SAT specification as explanations for a diverse of reasoning tasks

Offload planning and execution to SMT solver

Steering Textual Reasoning with Explanations



Post-Hoc Intervene



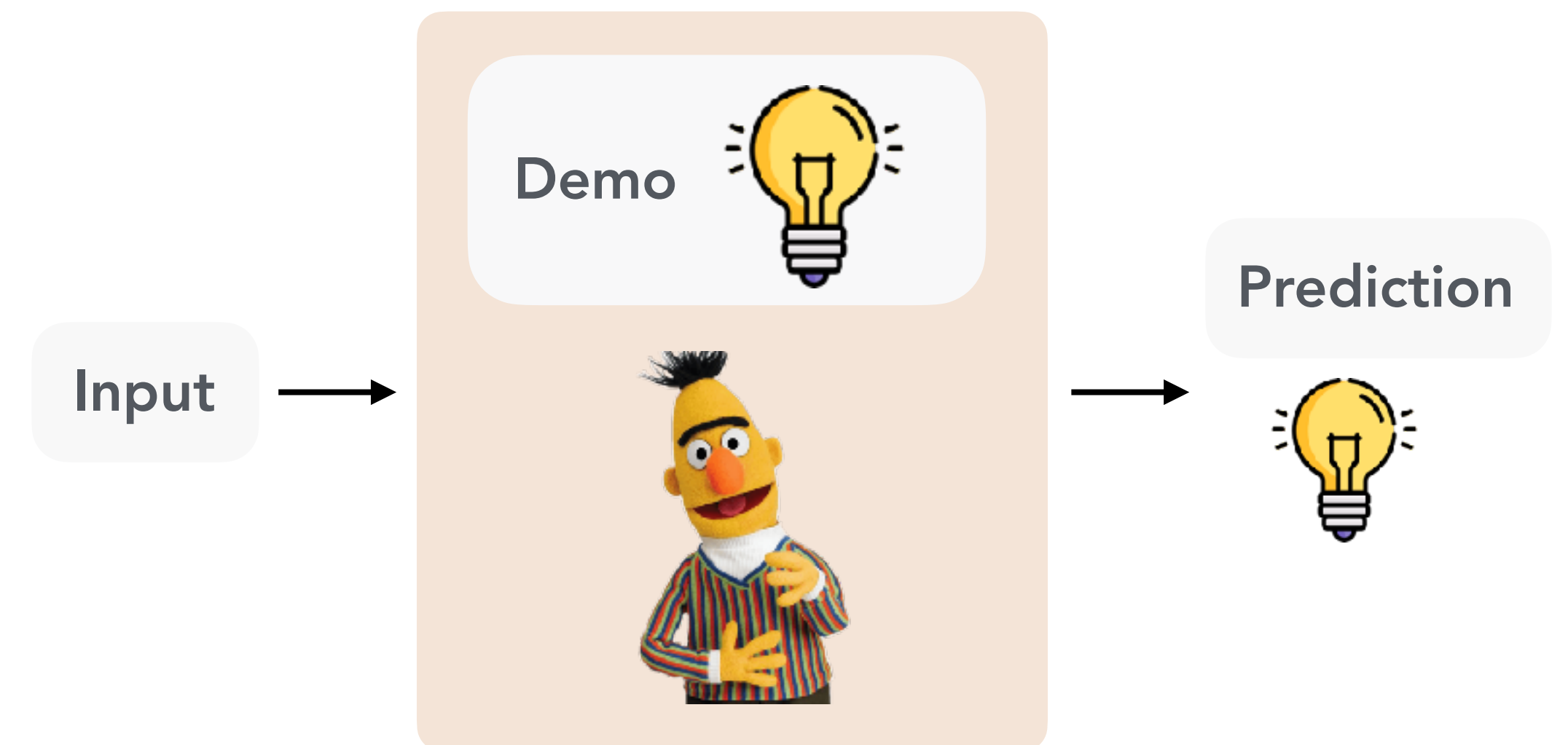
XY++ NeurIPS 22

XY++ ACL 22

XY++ EMNLP 21

PS*, JF*, **XY++** EACL 23

Teach with Explanations



XY++ NeurIPS 23

XY++ EMNLP 23

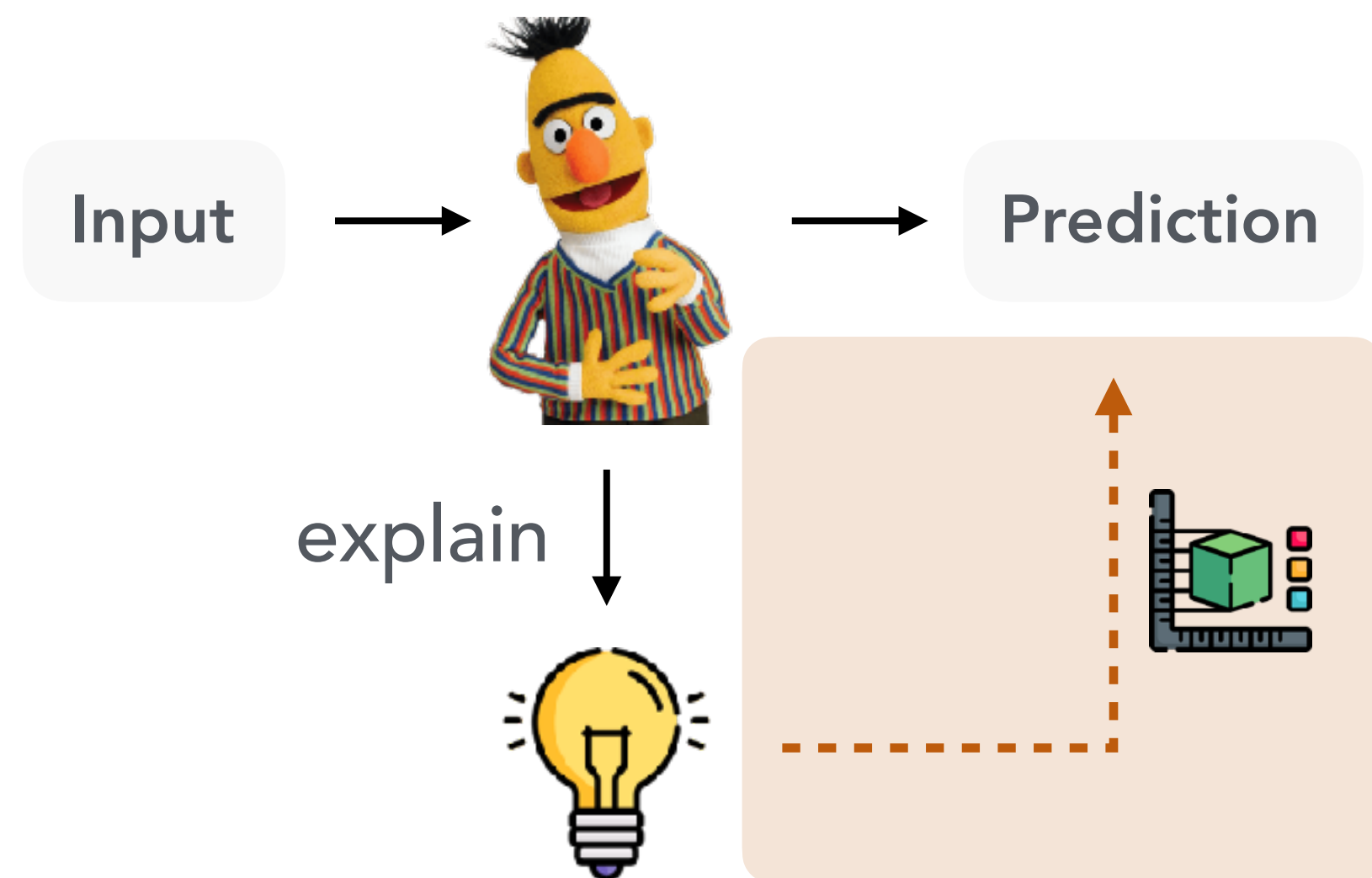
XY++ ACL Findings 23

ZS, **XY++** Arxiv 23 (in sub.)

Steering Textual Reasoning with Explanations

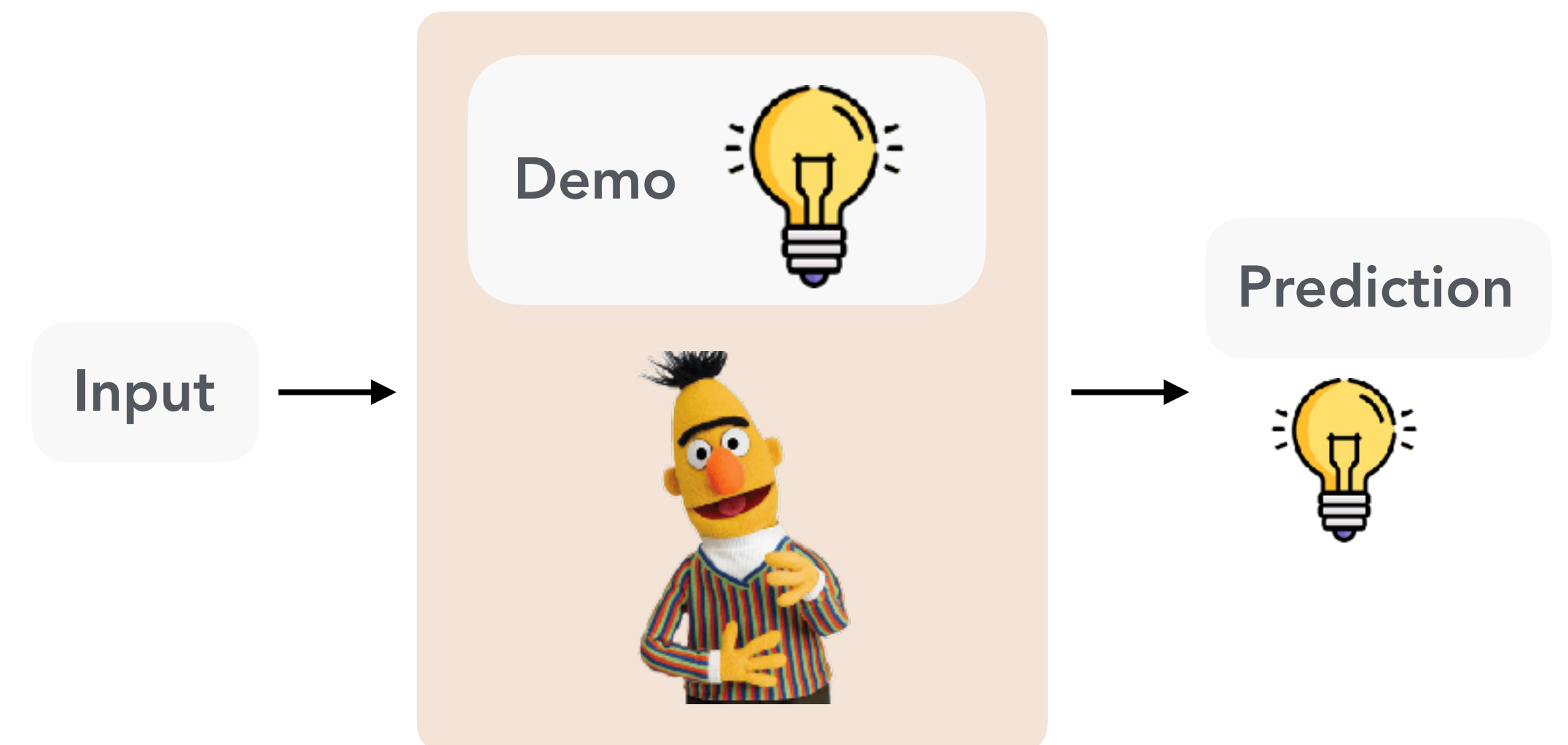


Post-Hoc Intervene



XY++ NeurIPS 22
XY++ ACL 22
XY++ EMNLP 21
PS*, JF*, **XY++** EACL 23

Teach with Explanations



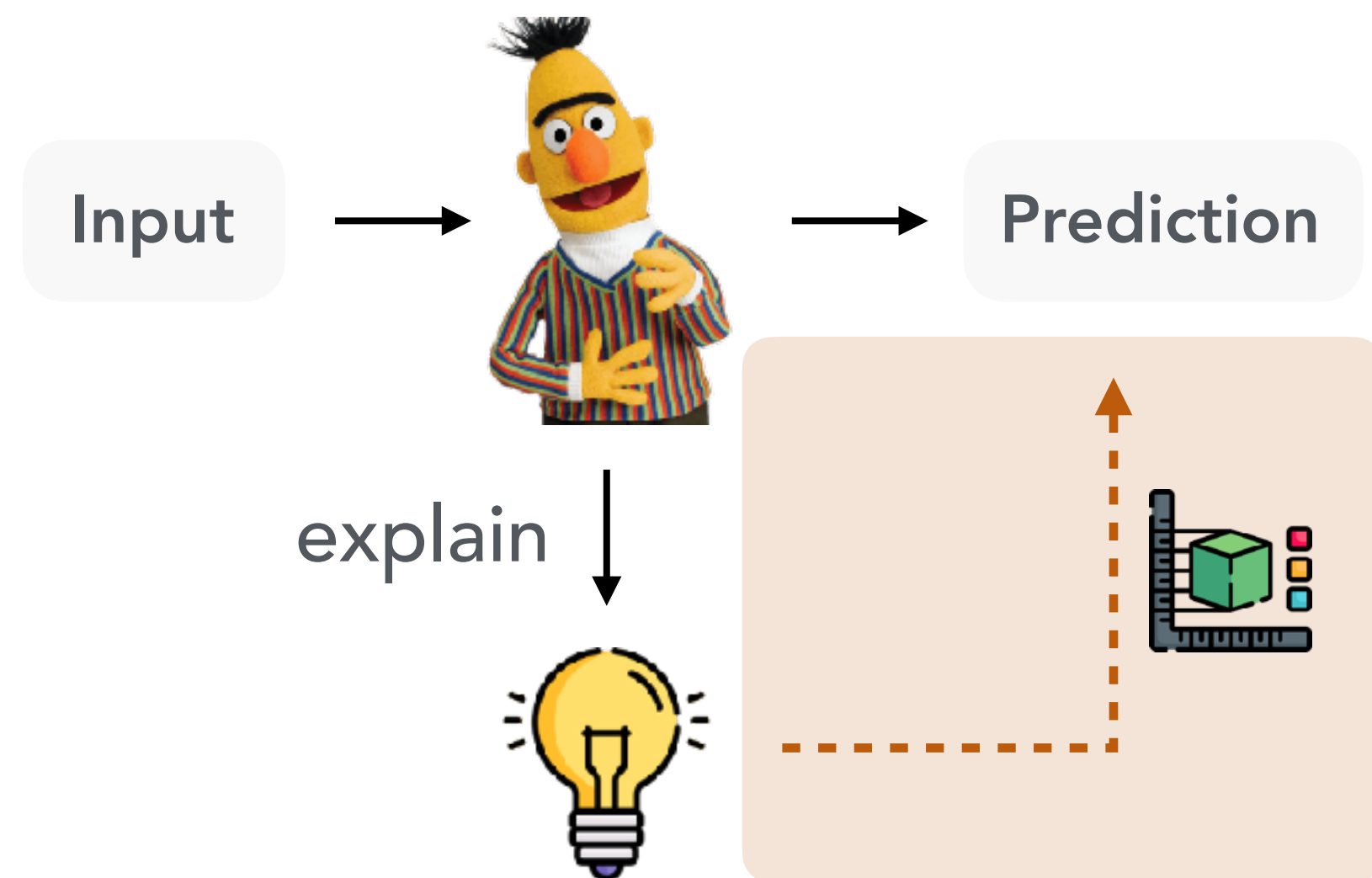
XY++ NeurIPS 23
XY++ EMNLP 23
XY++ ACL Findings 23
ZS, **XY++** Arxiv 23 (in sub.)

Empirical analysis on what makes explanations effective

Steering Textual Reasoning with Explanations

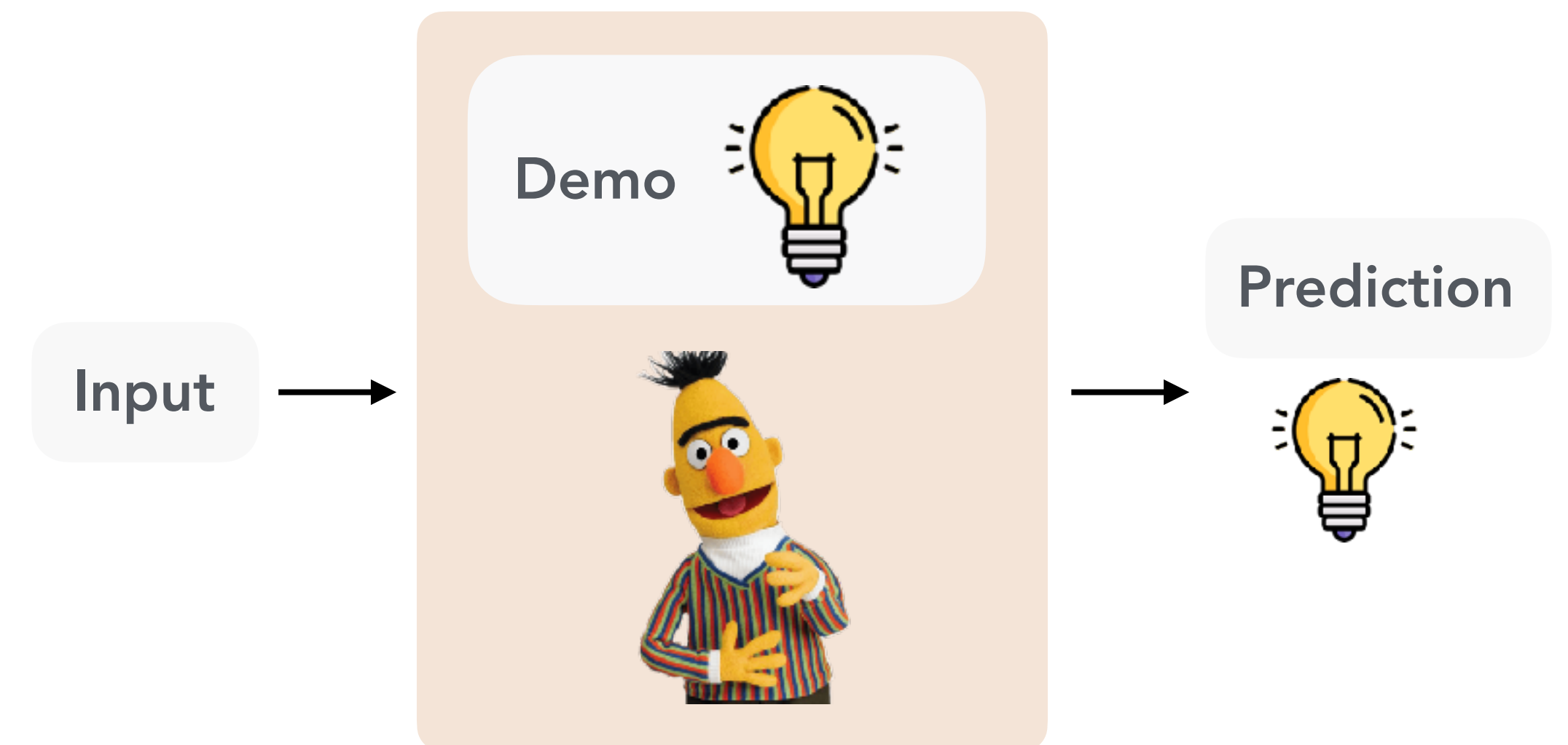


Post-Hoc Intervene



Use explanations to investigate reasoning process and calibrate model predictions post-hoc

Teach with Explanations

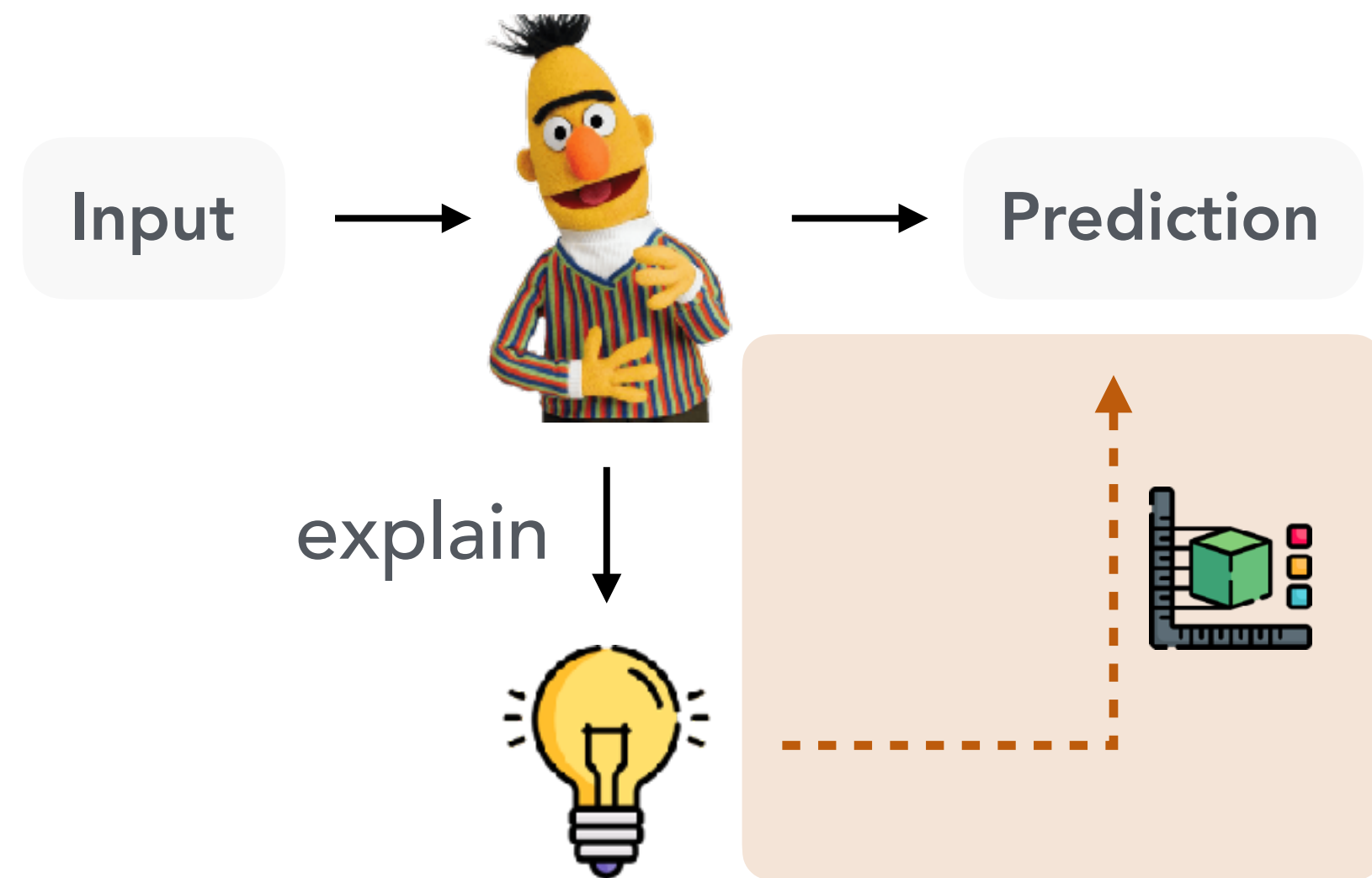


Construct effective explanations written in the right style and in the right form

Questions



Post-Hoc Intervene



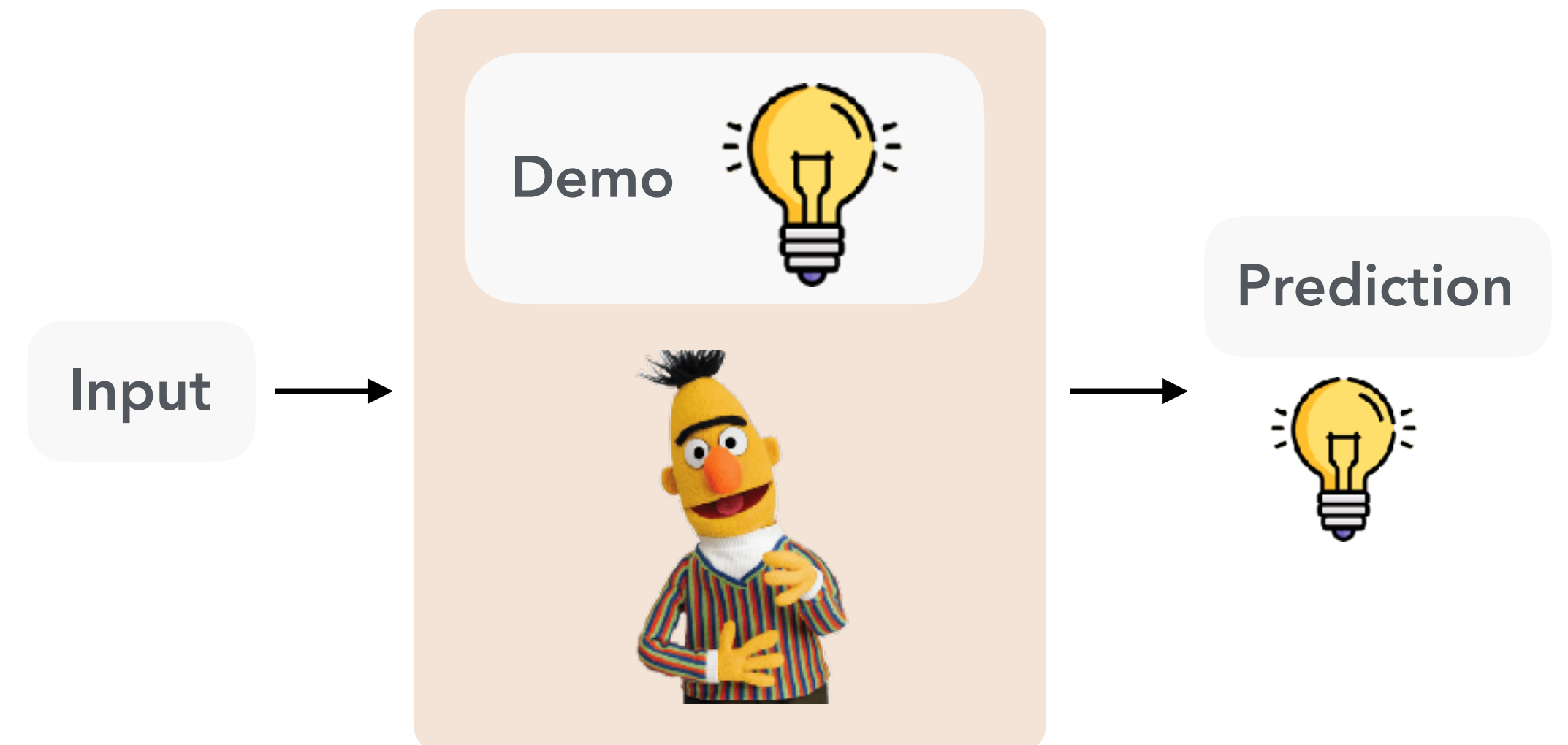
XY++ NeurIPS 22

XY++ ACL 22

XY++ EMNLP 21

PS*, JF*, **XY++** EACL 23

Teach with Explanations



XY++ NeurIPS 23

XY++ EMNLP 23

XY++ ACL Findings 23

ZS, **XY++** Arxiv 23 (in sub.)