Can Explanations Be Useful for Calibrating Black Box Models?

TEXAS

The University of Texas at Austin

TAUR LAB

Xi Ye and Greg Durrett

Black-Box Models

Black-Box Models

Perspective

OpenAI

Google
Natural Language
API

MonkeyLearn

- A growing number of black-box NLP models

Black-Box Models

- A growing number of black-box NLP models
- Performance degradation if deploying black-box models on a new domain

Calibrating Black-box Model

Adversarial SQuAD

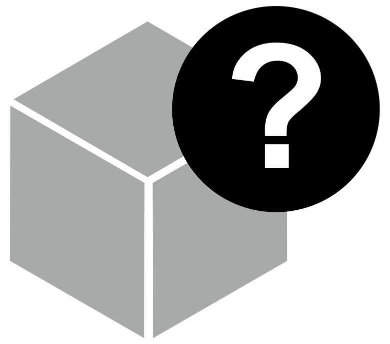
Question

Where did the Panthers practice ?

Context

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

Black-Box QA Model

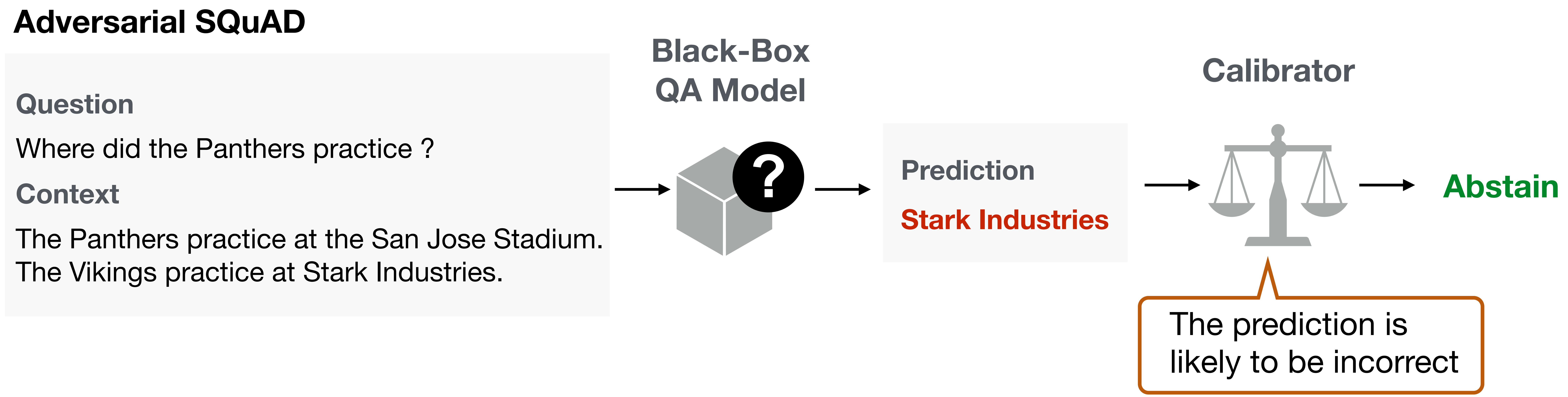


Prediction

Stark Industries

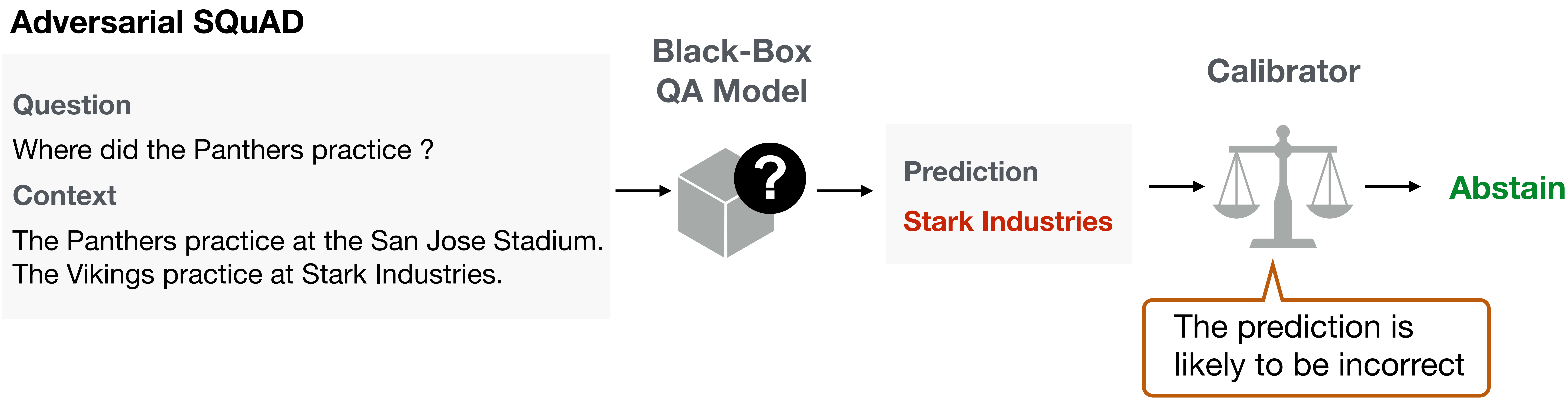
Calibrating Black-box Model

- Hard to calibrate black-box models due to extremely limited information available



Calibrating Black-box Model

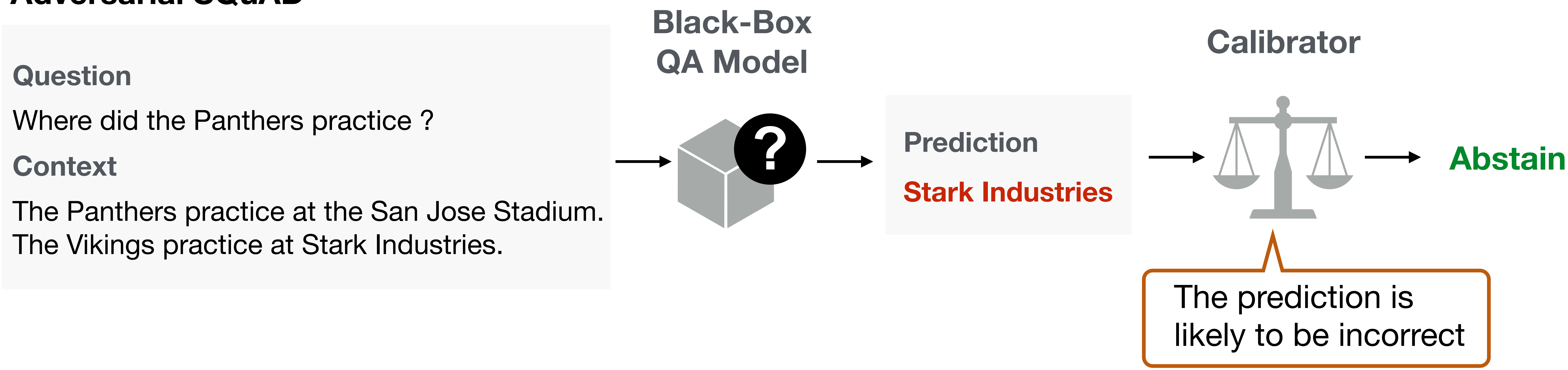
- Hard to calibrate black-box models due to extremely limited information available
- Use explanation techniques to reveal more information



Calibrating Black-box Model

- Hard to calibrate black-box models due to extremely limited information available
- Use explanation techniques to reveal more information
- Core question: can we leverage explanations to calibrate black box models?

Adversarial SQuAD



Ingredients for Calibration

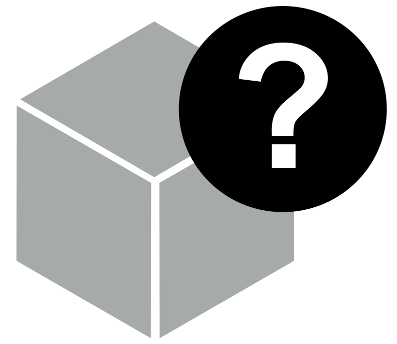
Question

Where did the Panthers practice ?

Context

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

Black-Box QA Model

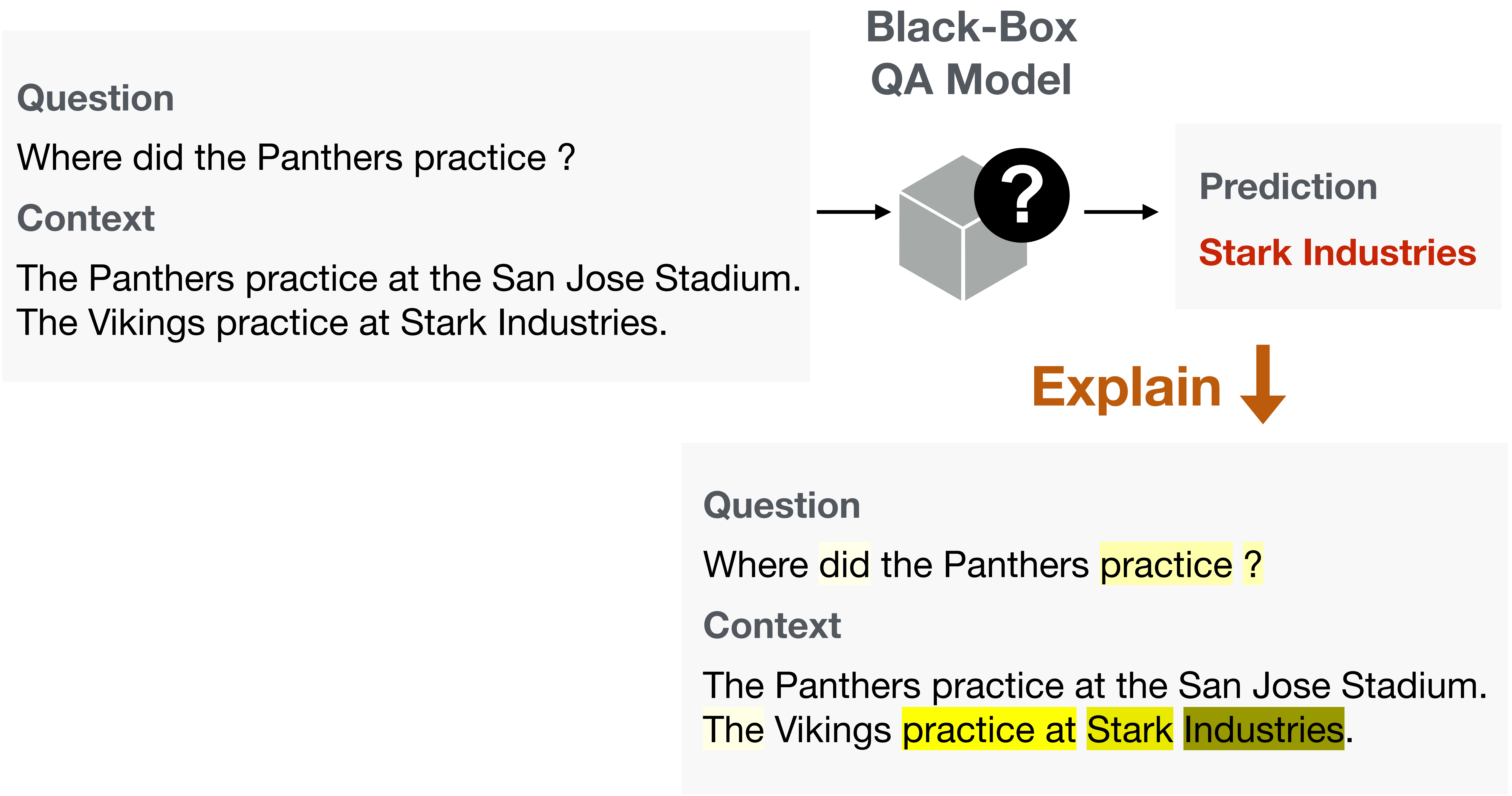


Prediction

Stark Industries

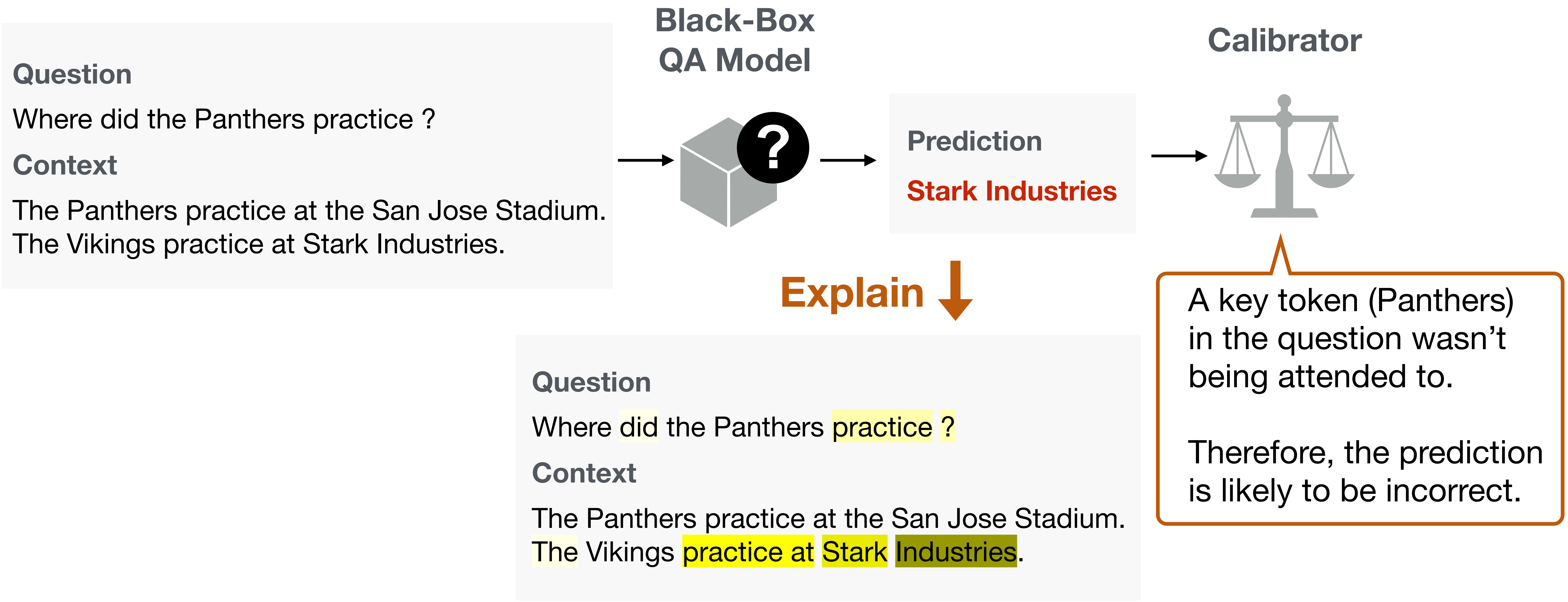
Ingredients for Calibration

- Explanations can tell important features that the model is relying on



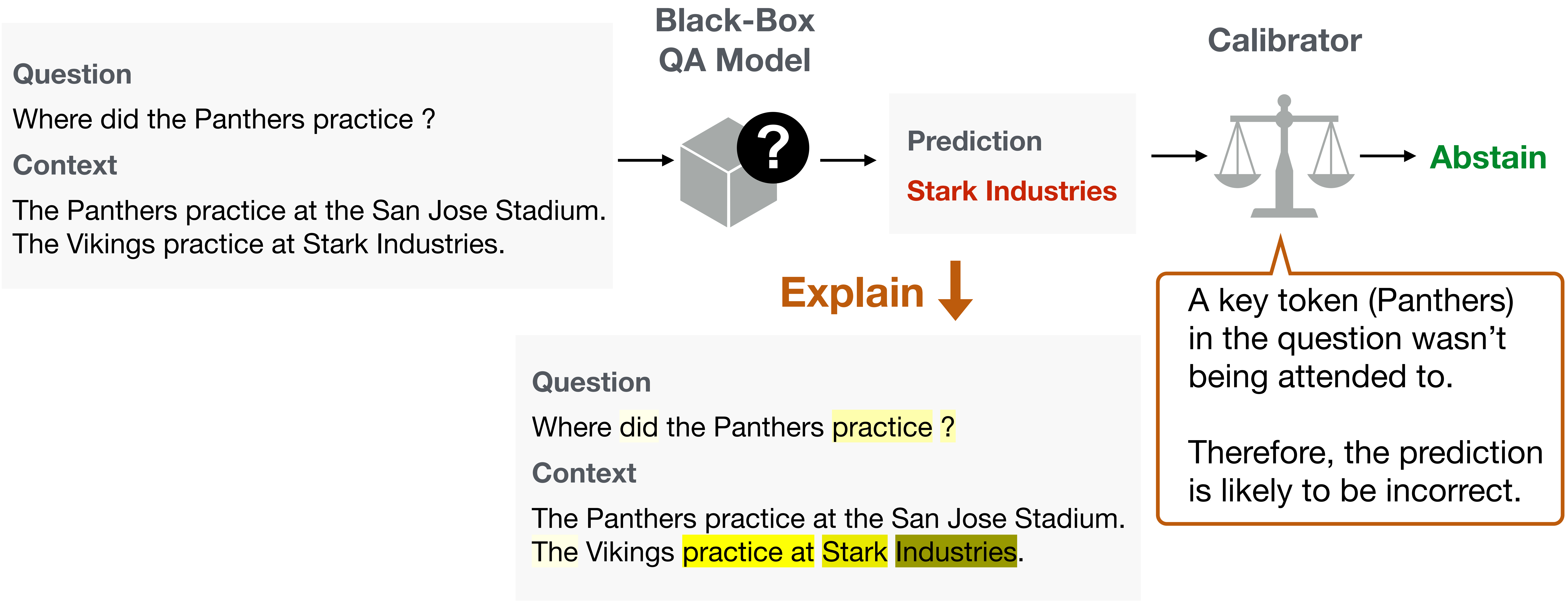
Ingredients for Calibration

- Explanations can tell important features that the model is relying on
- Calibrating based on whether the explanation is reasonable



Ingredients for Calibration

- Explanations can tell important features that the model is relying on
- Calibrating based on whether the explanation is reasonable



Calibrating with Explanations

Answer San Jose **Prediction** **Stark Industries**

Question *Where did the Panthers practice ?*

Context The Panthers practice at the San Jose Stadium.

The Vikings practice at Stark Industries .

Calibrating with Explanations

Answer San Jose **Prediction** **Stark Industries**

Question *Where did the Panthers practice ?*

Context The Panthers practice at the San Jose Stadium.

The Vikings practice at Stark Industries .

Feature **NNP** is not used by the model

- Extract features describing the "reasoning" of the model

Calibrating with Explanations

Answer San Jose **Prediction** **Stark Industries**

Question *Where did the Panthers practice ?*

Context The Panthers practice at the San Jose Stadium.

The Vikings practice at Stark Industries .

Feature **NNP** is not used by the model

↓

Calibrator

↓

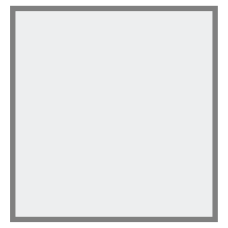
incorrect answer

- Extract features describing the “reasoning” of the model
- Use features to assess the correctness of the prediction

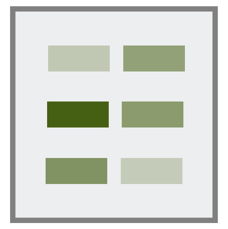
Calibration Framework

Calibration Framework

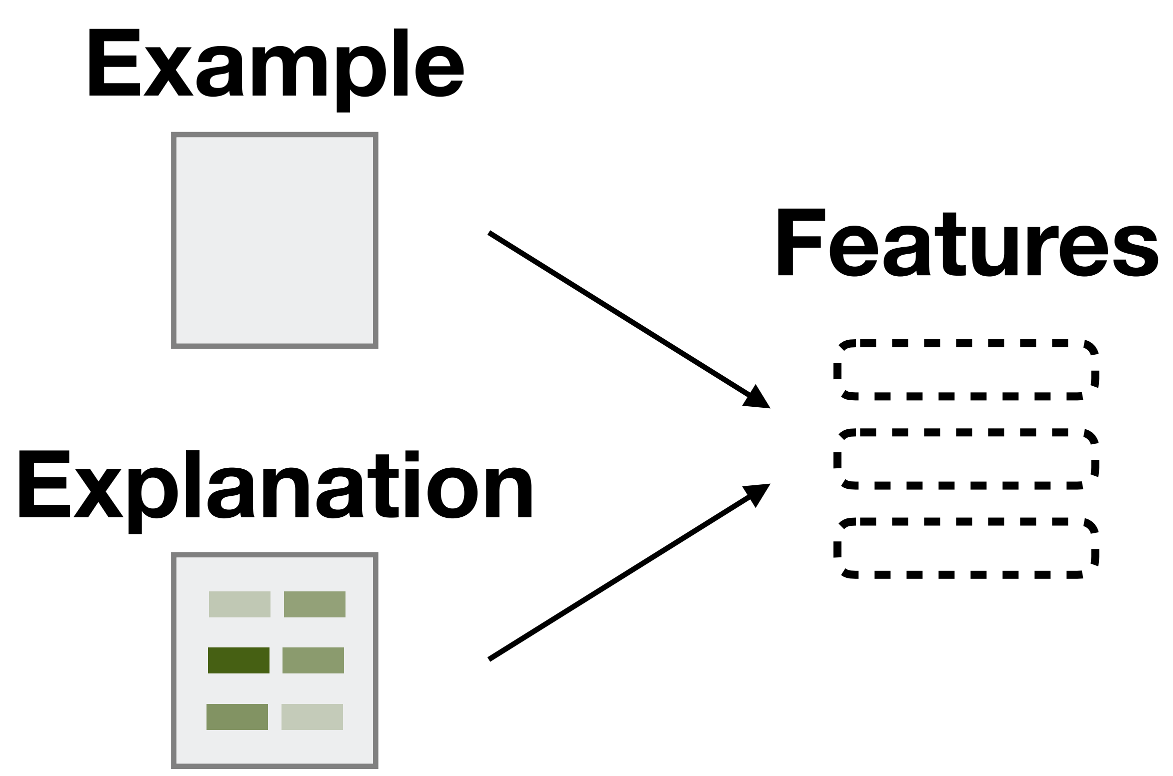
Example



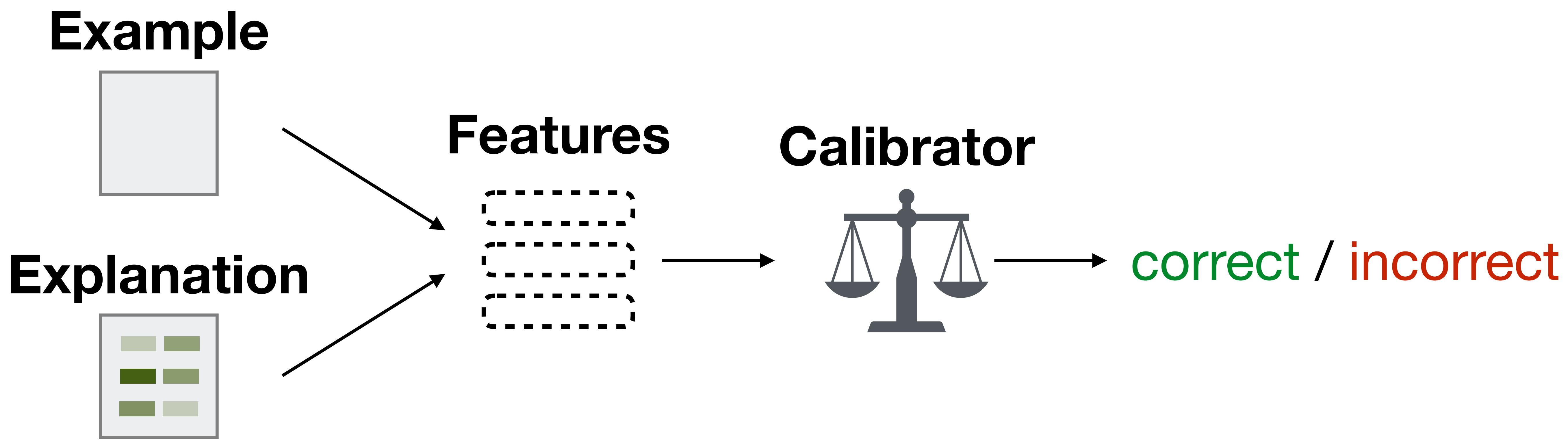
Explanation



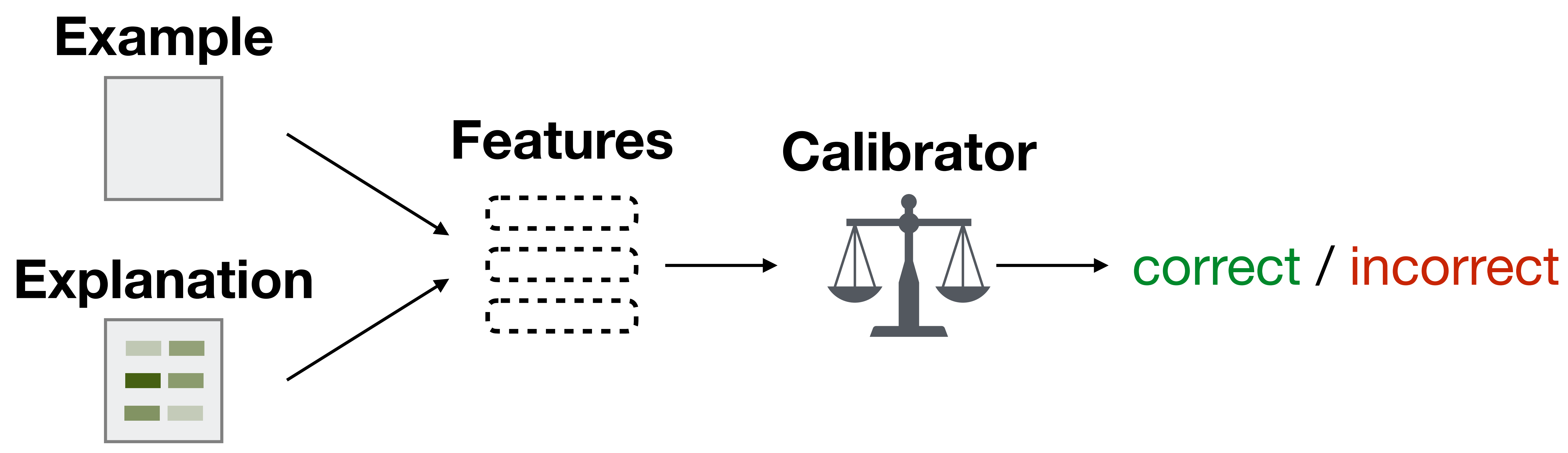
Calibration Framework



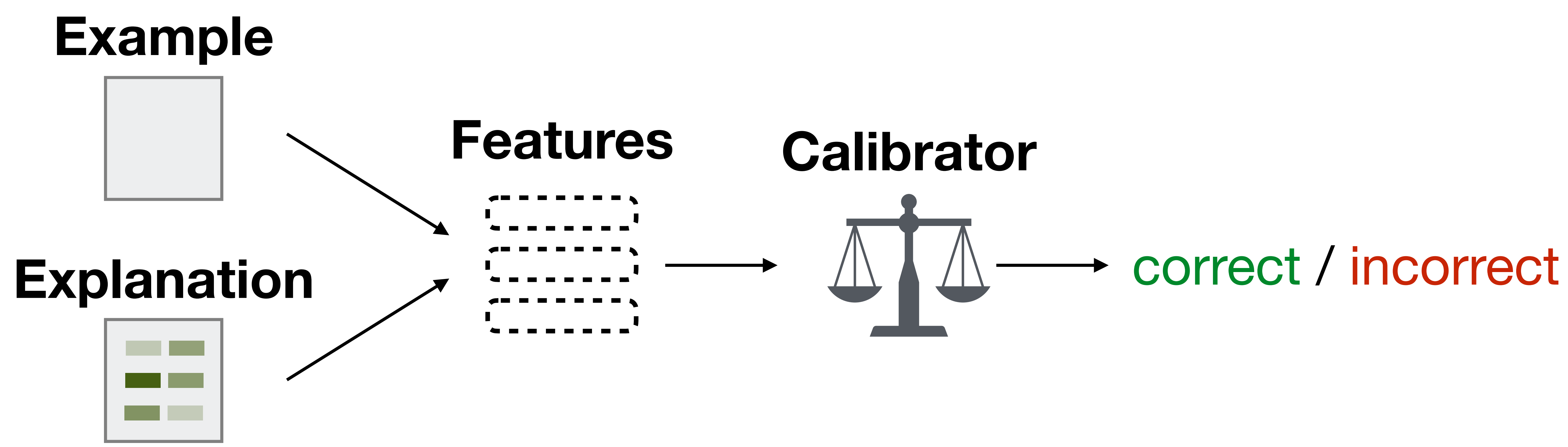
Calibration Framework



Generating Explanations

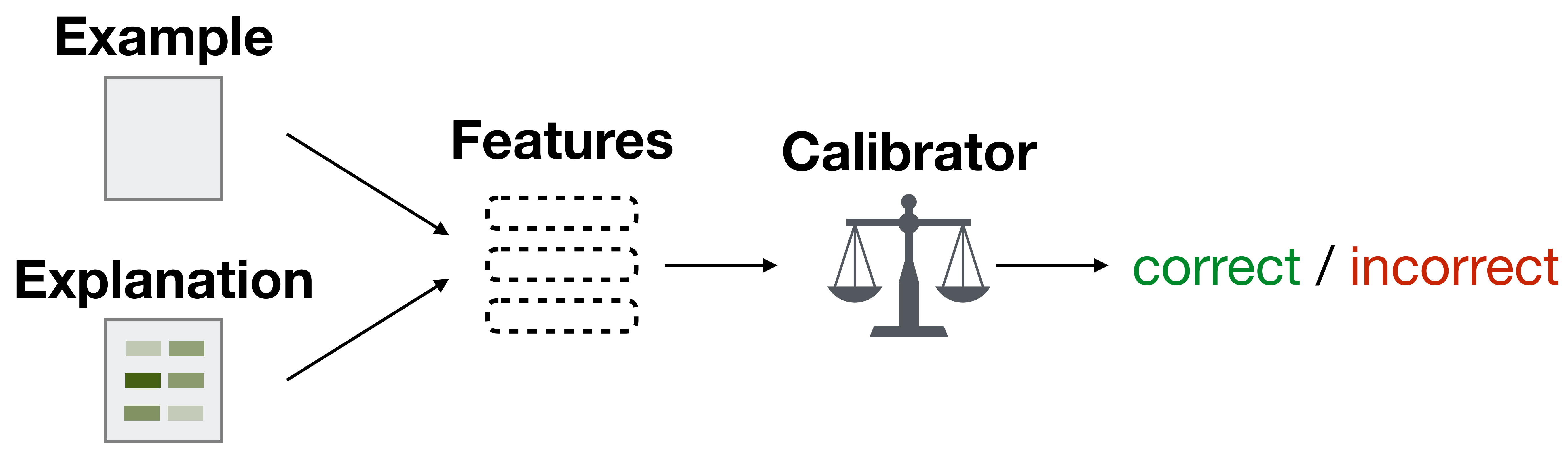


Generating Explanations



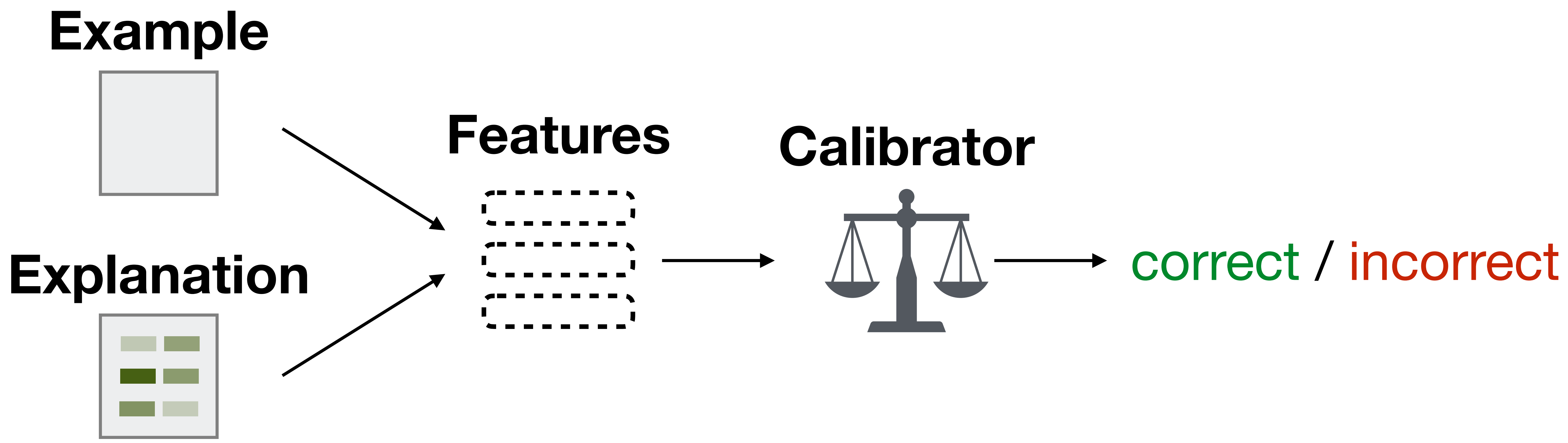
- Use **Lime** and **Shap** to generate interpretations

Generating Explanations

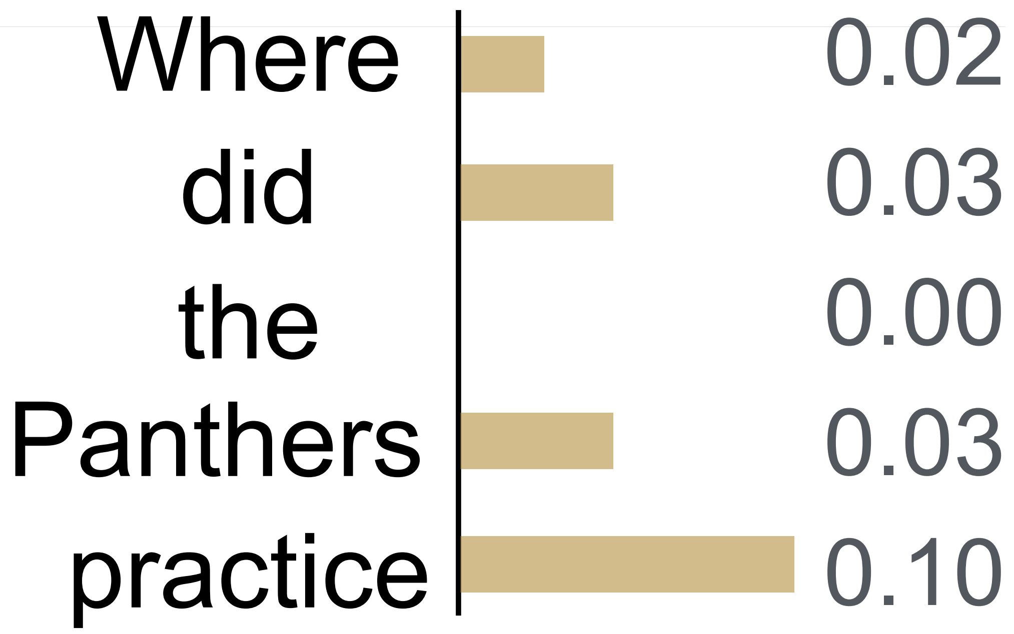


- Use **Lime** and **Shap** to generate interpretations
  - Do not require access to model parameters or gradients

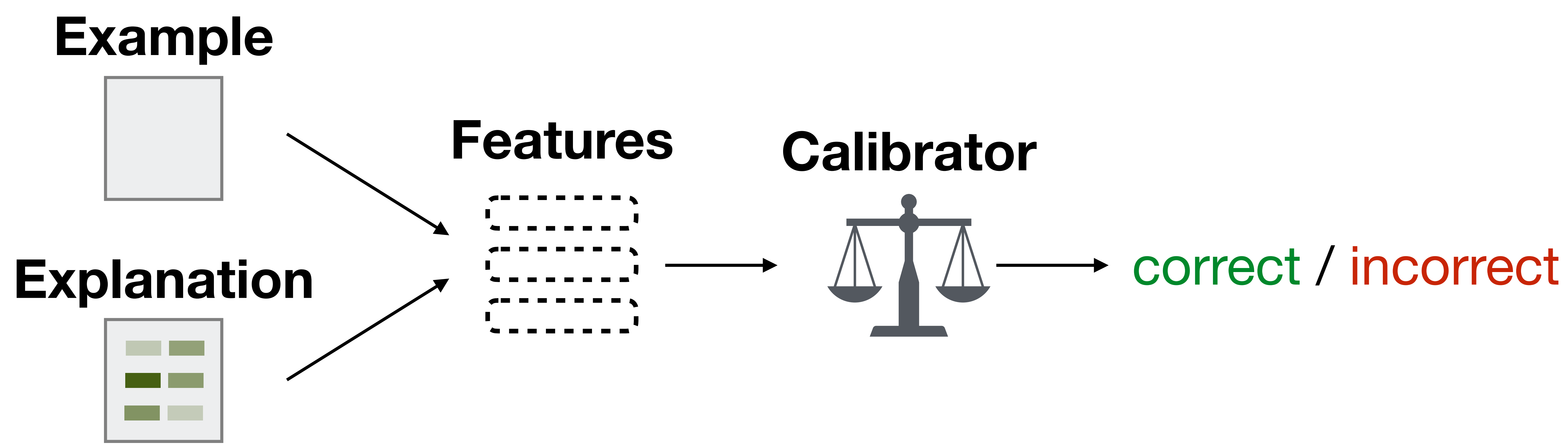
Generating Explanations



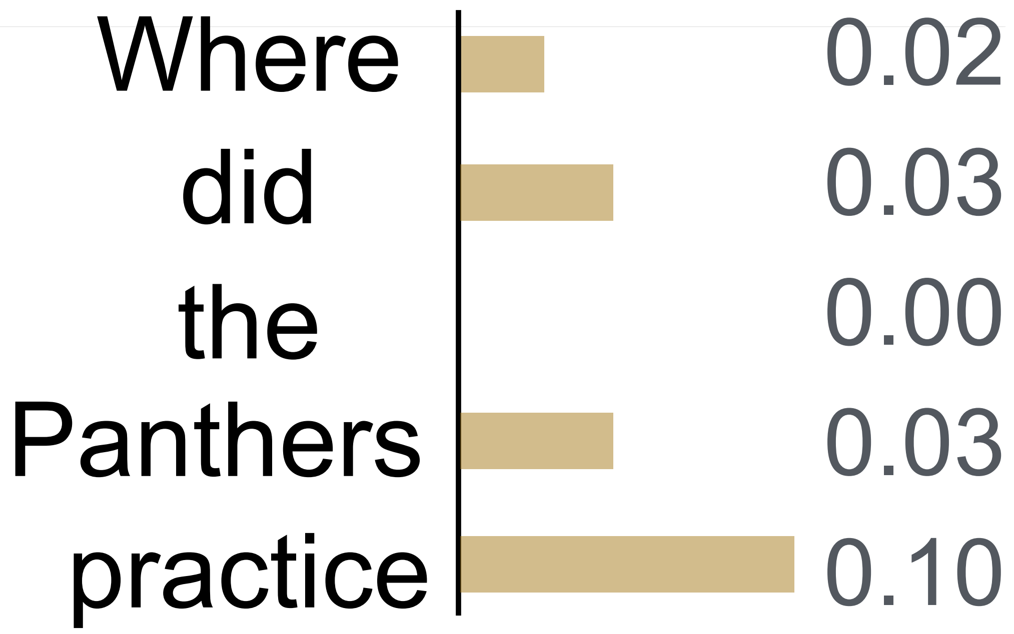
- Use **Lime** and **Shap** to generate interpretations
  - Do not require access to model parameters or gradients



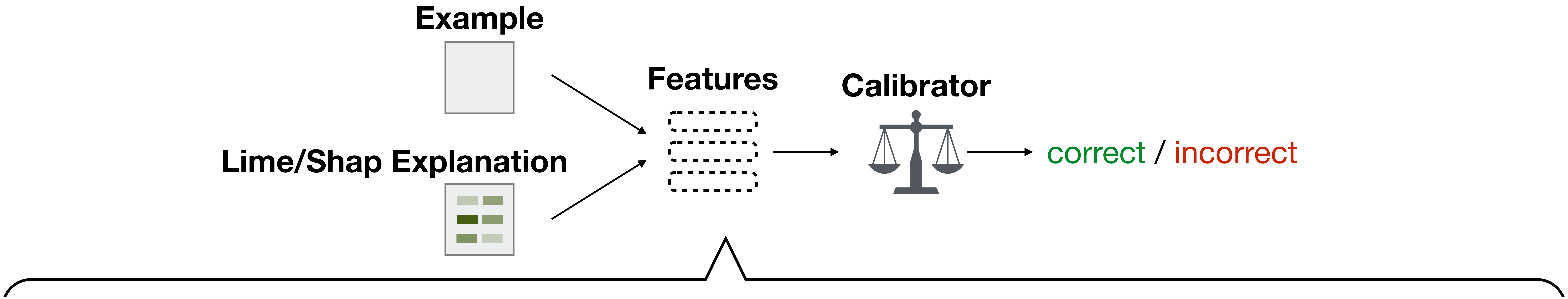
Generating Explanations



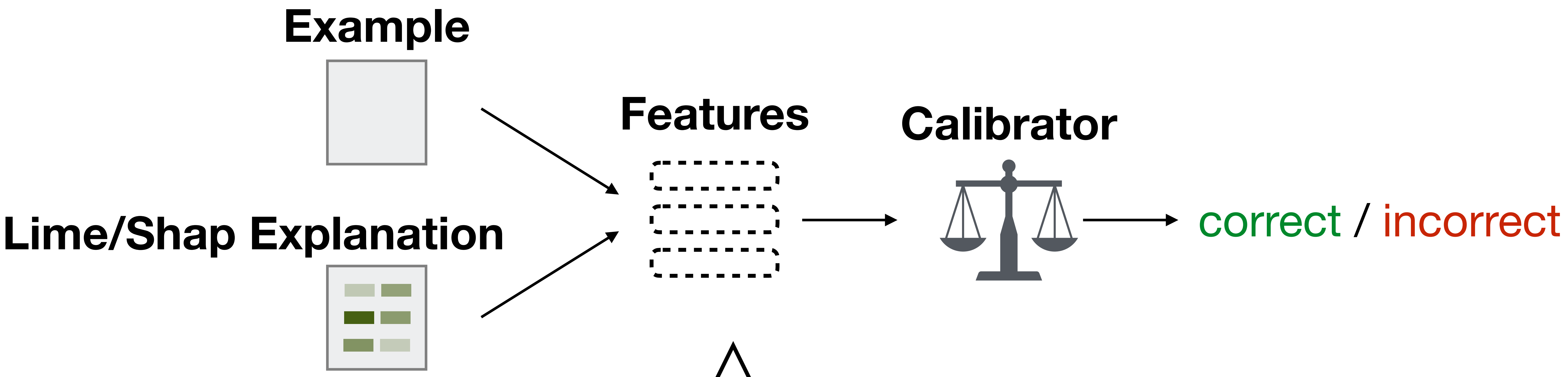
- Use **Lime** and **Shap** to generate interpretations
  - Do not require access to model parameters or gradients
  - Assign an attribution score (importance) to each input token



Extracting Features

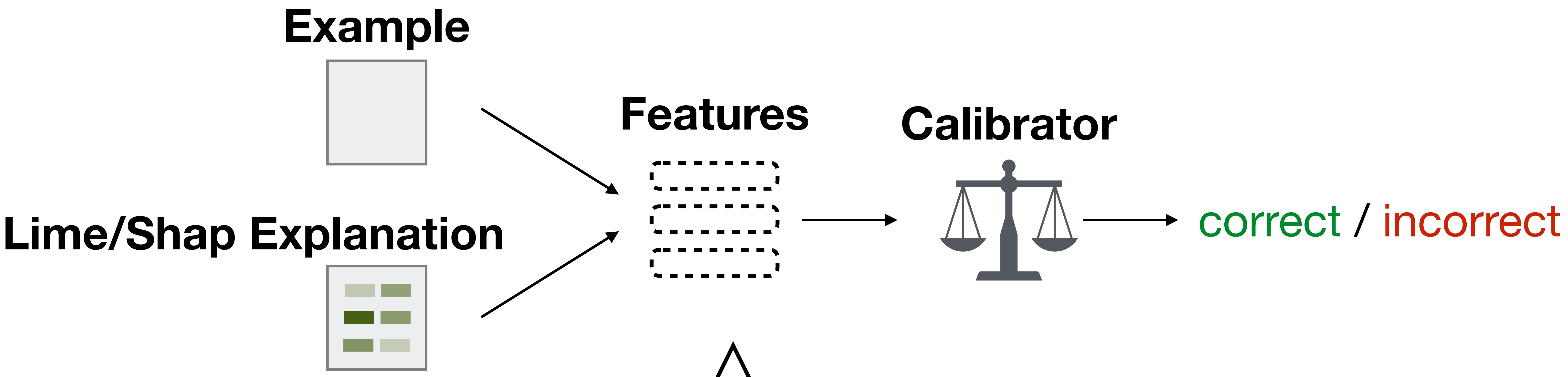


Extracting Features



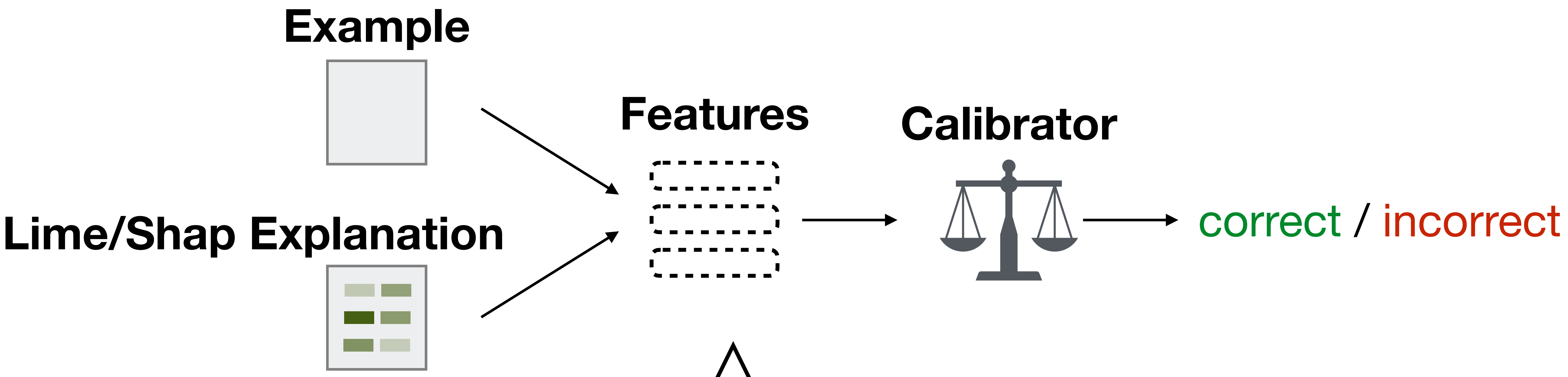
- Assign each token with a set of human-understandable properties (e.g., POS tags)

Extracting Features



- Assign each token with a set of human-understandable properties (e.g., POS tags)
- Extract a numeric feature by aggregating the attributions to tokens associated with each property

Extracting Features



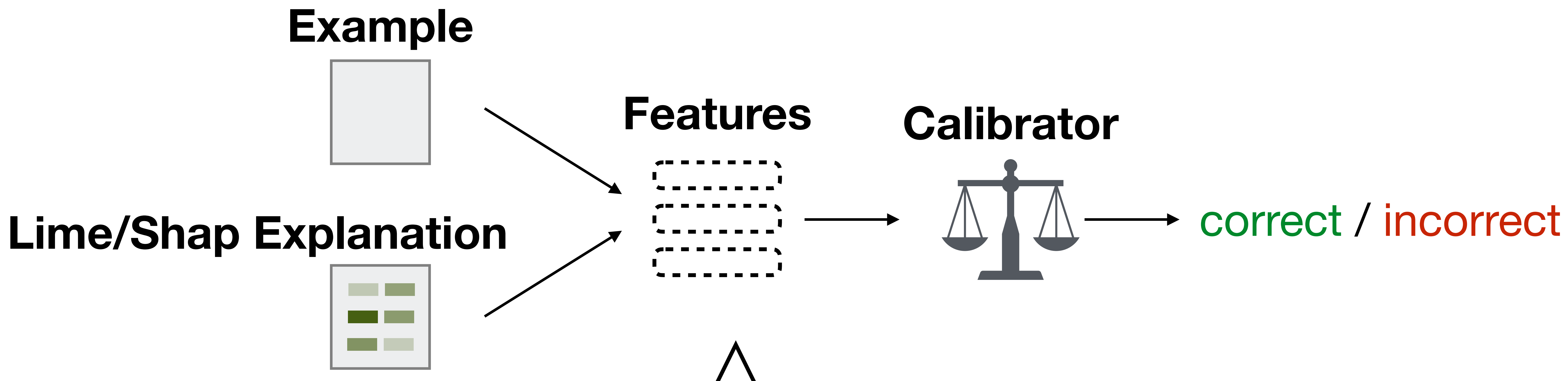
- Assign each token with a set of human-understandable properties (e.g., POS tags)
- Extract a numeric feature by aggregating the attributions to tokens associated with each property

Question *Where did the Panthers practice ?*

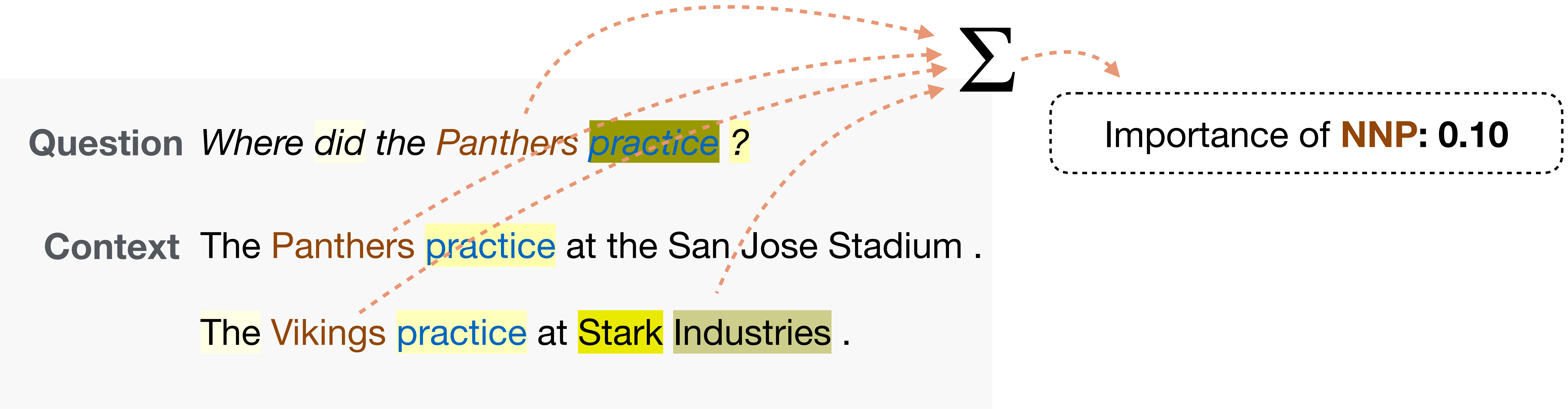
Context The Panthers practice at the San Jose Stadium .

The Vikings practice at Stark Industries .

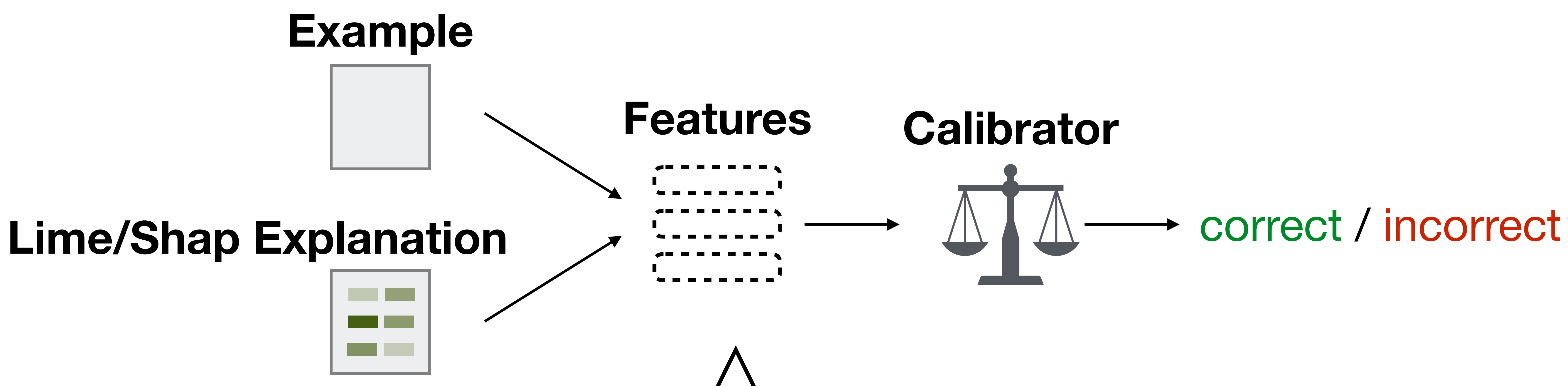
Extracting Features



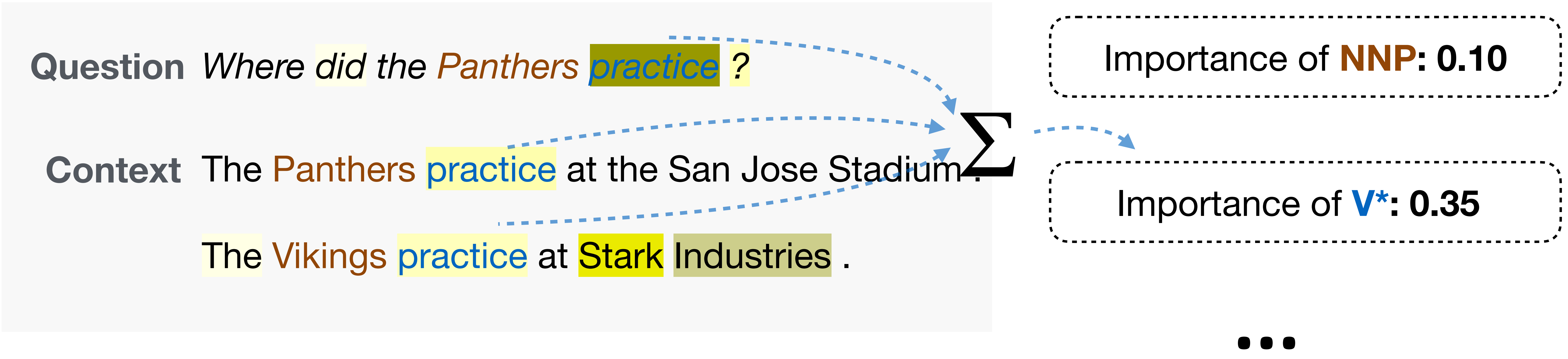
- Assign each token with a set of human-understandable properties (e.g., POS tags)
- Extract a numeric feature by aggregating the attributions to tokens associated with each property



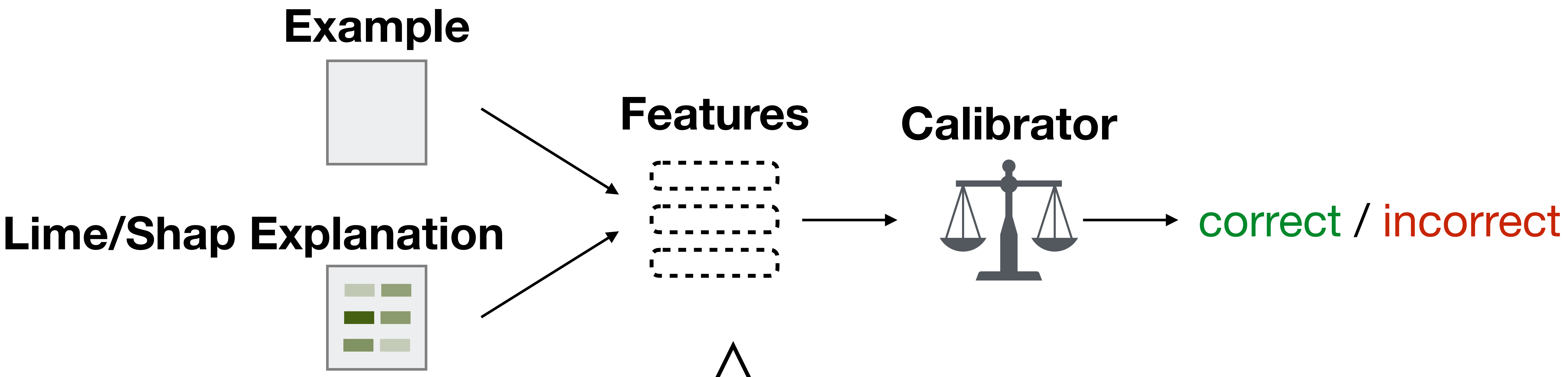
Extracting Features



- Assign each token with a set of human-understandable properties (e.g., POS tags)
- Extract a numeric feature by aggregating the attributions to tokens associated with each property



Extracting Features



- Assign each token with a set of human-understandable properties (e.g., POS tags)
- Extract a numeric feature by aggregating the attributions to tokens associated with each property
- Refer to the paper for details of the features used for calibrating QA and NLI models

Question *Where did the Panthers practice ?*

Context The Panthers practice at the San Jose Stadium .

The Vikings practice at Stark Industries .

Importance of **NNP: 0.10**

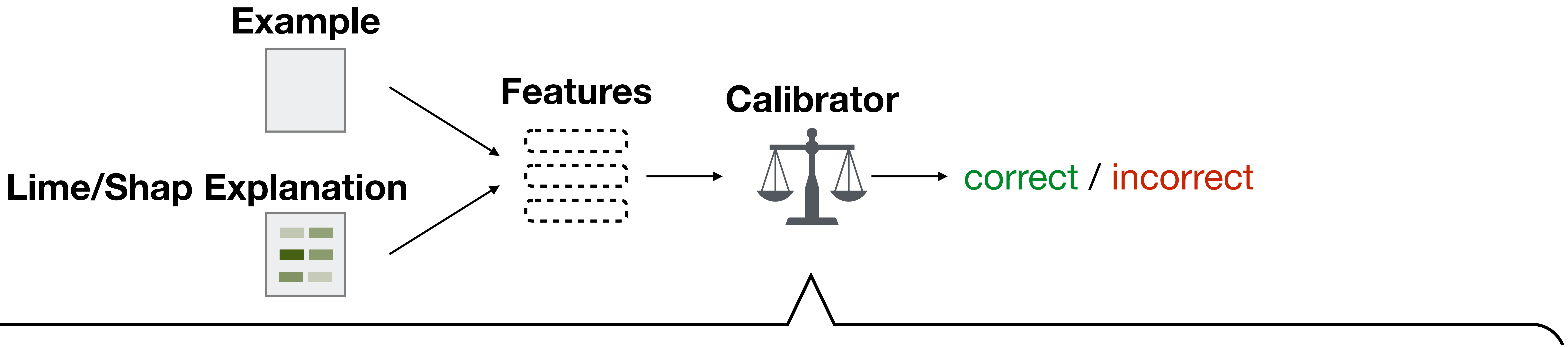
Importance of Question**: 0.27**

Importance of **V*: 0.35**

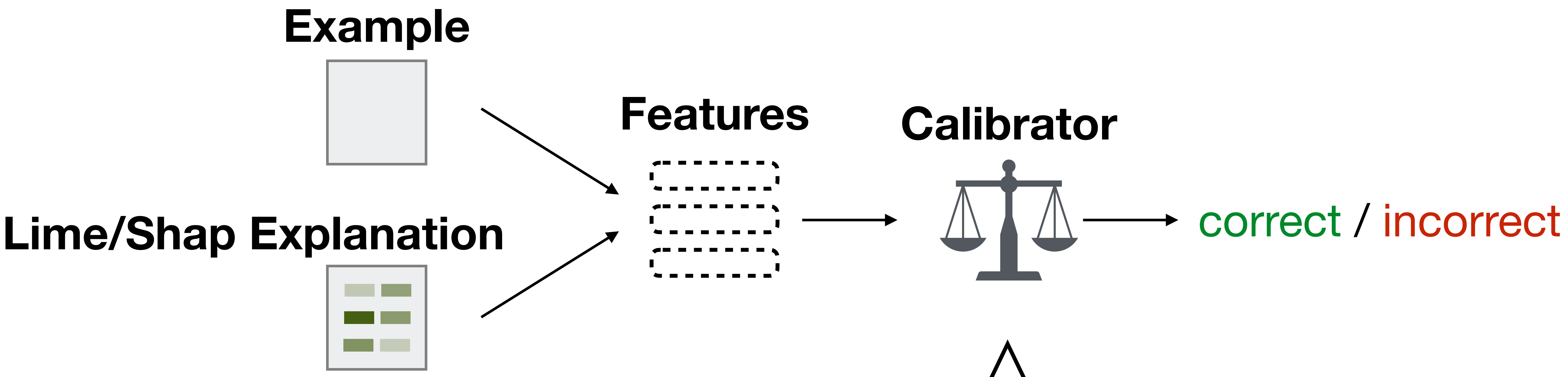
Importance of NNP in Context**: 0.07**

…

Calibrator

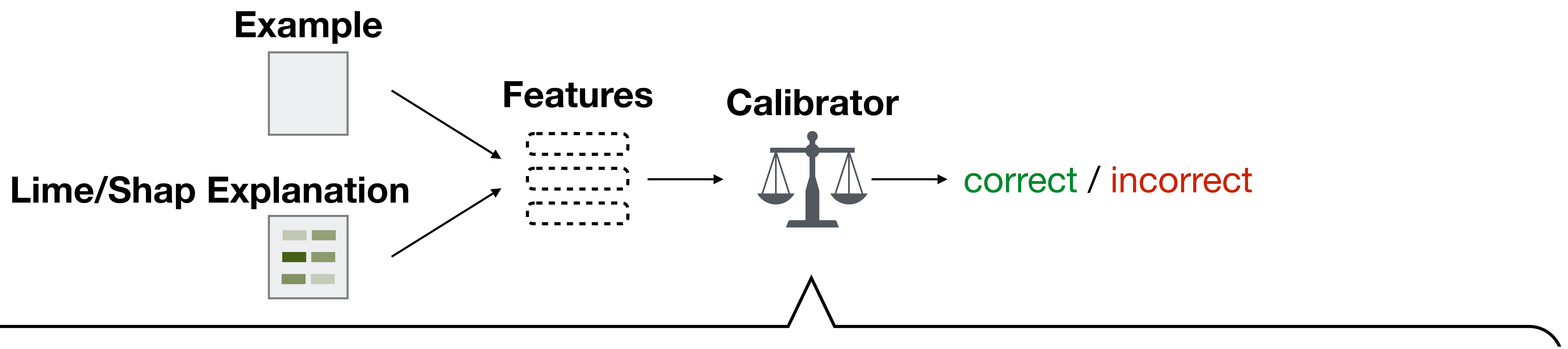

Example
Lime/Shap Explanation
Features
Calibrator
correct / incorrect

Calibrator



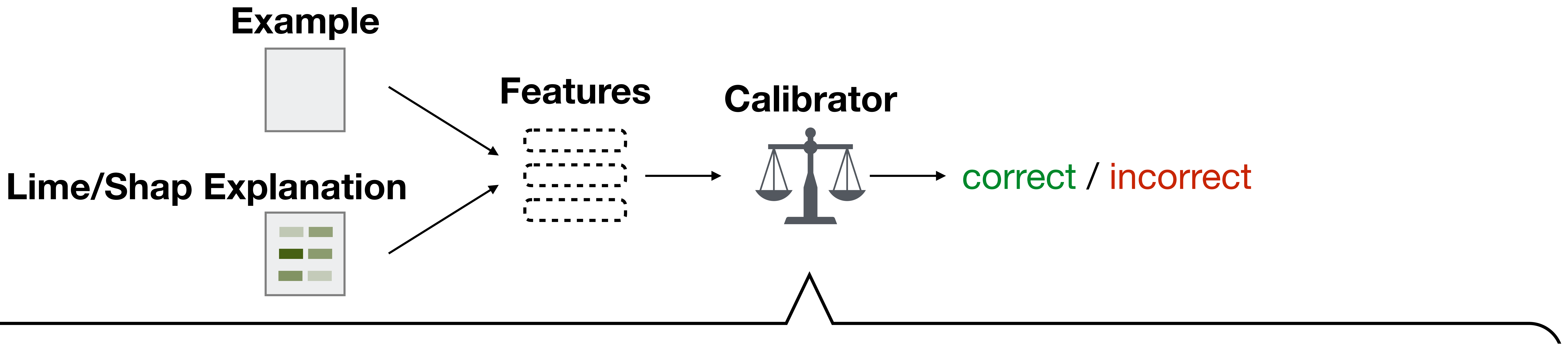
- Use **RandomForest** as the model class of calibrators

Calibrator

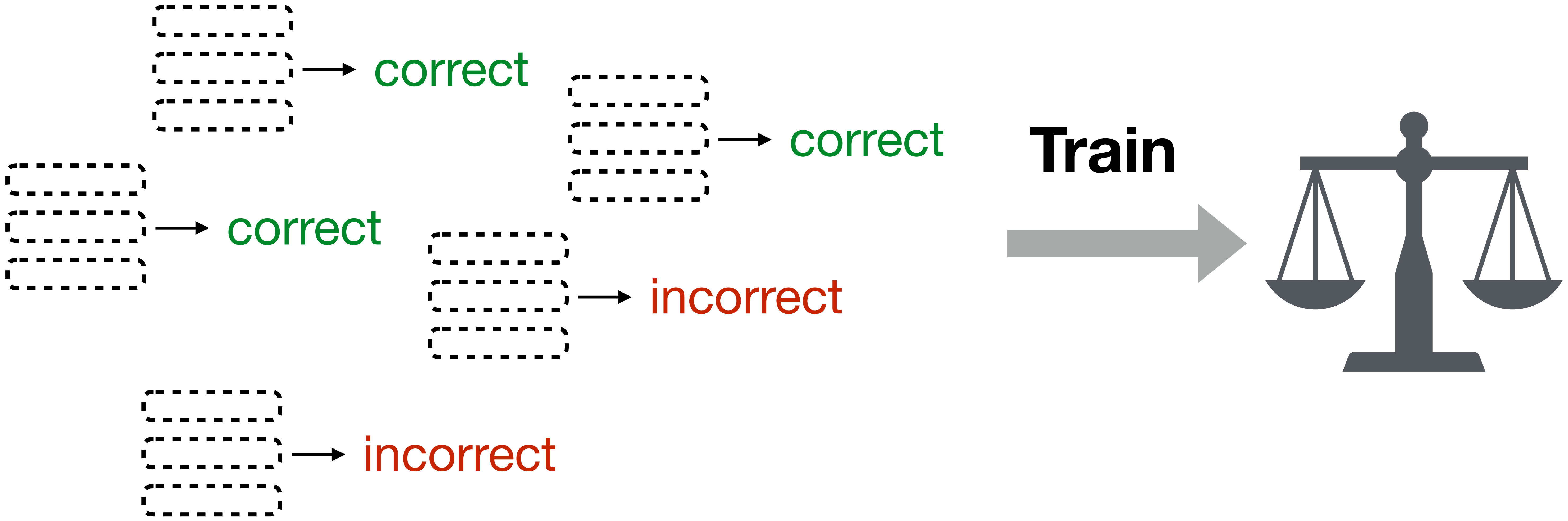


- Use **RandomForest** as the model class of calibrators
- Train the calibrator using a small number of feature-correctness pairs from the target domain

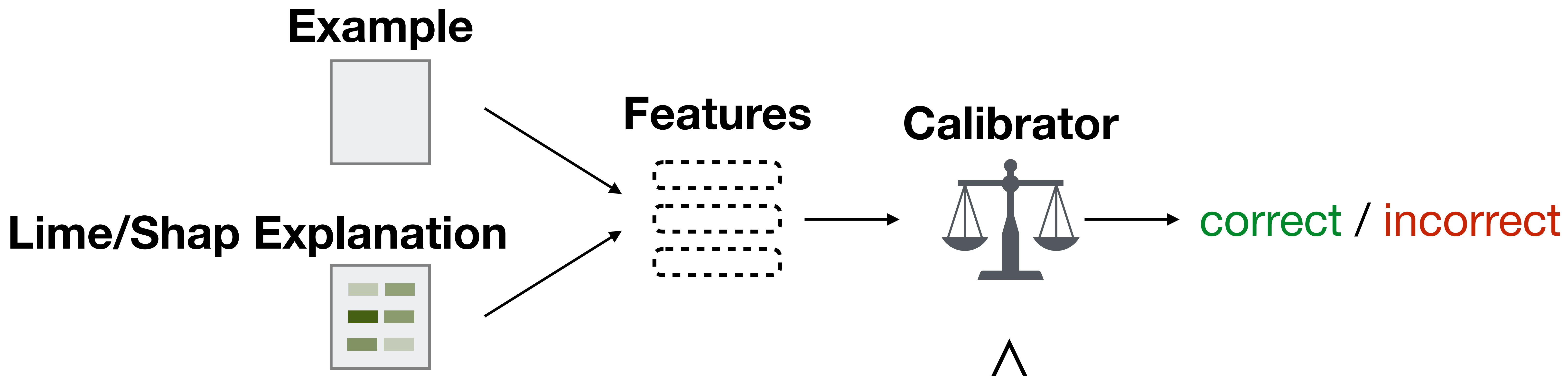
Calibrator



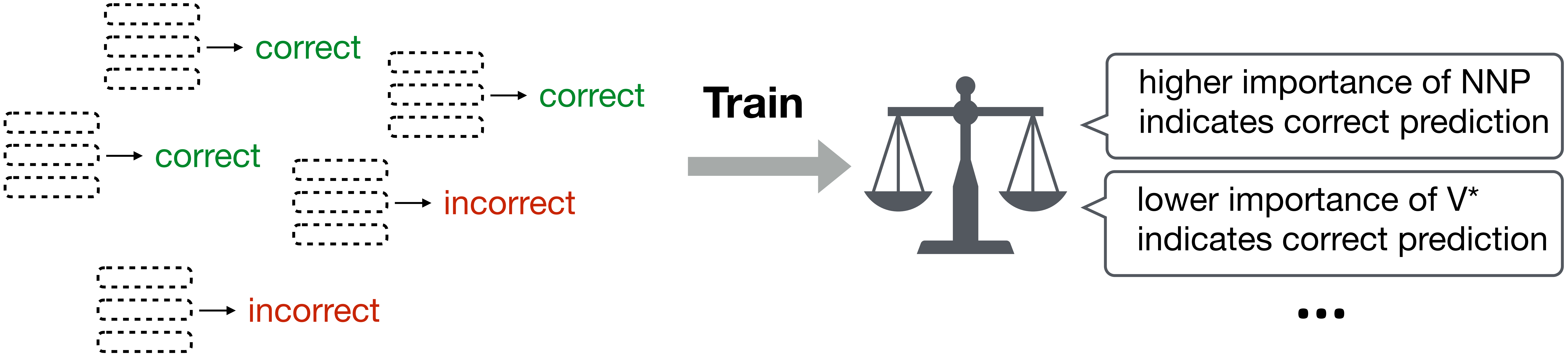
- Use **RandomForest** as the model class of calibrators
- Train the calibrator using a small number of feature-correctness pairs from the target domain



Calibrator

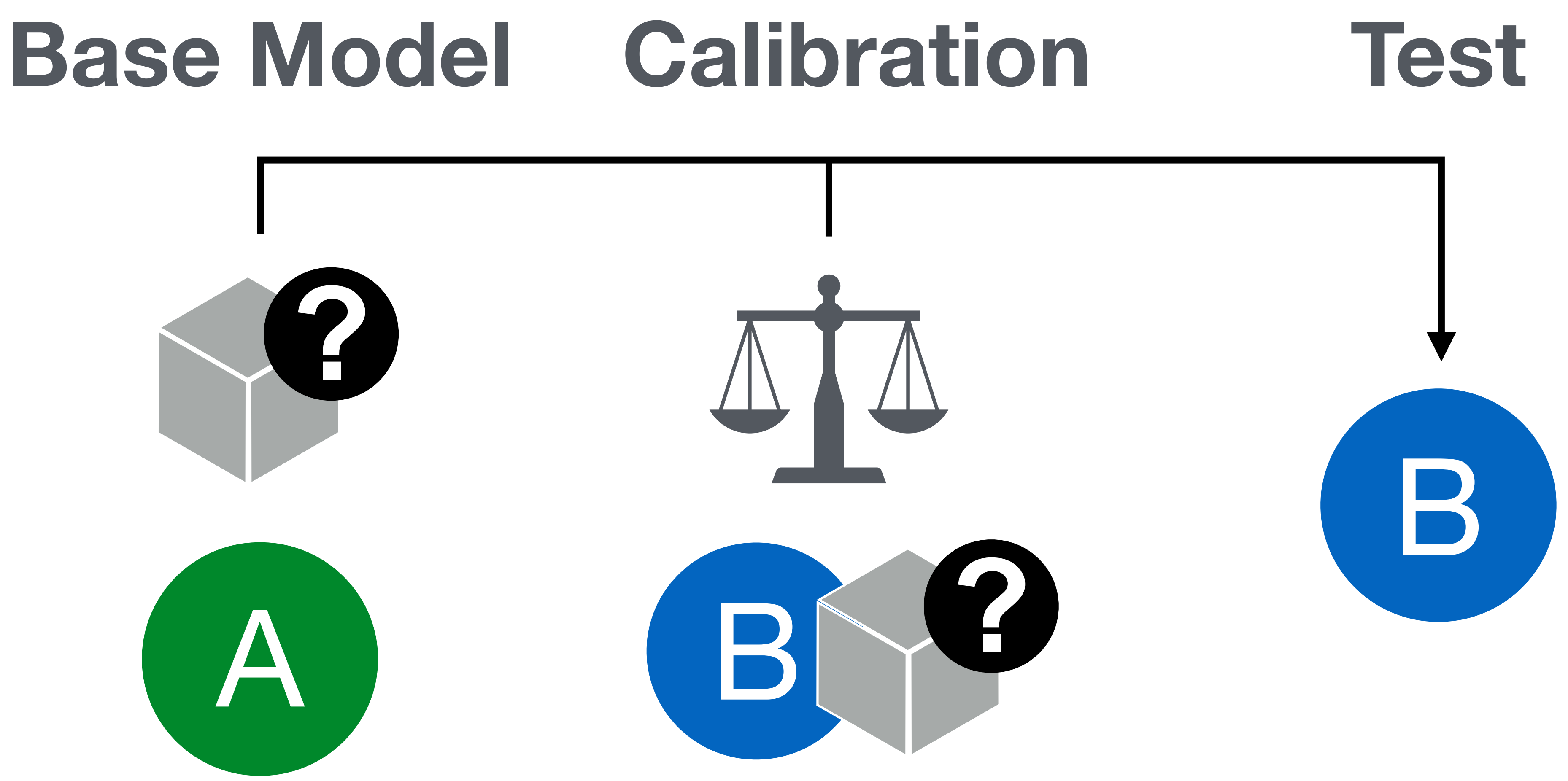


- Use **RandomForest** as the model class of calibrators
- Train the calibrator using a small number of feature-correctness pairs from the target domain

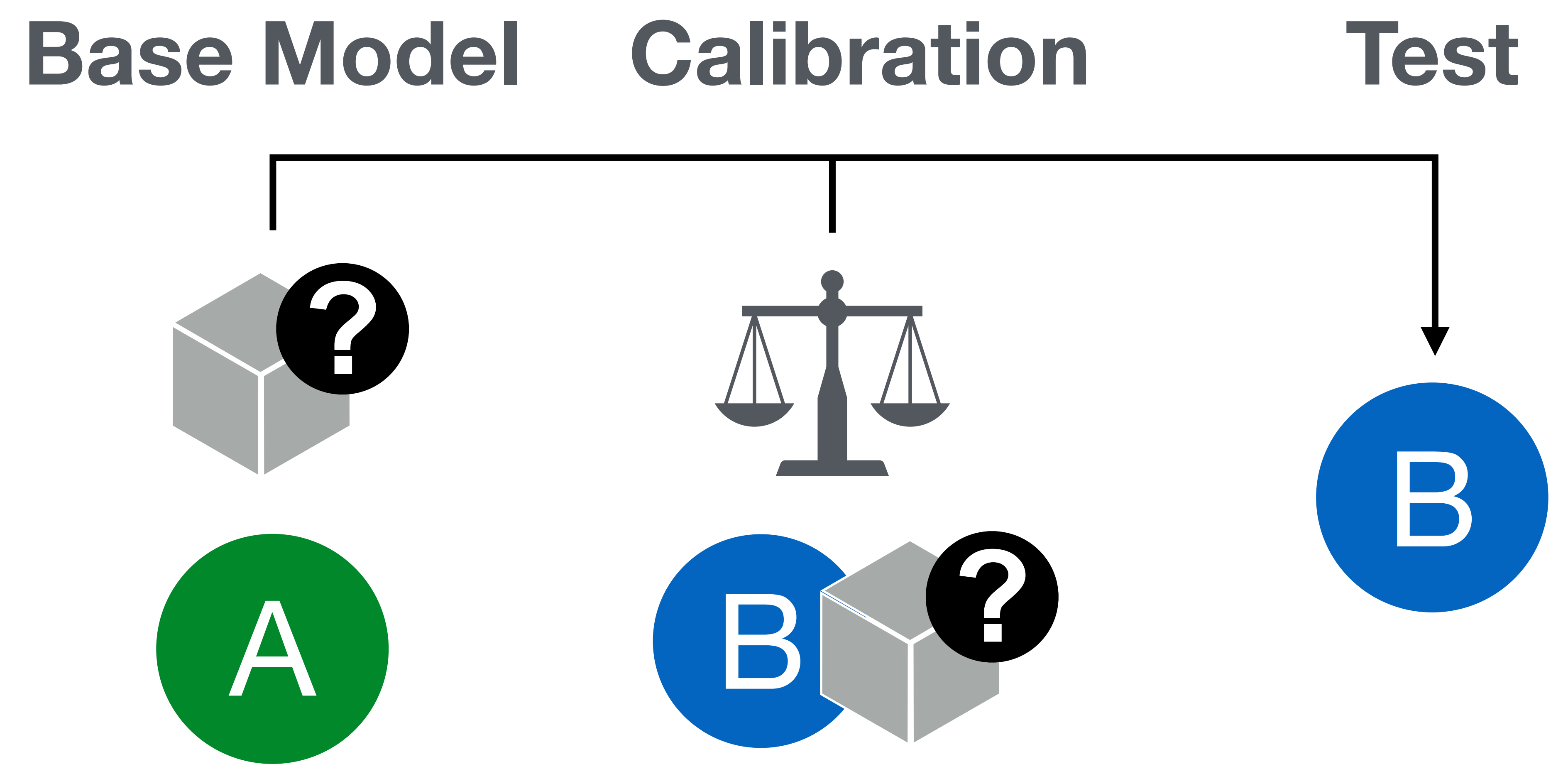


Experiments

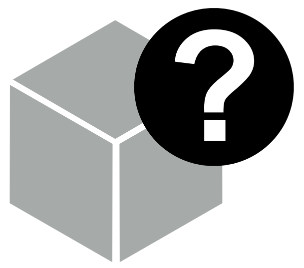
Experiments



Experiments

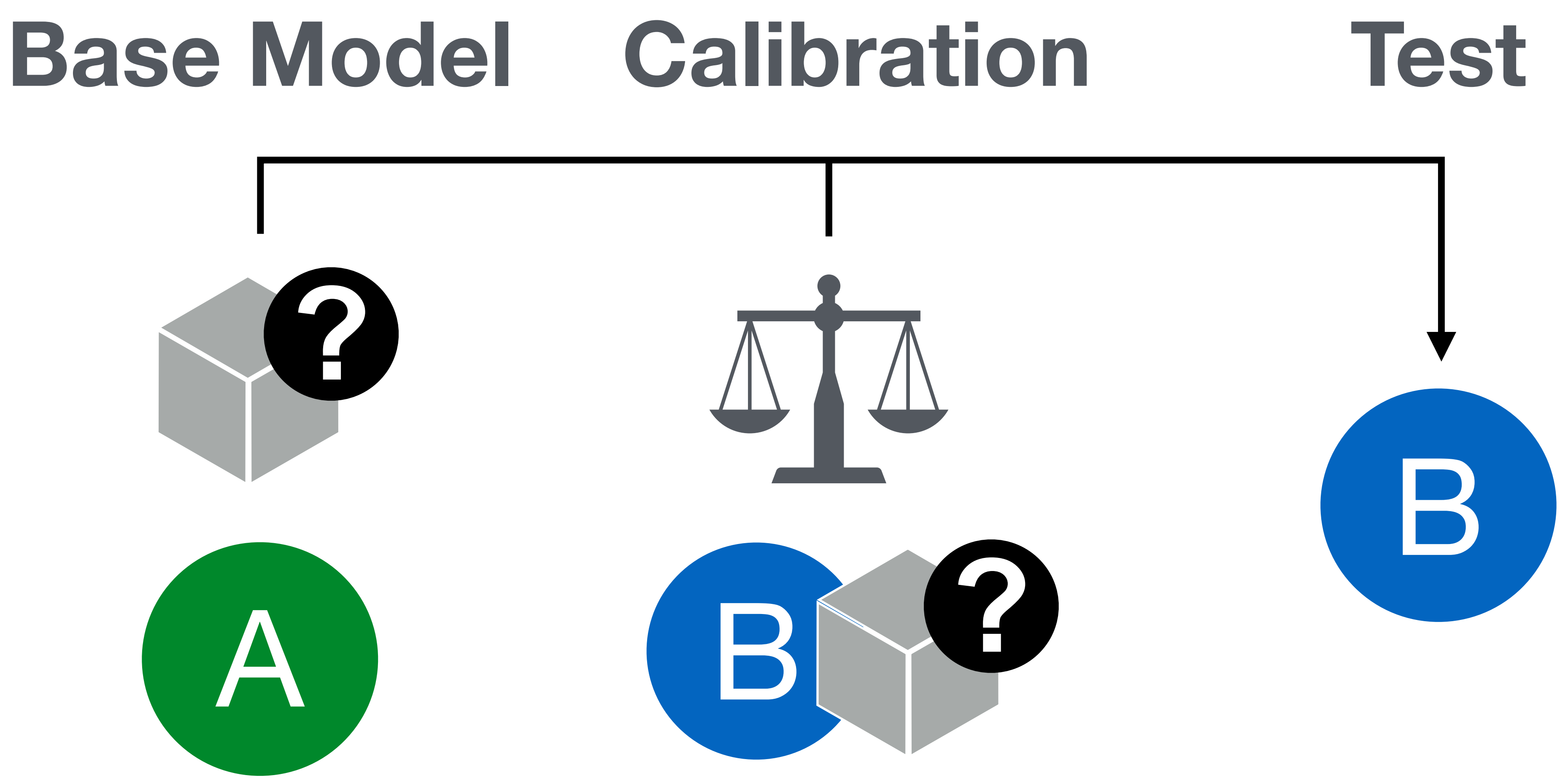


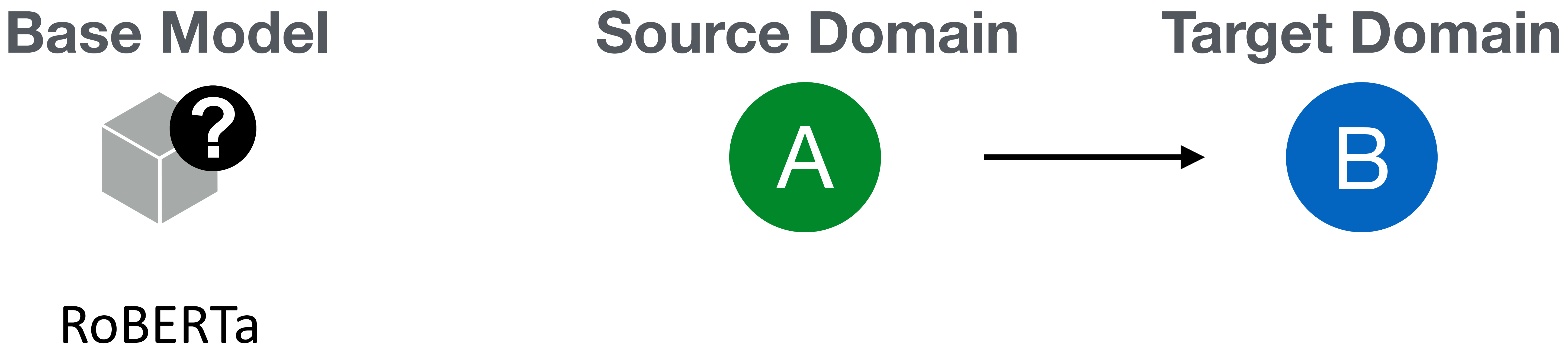
Base Model



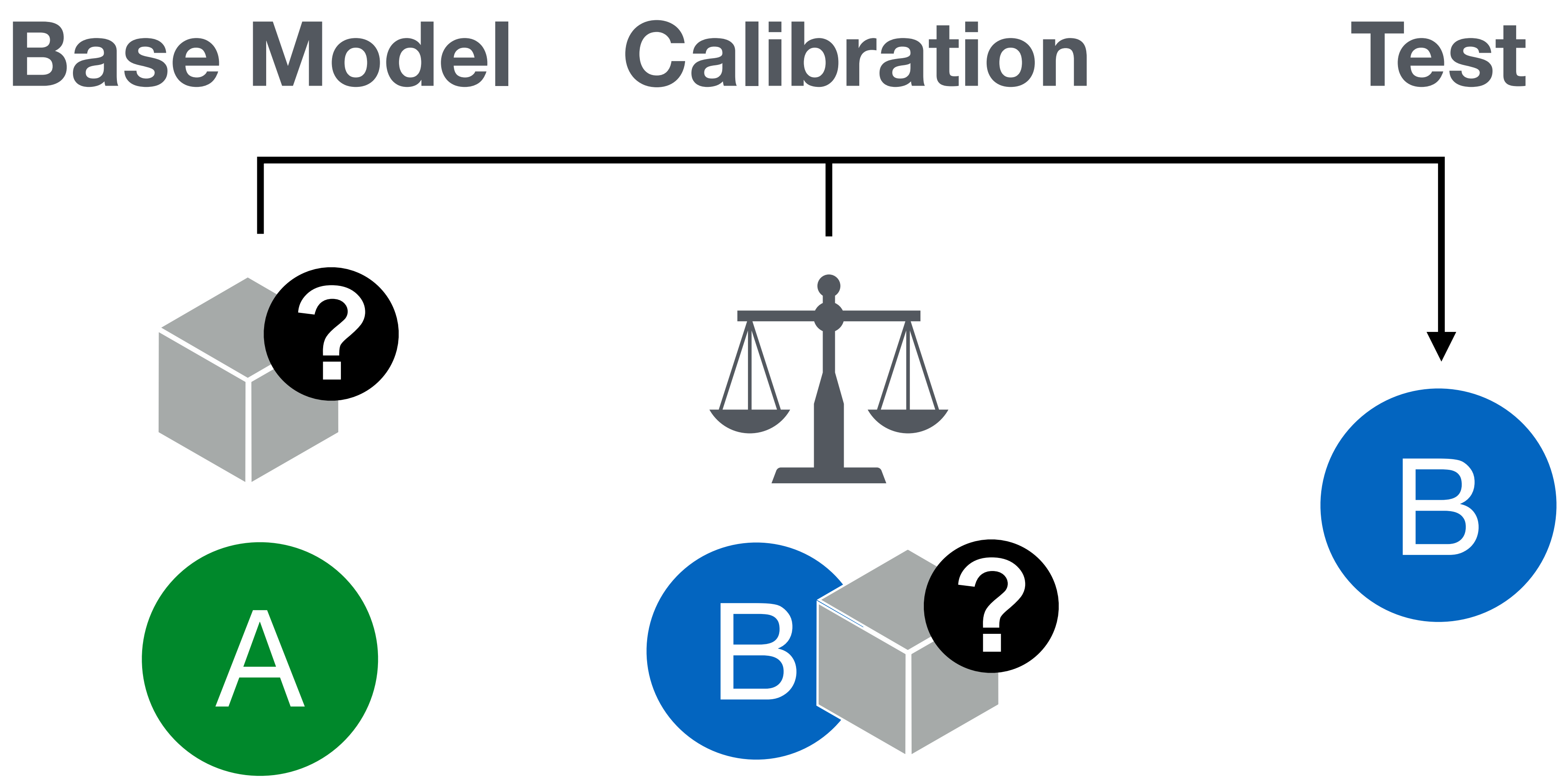
RoBERTa

Experiments





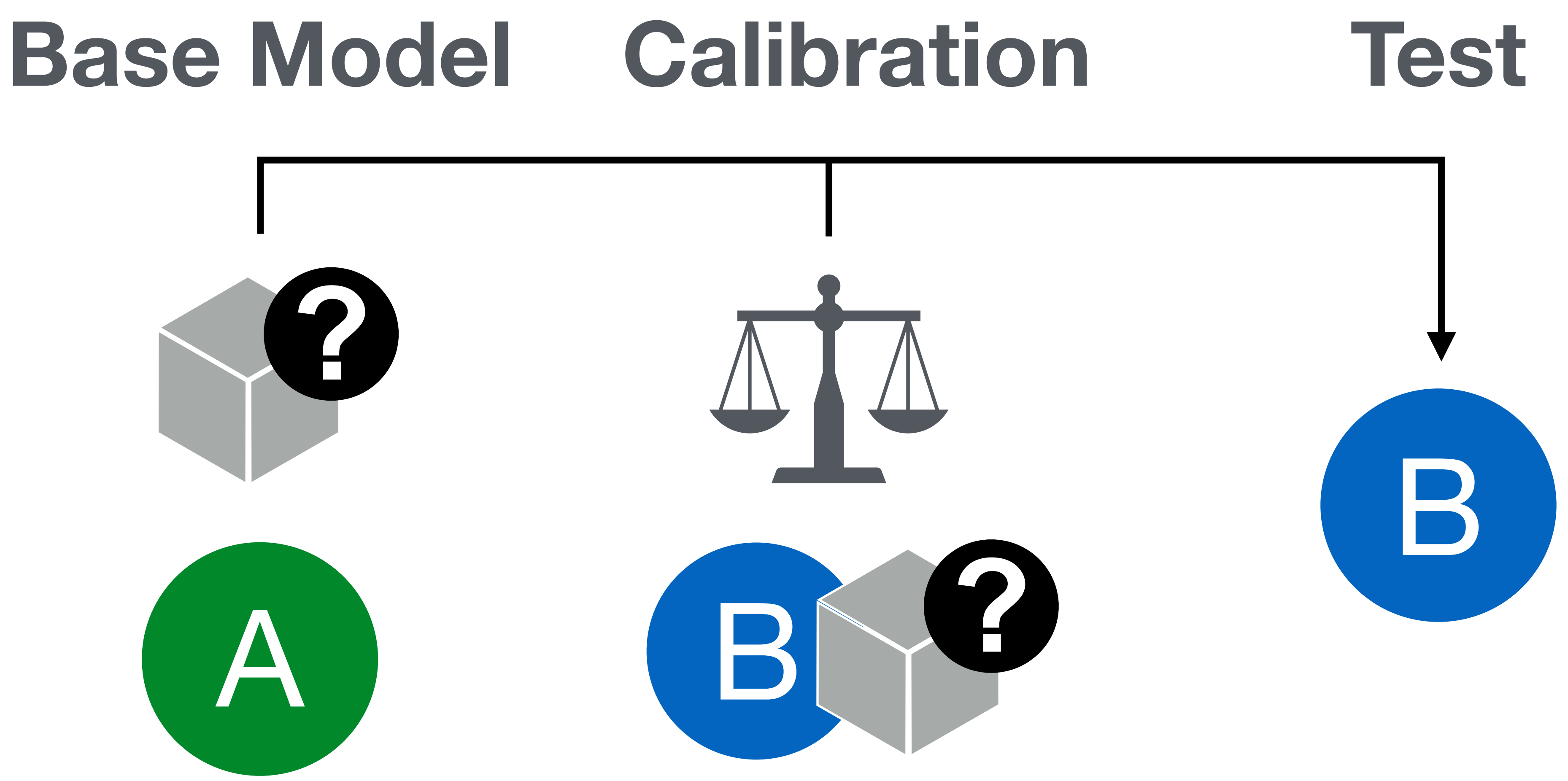
Experiments



RoBERTa

QA: SQuAD → SQuAD-Adv, TriviaQA, HotpotQA

Experiments



Base Model

RoBERTa

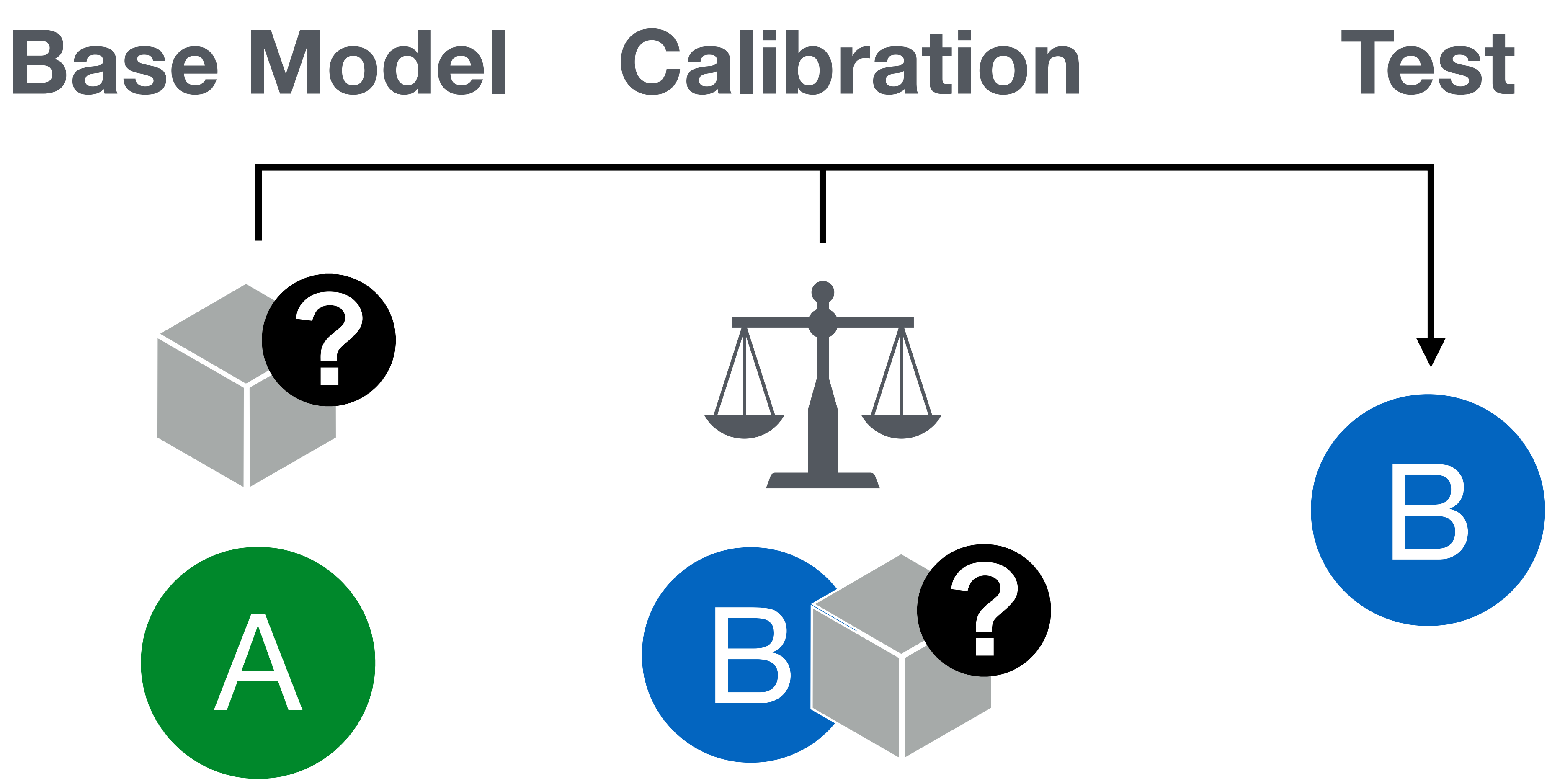
Source Domain → **Target Domain**

A → B

QA: SQuAD → SQuAD-Adv, TriviaQA, HotpotQA

NLI: MNLI → QNLI, MRPC

Experiments



Base Model

RoBERTa

Source Domain → **Target Domain**

A → B

QA: SQuAD → SQuAD-Adv, TriviaQA, HotpotQA

NLI: MNLI → QNLI, MRPC

Calibrator

RandomForest trained using 500 data points

Experiments (Cont'd)

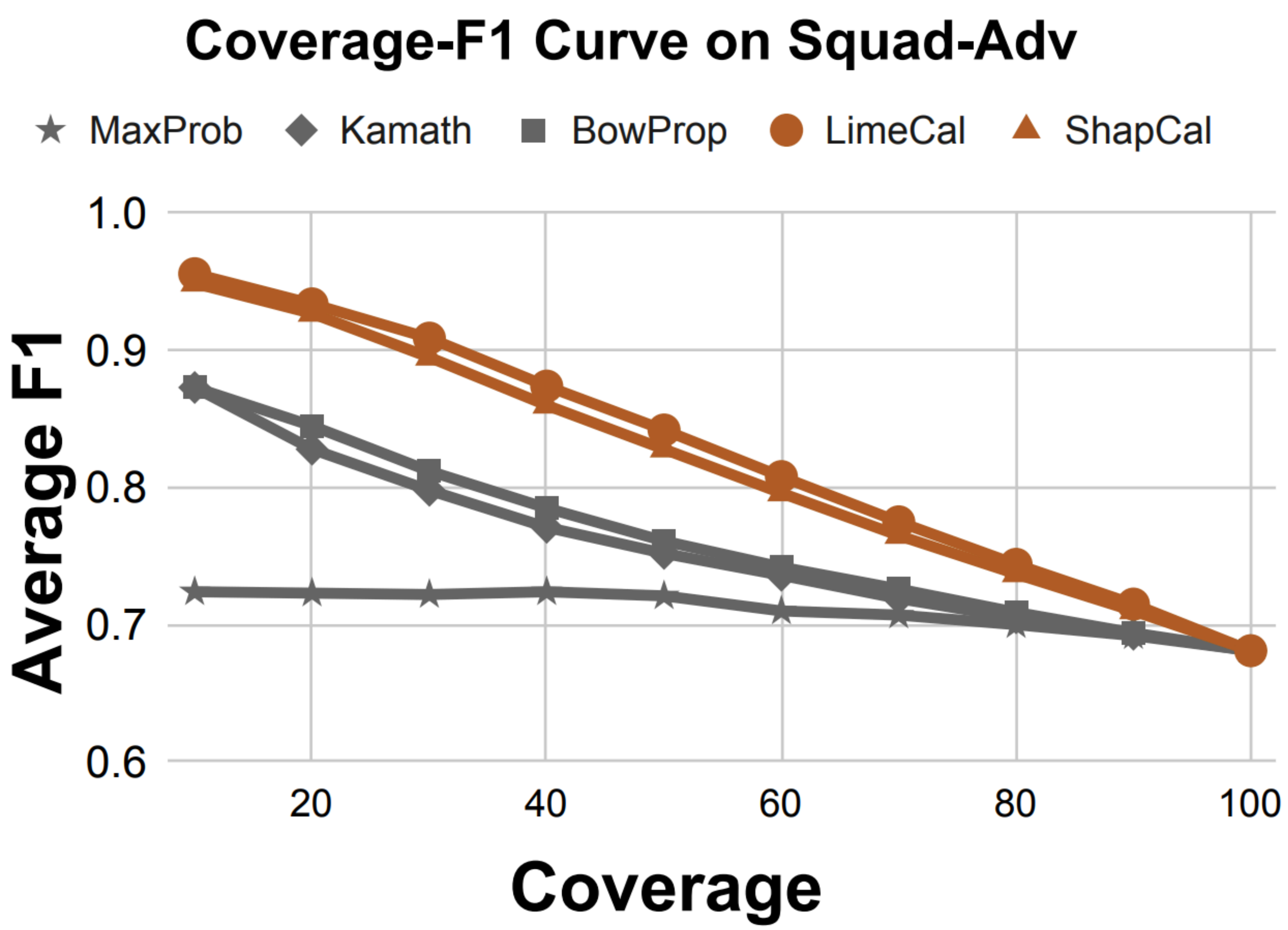
Experiments (Cont'd)

Metrics

Experiments (Cont'd)

Metrics

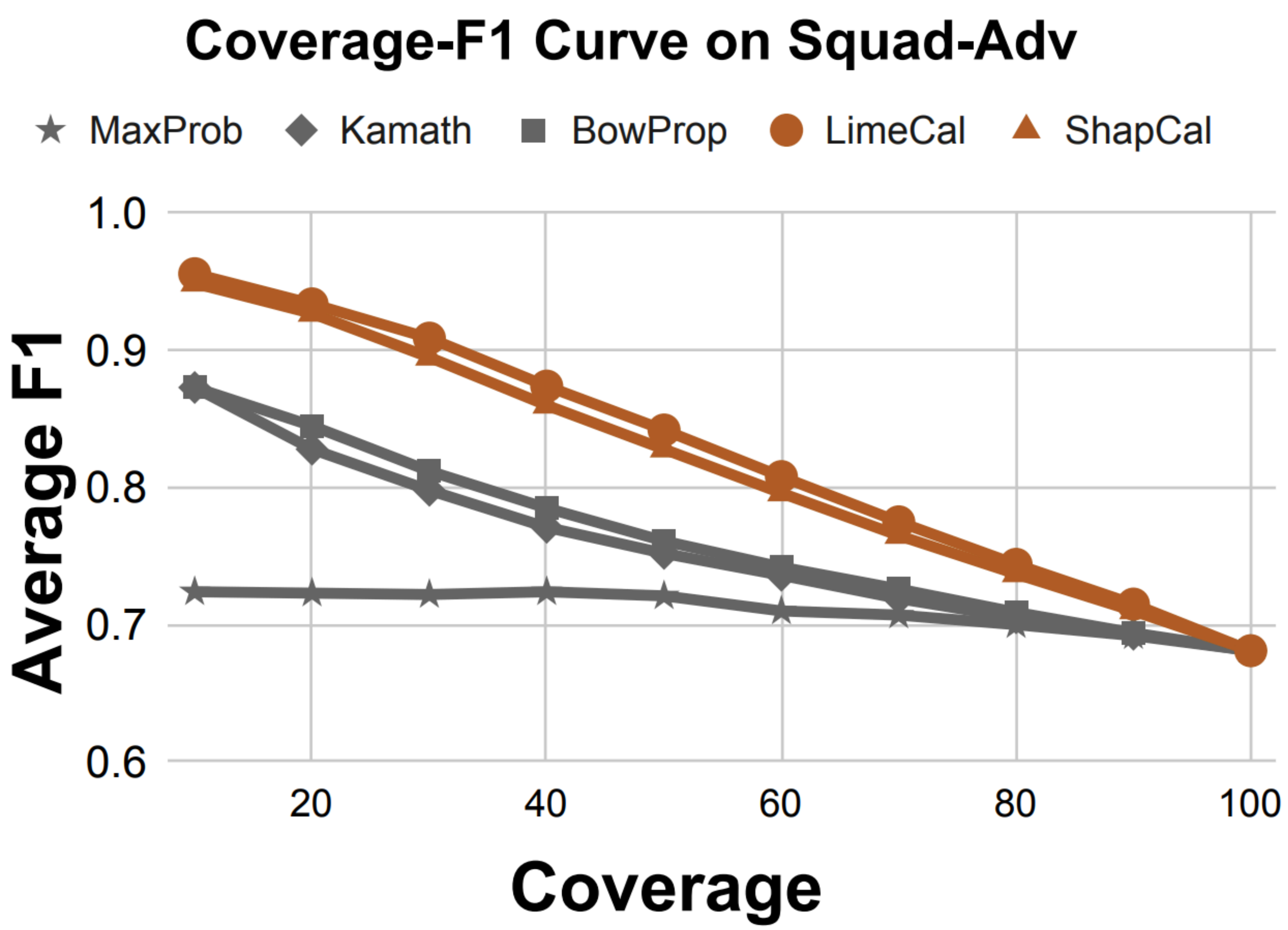
- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)



Experiments (Cont'd)

Metrics

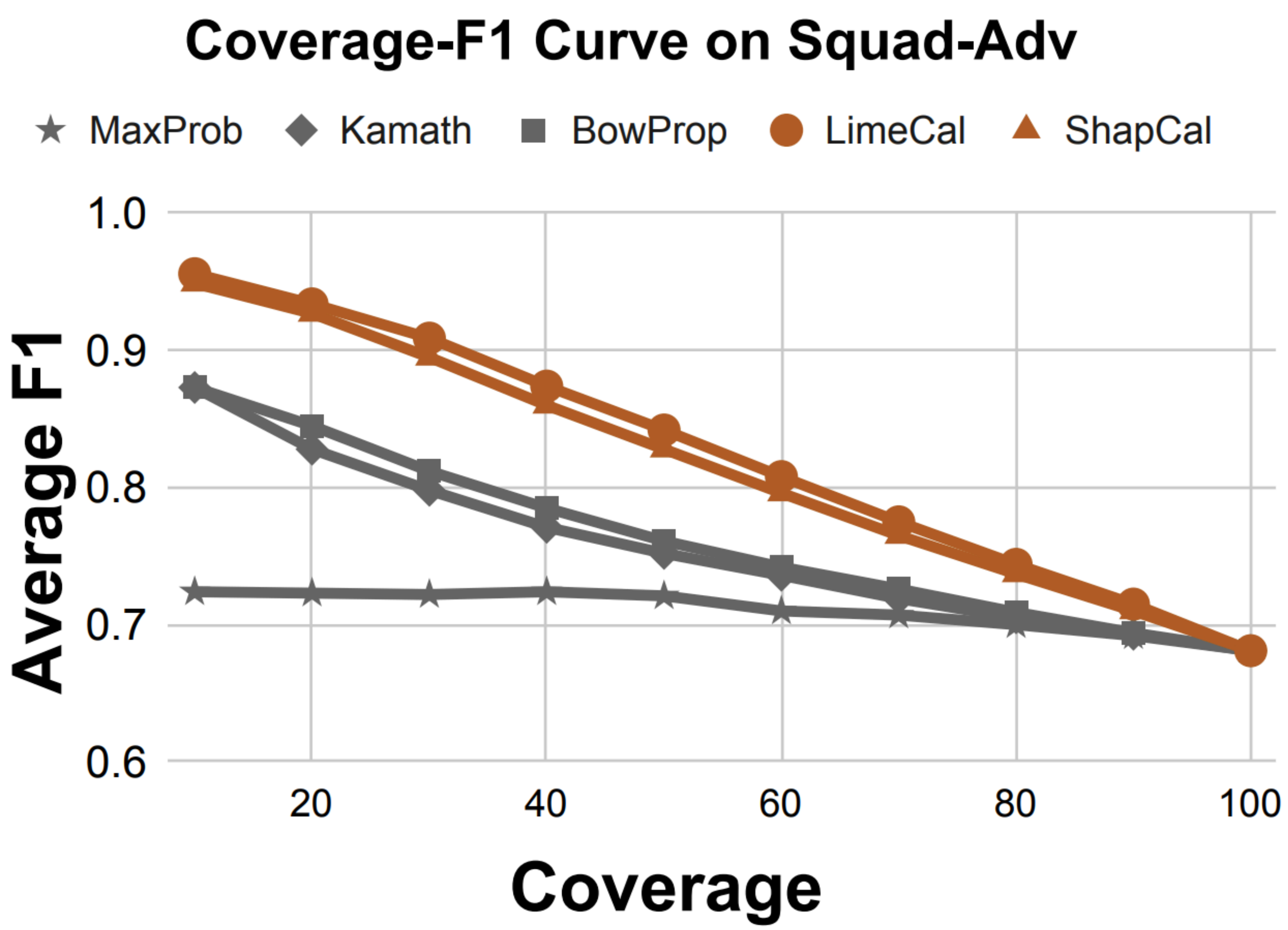
- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)
- Evaluate using area under coverage-F1 curve (**AUC**)



Experiments (Cont’d)

Metrics

- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)
- Evaluate using area under coverage-F1 curve (**AUC**)

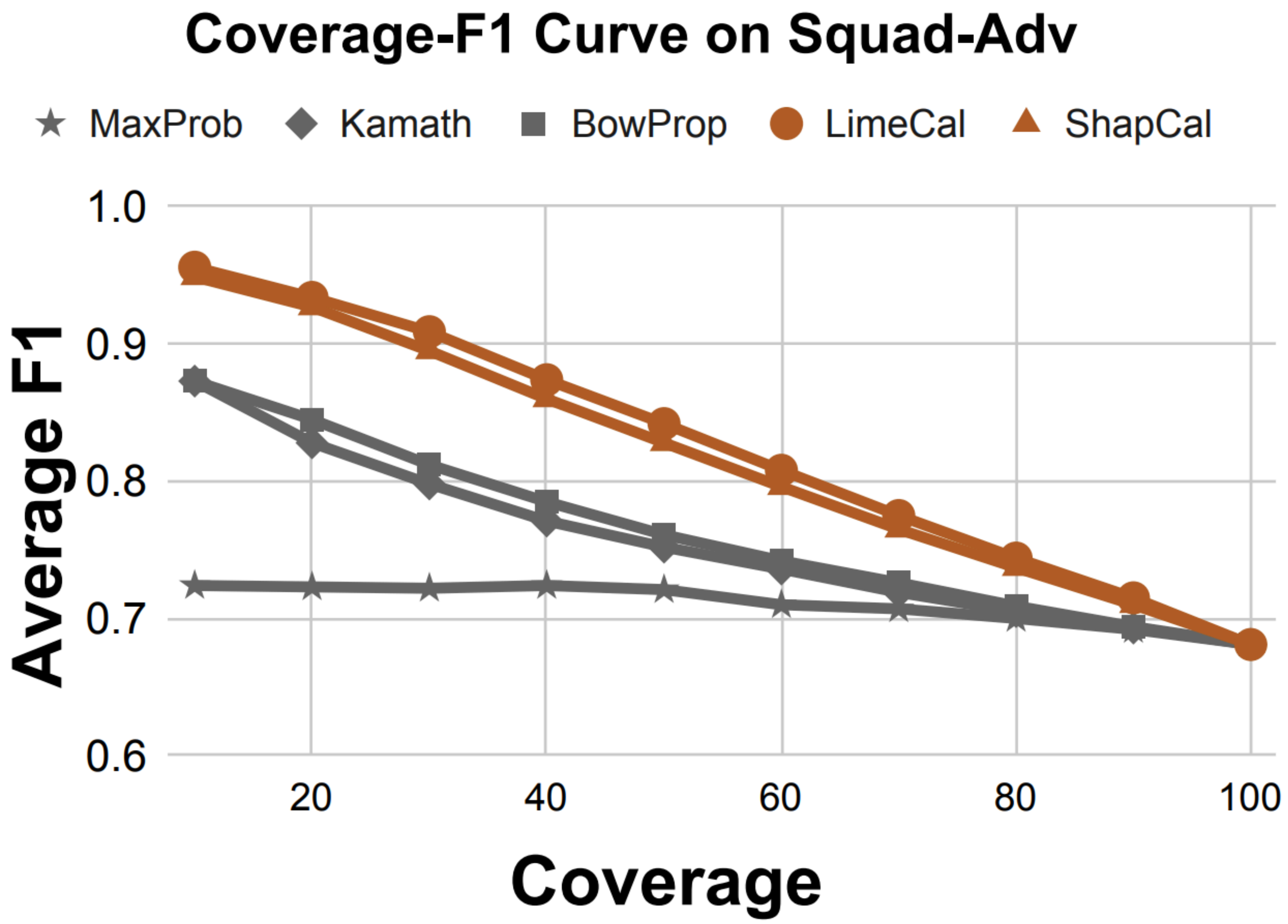


Methods

Experiments (Cont'd)

Metrics

- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)
- Evaluate using area under coverage-F1 curve (**AUC**)



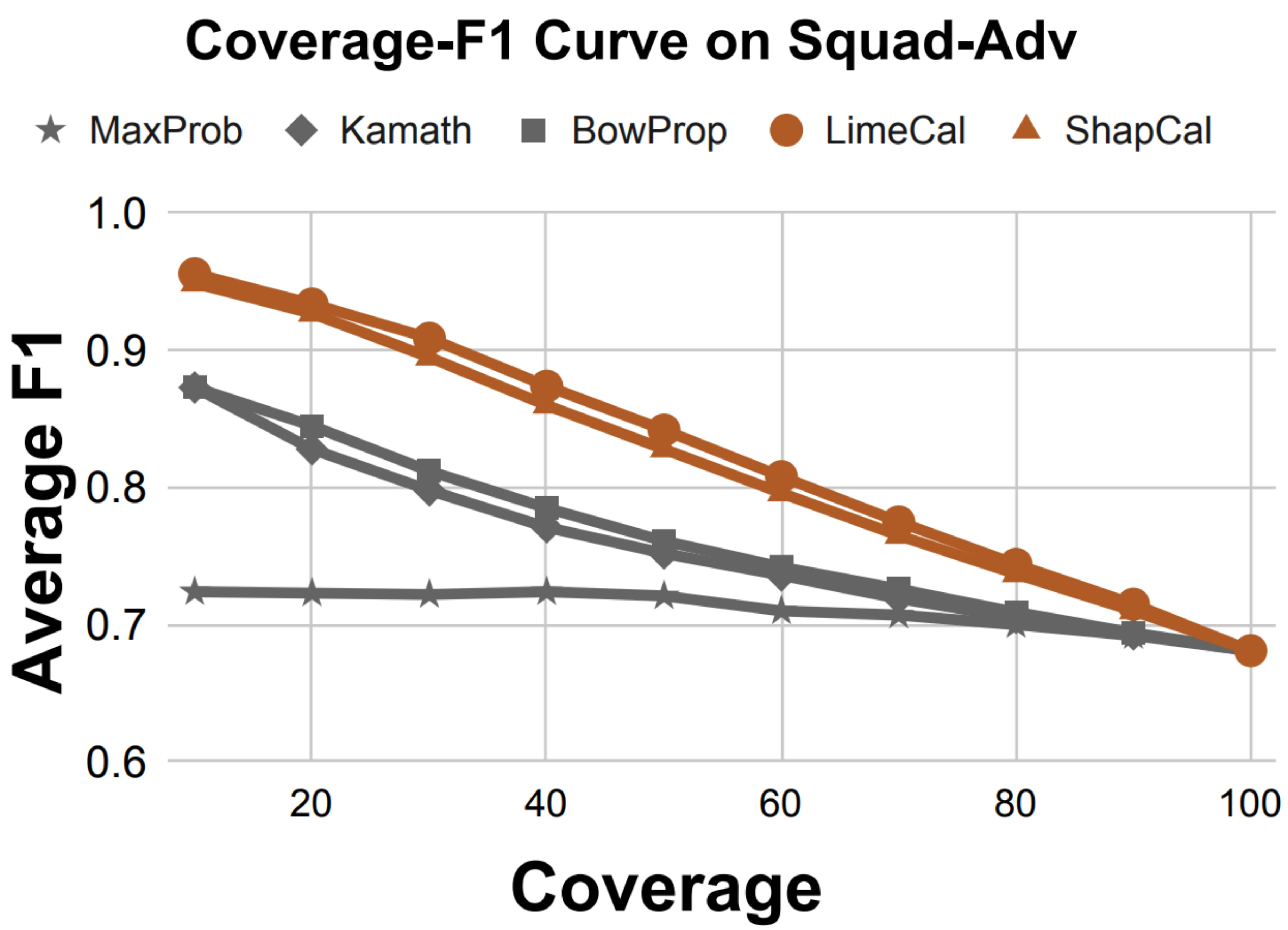
Methods

- **Prob**: confidence of prediction

Experiments (Cont'd)

Metrics

- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)
- Evaluate using area under coverage-F1 curve (**AUC**)



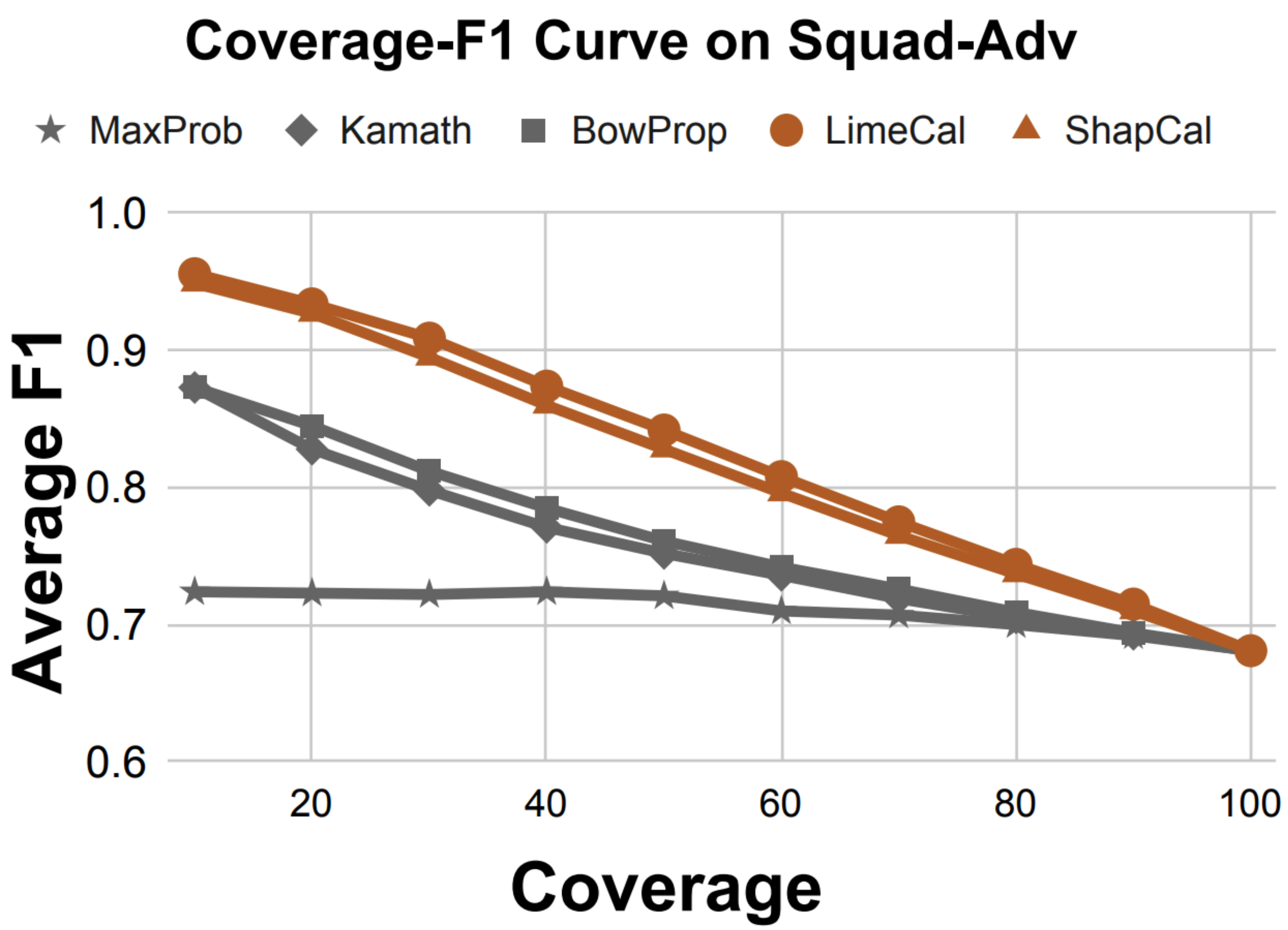
Methods

- **Prob**: confidence of prediction
- **Kamath**: (Kamath et al. 2020): calibrator using heuristic features (probabilities, length of context, length of answer)

Experiments (Cont'd)

Metrics

- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)
- Evaluate using area under coverage-F1 curve (**AUC**)



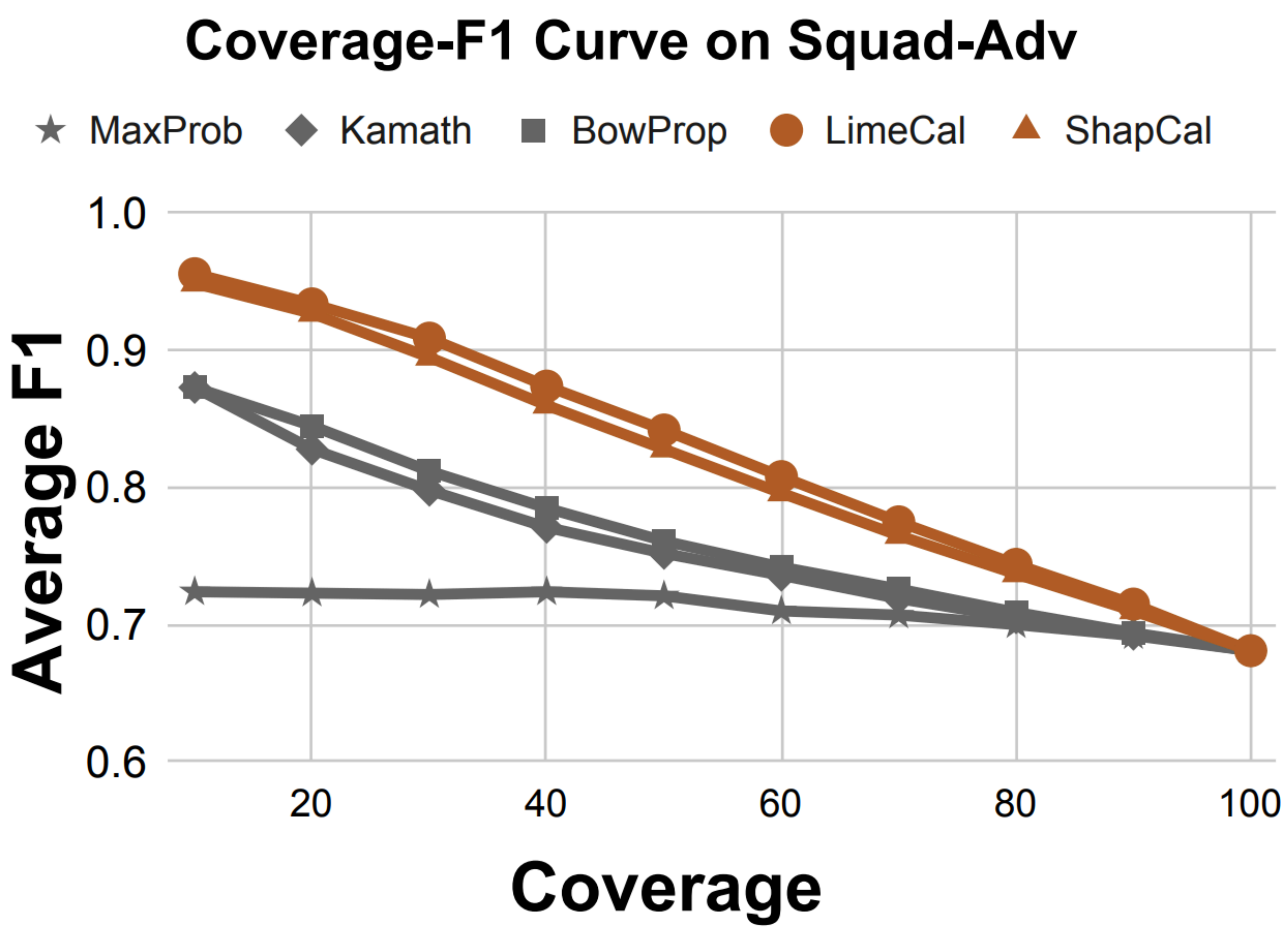
Methods

- **Prob**: confidence of prediction
- **Kamath**: (Kamath et al. 2020): calibrator using heuristic features (probabilities, length of context, length of answer)
- **BowCal**: calibrator using bag-of-word features without using explanations (e.g., count of NNP)

Experiments (Cont'd)

Metrics

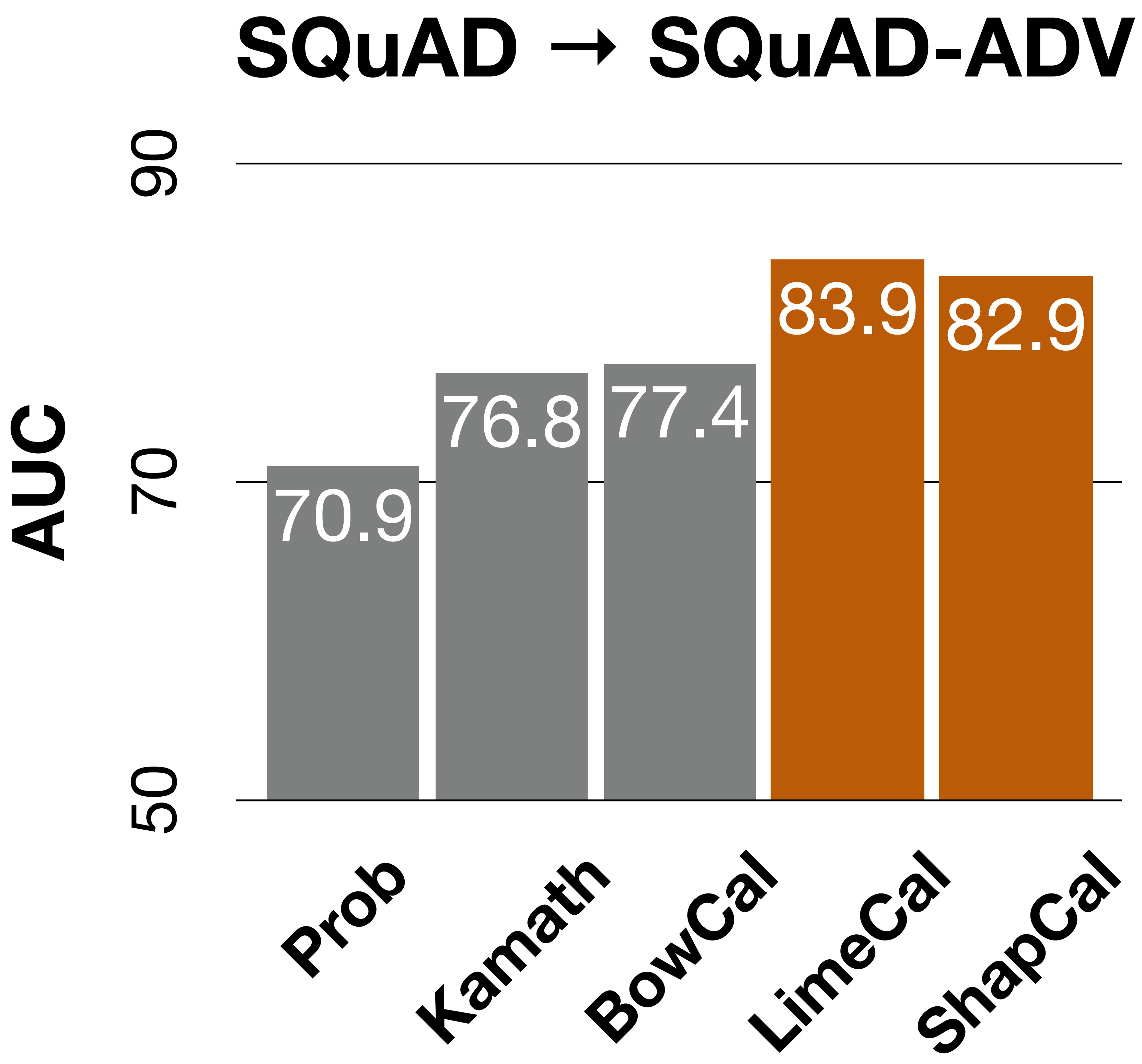
- Coverage-F1 Curve: average F1 scores with varying coverage (faction of most confident questions being answered)
- Evaluate using area under coverage-F1 curve (**AUC**)



Methods

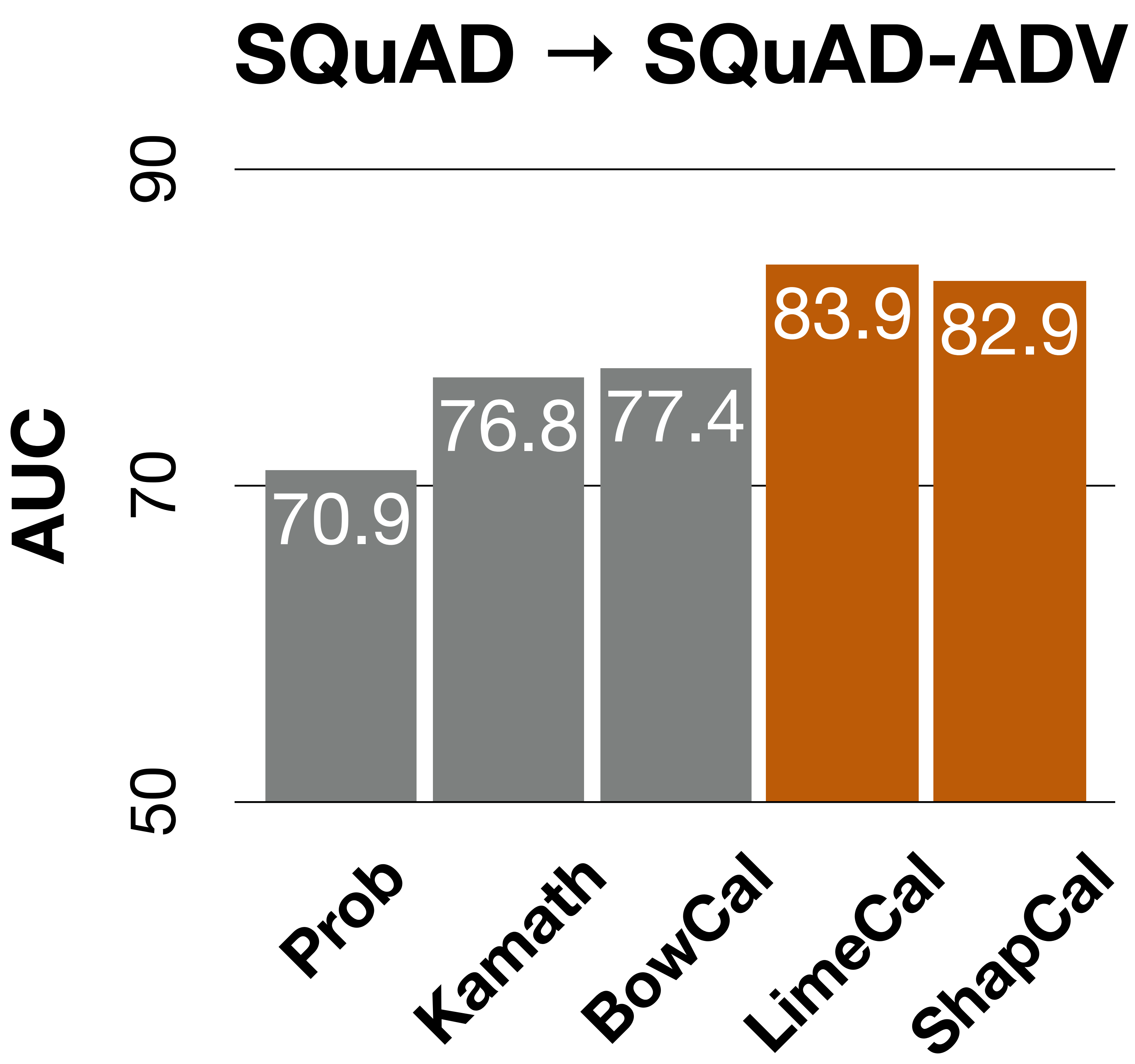
- **Prob**: confidence of prediction
- **Kamath**: (Kamath et al. 2020): calibrator using heuristic features (probabilities, length of context, length of answer)
- **BowCal**: calibrator using bag-of-word features without using explanations (e.g., count of NNP)
- **LimeCal & ShapCal**: calibrators using explanation-based features

Results



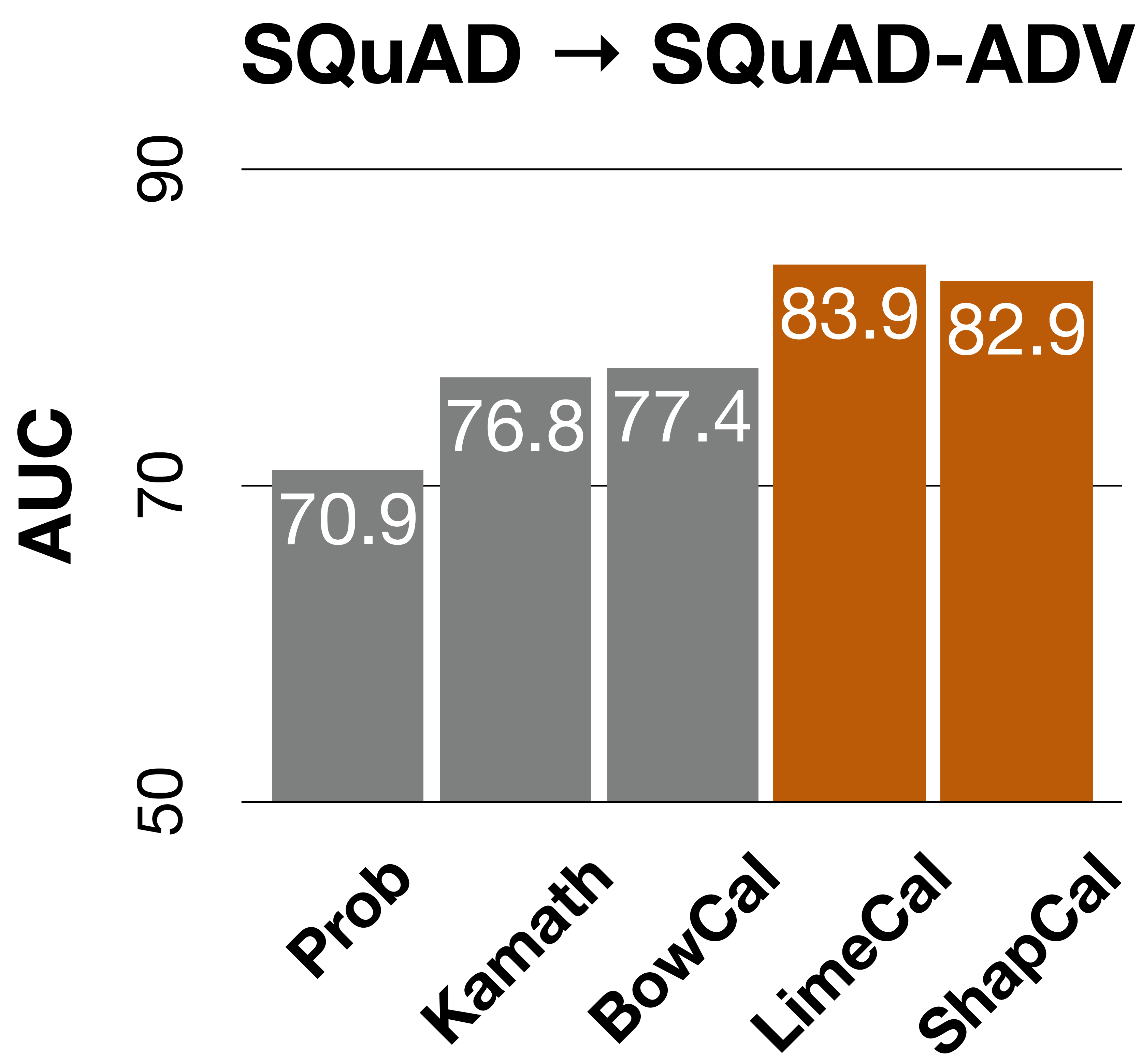
SQuAD → SQuAD-ADV
AUC
90
70
50
70.9
76.8
77.4
83.9
82.9
Prob
Kamath
BowCal
LimeCal
ShapCal

Results



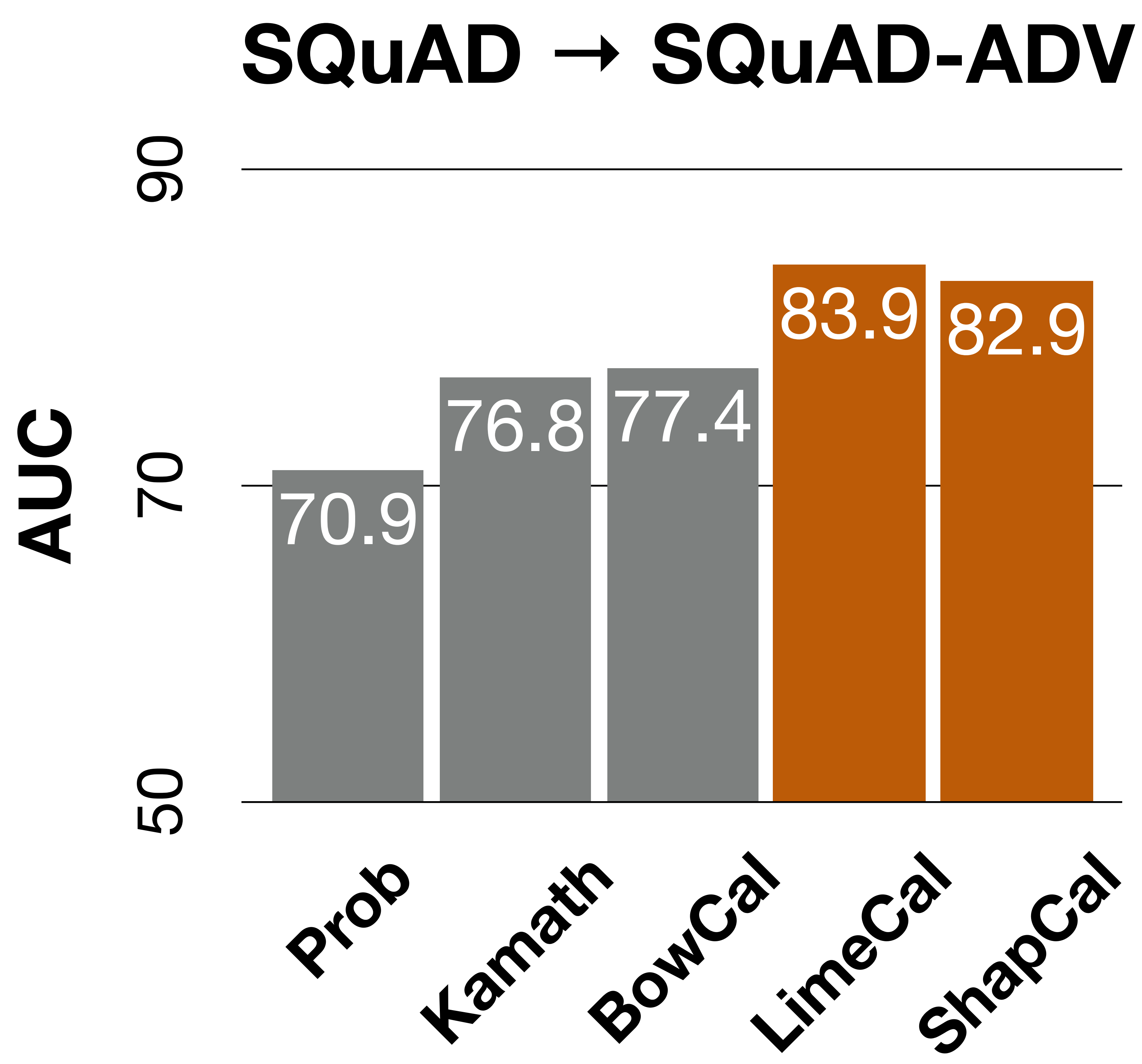
- **LimeCal** achieves the best performance

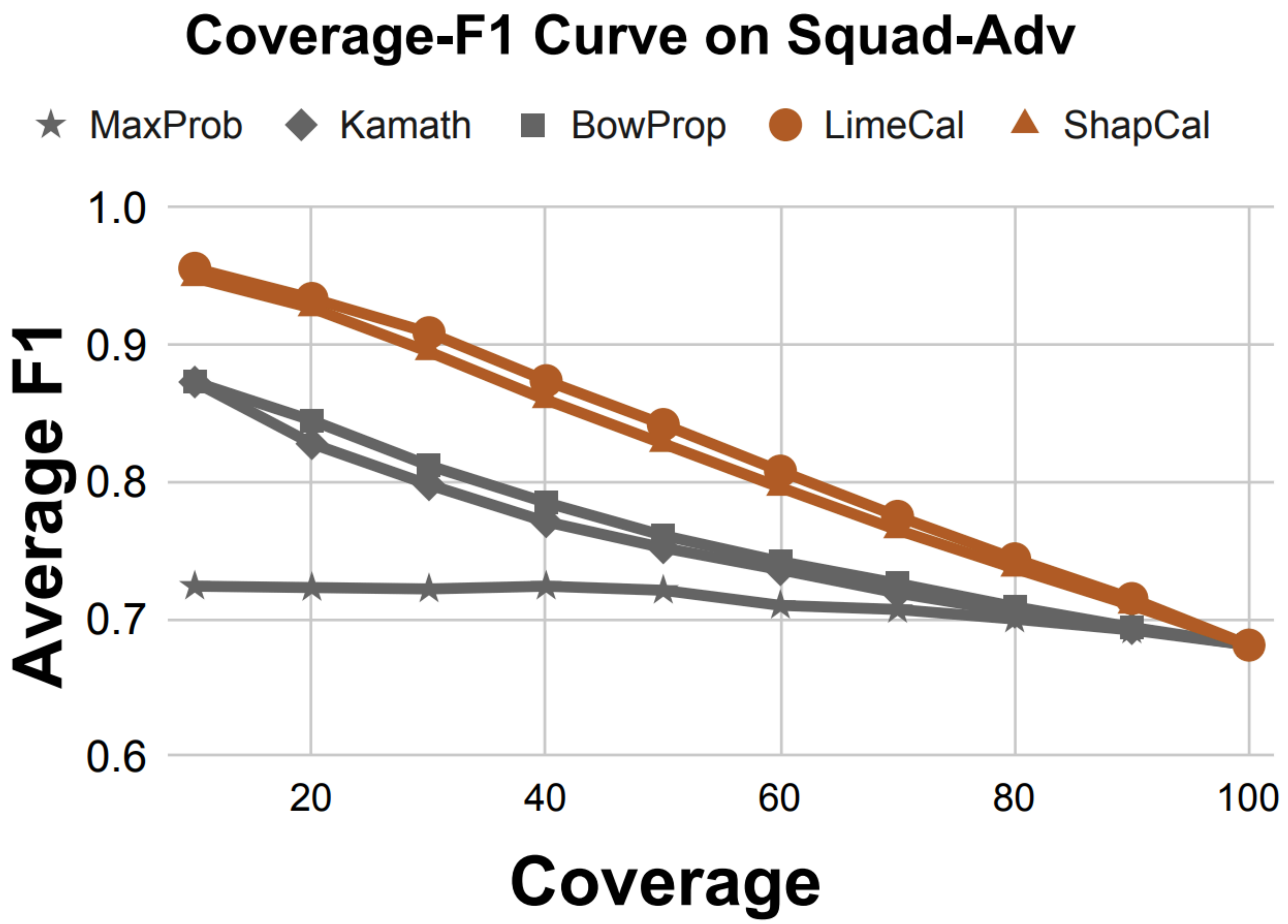
Results



- **LimeCal** achieves the best performance
- Explanations are helpful; **Lime/ShapCal** outperforms calibrators without using explanations

Results

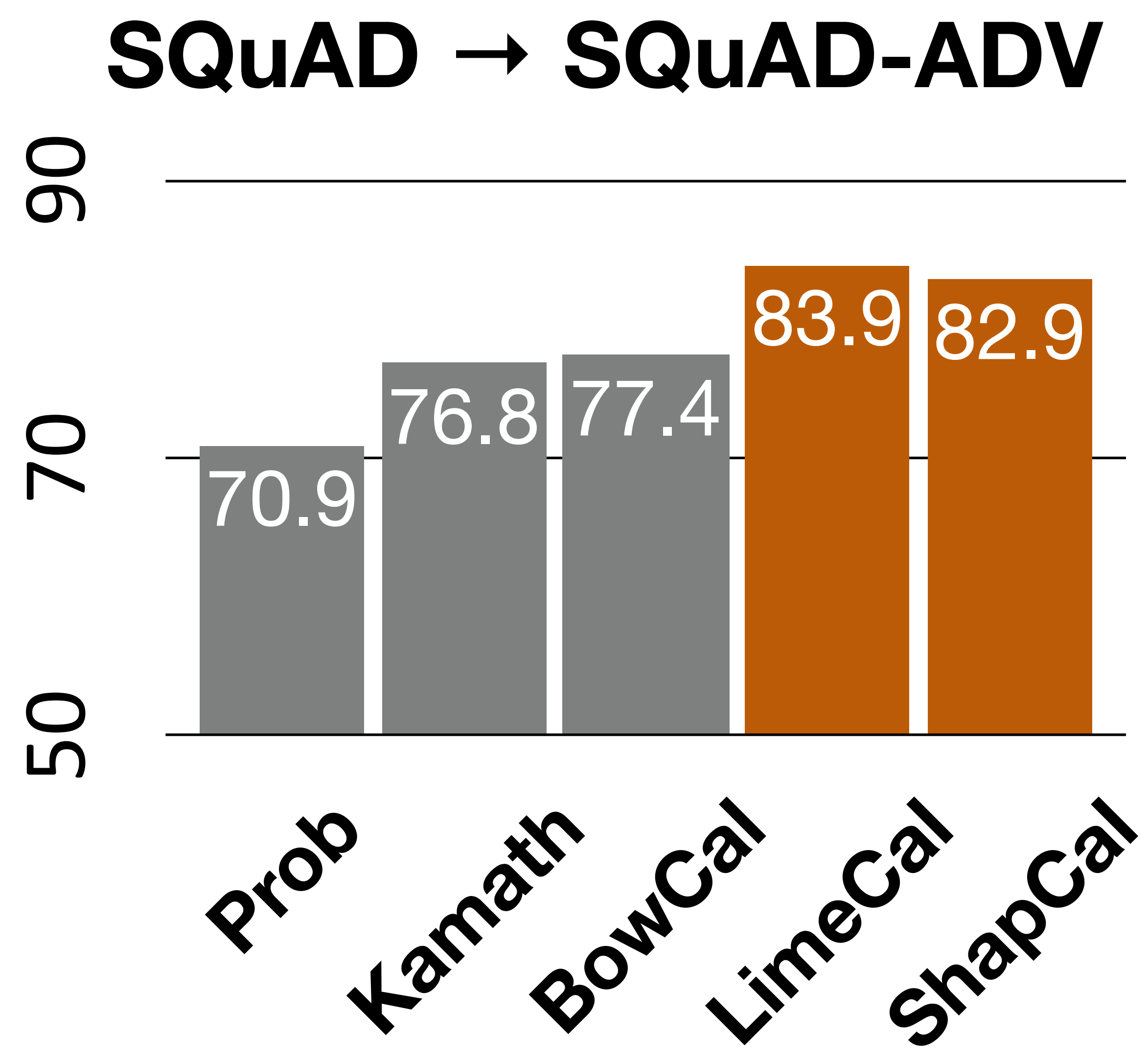


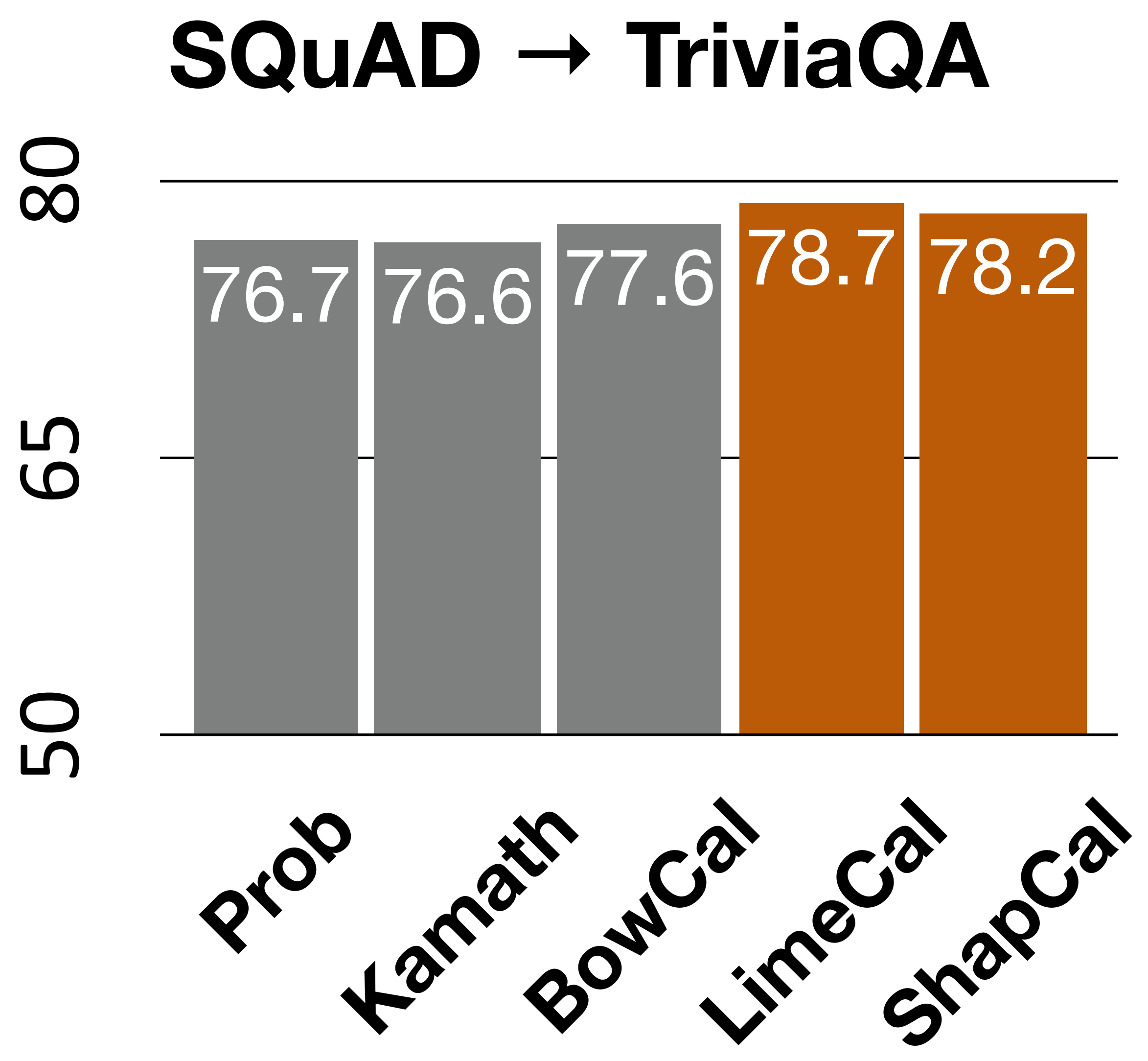


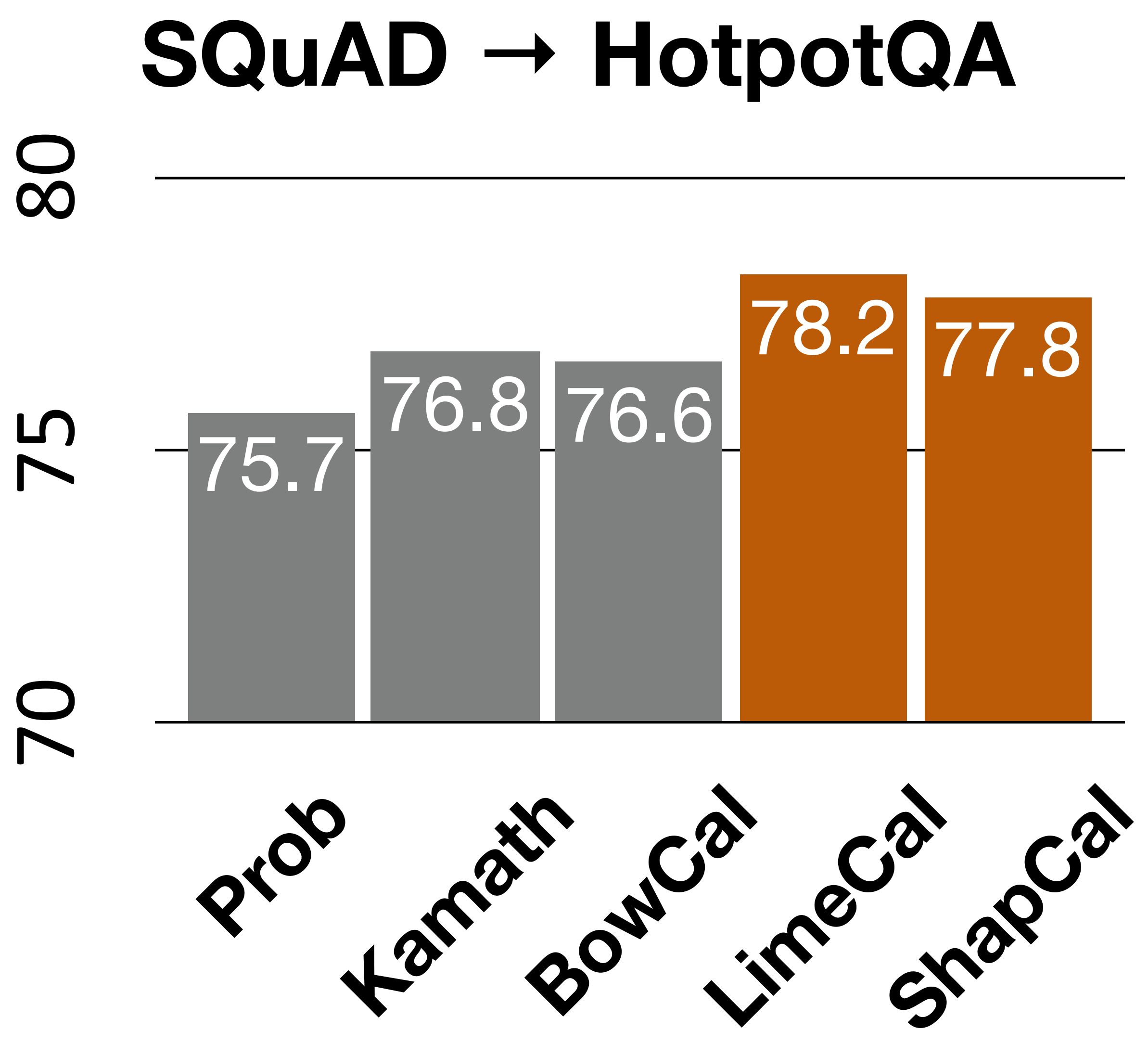
- **LimeCal** achieves the best performance
- Explanations are helpful; **Lime/ShapCal** outperforms calibrators without using explanations
- Substantial performance difference when selectively answering a part of the questions that the calibrator is most confident with

Results (Cont'd)

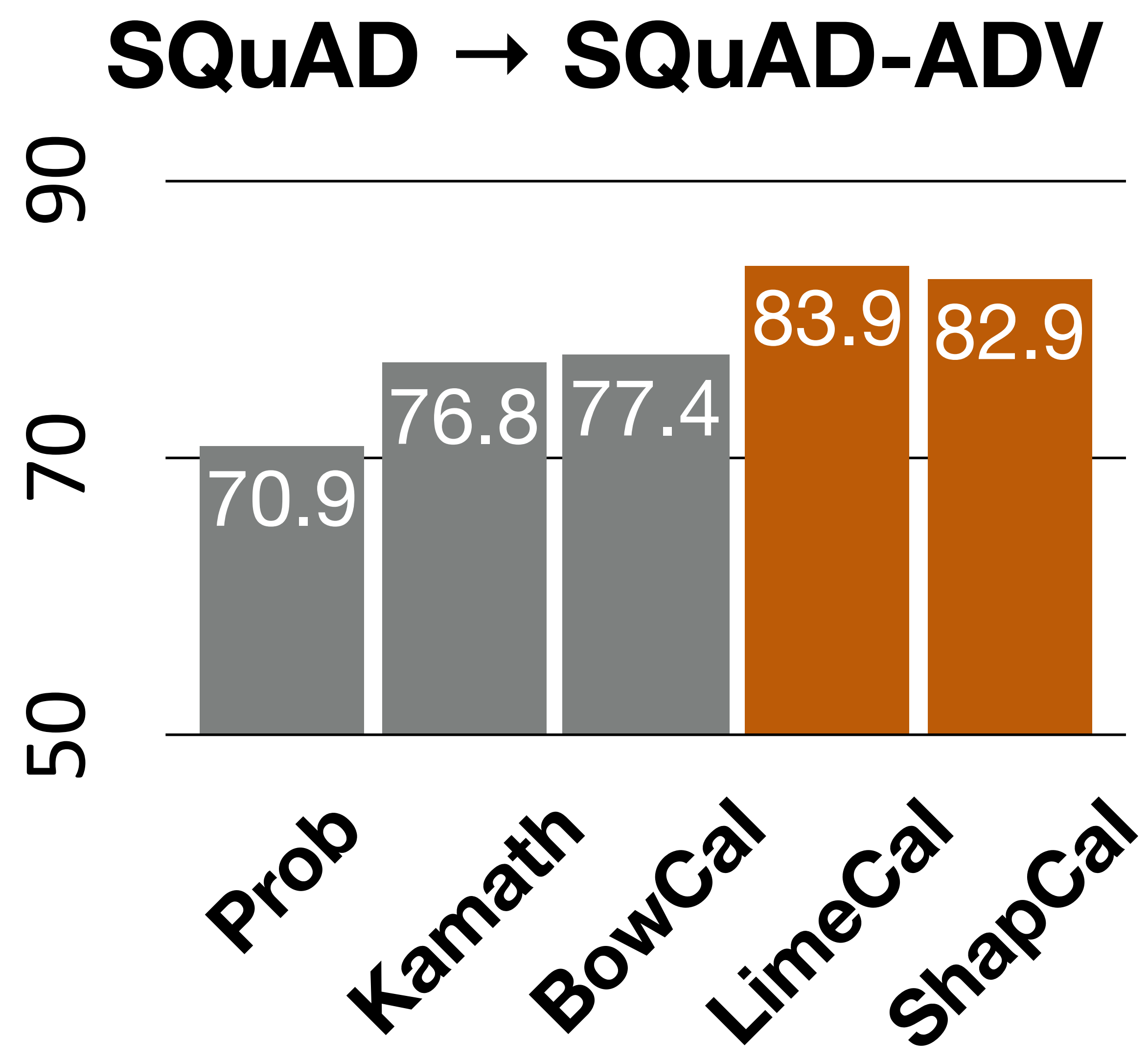
Results (Cont'd)

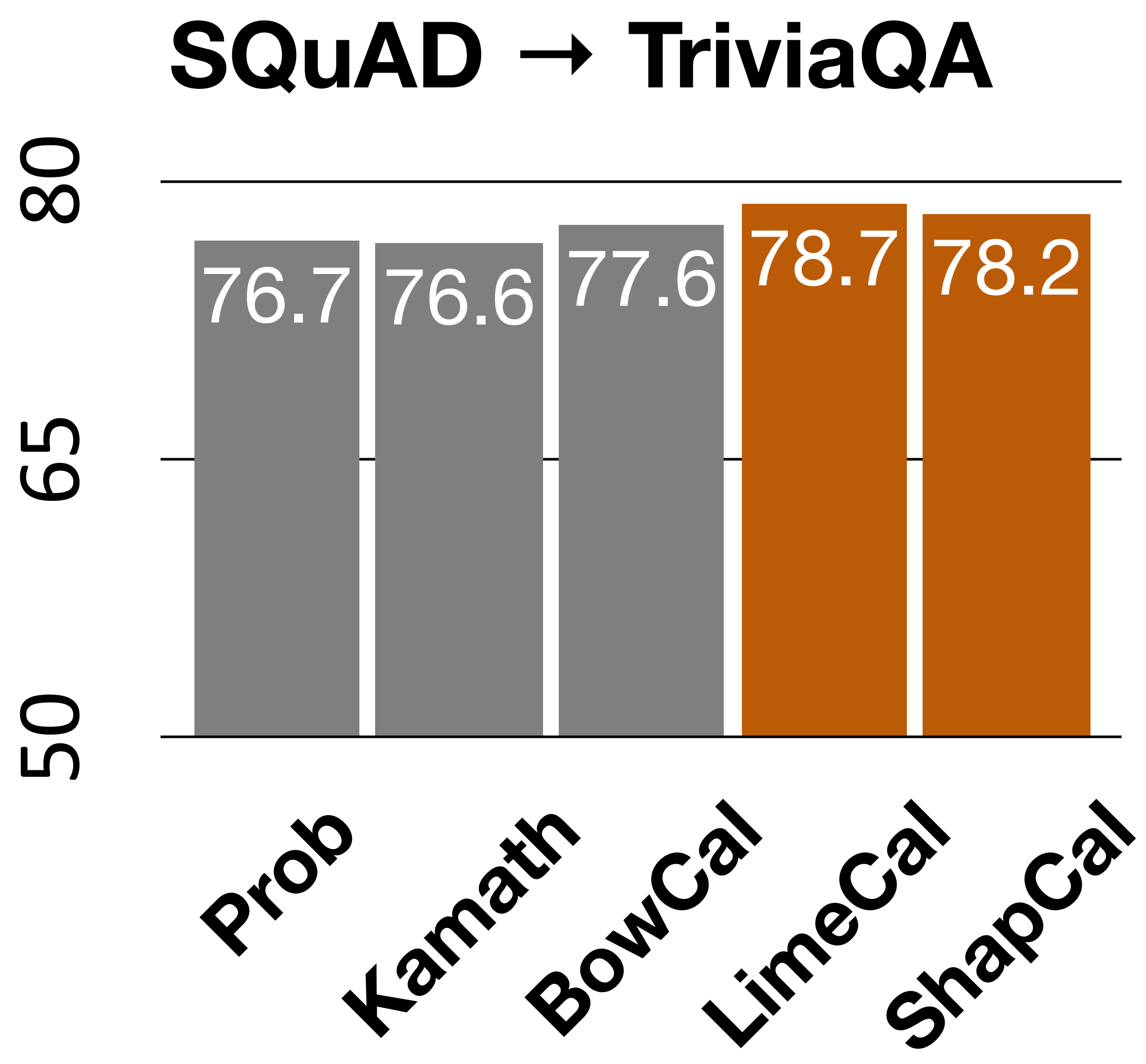


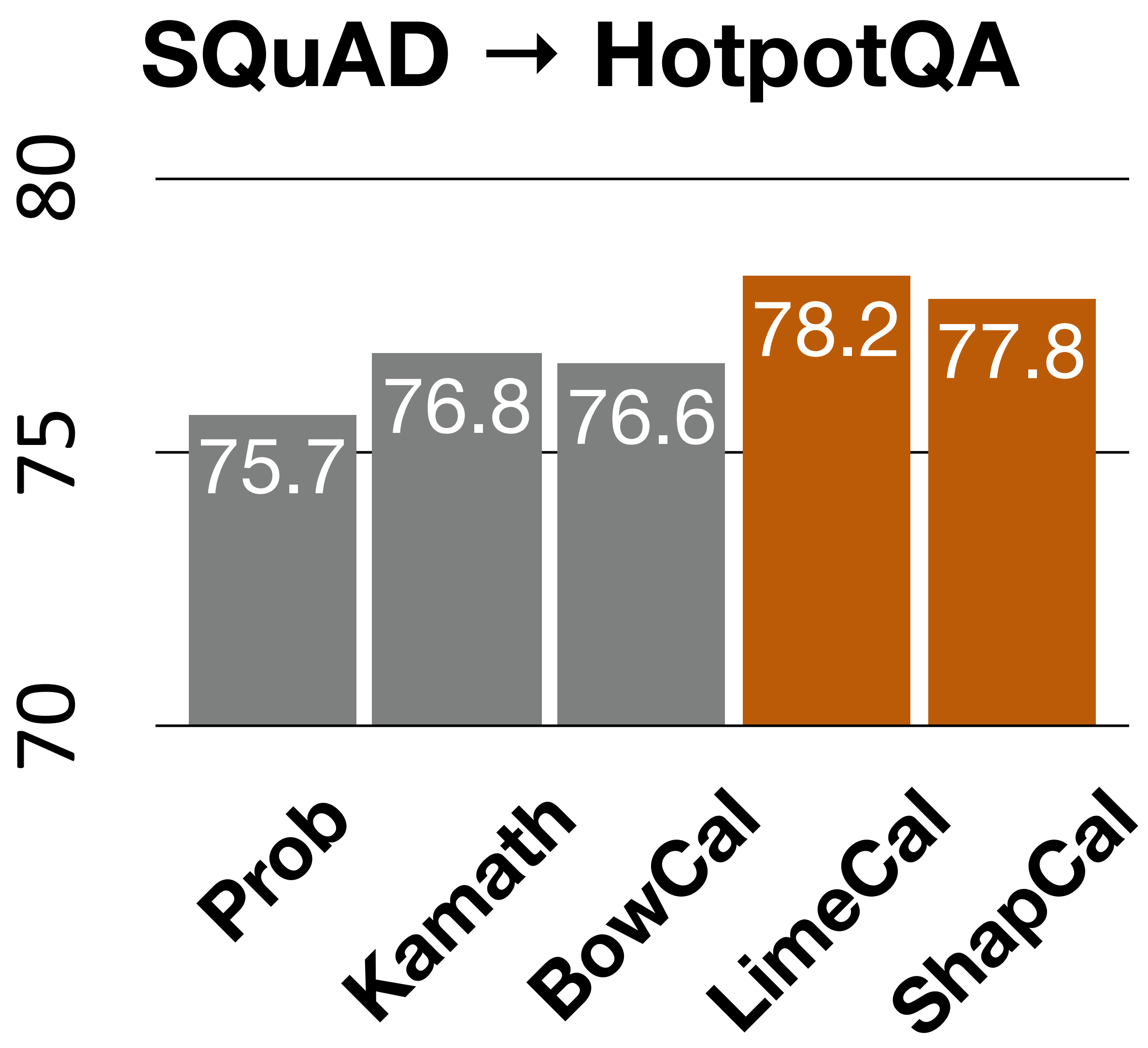


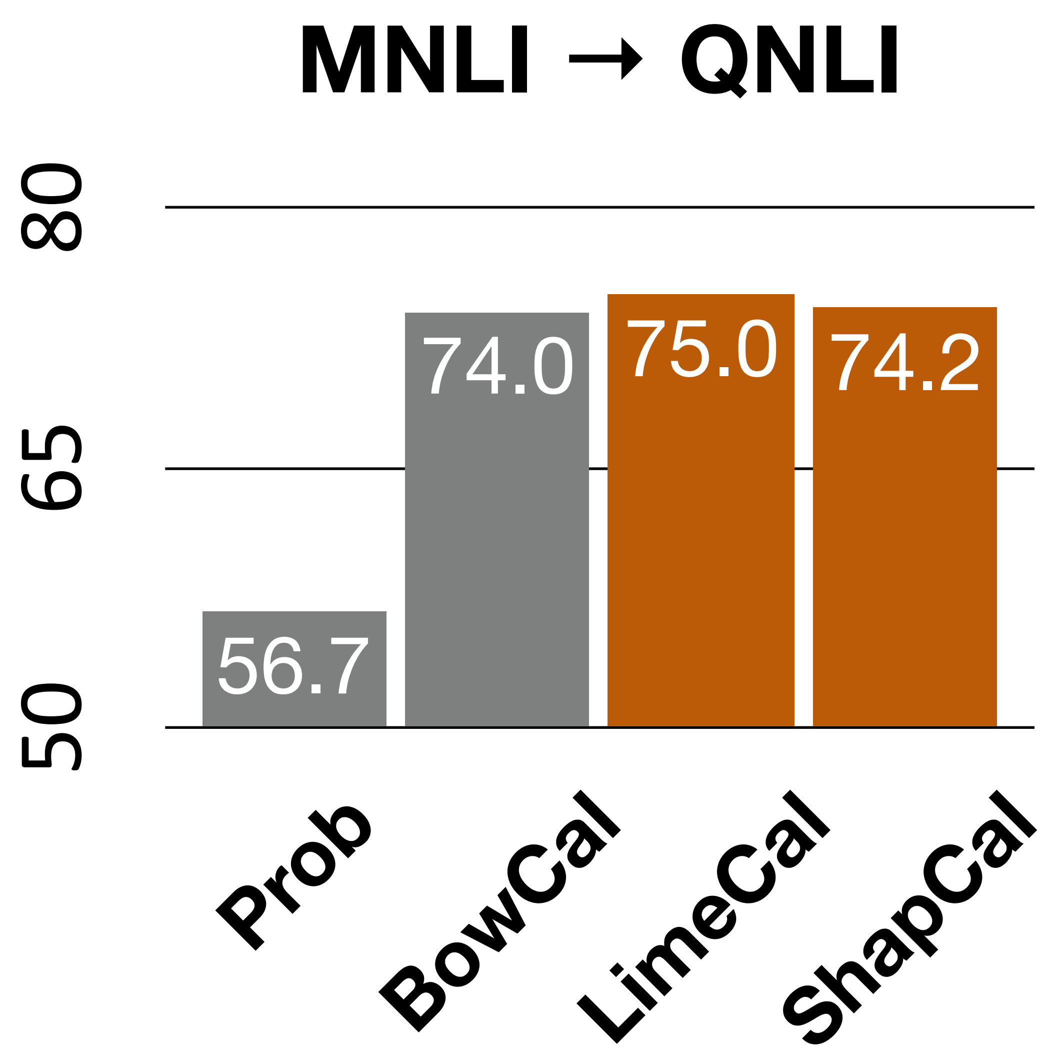


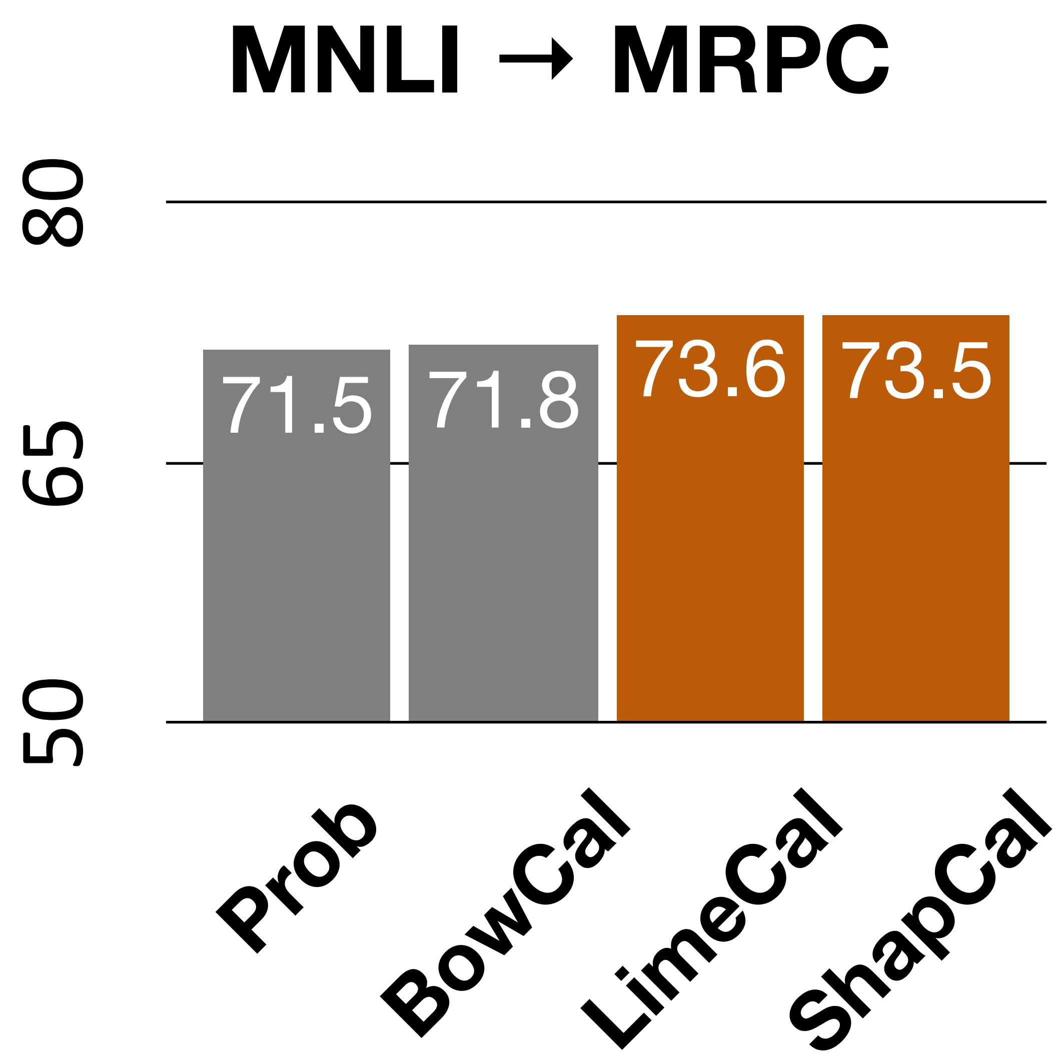
Results (Cont'd)



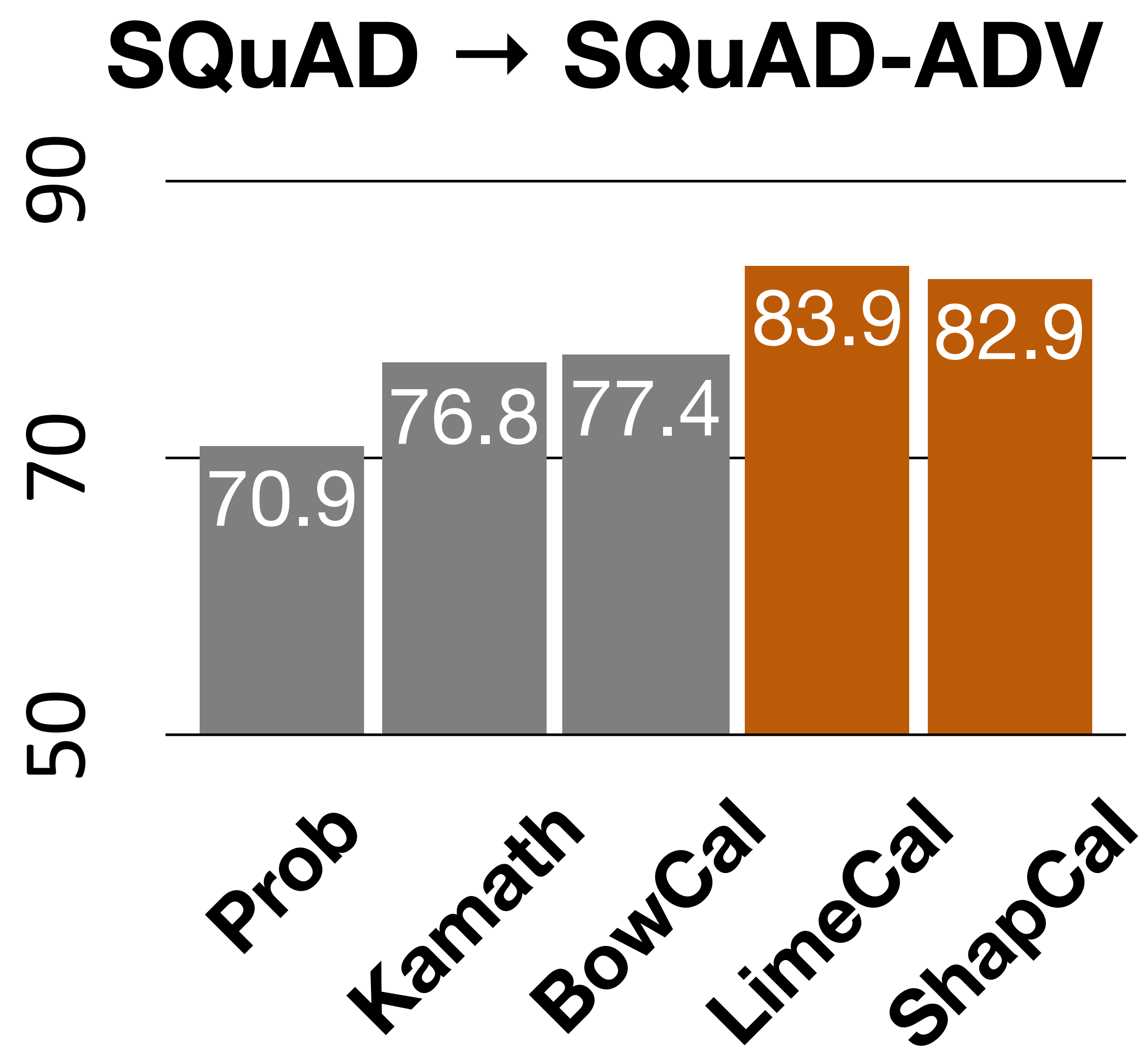


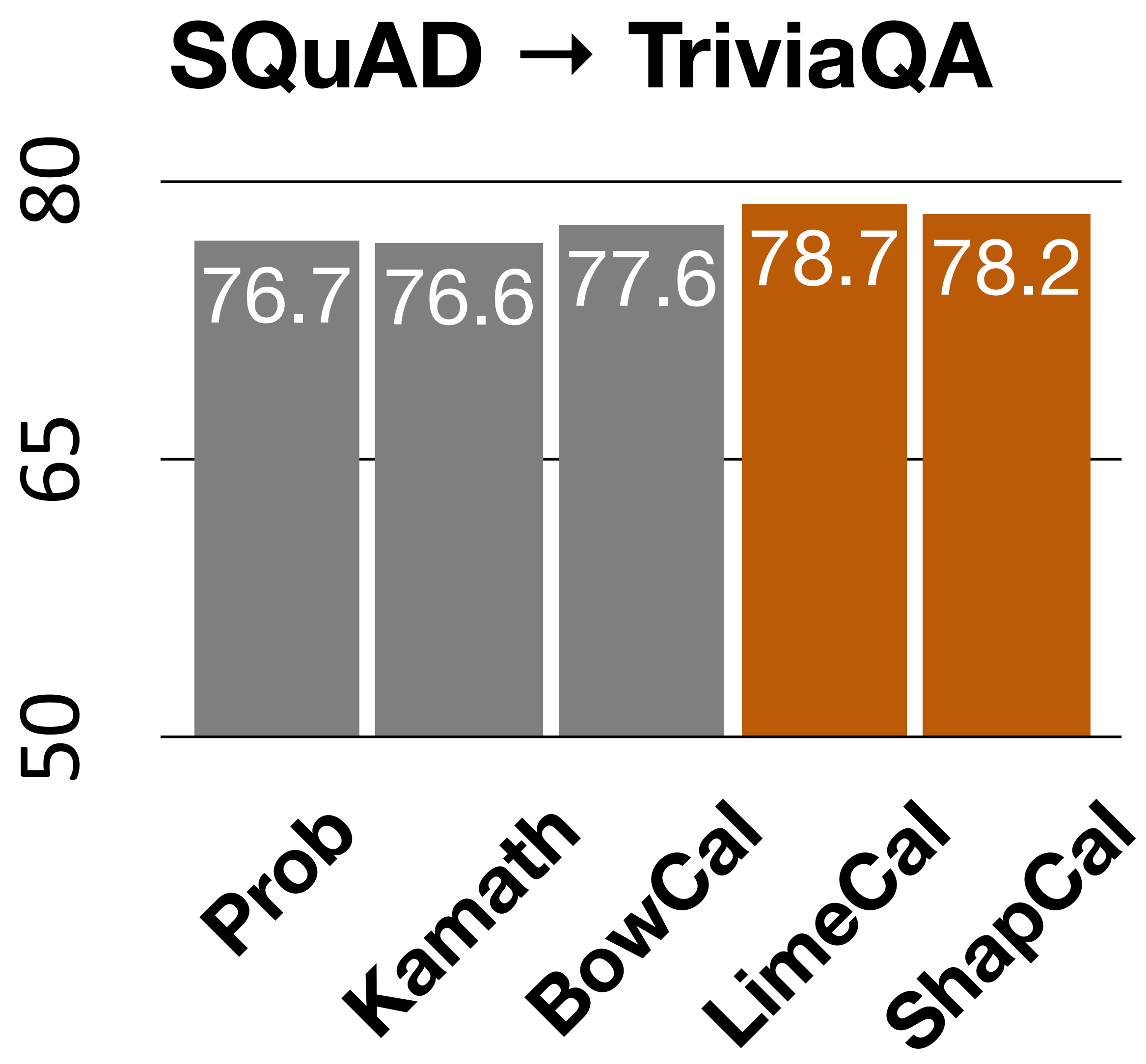


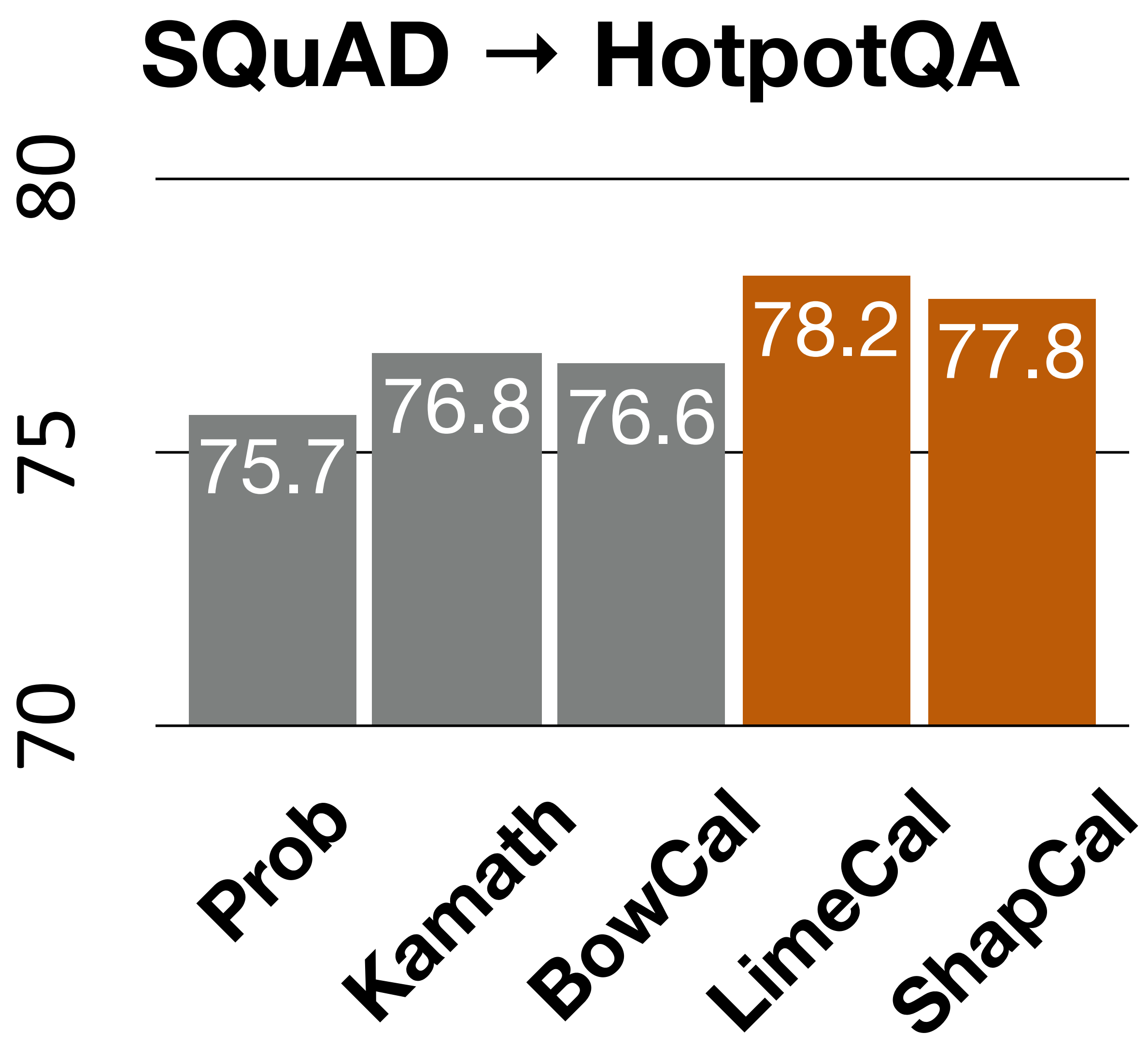


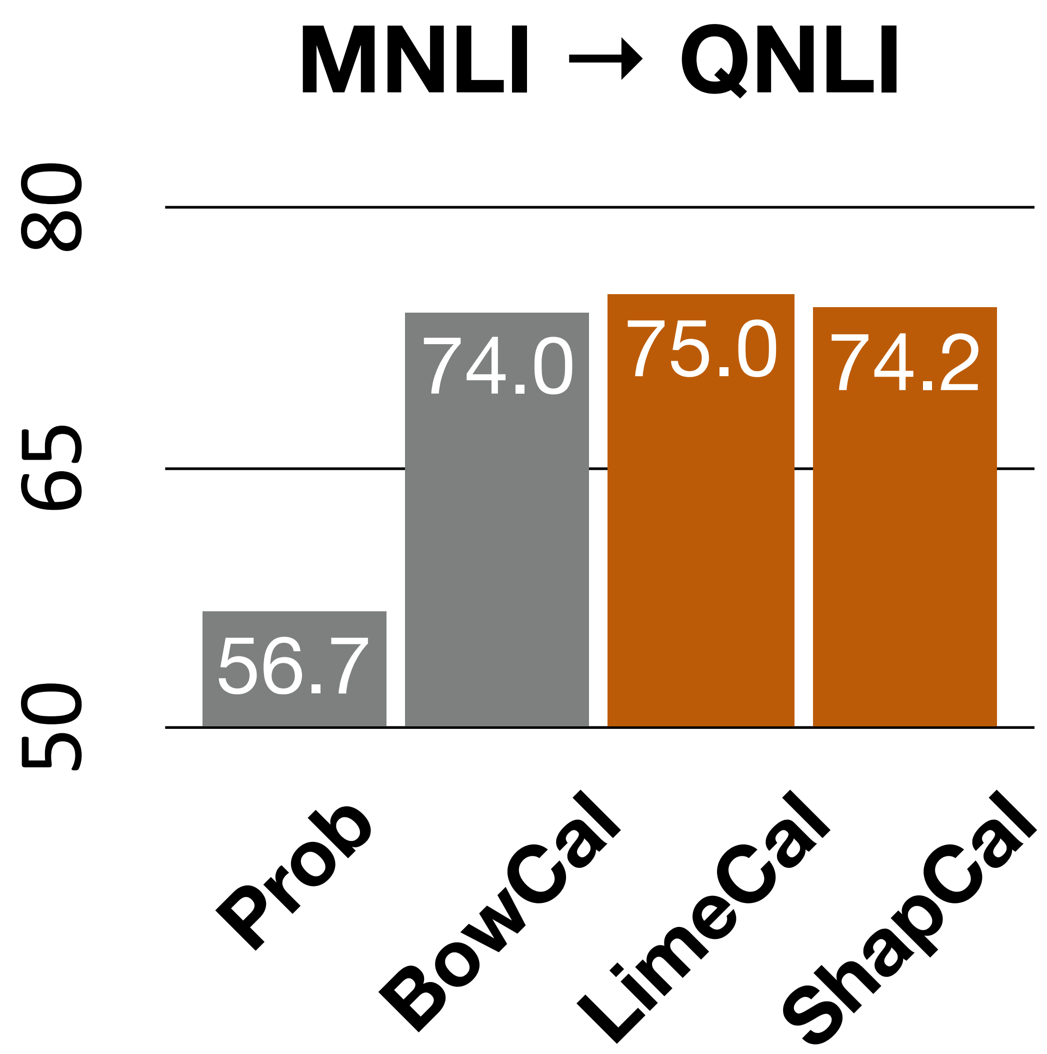


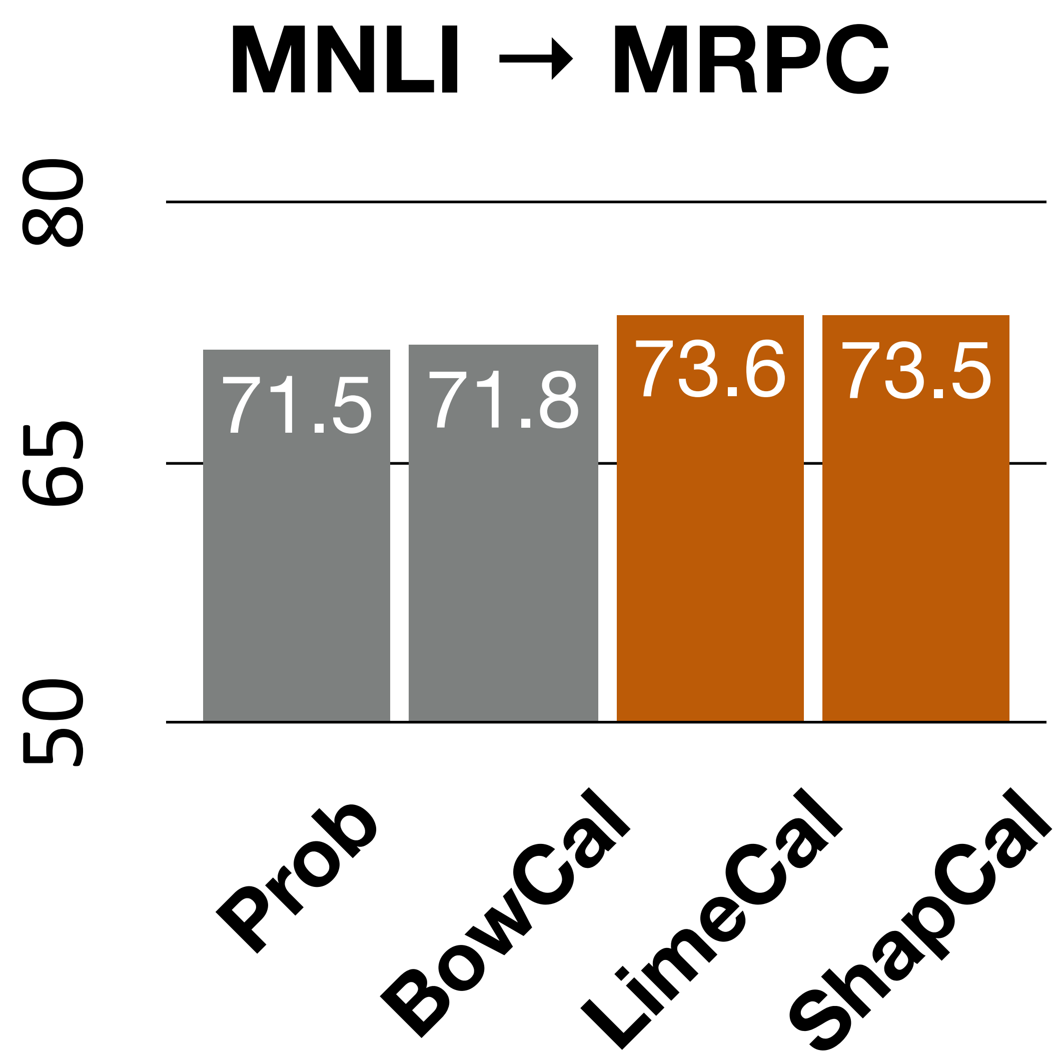
Results (Cont'd)









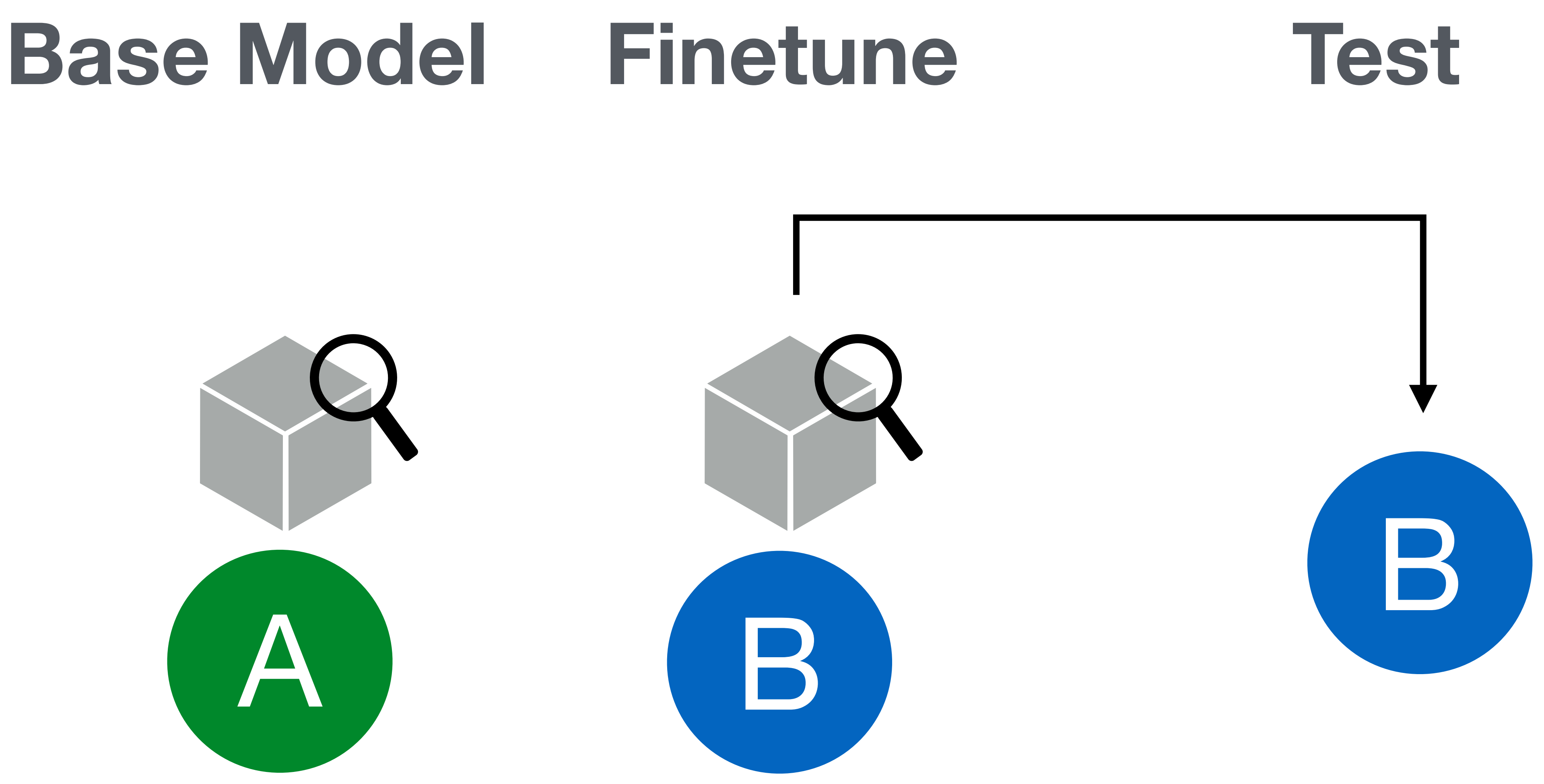


- Explanations improves the generalization performance across all pairs covering both QA and NLI tasks

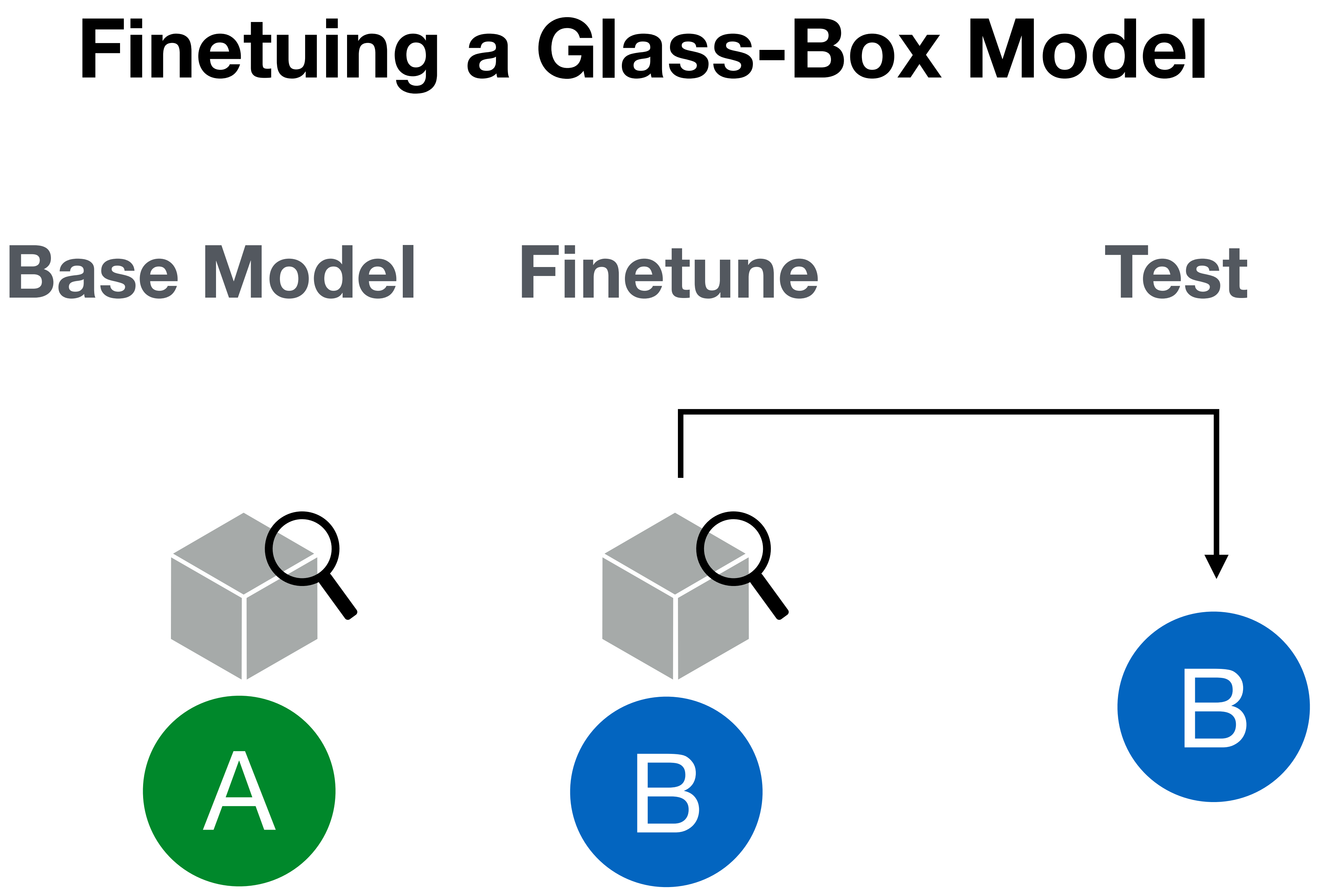
Comparison to Finetuned Models

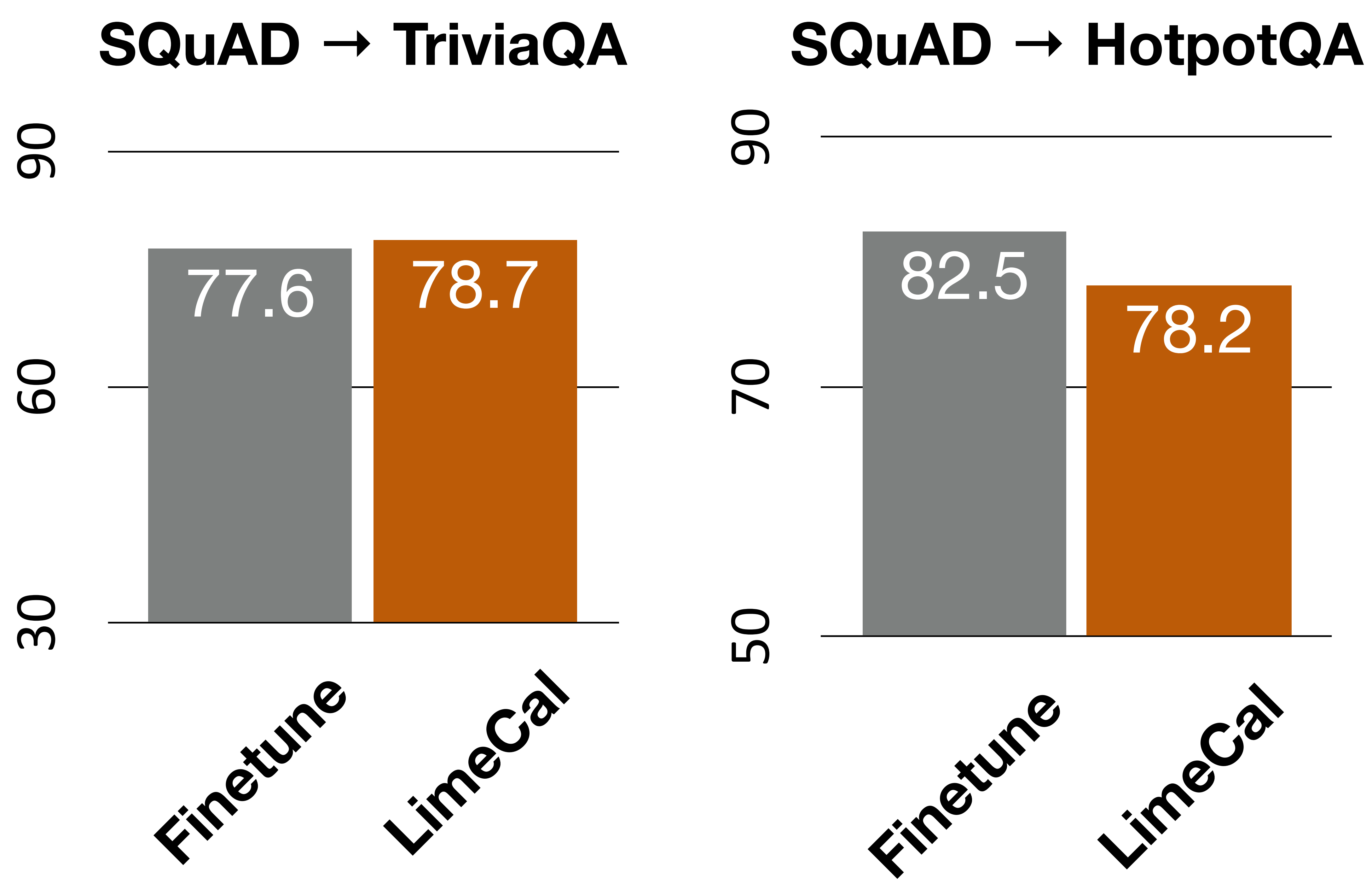
Comparison to Finetuned Models

Finetuing a Glass-Box Model



Comparison to Finetuned Models





- Explanation-based calibrator even outperforms a fine-tuned model on **SQuAD → TriviaQA**

Wrap-up

Wrap-up

- **Conclusion:**

Wrap-up

- **Conclusion:**
  - Can explanations be useful for calibrating black-box models? **YES!**

Wrap-up

- **Conclusion:**
  - Can explanations be useful for calibrating black-box models? **YES!**
  - Using explanations successfully improves model generalization on QA and NLI tasks

Wrap-up

- **Conclusion:**
  - Can explanations be useful for calibrating black-box models? **YES!**
  - Using explanations successfully improves model generalization on QA and NLI tasks

- **Limitations:**

Wrap-up

- **Conclusion:**
  - Can explanations be useful for calibrating black-box models? **YES!**
  - Using explanations successfully improves model generalization on QA and NLI tasks

- **Limitations:**
  - Generating explanations with **Lime** and **Shap** is computationally expensive

Wrap-up

- **Conclusion:**
  - Can explanations be useful for calibrating black-box models? **YES!**
  - Using explanations successfully improves model generalization on QA and NLI tasks

- **Limitations:**
  - Generating explanations with **Lime** and **Shap** is computationally expensive

How about Large Language Models?

The Unreliability of Explanations in Few-Shot In-Context Learning (Ye and Durrett, ArXiv 2022)

Free text explanations can also be useful for calibrating large LM (GPT-3) in some settings

Wrap-up

- **Conclusion:**
  - Can explanations be useful for calibrating black-box models? **YES!**
  - Using explanations successfully improves model generalization on QA and NLI tasks

- **Limitations:**
  - Generating explanations with **Lime** and **Shap** is computationally expensive

How about Large Language Models?

The Unreliability of Explanations in Few-Shot In-Context Learning (Ye and Durrett, ArXiv 2022)

Free text explanations can also be useful for calibrating large LM (GPT-3) in some settings

Code Available at https://github.com/xiye17/InterpCalib