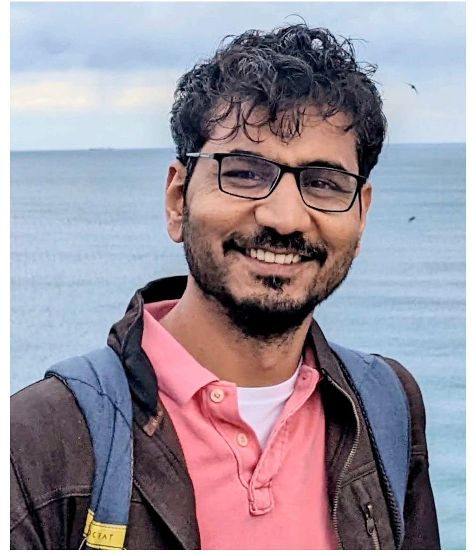
Prune 'n Predict

Optimizing LLM Decision-making with Conformal Prediction

ICML, 2025



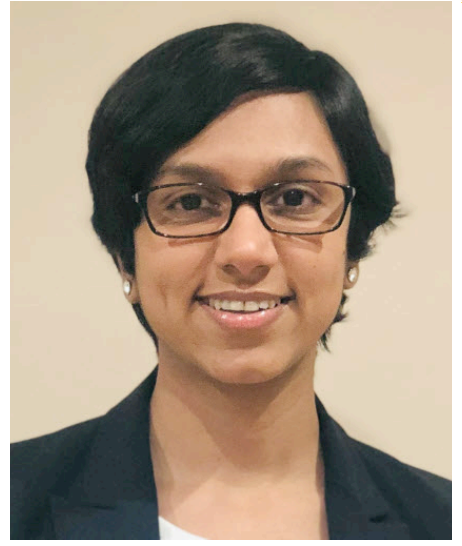
Harit Vishwakarma

Alan Mishler

Thomas Cook

Nic Dalmasso

Natraj Raman



Sumitra Ganesh

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

JPMORGAN CHASE & CO.

ICML
International Conference
On Machine Learning

Applications where LLM/Agent needs to decide from multiple choices.

Tool/API Selection

QUESTION: Given the API Lichess, and the following instruction, "I'm interested in joining a Lichess tournament. Can you show me a list of upcoming tournaments and their start times?" Which of the following functions should you call?

A. listTournaments Retrieve a list of ongoing and upcoming tournaments.

B. getPuzzle Retrieve a puzzle and its corresponding solution.

C. getUserInfo Retrieve all user information, including games played, ratings, and statistics.

D. getTournamentInfo Retrieve detailed information about a specific tournament.

QA with chatbot in consumer banking

What is the grace period for mortgage payment?

A. 1 day
B. 1 week
C. 15 days
D. 1 month

Standardized Tests

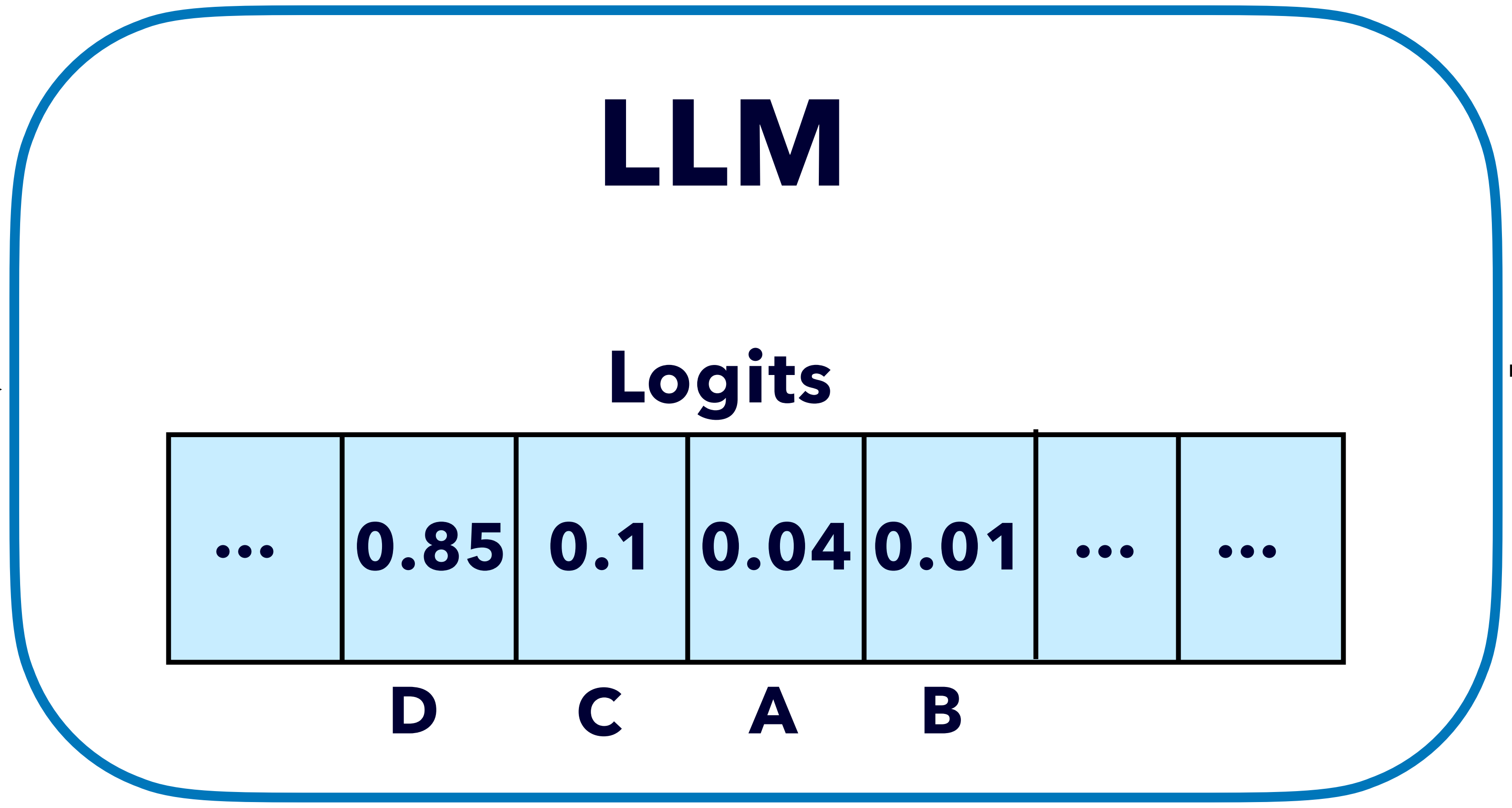
Common Benchmarks

LLMs may answer incorrectly and worse with high confidence

What is the grace period for mortgage payment?

A. 1 day
B. 1 week
C. 15 days
D. 1 month

The correct answer is :



Point Prediction

The correct answer is : D.

Confidence/Logit Score 85%

↓

Real consequences e.g. in Finance

Lose trust over time

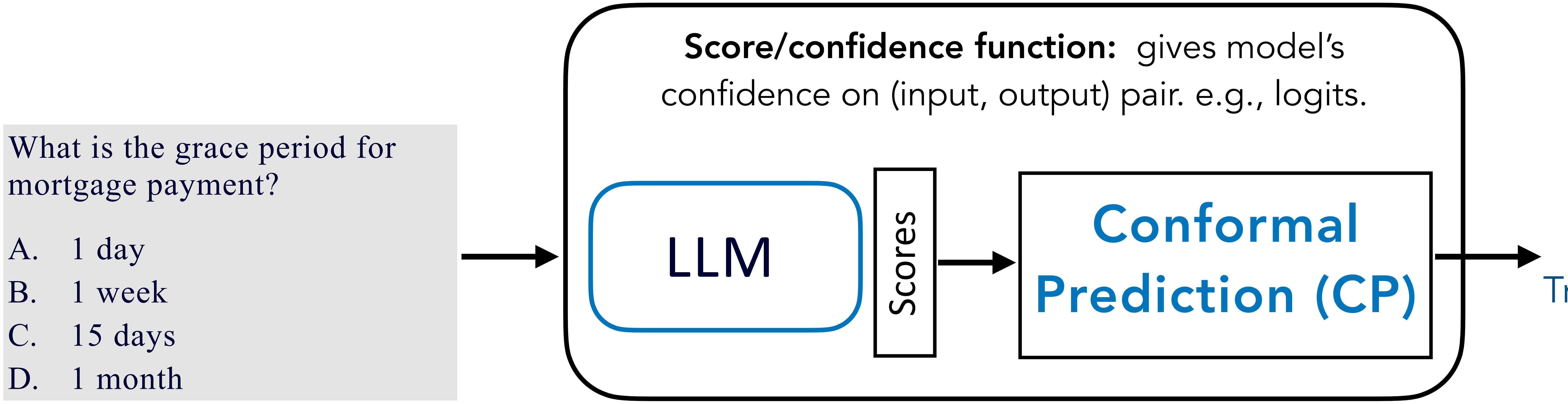
How to safeguard against such mistakes and improve accuracy?

Without Expensive Fine-tuning.

Moving from point prediction to set prediction for safety

Conformal Prediction gives confidence sets/intervals without distributional assumptions.

Set Prediction



Answer is in {C,D}.

True answer is in the predicted set 95% of the times.

→ ***Coverage guarantee***

Predicting sets is safer!

Kumar et al., Conformal Prediction with Large Language Models for Multi-Choice Question Answering, ICML WS 2023

Su et al., API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access arXiv, 2024

How does Conformal Prediction Work?

How does Conformal Prediction Work?

Question $x \in \mathcal{X}$

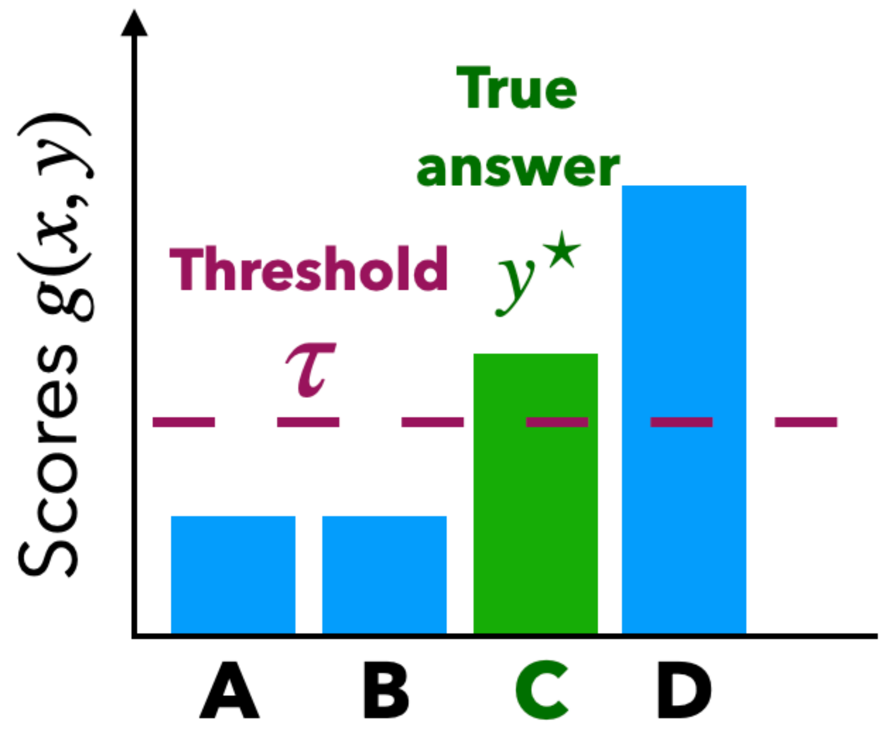
Option $y \in \mathcal{Y}$

$\mathcal{Y} = \{A, B, C, D\}$

$g(x, y)$: score for option y of question x

How does Conformal Prediction Work?

Prediction sets



Output
{C, D}

$$C(x; g, \tau) = \{y \in \mathcal{Y} : g(x, y) \geq \tau\}$$

Question $x \in \mathcal{X}$

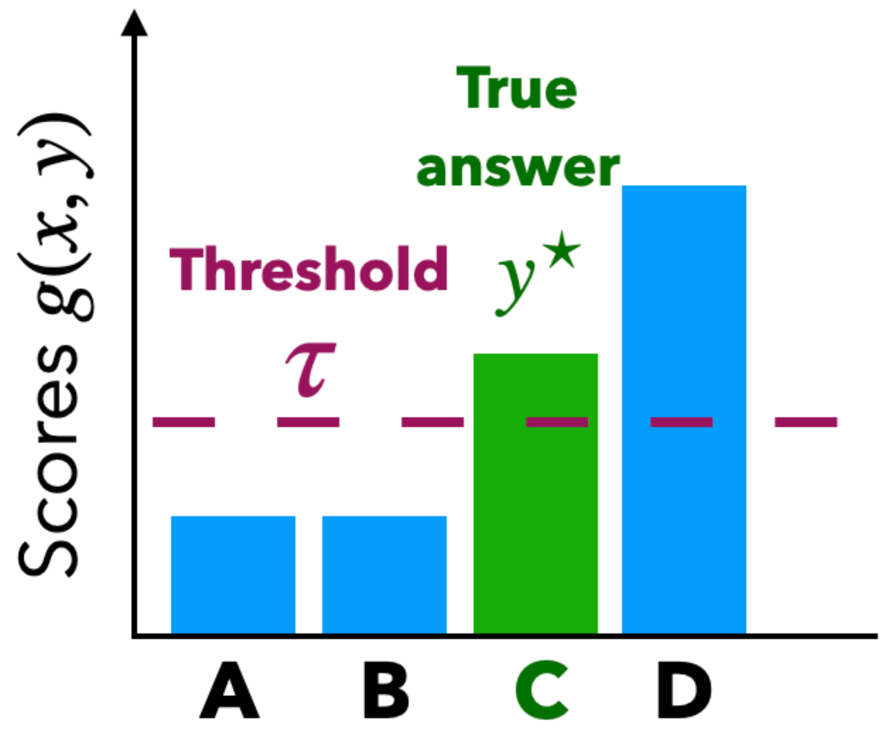
Option $y \in \mathcal{Y}$

$\mathcal{Y} = \{A, B, C, D\}$

$g(x, y)$: score for option y of question x

How does Conformal Prediction Work?

Prediction sets



Output
{C, D}

$$C(x; g, \tau) = \{y \in \mathcal{Y} : g(x, y) \geq \tau\}$$

Question $x \in \mathcal{X}$

Option $y \in \mathcal{Y}$

$\mathcal{Y} = \{A, B, C, D\}$

$g(x, y)$: score for option y of question x

Threshold Estimation

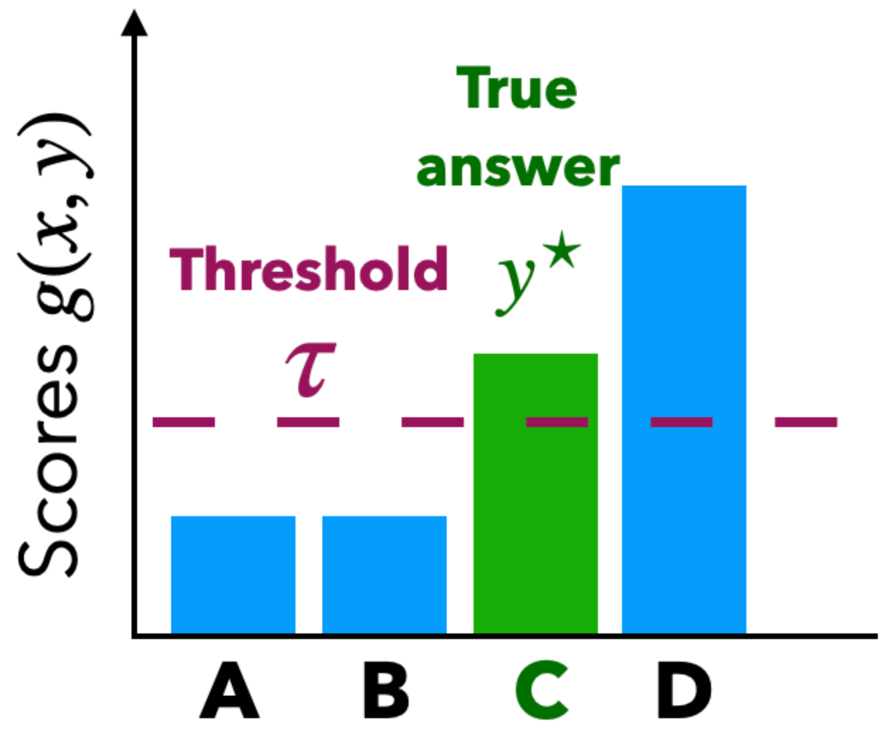
Calibration Data $D_{\text{cal}} = \{(x_i, y_i^\star)\}_{i=1}^n$

Coverage $\widehat{\mathcal{P}}(g, \tau) = \dfrac{\#\text{times } y_i^\star \in C(x_i; g, \tau)}{n}$

Avg. set size $\widehat{S}(g, \tau) = \dfrac{1}{n} \sum_{i=1}^{n} \left| C(x_i; g, \tau) \right|$

How does Conformal Prediction Work?

Prediction sets



Output
{C, D}

$$C(x; g, \tau) = \{y \in \mathcal{Y} : g(x, y) \geq \tau\}$$

Question $x \in \mathcal{X}$

Option $y \in \mathcal{Y}$

$\mathcal{Y} = \{A, B, C, D\}$

$g(x, y)$: score for option y of question x

Threshold Estimation

Calibration Data $D_{\text{cal}} = \{(x_i, y_i^\star)\}_{i=1}^{n}$

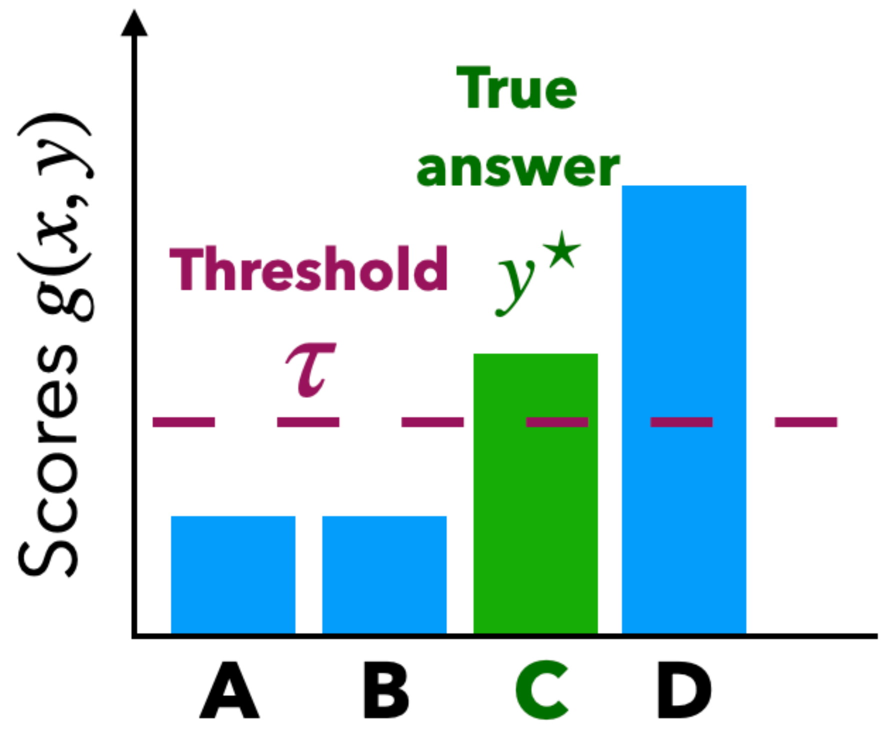
Coverage $\widehat{\mathcal{P}}(g, \tau) = \dfrac{\#\text{times } y_i^\star \in C(x_i; g, \tau)}{n}$

Avg. set size $\widehat{S}(g, \tau) = \dfrac{1}{n} \sum_{i=1}^{n} \left| C(x_i; g, \tau) \right|$

Find smallest $\hat{\tau}_\alpha$ such that $\widehat{\mathcal{P}}(g, \tau) \geq 1 - \alpha$

How does Conformal Prediction Work?

Prediction sets



Output
{C, D}

$$C(x; g, \tau) = \{y \in \mathcal{Y} : g(x, y) \geq \tau\}$$

Question $x \in \mathcal{X}$

Option $y \in \mathcal{Y}$

$\mathcal{Y} = \{A, B, C, D\}$

$g(x, y)$: score for option y of question x

Threshold Estimation

Calibration Data $D_{\text{cal}} = \{(x_i, y_i^\star)\}_{i=1}^n$

Coverage $\widehat{\mathcal{P}}(g, \tau) = \dfrac{\#\text{times } y_i^\star \in C(x_i; g, \tau)}{n}$

Avg. set size $\widehat{S}(g, \tau) = \dfrac{1}{n} \sum_{i=1}^{n} \left| C(x_i; g, \tau) \right|$

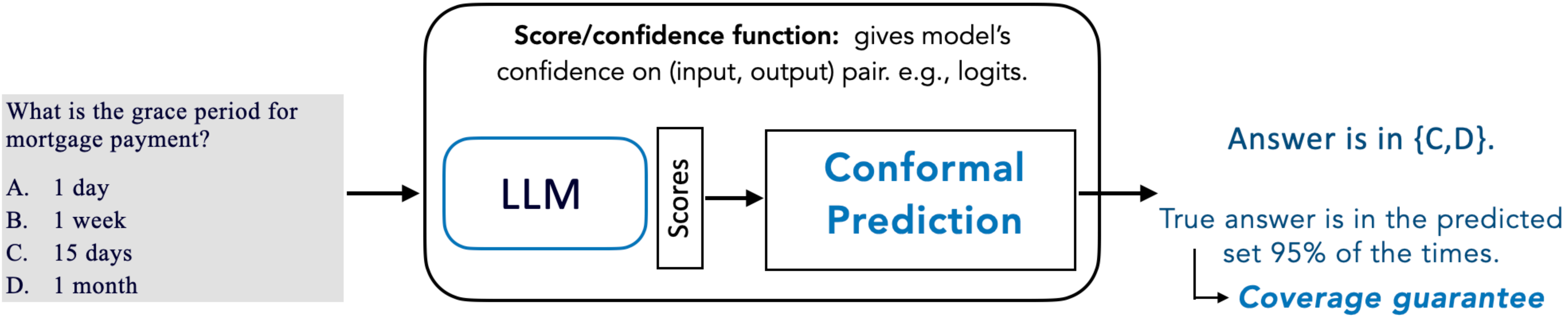
Find smallest $\hat{\tau}_\alpha$ such that $\widehat{\mathcal{P}}(g, \tau) \geq 1 - \alpha$

Coverage Guarantee

For x_{test} predict $C(x_{\text{test}}; g, \hat{\tau}_\alpha)$

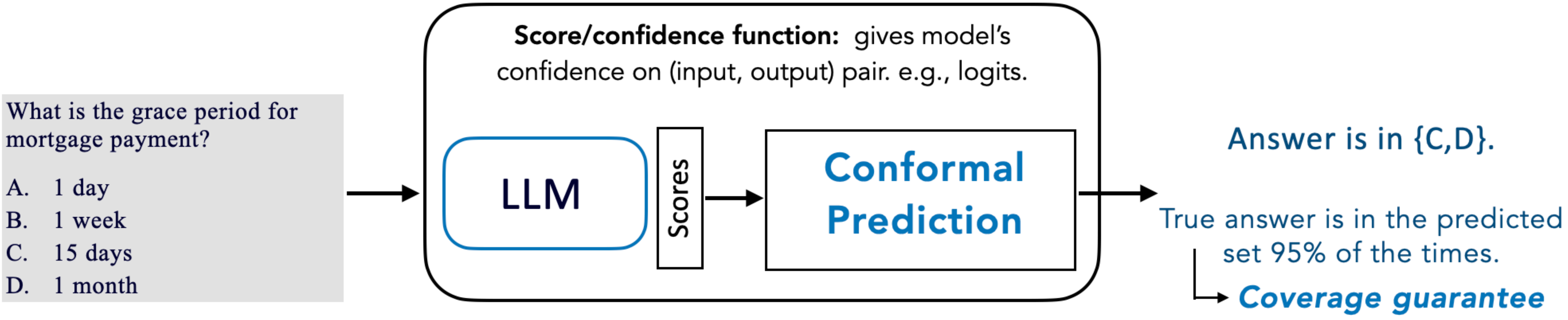
$$\mathbb{P}(\, y_{\text{test}}^\star \in C(x_{\text{test}}; g, \hat{\tau}_\alpha)\,) \;\geq 1 - \alpha$$

Set predictions are good but might want point predictions eventually



Set predictions are safer and quantify uncertainty!

Set predictions are good but might want point predictions eventually

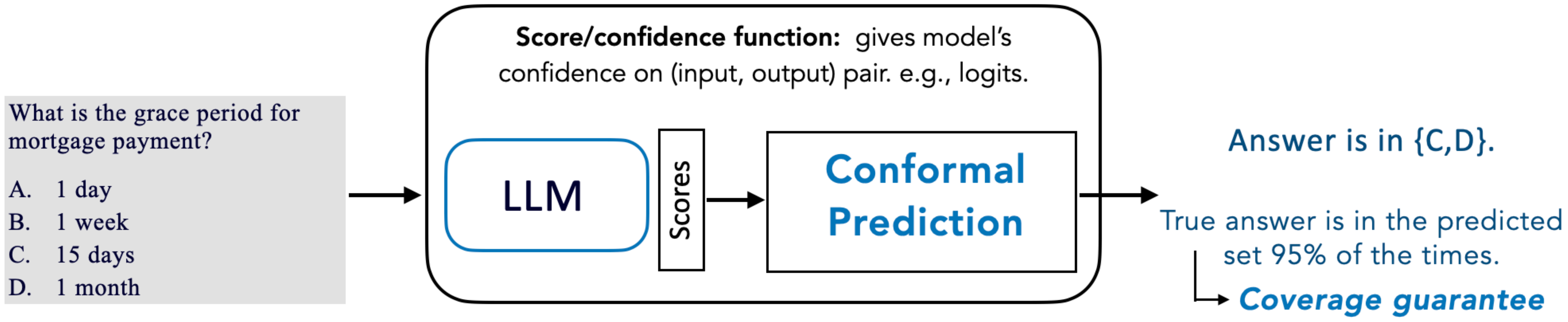


Set predictions are safer and quantify uncertainty!

Large set ⇒ High uncertainty ⇒ Reject the output /defer to expert.

Small set ⇒ Low uncertainty ⇒ Accept the output.

Set predictions are good but might want point predictions eventually



Set predictions are safer and quantify uncertainty!

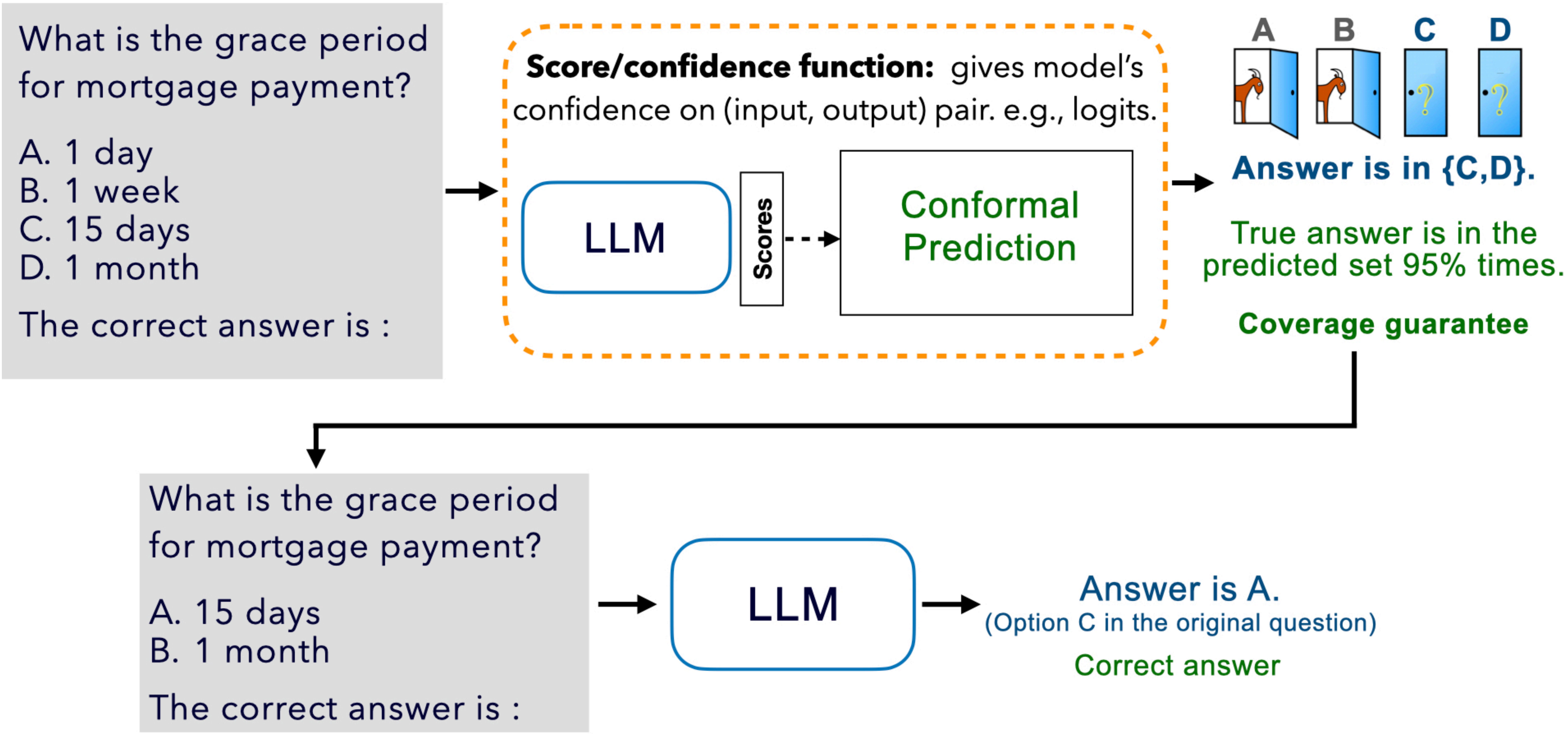
Large set ⇨ High uncertainty ⇨ Reject the output /defer to expert.

Small set ⇨ Low uncertainty ⇨ Accept the output.

More accurate point predictions might still be desired.

Conformal Revision of Question (CROQ)

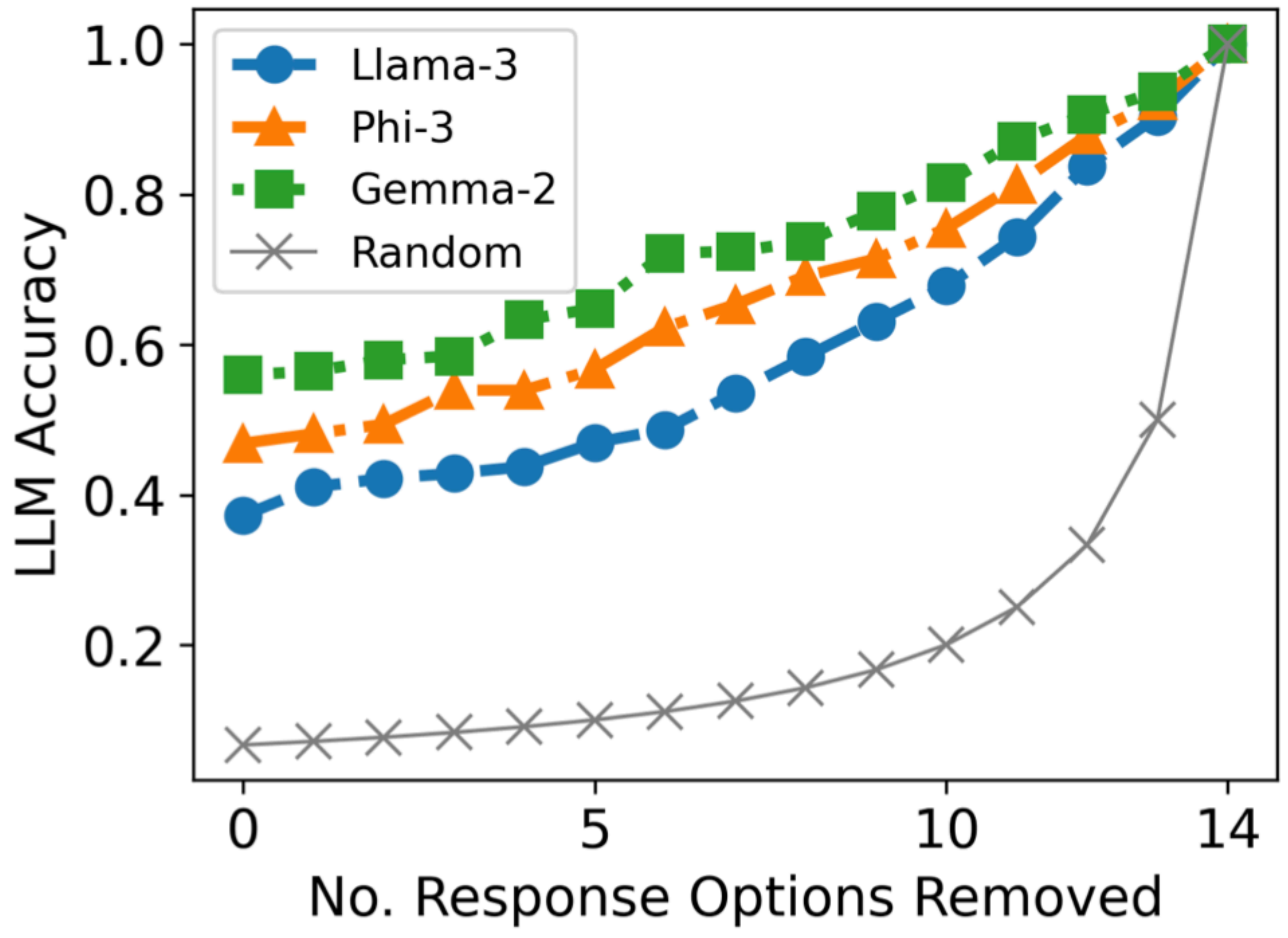
Expectation: Reduction in uncertainty should help LLM answer correctly.



Steps are fully automated.

Accuracy increases monotonically with the number of eliminated noisy options

Controlled Experiment

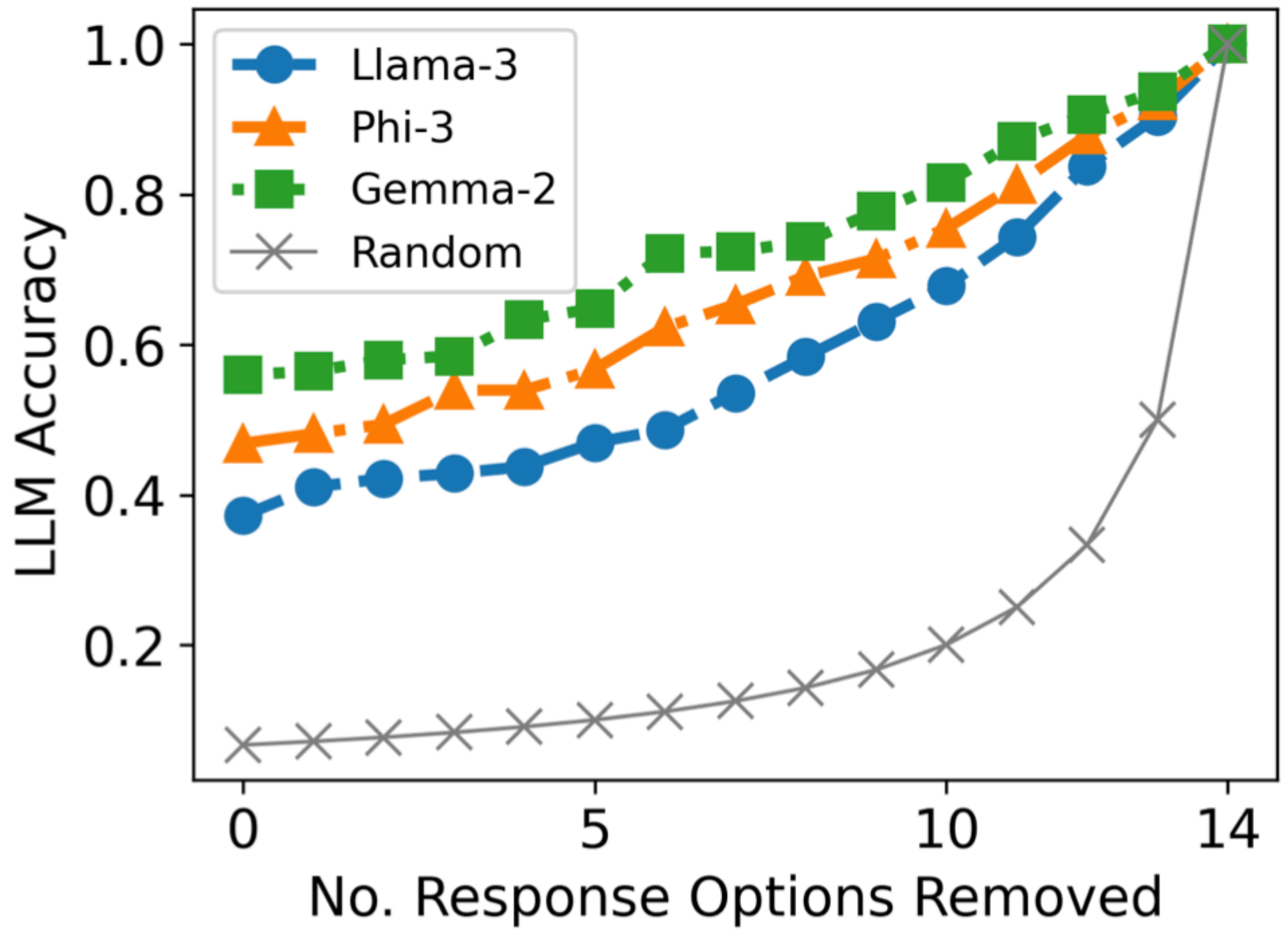


Dataset: Truthful QA

Remove $k \in \{0,1,\ldots,14\}$ **distractor** choices at random, ask the revised question to LLM and observe the accuracy.

Accuracy increases monotonically with the number of eliminated noisy options

Controlled Experiment



Dataset: Truthful QA

Remove $k \in \{0,1,\ldots,14\}$ **distractor** choices at random, ask the revised question to LLM and observe the accuracy.

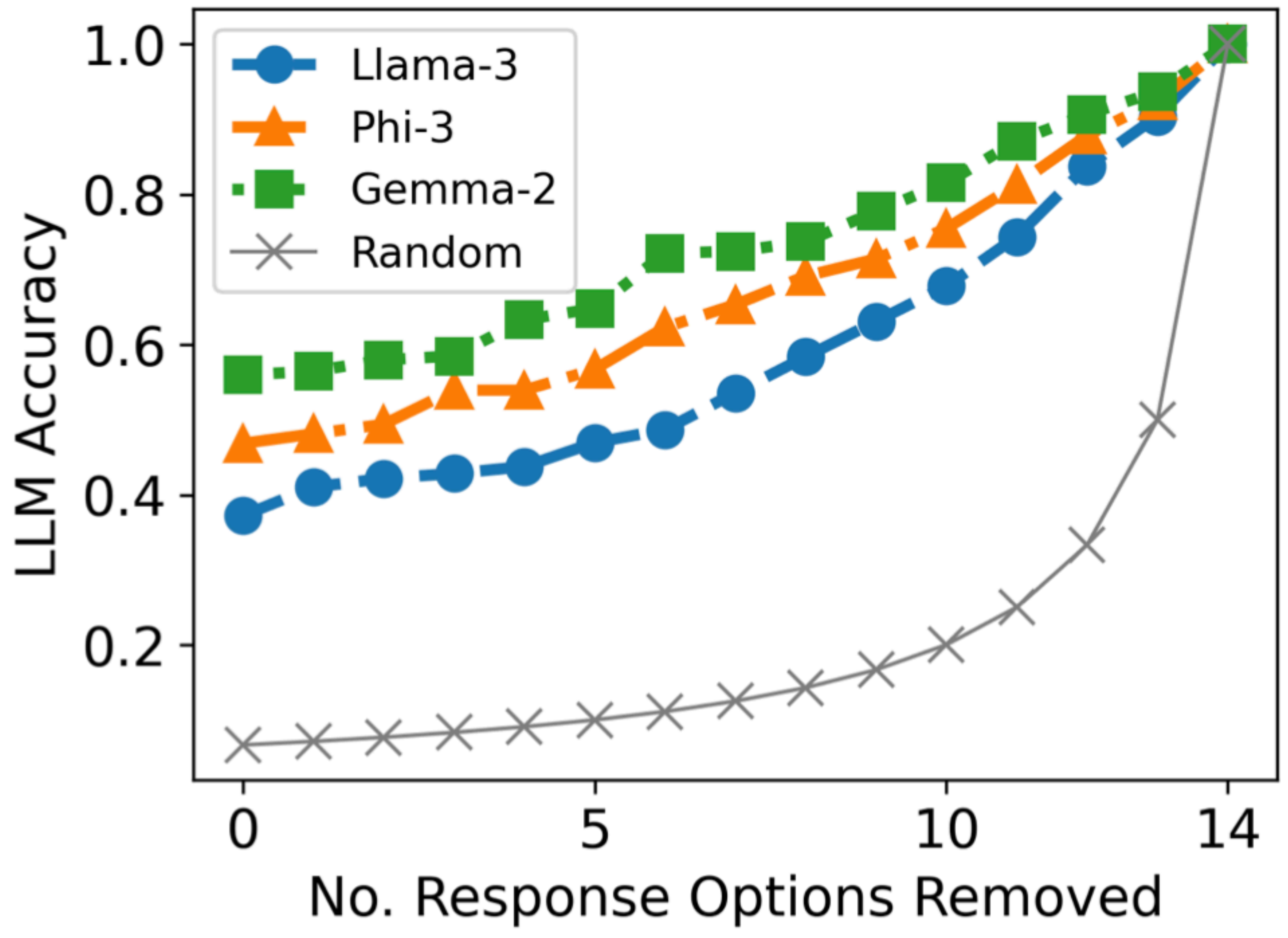
Smaller sets at high coverage

↓

Higher accuracy with CROQ.

Accuracy increases monotonically with the number of eliminated noisy options

Controlled Experiment



Dataset: Truthful QA

Remove $k \in \{0,1,\ldots,14\}$ **distractor** choices at random, ask the revised question to LLM and observe the accuracy.

Smaller sets at high coverage

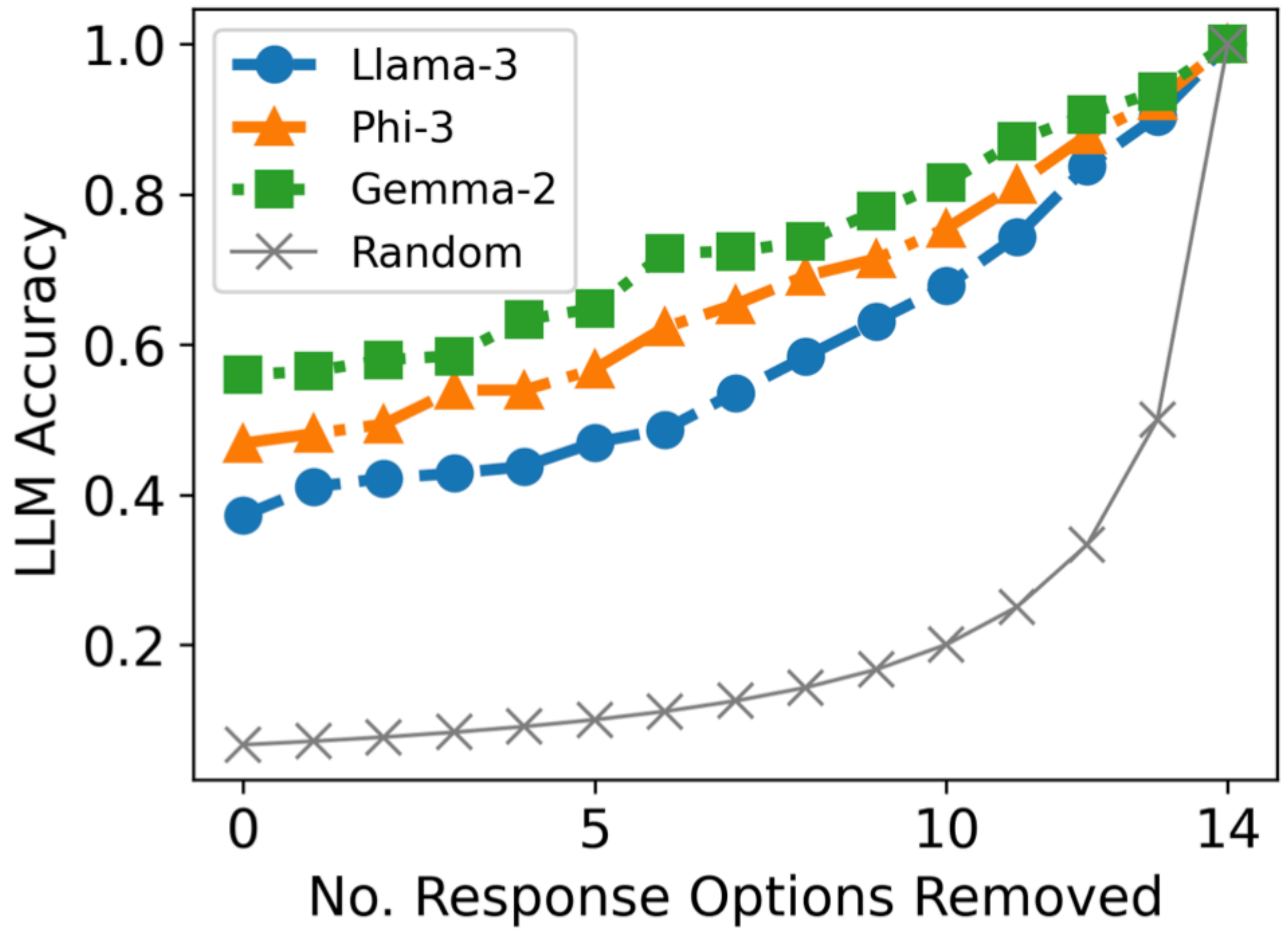
↓

Higher accuracy with CROQ.

But commonly used scores (e.g. logits) tend to give large sets.

Accuracy increases monotonically with the number of eliminated noisy options

Controlled Experiment



Dataset: Truthful QA

Remove $k \in \{0,1,\ldots,14\}$ **distractor** choices at random, ask the revised question to LLM and observe the accuracy.

Smaller sets at high coverage

↓

Higher accuracy with CROQ.

But commonly used scores (e.g. logits) tend to give large sets.

Optimize scores for CP

CP-OPT: Optimize Scores for Smaller Prediction Sets

Objective

1. Minimize set size,
2. Ensure coverage is at least $1 - \alpha$.

$\alpha \in (0,1)$.

CP-OPT: Optimize Scores for Smaller Prediction Sets

Objective

1. Minimize set size,
2. Ensure coverage is at least $1 - \alpha$.

$\alpha \in (0,1)$.

Avg. set size

$$\widehat{S}(g, \tau) = \frac{1}{n} \sum_{i=1}^{n} \left| C(x_i; g, \tau) \right|$$

Coverage

$$\widehat{\mathscr{P}}(g, \tau) = \frac{\text{\# times } y_i^\star \in C(x_i; g, \tau)}{n}$$

CP-OPT

$$\tilde{g}, \tilde{\tau}' \in \arg \min_{g:\mathscr{X}\times\mathscr{Y}\mathbb{R},\, \tau\in\mathbb{R}} \widetilde{S}(g, \tau) + \quad \lambda\left(\widetilde{\mathscr{P}}(g, \tau) - 1 + \alpha\right)^2$$

Smooth surrogates of avg. set size and coverage

Estimated on part of calibration data.

Solve using SGD.

CP-OPT: Optimize Scores for Smaller Prediction Sets

Objective

1. Minimize set size,
2. Ensure coverage is at least $1 - \alpha$.
   $\alpha \in (0,1)$.

Avg. set size

$$\widehat{S}(g,\tau) = \frac{1}{n}\sum_{i=1}^{n} \left| C(x_i; g, \tau) \right|$$

Coverage

$$\widehat{\mathscr{P}}(g,\tau) = \frac{\text{\# times } y_i^\star \in C(x_i; g, \tau)}{n}$$

CP-OPT

$$\tilde{g}, \tilde{\tau}' \in \arg \min_{g:\mathscr{X}\times\mathscr{Y}\mathbb{R},\, \tau\in\mathbb{R}} \widetilde{S}(g,\tau) + \quad \lambda\left(\widetilde{\mathscr{P}}(g,\tau) - 1 + \alpha\right)^2$$

Smooth surrogates of avg. set size and coverage

Estimated on part of calibration data.

Solve using SGD.

Expectation: using $\tilde{g}$ in CP will give smaller sets and maintain coverage at $1 - \alpha$.

Empirical Evaluation and Results

Models

Phi-3 Microsoft/Phi-3-4k-mini-Instruct

Llama-3 Meta/Llama-3-8B-Instruct

Gemma-2 Princeton-NLP/gemma-2-9b-it-SimPO

Datasets

MMLU, TruthfulQA, ToolAlpaca

Introduce noisy options to get 3 variations of each dataset with default (4 or 5), 10, 15 options

Total 27 settings

Empirical Evaluation and Results

Models

Phi-3	Microsoft/Phi-3-4k-mini-Instruct
Llama-3	Meta/Llama-3-8B-Instruct
Gemma-2	Princeton-NLP/gemma-2-9b-it-SimPO

Datasets

MMLU, TruthfulQA, ToolAlpaca

Introduce noisy options to get 3 variations of each dataset with default (4 or 5), 10, 15 options

Total 27 settings

CP-OPT reduces set sizes by up to 50% (relative) while maintaining coverage.

Empirical Evaluation and Results

Models

Phi-3	Microsoft/Phi-3-4k-mini-Instruct
Llama-3	Meta/Llama-3-8B-Instruct
Gemma-2	Princeton-NLP/gemma-2-9b-it-SimPO

Datasets

MMLU, TruthfulQA, ToolAlpaca

Introduce noisy options to get 3 variations of each dataset with default (4 or 5), 10, 15 options

Total 27 settings

CP-OPT reduces set sizes by up to 50% (relative) while maintaining coverage.

CROQ boosts accuracy by up to 6.4 % with logits and 7.2% with CP-OPT scores.

Limitations and Future work

Limitations and Future work

- CROQ was limited to two rounds and used the same LLM in both rounds.

Limitations and Future work

- CROQ was limited to two rounds and used the same LLM in both rounds.
- Multiple rounds of CROQ can yield more gains.

Limitations and Future work

- CROQ was limited to two rounds and used the same LLM in both rounds.
- Multiple rounds of CROQ can yield more gains.
- Intermediate rounds can use cheaper LLMs or other sources to get scores for CP-based pruning for computational efficiency.

Limitations and Future work

- CROQ was limited to two rounds and used the same LLM in both rounds.
- Multiple rounds of CROQ can yield more gains.
- Intermediate rounds can use cheaper LLMs or other sources to get scores for CP-based pruning for computational efficiency.
- CP-OPT used features from the last two layers of LLM.
  Using more expressive features could make it more effective.

Thank You!

Contact

hvishwakarma@cs.wisc.edu

alan.mishler@jpmorgan.com

Disclaimer

This presentation was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.