Monty Hall and Optimized Conformal Prediction to Improve Decision-Making with LLMs

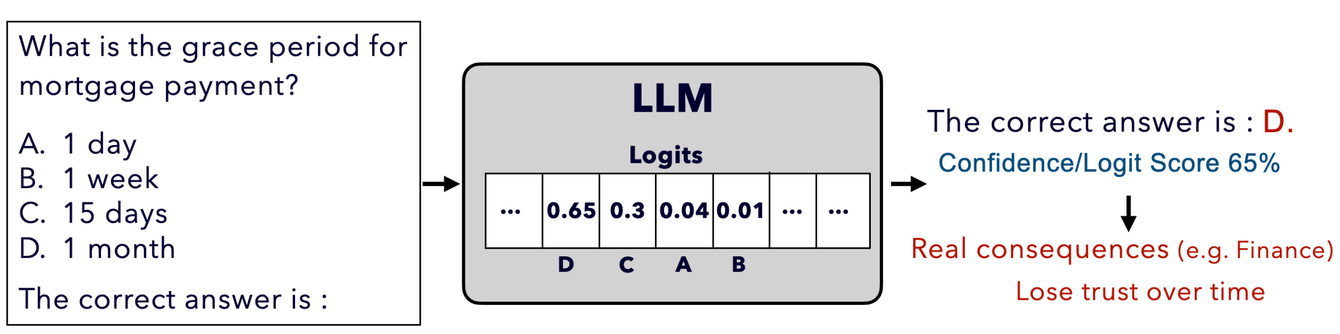
Harit Vishwakarma★ Alan Mishler Thomas Cook Niccoló Dalmasso
Natraj Raman Sumitra Ganesh

★University of Wisconsin-Madison, WI, USA
JPMorganChase AI Research New York, NY, USA
hvishwakarma@cs.wisc.edu

Motivation

Decide between finite choices e.g., MCQ, Tool Selection, etc.

LLMs can output incorrect answers with high confidence.



Can we improve uncertainty quantification (UQ) and accuracy without heavy fine-tuning?

Conformal prediction (CP) can help in UQ

Kumar et al., 2023, Su et al., 2024

Output high coverage sets e.g., the correct answer is in: **{C, D}**

Large set ⇨ *High uncertainty* **Small set ⇨ *Low uncertainty***

Summary

Prior works proposed conformal prediction (CP) for uncertainty quantification in LLMs but have limitations

Used ad-hoc scores (logits, self-consistency)
Unreliable, expensive to compute
Lead to large sets ⟹ Less useful

Utility of sets (CP) beyond UQ is underexplored

Our work

A principled method (CP-OPT) to learn scores

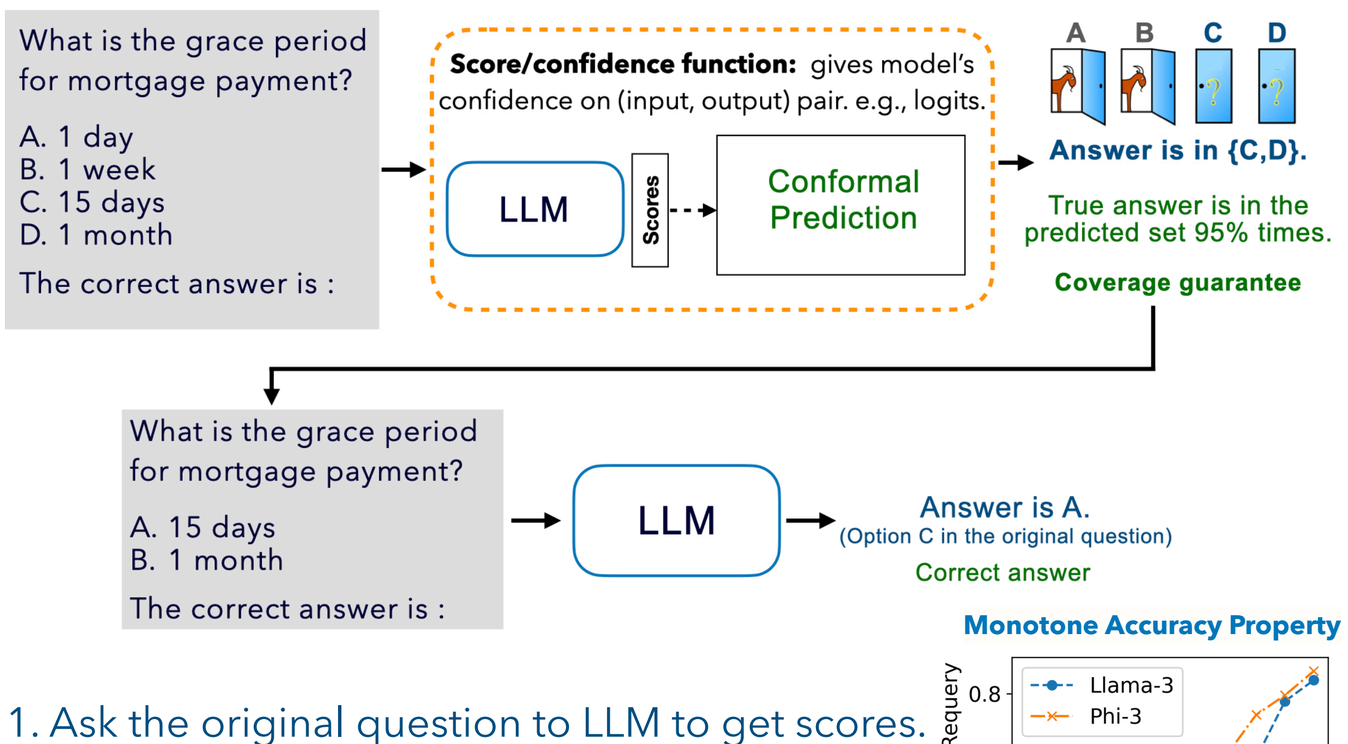
Revise question with the choices in the predicted set and ask LLM again (CROQ)

CP-OPT ⟹ Smaller sets + High coverage

CROQ + CP-OPT ⟹ Better accuracy

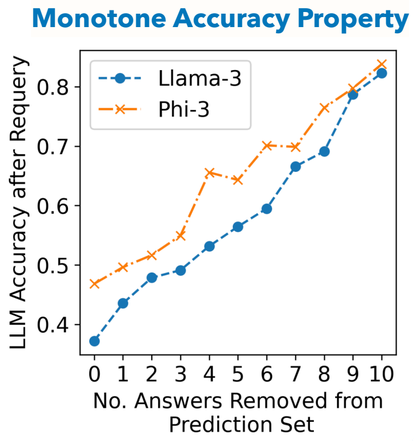
Conformal revision of question (CROQ)
Inspired from Monty Hall

Expectation: *Reduction in uncertainty should help LLM answer correctly just as in Monty Hall game.*

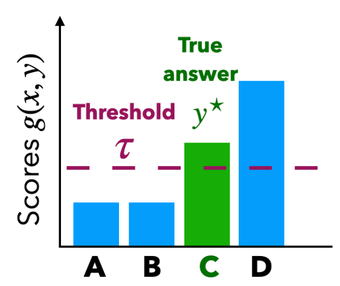


1. Ask the original question to LLM to get scores.
2. Run CP to get prediction set S.
3. Eliminate the choices that are not in the set S.
4. Ask the revised question to LLM to get final output.

Monotone Accuracy Property



Score optimization (CP-OPT)

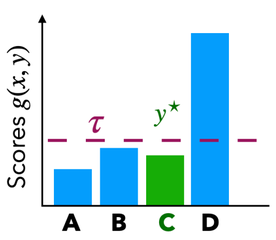


Prediction set

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

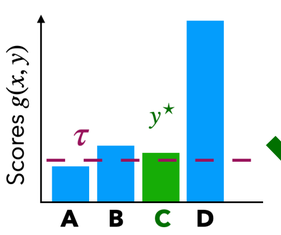
Output {C, D}

Effect of scores in CP



Output **{D}**

Miscoverage ✗



Output **{B, C, D}**

✓ Coverage

✗ Large set size

Avg. set size

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

Coverage

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

(CP-OPT)

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

Use $\hat{g}$ to get scores for CP

Empirical Results

(H1) *CP-OPT reduces set sizes while maintaining coverage*

* indicates statistically significant difference based on paired t-test.

		Llama-3				Gemma-2			
		Avg. Set Size		Coverage		Avg. Set Size		Coverage	
Dataset	# Opt.	Logits	Ours	Logits	Ours	Logits	Ours	Logits	Ours
MMLU	4	2.56	**2.53***	95.75	95.57	2.94	**2.40***	95.16*	94.23
	10	5.53	**4.90***	96.06*	95.45	7.79	**6.08***	95.00*	94.04
	15	7.69	**7.18***	95.42	95.06	11.71	**10.04***	94.58	94.58
ToolAlpaca	4	**1.17**	1.18	97.08	96.85	1.12	**1.05***	95.68	95.44
	10	1.51	**1.39***	95.21	95.56	2.05	**1.42***	95.56	94.51
	15	1.97	**1.67***	96.50	96.03	3.54	**1.77***	96.14	95.21
TruthfulQA	4	3.34	**2.69***	95.95*	92.41	2.74	**1.88***	96.46	95.44
	10	7.06	**6.41***	94.43	93.42	7.52	**5.64***	95.44	97.22
	15	**10.61**	10.62	94.68	94.68	11.23	**9.35***	95.44	96.46

(H2) *CROQ with logit and CP-OPT scores improves accuracy*

a_1 Baseline accuracy before CROQ
a_1' CROQ accuracy with logits

		Llama-3			Gemma-2		
Model	# Opt.	Accuracy Before (a_1)	Accuracy After (a_1')	Gain ($a_1' - a_1$)	Accuracy Before (a_1)	Accuracy After (a_1')	Gain ($a_1' - a_1$)
MMLU	4	**64.02**	63.83	-0.19	67.62	**67.70**	**0.07**
	10	54.82	**56.29**	**1.47***	53.80	**53.93**	**0.13**
	15	51.99	**54.11**	**2.11***	**50.78**	50.58	-0.20
ToolAlpaca	4	91.47	**91.94**	**0.47**	**93.46**	93.11	-0.35
	10	85.16	**88.67**	**3.50***	87.73	**89.60**	**1.87***
	15	81.43	**87.85**	**6.43***	87.97	**88.55**	**0.58**
TruthfulQA	4	54.43	**55.19**	**0.76**	74.68	**74.94**	**0.25**
	10	39.24	**40.76**	**1.52**	**56.46**	56.20	-0.25
	15	37.22	**37.22**	**0.00**	55.95	**56.96**	**1.01**

(H3) *CROQ with CP-OPT performs better than CROQ with logits.*

a_1' CROQ accuracy with logits
a_2' CROQ accuracy with CP-OPT

		LLama-3			Gemma-2		
Model	# Opt.	Accuracy Before (a_2)	Accuracy After (a_2')	Gain ($a_2' - a_2$)	Accuracy Before (a_2)	Accuracy After (a_2')	Gain ($a_2' - a_2$)
MMLU	4	**64.02**	63.67	-0.34	68.36	**69.56**	**1.20***
	10	54.82	**57.11**	**2.29***	53.99	**57.93**	**3.94***
	15	51.99	**54.77**	**2.78***	50.78	**51.31**	**0.52**
ToolAlpaca	4	91.47	**91.82**	**0.35**	93.46	**93.57**	**0.12**
	10	85.16	**89.02**	**3.86***	88.08	**90.42**	**2.34***
	15	81.43	**88.67**	**7.24***	88.08	**89.37**	**1.29**
TruthfulQA	4	54.43	**55.44**	**1.01**	74.94	**76.96**	**2.03**
	10	39.24	**42.28**	**3.04**	56.46	**60.76**	**4.30***
	15	37.22	**37.47**	**0.25**	55.95	**57.72**	**1.77**

★ This work was performed while at JPMorganChase.

This poster was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorganChase and its affiliates ("J.P. Morgan") and is not a product of the Research Department of JPMorganChase. J.P. 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.