Conformal Information Pursuit for Interactively Guiding Large Language Models

Kwan Ho Ryan Chan Yuyan Ge Edgar Dobriban Hamed Hassani René Vidal

University of Pennsylvania

Project

Motivation: Interactive Question Answering

- In Question-Answering (QA), context and information **may not be readily available all at once and require asking informative questions.**
- Can we guide Large Language Models (LLMs) to solve QA tasks **interactively** (Fig. 1)?
- We propose an information-theoretic framework to interactively guide an LLM to ask queries that are most informative for the QA task.
- The challenge lies in estimating uncertainty from LLMs.

Init. Info: *A 25-year-old woman comes to the physician for an examination.*

Uncertain → Ask a query

Doctor: *Is the patient's body temperature within the normal range?*

Patient: *The patient's body temperature is 36.6 ℃ (98.0 ℉).*

⋮

Uncertain → Ask a query

Doctor: *What is the serum ferritin level in ng/mL?*

Patient: *The patient's serum ferritin level is 170 ng/mL.*

Confident → Final Prediction

Prediction: (D) *Intravascular hemolysis*

Fig. 1: Example Interaction between Patient and Doctor LLM.

Challenge: Uncertainty Estimation for LLMs

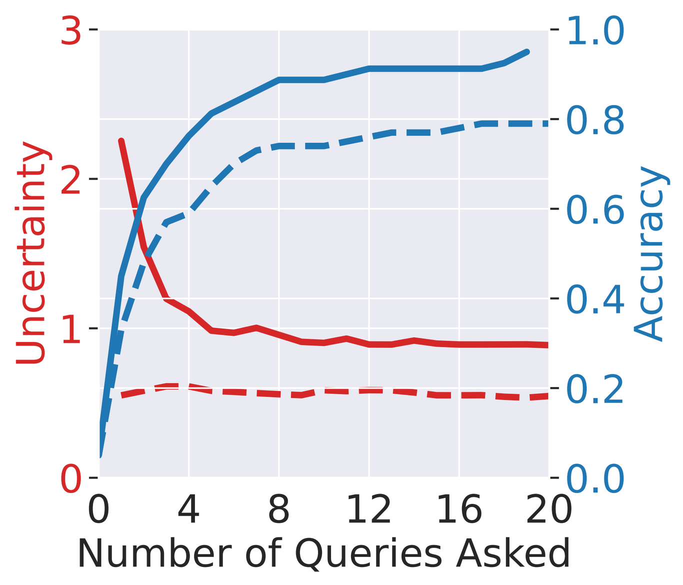


Fig. 2: Simulated 20Q Game with LLMs. (Dashed: uncalibrated; Solid: Calibrated)

- We focus on multiple-choice QA tasks, where an LLM answers a question via next token probability:

$$f(x)_y = \widehat{\mathbb{P}}_{\text{LLM}}(y \mid x)$$

- **Issue**: $\widehat{\mathbb{P}}_{\text{LLM}}$ might be **noisy/miscalibrated**, leading to poor estimates of uncertainty.
- **Proposal**: Leverage **Conformal Prediction!**

Proposed Method

- Let $\mathcal{Q} = \{q : \mathcal{X} \to \mathcal{A}\}$ be a set of task-relevant, textual queries about the data (e.g., $q =$"What is the temperature of the patient?").

Prior Work: Information Pursuit (IP)

- Given a test sample $\hat{x} \in \mathcal{X}$, IP interactively and sequentially selects queries whose answers are most informative for the task Y:

$$q_1 = \underset{q\in\mathcal{Q}}{\text{argmax}}\, I(Y; q(X)) = \underset{q\in\mathcal{Q}}{\text{argmin}}\, H(Y \mid q(X))$$

$$q_{k+1} = \underset{q\in\mathcal{Q}}{\text{argmin}}\, I(Y; q(X) \mid q_{1:k}(\hat{x})) = \underset{q\in\mathcal{Q}}{\text{argmin}}\, H(Y \mid q(X), q_{1:k}(\hat{x}))$$

- At each iteration, IP selects the query that minimizes entropy.
- IP terminates when the mutual information is less than ϵ.

Proposed: Conformal Information Pursuit (C-IP)

- Rather than entropy, we leverage (split) Conformal Prediction and use **average sizes of prediction sets** to estimate uncertainty.
- Conformalize IP as follows:

1. *Define the prediction set:*

$$\mathcal{C}_{\hat{\tau}}(q_{1:k}(X)) = \{y \in \mathcal{Y} \mid f(q_{1:k}(X))_y > \hat{\tau}\}$$

2. *Obtain calibration samples by running simulations with LLMs and construct prediction sets that satisfy the marginalized guarantee:*

$$\mathbb{P}_{X,Y,Q_{1:k}}(Y \in \mathcal{C}_k(q_{1:k}(X))) \approx 1 - \alpha \quad \text{for } k = 1, \ldots, L$$

3. *Select queries that minimize log-expected length at each iteration:*

$$q_1 = \text{argmin}_{q\in\mathcal{Q}} \log \mathbb{E}_X[|\mathcal{C}_{\hat{\tau}(1)}(q(X))|]$$

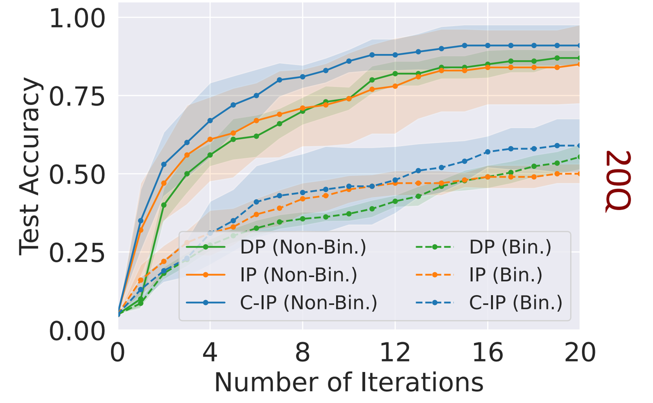
$$q_{k+1} = \text{argmin}_{q\in\mathcal{Q}} \log \mathbb{E}_X[|\mathcal{C}_{\hat{\tau}(k+1)}(q(X))| \mid q_{1:k}(\hat{x})]$$

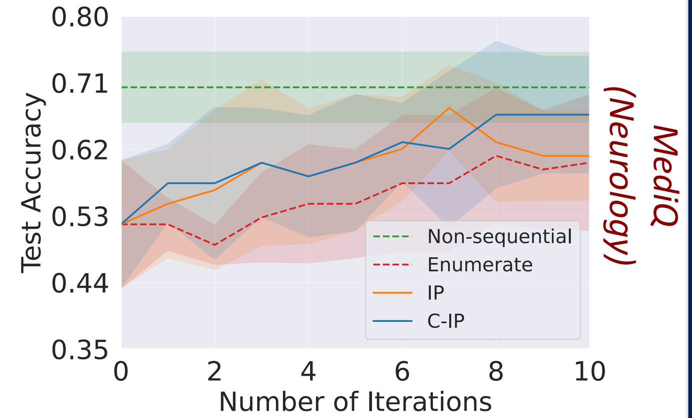
4. *Terminate algorithm when length < ϵ, then make a prediction:*

$$\hat{y} = \text{argmax}_{y\in\mathcal{Y}} \mathbb{P}(y \mid q_{1:k}(\hat{x}))$$

Experiments

- We evaluate C-IP on two tasks: **20 Questions** (20Q) and **Interactive Medical QA** (MediQ).
  - 20Q: *A "Querier" LLM samples queries at each iteration, and an "Answerer" LLM evaluates whether they are true.*
  - MediQ [Li et al. 24]: An "*Expert*" LLM asks questions about a patient and makes predictions, and a "*Patient*" LLM answers those questions.





Example of Interactive Medical Question Answering

Context

Initial Info: *An 84-year-old woman with an indwelling urinary catheter and a history of recurrent nephrolithiasis is brought to the emergency.*
Question: *Which of the following is most likely to be present on this patient's urine culture?*
Options: *(A) Gram-negative, oxidase-positive rods, (B) Gram-positive, novobiocin-resistant cocci, (C) Gram-positive, gamma-hemolytic cocci, (D) Gram-negative, oxidase-negative rods.*

History $q_{1:K}(x^{obs})$

q_1:	*What is the moisture level of the patient's mucous membranes?*
$q_1(x^{obs})$:	*She has dry mucous membranes.*
q_2:	*What is the white blood cell count in the urine?*
$q_2(x^{obs})$:	*The patient is oriented to person, and the urine studies show WBC of 40/HPF.*
q_3:	*What is the patient's current body temperature?*
$q_3(x^{obs})$:	*The patient's current body temperature is 38.3°C (100.9°F).*
q_4:	*Why was she brought to the hospital?*
$q_4(x^{obs})$:	*She was brought to the emergency department from her nursing home because of increasing confusion over the past day.*
q_5:	(Stopping Criteria Reached, Output Final Prediction)

Prediction

Final Prediction: *D.* **Explanation:** *The patient's symptoms and presentation suggest a urinary tract infection (UTI). The patient has an indwelling urinary catheter, which is a risk factor for UTIs. (more omitted)*

Theoretical Justification

- Why does our derivation make sense?
- How are entropy and expected length of the prediction set related?
- Proposition: [Correia et al 24]. If prediction the set $\mathcal{C}$ satisfies the marginal guarantee $\mathbb{P}_{X,Y}(Y \in \mathcal{C}_\tau(X)) \approx 1 - \alpha$, then

$$H(Y \mid X) \leq \text{constants} + (1 - \alpha) \log \mathbb{E}_X[|\mathcal{C}_\tau(X)|]$$