Conformal Information Pursuit for Interactively Guiding Large Language Models

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

University of Pennsylvania

Abstract

A significant use case of instruction-finetuned Large Language Models (LLMs) is to solve question-answering tasks interactively. In this setting, an LLM agent is tasked with making a prediction by sequentially querying relevant information from the user, as opposed to a single-turn conversation. This paper explores sequential querying strategies that aim to minimize the expected number of queries. One such strategy is Information Pursuit (IP), a greedy algorithm that at each iteration selects the query that maximizes information gain or equivalently minimizes uncertainty. However, obtaining accurate estimates of mutual information or conditional entropy for LLMs is very difficult in practice due to over- or under-confident LLM probabilities, which leads to suboptimal query selection and predictive performance. To better estimate the uncertainty at each iteration, we propose *Conformal Information Pursuit (C-IP)*, an alternative approach to sequential information gain based on conformal prediction sets. More specifically, C-IP leverages a relationship between prediction sets and conditional entropy at each iteration to estimate uncertainty based on the average size of conformal prediction sets. In contrast to conditional entropy, we find that conformal prediction sets are a distribution-free and robust method of measuring uncertainty. Experiments with 20 Questions show that C-IP obtains better predictive performance and shorter query-answer chains compared to previous approaches to IP and uncertainty-based chain-of-thought methods. Furthermore, extending to an interactive medical setting between a doctor and a patient on the MediQ dataset, C-IP achieves competitive performance with direct single-turn prediction while offering greater interpretability.

1 Introduction

Large Language Models (LLMs) [2, 3, 7, 39, 78] fine-tuned on instruction datasets are capable of generating human-like dialogues. Although many use-cases constrain these models to single-turn interactions [93, 105, 132], where a single prompt typically provides all context and examples, more realistic conversations behave interactively [12, 65]. In this work, we are interested in the setting where LLMs interactively solve a prediction task in a multi-turn, information-seeking environment by asking queries and obtaining information sequentially [66]. Popular examples of such settings include the game of 20 Questions (20Q) [11, 79, 137] and Interactive Medical Question Answering [65] (Figure 1).

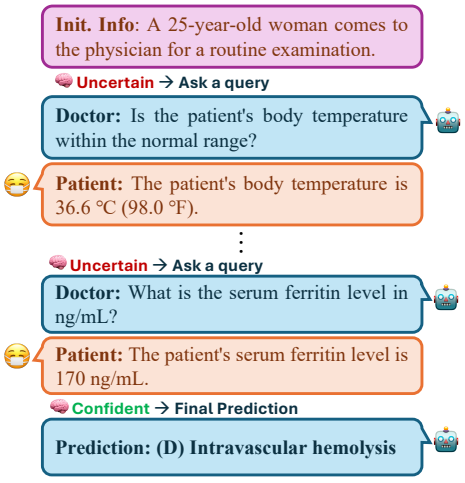


Figure 1: Example of diagnosis via Patient & Doctor LLM interaction.

While there are many strategies to interactively arrive at a final prediction, we consider strategies that yield a small number of queries. One such strategy is to always ask the most "informative"

*Corresponding Author. Email: ryanckh@seas.upenn.edu

query given the prior history, until the model is confident enough to make a prediction. This process is characterized in a framework called Information Pursuit (IP) [14, 35], which defines the most informative query as the one that *maximizes information gain* at each iteration. This requires selecting the query whose answer has the maximum mutual information [16, 19, 28, 72, 139] with the prediction, given the history of prior query-answer pairs. Equivalently, one can select the query whose answer minimizes the uncertainty of the prediction, as quantified by the conditional entropy.

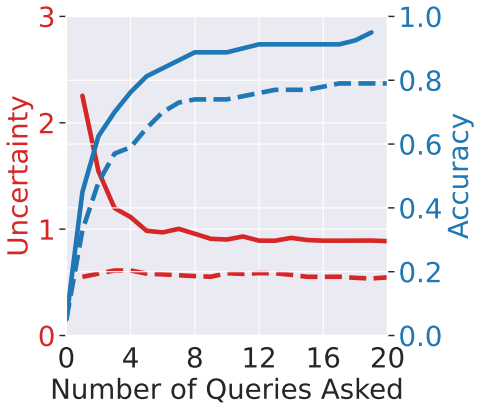


Figure 2: IP with calibrated (solid) versus uncalibrated (dashed) measures of uncertainty.

In practice, implementing IP for LLMs requires estimating the conditional entropy, which is very challenging because the probabilities of LLM outputs, such as logits of output tokens, are often *miscalibrated*: LLM predictions are often strongly over- or under-confident [6, 29, 36, 51, 73, 103, 107, 112, 115, 122, 126, 136, 138, 141, 142]. Consequently, using probability estimates to measure uncertainty leads to inaccurate estimates of information and lower predictive performance. We illustrate this phenomenon in Figure 2 using the example of two LLMs playing 20 Questions. In this setting, the Querier LLM has to guess the object that the Answerer LLM is thinking of by asking a sequence of most informative questions. Ideally, the uncertainty and accuracy of the prediction should behave like the solid red and solid blue curves. At the first iteration, since the Querier LLM has not gathered sufficient information, the uncertainty of the prediction should be large. As iterations proceed, uncertainty should diminish and converge to a certain level, where any new query would no longer offer any additional information than what the Querier LLM already knows. Then, the Querier LLM should stop and make a prediction. However, we observe that using the LLM's output probabilities to estimate uncertainty yields a flat and uninformative curve (dashed red curve). Naturally, this leads to suboptimal query selection and lower predictive performance (dashed blue curve). This motivates the need for better uncertainty estimation. To obtain more accurate measures of uncertainty for the outputs of the LLM, we propose an alternative approach to sequential information gain that leverages the notion of *prediction sets* from predictive inference and conformal prediction [5, 118]. Although the history of prediction sets dates back to the 1940s [97, 113, 114, 119, 127], the use of the expected size of a prediction set as a measure of uncertainty has been studied and popularized only recently [4, 26, 30, 54, 77, 94, 95]. In more recent applications, it has also been used to calibrate the LLMs' output probabilities for a more accurate measure of model confidence [21, 56, 75, 89, 134]. By and large, standard entropy as a measure of uncertainty relies heavily on proper estimation of the target distribution, while prediction sets from conformal prediction allow for a distribution-free and instance-wise measure, offering more flexibility, robustness, and interpretability. Given its dual purpose of being a tool for uncertainty quantification and calibration, we are motivated to use prediction sets as a measure of informativeness in lieu of conditional entropy.

Therefore, in this paper we propose **Conformal Information Pursuit (C-IP)**, a greedy sequential approach that asks informative queries in the order of information gain by minimizing prediction set size instead of the standard entropy. We first leverage the relationship between entropy and the expected size of prediction sets using results from Correia et al. [25], allowing us to upper bound the entropy of the target distribution given the history of query-answer pairs by the expected logarithm of the size of the prediction set for the ground-truth answer given the history. Then, utilizing conformal prediction, we propose two ways of constructing sequences of prediction sets, which allows us to provide accurate measures of informativeness for any given query at any iteration of the algorithm. To demonstrate the properties and behavior of our method, we perform a thorough study of our method in the setting of 20Q. Finally, we apply our method to the setting of interactive medical question answering on the MediQ [65] dataset. Our results demonstrate not only the applicability of our method in real-world settings, but also its competitive predictive performance.

2 Information Pursuit: A Framework for Sequential Information Gain

We first review the Information Pursuit (IP) framework [14, 35, 47], which will help us lay the groundwork for developing a greedy algorithm that asks queries using the information gain criterion [19, 28, 47, 139]. Throughout the paper, (X, Y) denotes an input-output pair, where $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ are random variables, Ω is the sample space, and $\mathcal{X}$ and $\mathcal{Y}$ are the sets of values that they can take on. We denote by $I(X;Y)$ the mutual information between X and Y, by $H(X)$ the entropy of the random variable X, and by $H(Y \mid X)$ the conditional entropy of X given Y.

Setup. The modern generative version of IP is a framework for making interpretable predictions: the user defines a set $\mathcal{Q}$ of task-relevant queries. Each query $q \in \mathcal{Q}$ is a function $q : \mathcal{X} \to \mathcal{A}$ that maps the textual input $x \in \mathcal{X}$ to an answer $q(x) \in \mathcal{A}$. For example, for the task of disease classification, $\mathcal{X}$ contains the set of patient features (e.g., clinical notes), $\mathcal{Y}$ is the set of disease labels (e.g., "pneumonia", "cardiomegaly"), $\mathcal{Q}$ is a set of queries about the patient such as "Are you coughing?" or "Do you have a fever?", and $\mathcal{A}$ to be the set of possible query answers (e.g. "Yes, I do", "The patient feels warm."). Last but not least, for our question-answering task, all query-answers $\mathcal{A}$ and labels $\mathcal{Y}$ are a subset of the space of natural language $\mathcal{X}$.

An important desired property of the query set is *sufficiency*, which states that for all x, y,

$$\mathbb{P}(y \mid x) = \mathbb{P}(y \mid \{x' \in \mathcal{X} : q(x) = q(x'),\ \forall q \in \mathcal{Q}\}). \tag{1}$$

That is, the answers to all queries about x, $\mathcal{Q}(x) := \{q(x) : q \in \mathcal{Q}\}$, determine the *posterior* $\mathbb{P}(y \mid x)$. In this work, we consider obtaining $\mathcal{Q}$ either from an offline set of answerable questions based on the task, or by prompting an LLM pre-trained with open-domain knowledge, hence capable of generating any necessary query. In either case, $\mathcal{Q}$ is expected to satisfy sufficiency.

IP Algorithm. Given an input $x^{\text{obs}} \in \mathcal{X}$, the IP algorithm makes a prediction for Y by sequentially selecting queries $q_1, q_2, \ldots \in \mathcal{Q}$ in order of *information gain*, i.e., by maximizing the conditional information between a query answer and the prediction given the *history* of prior query-answer pairs:

$$\begin{aligned} q_1 &= \arg\max_{q \in \mathcal{Q}} I(q(X); Y); \\ q_{k+1} &= \arg\max_{q \in \mathcal{Q}} I(q(X); Y \mid \mathcal{H}_k(x^{\text{obs}})). \end{aligned} \tag{IP}$$

When $k = 0$, the history of prior query-answer pairs is empty, i.e., $q_{1:k} := \emptyset$, and IP selects the query whose answer is most informative for predicting Y. When $k \geq 1$, IP selects the most informative query conditioned on the *history* $\mathcal{H}_k := \mathcal{H}_k(x^{\text{obs}}) = \{x' \in \mathcal{X} : q_i(x') = q_i(x^{\text{obs}}) \text{ for all } i = 1, \ldots, k\}$ induced by $q_{1:k}(x^{\text{obs}})$[1]. IP terminates at iteration K if $\forall q \in \mathcal{Q}, I(q(X); Y \mid \mathcal{H}_K) = 0$, meaning that no additional query provides extra information about the target Y. Moreover, at each iteration k, IP requires the estimation of the posterior $\mathbb{P}(Y \mid \mathcal{H}_k)$. In practice, this allows the user to observe how each query and answer affect the prediction at test time.

To carry out IP, we equivalently write it as conditional entropy minimization. For each k,

$$\max_{q \in Q} I(q(X); Y \mid \mathcal{H}_k) = \max_{q \in Q} \{H(Y \mid \mathcal{H}_k) - H(Y \mid q(X), \mathcal{H}_k)\} = \min_{q \in Q} H(Y \mid q(X), \mathcal{H}_k). \tag{2}$$

At every iteration, the query that maximizes the conditional mutual information is also the query that minimizes the conditional entropy. To find the query in (2), one would have to collect samples, estimate $H(Y \mid q(X), \mathcal{H}_k)$ for each q, and select the query that minimizes the conditional entropy.

IP for LLMs. Applying IP for LLMs requires directly estimating entropy terms and making predictions with the probability distribution $\mathbb{P}(Y \mid \mathcal{H}_k)$ induced by an LLM. We can construct a probabilistic predictor[2] $f : \mathcal{X} \to \Delta(\mathcal{Y})$,

[1] We say $q_{1:k}(x^{\text{obs}})$ as the query-answer chain to denote the set of query answers, and $\mathcal{H}_k$ as the history to denote the event induced by $q_{1:k}(x^{\text{obs}})$.

[2] Here $\Delta(\mathcal{Y})$ is the set of probability distributions over $\mathcal{Y}$.

which takes a text prompt as input and outputs a set of probability scores for each class, via an LLM. While there are many ways of obtaining such scores, in this work we first extract the logits of the output tokens of interest (such as class names), and then apply a softmax transform to obtain a predicted distribution (See more in Appendix §G). In contrast to previous versions of IP [14, 15, 17], our f is not trained to model $\mathbb{P}(Y, Q(X))$, but is extracted from an LLM that models $\mathbb{P}(X)$ as a general language distribution, where the output probabilities are often miscalibrated. This motivates the need for an alternative approach to IP.

3 Conformal Information Pursuit: Information Gain with LLMs

Next, we propose an alternative approach which leverages recent developments in predictive inference and conformal prediction [5, 118]. We first use predictive inference for uncertainty estimation, and then apply split conformal prediction to obtain prediction sets at test time, formalizing our proposed algorithm, Conformal Information Pursuit (C-IP).

3.1 Uncertainty Estimation via Predictive Inference

The results of Correia et al. [25] allow us to upper bound entropy by the size of certain prediction sets for the ground-truth answers, hence measuring uncertainty in a distribution-free and robust manner. We now turn to introducing this approach.

Preliminary on Predictive Inference. Let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i), i \in [N]\}$, where $(X_i, Y_i) \sim P$ and $[N] := \{1, \ldots, N\}$, be a calibration dataset of feature-label[3] pairs drawn from some unknown data distribution P. Moreover, let $(X_{\text{test}}, Y_{\text{test}}) \sim P$ be an independent test feature-label pair. We again consider f defined as a function that assigns to each value x the predicted probabilities $f(x) = (f(x)_1, \ldots, f(x)_{|\mathcal{Y}|})$ for the classes $1, \ldots, |\mathcal{Y}|$. We consider the scenario where f has been pre-trained on data that is separate from and independent of the calibration data. For a threshold $\tau \in [0, 1]$, one can define prediction sets depending on x as the labels with predicted probabilities $f(x)_y$ larger than τ:

$$\mathcal{C}_\tau(x) = \{y \in \mathcal{Y} \mid f(x)_y > \tau\}. \tag{3}$$

In predictive inference, we seek to provide a guarantee that the prediction set $\mathcal{C}_\tau(X_{\text{test}})$ contains the ground-truth label Y_{test} with some desired coverage probability $1 - \alpha \in (0, 1)$. Specifically, in split conformal prediction [83], we set the threshold $\hat{\tau}$ such that, over the randomness in the calibration data $\mathcal{D}_{\text{cal}}$ and the test data $(X_{\text{test}}, Y_{\text{test}})$, we have the coverage guarantee

$$\mathbb{P}_{X_{\text{test}}, Y_{\text{test}}, \mathcal{D}_{\text{cal}}} \left(Y_{\text{test}} \in \mathcal{C}_{\hat{\tau}}(X_{\text{test}})\right) \geq 1 - \alpha. \tag{4}$$

In words, we want the true label Y_{test} for the test feature X_{test} to be an element of the prediction set $\mathcal{C}_\tau(X_{\text{test}})$ with probability at least $1 - \alpha$. There are a number of ways to set the threshold τ, including using conformal prediction [96, 117]; and $\mathcal{C}_{\hat{\tau}}(x)$ is also known as a *conformal prediction set*. The approach leverages the exchangeability of the values $f(X_i)_{Y_i}, i \in [N] \cup \{\text{test}\}$ and sets $\hat{\tau}$ as the $\lceil (1 - \alpha)(N + 1) \rceil$-th smallest of the values $f(X_i)_{Y_i}, i \in [N]$. This method satisfies (4). If the distribution of $f(X)_Y$ when $(X, Y) \sim P$ does not have any point masses, then the coverage probability is also upper bounded by $1 - \alpha_N$ for $\alpha_N = \alpha - 1/(N + 1)$.

Contrasting Notions of Uncertainty. There is a plethora of work that uses notions from predictive inference, such as the size of a prediction set, as a means to measure uncertainty [4, 26, 30, 54, 77, 94, 95]. We seek to mathematically characterize the relation between entropy and the size of prediction sets by leveraging a connection provided by Correia et al. [25].

From now on, we will write X, Y instead of $X_{\text{test}}, Y_{\text{test}}$ for brevity. Moreover, we will consider a prediction set $\mathcal{C}_{\mathcal{D}_{\text{cal}}}$ that depends on the calibration dataset $\mathcal{D}_{\text{cal}}$, but we will write $\mathcal{C}$ for simplicity.

[3] Text can also be treated as features.

Proposition 3.1 (Correia et al. [25])**.** For $\alpha \in (0, 0.5)$, consider any prediction set method $\mathcal{C} = \mathcal{C}_{\mathcal{D}_{\text{cal}}}$ satisfying

$$1 - \alpha \leq \mathbb{P}_{X,Y,\mathcal{D}_{\text{cal}}} \left(Y \in \mathcal{C}(X) \right) \leq 1 - \alpha_N. \tag{5}$$

Let $\lambda_\alpha := h_b(\alpha) + \alpha \log|\mathcal{Y}| - (1-\alpha_N)\log(1-\alpha)$, where h_b is the binary entropy function. For the true distribution P, we have

$$H(Y \mid X) \leq \lambda_\alpha + (1 - \alpha_N) \log \mathbb{E}_X[|\mathcal{C}(X)|]. \tag{6}$$

In words, the uncertainty of Y given data X is upper bounded by the expected size of the prediction set. The more uncertain one is about Y, the higher the entropy and the larger the prediction set size, and vice versa. Therefore, using a well-constructed prediction set $\mathcal{C}$ ensures the bound is not vacuous. Next, we will derive a conformalized formulation for IP.

3.2 Conformal Information Pursuit Algorithm

We now formalize our proposed method. Since we leverage split conformal prediction to obtain prediction sets at test time, we refer to our new approach to IP as **Conformal Information Pursuit (C-IP)**[4]. We will use the results above to derive a new method for choosing the next query given history. We emphasize that query answers $q(x)$ and the query-answer chains $q_{1:k}(x)$ are represented as text, hence also live in the domain of $\mathcal{X}$.

To begin, let us look at the base case. Due to (2), $\operatorname{argmax}_{q \in \mathcal{Q}} I(q(X); Y) = \operatorname{argmin}_{q \in \mathcal{Q}} H(Y \mid q(X))$. From Proposition 3.1, we construct a $\mathcal{C}_{\hat{\tau}(q)}$ for every $q \in \mathcal{Q}$ that satisfies the marginal coverage guarantee (5):

$$1 - \alpha \leq \mathbb{P}_{X,Y,\mathcal{D}_{\text{cal}}} \left(Y \in \mathcal{C}_{\hat{\tau}(q)}(q(X)) \right) \leq 1 - \alpha_N. \tag{7}$$

Then,

$$\min_{q \in \mathcal{Q}} H(Y \mid q(X)) \leq \min_{q \in \mathcal{Q}} \left\{ \lambda_\alpha + (1 - \alpha_N) \log \mathbb{E}_X[|\mathcal{C}_{\hat{\tau}(q)}(q(X))|] \right\}. \tag{8}$$

Our C-IP method aims to choose the first question q_1 by minimizing the above upper bound on the entropy. Similarly, for the subsequent iterations $k \geq 1$, due to (2), $\operatorname{argmax}_{q \in \mathcal{Q}} I(q(X); Y \mid \mathcal{H}_k) = \operatorname{argmin}_{q \in \mathcal{Q}} H(Y \mid q(X), \mathcal{H}_k)$. Again, from Proposition 3.1, we construct a $\mathcal{C}_{\hat{\tau}(q,\mathcal{H}_k)}$ for every $q \in \mathcal{Q}$ that satisfies the coverage guarantee (5) conditioned on *the given* $\mathcal{H}_k$ *induced by* $q_{1:k}(x^{\text{obs}})$, i.e.,

$$1 - \alpha \leq \mathbb{P}_{X,Y,\mathcal{D}_{\text{cal}}} \left(Y \in \mathcal{C}_{\hat{\tau}(q,\mathcal{H}_k)}(q(X)) \mid \mathcal{H}_k \right) \leq 1 - \alpha_N. \tag{9}$$

Then,

$$\min_{q \in \mathcal{Q}} H(Y \mid q(X), \mathcal{H}_k) \leq \min_{q \in \mathcal{Q}} \left\{ \lambda_\alpha + (1 - \alpha_N) \log \mathbb{E}_X \left[|\mathcal{C}_{\hat{\tau}(q,\mathcal{H}_k)}(q(X))| \mid \mathcal{H}_k \right] \right\}. \tag{10}$$

Our C-IP method aims to choose q_k by minimizing the above upper bound. We will refer to the algorithm that uses empirical estimates of the entropy on the left-hand side of (10) as IP and to the algorithm that uses empirical estimates of the upper bound on the right-hand side of (10) as C-IP.

From our derivation above, directly applying Proposition 3.1 requires coverage guarantees in (7) and (9) that condition on the *one instantiation of the query chain* $q_{1:k}$ obtained from running C-IP. Unfortunately, this is intractable in general [55, 59], as the number of possible $q_{1:k}(x^{\text{obs}})$ is exponentially large, of size $\sim |\mathcal{Q}|^k$. Even when the maximum number of iterations L is small, ensuring this would also in general require a large enough subset of the calibration dataset X_i, $i \in [N]$, with the same history $\mathcal{H}_k$.

[4]We remark that "conformalized" often refers to a wrapper of a method that enjoys a conformal guarantee; this is not the sense in which we use it.

Obtaining Prediction Sets $\mathcal{C}$. To ensure tractability, we propose the following guarantee:

$$1-\alpha \leq \mathbb{P}_{X,Y,Q_{1:k},\mathcal{D}_{\text{cal}}}\left(Y \in \mathcal{C}_{\hat{\tau}(k)}(Q_{1:k}(X))\right) \leq 1-\alpha_N, \tag{11}$$

which means to offer a guarantee over distribution of random query-answer chains $Q_{1:k}(X)$ with length k rather than a single choice of $q_{1:k}(X)$. Our marginalized guarantee only requires us to construct L prediction set functions. Next, we discuss how to obtain samples from $Q_{1:k}$ to construct $\mathcal{C}_{\hat{\tau}(k)}$ and satisfy (11).

The question of how to efficiently sample histories arises in many occasions when implementing IP (see Related Work in Appendix §B for discussion). In this work, we consider two possible parameterizations:

- **Sampling Histories from a Uniform Parameterization:** When the query set $\mathcal{Q}$ is a closed set with predetermined queries, an effective strategy is to parameterize queries at each iteration with a uniform distribution over the query set [15]. As for V-IP, we propose to marginalize our coverage over *uniformly sampled histories of length k*. For every iteration $k = 1, \ldots, L$, where L is the maximum number of iterations allowed, let $Q_{1:k} = (Q_1, \cdots, Q_k)$ be a distribution of uniformly constructed histories. Then,

$$Q_{1:k} \sim \text{Unif}(\mathcal{Q})^k \quad \text{for} \quad k = 1, \ldots, L. \tag{12}$$

- **Sampling of Histories from LLM Simulations:** If $\mathcal{Q}$ is an open query set with the space of questions in natural language, another option is to consider the distribution of histories obtained from LLMs. For a given observation x^{obs} and query chain $q_{1:k}$, q_{k+1} can be obtained by directly prompting for the next query, i.e., $q_{k+1} \sim \text{LLM}(\mathcal{Q}, q_{1:k}(x^{\text{obs}}))$. We refer this strategy as Direct Prompting (DP) (see Appendix §I for the algorithm), where the distribution of queries is:

$$Q_{1:k} \sim \text{DP}(X, Y) \quad \text{for} \quad k = 1, \ldots, L. \tag{13}$$

Hence, we obtain a list of prediction set functions $\mathcal{C}_{\hat{\tau}(1)} \ldots \mathcal{C}_{\hat{\tau}(L)}$ for every length. The precise procedure is described in Appendix §I, Algorithm 1.

C-IP Algorithm. To summarize, the C-IP algorithm is as follows. First, we obtain a list of prediction set functions $\mathcal{C}_{\hat{\tau}_1} \ldots \mathcal{C}_{\hat{\tau}_L}$ based on the desired coverage $1-\alpha$. Then, for any given observation x^{obs},

$$\begin{aligned} q_1 &= \operatorname*{argmin}_{q \in \mathcal{Q}} \log \mathbb{E}_X[|\mathcal{C}_{\hat{\tau}(1)}(q(X))|]; \\ q_{k+1} &= \operatorname*{argmin}_{q \in \mathcal{Q}} \log \mathbb{E}_X[|\mathcal{C}_{\hat{\tau}(k+1)}(q(X))| \mid \mathcal{H}_k]. \end{aligned} \tag{C-IP}$$

Similar to IP, the quantities in expectation are estimated by their empirical expectations. See Appendix §I for pseudocode for the IP, C-IP and DP algorithms.

Remark 1: Stopping Criteria. As mentioned in §2, IP stops when $\forall q \in \mathcal{Q}, I(Y; q(X) \mid \mathcal{H}_k) = 0$. From (2), this corresponds to $\forall q \in \mathcal{Q}, H(Y \mid q(X), \mathcal{H}_k) = 0$. This also nicely extends to C-IP. When the model is highly confident, the prediction set should be on average a *singleton*, which implies $\forall q \in \mathcal{Q}, \log \mathbb{E}_X[|\mathcal{C}_{\hat{\tau}(q)}(q(X))| \mid \mathcal{H}_k] = 0$. In other words, the stopping criteria are equivalent in IP and C-IP. In practice, this stringent stopping criterion may not be reached. Hence, we stop querying once the standard deviation of the estimated entropy for every query is smaller than a chosen threshold $\varepsilon > 0$. See Appendix §G.10 for more details.

Remark 2: Risk Control. Our marginalization approach is a means to estimate uncertainty. Ideally, one would characterize the full distribution over IP histories. We present our work as a first step in this direction, and consider the aspect of error control as future work.

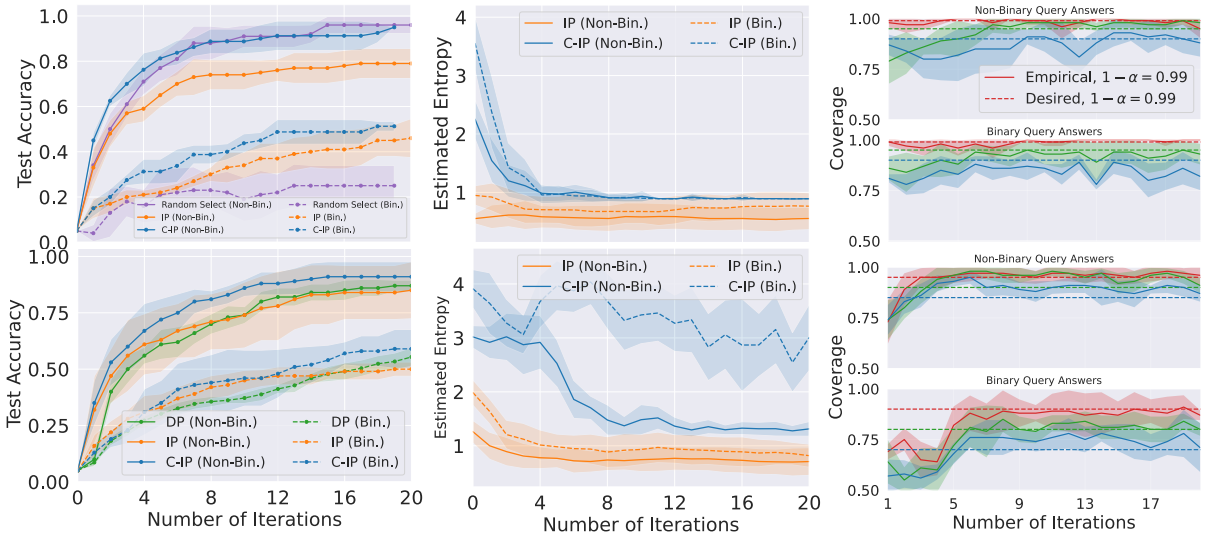


Figure 3: Evaluations of C-IP with a closed query set $\mathcal{Q}_{\text{closed}}$ *(top row)* and open query set $\mathcal{Q}_{\text{open}}$ *(bottom row)*. Each curve shows the average performance, with the shaded area presenting the standard deviation (std.). *Left column*: Performance on the 20 Questions task of C-IP and baselines with binary (dashed) and non-binary (solid) query answers using Llama-3.1-8b. *Middle column*: Uncertainty estimated by IP and C-IP, with each curve evaluating the uncertainty of the selected query at each iteration. *Right column*: Desired (dashed) and empirical (solid) coverage of C-IP. Hyperparameter choice is marked with different colors, with desired and empirical coverage as dashed and solid.

4 Exploration with 20 Questions

To create a simple synthetic setting for validating our framework, we explore the properties of IP and C-IP by playing 20Q with two LLMs, one *Querier* LLM and one *Answerer* LLM. The task is simple: The Answerer LLM thinks of an animal and the Querier LLM is tasked to guess it within 20 queries.

4.1 Experiment Setup

Dataset. We obtain 20 common animal names from the Animals with Attributes 2 (AwA2) [130] dataset. This is done for two main reasons: 1) to ensure that the LLMs we consider have sufficient knowledge about each class, allowing us to accurately evaluate our algorithm in a setting where the LLM performance is high. 2) The AwA2 dataset provides an annotated set of binary attributes about each animal. This allows us to obtain a binary query set with expert-labeled answers. We consider a closed query set $\mathcal{Q}_{\text{closed}}$, in which 85 annotated attributes from AwA2 are converted into textual questions, and an open query set $\mathcal{Q}_{\text{open}}$, where queries are free-form questions obtained by prompting LLMs. We consider binary query-answers with $\mathcal{A}_{\text{binary}} = \{\text{“Yes”}, \text{“No”}\}$ based on the AwA2 expert labels, and open query-answers with $\mathcal{A}_{\text{free-text}}$ being the set of English sentences (Examples in Appendix §J.1). In the 20Q setting, the data space is equal to the label space ($\mathcal{X} = \mathcal{Y}$), thus reducing the problem's complexity to help establish the effectiveness of using info-theoretic quantities in guiding LLMs.

Models and Implementation. We consider three open-source instruction-tuned models: Llama-3.1-8B [39], Qwen2.5-7B [7] and Phi-3-small [2]. We present results for Llama-3.1-8B in the main text, and other results in Appendix §E.1. To achieve the target coverage for both closed and open query sets, we calibrate using 100 random histories and labels. Every entropy term is estimated with 4 randomly drawn samples. We obtain our prediction sets $\mathcal{C}$'s by sampling via (12) for the setting with closed query set $\mathcal{C}_{\text{closed}}$ and via (12) for the setting with open

query set $\mathcal{C}_{\text{open}}$. For LLM hyperparameters during generation, prompts, and further implementation details, refer to Appendix §G.

Baselines. To compare the performance of C-IP in the closed query set setting, we consider two baselines: *Random*, which uniformly selects a query from $\mathcal{Q}_{\text{closed}}$ at each iteration, and *IP*, which evaluates (2) at each iteration. For the open query set setting, the baselines are *Direct Prompting (DP)* and *Uncertainty of Thought (UoT)* [45]. Direct Prompting, as described in §3.2 and Appendix §I, directly prompts the LLM for a single query at each iteration.

4.2 Results

Predictive Performance. We evaluate C-IP against baselines on predictive performance in Figure 3 (top left). For each iteration up to a maximum of 20, we evaluate the test accuracy given $q_{1:k}(x^{\text{obs}})$. We use a target coverage of $1-\alpha = 0.9$. For the results with binary answers, we observe a sharp difference between random selection and IP-based methods, with C-IP performing better at each iteration. For non-binary answers, C-IP outperforms IP at every iteration, showing that early query selection weighs greatly for performance. While outperforming random selection in iterations 1-6, it performs comparably in iterations 7-20. We argue that this occurs due to the simplicity of the task, where the combination of open query-answers and random selection on task-relevant queries pre-determined by AwA2 provides sufficient information to achieve high accuracy. Overall, this result shows that C-IP is a better method for selecting queries from a pre-defined, closed query set.

We perform a similar performance evaluation for C-IP with $\mathcal{Q}_{\text{open}}$ in Figure 3 (bottom left). Here, we observe that C-IP with $\alpha = 0.15$ outperforms both baselines IP and DP. This shows that direct prompting may not always suggest the most informative query, and C-IP as an uncertainty estimate can potentially generalize to previously unseen queries.

Uncertainty Estimation. To ensure that C-IP selects informative queries, we evaluate the estimation of the entropy $\hat{H}(Y \mid q_{k+1}(X), \mathcal{H}_k)$ at each iteration. C-IP bounds this via the size of prediction sets in (10) and IP directly estimates it in (2). In Figure 3 (top middle), we observe that, for both binary $\mathcal{A}_{\text{binary}}$ and free-text query answers $\mathcal{A}_{\text{free-text}}$, the C-IP bounds (blue curves) decrease as the algorithm progresses, whereas entropy estimates (orange curves) revolve around the same level. This shows that C-IP uncertainty measures are more informative than entropy estimates. The results for C-IP with $\mathcal{Q}_{\text{open}}$ (Figure 3, bottom middle) also show a decrease in the estimated entropy as number of iterations increase. In the case of binary query answers $\mathcal{A}_{\text{binary}}$, while we observe a less stable decay, C-IP still performs comparably with our baselines, showing that the estimation of uncertainty is robust despite noisy probabilities.

Empirical Coverage. To evaluate the coverage of our prediction sets, we choose three target values of α and evaluate the empirical probability that the true label is contained in the prediction set at each iteration over our test set (See Appendix §G.11 for empirical formula). Our results are shown in Figure 3 (right column). For $\mathcal{C}_{\text{closed}}$ with both binary $\mathcal{A}_{\text{binary}}$ and free-text query answers $\mathcal{A}_{\text{free-text}}$, the realized coverage is close to the target for $1-\alpha = 0.99$, whereas our method undercovers for $1-\alpha = 0.95$ and 0.99. Overall, the choice of α largely influence the predictive performance of C-IP, which explains the superior performance for C-IP over IP and random query selection when $1-\alpha = 0.99$. On the other hand, for $\mathcal{C}_{\text{open}}$ with both binary $\mathcal{A}_{\text{binary}}$ and free-text query answers $\mathcal{A}_{\text{free-text}}$, we observe a slight over-coverage for smaller choices of α and under-coverage for early iterations. This is likely due to the wide range of possible queries in few initial steps, leading to under-coverage and misspecification. See Appendix §F.1 for ablation studies regarding choices of α.

Comparison with UoT. We compare our method against a state-of-the-art LLM reasoning method, Uncertainty-of-Thought (UoT) [45]. UoT handles multi-turn conversations by propagating rewards based on information gain from future steps. At each iteration, it builds a shallow tree computing the information gain of a few potential trajectories, akin to seeing a few steps ahead. Importantly, UoT applies only to the setting with open query set and binary query answers. We will compare UoT with IP and C-IP, as both are entropy-based uncertainty quantification methods for interactive question answering. Further details are provided in Appendix G.

Our results are shown in Table 1. For binary query answers, UoT achieves higher success rate than IP or C-IP. This is reasonable as accumulating information gain a few iterations ahead is likely more advantageous than using immediate rewards. However, for free-text query answers, C-IP achieves a near 45% accuracy improvement, with a smaller average number of queries. Thus, using an immediate reward with free-text query answers works better than accumulated entropy-based rewards with binary answers.

Table 1: Comparison of C-IP with UoT reasoning, averaged across 5 runs with std..

Method	Query Answers $\mathcal{A}$	Avg. Len	Accuracy
UoT	binary	13.57 ± 1.69	0.57 ± 0.13
IP	binary	11.40 ± 0.78	0.31 ± 0.04
IP	free-text	10.04 ± 0.45	0.65 ± 0.18
C-IP ($\alpha = 0.1$)	binary	11.87 ± 0.83	0.37 ± 0.11
C-IP ($\alpha = 0.1$)	free-text	10.58 ± 0.58	0.83 ± 0.05

5 Application to Interactive Medical Question Answering

In this section, we apply C-IP to an interactive medical question-answering task on the MediQ dataset. We provide the basic setup here and other details (such as prompts) can be found in Appendix §G.

5.1 Setup

Dataset. MediQ [65] is an interactive medical question-answering dataset extended from MedQA [49], a large-scale medical question-answering dataset designed to evaluate the professional knowledge and clinical decision-making abilities of LLMs. All data is drawn from medical licensing examinations in the United States. Each datapoint in MediQ contains a medical question, four multiple-choice answers, a label for the specialty/area, and a list of facts from a context paragraph about the patient[5]. For IP in this context, $\mathcal{X}$ is the set of patients' information, $\mathcal{Q}$ is the set of (closed) queries one can ask about the patient, and $\mathcal{Y}$ is the set of multiple-choice answers.

The task simulates an interactive setting between an Expert LLM (the predictor and the querier in our nomenclature) and a Patient LLM (the query-answerer), and the goal is that the Expert LLM diagnoses the patient by answering the provided medical questions. For our application, we consider medical questions from the following three categories: "Internal Medicine (IM)" (290 samples), "Pediatrics (P)" (217 samples) and "Neurology (N)" (108 samples). We treat each category as a single task, and the

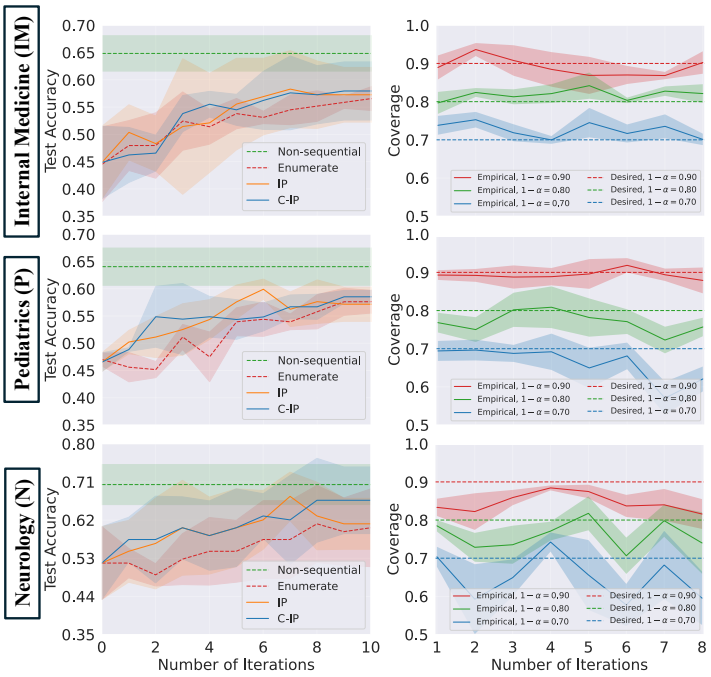


Figure 4: *Top:* Predictive Performance in Medical Interactive Question Answering on the MediQ dataset. The tasks are divided based on specialty. Desired coverage is $1 - \alpha = 0.8$ for IM and P and $1 - \alpha = 0.7$ for N. *Bottom:* Comparison of empirical and desired coverage for C-IP. Desired coverage (dashed curve) is shown for $1 - \alpha \in \{0.7, 0.8, 0.9\}$. Since the number of queries may depend on the datapoints, Empirical Coverage (solid curve) at iteration k is evaluated for all test datapoints that have stopped at or before iteration k. Each curve is averaged over three splits, with the shaded area denoting their standard deviation.

[5]The number of facts may vary depending on the number of sentences in each sample.

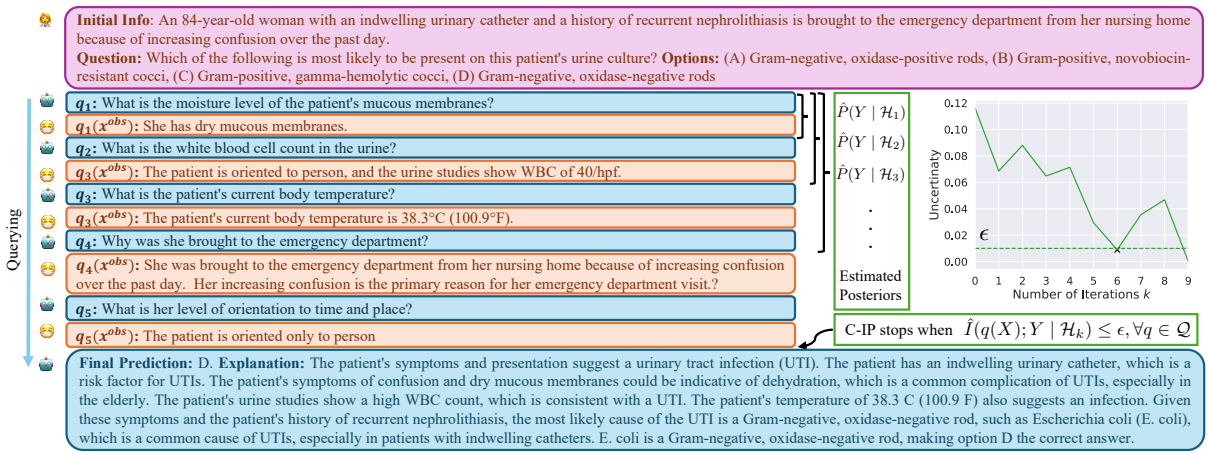


Figure 5: Example query-answer chain produced by C-IP for a patient in Internal Medicine. C-IP sequentially selects queries that approximately reduce uncertainty (i.e. average prediction set sizes from (10)) at each iteration. The posterior $\hat{P}(Y \mid \mathcal{H}_k)$ is also estimated and a prediction is made. Once uncertainty drops below a chosen threshold ϵ, the C-IP algorithm stops and makes a final prediction. Based on the obtained query-answer chain and the prediction, an explanation is also generated.

coverage is calculated using calibration data
obtained from the same category.

Models and Evaluation. To ensure a fair evaluation, we divide each category into three equal sets: $\mathcal{D}_{\text{est}}$ for estimating entropy, $\mathcal{D}_{\text{cal}}$ for calibration, and $\mathcal{D}_{\text{test}}$ for test-set evaluation. We perform three-fold cross validation and evaluate the average performance and its standard deviation. In our main text, we evaluate Llama-3.1-8B as both the Expert LLM and the Patient LLM. Similar to 20Q, we first obtain the output token's logit score based on the four options (A, B, C, D), then apply softmax to obtain an estimate of the posterior distribution $\mathbb{P}(Y \mid \mathcal{H}_k)$. For implementation details, please refer to the Appendix §G.

Baselines. We compare with two baselines. In *Non-sequential* prediction, the Expert LLM makes a single prediction based on all necessary information. This serves as an upper bound on the predictive performance, and has been shown to outperform interactive methods [65]. In *Enumerate*, we provide the facts in the order of decomposition from the original context paragraph. Since our LLMs are inherently sequential, this acts as if the LLM is reading the context paragraph in order.

Query Set Construction. We find that constructing a single query set for the entire test set often leads to queries with unknown query answers. Hence, we construct a query set for each test datapoint based on the existing atomic facts in the input, by converting each fact into a question. Therefore, the goal of the IP algorithm is to select the queries with answers obtainable from the context. We find this setting more constructive compared to the setting with a single query set for all evaluations.

5.2 Results

Predictive Performance and Coverage. The performance of C-IP and the baselines for the three specialties is shown in Figure 4. C-IP consistently obtains higher test accuracy at each iteration than IP for IM and P while achieving comparable performance for N. This suggests that C-IP selects more informative queries during the interactive process. To explain the effectiveness of C-IP and why it did not strongly outperform baselines for N, we analyze their empirical coverage at each iteration. We observe that C-IP has valid coverage IM and P but undercovers for N. We argue that this is likely due to the small calibration set size for N (36 datapoints).

Explanations. We show a run-down of the C-IP algorithm in Figure 5. Before any query selection, the Expert LLM

is provided with initial information, the medical question, and possible options for the answer. Then, C-IP selects the first query q_1 = "What is the moisture level of the patient's mucous membranes?", to which the Patient LLM responds $q_1(x^{\text{obs}})$ = "She has dry mucous membranes". This process continues until the uncertainty measured based on the posterior of the answer options from the Expert LLM drops below 0.01. Once C-IP stops, it makes the final prediction "D", and summarizes the query-answer chain into a paragraph explaining the decision. More examples (including other specialties) are shown in Appendix §J.2.

6 Conclusion and Future Work

In this work, we proposed C-IP, a novel sequential information pursuit algorithm that uses prediction set sizes as an alternative to conditional entropy to estimate uncertainty. We leveraged a mathematical connection between prediction sets and entropy, and proposed two sampling methods for obtaining histories. Through our empirical studies on 20Q and MediQ datasets, we demonstrated that our proposed approach performs well in practice, and that the size of the prediction sets is an effective measure of uncertainty. As we suggested in §3 Remark 2, an interesting avenue for future work is to provide guarantees for the correctness of the prediction given query-answer chains, $\mathbb{P}(Y \mid \mathcal{H}_k)$, obtained by C-IP and for the query-answer chains $q_{1:k}(x^{\text{obs}})$ themselves. Doing so would require a careful characterization of the distribution $Q_{1:k} \sim \text{IP}(\{q_i, q_i(X)\}_{1:k})$.

Acknowledgments

K. H. R. C. would personally like to thank Tianjiao Ding, Uday Kiran Reddy Tadipatri, Liangzu Peng, Darshan Thaker, Kyle Poe, Buyun Liang, Ryan Pilgrim, and Aditya Chattopadhyay for fruitful discussions and comments on the draft. K. H. R. C., Y. G., and R. V. acknowledge the support from the Penn Engineering Dean's Fellowship, the startup funds, and the NIH grant R01NS135613. E. D. was partially supported by the NSF, ARO, AFOSR, ONR, and the Sloan Foundation. H. H. was supported by the NSF CAREER award CIF-1943064.

References

[1] Z. Abbasiantaeb, Y. Yuan, E. Kanoulas, and M. Aliannejadi. Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2023.

[2] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[5] A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[6] Y. Ba, M. Mancenido, and R. Pan. Fill in the gaps: Model calibration and generalization with synthetic data. In *Conference on Empirical Methods in Natural Language Processing*, pages 17211–17225, 2024.

[7] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[8] S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

[9] A. M. Bean, R. Payne, G. Parsons, H. R. Kirk, J. Ciro, R. Mosquera, S. H. Monsalve, A. S. Ekanayaka, L. Tarassenko, L. Rocher, et al. Clinical knowledge in llms does not translate to human interactions. *arXiv preprint arXiv:2504.18919*, 2025.

[10] A. Ben Abacha and D. Demner-Fushman. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23, 2019.

[11] L. Bertolazzi, D. Mazzaccara, F. Merlo, and R. Bernardi. ChatGPT's information seeking strategy: Insights from the 20-questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162, 2023.

[12] G. M. Biancofiore, Y. Deldjoo, T. D. Noia, E. Di Sciascio, and F. Narducci. Interactive question answering systems: Literature review. *ACM Computing Surveys*, 56(9):1–38, 2024.

[13] A. Bodrova, P. Xu, J. Varley, A. Zeng, A. Majumdar, D. Sadigh, S. Singh, F. Xia, N. Brown, L. Takayama, A. Z. Ren, A. Dixit, Z. Xu, and S. Tu. Robots that ask for help: Uncertainty alignment for large language model planners. *ArXiv*, abs/2307.01928, 2023. URL https://api.semanticscholar.org/CorpusId:259342058.

[14] A. Chattopadhyay, S. Slocum, B. D. Haeffele, R. Vidal, and D. Geman. Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[15] A. Chattopadhyay, K. H. R. Chan, B. D. Haeffele, D. Geman, and R. Vidal. Variational information pursuit for interpretable predictions. *arXiv preprint arXiv:2302.02876*, 2023.

[16] A. Chattopadhyay, R. Pilgrim, and R. Vidal. Information maximization perspective of orthogonal matching pursuit with applications to explainable ai. *Advances in Neural Information Processing Systems*, 36:2956–2990, 2023.

[17] A. Chattopadhyay, K. H. R. Chan, and R. Vidal. Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] J. Chen and J. Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusId:263611057.

[19] Y. Chen, S. H. Hassani, A. Karbasi, and A. Krause. Sequential information maximization: When is greedy near-optimal? In *Conference on Learning Theory*, pages 338–363. PMLR, 2015.

[20] Z. Chen, K. Zhou, B. Zhang, Z. Gong, W. X. Zhao, and J. rong Wen. Chatcot: Tool-augmented chain-of-thought reasoning on chat-based large language models. *ArXiv*, abs/2305.14323, 2023. URL https://arxiv.org/pdf/2305.14323.pdf.

[21] J. Cherian, I. Gibbs, and E. Candes. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, volume 37, pages 114812–114842, 2024.

[22] J. J. Cherian, I. Gibbs, and E. Candès. Large language model validity via enhanced conformal prediction methods. *ArXiv*, abs/2406.09714, 2024. URL https://api.semanticscholar.org/CorpusId:270521658.

[23] V. Chernozhukov, K. Wuthrich, and Y. Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference On Learning Theory*, 2018.

[24] A. Comas, A. Chattopadhyay, F. Formosa, C. Liu, O. Camps, and R. Vidal. Incode: Interpretable compressed descriptions for image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[25] A. H. Correia, F. V. Massoli, C. Louizos, and A. Behboodi. An information theoretic perspective on conformal prediction. *arXiv preprint arXiv:2405.02140*, 2024.

[26] C. Cortes, A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Cardinality-aware set prediction and top-k classification. *arXiv preprint arXiv:2407.07140*, 2024.

[27] I. C. Covert, W. Qiu, M. Lu, N. Y. Kim, N. J. White, and S.-I. Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pages 6424–6447. PMLR, 2023.

[28] S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems*, 17, 2004.

[29] G. Detommaso, M. Bertran, R. Fogliato, and A. Roth. Multicalibration for confidence scoring in LLMs. *arXiv preprint arXiv:2404.04689*, 2024.

[30] G. S. Dhillon, G. Deligiannidis, and T. Rainforth. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pages 1549–1557. PMLR, 2024.

[31] R. Dunn, L. Wasserman, and A. Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, pages 1–12, 2022.

[32] B.-S. Einbinder, Y. Romano, M. Sesia, and Y. Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 2022.

[33] S. Feldman, L. Ringel, S. Bates, and Y. Romano. Achieving risk control in online learning settings. *Trans. Mach. Learn. Res.*, 2023, 2022. URL https://arxiv.org/pdf/2205.09095.pdf.

[34] O. F'eron, J. Josse, Y. Goude, M. Zaffran, and A. Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusId:246863519.

[35] D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):1–14, 1996.

[36] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych. A survey of confidence estimation and calibration in large language models. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6577–6595, 2024.

[37] I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift. *ArXiv*, abs/2106.00170, 2021. URL https://api.semanticscholar.org/CorpusId:235266057.

[38] I. Gibbs and E. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *J. Mach. Learn. Res.*, 25:162:1–162:36, 2022. URL https://api.semanticscholar.org/CorpusId:251622480.

[39] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[40] L. Guan. A conformal test of linear models via permutation-augmented regressions. *arXiv preprint arXiv:2309.05482*, 2023.

[41] L. Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

[42] L. Guan and R. Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):524–546, 2022.

[43] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[44] B. Hou, Y. Liu, K. Qian, J. Andreas, S. Chang, and Y. Zhang. Decomposing uncertainty for large language models through input clarification ensembling. *ArXiv*, abs/2311.08718, 2023. URL https://api.semanticscholar.org/CorpusId:265213190.

[45] Z. Hu, C. Liu, X. Feng, Y. Zhao, S.-K. Ng, A. T. Luu, J. He, P. W. Koh, and B. Hooi. Uncertainty of Thoughts: Uncertainty-aware planning enhances information seeking in large language models. *arXiv preprint arXiv:2402.03271*, 2024.

[46] Z. Huang, S. Rossi, R. Yuan, and T. Hannagan. From predictions to confidence intervals: an empirical study of conformal prediction methods for in-context learning. In *unknown*, 2025. URL https://api.semanticscholar.org/CorpusId:277994329.

[47] E. Jahangiri, E. Yörük, R. Vidal, L. Younes, and D. Geman. Information Pursuit: A Bayesian framework for sequential scene parsing. *arXiv preprint arXiv:1701.02343*, 2017.

[48] P. Jiang, J. Rayan, S. P. Dow, and H. Xia. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

[49] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have. *A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv [cs. CL]*, 2020.

[50] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[51] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[52] R. Kaur, S. Jha, A. Roy, S. Park, E. Dobriban, O. Sokolsky, and I. Lee. iDECODe: In-distribution equivariance for conformal out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[53] R. Kaur, C. Samplawski, A. D. Cobb, A. Roy, B. Matejek, M. Acharya, D. Elenius, A. M. Berenbeim, J. Pavlik, N. D. Bastian, and S. Jha. Addressing uncertainty in llms to enhance reliability in generative ai. *ArXiv*, abs/2411.02381, 2024. URL https://api.semanticscholar.org/CorpusId:273822044.

[54] S. Kiyani, G. J. Pappas, and H. Hassani. Length optimization in conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 99519–99563, 2024.

[55] C. Koch, C. Strassle, and L.-Y. Tan. Superconstant inapproximability of decision tree learning. *Proceedings of Machine Learning Research vol*, 196:1–32, 2024.

[56] B. Kumar, C. Lu, G. Gupta, A. Palepu, D. Bellamy, R. Raskar, and A. Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

[57] B. Kumar, C.-C. Lu, G. Gupta, A. Palepu, D. R. Bellamy, R. Raskar, and A. Beam. Conformal prediction with large language models for multi-choice question answering. *ArXiv*, abs/2305.18404, 2023. URL https://api.semanticscholar.org/CorpusId:258967849.

[58] P. Laban, H. Hayashi, Y. Zhou, and J. Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

[59] H. Laurent and R. L. Rivest. Constructing optimal binary decision trees is np-complete. *Information processing letters*, 5(1):15–17, 1976.

[60] J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

[61] J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[62] J. Lei, A. Rinaldo, and L. Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43, 2015.

[63] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[64] S. Li, X. Ji, E. Dobriban, O. Sokolsky, and I. Lee. PAC-Wrap: Semi-supervised pac anomaly detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[65] S. Li, V. Balachandran, S. Feng, J. Ilgen, E. Pierson, P. W. W. Koh, and Y. Tsvetkov. MediQ: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

[66] Y. Li, X. Shen, X. Yao, X. Ding, Y. Miao, R. Krishnan, and R. Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025.

[67] Z. Liang, Y. Zhou, and M. Sesia. Conformal inference is (almost) free for neural networks trained with early stopping. In *International Conference on Machine Learning*, 2023.

[68] Z. Liang, M. Sesia, and W. Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad138, 01 2024.

[69] Z. Lin, S. Trivedi, and J. Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024, 2023. URL https://api.semanticscholar.org/CorpusId:258967487.

[70] C. Ling, X. Zhao, X. Zhang, W. Cheng, Y. Liu, Y. Sun, M. Oishi, T. Osaki, K. Matsuda, J. Ji, et al. Uncertainty quantification for in-context learning of large language models. *arXiv preprint arXiv:2402.10189*, 2024.

[71] O. Liu, D. Fu, D. Yogatama, and W. Neiswanger. DeLLMa: Decision making under uncertainty with large language models. *arXiv preprint arXiv:2402.02392*, 2024.

[72] S. Luttrell. The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems*, 1(3):199, 1985.

[73] Q. Lyu, K. Shridhar, C. Malaviya, L. Zhang, Y. Elazar, N. Tandon, M. Apidianaki, M. Sachan, and C. Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268, 2025.

[74] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[75] C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, pages 36029–36047. PMLR, 2024.

[76] C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. *ArXiv*, abs/2402.10978, 2024. URL https://api.semanticscholar.org/CorpusId:267750963.

[77] U. K. Müller and M. W. Watson. Measuring uncertainty about long-run predictions. *Review of Economic Studies*, 83(4):1711–1740, 2016.

[78] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[79] D. Noever and F. McKee. Chatbots as problem solvers: Playing twenty questions with role reversals. *arXiv preprint arXiv:2301.01743*, 2023.

[80] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

[81] A. Panchenko, E. Fadeeva, M. Panov, D. Vasilev, T. Baldwin, A. Tsvigun, K. Fedyanin, A. Shelmanov, E. Goncharova, S. Petrakov, A. Vazhentsev, and R. Vashurin. Lm-polygraph: Uncertainty estimation for language models. *ArXiv*, abs/2311.07383, 2023. URL https://api.semanticscholar.org/CorpusId:265149591.

[82] J.-C. Pang, H.-B. Fan, P. Wang, J.-H. Xiao, N. Tang, S.-H. Yang, C. Jia, S.-J. Huang, and Y. Yu. Empowering language models with active inquiry for deeper understanding. *arXiv preprint arXiv:2402.03719*, 2024.

[83] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

[84] S. Park, O. Bastani, N. Matni, and I. Lee. PAC confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020.

[85] S. Park, E. Dobriban, I. Lee, and O. Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022.

[86] S. Park, E. Dobriban, I. Lee, and O. Bastani. PAC prediction sets for meta-learning. In *Advances in Neural Information Processing Systems*, 2022.

[87] H. Qiu, E. Dobriban, and E. Tchetgen Tchetgen. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad069, 07 2023.

[88] V. Quach, A. Fisch, T. Schuster, A. Yala, J. Sohn, T. Jaakkola, and R. Barzilay. Conformal language modeling. *ArXiv*, abs/2306.10193, 2023. URL https://api.semanticscholar.org/CorpusId:259203582.

[89] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

[90] H. Rajabzadeh, S. Wang, H. J. Kwon, and B. Liu. Multimodal multi-hop question answering through a conversation between tools and efficiently finetuned large language models. *arXiv preprint arXiv:2309.08922*, 2023.

[91] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. *ArXiv*, abs/2307.01928, 2023. URL https://arxiv.org/pdf/2307.01928.pdf.

[92] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024.

[93] J. Robinson and D. Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023.

[94] Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 2020.

[95] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[96] C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *Sixteenth International Joint Conference on Artificial Intelligence*, pages 722–726, 1999.

[97] H. Scheffe and J. W. Tukey. Non-parametric estimation. I. Validation of order statistics. *The Annals of Mathematical Statistics*, 16(2):187–192, 1945.

[98] T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Q. Tran, Y. Tay, and D. Metzler. Confident adaptive language modeling. *ArXiv*, abs/2207.07061, 2022. URL https://arxiv.org/pdf/2207.07061.pdf.

[99] M. Sesia, S. Favaro, and E. Dobriban. Conformal frequency estimation using discrete sketched data with coverage for distinct queries. *Journal of Machine Learning Research*, 24(348):1–80, 2023.

[100] H. Shahrokhi, D. Roy, Y. Yan, V. Arnaoudova, and J. R. Doppa. Conformal prediction sets for deep generative models via reduction to conformal regression. *ArXiv*, abs/2503.10512, 2025. URL https://api.semanticscholar.org/CorpusId:276961510.

[101] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[102] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[103] M. Shen, S. Das, K. Greenewald, P. Sattigeri, G. Wornell, and S. Ghosh. Thermometer: Towards universal calibration for large language models. *arXiv preprint arXiv:2403.08819*, 2024.

[104] W. Si, S. Park, I. Lee, E. Dobriban, and O. Bastani. PAC prediction sets under label shift. In *The Twelfth International Conference on Learning Representations*, 2024.

[105] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

[106] J. Song, Z. Wang, H. Chen, Y. Huang, and L. Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *ArXiv*, abs/2307.10236, 2023. URL https://api.semanticscholar.org/CorpusId:259991714.

[107] C. Spiess, D. Gros, K. S. Pai, M. Pradel, M. R. I. Rabin, A. Alipour, S. Jha, P. Devanbu, and T. Ahmed. Calibration and correctness of language models for code. In *IEEE/ACM 47th International Conference on Software Engineering*, pages 495–507. IEEE Computer Society, 2025.

[108] J. Su, J. Luo, H. Wang, and L. Cheng. Api is enough: Conformal prediction for large language models without logit-access. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusId:268230867.

[109] H. Suri, Q. Zhang, W. Huo, Y. Liu, and C. Guan. Mediaqa: A question answering dataset on medical dialogues. *arXiv preprint arXiv:2108.08074*, 2021.

[110] J. H. Szabadváry. Beyond conformal predictors: Adaptive conformal inference with confidence predictors. *arXiv preprint arXiv:2409.15548*, 2024.

[111] S. Tayebati, D. Kumar, N. Darabi, D. Jayasuriya, R. Krishnan, and A. R. Trivedi. Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *ArXiv*, abs/2502.06884, 2025. URL https://arxiv.org/pdf/2502.06884.pdf.

[112] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

[113] J. W. Tukey. Non-parametric estimation II. Statistically equivalent blocks and tolerance regions–the continuous case. *The Annals of Mathematical Statistics*, 18(4):529–539, 1947.

[114] J. W. Tukey. Nonparametric estimation, III. Statistically equivalent blocks and multivariate tolerance regions–the discontinuous case. *The Annals of Mathematical Statistics*, 19(1):30–39, 1948.

[115] D. T. Ulmer, M. Gubri, H. Lee, S. Yun, and S. J. Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15440–15459. Association for Computational Linguistics, 2024.

[116] V. Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.

[117] V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, 1999.

[118] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

[119] A. Wald. An extension of Wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1):45–55, 1943.

[120] B. Wang, X. Yue, and H. Sun. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

[121] K. Wang, F. Duan, S. Wang, P. Li, Y. Xian, C. Yin, W. Rong, and Z. Xiong. Knowledge-Driven CoT: Exploring faithful reasoning in LLMs for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023.

[122] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

[123] Q. Wang, T. Geng, Z. Wang, T. Wang, B. Fu, and F. Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. *ArXiv*, abs/2410.08174, 2024. URL https://api.semanticscholar.org/CorpusId:273233940.

[124] X. Wang and R. J. Hyndman. Online conformal inference for multi-step time series forecasting. *arXiv preprint arXiv:2410.13115*, 2024.

[125] Z. Wang, Q. Wang, Y. Zhang, T. Chen, X. Zhu, X. Shi, and K. Xu. Sconu: Selective conformal uncertainty in large language models. In *unknown*, 2025. URL https://api.semanticscholar.org/CorpusId:277954950.

[126] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[127] S. S. Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

[128] S. Wu, X. Shen, and R. Xia. A new dialogue response generation agent for large language models by asking questions to detect user's intentions. *arXiv preprint arXiv:2310.03293*, 2023.

[129] H. Xi, K. Liu, H. Zeng, W. Sun, and H. Wei. Robust online conformal prediction under uniform label noise. *ArXiv*, abs/2501.18363, 2025. URL https://api.semanticscholar.org/CorpusId:275994111.

[130] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2251–2265, 2018.

[131] G. Xiong, J. Bao, and W. Zhao. Interactive-KBQA: Multi-turn interactions for knowledge base question answering with large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[132] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[133] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[134] F. Ye, M. Yang, J. Pang, L. Wang, D. Wong, E. Yilmaz, S. Shi, and Z. Tu. Benchmarking LLMs via uncertainty quantification. In *Advances in Neural Information Processing Systems*, volume 37, pages 15356–15385, 2024.

[135] F. Ye, M. Yang, J. Pang, L. Wang, D. F. Wong, E. Yilmaz, S. Shi, and Z. Tu. Benchmarking llms via uncertainty quantification. *ArXiv*, abs/2401.12794, 2024. URL https://arxiv.org/pdf/2401.12794.pdf.

[136] M. Zhang, M. Huang, R. Shi, L. Guo, C. Peng, P. Yan, Y. Zhou, and X. Qiu. Calibrating the confidence of large language models by eliciting fidelity. In *Conference on Empirical Methods in Natural Language Processing*, pages 2959–2979, 2024.

[137] Y. Zhang, J. Lu, and N. Jaitly. Probing the multi-turn planning capabilities of LLMs via 20 question games. *arXiv preprint arXiv:2310.01468*, 2023.

[138] T. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706, 2021.

[139] A. X. Zheng, I. Rish, and A. Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. *arXiv preprint arXiv:1207.1418*, 2012.

[140] Z. Zhi, C. Feng, A. Daneshmend, M. Orlu, A. Demosthenous, L. Yin, D. Li, Z. Liu, and M. Rodrigues. Seeing and reasoning with confidence: Supercharging multimodal llms with an uncertainty-aware agentic framework. *ArXiv*, abs/2503.08308, 2025. URL https://arxiv.org/pdf/2503.08308.pdf.

[141] H. Zhou, X. Wan, L. Proleev, D. Mincu, J. Chen, K. A. Heller, and S. Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*, 2024.

[142] C. Zhu, B. Xu, Q. Wang, Y. Zhang, and Z. Mao. On the calibration of large language models and alignment. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Appendix

A Outline

- Appendix §B is the Related Work section.
- Appendix §C discusses the limitations of this work.
- Appendix §D discusses broader impacts.
- Appendix §E provides additional results, which include results for other models, etc.
- Appendix §F provides further ablation studies.
- Appendix §G provides implementation details including hyperparameters.
- Appendix §H provides the prompts used for generating responses with LLMs.
- Appendix §I is the pseudocode for the different algorithms including IP, C-IP and DP.
- Appendix §J provides additional examples of query-answer chains.
- Appendix §K provides examples of sampled queries in open query set settings.

B Related Work

B.1 Large Language Models and Interactive Settings

Information-Seeking Environments of LLMs. In information seeking environments, LLM agents seek information by constructing well-design prompts and designing proper rewards with reinforcement learning (RL) to encourage proper information-seeking behavior. Prompting-based methods [1, 82, 120, 120, 121, 128] typically involve LLM agents interacting and solving a task together. Advanced techniques include building graphs [48], using tools such as APIs and databases [90, 131], and leveraging Chain-of-Thought (CoT) methods [20, 45, 70, 121] to ensure interaction is valid and improve performance. However, they do not explicit attempt to minimize the average number of interactions needed. On the other hand, RL-based methods such as ReAct and DeLLMa [71, 133] usually involves designing a good reward or utility function (can be non-entropy-based) that balances between exploration and exploitation, which our method can be classified in the pure exploitation regime. By and large, prompting-based methods and RL-based methods consider a different strategy that is not as explicit as our work here, and are not always plug-and-play to the interactive question-answering setting.

Interactive Medical Question Answering Setting. As our motivation (see Fig. 1), interactive medical question answering is one potentially impactful application of our work. While there are many existing benchmarks of medical question answering on LLMs such as MedQA [49], MedQuAD [10], MedMCQA [80], PubMedQA [50], they mostly focus on single-turn one-input-one-response type interaction. Instead, we focus on a more recently trending setting of interactive medical question answering and diagnosis. Datasets in this setting are still rare, with the two being the WMeDiaQA [109], which is a chinese interactive dataset, and the MediQ [65]. Importantly furthermore, it has been shown that LLMs in interactive human-LLM medical/clinical settings are subpar: as shown in Bean et al. [9], they perform nearly 30% worse than LLMs acting alone. Similarly, Li et al. [65] has also shown that directly applying LLMs for interactive settings perform worse without additional guidance to the iterative process.

To summarize, the research question of why LLMs in a multi-turn, interactive environments perform worse than single-turn use-cases remains active and highly relevant in today's use-cases [58, 65]. Our work seeks to provide some answer via the lens of information gain.

B.2 Uncertainty Quantification

Predictive Inference and Conformal Prediction. Our proposed formulation uses entropy and mutual information, which are fundamental measurements of uncertainty that dates back to Shannon's Information Theory [74, 101, 102] and predictive inference [97, 113, 114, 119, 127]. Amongst existing methods for predictive inference, conformal prediction has gained popularity due to its flexibility for constructing prediction intervals with a marginal coverage guarantee under the exchangeability of data points. While distribution-free inference and the conformal prediction framework have been extensively studied in recent works, with many going beyond standard assumptions [see, e.g., 5, 8, 23, 31, 32, 40–42, 52, 60–64, 67, 68, 83–87, 94, 96, 99, 104, 116–118], few has formalized the framework in the interactive and iterative setting. Arguably, the most similar setting is the online or adaptive setting [33, 34, 37, 38, 110, 124, 129], where distribution shift exists and the exchangeability assumption no longer holds. This line of research is different from our work in two major ways: 1) they focus on an online setting where data points are sequentially provided, such as time series, with assumptions such as affine transformations for the distribution shifts. In contrast, our work focuses on an interactive question answering setting that depends on multiple factors, such as the query set $\mathcal{Q}$ (open and closed setting), the query answers $q_{1:k}(x^{\text{obs}})$, and the target distribution Y; 2) they focus on providing tighter risk control in an online setting, whereas our work focuses on how to use uncertainty measured from prediction sets to drive decision making. While we agree adaptive conformal prediction and sequential information gain via conformal prediction draw some resemblances, we reserve the research of their precise relationship in a future work.

Uncertainty Quantification for LLMs. Methods for quantifying uncertainty in LLMs largely revolve around controlling the quality and ensuring correctness of LLM generations, either via minimizing entropy with a constructed distribution [13, 18, 44, 69, 81, 106, 135] or leveraging conformal prediction [22, 46, 53, 57, 76, 88, 91, 98, 100, 108, 111, 123, 125, 135, 140]. Here we highlight the ones that are highly relevant to our method: Ling et al. [70], which considers the problem of uncertainty quantification in the context of in-context learning. Similar to our IP baseline, it also measures uncertainty by computing the entropy of distributions estimated from the token distribution. On the other hand, Kumar et al. [56] applies conformal prediction to question answering with multiple choices. Their method also extracts probability of each choice by obtaining the logits of each answer token, but their setting is single-turn rather than multi-turn. As aforementioned, there exist multiple methods to measure uncertainty in LLMs, and we reserve how different choices might affect the interactive question-answering process in this paper's setting for future work.

Uncertainty Quantification in Interactive Environments. The two closely related works are Uncertainty of Thought (UoT) [45] and EQA work by Ren et al. [92]. UoT is in the setting of efficient reasoning for LLMs, in which the method formulates its exploration as a tree. At each iteration, UoT explores multiple branches of solutions with up to a depth at three, then evaluates the accumulated information gain at the leaf node, ultimately choosing the leaf with the highest information gain. Compared to C-IP, UoT considers only binary query-answers, considers accumulated rewards, and computes entropy directly, whereas C-IP considers free-text query answers, immediate rewards at each iteration, and utilizes calibrated information. On the other hand, Ren et al. [92] is positioned the setting of robotics, where the Vision-Language Model (VLM) agent has to answer a question by exploring different parts of the room. While this work also uses conformal inference and prediction set size to estimate confidence, the connection to entropy was not presented. It also differs from our language-domain setting by using a custom, domain-specific score function called "relevance" rather than the probability score as in the case of C-IP.

B.3 Sequential Information Gain via Information Pursuit

Previous Iterations of IP. Information Pursuit framework was first proposed by Jahangiri et al. [47] as an active-testing algorithm for scene parsing. Later learning-based approaches, such as generative approach [14] and variational approaches [15, 17, 24], require having access to the data distribution and learning a model to estimate the posterior distribution. Our work differs from all of the above in multiple aspects: 1) learning-based approaches of IP largely focus on computer vision tasks such as image classification and image generation, whereas our work focuses on the language domain; 2) their setting often provides sufficient samples that allows for a good estimation of the posterior distribution $\mathbb{P}(Y \mid \mathcal{H}_k)$, which enables efficient implementation with faster and scalable inference and make it applicable to large-scale tasks. In contrast, LLMs may not provide good estimates of the posterior distribution because they are pretrained models; 3) IP has mostly been written as an interpretability framework rather than a framework for interactive guidance. It's unclear yet how IP will perform in settings where multiple off-the-shelf models are involved; 4) query set in previous iterations of IP is always a pre-determined fixed set, whereas our work explores explores both the closed setting and an open generation setting where queries are obtained from LLMs iteratively; and 5) query answers in previous iterations of IP are either concept-based or raw features such as pixel-values. All the reasons above provide strong motivation to our current work.

Efficient Sampling. The challenge of efficiently obtaining samples for estimating mutual information often arises in IP. The generative approach [14] learns a joint distribution of $\mathbb{P}(Q(X), Y)$, then uses Langevin Dynamics to compute the conditional distribution. Alternative, V-IP [15], the current state-of-the-art implementation of IP, learns a query selection function g parameterized by a neural network g_θ, which also requires efficient sampling of random histories. In this work, our uniform parameterization is inspired from V-IP, whereas our sampling via LLM prompting provides a new domain-specific way to sample query-answer chains.

C Limitations

Our work explores the areas of using LLMs interactively and using information gain to obtain further information. While efficient, it still requires sampling in order to estimate quantities such as conditional entropy and prediction sets. One potential exploration is to find ways to estimate posteriors via a single pass, for example, via tractable approximations. Moreover, the success of our method also relies on LLMs that can follow instructions well, which today's LLMs are still not fully reliable yet and instruction-following in general is still an active area of research. Furthermore, as stated in our Section 6, we are not fully leveraging the potential of conformal prediction as we are only using it as a measure of uncertainty but not using it as means to control risk and evaluate the correctness of the overall query-answer chains. We reserve these interesting directions for future work.

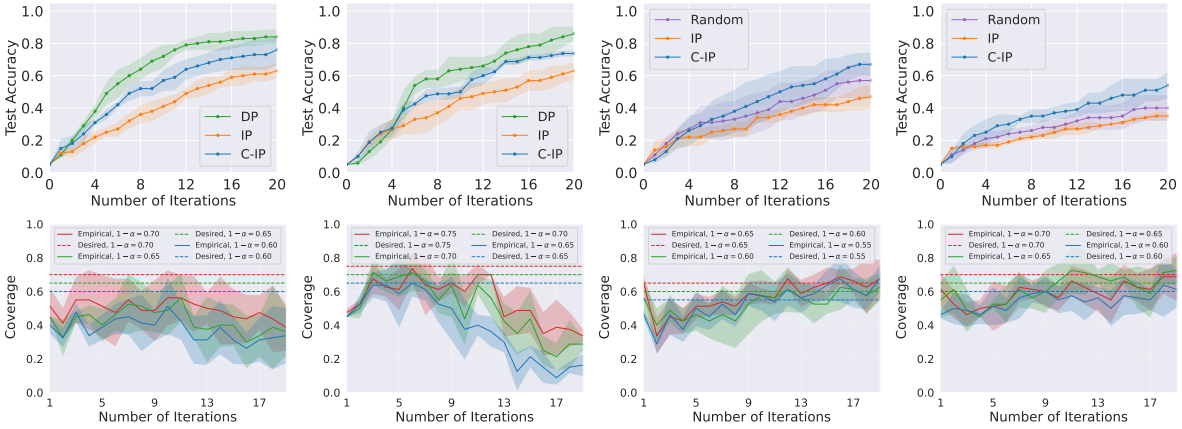
D Broader Impacts

Our work studies the setting of using LLMs interactively, where information is not provided all at once but sequentially depending on the previous responses of LLMs. While currently understudied, this research direction aligns closely with how human-human or human-LLM behaves in the real world, and potentially describes the predominant way of how human would use LLMs as the capabilities of LLMs grows. Hence, it is crucial to understand the behavior through careful mathematical formulation and experimental designs. We argue that the formulation in this work can serve as the foundational ground work for further analysis when interactively using LLMs. Our empirical study with interactive medical question answering also demonstrates potential applications of our method in real-world medical settings.

E Additional Results

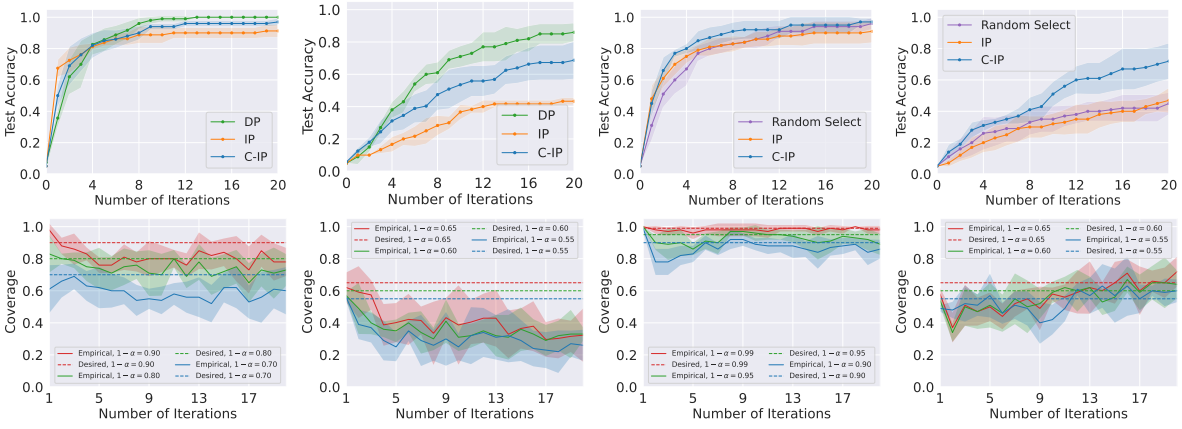
E.1 20Q

The predictive performance of 20Q on two additional models (Qwen2.5-7B [7] and Phi-3-small [2]) are shown in Figure 6 and Figure 7, respectively. Moreover, we visualize the different thresholds obtained from split conformal prediction in Figure 8, with thresholds for each model in each row. We will discuss the performance of C-IP with respect to the thresholds obtained together, as they demonstrate a picture of when C-IP would be successful and unsuccessful in guiding LLMs.



(a) $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{free-text}}$ (b) $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{binary}}$ (c) $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{free-text}}$ (d) $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{binary}}$

Figure 6: Predictive Performance of 20Q with Qwen-2.5-7B model.



(a) $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{free-text}}$ (b) $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{binary}}$ (c) $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{free-text}}$ (d) $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{binary}}$

Figure 7: Predictive Performance of 20Q with Phi-3-small model.

Successful Cases of C-IP. We first discuss the setting of 20Q with closed query set $\mathcal{Q}_{\text{closed}}$. It is clear that for all three models (including Llama-3.1-8B from Figure 3), C-IP outperforms IP and random query selection, where it is able to achieve a higher accuracy at every iteration. One sign of this success can be attributed to the fact that our empirical marginal coverage guarantee follows closely to the desired guarantee at each iteration.

Alternatively, one can attribute this success by looking at the thresholds obtained through split conformal prediction from Figure 8. In the closed query set setting (right two columns), we observe a progressive increase in the

thresholds also as the number of iterations increase. This matches with our intuition: in the first few iterations, the model is not confident about its prediction, hence the threshold τ remains small, and a large number of classes y would be in the prediction set. As the number of iterations increase, the predictor (on average relative to the calibration set) becomes more confident, hence the threshold τ increases.

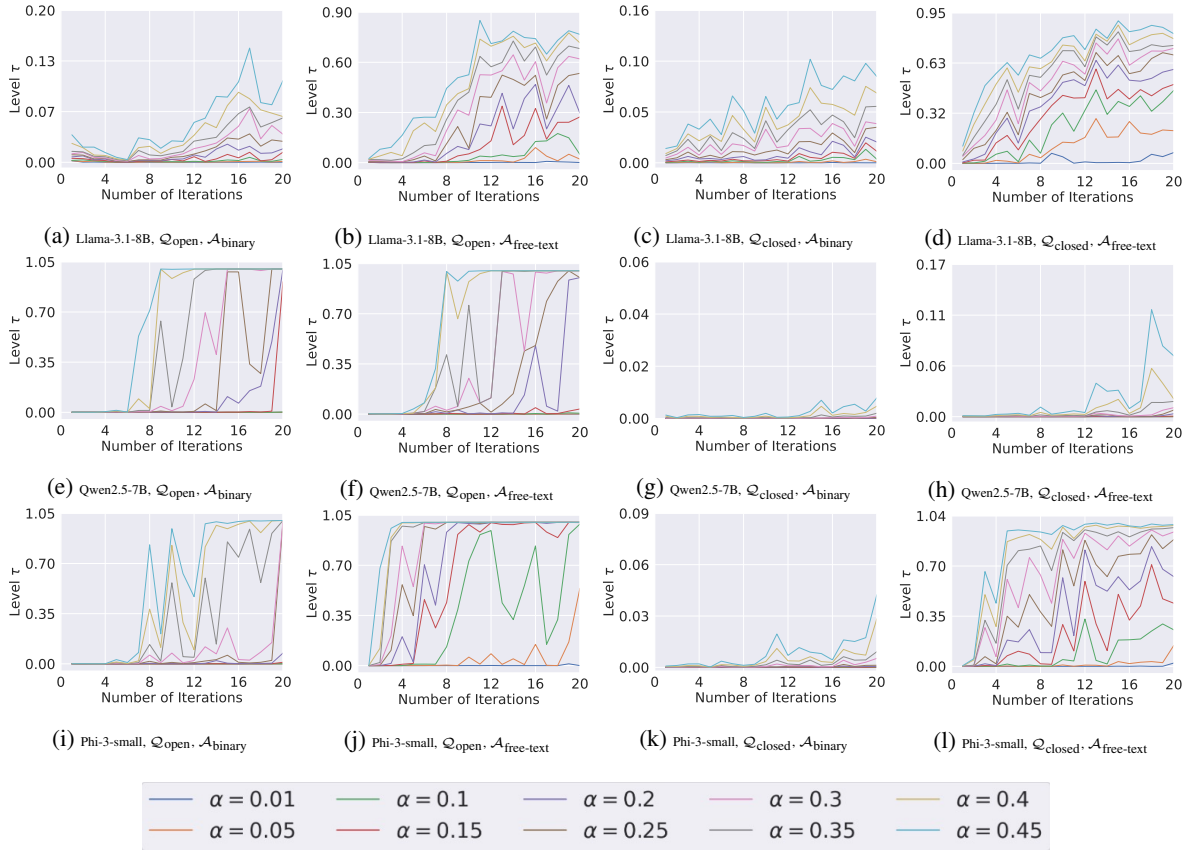


Figure 8: Thresholds obtained through split conformal prediction for 20Q. The legend below shows different colors Each color corresponds to different choices of target coverage α.

Unsuccessful Cases of C-IP. We now turn to the unsuccessful cases of C-IP, which largely pertains the open query set setting for Qwen2.5-7B and Phi-3-small. We observe that C-IP in fact does not outperform the DP method. To understand why, we argue it from the perspective of *meaningful calibration*.

Recall that in the open query set setting, C-IP is calibrated on query-answer chains obtained from DP. Turning to Figure 6 (a, b) and Figure 7 (a, b), we observe that the empirical coverage is below the desired coverage a majority of the iterations, a sharp contrast to successful cases of C-IP where empirical and desired coverage is mostly coincides. To explain the misalignment between the empirical coverage and desired coverage, we can also infer from Figure 8, where we observe sudden jumsp in the thresholds in the open query set setting for Qwen2.5-7B and Phi-3-small. From our results, we observe the following: Given Qwen2.5-7B and Phi-3-small models are relatively strong models and our task is relatively simple, it is often the case where the label is guessed correctly within a few iteration. Once the correct prediction is made, it becomes part of the history $\mathcal{H}_k$ (e.g. "Is the animal a dolphin? Yes.") and any new query no longer provides any meaningful information gain. This ultimately leads to an over-confident prediction for every data point, yielding an uninformative threshold. As a result, every prediction now requires a very high probability (near 1) to be an element in the prediction set. Consequently, the prediction sets cannot quantify uncertainty in a meaningful way, leading to suboptimal results.

While there exist cases where C-IP does not outperform other methods, we argue it is more interpretable than IP in

that one can observe its failure modes from the thresholds obtained and empirical coverage. As discussed in our Future Work (Section §6) and Limitations (Appendix §C), this direction of obtaining the *proper* conformal coverage guarantee over distributions of natural language deserves further study, and we reserve this for future work.

Additional Comparisons with UoT. We provide the full comparison with UoT for Llama-3.1-8B in Table 2. Focusing on IP and C-IP, we can see that C-IP is able to outperform IP in all four settings, which aligns with our previous results regarding the predictive performance. Comparing C-IP with UoT, C-IP is able to outperform UoT in free-text query answers, but not in closed binary query answer setting. This can be explained by the design of the respective methods: C-IP relies on immediate reward (information gain) whereas UoT looks ahead, accumulates and propagates rewards up to the certain depth before selecting the next query.

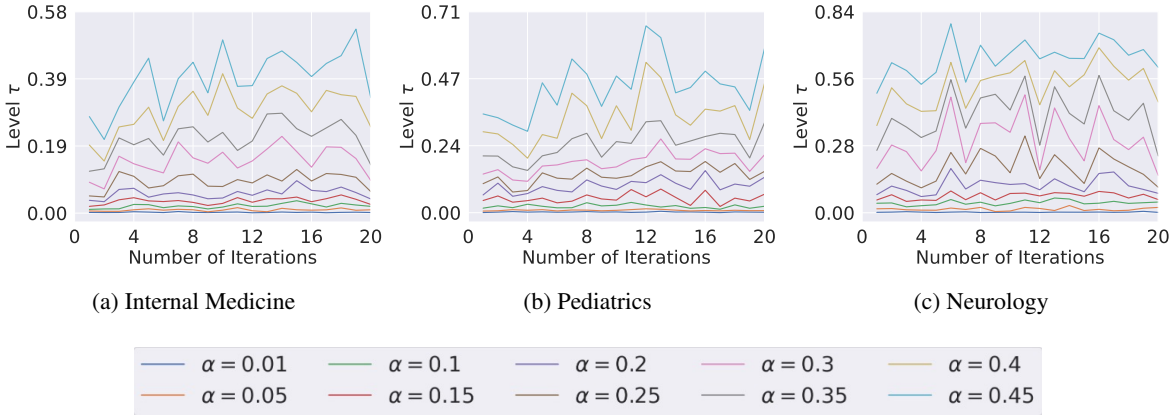
As discussed previously in Section §E.1, while having binary query answers allow UoT to look forward and accumulate rewards (akin to a beam search), free-text query answers provide a much larger marginal gain over binary query answers and is a more realistic setting to consider.

Table 2: Full Comparison of C-IP with UoT reasoning, averaged across 5 runs with shaded area as standard deviation.

Method	**Query Set $\mathcal{Q}$**	**Query Answers $\mathcal{A}$**	**Avg. Len.**	**Avg. Success Len.**	**Success Rate**
UoT	open	binary	13.57 ± 1.69	8.75 ± 0.82	0.57 ± 0.13
IP	open	binary	11.40 ± 0.78	10.48 ± 0.98	0.31 ± 0.04
IP	open	free-text	10.04 ± 0.45	9.66 ± 0.68	0.65 ± 0.18
IP	closed	binary	10.74 ± 0.59	10.98 ± 0.86	0.29 ± 0.11
IP	closed	free-text	13.81 ± 0.16	13.73 ± 0.24	0.71 ± 0.08
C-IP ($\alpha = 0.1$)	open	binary	11.87 ± 0.83	12.12 ± 0.26	0.37 ± 0.11
C-IP ($\alpha = 0.1$)	open	free-text	10.58 ± 0.58	10.45 ± 0.63	0.83 ± 0.05
C-IP ($\alpha = 0.1$)	closed	binary	11.26 ± 0.84	10.58 ± 0.61	0.29 ± 0.04
C-IP ($\alpha = 0.1$)	closed	free-text	8.87 ± 0.31	8.24 ± 0.51	0.83 ± 0.07

E.2 MediQ

The thresholds τ obtained through split conformal prediction is shown in Figure 9. Similar to our observations in the 20Q case, we observe that a useful and successful case of C-IP can be attributed to having good calibration and finding meaningful thresholds.



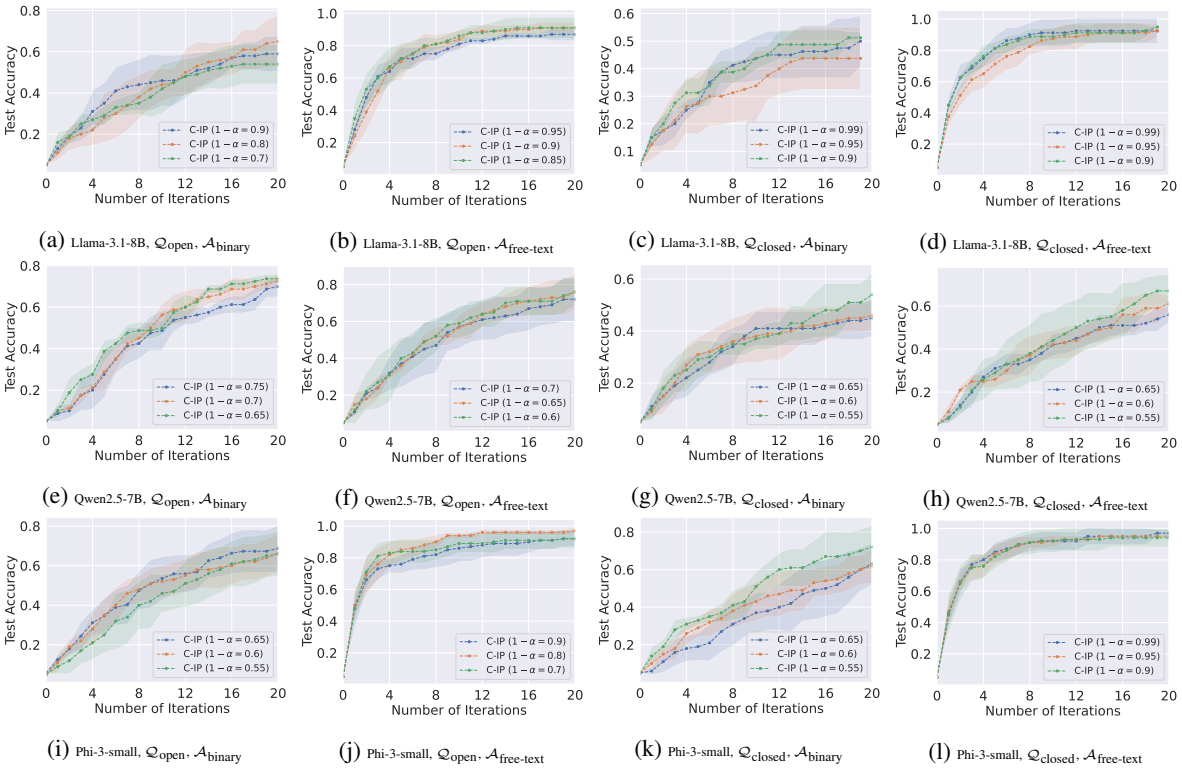
(a) Internal Medicine (b) Pediatrics (c) Neurology

Figure 9: Thresholds obtained through split conformal prediction for MediQ. The legend below shows different colors Each color corresponds to different choices of target coverage α.

F Ablation Studies

F.1 Varying Desired Coverage

We compare predictive performance of 20Q (Figure 10) and MediQ (Figure 11). In nearly all cases, we observe only minor differences between different levels of α, which indicates C-IP is fairly robust with the appropriate choices of α.

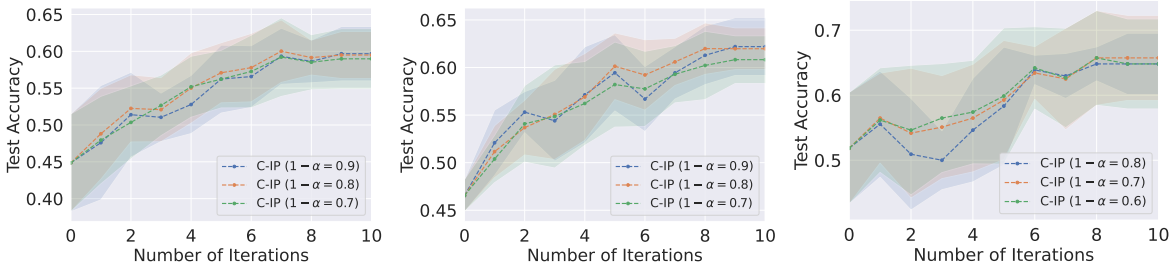


(a) Llama-3.1-8B, $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{binary}}$ (b) Llama-3.1-8B, $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{free-text}}$ (c) Llama-3.1-8B, $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{binary}}$ (d) Llama-3.1-8B, $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{free-text}}$

(e) Qwen2.5-7B, $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{binary}}$ (f) Qwen2.5-7B, $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{free-text}}$ (g) Qwen2.5-7B, $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{binary}}$ (h) Qwen2.5-7B, $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{free-text}}$

(i) Phi-3-small, $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{binary}}$ (j) Phi-3-small, $\mathcal{Q}_{\text{open}}$, $\mathcal{A}_{\text{free-text}}$ (k) Phi-3-small, $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{binary}}$ (l) Phi-3-small, $\mathcal{Q}_{\text{closed}}$, $\mathcal{A}_{\text{free-text}}$

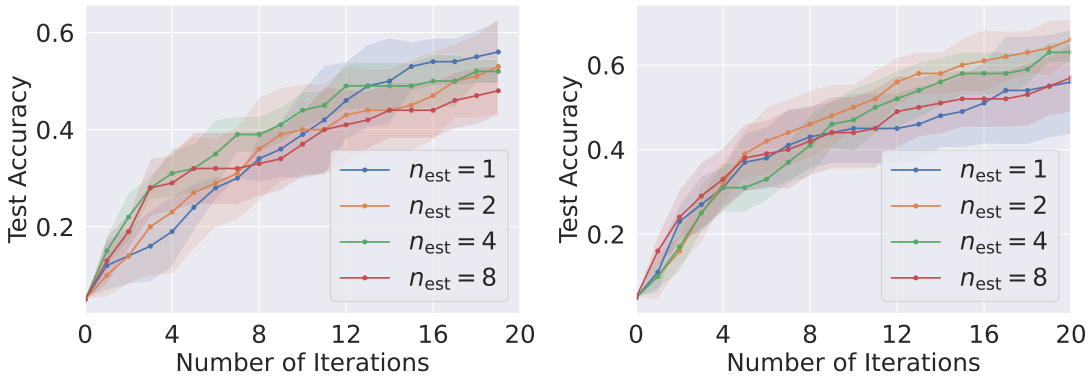
Figure 10: Predictive Performance of 20Q with different choices of α.



(a) Internal Medicine (IM) (b) Pediatrics (P) (c) Neurology (N)

Figure 11: Predictive Performance of MediQ with different choices of α.

F.2 Varying Number of Estimation Samples



(a) Closed Query Set $\mathcal{Q}_{\text{closed}}$

(b) Closed Query Set $\mathcal{Q}_{\text{open}}$

Figure 12: Performance of 20Q with different number of estimation samples. Each curve corresponds to the averaged accuracy over 5 runs, with shaded areas representing their standard deviation.

We evaluate whether C-IP is sensitive to the number of samples n_{est} used during estimation for the task of 20Q on Llama-3.1-8B (Figure 12). Recall that by the number of estimation samples n_{est}, we refer to the number of samples used to estimate the upper bound (C-IP). In our ablation study for 20Q on Llama-3.1-8b with closed query set $\mathcal{Q}_{\text{closed}}$, binary query answers $\mathcal{A}_{\text{binary}}$ and open query set $\mathcal{Q}_{\text{open}}$, binary query answers $\mathcal{A}_{\text{binary}}$, we find that the number of samples do have impact on the performance. While one would expect the more is better, but we find that using $n_{\text{est}} \in \{1, 2, 4, 8\}$ performs somewhat similarly. One possible explanation that aligns with empirical observations of LLMs is that stochasticity plays a role in obtaining better predictive performance in LLMs, akin to how LLMs decoding via sampling based on output token distribution often outperforms greedy decoding where tokens are selected deterministically by choosing the most likely token [43].

F.3 Varying Number of Queries Sampled for Open Query Set.

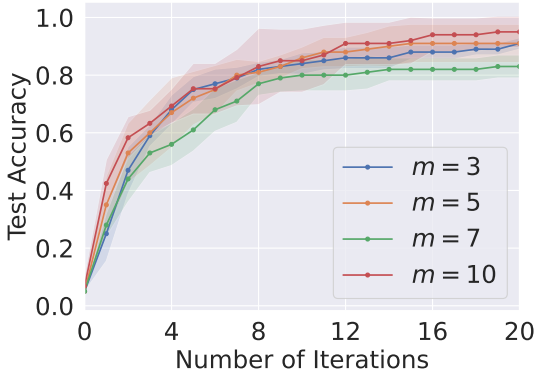


Figure 13: Performance of 20Q in the open query setting with different number of sampled queries m. Here we focus on Free-text Query Answer $\mathcal{A}_{\text{free-text}}$. Each curve corresponds to the averaged accuracy over 5 runs, with shaded areas representing their standard deviation.

We evaluate whether C-IP on Llama-3.1-8B for 20Q with the open query set $\mathcal{Q}_{\text{open}}$ is sensitive to the number of queries m asked at each iteration (Figure 13). We observe a comparable performance for all choices of m.

G Implementation Details

G.1 Code Release

We will release the code in the near future.

G.2 Computational Resources

The experiments are conducted on a workstation of 8 NVIDIA A5000 GPUs. Each LLM used in this work is able to be loaded and run on a single A5000 GPU.

G.3 Code

All experiments are implemented in Python 3.12. The main packages used are huggingface, PyTorch, Numpy, and TogetherAI API (for UoT baseline). The three models used are `meta-llama/Meta-Llama-3.1-8B-Instruct`, `microsoft/Phi-3-small-128k-instruct`, and `Qwen/Qwen2.5-7B-Instruct`. Unless stated otherwise, we use the default hyperparameters from huggingface. We use the following LLM hyperparameters every time we inference: `do_sample=True`, `temperature=0.7`, and `max_new_tokens=1024`.

G.4 Classes for 20Q and MediQ

For 20Q, we select 20 classes from all 50 classes that are available in the Animals with Attributes 2 [130] dataset. The 20 classes are: giraffe, zebra, elephant, killer whale, dalmatian, polar bear, giant panda, hippopotamus, rhinoceros, lion, tiger, blue whale, walrus, grizzly bear, siamese cat, cow, german shepherd, gorilla, dolphin, and moose.

For MediQ, each medical question is a multiple-choice question with four options: A, B, C, and D. Note that the possible choices and order of the options are different for each problem.

G.5 Obtaining Probabilities from LLM

In this work, we focus on LLMs where logit scores of output tokens are accessible. While there are many methods for obtaining a posterior distribution $\mathbb{P}(Y \mid \mathcal{H}_k)$ for a given x, we estimate the posterior by first obtaining the LLMs' output token logits based on the class labels' tokens, then applying softmax function. When certain class names consist of multiple tokens, we select the first token's probability to represent the class's probability. For all of our models, none of the classes consist of the same *first* token. In the case that happens, one may consider adding a prefix such as enumerations ("1.", "2.", ...) or leveraging special symbols trained with large language models.

G.6 Entropy Estimation

Estimation of terms such as entropy in (IP) or upper bounds in (C-IP) requires samples of (X, Y) pairs. We use 4 randomly drawn samples for all 20Q experiments whenever a entropy term needs to be estimated and use 12 randomly drawn samples from the estimation set $\mathcal{D}_{\text{est}}$ for all MediQ experiments.

G.7 Calibration

Achieving marginal coverage in (11) requires calibration samples. For 20Q, we sample 100 randomly drawn labels. For MediQ, we sample 200 randomly drawn datapoints with replacement in the calibration set $\mathcal{D}_{\text{cal}}$.

G.8 Preprocessing Query Answers for 20Q

To avoid generating query-answers from LLMs at each iteration for every query, in the case of closed query set $\mathcal{Q}_{\text{closed}}$ and non-binary query-answers, we generate query-answers offline by sampling 10 possible query-answers for every class and query. Then, in our experiments, for any given query and class, we select a query answer by uniformly sampling one out of the ten generated answers. For binary answers, we only consider query-answers with "Yes" and "No", without any variability.

G.9 Evaluation of Predictive Accuracy

We discuss the two evaluation methods for 20Q and MediQ separately as they are different. 20Q is often played in a manner that the Querier LLM keeps guessing until either it guesses correctly or until the maximum number of iterations L is reached[6]. Here, we following the same protocol. If the Querier LLM guesses the correct prediction $\hat{y}_k^*$ at iteration k, then $\hat{y}_k = \hat{y}_i^*$ for $i = k+1, \ldots, L$. For MediQ, since the neither the Expert LLM or Patient LLM knows what the correct prediction is, we require C-IP to reach the stopping criteria (See Appendix §G.10). For any given datapoint x^{obs}, suppose the stopping criterion is reached at iteration $\tilde{k} < L$ and $\hat{y}_{\tilde{k}}$ is the prediction. Then we assume the prediction does not change after iteration $\tilde{k}$ and set $\hat{y}_i = \hat{y}_{\tilde{k}}$ for $i = \tilde{k}+1, \ldots, L$.

G.10 Stopping Criteria

As mentioned in Section 3.2 Remark 1, arriving at $I(q(X); Y \mid \mathcal{H}_k) = 0$ in practice is highly unlikely. Here, we explain how one goes about arrive at an approximation $I(q(X); Y \mid \mathcal{H}_k) \approx 0$.

One approach is to utilize the model confidence $\mathbb{P}(Y \mid \mathcal{H}_k)$. In previous variations of IP, including the generative version [14] and the variational approaches [15, 27], the algorithm stops querying once the posterior $\mathbb{P}(y \mid \mathcal{H}_k) > \epsilon$ for any $y \in \mathcal{Y}$. This is suited for previous cases because the predictor model f is learned from training data, which provides better estimates for the posterior. Unfortunately, in the case where model confidence is miscalibrated, this process becomes unstable and unreliable.

Empirically, we observe that while estimated conditional entropy (either with direct computation or with approximation via prediction sets) for every q, i.e. $\hat{H}(Y \mid q(X), \mathcal{H}_k)$, does not converge to 0 as the sequence length increases. However, it tends to converge to some constant number when the algorithm is confident enough to some number. Hence we propose to calculate the standard deviation between obtained estimations. Precisely, given a query set $\mathcal{Q}$ (closed or open), the query-answer chain $q_{1:k}(x^{\text{obs}})$, and a stopping threshold ϵ, the stopping criterion can be described as:

- *Step 1.* Let

$$c_i = \hat{H}(Y \mid q(X), \mathcal{H}_k) \quad \text{for } q \in \mathcal{Q}. \tag{14}$$

[6] This is also consistent with UoT [45].

- *Step 2.* Compute the estimated standard deviation

$$\bar{c} = \sum_{i=1}^{|\mathcal{Q}|} c_i, \quad \hat{\sigma} = \sqrt{\frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} (c_i - \bar{c})^2}. \tag{15}$$

- *Step 3.* Stop if $\hat{\sigma} < \epsilon$, else continue the algorithm.

G.11 Evaluation of Empirical Coverage

Suppose we have a test set $\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{test}}}$ with size n_{test} for evaluation. The empirical coverage at iteration k is evaluated by

$$\mathbb{P}_{X,Y,Q_{1:k},\mathcal{D}_{\text{cal}}} \left(Y \in \mathcal{C}_{\hat{\tau}_k}(Q_{1:k}(X)) \right) \approx \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}\{y_i \in \mathcal{C}_{\hat{\tau}_k}(q_{1:k}(x_i))\}, \tag{16}$$

where $(x_i, y_i) \in \mathcal{D}_{\text{test}}$ is the i-th test sample, $\mathcal{C}_{\hat{\tau}_k}$ is the prediction set for query-answer chains of length k, $q_{1:k}(x_i)$ is the obtained query answer chain (e.g. from C-IP), and $\mathbb{1}(\cdot)$ is a binary indicator function, which equals to 1 when the condition in the parameter is true and 0 otherwise.

G.12 Hyperparameters for UoT Baseline

Unless specified below, we follow the default settings as suggested in UoT's main manuscript [45] and their Github repository `https://github.com/zhiyuanhubj/UoT/tree/main/src/uot`.

For results in Table 1, to ensure the fairness of the experiments, we set the number of branches in their method (`n_potential_actions` in their code) to 5. Their experimental results are produced with API from TogetherAI. Please see their website `https://www.together.ai/` for more details. The total charges to produce our results for the three models cost no more than $50.

H Prompts

Here we provide the prompts used in our experiments.

H.1 20Q

System Instruction for Querier LLM

```
You are an expert on animals. Your goal is to predict the animal given the information you have gathered. The animal must be one of the
↪  following: {class_names}. Be as precise and as direct as possible. You may provide your reasoning first before making a prediction. End
↪  your response by making a guess and saying 'The animal is: X' (e.g. The animal is: dog) at the end of your reasoning, where X is your
↪  guess. Do not provide any additional information. Your response should all fit in a single paragraph. If you are not provided any
↪  information, make a random guess.
```

Input Prompt for Querier LLM Obtaining the Posterior

```
You have not gathered any information yet. Please make a random guess. # If history is empty

Here is the information you have gathered. # If history is not empty
1. {q1} {q1(x)}
2. {q2} {q2(x)}
3. {q3} {q3(x)}
4. {q4} {q4(x)}
...

Given the information you have gathered, make an intermediate SINGLE prediction of what you think the animal is. First make your guess in the
↪  format 'The animal is: X' (e.g. The animal is: dog), where X is your guess, then provide your reasoning. Do not provide any additional
↪  information. Your response should all fit in a single paragraph. Make sure your prediction is one of the classes: {class_names}.
```

Input Prompt for Querier LLM Suggesting Queries in Open Query Set Setting

```
You have not gathered any information yet. Please make a random guess. # If history is empty

Here is the information you have gathered. # If history is not empty
1. {q1} {q1(x)}
2. {q2} {q2(x)}
3. {q3} {q3(x)}
4. {q4} {q4(x)}
...

Now suggest {n_queries_per_step} questions.\nReturn the question in this format:\n{{"questions": ["QUESTION_1", "QUESTION_2", "QUESTION_3"]}}
```

System Instruction for Expert LLM for Non-binary Query Answers

```
You are an expert on {label}. Based on the question provided, answer truthfully about the question. Do not directly tell the other player
↪  what you are thinking. Be as precise and as direct as possible, and answer in complete sentence. For example, if the question is "Does
↪  the animal have a tail?", you can answer "The animal has a tail." without saying yes or no. Do not say the name of the animal in your
↪  answer.
```

System Instruction for Expert LLM for Binary Query Answers

```
You are an expert on {label}. Based on the question provided, answer truthfully about the question. Do not directly tell the other player
↪  what you are thinking. Be as precise and as direct as possible, and answer with a single word. For example, if the question is "Does the
↪  animal have a tail?", you can answer "Yes." or "No.". Do not say the name of the animal in your answer. If you don't know the answer,
↪  make a guess. Do not answer anything other than "Yes." or "No.".
```

H.2 MediQ

System Instruction for Expert LLM Obtaining the Posterior

```
You are a medical doctor specialized in {specialty}, trained to provide accurate, evidence-based responses to medical inquiries. Your goal is
↪  to answer questions with clarity, precision, and professionalism while ensuring your responses align with established medical guidelines.
↪  Answer concisely, accurately, and compassionately. Make a prediction and provide your reasoning as explanation. Respond in the following
↪  format:

{{"answer": "A/B/C/D", "explanation": "YOUR EXPLANATION HERE"}}
```

Input Prompt for Expert LLM Obtaining the Posterior

```
Answer the multiple choice based on the context.

Context: {context}

Question: {question}

Options:
    A - {option_a}
    B - {option_b}
    C - {option_c}
    D - {option_d}

Please select the most appropriate answer (A/B/C/D).
```

System Prompt for Converting Facts into Queries

```
Convert the medical fact into a question, in which the answer is the fact itself. The question should be specific and relevant to the
↪  patient's condition. Please do not ask any questions that are not related to the patient's medical history or condition. Suggest one
↪  question only. Return only your question and nothing else.

Medical fact: He has a non-productive cough for 4 months.
Question: What are some preliminary symptoms?

Medical fact: He complains of nausea and 1 episode of vomiting during the past day.
Question: Did the patient complain about nausea?
```

System Prompt for Converting Facts into Queries

```
Medical fact: {fact}
```

I Pseudocode for Implementation

Here we provide the pseudocode for algorithms used in this paper.

Algorithm 1 Constructing Prediction set functions for each length $\mathcal{C}_{\hat{\tau}_1} \dots \mathcal{C}_{\hat{\tau}_L}$.

Input: Calibration set $\mathcal{D}_{\text{cal}}$ with n_{cal} samples, Desired Coverage $1-\alpha$, maximum iteration L, LLM-based predictor f.

for each length $k = 1 \dots L$ **do**

Step 1: Obtain conformal scores

for $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ **do**

Option 1.1: Uniform Sampling

Sample random history $q_{1:k} \sim Q_{1:k}$ based on Equation 12.

Option 1.2: DP Sampling

Sample histories from direct prompting $q_{1:k} \sim \text{DP}(X, Y)$ based on Equation 13.

Obtain scores based on LLM's output token logits for each class

$$s_i(x_i, y_i) = f(q_{1:k}(x_i))_{y_i}$$

end for

Step 2: Quantile Estimation

Estimate quantiles using the obtained scores:

$$\hat{\tau}_k = \text{Quantile}\left(\{s_i\}_{i=1}^{n_{\text{cal}}}; \frac{\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil}{n_{\text{cal}}}\right)$$

Step 3: Prediction Set Construction

Construct and define the prediction set function using the calculated quantile

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

end for

return Prediction set functions for each length $\mathcal{C}_{\hat{\tau}_1} \dots \mathcal{C}_{\hat{\tau}_L}$.

Algorithm 2 Information Pursuit with a closed query set $\mathcal{Q}_{\text{closed}}$.

Input: Observation or test sample x^{obs}, maximum iteration L, LLM-based predictor f, query set $\mathcal{Q} = \mathcal{Q}_{\text{closed}}$, estimation set $\mathcal{D}_{\text{est}}$

Output: Selected queries $q_{1:k}$ and predictions $\hat{y}_{1:k}$

1: Set iteration $k \leftarrow 0$
2: Initialize an empty history $\mathcal{S}_0 = \{\}$
3: **while** stopping criteria is met or $k < L$ **do**
4: Predict $y_{k+1} = \text{argmax}_{y \in \mathcal{Y}} f(y \mid S_k)$
5: Estimate the entropy of Y given each query q using $\mathcal{D}_{\text{est}}$ select the most informative query

$$q_{k+1} = \underset{q \in \mathcal{Q}}{\text{argmin}}\, H(Y \mid q(X), \mathcal{S}_k)$$

6: Compute query answer $q_{k+1}(x^{\text{obs}})$ and update current history

$$\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{q_{k+1}(x^{\text{obs}}\}$$

7: $k \leftarrow k + 1$
8: **end while**

Algorithm 3 Information Pursuit with an open query set $\mathcal{Q}_{\text{open}}$.

Input: Observation or test sample x^{obs}, maximum iteration L, LLM-based predictor f, an LLM $\text{LLM}(\cdot)$, m queries to sample at each iteration, estimation set $\mathcal{D}_{\text{est}}$

Output: Selected queries $q_{1:k}$ and predictions $\hat{y}_{1:k}$

1: Set iteration $k \leftarrow 0$
2: Initialize an empty history $\mathcal{S}_0 = \{\}$
3: **while** stopping criteria is met or $k < L$ **do**
4: Predict $\hat{y}_{k+1} = \text{argmax}_{y \in \mathcal{Y}} f(y \mid S_k)$
5: Prompt language model for m queries based on current history

$$\mathcal{Q} = \{q_j\}_{j=1}^{m} \leftarrow \text{LLM}(\mathcal{Y}, q_{1:k}(x^{\text{obs}}), m)$$

6: Estimate the entropy of Y given each query q using $\mathcal{D}_{\text{est}}$ select the most informative query

$$q_{k+1} = \underset{q \in \mathcal{Q}}{\text{argmin}}\, H(Y \mid q(X), \mathcal{S}_k)$$

7: Compute query answer $q_{k+1}(x^{\text{obs}})$ and update current history

$$\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{q_{k+1}(x^{\text{obs}}\}$$

8: $k \leftarrow k + 1$
9: **end while**

Algorithm 4 Conformal Information Pursuit with a closed query set $\mathcal{Q}_{\text{closed}}$.

Input: Observation or test sample x^{obs}, maximum iteration L, LLM-based predictor f, query set $\mathcal{Q} = \mathcal{Q}_{\text{closed}}$, Prediction set functions for each length $\mathcal{C}_{\hat{\tau}_1} \dots \mathcal{C}_{\hat{\tau}_L}$, estimation set $\mathcal{D}_{\text{est}}$

Output: Selected queries $q_{1:k}$ and predictions $\hat{y}_{1:k}$

Set iteration $k \leftarrow 0$

Initialize an empty history $\mathcal{S}_0 = \{\}$

while stopping criteria is met or $k < L$ **do**

  Predict $\hat{y}_{k+1} = \text{argmax}_{y \in \mathcal{Y}} f(y \mid S_k)$

  Estimate the entropy of Y given each query q using $\mathcal{D}_{\text{est}}$ select the most informative query

$$q_{k+1} = \min_{q \in Q} \left\{ \lambda_\alpha + (1 - \alpha_N) \log \mathbb{E}_X [|\mathcal{C}_{\hat{\tau}_{k+1}}(q(X), \mathcal{S}_k)|] \right\}$$

  Compute query answer $q_{k+1}(x^{\text{obs}})$ and update current history

$$\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{q_{k+1}(x^{\text{obs}})\}$$

  $k \leftarrow k + 1$

end while

Algorithm 5 Conformal Information Pursuit with an open query set $\mathcal{Q}_{\text{open}}$.

Input: Observation or test sample x^{obs}, maximum iteration L, LLM-based predictor f, an LLM $\text{LLM}(\cdot)$, m queries to sample at each iteration, Prediction set functions for each length $\mathcal{C}_{\hat{\tau}_1} \dots \mathcal{C}_{\hat{\tau}_L}$, estimation set $\mathcal{D}_{\text{est}}$

Output: Selected queries $q_{1:k}$ and predictions $\hat{y}_{1:k}$

Set iteration $k \leftarrow 0$

Initialize an empty history $\mathcal{S}_0 = \{\}$

while stopping criteria is met or $k < L$ **do**

  Predict $\hat{y}_{k+1} = \text{argmax}_{y \in \mathcal{Y}} f(y \mid S_k)$

  Prompt language model for m queries based on current history

$$\mathcal{Q} \leftarrow \text{LLM}(\mathcal{Y}, q_{1:k}(x^{\text{obs}}), m)$$

  Estimate the entropy of Y given each query q using $\mathcal{D}_{\text{est}}$ select the most informative query

$$q_{k+1} = \min_{q \in Q} \left\{ \lambda_\alpha + (1 - \alpha_N) \log \mathbb{E}_X [|\mathcal{C}_{\hat{\tau}_{k+1}}(q(X), \mathcal{S}_k)|] \right\}$$

  Compute query answer $q_{k+1}(x^{\text{obs}})$ and update current history

$$\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{q_{k+1}(x^{\text{obs}})\}$$

  $k \leftarrow k + 1$

end while

Algorithm 6 Direct Prompting

Input: Observation or test sample x^{obs}, maximum iteration L, a Querier LLM QuerierLLM$(\cdot)$, a Expert LLM ExpertLLM$(\cdot)$, LLM-based predictor f

Output: Selected queries $q_{1:k}$ and predictions $\hat{y}_{1:k}$

1: Initialize an empty history $\mathcal{S}_0 = \{\}$

2: **for** $k = 0, \ldots, L$ **do**

3: Predict with f

$$\hat{y}_k = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, f(y \mid S_k)$$

4: Prompt Querier LLM for one single query given history

$$q_{k+1} = \text{QuerierLLM}(\mathcal{S}_k)$$

5: Obtain query answer and update history

$$q_{k+1}(x^{\text{obs}}) = \text{ExpertLLM}(q_{k+1}, x^{\text{obs}})$$
$$S_{k+1} = \mathcal{S}_k \cup \{q_{k+1}(x^{\text{obs}})\}$$

6: **end for**

J Additional Examples of Query-Answer Chains

In this section, we provide additional examples of query-answer chains obtained using C-IP. All the examples we provide in the following are correct predictions. Here we only show the iteration that the method stopped.

J.1 20Q

Here we provide one example for each setting, which is either open or closed query set and either binary or free-text query answers.

Example of 20Q with Closed Query Set and Binary Query Answers

```
## System Instruction
You are an expert on animals. Your goal is to predict the animal given the information you have gathered. The animal must be one of the
↪  following: 'giraffe', 'zebra', 'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros',
↪  'lion', 'tiger', 'blue whale', 'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'. Be as
↪  precise and as direct as possible. You may provide your reasoning first before making a prediction. End your response by making a guess
↪  and saying 'The animal is: X' (e.g. The animal is: dog) at the end of your reasoning, where X is your guess. Do not provide any
↪  additional information. Your response should all fit in a single paragraph. If you are not provided any information, make a random guess.

## Input Prompt
Here is the information you have gathered.
1. Does the animal live in the Arctic? No.
2. Does the animal have stripes? No.
3. Is the animal domestic? Yes.

Given the information you have gathered, make an intermediate SINGLE prediction of what you think the animal is. First make your guess in the
↪  format 'The ansmal is: X' (e.g. The animal is: dog), where X is your guess, then provide your reasoning. Do not provide any additional
↪  information. Your response should all fit in a single paragraph. Make sure your prediction is one of the classes: 'giraffe', 'zebra',
↪  'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros', 'lion', 'tiger', 'blue whale',
↪  'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'.

## Response
The animal is: cow. Given the information that the animal lives in the Arctic is no, which eliminates polar bear, walrus, and grizzly bear,
↪  and that it is domestic, which eliminates most wild animals, leaving domesticated options such as a cow.
```

Example of 20Q with Closed Query Set and Free-text Query Answers

```
## System Instruction
You are an expert on animals. Your goal is to predict the animal given the information you have gathered. The animal must be one of the
↪  following: 'giraffe', 'zebra', 'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros',
↪  'lion', 'tiger', 'blue whale', 'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'. Be as
↪  precise and as direct as possible. You may provide your reasoning first before making a prediction. End your response by making a guess
↪  and saying 'The animal is: X' (e.g. The animal is: dog) at the end of your reasoning, where X is your guess. Do not provide any
↪  additional information. Your response should all fit in a single paragraph. If you are not provided any information, make a random guess.

## Input Prompt
Here is the information you have gathered.
1. This animal primarily inhabits grasslands and savannas, rather than dense forests.

Given the information you have gathered, make an intermediate SINGLE prediction of what you think the animal is. First make your guess in the
↪  format 'The ansmal is: X' (e.g. The animal is: dog), where X is your guess, then provide your reasoning. Do not provide any additional
↪  information. Your response should all fit in a single paragraph. Make sure your prediction is one of the classes: 'giraffe', 'zebra',
↪  'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros', 'lion', 'tiger', 'blue whale',
↪  'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'.

## Response
The animal is: zebra. This is because the information provided suggests a grassland or savanna habitat, which aligns with the typical
↪  habitats of zebras, as they are commonly found in grasslands and savannas.
```

J.2 MediQ

Here we provide one example for each specialty. Recall that, query answers (denoted as `Patient's Fact` in the prompt), is obtained in the order that C-IP obtains. When formatting as input, we ignore the query as that provides repetitive information as the query answer.

Example of 20Q with Open Query Set and Free-text Query Answers

```
## System Instruction
You are an expert on animals. Your goal is to predict the animal given the information you have gathered. The animal must be one of the
↪  following: 'giraffe', 'zebra', 'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros',
↪  'lion', 'tiger', 'blue whale', 'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'. Be as
↪  precise and as direct as possible. You may provide your reasoning first before making a prediction. End your response by making a guess
↪  and saying 'The animal is: X' (e.g. The animal is: dog) at the end of your reasoning, where X is your guess. Do not provide any
↪  additional information. Your response should all fit in a single paragraph. If you are not provided any information, make a random guess.

## Input Prompt
Here is the information you have gathered.
1. This animal primarily lives on land.
2. It is native to Europe and Asia, but was later introduced to the Americas and Africa.
3. This animal is a carnivore by natural instinct and primarily feeds on a diet rich in meat.

Given the information you have gathered, make an intermediate SINGLE prediction of what you think the animal is. First make your guess in the
↪  format 'The ansmal is: X' (e.g. The animal is: dog), where X is your guess, then provide your reasoning. Do not provide any additional
↪  information. Your response should all fit in a single paragraph. Make sure your prediction is one of the classes: 'giraffe', 'zebra',
↪  'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros', 'lion', 'tiger', 'blue whale',
↪  'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'.

## Response
The animal is: dalmatian. The animal primarily lives on land, which matches with the fact that it is native to Europe and Asia, but was later
↪  introduced to the Americas and Africa, indicating a terrestrial habitat. This animal is a carnivore by natural instinct and primarily
↪  feeds on a diet rich in meat, which aligns with the characteristics of a dalmatian.
```

Example of 20Q with Open Query Set and Binary Query Answers

```
## System Instruction
You are an expert on animals. Your goal is to predict the animal given the information you have gathered. The animal must be one of the
↪  following: 'giraffe', 'zebra', 'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros',
↪  'lion', 'tiger', 'blue whale', 'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'. Be as
↪  precise and as direct as possible. You may provide your reasoning first before making a prediction. End your response by making a guess
↪  and saying 'The animal is: X' (e.g. The animal is: dog) at the end of your reasoning, where X is your guess. Do not provide any
↪  additional information. Your response should all fit in a single paragraph. If you are not provided any information, make a random guess.

## Input Prompt
Here is the information you have gathered.
1. Does the animal primarily eat plants? No.
2. Does the animal primarily live on land? Yes.
3. Does the animal have four legs? Yes.
4. Does the animal have a large body size (over 400 kg)? Yes.
5. Does the animal live in a herd? No.
6. Is the animal known to be aggressive? Yes.
7. Does the animal have a long neck? No.
8. Does the animal have sharp claws? Yes.
9. Is the animal a carnivore? Yes.
10. Is the animal a mammal? Yes.

Given the information you have gathered, make an intermediate SINGLE prediction of what you think the animal is. First make your guess in the
↪  format 'The ansmal is: X' (e.g. The animal is: dog), where X is your guess, then provide your reasoning. Do not provide any additional
↪  information. Your response should all fit in a single paragraph. Make sure your prediction is one of the classes: 'giraffe', 'zebra',
↪  'elephant', 'killer whale', 'dalmatian', 'polar bear', 'giant panda', 'hippopotamus', 'rhinoceros', 'lion', 'tiger', 'blue whale',
↪  'walrus', 'grizzly bear','siamese cat', 'cow', 'german shepherd', 'gorilla', 'dolphin','moose'.

## Response
The animal is: lion. I based my prediction on the fact that the animal is a carnivore (9), lives on land (2), has a large body size (4), and
↪  is known to be aggressive (6). These characteristics are commonly associated with big cats, and the lion is one of the most aggressive
↪  and large land carnivorous mammals.
```

Some text that involves characters α, β, μ are unable to render in our example, so we replace with with spelled out words, i.e. $\alpha \rightarrow$ "alpha".

Example of Diagnosis in Internal Medicine

```
## Input
You are a medical doctor answering real-world medical entrance exam questions. Based on your understanding of basic and clinical science,
↪  medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy, answer the following multiple choice
↪  question. Select one correct answer from A to D. Base your answer on the current and standard practices referenced in medical guidelines.
↪  Repond in the following format:

{{"answer": "A/B/C/D", "explanation": "YOUR EXPLANATION HERE"}}

Answer the multiple choice based on the context.

Initial Info: A 54-year-old woman presents to the emergency department due to recent onset of a high fever, accompanied by severe headache
↪  and neck stiffness.

Conversation Log between doctor and patient:
Here is the information you have gathered.
Patient's Fact: The patient's current blood pressure reading is 112/71 mmHg.
Patient's Fact: She is currently on immunosuppressants after a kidney transplant 6 months ago. Her current medication regimen includes
↪  immunosuppressants.
Patient's Fact: The patient is 54 years old.
Patient's Fact: The patient has a history of rapidly progressive glomerulonephritis. She also underwent a kidney transplant 6 months ago and
↪  is currently on immunosuppressants.
Patient's Fact: The patient's current body temperature is 102.2°F (39°C).
Patient's Fact: The patient's neck pain is severe with neck flexion and reciprocal knee flexion. Facts 9 and 9 are the answer to this
↪  question.
Patient's Fact: The patient cannot answer the question about the typical triad of symptoms associated with meningitis.
Patient's Fact: Her respirations are 22/min.


Question: Which of the following characteristics is true of the most likely cause of this patient's disorder?
Options:
    A - Anaerobic spore forming rods
    B - Catalase-negative
    C - Coagulase-positive
    D - Tumbling motility in broth

Please select the most appropriate answer (A/B/C/D).

Medical conditions are complex, so you should seek to understand their situations across many features. First, consider which medical
↪  specialty is this patient's case; then, consider a list of necessary features a doctor would need to make the right medical judgment.
↪  Think step by step, reason about the patient information, the inquiry, and the options.

## Response
{"answer": "D", "explanation": "The patient's presentation of high fever, severe headache, neck stiffness, and neck pain with neck flexion
↪  and reciprocal knee flexion (positive Brudzinski's sign) suggests meningitis, which is an inflammation of the protective membranes
↪  covering the brain and spinal cord, known as the meninges. Given the patient's history of immunosuppression following a kidney
↪  transplant, she is at increased risk for opportunistic infections. The patient's inability to answer questions about the typical triad of
↪  symptoms associated with meningitis (headache, fever, and stiff neck) further supports this diagnosis.

The patient's age and immunosuppressed state make her susceptible to various pathogens, but the most likely cause of meningitis in this
↪  context is Listeria monocytogenes. Listeria monocytogenes is a gram-positive, catalase-negative bacterium that exhibits tumbling motility
↪  in broth. This characteristic is unique to Listeria species, making option D the correct answer.

Listeria monocytogenes is a common cause of meningitis in immunocompromised individuals, particularly those with a history of organ
↪  transplantation. It is also a cause of meningitis in pregnant women, the elderly, and individuals with certain underlying medical
↪  conditions. The other options do not accurately describe the typical characteristics of Listeria monocytogenes. Anaerobic spore-forming
↪  rods (option A) are more commonly associated with Clostridium species. Coagulase-positive (option C) is characteristic of Staphylococcus
↪  aureus, and while some anaerobic bacteria may exhibit catalase-negative (option B), this is not a specific characteristic of the most
↪  likely cause of this patient's disorder."}
```

K Examples of Sampled Queries

Here we provide some brief examples of queries sampled in the 20Q open query set setting. Here the number of queries to sample $m = 3$.

Example of Diagnosis in Pediatrics

```
## Input
You are a medical doctor answering real-world medical entrance exam questions. Based on your understanding of basic and clinical science,
↪  medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy, answer the following multiple choice
↪  question. Select one correct answer from A to D. Base your answer on the current and standard practices referenced in medical guidelines.
↪  Repond in the following format:

{{"answer": "A/B/C/D", "explanation": "YOUR EXPLANATION HERE"}}

Answer the multiple choice based on the context.

Initial Info: An 8-year-old girl of Asian descent is brought to the physician because of fatigue.

Conversation Log between doctor and patient:
Here is the information you have gathered.
Patient's Fact: The patient's eye color and appearance is that she has pale conjunctivae.
Patient's Fact: The patient's mean corpuscular volume is 74 um3.
Patient's Fact: The chief complaint in this case is fatigue. She is not able to keep up with the rest of her classmates in gym class because
↪  she tires easily.
Patient's Fact: The patient tires easily during physical activities because she is not able to keep up with her classmates in gym class and
↪  is brought to the physician because of fatigue. This is also due to her pale conjunctivae and small, pale red blood cells.
Patient's Fact: She has a serum ferritin level of 77 ng/mL.
Patient's Fact: The patient's current Hemoglobin level is 11.0 g/dL.
Patient's Fact: She has normal hemoglobin electrophoresis. Hemoglobin electrophoresis is normal.


Question: Which of the following best describes the pathogenesis of the disease process in this patient?"
Options:
    A - Decreased production of beta-globin proteins
    B - Amino acid substitution in the beta-globin protein
    C - Cis deletion of alpha-globin genes
    D - Acquired inhibition of heme synthesis

Please select the most appropriate answer (A/B/C/D).

Medical conditions are complex, so you should seek to understand their situations across many features. First, consider which medical
↪  specialty is this patient's case; then, consider a list of necessary features a doctor would need to make the right medical judgment.
↪  Think step by step, reason about the patient information, the inquiry, and the options.


## Response
{"answer": "C", "explanation": "This patient's presentation of fatigue, pale conjunctivae, and small, pale red blood cells (microcytosis) is
↪  suggestive of anemia. The patient's mean corpuscular volume (MCV) of 74 um3 indicates a microcytic anemia. The normal hemoglobin
↪  electrophoresis rules out hemoglobinopathies such as sickle cell disease or thalassemia major. The serum ferritin level of 77 ng/mL is
↪  elevated, which suggests iron overload. However, the patient's hemoglobin level is low, which is not consistent with iron overload. The
↪  most likely diagnosis is beta-thalassemia trait, which is a mild form of thalassemia. Beta-thalassemia trait is caused by a cis deletion
↪  of one of the beta-globin genes, leading to reduced production of the beta-globin protein. This results in a mild anemia, which is
↪  consistent with the patient's presentation. The patient's fatigue and pale conjunctivae are due to the anemia, and the small, pale red
↪  blood cells are consistent with microcytic anemia. The normal hemoglobin electrophoresis and elevated ferritin level support this
↪  diagnosis. Therefore, the correct answer is C - Cis deletion of alpha-globin genes is incorrect, as the patient has beta-thalassemia
↪  trait, not alpha-thalassemia. The correct answer is actually Cis deletion of beta-globin genes but the options provided do not have
↪  beta-globin, so the closest option is C."}
```

Example of Diagnosis in Neurology

```
## Input
You are a medical doctor answering real-world medical entrance exam questions. Based on your understanding of basic and clinical science,
↪  medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy, answer the following multiple choice
↪  question. Select one correct answer from A to D. Base your answer on the current and standard practices referenced in medical guidelines.
↪  Repond in the following format:

{{"answer": "A/B/C/D", "explanation": "YOUR EXPLANATION HERE"}}

Answer the multiple choice based on the context.

Initial Info: A 25-year-old man presents to the ER after a street fight.

Conversation Log between doctor and patient:
Here is the information you have gathered.
Patient's Fact: He has been rapidly decompensating since several hours after the initial hit. Since then, he has been rapidly decompensating.
Patient's Fact: The patient cannot answer the question about the exact cause of the altered mental status and head pain. However, the patient
↪  is experiencing head pain and altered mental status. The atomic factual statements that answer the question partially are: - He is
↪  grabbing his head in pain.- His mental status is altered.
Patient's Fact: The patient's right pupil is dilated. The information comes from fact #7.

Question: If this patient's symptoms are allowed to progress without intervention, which of the following would be expected to occur?
Options:
    A - Right sided hemiparesis
    B - Down and in gaze of the right pupil
    C - Right sided homonymous hemianopsia.
    D - Right eye with an inability to abduct

Please select the most appropriate answer (A/B/C/D).

Medical conditions are complex, so you should seek to understand their situations across many features. First, consider which medical
↪  specialty is this patient’s case; then, consider a list of necessary features a doctor would need to make the right medical judgment.
↪  Think step by step, reason about the patient information, the inquiry, and the options.

## Response
{"answer": "A", "explanation": "Based on the patient's symptoms of rapidly decompensating altered mental status and head pain, along with a
↪  dilated pupil on the right side, the most likely diagnosis is a mass lesion in the brain, such as a hemorrhage or an ischemic stroke. The
↪  dilated pupil is a sign of third cranial nerve (oculomotor nerve) compression, which is often seen in patients with a space-occupying
↪  lesion in the brain.

If this patient's symptoms are allowed to progress without intervention, it is likely that the mass lesion will continue to increase in size
↪  and cause further brain damage. This could lead to a range of complications, including right-sided hemiparesis, which is weakness on the
↪  opposite side of the body from the lesion.

Right-sided hemiparesis is a common consequence of a left-sided brain lesion, as the motor tracts in the brain cross over to the opposite
↪  side of the body. The other options are not directly related to the expected outcome of a space-occupying lesion in the brain. Down and
↪  in gaze of the right pupil would be more consistent with a third cranial nerve palsy, but the patient's symptoms are not consistent with
↪  this diagnosis. Right-sided homonymous hemianopsia would be more consistent with a lesion in the visual pathway, and an inability to
↪  abduct the right eye would be more consistent with a sixth cranial nerve palsy, which is not suggested by the patient's symptoms."}
```

Example of Sampled Queries at Each Iteration

```
[
    "Does the animal have four legs?",
    "Does the animal live on land?",
    "Is the animal a mammal?"
],
[
    "Does the animal have spots or stripes?",
    "Does the animal primarily live on land?",
    "Does the animal have a long neck?"
],
[
    "Does the animal primarily eat plants?",
    "Does the animal live in a group?",
    "Does the animal have a distinctive coat pattern?"
],
[
    "Does the animal have a long neck?",
    "Does the animal primarily eat bamboo?",
    "Does the animal give birth to a single cub at a time?"
],
[
    "Does the animal have a large body size?",
    "Does the animal primarily eat bamboo?",
    "Does the animal have a distinctive coat pattern?"
],
[
    "Does the animal primarily consume bamboo?",
    "Does the animal have retractable claws?",
    "Does the animal have a distinctive pattern of black and white stripes rather than spots?"
],
[
    "Does the animal primarily feed on plants?",
    "Does the animal have a herbivorous diet?",
    "Is the animal native to Asia?"
],
[
    "Does the animal primarily feed on bamboo?",
    "Is the animal a carnivore?",
    "Does the animal have a large body size compared to other land animals?"
],
[
    "Does the animal primarily walk on all fours?",
    "Does the animal have a long gestation period?",
    "Is the animal a member of the Ursidae family?"
],
[
    "Does the animal have a long snout?",
    "Is the animal native to Africa?",
    "Does the animal have a large body size compared to its neck?"
],
[
    "Does the animal primarily live on land?",
    "Is the animal a carnivore?",
    "Does the animal have a large body size compared to other animals on the list?"
]
```