ADAPTIVE ELICITATION OF LATENT INFORMATION USING NATURAL LANGUAGE

Jimmy Wang* jw4209@columbia.edu
Thomas Zollo* tpz2105@columbia.edu
Richard Zemel zemel@cs.columbia.edu
Hongseok Namkoong namkoong@gsb.columbia.edu

ABSTRACT

Eliciting information to reduce uncertainty on a latent entity is a critical skill in many application domains, e.g., assessing individual student learning outcomes, diagnosing underlying diseases, or learning user preferences. Though natural language is a powerful medium for this purpose, large language models (LLMs) and existing fine-tuning algorithms lack mechanisms for strategically gathering information to refine their own understanding of the latent entity. We propose an adaptive elicitation framework that *actively* reduces uncertainty on the latent entity by simulating counterfactual responses. Since probabilistic modeling of an abstract latent entity is difficult, we validate and finetune LLM-based uncertainty quantification methods using perplexity over masked future observations produced by the latent entity. Our framework enables the development of sophisticated information-gathering strategies, and we demonstrate its versatility through experiments on dynamic opinion polling and adaptive student assessment.

1 INTRODUCTION

The performance of many valuables services and systems depends on the ability to efficiently elicit information and reduce uncertainty about a new environment or problem instance. For example, before an optimal lesson plan can be prepared for a particular student, information must first be gathered about their underlying skills and abilities. Similarly, a patient's health status must be quickly assessed upon intake, while an online service seeking retention aims to gain a fast understanding of a new customer's preferences.

Notably, in these (and many other) cases, the object of interest is *latent*, meaning it cannot be directly measured or observed but can only be queried indirectly. This makes gathering information particularly challenging, as it requires carefully designed strategies to infer the latent entity's characteristics through indirect signals. To achieve efficiency, these strategies must be *adaptive*, dynamically tailoring subsequent queries based on the information gained so far. In the context of student assessment, an adaptive approach might start with a broad math question covering multiple skills. If the student demonstrates strength in algebra, the system would follow up with more challenging algebra questions to determine the limits of their proficiency. Conversely, if the student struggles with geometry, the system would present easier geometry questions to pinpoint the exact concepts they have yet to master. By progressively refining its queries in this way, the system efficiently maps out the student's knowledge boundaries, gaining a clearer picture of their individual skill profile (see Figure 1).

As natural language is a particularly powerful and flexible medium for eliciting such latent information, one might assume that modern large language models (LLMs) (Brown et al., 2020; Bai et al., 2022; DeepSeek-AI et al., 2025) could be helpful in such dynamic information-gathering efforts. To do so would require the language model to quantify epistemic uncertainty, refine it given additional information, and/or act to reduce uncertainty in an optimal way. However, LLMs and existing fine-tuning algorithms often treat uncertainty passively, and lack mechanisms for strategically gathering information to refine their own understanding of the latent entity.

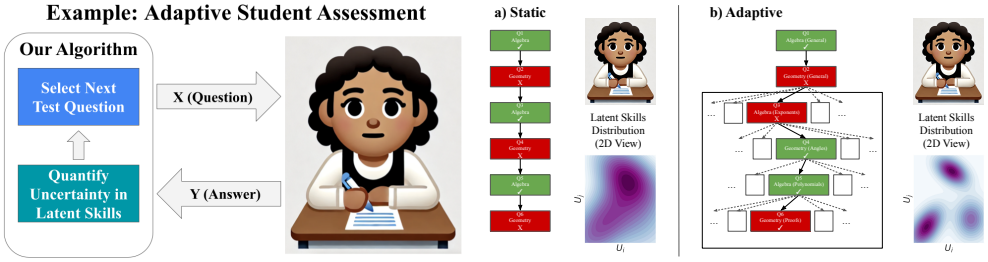

Figure 1: **Left:** Our algorithm can adaptively elicit information about a latent entity via natural language interaction. For example, in assessing a new student, the system may ask questions in areas where the student's abilities are not yet known, to maximize the information gained from each question and efficiently reduce uncertainty about the student's individual skill profile. **Right:** An example of how adaptive elicitation can improve over static strategies in student assessment. Each square represents a question asked to the student, and is marked green (answered correctly) or red (answered incorrectly). Here, a static strategy **(a)** would only expose that the student is strong in algebra and weak in geometry. An adaptive strategy **(b)**, on the other hand, will find the limits of their knowledge in both, thus reducing more uncertainty about their true latent skills U. Though we visualize uncertainty via distributions over latent features, these objects are high-dimensional and ill-defined in nature and cannot be easily modeled.

To overcome this challenge, we introduce an *adaptive elicitation framework* that uses natural language to *actively* reduce uncertainty on the latent entity by simulating counterfactual responses. Rather than modeling a latent entity probabilistically (Blei et al., 2003; Salakhutdinov & Mnih, 2007)—a difficult and often intractable step—we validate and finetune LLM-based uncertainty quantification methods through *perplexity over masked future observations* the latent entity might produce (Ye et al., 2024; Fong et al., 2023). By aligning the LLM perplexity objective with the goal of predicting future observations from a previously unseen latent entity, our approach enables the model to identify epistemic uncertainty and facilitate sophisticated information-gathering strategies as it updates its understanding of the latent. In doing so, we enable a wide range of exciting and impactful applications, e.g., constructing a dynamic diagnostic questionnaire that maximizes the information gained about a patient's health or generating a personalized set of test questions that yield the most insight into a student's learning needs.

In the remainder of this paper, we introduce our framework for latent uncertainty reduction using natural language and demonstrate its effectiveness across several key applications. Our work contributes a key conceptual and algorithmic insight to the accelerating field of LLMs: by obviating the need for directly modeling the latent and instead employing a predictive view of uncertainty, we enable the development of adaptive information-gathering strategies. Through experiments on tasks such as dynamic opinion polling and adaptive student assessments, we illustrate the versatility and significant potential of our framework to enable more efficient and targeted information elicitation in critical domains and applications. Overall, we aim to lay the foundation for future research into rigorous uncertainty quantification and adaptive decision-making in LLMs, highlighting the promise of active, context-aware strategies in advancing real-world AI systems.[1]

2 Adaptive Elicitation Framework

In this section, we present an approach to uncertainty quantification and adaptive question selection in scenarios where the latent entity cannot be directly modeled. Our method: (1) *Meta-learns* a predictive language model from historical question–answer data (2) Uses this model to *quantify uncertainty* about future or unobserved answers via simulation (3) *Adapts question selection* in real time to reduce uncertainty about the latent entity.

Throughout the rest of this paper, we adopt a *predictive view* of uncertainty: rather than specifying a direct prior or complete model of the latent entity, we focus on how well the model can predict future observations of that entity, interpreting predictive uncertainty as a strong proxy for epistemic uncertainty. This approach echoes classical missing-data and Bayesian predictive viewpoints (Ru-

[1]Because of space constraints, we defer a thorough discussion of other related work to Appendix Section A.

bin, 1976; Lindley, 1965; Hill, 1968; Dawid, 1984) as well as modern treatments of exchangeable models in deep learning (Fong et al., 2023; Ye et al., 2024). We emphasize, however, that fully rigorous exchangeability assumptions are often unrealistic in natural-language settings. Instead, our aim is a practical heuristic that remains robust for complex, human-generated text.

2.1 PRELIMINARIES

Latent Entities and Observations. We consider an unobservable latent entity $U \in \mathcal{U}$ (e.g., a student's skill profile or a patient's health status). We query U by posing a question $X \in \mathcal{X}$ (in natural language) and observing an answer $Y \in \mathcal{Y}$, $Y \sim \mathbb{P}(\cdot|\text{ Question } X, \text{ Latent } U)$. Our two primary goals are to: (1) *Quantify* our uncertainty about U based on observed question–answer pairs. (2) *Reduce* that uncertainty by adaptively choosing which questions X to ask next.

Following classical views that treat latent variables as unobserved data (Rubin, 1976; Lindley, 1965), we interpret "knowing U" as being able to predict all future answers Y with high accuracy. Other approaches might model U directly (e.g., by assigning a probability distribution over a structured latent space) (Blei et al., 2003; Salakhutdinov & Mnih, 2007); yet specifying such models for complex human-generated responses can be both restrictive and infeasible. In contrast, our predictive focus aligns naturally with the goal of adaptive elicitation in a flexible domain (i.e., open-ended natural language) where strict parametric assumptions about U are difficult to justify and often incomplete.

Predictive Uncertainty. Rather than directly modeling U with a parametric prior, we focus on the model's ability to predict future answers $Y_{t+1:\infty}$ given observed data $\mathcal{H}_t = \{(X_i, Y_i)\}_{i=1}^t$: $\mathbb{P}(Y_{t+1:\infty}|\mathcal{H}_t) = \mathbb{P}(\text{Future answers} \mid \text{Current info})$. Any uncertainty in these predictions serves as a practical measure of our epistemic uncertainty about U. In formal Bayesian terms, this often corresponds to a posterior-predictive distribution $\int P(Y_{t+1:\infty} \mid U)\, \pi(U \mid \mathcal{H}_t)\, dU$ (Ye et al., 2024; Zhang et al., 2024). Our framework aims to approximate this distribution directly, without requiring an explicit prior $\pi(\cdot)$.

2.2 META-LEARNING A PREDICTIVE MODEL

We assume access to historical data from a collection of latent entities $\mathcal{U}_{\text{train}}$. Each entity $U \in \mathcal{U}_{\text{train}}$ is associated with a sequence of question–answer pairs $\{(X_{1:N}^{(U)}, Y_{1:N}^{(U)})\}$. Our first step is to meta-train an autoregressive language model on this historical data consisting of sequences of questions and answers from various latent entities $\mathcal{D}_{\text{train}} := \{X_{1:N}^{(U)}, Y_{1:N}^{(U)} : U \in \mathcal{U}_{\text{train}}\}$. In the student assessment example, $\mathcal{D}_{\text{train}}$ may be a historical dataset of past students, each with an associated sets of test questions and answers. For simplicity, we assume that each sequence is of length N, but our framework is agnostic to differing sequence lengths.

Objective Define a sequence of previous observations $\mathcal{H}_t := \{X_{1:t}, Y_{1:t}\}$. We train our autoregressive language model, parameterized by $\theta \in \Theta$, to output one-step probabilities over future answers conditioned on previous observations (i.e., question/answer pairs) $p_\theta(Y_{t+1}|\ \mathcal{H}_t, X_{t+1} = x)$, inducing a joint distribution over future outcomes

$$p_\theta(Y_{t+1:\infty}|\mathcal{H}_t, X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, ...) = \prod_{s=t+1}^{\infty} p_\theta(Y_s|\mathcal{H}_{s-1}, X_s = x_s). \tag{1}$$

The training objective for our model is then to optimize the joint log likelihood/marginal likelihood of the observed sequence within the historical dataset

$$\max_{\theta\in\Theta} \left\{ \frac{1}{|\mathcal{U}_{\text{train}}|} \sum_{U\in\mathcal{U}} \sum_{t=1}^{T} \log p_\theta(Y_t^U|\mathcal{H}_{t-1}, X_t^U = x_t) \right\}.$$

Training. For training, we process each sequence of questions and answers $\{X_{1:N}^{(U)}, Y_{1:N}^{(U)}\}$ corresponding to a latent entity U by sequentially arranging them into one long natural language string $(X_1^{(U)}, Y_1^{(U)}, X_2^{(U)}, Y_2^{(U)}, ...)$. Then we optimize a language model to predict each answer Y_t conditioned on the current question X_t and previous observations $\mathcal{H}_{t-1}$. To do so, we apply a gradient mask that masks out tokens which do not correspond to any Y_i. We use stochastic gradient descent procedures to optimize the training loss.

Exchangeability. While human-generated language data may generally fail to be exchangeable, a condition necessary for valid predictive uncertainty estimates in classical treatments of Bayesian predictive inference (de Finetti, 1937), our question-answer setting is conducive to this exchangeability condition. That is, the order in which someone answers questions should not affect their answers. To enforce this condition during training, we *randomly permute* the order of pairs within each entity's sequence during training. This helps ensure the model remains relatively agnostic to a particular question ordering and can better generalize to a new entity presented in arbitrary query orders.

2.3 Uncertainty Quantification by Simulated Futures

Given a new entity for which we have observed $\mathcal{H}_t = \{(X_i, Y_i)\}_{i=1}^{t}$, we interpret our predictive model's distribution over $\{Y_{t+1:\infty}\}$ as our *uncertainty* about the entity. Concretely:

$$p_\theta(Y_{t+1:\infty} \mid \mathcal{H}_t) = \prod_{s=t+1}^{\infty} p_\theta\Big(Y_s \mid X_s, \mathcal{H}_{s-1}\Big),$$

where each one-step conditional probability is given by our meta-learned model.

Simulation for Inference. To quantify or visualize uncertainty, one can draw samples of "simulated futures": $Y_{t+1:\infty}^{(k)} \sim p_\theta(\cdot \mid \mathcal{H}_t)$ for $k = 1, \ldots, K$. Each sample represents one plausible realization of the entity's future answers. We can then treat the variability in these simulated futures as a measure of epistemic uncertainty.

2.4 Adaptive Question Selection

Finally, we aim to reduce our predictive uncertainty by posing additional questions that maximally *inform* the model about the new entity. In practice, this corresponds to an *active learning* or *optimal design* paradigm where, at each step t, we choose X_{t+1} (e.g., a test question) to maximize the expected information gain about some target information Z (see Figure 1, right). In our case, Z will be the set of answers to future questions $Y_{t+1:\infty}^{U}$

Information Gain. Let Z denote information of interest (e.g., "What answers would the student select to this new set of questions?"). We measure current uncertainty by $H(Z \mid \mathcal{H}_t)$ (entropy). After observing (X_{t+1}, Y_{t+1}), the posterior entropy is $H(Z \mid \mathcal{H}_t, X_{t+1}, Y_{t+1})$. The one-step expected *information gain* is:

$$\mathrm{EIG}_{t:t+1}(Z; X_{t+1}) \;=\; H(Z \mid \mathcal{H}_t) \;-\; \mathbb{E}\big[H\big(Z \mid \mathcal{H}_t, X_{t+1}, Y_{t+1}\big)\big].$$

To choose the optimal question, ideally we would solve

$$\arg\max_{X_{t+1}} \; \mathrm{EIG}_{t:\infty}(Z; X_{t+1}) = H(Z \mid \mathcal{H}_t) \;-\; \mathbb{E}_{Y_{t+1:\infty} \sim p_\theta}\big[H\big(Z \mid \mathcal{H}_t, X_{t+1}, Y_{t+1:\infty}\big)\big] \tag{2}$$

In practice, it is intractable to simulate $Y_{t+1:\infty}$ and to calculate the expected information gain as it is combinatorial in the number of steps. Instead, we introduce two procedures that show strong practical performance while having feasible computational cost.

Greedy Selection. A simple heuristic is to first enumerate the candidate questions $x \in x_1, ..., x_k$. Then for each x_j, calculate the one-step expected information gain $\mathrm{EIG}_{t:t+1}(Z; X_{t+1})$. Finally, choose the x_j that maximizes this quantity. Although greedy, this often performs well in practice and is computationally simpler than globally optimal planning. We prove the theoretical validity of this procedure in Section B, where Theorem B.4 bounds the performance gap between a full combinatorial planning approach and the greedy selection procedure.

Lookahead / Monte Carlo Planning. To account for multi-step effects (e.g., a question that might not immediately reduce much uncertainty but paves the way for more informative follow-ups) and to better approximate the intractable quantityin Equation 2, we can apply standard *Monte Carlo Tree Search* (MCTS) techniques from reinforcement learning. With MCTS, we sample entire future question–answer sequences using the meta-learned model p_θ up to depth d to estimate the cumulative information gain. Though more expensive computationally, this can find better long-horizon query strategies.

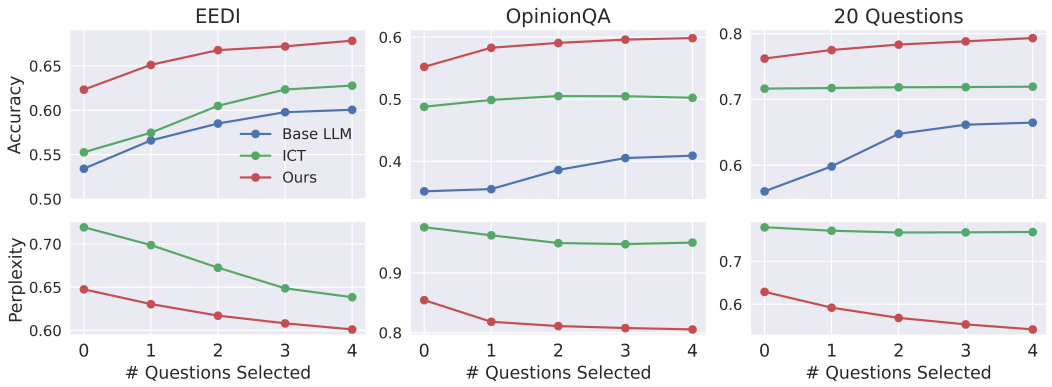

Figure 2: Accuracy (top) and perplexity (bottom) of our adaptive elicitation framework compared to baseline methods across three datasets: OpinionQA, EEDI student assessment, and 20 Questions. The x-axis represents the number of questions selected. Our method works best to gather information and accurately characterize the latent in each case.

3 Experiment Details

Our experiments considers 3 applications: opinion polling, student assessment, and the 20 questions game. In each case, the goal is to adaptively select questions that are as informative as possible with respect to the answers to another set of target questions. Questions are selected one at a time, and each is added to the LLM context before selecting the next.

For each experiment, we have a dataset containing a collection of latent entities U, which are associated with questions X^U and answers Y^U. To train our meta-learned model, we split each dataset by groups of latent entities into train, validation, and test sets. We first meta-learn our model p_θ on questions and answers corresponding to latent entities in the training dataset. Then, we evaluate the model's ability to quantify and reduce uncertainty on questions and answers corresponding to entities in the test set. More details on datasets, training, baselines, and evaluation are below.

3.1 Datasets

OpinionQA Santurkar et al. (2023) Originally created to evaluate the alignment of LLM opinions to those of 60 US demographic groups, this dataset contains 1498 multiple choice political questions answered by a diverse collection of survey respondents. These questions target various political issues ranging from abortion to automation. Here, each question corresponds to a question X, the multiple choice answer corresponds to the observable feedback Y, and the survey respondent's latent political preference corresponds to the unobservable U.

EEDI Tutoring Dataset Wang et al. (2020) Eedi is an online educational and tutoring platform that serves millions of students around the globe. This dataset includes a collection of 938 math questions focusing on various areas such as algebra, number theory, and geometry, as well as individual responses from many students. Each question is a multiple choice question with four answers that includes a visual diagram as well as associated text. The student's true mathematical ability U generates the student's answer Y to the math question X.

Twenty Questions We create a synthetic twenty questions dataset, where the objective is to ask relevant questions in order to determine the underlying object or certain traits. We first retrieve a collection of objects from the THINGS Hebart et al. (2019) dataset. Then we use Claude 3.5 Sonnet to generate potential questions, and answer each question given each object. We end up with 800 objects and 1200 questions and answers for each. Here, each object corresponds to the latent entity U, which generates the answers Y to each question X. While Claude 3.5 Sonnet may generate wrong answers, we emphasize the correctness of the answers is not important, but rather that our model learns the underlying data generating process according to which Claude produces the answers (i.e., a conditional language model).

3.2 META-TRAINING DETAILS

We first split the training datasets by entity into train, validation, and test with a 70%, 15%, 15% split. To meta-train our model, we initialize a pre-trained Llama-3.1-8B model in FP16 precision and use LoRA Hu et al. (2021) to finetune our model with parameters $\alpha = 24$, rank$= 8$, and dropout$= 0.1$. We initialize the AdamW Loshchilov & Hutter (2019) optimizer with learning rate of 0.0001 and $\beta = (0.9, 0.95)$, weight decay of 0.1, and we use a linear warmup for the learning rate after which we use a cosine scheduler. We train our model for $10,000$ epochs with a batch size of 4 and block size of 1024, after which we take the checkpoint with the lowest validation loss.

3.3 BASELINES

Base LLM First, we consider a simple baseline. For an LLM we use Llama-3.1-8B, from which our meta-trained model is initialized; question selection is performed randomly.

In-Context Tuning (ICT) Next, we consider a typical in-context learning (ICL) baseline. First, we meta-train the model via In-Context-Tuning (Chen et al., 2022), where the objective is to predict the label for a query example given some number of in-context support examples. Then, questions are selected based on embedding similarity to the target questions that we aim to answer (Liu et al., 2021). We use the same model and parameters as 3.2, and we use Alibaba-NLP/gte-large-en-v1.5 as our embedding model.

3.4 EVALUATION

To evaluate how well each method can ask targeted questions to reduce uncertainty about the latent entity, we randomly select 10,000 entities. For each entity, we randomly select a pool of N questions from which the methods can sequentially choose questions to ask, and randomly select K held-out target questions. The objective is to sequentially choose optimal questions from the N questions to reduce the most uncertainty about the K held-out targets for each entity. In our experiments, we choose $N = 20$ and $K = 5$, but we include ablations that vary these quantities in 4.4. We evaluate the performance on the target questions with four metrics. (1) Accuracy, (2) Perplexity, (3) Expected Calibration Error, and (4) Brier Score.

4 RESULTS AND DISCUSSION

In this section, we empirically study the following questions: (1) Can our framework be used to adaptively select questions to reduce uncertainty and elicit information about the latent? (2) Do we generate reasonable posterior probability updates and reduce uncertainty as more information is gathered? (3) When is this adaptive procedure particularly helpful? (4) How important is our training procedure for producing actionable uncertainty quantification? Throughout, we connect these findings to the paper's broader motivation: eliciting information efficiently in real-world scenarios.

4.1 OVERALL GAINS FROM ADAPTIVE ELICITATION

Overall results for our method and 2 baselines across all 3 datasets are shown in Figure 2. The top row of plots record accuracy on the target questions, while the bottom row record perplexity (or negative log-likelihood loss). The Base LLM is omitted on bottom for ease of visualization. In both figures, the X-axis records the number of questions that have been selected so far.

Across all 3 datasets and both metrics, our algorithm most effectively characterizes the latent by predicting the answers to target questions (we show similar results for Brier Score in Figure 6). Further, our algorithm consistently improves its characterization as more information is gathered, whereas gathering more questions based on embedding distance does not always help. Overall, our adaptive elicitation framework proves effective in gathering information and reducing uncertainty across 3 diverse domains.

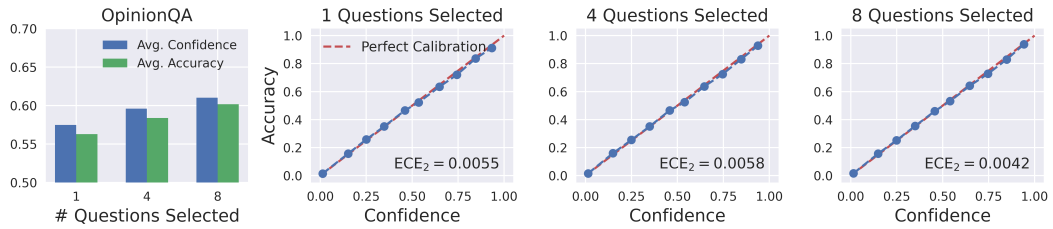

Figure 3: Reliability diagrams comparing confidence and accuracy after different numbers of selected questions (and observed answers). Our model maintains well-calibrated uncertainty estimates, increasing both confidence and accuracy as more questions are asked.

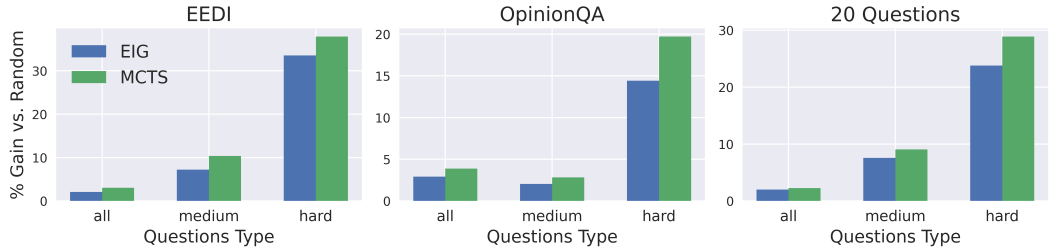

Figure 4: Relative accuracy gain from adaptive question selection (EIG and MCTS) over random selection for different subsets of target questions: all, medium difficulty (answered correctly by $< 50\%$ of the population), and hard (answered correctly by $< 30\%$). Adaptivity provides the greatest benefits when identifying rare latent traits, demonstrating when strategic question selection is most advantageous.

4.2 Uncertainty Quantification

A cornerstone of our approach is using *predictive* perplexity as an indicator of uncertainty; this makes sense only if our model's probabilities reflect genuine confidence about unseen data. To assess this, we examine calibration, roughly the extend to which the model's confidence reflects its accuracy.

For each dataset, we plot reliability diagrams (Guo et al., 2017) of confidence vs. accuracy, where perfect calibration lies on the $y = x$ line, and record Expected Calibration Error (ECE). Both the reliability diagram and ECE are produced by separating predictions into 10 bins by confidence, and comparing the average confidence and accuracy for each bin. Results are shown after 1, 4, and 8 questions are selected, and the far right figure plots overall average confidence and accuracy for each setting.

Results for OpinionQA are shown in Figure 3, while EEDI and 20 Questions are shown in Appendix Figures 7 and 8. For all 3 datasets, we observe that the predicted probabilities lie close to the diagonal of perfect calibration—our model's confidence aligns well with actual accuracy. As more questions are observed, the model's average confidence (and accuracy) both go up, confirming that uncertainty diminishes in an intuitive way. In the motivating student-assessment scenario, this means that by asking just a few strategically chosen questions, the model not only improves its predictions but also becomes *more certain* in them. For a high-stakes application such as medical diagnostics or skill placement exams, it is crucial to know when a model has enough data to be sure in its predictions, versus when it is still uncertain; these calibration results confirm our framework performs well in this sense.

4.3 When is Adaptivity Most Helpful?

Having established that our adaptive question selection method is generally effective at quantifying uncertainty and eliciting information about some latent, we next examine *when* such a procedure is most helpful. In particular, we hypothesize that adaptive strategies are most important in characterizing features of the latent which are relatively rare in the population. As a concrete example, while many students may have overlapping weaknesses (e.g., many get the same test question wrong), it can be harder to learn that a particular student is struggling in an area where other students generally

do not. An adaptive strategy could help by selecting a test question that most find easy but this student may answer incorrectly.

To investigate this hypothesis, we specify two different subgroups of questions as targets by running our evaluation where all target questions for each entity have probability less than either 50% ("medium") or 30% ("hard") across the population. We use our meta-trained model with random, EIG, and MCTS question selection, and record results after N questions have been selected. Results are shown in Figure 4. For each question subgroup (as well as all questions from the previous experiment), we record on the y-axis the relative accuracy gain from using EIG or MCTS, compared to selecting questions randomly.

First, we notice that the more advanced MCTS planning strategy outperforms EIG in all cases, and both always outperform random. This means that given a good model for uncertainty quantification, we can improve our results by spending more compute, indicating good scaling behavior in our algorithm. Next, we observe trends across different subgroups of questions. In all 3 example applications, adaptivity and planning have a massive impact on the ability to answer hard questions compared to random question selection. For EEDI and 20 Questions the percent gain over random with EIG or MCTS is more than 10x higher for hard questions than for all questions; for OpinionQA, it is 5x higher. We thus have strong evidence that our adaptive information elicitation strategy is most important when characterizing the latent features which are most atypical with respect to the population. If the latent entity exhibits atypical behavior (a student struggles with a concept that most find easy, or an opinion respondent holds a rare viewpoint), an adaptive method can target precisely those concepts that discriminate such cases. Conversely, random or fixed questionnaires fail to unearth those nuances in a limited query budget.

4.4 Planning Ablation

Our results in Figure 2 show the effectiveness of our full adaptive elicitation framework of meta-training and adaptive information gathering via planning. Those in Figure 4 establish the significant performance gains from applying planning algorithms on top of our model, as compared to random question selection. With our final experiment, we aim to understand the importance of our meta-training procedure. To do so, we compare the results of applying a planning strategy atop our model, to those produced when applying the same strategy to the ICT and base LLM models. We use the 20 Questions dataset, and the same splits of all, medium, and hard questions as the previous experiment. For each question type and each of 3 underlying models, (Base, ICT, and ours), we record the accuracy on target questions after selecting 3 questions with either random selection or the EIG strategy. To measure what is gained from planning, we record the ratio of target question accuracy with planning to that with random selection (a value above 1 indicates some accuracy gain from planning).

Results are shown in Appendix Figure 5. First, we see that planning performs poorly using the Base LLM, reducing accuracy almost 15% on hard questions compared to random question selection. The ICT model performance is largely unchanged by planning, across all 3 question types. On the other hand, our model's performance is greatly improved when question selection is guided by planning, highlighting that our training procedure is essential to enable such strategic information gathering with LLMs.

4.5 Other Ablations

We first ablate the number of targets and questions to choose from. Our experiments were run with the models being able to select from 20 questions in order to accurately predict 5 targets. In Table 2 in Appendix E, we find that our method gains more accuracy as the question bank becomes larger. In Table 1, we find that performance stays roughly the same as the number of target questions changes. Finally, we study the effect of the base model for our meta-training procedure. We test GPT2, Llama-3.2-1B, and Llama-3.1-8B, and find in Table 3 that performance increases as the model is larger.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003. ISSN 1532-4435.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7f48f642a0ddb10272b5c31057f0663-Paper.pdf.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning, 2022. URL https://arxiv.org/abs/2110.07814.

A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:278–292, 1984.

Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying

Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. Active prompting with chain-of-thought for large language models, 2023.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2024. Google Deepmind, Massachusetts Institute of Technology, UC Berkeley.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models, 2024.

Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society, Series B*, 2023.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.330. URL http://dx.doi.org/10.18653/v1/2021.findings-emnlp.330.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. Bayesian preference elicitation with language models, 2024. URL https://arxiv.org/abs/2403.05534.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24, 10 2019. doi: 10.1371/journal.pone.0223792. URL https://doi.org/10.1371/journal.pone.0223792.

Bruce M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322):677–691, 1968.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models, 2024.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in neural information processing systems*, volume 34, pp. 1273–1286, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Andreas Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. In-context decision-making from supervised pretraining. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. URL https://openreview.net/forum?id=WIzyLD6j6E.

Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, and Igor Mordatch. Multi-game decision transformers. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=yN4Wv17ss3.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024b.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4160–4173, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.328. URL https://aclanthology.org/2022.findings-acl.328.

Dennis V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, 1965.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3?, 2021. URL https://arxiv.org/abs/2101.06804.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021.

Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks, 2023.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2335739.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, pp. 1257–1264, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm, 2024.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Naimeng Ye, Hanming Yang, Andrew Siah, and Hongseok Namkoong. Exchangeable sequence models can naturally quantify uncertainty over latent concepts. *arXiv preprint arXiv:2408.03307*, 2024. Available at https://arxiv.org/abs/2408.03307.

Kelly W. Zhang, Tiffany (Tianhui) Cai, Hongseok Namkoong, and Daniel Russo. Posterior sampling via autoregressive generation. *arXiv preprint arXiv:2405.19466*, 2024. Published on arXiv, 8 October 2024.

A Related Work

Reinforcement Learning with Sequence Models A number of works propose to train or use powerful pre-trained models in order to solve complex reinforcement learning (RL) tasks, focusing on how these models can make decisions using vast amounts of offline data Janner et al. (2021); Yang et al. (2023); Chen et al. (2021); Du et al. (2024); Lee et al. (2022). Another line of works show that using meta-learned sequence models to predict the next action can approximate standard bandit algorithms Lin et al. (2024a); Lee et al. (2023); Zhang et al. (2024). We extend these ideas to natural language while focusing on how our meta-learned model can quantify uncertainty to make a decision.

Uncertainty Quantification over Natural Language. There has been a recent class of works focusing on developing uncertainty measures to augment the reliability of model responses. Kuhn et al. (2023); Lin et al. (2024b); Malinin & Gales (2021); Duan et al. (2024) focus on predictive entropy measures with off-the-shelf language models, while other approaches focus on self-consistency in the generation space Lin et al. (2022); Si et al. (2023); Kadavath et al. (2022); Diao et al. (2023). Another class of works focuses instead on detecting *epistemic* uncertainty from *aleatoric* uncertainty in model outputs Yadkori et al. (2024); Osband et al. (2023); Hou et al. (2024); Glushkova et al. (2021). Our meta-learning uncertainty quantification framework is complimentary to these works, as these measures are designed to be applied on top of pre-trained foundation models.

Planning and Information Gathering with LLMs Our work is related to Uncertainty of Thoughts (UoT) Hu et al. (2024) and OPEN Handa et al. (2024). While these methods build elicitation procedures on top of off-the-shelf language models, we use a meta-learning procedure in order to accurately quantify uncertainty over new environments. Other works introduce methods to enhance general reasoning or planning capabilities by using natural language reasoning steps Wei et al. (2022); Wang et al. (2022); Yao et al. (2023).

B THEORETICAL VALIDITY

B.1 GREEDY EIG SELECTION

In this section, we show the theoretical validity of using a greedy procedure to select actions with the highest expected information gain. Consider a set of observable feedback $Z = Y_{t+1:T}$ corresponding to a set of target designs $X_{t+1:T} = x_{t+1:T}$ that we would like to minimize our uncertainty over. We first formally define the Expected Information Gain (EIG).

Definition B.1 (Expected Information Gain)**.** Let $\mathfrak{X}_t = (x_1, x_2, ..., x_t) \subseteq \mathcal{X}$. Given a distribution p_θ and targets $Y_{t:T}$, the EIG is defined as

$$\text{EIG}(Z = Y_{t:T}, \mathfrak{X}) = \mathbb{E}_{Y \sim p_\theta}[H(Y_{t+1:T}) - H(Y_{t+1:T}|\mathfrak{X})], \tag{3}$$

where $H(Y_{t+1:T}|\mathfrak{X}) = H(Y_{t+1:T}|X_{1:t} = x_{1:t}, Y_{1:t})$. Note that $Y_{1:t}$ are random variables and not deterministic quantities.

Ultimately, our goal is to choose a set of designs $X_{1:t} = x_{1:t}$ that approximates the set of designs that provide the most amount of information possible.

We first begin with the assumption

Assumption B.2. $Y_{1:\infty}$ are conditionally, identically distributed.

As a consequence, $Y_{1:\infty}$ are exchangeable, which is a reasonable assumption given that someone's answers is likely not to depend on the order in which they are presented.

Define $\mathfrak{X}_t^*$ and $\mathfrak{X}_{p_\theta}^*$ to be the optimal set of designs X under the true distribution q and the model p_θ such that

$$\mathfrak{X}^* := \underset{\{x_1, x_2, ..., x_t\} \in \mathcal{X}}{\arg\max} \; \mathbb{E}_{Y \sim q}[\log q(Y_{t:T}|X_{1:t} = x_{1:t}, Y_{1:t})].$$

$$\mathfrak{X}_{p_\theta}^* := \underset{\{x_1, x_2, ..., x_t\} \in \mathcal{X}}{\arg\max} \; \mathbb{E}_{Y \sim q}[\log p_\theta(Y_{t:T}|X_{1:t} = x_{1:t}, Y_{1:t})].$$

Next, define $\mathfrak{X}_{\text{greedy}}$ to be the set of designs chosen through the greedy, information gain procedure $\mathfrak{X}_{\text{greedy}} := (x_1, x_2, ..., x_t)$, where each x_i is chosen as

$$x_i = \underset{x \in \mathcal{X}}{\arg\max} \; \text{EIG}(Y_{t:T}, \{x_1, x_2, ...x_{i-1}\} \cup x)$$

In order to show that the greedy procedure is able to perform close to the optimal solution, we rely on the following assumption.

Assumption B.3 (Submodularity)**.** Consider any $\mathfrak{X}', \mathfrak{X} \subseteq \mathcal{X}$ where $\mathfrak{X}' \subseteq \mathfrak{X}$. Consider any target Z and a random variable $X : \Omega \to \mathcal{X}$. Then for an $\varepsilon > 0$,

$$H_{p_\theta}(Z|X \cup \mathfrak{X}) - H_{p_\theta}(Z|X) + \varepsilon \geq H_{p_\theta}(Z|X \cup \mathfrak{X}') - H_{p_\theta}(Z|X).$$

The submodularity assumption simply says that the entropy over future observations decreases as the model conditions on more information. We assume an approximate version of submodularity to account for training instabilities or inaccuracies in the model p_θ. While full submodularity would imply that conditioning on more context would strictly reduce entropy, we take a more relaxed stance.

We quantify the information gained throughout the selection process in terms of the optimal perplexity under the true environment.

Theorem B.4. *Under the greedy information gain selection procedure, the KL divergence between the meta-learned model conditioned on the information gathered and the optimal distribution is bounded as*

$$\text{KL}\left(q(Y_{t:T} \mid \mathfrak{X}_t^*)||p_\theta(Y_{t:T} \mid \mathfrak{X}_{\textit{greedy}})\right) \leq \text{KL}(q(\cdot|\mathfrak{X}_t^*)||p_\theta(\cdot|\mathfrak{X}_{p_\theta}^*)) + \log(|\mathcal{Y}|(T-t))\sqrt{\frac{1}{2}\text{KL}(q(\cdot|\mathfrak{X}_{p_\theta}^*)||p_\theta(\cdot|\mathfrak{X}_{p_\theta}^*))}$$
$$+ \frac{1}{e}(\text{EIG}(\cdot|\mathfrak{X}_{p_\theta}^*) + t\varepsilon.$$

This bound can be broken up into three intuitive terms. The first term $\mathrm{KL}(q(\cdot|\mathfrak{X}_t^*)||p_\theta(\cdot|\mathfrak{X}_{p_\theta}^*))$ is equivalent to the difference between the true environment and using the optimal selection procedure under the meta-learned model in the true environment. The next term $\log(|\mathcal{Y}|(T-t))\sqrt{\frac{1}{2}\mathrm{KL}(q(\cdot|\mathfrak{X}_{p_\theta}^*)||p_\theta(\cdot|\mathfrak{X}_{p_\theta}^*))}$ quantifies the difference between choosing under the true environment distribution, and the distribution generated by the meta-learned model. Finally, the third term $\frac{1}{e}(\mathrm{EIG}(\cdot|\mathfrak{X}_{p_\theta}^*) + t\varepsilon$ quantifies the difference between using the greedy procedure versus the full optimal selection procedure.

B.2 Proof of Theorem B.4

Then, we can decompose

$$\mathbb{E}_{Y\sim q}[\log q(Y_{t:T} \mid \mathfrak{X}_t^*) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{\text{greedy}})] = \mathbb{E}_{Y\sim q}[\log q(Y_{t:T} \mid \mathfrak{X}_t^*) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] \tag{4}$$

$$+ \mathbb{E}_{Y\sim q}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] - \mathbb{E}_{Y\sim p_\theta}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] \tag{5}$$

$$+ \mathbb{E}_{Y\sim p_\theta}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{\text{greedy}})] \tag{6}$$

First, $\mathbb{E}_{Y\sim q}[\log q(Y_{t:T} \mid \mathfrak{X}_t^*) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] = \mathrm{KL}(q(\cdot|\mathfrak{X}_t^*)||p_\theta(\cdot|\mathfrak{X}_{p_\theta}^*)$.

Bounding the second term equation 5 To characterize the second term, we establish the following facts For any $S \in \mathcal{S}$,

$$0 \le H(S) \le \log|\mathcal{S}|.$$

[Russo & Van Roy (2016)] For any distributions P and Q such that P is absolutely continuous with respect to Q, any random variable $Z : \Omega \to \mathcal{Z}$, and any $f : \mathcal{Z} \to \mathbb{R}$ such that $f_\infty \le 1$,

$$\mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] \le \sqrt{\frac{1}{2}\mathrm{KL}(P||Q)}$$

First, by making use of Fact equation B.2, we can see that because $Y_{t:T}$ is exchangeable, then both $H_q(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)$ and $H_{p_\theta}(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)$ are bounded by $\log(|\mathcal{Y}|(T-t))$. Then we can bound the term equation 5 as

$$\begin{aligned}
&\mathbb{E}_{Y\sim q}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] - \mathbb{E}_{Y\sim p_\theta}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] \\
&= \mathbb{E}_{Y_{1:t}\sim q}[H_q(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] - \mathbb{E}_{Y_{1:t}\sim p_\theta}[H_{p_\theta}(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*)] \\
&\le \log(|\mathcal{Y}|(T-t))\sqrt{\frac{1}{2}\mathrm{KL}(q(\cdot|\mathfrak{X}_{p_\theta}^*)||p_\theta(\cdot|\mathfrak{X}_{p_\theta}^*))}.
\end{aligned}$$

Bounding the third term equation 6 It follows that

$$\begin{aligned}
&\mathbb{E}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{\text{greedy}})] \\
=&\mathbb{E}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*) - \log p_\theta(Y_{t:T})] \\
+&\mathbb{E}[\log p_\theta(Y_{t:T}) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{\text{greedy}})] \\
=&\mathrm{EIG}(Y_{t:T}; \mathfrak{X}_{p_\theta}^*) - \mathrm{EIG}(Y_{t:T}; \mathfrak{X}_{\text{greedy}})
\end{aligned}$$

The ϵ submodularity of entropy directly implies the ϵ submodularity of the Expected Information Gain. Directly using results from Krause et al. (2008), we have

$$\mathbb{E}[\log p_\theta(Y_{t:T} \mid \mathfrak{X}_{p_\theta}^*) - \log p_\theta(Y_{t:T} \mid \mathfrak{X}_{\text{greedy}})] \le \frac{1}{e}\mathrm{EIG}(\cdot|\mathfrak{X}_{p_\theta}^*) + t\varepsilon,$$

showing the result.

C Experiment Details

For ease of reproducibility, our code will be made public upon release of this paper.

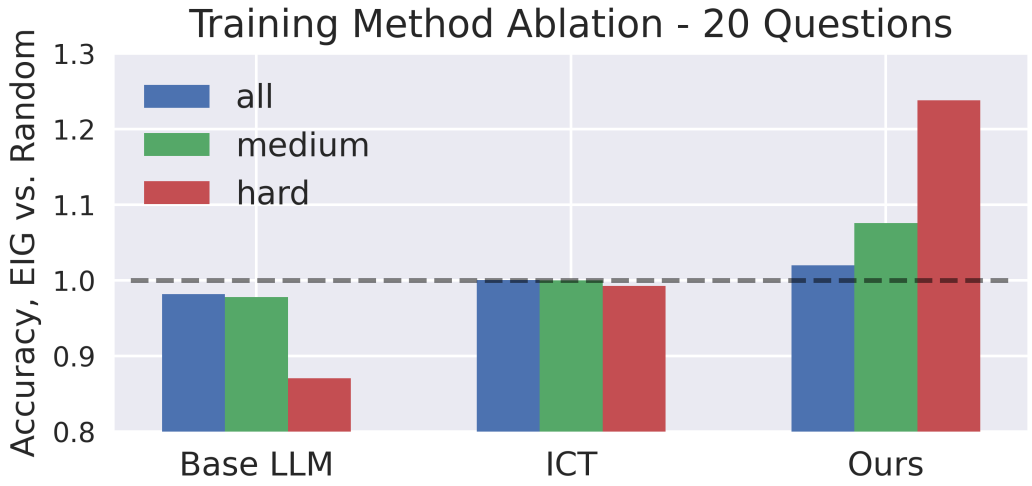

Figure 5: Comparison of performance gains from planning (EIG-based selection) using different models: a base LLM, an in-context tuning (ICT) model, and our meta-trained model. The y-axis represents the ratio of accuracy with planning versus random selection; our model benefits the most from planning, while the base and ICT models show accuracy loss or no improvement.

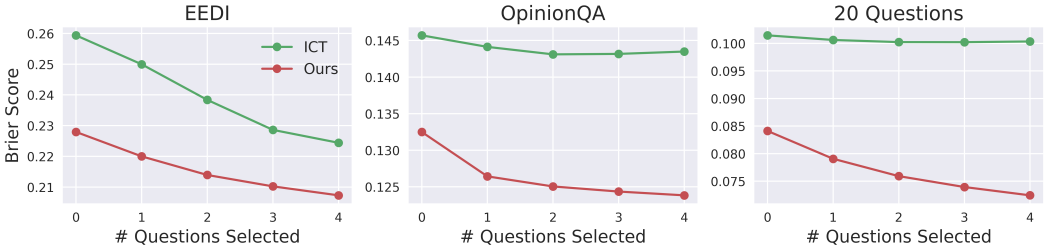

Figure 6: Brier score results in our overall setting across 3 datasets.

D Experiment Results

Here we include additional experiment results. Figure 6 shows results for the overall experiments with the Brier Score metric. Figure 7 shows calibration results for EEDI, and Figure 8 shows calibration results for 20 Questions.

E Ablations

Table 1: Ablating Number of Targets on EEDI Conditioned on 4 Questions

Accuracy	**1**	**5**	**10**	**20**
Base	0.6042	0.6005	0.6066	0.5987
Ictx	0.6269	0.6278	0.6295	0.6255
Ours	**0.6759**	**0.6784**	**0.6871**	**0.6832**

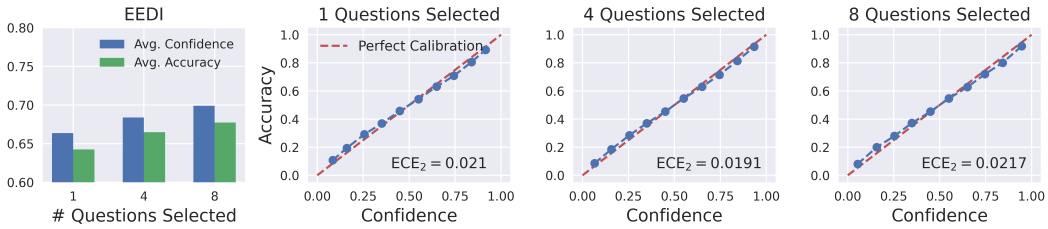

Figure 7: Calibration results with EEDI.

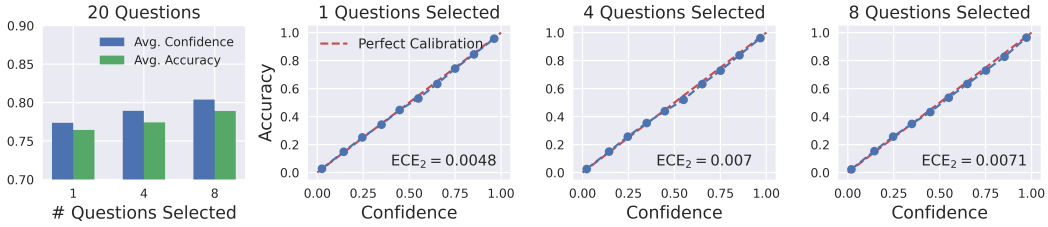

Figure 8: Calibration results with 20 Questions.

Table 2: Ablating Number of Possible Questions on OpinionQA Conditioned on 4 Questions

Accuracy	**10**	**15**	**20**	**25**
Base	0.4030	0.4042	0.4089	0.4093
Ictx	0.4988	0.4993	0.5023	0.5009
Ours	0.**5933**	**0.5953**	**0.5987**	**0.6068**

Table 3: Ablating Base Model: Twentyq performance conditioned on 4 questions

	GPT2	**Llama-3.2-1B**	**Llama-3.1-8B**
Accuracy	0.5201	0.6131	0.7382