Demystifying Uncertainty in LLMs: Active Calibration between Concepts and Human Evaluations

Anonymous ACL submission

Abstract

Hallucinations arise when large language models (LLMs) guess rather than acknowledge their underlying uncertainty. Existing static strategies for mitigating hallucinations have been only partially successful, largely because they do not explicitly model the information gain from interacting with the external environment. Researchers need a general method to proactively steer users toward informative clarifications, thereby unlocking the model's effective capacity under underspecified inputs. We model the uncertainty of LLMs in interactive settings and uncover the mechanism of active calibration between model concepts and human evaluations, improving reliability. We show that calibration error in LLMs density estimation admits a non-vanishing lower bound under non-interactive learning, while interaction empirically reduces it. We further characterize that calibration error identifies informative queries and that calibration can be accelerated by shifting query distributions from imbalanced to balanced regimes. Guided by these insights, we propose a calibration-driven Interactive Learning Strategy (ILS) that selects clarification queries by optimizing calibration error, providing both theoretical guarantees and empirical gains for reliability. Code and data are available at https://anonymous.4open.science/r/DemystifyingUncertainty/.

1 Introduction

Human interaction with the world is fundamentally guided by *beliefs*. To make effective decisions, these beliefs must align with appropriate "world models" (Ha and Schmidhuber, 2018; LeCun, 2022; Wong et al., 2023), which capture the latent structure of reality. Since our knowledge of the external world is rarely complete, such beliefs are probabilistic, encoding uncertainty about the environment. As the world changes or new evidence emerges, these beliefs must be updated accordingly to remain consistent with reality. Recent advances in large language models (LLMs) (Brown et al., 2020; Li et al., 2023; Guo et al., 2025) invite a natural extension of this perspective:

> What forms of uncertainty underlie the interaction between LLMs and humans?

LLMs have demonstrated remarkable capabilities across a variety of tasks, including text generation, question answering (Ouyang, 2022; Wei et al., 2022; Havlík et al., 2025), code generation (Zhao et al., 2025), information retrieval (Dai et al., 2024), and tool use (Qin et al., 2023; Wang et al., 2025). By analogy, LLMs interact with humans through outputs that implicitly encode beliefs about knowledge and reasoning. However, LLMs often hallucinate (Manakul et al., 2023; Ji et al., 2023), producing fluent but incorrect statements even in the latest models (OpenAI, 2025), limiting adoption in safety-critical domains. To mitigate these risks and encourage LLMs to acknowledge uncertainty rather than guess, uncertainty quantification (UQ) has been explored as a key strategy for assessing model reliability. Yet, existing UQ methods for LLMs primarily rely on numerical estimates (Malinin and Gales, 2020; Ling et al., 2024), instead of leveraging the language space to actively communicate and elicit clarifications about the sources of uncertainty, even though such language-based uncertainty articulation is typically more intuitive for humans (Zimmer, 1983; Windschitl and Wells, 1996). Besides, some studies propose leveraging an LLMs' own ability to evaluate the uncertainty of its responses without relying on external knowledge (Kadavath et al., 2022; Miao et al., 2023). However, these approaches suffer from *overconfidence* issues due to the model's inherent bias to trust its own outputs (Xiong et al., 2023; Li et al., 2024). These limitations highlight the need for new frameworks that model **how LLMs represent and**

update uncertainty in interaction with humans.

To this end, we decompose uncertainty into aleatoric and epistemic components (Section 2), and show that they are tightly entangled in LLMs rather than cleanly separable. We introduce calibration error as a key indicator of this entanglement by measuring the model's density-estimation mismatch. In particular, a larger calibration error suggests a greater discrepancy between these uncertainty components. This perspective motivates an interactive learning paradigm in which LLMs move beyond static numerical evaluation and instead adopt an active calibration framework, actively querying humans to reduce calibration error and align their probabilistic concepts with human evaluations. Crucially, enabling LLMs to ask humans—rather than merely answer them—captures an equally important dimension of intelligence, highlighting the model's capacity to guide its own knowledge acquisition.

In this work, we advance the theory of uncertainty in LLM–human interaction. We show that calibration error cannot vanish in the non-interactive setting (Theorem 1 in Section 3.1), admitting a monofact-rate–controlled lower bound even under zero aleatoric uncertainty, and validate this relation via controlled n-gram simulations (Section 3.2). Empirically, multi-turn interaction consistently reduces calibration error (Section 4). We further show that calibration error pinpoints informative queries and that calibration can be accelerated by shifting the query distribution from imbalanced to balanced (Theorem 2 in Section 5.1). Guided by these insights, we propose ILS (Section 5.2, with a case study in Figure 1), which selects clarifications by maximizing calibration error and adaptively reweights the query distribution, with theoretical guarantees and consistent gains over prompt-only baselines on free-form generation benchmarks.

2 Backgrounds

We consider the LLM π_θ that maps an input $\mathbf{x} \in \mathcal{X}$ to a distribution over sequences $\mathcal{Y} := \mathcal{V}^L$, where $\mathcal{V}$ denotes the vocabulary of size V and L represents the max length of windows. Given the question $\mathbf{x}$, a concrete prediction $\mathbf{y} = [y_1, \ldots, y_L] \in \mathcal{Y}$ is then sampled from this distribution. The sequential nature of language modeling requires an autoregressive factorization: $\pi_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{l=1}^{L} \pi_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{<l})$, where $\mathbf{y}_{<l}$ denotes the prefix of y_l, which makes the analysis more complicated. To solve this problem, we can merge this factorization into the definition of the backbone h_θ, similar to the formulation in Ren and Sutherland (2024). Letting $\boldsymbol{\chi}$ denote the concatenation of $\mathbf{x}$ and $\mathbf{y}$, the prediction of all tokens in $\mathbf{y}$ can then be written as:

$$\mathbf{z} = h_\theta(\boldsymbol{\chi}), \qquad \pi_\theta(\mathbf{y} \mid \mathbf{x}) = \text{Softmax\_column}(\mathbf{z}).$$

Here $\mathbf{z}$ is a $V \times L$ matrix where the l-th column corresponds to the logits used to predict the l-th token. Our h_θ, even though it receives the entire sequence $\boldsymbol{\chi}$ as input, will force the model not to refer to the future tokens $\mathbf{y}_{>l}$ when making predictions on the l-th token y_l, which is typically implemented via "causal masking" (Vaswani et al., 2017).

Probability measure. Let $(\mathcal{X}, \mathcal{T}_\mathcal{X})$ be a topological space. We write $\mathcal{M}^+(\mathcal{X})$ for the set of all Borel probability measures on $\mathcal{X}$. We assume that an input $\mathbf{x}$ is drawn from a query distribution $\mu \in \mathcal{M}^+(\mathcal{X})$. Formally, a parameterized LLM with parameters $\theta \in \Theta$ assigns, for each input $\mathbf{x} \in \mathcal{X}$, a probability distribution over the output space $\mathcal{Y}$, which is denoted by $\pi_\theta(\cdot \mid \mathbf{x}) \in \mathcal{M}^+(\mathcal{Y})$. Therefore, it can be equivalently viewed as an operator, i.e. a **concept**:

$$\pi_\theta : \mathcal{X} \to \mathcal{M}^+(\mathcal{Y}), \qquad \mathbf{x} \mapsto \pi_\theta(\cdot \mid \mathbf{x}).$$

with a concrete prediction $\mathbf{y} \sim \pi_\theta(\cdot \mid \mathbf{x})$.

The target distribution over the entire output space $\mathcal{Y}$ can be denoted as $\mathbf{p}(\cdot \mid \mathbf{x})$ where each element $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$ represents the probability of a specific sequence $\mathbf{y}$. A language model learns a parameterized distribution π_θ to approximate $\mathbf{p}$, which is a classical density estimation problem typically optimized via the auto-regressive loss $\mathcal{L}(\mathbf{x}; \theta) = \mathbb{E}_{\mathbf{y} \sim \mathbf{p}}[-\sum_{l=1}^{L} \log \pi_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{<l})]$. The entropy of the learned distribution π_θ can be formulated as $H_\mathbf{x}(\pi_\theta) = -\sum_{\mathbf{y} \in \mathcal{Y}} \pi_\theta(\mathbf{y} \mid \mathbf{x}) \log(\pi_\theta(\mathbf{y} \mid \mathbf{x}))$. If the target distribution $\mathbf{p}$ is known, the KL divergence between π_θ and $\mathbf{p}$ can also be computed as $\mathrm{KL}_\mathbf{x}(\mathbf{p} \mid \pi_\theta) = \sum_{\mathbf{y} \in \mathcal{Y}} \mathbf{p}(\mathbf{y} \mid \mathbf{x}) \log[\mathbf{p}(\mathbf{y} \mid \mathbf{x}) / \pi_\theta(\mathbf{y} \mid \mathbf{x})]$.

Aleatoric uncertainty. Intuitively, aleatoric uncertainty reflects the inherent randomness in the data-generating process, arising from the true conditional distribution $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$, such that even an ideal model with unlimited capacity cannot eliminate it. If a sample $\mathbf{x}$ admits a unique correct answer $\mathbf{y}^*$, then $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$ is degenerate and the aleatoric uncertainty is zero. Conversely, if the

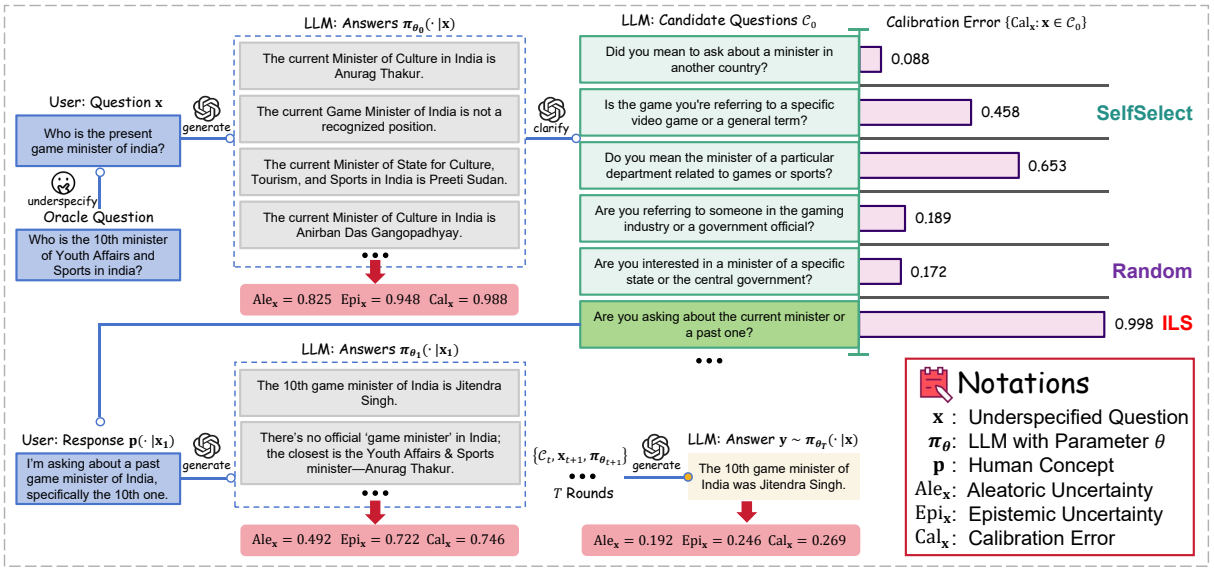

Figure 1: **Overview of Interactive Learning Strategy (ILS).** Given an underspecified user query $\mathbf{x}$, the LLM produces an initial answer distribution $\pi_{\theta_0}(\cdot \mid \mathbf{x})$ and we estimate $\mathrm{Ale}_\mathbf{x}$, $\mathrm{Epi}_\mathbf{x}$, and $\mathrm{Cal}_\mathbf{x}$. ILS then generates a candidate clarification set $\mathcal{C}_0$ and selects the most informative query by maximizing the calibration error, contrasting prompt-only baselines (Random, SelfSelect). After T turns, ILS progressively enhances calibration.

task itself is intrinsically ambiguous (e.g., multiple valid translations or open-ended QA), $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$ has high entropy, yielding irreducible aleatoric uncertainty. In practice, since $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$ is inaccessible, we approximate this quantity by combining ground-truth signals with the model's output distribution obtained from sampling (Kadavath et al., 2022). Inspired by Kalai et al. (2025), we partition answer space $\mathcal{Y}$ into valid answers $\mathcal{A}_\mathbf{x} = \mathrm{supp}_\mathbf{x}(\mathbf{p}) = \{\mathbf{y} \in \mathcal{Y} \mid \mathbf{p}(\mathbf{y} \mid \mathbf{x}) > 0\}$ and errors $\mathcal{E}_\mathbf{x} = \mathcal{Y} \setminus \mathcal{A}_\mathbf{x}$ that are inconsistent with $\mathbf{p}$. The aleatoric uncertainty of the LLM π_θ w.r.t. query $\mathbf{x}$ is denoted by:

$$\mathrm{Ale}_\mathbf{x} := \pi_\theta(\mathcal{E}_\mathbf{x} \mid \mathbf{x}) = \Pr_{\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})}[\mathbf{y} \in \mathcal{E}_\mathbf{x}].$$

Moreover, the aleatoric uncertainty of π_θ w.r.t. query distribution μ is given by $\mathrm{Ale} := \mathbb{E}_{\mathbf{x} \sim \mu} \mathrm{Ale}_\mathbf{x}$. Here, a single query inherently carries uncertainty, yet our analysis elevates this notion to the distributional level by modeling uncertainty over the query space with μ. In experiments we report the uncertainty of π_θ by averaging across 100 queries.

Epistemic uncertainty. Unlike aleatoric uncertainty, which arises from irreducible randomness in the data distribution, epistemic uncertainty reflects the limitations of the model class in approximating the true preference distribution. Even when sufficient training signal is available, a model with restricted capacity may fail to capture the underlying structure. To formalize this, we define epistemic uncertainty w.r.t. query $\mathbf{x}$ as the entropy induced by the distribution π_θ, i.e.

$$\mathrm{Epi}_\mathbf{x} := H_\mathbf{x}(\pi_\theta) = -\sum_{\mathbf{y} \in \mathcal{Y}} \pi_\theta(\mathbf{y} \mid \mathbf{x}) \log(\pi_\theta(\mathbf{y} \mid \mathbf{x})),$$

with further uncertainty estimation methods (e.g., confidence, margin score, perplexity, semantic entropy) detailed in Appendix B.2. Similarly, the epistemic uncertainty of π_θ w.r.t. query distribution μ is denoted by $\mathrm{Epi} := \mathbb{E}_{\mathbf{x} \sim \mu} \mathrm{Epi}_\mathbf{x}$.

Calibration Error. Calibration asks whether probabilities assigned by a language model match empirical frequencies. For example, in weather forecasting, if an LLM predicts a 70% chance of rain on a set of days, then approximately 70% of those days should indeed experience rain. More generally, an LLM is calibrated if, among responses to which it assigns confidence near $\pi_\theta(\mathbf{y} \mid \mathbf{x})$, the empirical frequency of correctness is also $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$. We define the calibration error as the distance between π_θ and $\mathbf{p}$, i.e.

$$\begin{aligned} \mathrm{Cal}(\pi_\theta, \mathbf{p}) &:= \mathbb{E}_{\mathbf{x} \sim \mu} \mathrm{Cal}_\mathbf{x}(\pi_\theta, \mathbf{p}) \\ &= \mathbb{E}_{\mathbf{x} \sim \mu} d\left(\pi_\theta(\cdot \mid \mathbf{x}), \mathbf{p}(\cdot \mid \mathbf{x})\right), \end{aligned}$$

where d may be instantiated as total variation distance or Jensen–Shannon divergence. We derive lower bounds for calibration error in the next section and show that the relation implied by Theorem 1 holds for both choices of d (Appendix B.1).

3 Calibration Error Cannot Vanish without Interaction

In this section, we show that calibration error admits a non-vanishing lower bound: without interaction, it cannot be eliminated, even under zero aleatoric uncertainty. The key is the monofact rate, inspired by Turing's "missing-mass" estimator (Good, 1953; Kalai et al., 2025). n-gram experiments show that high monofact rate induces aleatoric uncertainty, epistemic uncertainty and hence calibration error. Combining theoretical analysis and empirical validation, we highlight that actively asking questions is essential for LLMs to close the calibration gap.

3.1 Why Calibration Error Persists

Our lower bound for calibration error is based on the fraction of facts appearing just once in the training data:

Definition 1 (Monofact rate (Kalai et al., 2025))**.** *A query* $\mathbf{x} \in \mathcal{X}$ *is a monofact if it appears exactly once in the* N *training data* $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ *without abstention. Formally, denote the training count by* $c_N(\mathbf{x}) := \sum_{i=1}^{N} \mathbf{1}\{\mathbf{x}_i = \mathbf{x}\}$*, and define the monofact set* $\mathcal{S} := \{(\mathbf{x} \in \mathcal{X} : c_N(\mathbf{x}) = 1\}$*. Then, the monofact rate is denoted as the fraction of training monofacts:*

$$\mathrm{MFR} := \frac{|\mathcal{S}|}{N}.$$

The monofact rate gauges the remaining probability mass on unseen outcomes, exploiting the symmetry between zero and one occurrences. Following Turing's estimate, the probability of unseen events is approximated by the fraction of monofacts, which thus serves as a proxy for the missing portion of the distribution. Now we establish the lower bound of calibration error.

Theorem 1 (Non-vanishing Lower Bound for Calibration Error)**.** *With probability at least* $1 - \delta$ *over the training samples,*

$$\mathrm{Cal}(\pi_\theta, \mathbf{p}) \geq \mathrm{MFR} - \mathrm{Ale} - \frac{3e^{-m}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{N}},$$

where m *is a sparsity parameter. For realistic data size, the final two terms are negligible, leaving calibration error and uncertainty essentially governed by the monofact rate.*

The complete proof of Theorem 1 is provided in Appendix C. The core idea is that generating outputs with low aleatoric uncertainty—i.e., producing responses consistent with $\mathbf{p}$—is strictly harder than merely classifying whether an output belongs to $\mathbf{p}$. Consider questions of the form "Is this a valid language model output?" Generation implicitly requires answering this Yes/No query for every candidate response, so any failure to distinguish good from bad inevitably propagates to the generative setting. Consequently, if a model cannot reliably perform the classification task, it cannot be expected to generate only valid outputs. This reduction aligns our setting with computational learning theory (Kearns and Vazirani, 1994), providing a principled lens for analyzing error mechanisms in generative models.

Remark. Theorem 1 highlights a structural limitation of density estimation in LLMs: calibration error with respect to the target distribution $\mathbf{p}$ cannot vanish when portions of the distribution are absent from training data. Even with zero aleatoric uncertainty, the model inevitably hallucinates on facts outside the training support, since **it approximates a distribution rather than memorizing ground truth**. In pretraining, an LLM must approximate the entire language distribution, which inevitably includes arbitrary and patternless components that are hard to learn. Because the model cannot perfectly classify the validity of such facts, the inequality implies that generation necessarily incurs both uncertainty and calibration error. This further suggests that interactive mechanisms which reduce the monofact rate provide a principled means to lower uncertainty and calibration error in practice.

3.2 n-gram Study of MFR, Cal, Ale, and Epi

To validate the quantitative relationship among monofact rate, calibration error, aleatoric uncertainty, and epistemic uncertainty, we begin with classical n-gram models, which provide a controlled setting with full authority over data generation, model architecture, training, and evaluation. This allows us to test Theorem 1 before moving to large-scale LLMs. Let T denote the base fact set, consisting of unique, structured, comma-separated six-tuples—⟨*Actor*, *Costar*, *Movie*, *Director*, *Genre*, *Year*⟩—derived from the non-commercial IMDb corpus (IMDb, 2024).

Data Generating. Let $\mathbf{p}$ denote the distribution over six-tuples $\mathbf{x} = (x_1, \ldots, x_6)$. For any pair of tokens (x_i, x_{i+1}), the bigram model learns $\mathbf{p}(x_1, x_2) = \mathbf{p}(x_1)\mathbf{p}(x_2 \mid x_1)$, where $\mathbf{p}(x_1)$ is the

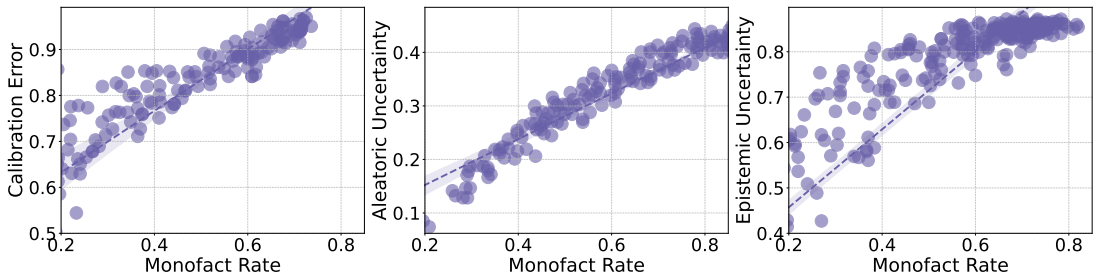

Figure 2: n-gram analysis of MFR, Cal, Ale, and Epi. Each dot corresponds to a sample of 5,000 facts, showing that Cal, Ale, and Epi all increase with the monofact rate.

marginal over first tokens and $\mathbf{p}(x_2 \mid x_1)$ the conditional over second tokens. Generating a new six-tuple proceeds sequentially:

$$\mathbf{p}(\mathbf{x}) = \mathbf{p}(x_1) \prod_{i=1}^{5} \mathbf{p}(x_{i+1} \mid x_i).$$

Concretely, we first draw $x_1 \sim \mathbf{p}(\cdot)$ and then sample $x_{i+1} \sim \mathbf{p}(\cdot \mid x_i)$ for each position $i = 1 : 5$. We use a schema-agnostic bigram generator that applies the same transition mechanism to all adjacent positions, without conditioning on slot identity in the six-tuple. Let S denote the training sample set, obtained by sampling with replacement from T according to $\mathbf{p}$. Each model is trained on $|S| = 5,000$ facts. Thus, S is a multiset that may contain duplicates, and empirical statistics such as the monofact rate are always computed with respect to S. Here, $\mathrm{Cal}(\pi_\theta, \mathbf{p})$ is computed using the total variation distance.

MFR **Controlling.** To construct the target distribution and vary the monofact rate, we replicate each fact in the base set T a random number of times, drawing counts from candidate distributions. Gaussian and Poisson choices concentrate mass near the mean and fail to produce diverse ranges of monofact rates. Instead, for each fact $\mathbf{x} \in T$ we sample a replication count:

$$r \sim f(r; \gamma, r_m) = \frac{\gamma r_m^\gamma}{r^\gamma}, \quad r \geq r_m,$$

according to a heavy-tailed Pareto distribution with shape γ and scale r_m, where $r_m = 1$ ensures every fact appears at least once. The replication count is rounded to an integer, giving multiplicities $\{c_i\}_{i=1}^{|T|}$. These multiplicities define a weighted multiset, which induces the normalized target distribution:

$$\mathbf{p}(\mathbf{x}_i) = \frac{c_i}{\sum_{j=1}^{|T|} c_j}, \quad i = 1, \ldots, |T|.$$

We then sample with replacement from this expanded pool to form S, ensuring i.i.d. sampling. Adjusting γ thus directly tunes the empirical monofact rate MFR.

Preliminary Insight. Our controlled simulations validate the relationships among MFR, Cal, Ale, and Epi predicted by Theorem 1. As shown in Figure 2, all three quantities increase with the monofact rate, suggesting that calibration error reflects misalignment in uncertainty decomposition and serves as a proxy for density-estimation mismatch in LLMs. Together, these results underscore the need for mechanisms that reduce the monofact rate.

Insight 1: Reducing the monofact rate provides an effective way to mitigate calibration error.

4 Active Calibration with Human

Given that calibration error admits a non-vanishing lower bound with respect to the monofact rate in the non-interactive setting, a natural question is whether interaction can eliminate this gap.

Benchmarks. To investigate, we employ AR-Bench (Zhou et al., 2025a), which simulates real-world information-seeking scenarios. We evaluate on Detective Cases (DC), simulating criminal investigations that require commonsense reasoning, and Situation Puzzles (SP), or "lateral thinking puzzles," which probe logical and divergent reasoning through Yes/No questions. Both tasks exhibit high aleatoric uncertainty, as the initial query lacks sufficient information and demands active interaction with an oracle. Following prompts (Appendix G.1), a Policy model generates short, verifiable questions ($\leq$20 words), while a Response model (oracle) returns concise answers—preferring Yes/No, otherwise brief facts—and outputs "It doesn't matter" to filter irrelevant queries. This design assesses

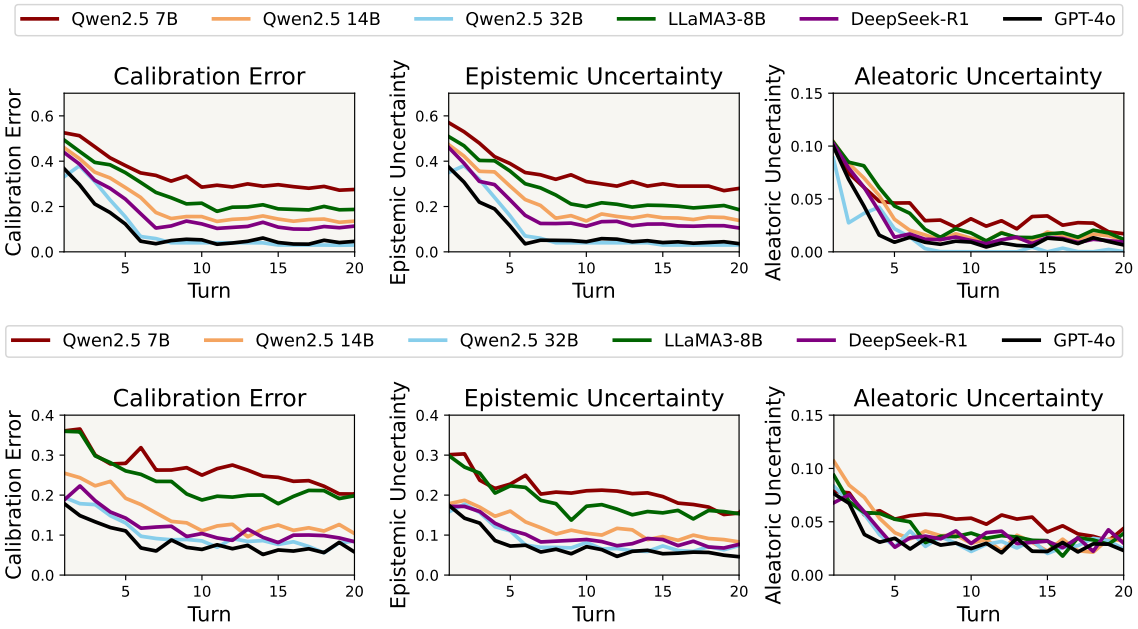

Figure 3: $\mathrm{Cal}(\pi_\theta, \mathbf{p})$ consistently decreases with interaction turns across LLMs in both DC (above) and SP (bottom).

not only final correctness but also the quality and informativeness of the model's queries.

Evaluation Metrics. Given a query $\mathbf{x}$, at each interaction round we sample $K = 8$ and estimate $\mathrm{Cal}_{\mathbf{x}}(\pi_\theta, \mathbf{p})$ via Jensen–Shannon divergence. We fix $K = 8$ for efficiency; in a sensitivity check, varying K from 5 to 20 does not affect model comparisons. Note that JS divergence ranges from $[0, 1]$, and the maximum number of interaction turns is capped at 20. To obtain the overall calibration error, we evaluate $\mathrm{Cal}_{\mathbf{x}}$ on a set of 100 queries and report the mean across all queries, i.e. Cal. For aleatoric uncertainty and epistemic uncertainty, we use error rate and entropy, as illustrated in Section 2. Following prior work (Kirchhof et al., 2025a), we use `Qwen2.5-7B` as the unified Response LLM, while the Policy LLM spans `LLaMA3-8B`, `DeepSeek-R1`, `Qwen2.5`, and `GPT-4o`.

Empirically, calibration error starts above 0.5, decreases steadily with interaction, and converges to around 0.2 after roughly forty turns, supporting the conclusion that interaction is able to reduce calibration error. As shown in Figure 3, calibration error converges before the 20-turn cap.

Observation 1: As the interaction progresses, the calibration error and uncertainty of all LLMs decrease and eventually converge.

Remark. Recent works (Kirchhof et al., 2025b) have highlighted that the two forms of uncertainty (aleatoric and epistemic) can be ambiguous in multi-turn interaction settings: the aleatoric uncertainty from prior turns can also be treated as epistemic uncertainty when furnished with information from subsequent turns. Our observations corroborate this finding: as the number of interaction turns increases, these two forms of uncertainty become mutually coupled, and the calibration error precisely captures this uncertainty-induced discrepancy. Beyond this, different models vary in both convergence speed and asymptotic error. For example, `Qwen2.5-7B` converges near zero calibration error by turn 15, whereas another model of the same family only begins to converge around turn 35 and stabilizes near 0.23. These differences motivate us to reconsider the **mechanism underlying interactive learning**: what makes multi-turn interaction **informative** and **efficient**?

Observation 2: LLMs that converge faster also achieve lower asymptotic calibration error.

5 LLM Learning to Query Human

Building on our theoretical and empirical results that validate the metric and establish the necessity of interaction, we now ask how to make interaction more informative and efficient—teaching the model how to query humans rather than teaching humans how to query the model.

5.1 What Makes Interactions Informative and Efficient

Informative Interactions. According to Observation 1, on the one hand, calibration error faithfully characterizes the uncertainty bias of a query by quantifying the mismatch between the model's predictive distribution and the ground truth. On the other hand, this very characterization provides actionable guidance during interaction: the query with the largest calibration error is the most informative to ask, as it promises the greatest reduction in uncertainty once resolved. That is, $\mathrm{Cal}_{\mathbf{x}}$ serves both as a **diagnostic metric** and as an **optimization target** for identifying and mitigating inputs that are likely to induce hallucinated generations.

Insight 2: Calibration error pinpoints the most informative query to ask during interaction.

Efficient Interactions. Extending Insight 1, we observe that the monofact rate is minimized by the imbalanced natural distribution $\mu(\mathbf{x})$, while for large N, it is minimized under a balanced $\mu(\mathbf{x})$. The detailed proof is provided in Appendix D.

Theorem 2. *Let μ be a distribution on $\mathcal{X}$ with $|\mathcal{X}| = K$, then the expected monofact rate is*

$$\mathbb{E}_{S\sim\mu^N}[\mathrm{MFR}] \;\approx\; \frac{1}{N}\sum_{i=1}^{K}\phi(\lambda_i),$$

where $\lambda_i := N\mu_i$ and $\phi(t) = te^{-t}$. If $N\max_i \mu_i \le 2$, then ϕ is concave on $[0,2]$, and $\mathbb{E}[\mathrm{MFR}]$ is Schur-concave in μ, minimized by imbalanced μ. If $N\min_i \mu_i \ge 2$, then ϕ is convex on $[2,\infty)$, and $\mathbb{E}[\mathrm{MFR}]$ is Schur-convex in μ, minimized by balanced μ.

Remark. Prior work (Zucchet et al., 2025) has observed that LLMs often exhibit a performance plateau before acquiring precise factual knowledge, and that training data distribution strongly shapes this dynamic: imbalanced distributions shorten the plateau but increase overfitting risk and slow long-term knowledge acquisition, while balanced distributions prolong the plateau yet yield faster accumulation and better generalization. Theorem 1 and 2 provide a unified explanation: throughout training, lower monofact rates induce smaller calibration error and better learning dynamics. Hence, imbalanced distributions in early stages help the model escape the plateau quickly, whereas uniform distributions in later stages promote efficient knowledge gain. Dynamically adjusting the data distribution to maintain a low monofact rate thus both shortens the plateau and maximizes learning speed.

Insight 3: Efficient interactions arise from sampling imbalanced queries early and shifting to uniform ones later.

5.2 Interactive Learning Strategy

Combining Insight 3 with Insight 2, we propose an Interactive Learning Strategy (ILS), formulated in Algorithm 1 (Appendix E). Given a query $\mathbf{x}$, the LLM first samples a pool of candidate questions. From this pool, it selects the question with the largest calibration error—our first experiment confirms that this metric faithfully captures the information value of a question. The answer from the human is then incorporated via implicit parameter updates (i.e., contextual updates), and after T rounds the model produces its final response to $\mathbf{x}$.

To make interaction efficient, we build on Insight 3 and curriculum learning (Bengio et al., 2009). The candidate pool is first sampled from the natural distribution μ, reflecting the long-tailed nature of real-world questions, but gradually shifts toward balanced exploration as interaction progresses. To this end, we introduce a unified scheduling function $\beta(t)$ that gradually shifts the sampling distribution, consistent with our Insight 1 that effective calibration requires prioritizing common cases first while progressively covering rare events. A case study of ILS is shown in Figure 1.

Baselines. To compare our calibration-based selection with alternative prompt-only selection schemes, we adopt two query-selection heuristics as baselines. **Random** first uses a PolicyLLM to generate M diverse clarification questions and then uniformly samples one query from the candidate set. **SelfSelect** uses the same candidate set but asks the LLM to select the best follow-up question. In contrast, our **ILS** selects queries by estimating calibration error. All prompts used in our experiments are provided in Appendix G.2.

Benchmarks. To evaluate ILS in realistic free-form generation, we use the ambiguous queries from AmbigQA (Min et al., 2020) and construct an ambiguous variant of HotpotQA (Yang et al., 2018) by injecting ambiguity following (Liu et al., 2025). Specifically, we convert each query into an AMR graph (Shi et al., 2023), prompt `GPT-4o` to remove key modifiers/constraints and perturb salient relations (Appendix G.1), and then linearize

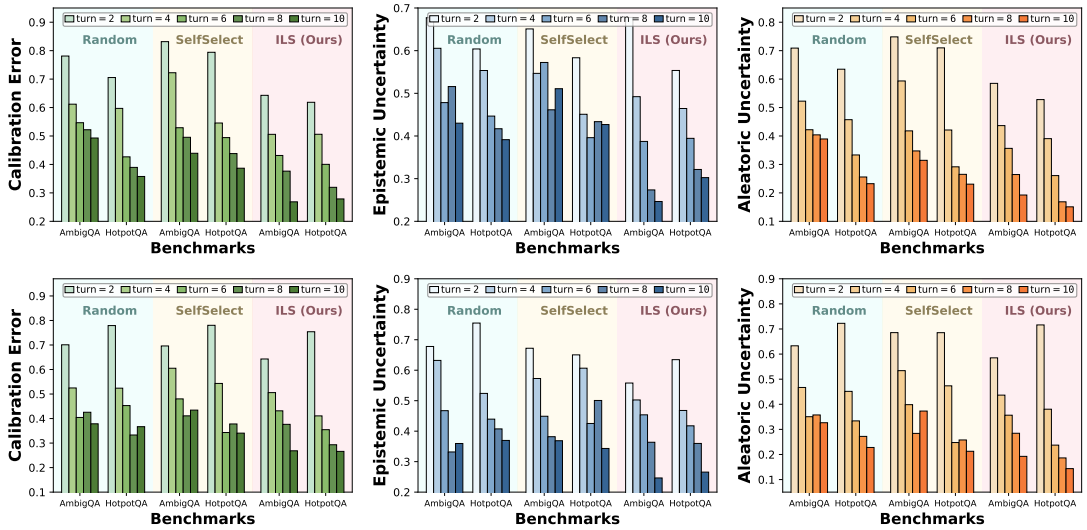

Figure 4: Convergence of calibration error, epistemic uncertainty, and aleatoric uncertainty under different interaction strategies. Across benchmarks, our calibration-based Interactive Learning Strategy (ILS) yields faster convergence than prompt-only baselines (Random, SelfSelect) for `Qwen2.5-7B` (top) and `LLaMA3-8B` (bottom).

the edited AMR back into an ambiguous query with corresponding clarification questions.

Evaluation metrics. Entropy in free-form generation is notoriously difficult to estimate because answers can be semantically equivalent while remaining lexically or syntactically distinct. Consequently, naively aggregating token-level probabilities as in the multiple-choice tasks in Section 4 can overestimate uncertainty (Xiao et al., 2020). To address this, we adopt semantic entropy (Farquhar et al., 2024) to measure epistemic uncertainty for long-form QA here. Detailed estimation procedures and prompts are provided in Appendix G.3.

Main Results. The results are summarized in Figure 4. The experiments in this section are conducted using `Qwen-2.5-7B` and `LLaMA3-8B` to enable efficient computation of the proposed metrics. Across benchmarks, our calibration-based ILS exhibits faster convergence than prompt-only baselines (Random and SelfSelect), suggesting that informatively guiding queries with calibration signals provides a robust advantage over naïve prompting. Moreover, the concurrent reduction of epistemic uncertainty, aleatoric uncertainty, and calibration error indicates substantial overlap among these metrics, and supports the view that calibration-driven interaction can also mitigate uncertainty. Finally, we observe that SelfSelect occasionally underperforms Random, consistent with *overconfidence* arising from an LLM's tendency to trust its own outputs; ILS alleviates this issue by selecting queries using an explicit calibration signal.

In Appendix F, we additionally report the valid solutions rate in Figure 5 and conduct a sensitivity analysis on the candidate set size M. In Figure 4 we report the main results using the respective optimal M for each LLM backbone.

Remark. ILS provides a systematic procedure for adaptive query selection that couples calibration-driven informativeness with curriculum scheduling. By dynamically reweighting the query distribution, ILS connects the theoretical insights of Section 5.1 with practical interaction design.

6 Conclusion

In this paper, we develop a calibration-centric view of uncertainty in LLM–human interaction. We establish that calibration error in LLMs admits a non-vanishing lower bound in the non-interactive setting, governed by the monofact rate. Controlled n-gram simulations and modern LLMs experiments confirm this theoretical link, showing that hallucinations persist without interaction. We further demonstrate that multi-turn interaction reduces calibration error. Therefore, we characterize how query distributions affect the monofact rate and propose Interactive Learning Strategy (ILS) that selects clarification questions by maximizing calibration error and schedules candidate sampling. We hope our theoretical insights and proposed interactive calibration strategy represent a step towards modeling uncertainty in LLM–human interaction and making hallucinations more precise.

Limitations

While this work establishes the theoretical foundations of interactive calibration, our experiments rely on large language models to simulate human interaction. Although such models can approximate ground-truth responses, they may still fail on intermediate or underspecified queries posed by another LLM, reflecting inherent capacity limitations. Addressing this limitation may require augmenting LLMs with external tools or interactive capabilities. In this work, we restrict our focus to standalone LLMs; extending the framework to multi-agent systems that incorporate fact-checking components and modeling uncertainty at the system level are left for future research.

This paper provides preliminary analyses of the proposed ILS algorithm, but the current evaluation remains limited in scope. We plan to conduct more comprehensive experiments to empirically validate the effectiveness and convergence behavior of ILS across diverse datasets, interaction settings, and model scales. Future work will also explore integrating quantitative convergence diagnostics and ablation studies to better understand how different scheduling strategies and uncertainty estimators influence stability and calibration performance.

References

Anastasios N Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. In *Foundations and Trends in Machine Learning*.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. In *International Conference on Machine Learning (ICML)*.

Catarina G Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chengliang Chai and Guoliang Li. 2020. Human-in-the-loop techniques in machine learning. volume 43, pages 37–52.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Görür. 2024. Finetuning language models to emit linguistic expressions of uncertainty. In *arXiv*.

Imre Csiszár and János Körner. 2011. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2 edition. Cambridge University Press.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.

Yihe Dong, Yizhong Zhang, Tianyi Zhang, Jindong Chen, Tianyu Gao, and Danqi Chen. 2023. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Irving J Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3).

G. H. Hardy, J. E. Littlewood, and G. Pólya. 1952. *Inequalities*, 2nd edition. Cambridge University Press.

Václav Havlík and 1 others. 2025. Why are llms' abilities emergent? *arXiv preprint arXiv:2508.04401*. Preprint.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. In *Findings of EMNLP*.

IMDb. 2024. Imdb non-commercial datasets. https://developer.imdb.com/non-commercial-datasets/. Accessed: 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Enkelejda Kasneci, Kathrin Seßler, and Stefan Küchemann. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. volume 103, page 102274.

Michael J Kearns and Umesh Vazirani. 1994. *An introduction to computational learning theory*. MIT press.

Michael Kirchhof, Luca Füger, Adam Goliński, Eeshan Gunesh Dhekane, Arno Blaas, and Sinead Williamson. 2025a. Self-reflective uncertainties: Do llms know their internal answer distribution? In *arXiv*.

Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. 2023. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *International Conference on Machine Learning (ICML)*.

Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. 2025b. Position: Uncertainty quantification needs reassessment for large-language model agents. *arXiv preprint arXiv:2505.22655*.

Kasia Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. 2025. Active task disambiguation with LLMs. In *The Thirteenth International Conference on Learning Representations*.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. *arXiv preprint arXiv:2403.09972*.

Stephanie C Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. In *Transactions on Machine Learning Research (TMLR)*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, and 1 others. 2024. Uncertainty quantification for in-context learning of large language models. *arXiv preprint arXiv:2402.10189*.

Jingyu Liu, Jingquan Peng, Xiaopeng Wu, Xubin Li, Tiezheng Ge, Bo Zheng, and Yong Liu. 2025. Do not abstain! identify and solve the uncertainty. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17177–17197.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

David McAllester and Luis E. Ortiz. 2003. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911. Earlier version in NIPS 2002.

Colin McDiarmid. 1989. On the method of bounded differences. In Johannes Siemons, editor, *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, Cambridge, UK.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Accessed: 2025-09-02.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Long Ouyang. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Yi Ren and Danica J Sutherland. 2024. Learning dynamics of llm finetuning. *arXiv preprint arXiv:2407.10490*.

Burr Settles. 2009. *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Kaize Shi, Xueyao Sun, Li He, Dingxian Wang, Qing Li, and Guandong Xu. 2023. Amr-tst: Abstract meaning representation-based text style transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4231–4243.

Anne M Van der Bles, Sander van der Linden, Alexandra L J Freeman, Judith Mitchell, Lisa Zaval, Ana B Galvão, and David J Spiegelhalter. 2019. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119. IEEE.

Peiqi Wang, Barbara D Lam, Yingcheng Liu, Ameneh Asgari-Targhi, Rameswar Panda, William M Wells, Tina Kapur, and Polina Golland. 2024. Calibrating expressions of certainty. In *arXiv*.

Renxi Wang, Xudong Han, Lei Ji, Shu Wang, Timothy Baldwin, and Haonan Li. 2025. Toolgen: Unified tool retrieval and calling via generation. In *ICLR 2025*. Poster.

Jason Wei, Yi Tay, Rishi Bommasani, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Chi, and 1 others. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Paul D Windschitl and Gary L Wells. 1996. Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, 2(4):343.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Tim Z Xiao, Aidan N Gomez, and Yarin Gal. 2020. Wat zei je? detecting out-of-distribution translations with variational transformers. *arXiv preprint arXiv:2006.08344*.

Sang Michael Xie, Yuhuai Lu, Tianle Gao, and Tengyu Chen. 2021. Explanations for in-context learning as implicit bayesian inference. In *International Conference on Learning Representations (ICLR)*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. 2025. Uncle: Uncertainty expressions in long-form generation. In *arXiv*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Empirical Methods in Natural Language Processing (EMNLP)*.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning models better express their confidence. In *arXiv*.

Boxuan Zhang and Ruqi Zhang. 2025. Cot-uq: Improving response-wise uncertainty quantification in llms with chain-of-thought. In *Findings of the Association for Computational Linguistics (ACL Findings)*.

Yicong Zhao, Shisong Chen, Jiacheng Zhang, and Zhixu Li. 2025. Recode: Improving llm-based code repair with fine-grained retrieval-augmented generation. *arXiv preprint arXiv:2509.02330*. Accepted by CIKM 2025.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. 2025a. From passive to active reasoning: Can large language models ask the right questions under incomplete information? In *Proceedings of the International Conference on Machine Learning (ICML)*. ArXiv preprint arXiv:2506.08295.

Zhi Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Lan-Zhe Guo, Xiaoxing Ma, and Yu-Feng Li. 2025b. Bridging internal probability and self-consistency for effective and efficient llm reasoning. *arXiv preprint arXiv:2502.00511*.

Alf C Zimmer. 1983. Verbal vs. numerical processing of subjective probabilities. In *Advances in psychology*, volume 16, pages 159–182. Elsevier.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. 2025. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*.

A Related Work

A.1 Output Uncertianty

Early work on output uncertainty in machine learning focused on numerical representations such as predicted probabilities, calibrated scores, or binary uncertainty flags (Lin et al., 2022; Band et al., 2024; Yona et al., 2024). While effective for classification and calibration tasks (Huang et al., 2024; Band et al., 2024), these methods confine uncertainty communication to a narrow scalar space.

Recent studies argue that LLMs should express uncertainty in text rather than numbers, leveraging the expressiveness of natural language (Angelopoulos and Bates, 2023; Kirchhof et al., 2023; Xu et al., 2024; Kirchhof et al., 2025b). This line of work highlights that models should not only report uncertainty levels but also explain which alternatives exist, why ambiguity arises, and what could resolve it (e.g., "*I'm not sure, but I think...*"). Approaches such as SelfReflect (Kirchhof et al., 2025a) and summarization-based methods (Zhang and Zhang, 2025; Yang et al., 2025; Yoon et al., 2025) generate textual uncertainty explanations aligned with internal belief states, while others explore verbal cues like "most likely" or "perhaps" (Chaudhry et al., 2024; Wang et al., 2024), raising challenges in aligning linguistic hedges with human interpretation (Van der Bles et al., 2019; Belém et al., 2024).

Despite this progress, output-level uncertainty research remains limited, especially beyond scalar confidence indicators. Our work advances this direction by introducing a metric that captures the bias between internal and expressed uncertainty in LLMs, providing a foundation for systematic, quantitative modeling of uncertainty expression.

A.2 Interactive Learning

A defining feature that distinguishes LLM agents from traditional machine learning models is their ability to interact with users to acquire information that reduces uncertainty. Such interaction manifests in two complementary ways (Kirchhof et al., 2025b; Kobalczyk et al., 2025): (1) users may provide information outside the model's training distribution (e.g., post-cutoff events or private data), thereby expanding the concept space with previously unseen possibilities; and (2) users may clarify underspecified aspects within the known space (e.g., disambiguating intentions or adding context), thus contracting uncertainty by resolving internal ambiguity without altering the outcome space. Recent work on task disambiguation operationalizes this paradigm by having an LLM proactively ask clarification questions to resolve underspecified tasks and improve downstream performance (Kobalczyk et al., 2025)

Compared with active learning (Settles, 2009), interactive learning differs in both goal and target. Active learning seeks to improve the global model θ through informative samples for generalization, whereas interactive learning is local, aiming to reduce uncertainty for the current instance $\mathbf{x}$. Moreover, active learning queries an oracle for labels, while interactive learning queries the user during problem solving, introducing unique challenges in human–computer interaction (Kasneci et al., 2023; Chai and Li, 2020). It also differs from in-context learning (ICL) (Brown et al., 2020; Xie et al., 2021; Min et al., 2022; Wei et al., 2023; Dong et al., 2023), which passively incorporates external context via prompts but lacks the ability to actively seek missing information.

In this work, we take a first step toward a formal theory of interactive learning, defining interaction in terms of uncertainty and showing how it reduces uncertainty through both concept-space expansion and contraction. This formulation extends beyond static prompt-based conditioning and captures the dynamic, epistemic role of LLM–user interaction.

B Auxiliary Lemmas

In this section, we justify several of our definitions.

B.1 Calibration Error

We begin with the definition of calibration error Cal in Section 2, where the distance d may be instantiated either as total variation distance or as Jensen–Shannon divergence. We show that the relation implied by Theorem 1 holds under both choices of d.

Lemma 1 (TV vs. JS under the IIV gate)**.** *Let* $\widehat{\mathcal{G}} := \{(\mathbf{x}, \mathbf{y}) : \hat{h}(\mathbf{x}, \mathbf{y}) = +1\}$ *be the high-probability region induced by the IIV classifier, and define*

$$\mathrm{Cal}(\pi_\theta, \mathbf{p}) := |\, \pi_\theta(\widehat{\mathcal{G}}) - \mathbf{p}(\widehat{\mathcal{G}})|.$$

Let $\mathrm{Cal}_{\mathrm{TV}}(\pi_\theta, \mathbf{p}) := \mathbb{E}_{\mathbf{x}\sim\mu}[|\pi_\theta(\cdot \mid \mathbf{x}) - \mathbf{p}(\cdot \mid \mathbf{x})|] = \mathrm{TV}(\pi_\theta \| \mathbf{p})$ *denote the total variation distance and* $\mathrm{Cal}_{\mathrm{JS}}(\pi_\theta, \mathbf{p}) := \mathbb{E}_{\mathbf{x}\sim\mu}[\mathrm{JS}(\pi_\theta(\cdot \mid \mathbf{x}) \| \mathbf{p}(\cdot \mid \mathbf{x}))] = \mathrm{JS}(\pi_\theta \| \mathbf{p})$ *denote the Jensen–Shannon divergence. Then*

$$\mathrm{Cal}(\pi_\theta, \mathbf{p}) \;\le\; \mathrm{Cal}_{\mathrm{TV}}(\pi_\theta, \mathbf{p}) \;\le\; \sqrt{2\,\mathrm{Cal}_{\mathrm{JS}}(\pi_\theta, \mathbf{p})}.$$

Proof. The first inequality is immediate from the variational form of total variation: $\mathrm{Cal}(P, Q) = \sup_A |P(A) - Q(A)|$, hence taking $A = \widehat{\mathcal{G}}$ gives $\mathrm{Cal}(\pi_\theta, \mathbf{p}) \le \mathrm{Cal}_{\mathrm{TV}}(\pi_\theta, \mathbf{p})$.

For the second inequality, by Cauchy–Schwarz,we have:

$$\begin{aligned}\mathrm{Cal}_{\mathrm{TV}}(\pi_\theta, \mathbf{p})^2 &= \Big(\mathbb{E}_{\mathbf{x}}\, d_{\mathrm{TV}}(\pi_\theta(\cdot \mid \mathbf{x}), \mathbf{p}(\cdot \mid \mathbf{x}))\Big)^2 \\ &\le \mathbb{E}_{\mathbf{x}}\, d_{\mathrm{TV}}(\pi_\theta(\cdot \mid \mathbf{x}), \mathbf{p}(\cdot \mid \mathbf{x}))^2.\end{aligned}$$

For each fixed $\mathbf{x}$, Pinsker's inequality (Csiszár and Körner, 2011) yields: $\mathrm{TV}(\pi_\theta(\cdot \mid \mathbf{x}), \mathbf{p}(\cdot \mid \mathbf{x}))^2 \le 2\,\mathrm{JS}(\pi_\theta(\cdot \mid \mathbf{x}) \| \mathbf{p}(\cdot \mid \mathbf{x}))$. Taking expectation over $\mathbf{x}$ concludes

$$\begin{aligned}\mathrm{Cal}_{\mathrm{TV}}(\pi_\theta, \mathbf{p})^2 &\le 2\,\mathbb{E}_{\mathbf{x}}\Big[\mathrm{JS}(\pi_\theta(\cdot \mid \mathbf{x}) \| \mathbf{p}(\cdot \mid \mathbf{x}))\Big] \\ &= 2\,\mathrm{Cal}_{\mathrm{JS}}(\pi_\theta, \mathbf{p}).\end{aligned}$$

□

B.2 Epistemic Uncertainty

Then we introduce other methods to represent the epistemic uncertainty of LLMs.

The first choice is based on aggregated probabilities, which typically aggregate output token probabilities of the generated text tokens $\mathbf{y} = [y_1, \ldots, y_L]$ to measure the LLMs' confidence for each response (Kadavath et al., 2022; Huang et al., 2023; Varshney et al., 2023). Formally, given an aggregation function $\mathrm{Aggr}(\cdot)$ and query $\mathbf{x}$, the epistemic uncertainty based on confidence can be represented as:

$$\mathrm{Epi}_{\mathbf{x}} = 1 - \mathbb{E}_{\mathbf{y}} \mathrm{Aggr}_{l=1}^{L} \left(\Pr(y_i \mid \mathbf{x}, \mathbf{y}_{<l})\right).$$

Common uses of $\mathrm{Aggr}(\cdot)$ include the mean and minimum (Orgad et al., 2024).

The other choice is the Perplexity (PPL) (Zhou et al., 2025b), which is defined as:

$$\mathrm{PPL}_{\mathbf{x}} = \mathbb{E}_{\mathbf{y}} \exp\left\{-\frac{1}{|L|}\sum_{l=0}^{L} \log \pi_\theta(y_l \mid \mathbf{x}, \mathbf{y}_{<l})\right\}$$

Besides, we can also use Margin Score (MS) to represent epistemic uncertainty, which is detailed by Wang and Shang (2014).

In both short- and long-form QA, prior work (Farquhar et al., 2024) has shown that token-/sequence-probability–based confidence estimators can be sub-optimal, since a single underlying meaning may admit many surface realizations. Consequently, uncertainty is more faithfully quantified over semantic equivalence classes rather than over specific word sequences. Accordingly, for long-form generations we use semantic entropy as a meaning-level proxy for epistemic uncertainty.

B.3 Interactive Learning Strategy (ILS)

Finally, we specify the Interactive Learning Strategy (ILS) mentioned in Section 5.2. At each interaction round t, we draw a set of candidate queries C_t i.i.d. from the scheduled distribution $\mathbf{q}_t(\mathbf{x}) \propto \mu(\mathbf{x})^{\beta(t)}$, where the schedule function $\beta(t)$ controls the transition from the natural long-tailed distribution $\mu(\mathbf{x})$ to a balanced one. Below we provide several illustrative examples:

- **Linear schedule.** A simple choice is $\beta(t) = 1 + \frac{t}{T}(\beta_{\max} - 1)$, which linearly increases from $\beta(0) = 1$ (fully natural sampling) to $\beta(T) = \beta_{\max}$ (nearly uniform). This schedule offers a smooth and interpretable transition and is effective when the total number of turns T is fixed in advance.
- **Logistic schedule.** To emphasize early exploration and later stabilization, one can use $\beta(t) = 1 + (\beta_{\max} - 1)\,\frac{1}{1 + \exp[-k(t - t_0)]}$, where k controls steepness and t_0 the midpoint. This form allows a gradual warm-up phase followed by rapid balancing, consistent with the idea of curriculum learning.
- **Exponential schedule.** For more aggressive late-stage balancing, $\beta(t) = 1 + (\beta_{\max} - 1)\left(1 - e^{-\lambda t}\right)$ with $\lambda > 0$ ensures a fast saturation to $\beta_{\max}$ and is suitable when early interactions quickly correct coarse biases.

In practice, $\beta_{\max}$ is typically set between 2 and 5, and the precise schedule shape is not critical as long as $\beta(t)$ is monotone nondecreasing and sufficiently smooth. Both the linear and logistic variants yield stable performance across datasets.

The model then selects the most informative query at round t:

$$\mathbf{x}_t = \arg\max_{\mathbf{x} \in C_t} D_t(\mathbf{x}),$$

where $D_t(\mathbf{x}) = \mathrm{Cal}_{\mathbf{x}}(\pi_{\theta_t}, p)$ denotes the local calibration error measuring the misalignment between the model's belief and the target distribution. The selected query $\mathbf{x}_t$ is presented to the human annotator, who provides a response $\mathbf{y}_t \sim \mathbf{p}(\cdot|\mathbf{x}_t)$. The LLM then performs an implicit in-context update

$\theta_{t+1} = \mathcal{I}(\theta_t; \mathbf{x}_t, \mathbf{y}_t)$, thereby refining its internal belief representation without explicit parameter optimization.

C Proof of Theorem 1

Background. Inspired by (Kalai et al., 2025), we partition answer space $\mathcal{Y}$ into valid answers $\mathcal{A}_\mathbf{x} = \text{supp}_\mathbf{x}(\mathbf{p}) = \{\mathbf{y} \in \mathcal{Y} \mid \mathbf{p}(\mathbf{y} \mid \mathbf{x}) > 0\}$ and errors $\mathcal{E}_\mathbf{x} = \mathcal{Y} \setminus \mathcal{A}_\mathbf{x}$ that are inconsistent with $\mathbf{p}$, for nonempty disjoint sets $\mathcal{E}_\mathbf{x}, \mathcal{A}_\mathbf{x}$. The aleatoric uncertainty of the LLM π_θ w.r.t. query $\mathbf{x}$ is denoted by:

$$\text{Ale}_\mathbf{x} := \pi_\theta(\mathcal{E}_\mathbf{x} \mid \mathbf{x}) = \Pr_{\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})}[\mathbf{y} \in \mathcal{E}].$$

Moreover, the aleatoric uncertainty of π_θ w.r.t. query distribution μ is given by $\text{Ale} := \mathbb{E}_{\mathbf{x}\sim\mu}\text{Ale}_\mathbf{x}$.

Is-It-Valid (IIV) reduction. We now formalize the Is-It-Valid (IIV) binary classification problem. IIV is specified by the target function $h : \mathcal{X} \times \mathcal{Y} \to \{-1, +1\}$ to be learned and the synthetic distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ (a 50/50 mix of samples from $\mathbf{p}$ and uniformly random errors):

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{1}{2}\,\mu(\mathbf{x})\,\mathbf{p}(\mathbf{y} \mid \mathbf{x}), & (\mathbf{x}, \mathbf{y}) \in \mathcal{A}, \\ \frac{1}{2}\,\mu(\mathbf{x})\,\frac{1}{|\mathcal{E}_\mathbf{x}|}, & (\mathbf{x}, \mathbf{y}) \in \mathcal{E}. \end{cases}$$

Thus $\mathcal{D}$ selects, with equal probability, either a pair from the preference distribution $\mathbf{p}$ or a pair where $\mathbf{x} \sim \mu$ and $\mathbf{y}$ is drawn uniformly from $\mathcal{E}_\mathbf{x}$. The classifier $\hat{h}(\mathbf{x}, \mathbf{y})$ predicts $+1$ whenever $\pi_\theta(\mathbf{y} \mid \mathbf{x}) > 1/\min_\mathbf{x}|\mathcal{E}_\mathbf{x}|$, and -1 otherwise.

We lower bound calibration error Cal in terms of IIV's aforementioned misclassification rate err:

$$\text{err} := \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\hat{h}(\mathbf{x}, \mathbf{y}) \neq h(\mathbf{x}, \mathbf{y})].$$

Lemma 2. *For any LLM π_θ and target distribution $\mathbf{p}$,*

$$\text{Cal}(\pi_\theta, \mathbf{p}) \;\geq\; 2 \cdot \text{err} - \frac{\max_\mathbf{x}|\mathcal{A}_\mathbf{x}|}{\min_\mathbf{x}|\mathcal{E}_\mathbf{x}|} - \text{Ale}.$$

Here, $\text{Cal}(\pi_\theta, \mathbf{p}) = |\pi_\theta(\widehat{\mathcal{G}}) - \mathbf{p}(\widehat{\mathcal{G}})|$ for $\mathcal{G} := \{(\mathbf{x}, \mathbf{y}) \mid \hat{h}(\mathbf{x}, \mathbf{y}) = +1\}$ measures the deviation between the model distribution and target distribution on the high-probability region.

Proof. Let $m := \min_\mathbf{x}|\mathcal{E}_\mathbf{x}|$ and $M := \max_\mathbf{x}|\mathcal{A}_\mathbf{x}|$. Define the high- and low-probability regions with respect to the threshold $1/m$:

$$\mathcal{G} := \{(\mathbf{x}, \mathbf{y}) : \pi_\theta(\mathbf{y} \mid \mathbf{x}) > 1/m\},$$
$$\mathcal{K} := \{(\mathbf{x}, \mathbf{y}) : \pi_\theta(\mathbf{y} \mid \mathbf{x}) \leq 1/m\}.$$

We partition the aleatoric uncertainty and misclassification rates into contributions above and below the threshold:

$$\text{Ale} = \pi_\theta(\mathcal{G} \setminus \mathcal{A}) + \pi_\theta(\mathcal{K} \setminus \mathcal{A}),$$
$$\text{err} = \mathcal{D}(\mathcal{G} \setminus \mathcal{A}) + \mathcal{D}(\mathcal{K} \cap \mathcal{A}).$$

For $(\mathbf{x}, \mathbf{y}) \in \mathcal{G} \setminus \mathcal{A}$, each term contributes

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \frac{\mu(\mathbf{x})}{2|\mathcal{E}_\mathbf{x}|} \;\leq\; \frac{\mu(\mathbf{x})}{2m},$$

while simultaneously contributing at least $\mu(\mathbf{x})/m$ to $\pi_\theta(\mathcal{G} \setminus \mathcal{A})$. Hence,

$$\pi_\theta(\mathcal{G} \setminus \mathcal{A}) \;\geq\; 2 \cdot \mathcal{D}(\mathcal{G} \setminus \mathcal{A}).$$

We aim to show

$$\pi_\theta(\mathcal{K} \setminus \mathcal{A}) \;\geq\; 2 \cdot \mathcal{D}(\mathcal{K} \cap \mathcal{A}) - \frac{M}{m} - \delta.$$

By construction,

$$2 \cdot \mathcal{D}(\mathcal{K} \cap \mathcal{A}) = \mathbf{p}(\mathcal{K} \cap \mathcal{A}).$$

Since $|\mathcal{A}_\mathbf{x}| \leq M$ and $\pi_\theta(\mathbf{y} \mid \mathbf{x}) \leq 1/m$ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{K} \cap \mathcal{A}$, we have

$$\pi_\theta(\mathcal{K} \cap \mathcal{A}) \;\leq\; \sum_\mathbf{x} \mu(\mathbf{x}) \cdot \frac{M}{m} \;=\; \frac{M}{m}.$$

Therefore,

$$\begin{aligned} &2 \cdot \mathcal{D}(\mathcal{K} \cap \mathcal{A}) - \pi_\theta(\mathcal{K} \setminus \mathcal{A}) \\ &= \mathbf{p}(\mathcal{K}) - \pi_\theta(\mathcal{K} \setminus \mathcal{A}) \\ &= \mathbf{p}(\mathcal{K}) - \big(\pi_\theta(\mathcal{K}) - \pi_\theta(\mathcal{K} \cap \mathcal{A})\big) \\ &\leq \text{Cal}(\pi_\theta, \mathbf{p}) + \pi_\theta(\mathcal{K} \cap \mathcal{A}) \\ &\leq \text{Cal}(\pi_\theta, \mathbf{p}) + \frac{M}{m}. \end{aligned}$$

Combining the above- and below-threshold contributions yields the claimed bound. □

We then review the Good–Turing (GT) estimator of missing mass (Good, 1953) and its finite-sample guarantees (McAllester and Ortiz, 2003). In our setting, prompts are drawn i.i.d. from a distribution μ over $\mathcal{X}$. Let the training multiset be $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ and denote by $c_N(\mathbf{x}) := \sum_{i=1}^N \mathbf{1}\{\mathbf{x}_i = \mathbf{x}\}$ the prompt count in S. Define the (population) *missing mass* of prompts as

$$R \;:=\; \Pr_{\mathbf{x}\sim\mu}[\,\mathbf{x} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}\,].$$

The Good–Turing estimator of R equals the fraction of monofacts in S, which in our notation is precisely the empirical monofact rate:

$$\widehat{R} \;=\; \tfrac{|\{\mathbf{x}\in\mathcal{X}:\, c_N(\mathbf{x})=1\}|}{N} \;=\; \text{MFR}.$$

Corollary 1 (McAllester and Ortiz (2003))**.** *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_N \overset{i.i.d.}{\sim} \mu$ *and let* R *and* MFR *be defined as above. For any* $\gamma \in (0, 1]$*,*

$$\Pr\Big[\, |\, R - \mathrm{MFR}\,| \;\le\; \tfrac{1}{N} \;+\; 2.42\sqrt{\tfrac{\ln(4/\gamma)}{N}}\,\Big] \;\ge\; 1-\gamma.$$

Proof. Let $\overline{R} := \mathbb{E}[R]$ and $\overline{\mathrm{MFR}} := \mathbb{E}[\mathrm{MFR}]$, where the expectation is taken over the training sample $S \sim \mu^N$. By the classical Good–Turing analysis (Good, 1953), the bias between MFR and $\overline{R}$ is bounded:

$$0 \;\le\; \mathrm{MFR} - \overline{R} \;\le\; \frac{1}{N}. \tag{1}$$

Next, McDiarmid's inequality (McDiarmid, 1989) implies concentration of R:

$$\Pr\big[\,|R - \overline{R}| \;\ge\; \varepsilon_1\,\big] \;\le\; 2\exp(-2N\varepsilon_1^2). \tag{2}$$

Similarly, Theorems 10 and 16 of McAllester and Ortiz (2003) yield concentration of MFR:

$$\Pr\big[\,|\mathrm{MFR} - \overline{\mathrm{MFR}}| \;\ge\; \varepsilon_2\,\big] \;\le\; 2\exp\Big(-\tfrac{N\varepsilon_2^2}{2}\Big). \tag{3}$$

Set $\varepsilon_1 = \sqrt{\frac{\ln(4/\gamma)}{N}}$ and $\varepsilon_2 = 2\sqrt{\frac{\ln(4/\gamma)}{N}}$. By a union bound, with probability at least $1-\gamma$ both (2) and (3) hold. Combining with (1) and the triangle inequality gives

$$|R - \mathrm{MFR}| \;\le\; \tfrac{1}{N} + (1+\sqrt{2})\sqrt{\tfrac{\ln(4/\gamma)}{N}}.$$

Finally, note that $1 + \sqrt{2} \le 2.42$, yielding the stated bound. □

We next extend the Good–Turing analysis to the case where the model may abstain, producing an IDK response that is not counted in calibration. Specifically, for a prompt $\mathbf{x}$, we say it is *answered* in the training data if there exists a training example $(\mathbf{x}, \mathbf{y})$ with $\mathbf{y} \neq \texttt{IDK}$, and *unanswered* otherwise. Let

$$\mathcal{U} \;:=\; \{\mathbf{x} \in \mathcal{X} : \nexists\, \mathbf{y} \in S \;\text{ with }\; \mathbf{y} \neq \texttt{IDK}\}$$

denote the set of unanswered prompts. The missing mass with abstentions is then defined as

$$R^* \;:=\; \Pr_{\mathbf{x}\sim\mu}[\,\mathbf{x} \in \mathcal{U}\,].$$

The empirical estimator in this setting is simply the monofact rate restricted to answered samples, denoted $\widehat{R}^*$. Intuitively, memorizing all answered queries suffices for perfect classification, so the only uncertainty comes from $\mathcal{U}$. We now show that $\widehat{R}^*$ remains a good estimator of R^*.

Lemma 3. *For all* N *and* $\gamma \in (0, 1]$*,*

$$\Pr\Big[\,|R^* - \widehat{R}^*| \;\le\; \tfrac{1}{N} + \tfrac{\ln(5/\gamma)}{N} + 2.42\sqrt{\tfrac{\ln(5/\gamma)}{N}}\,\Big] \;\ge\; 1-\gamma.$$

Proof. The only difference from the standard Good–Turing estimator is that we collapse all IDK responses into identical samples. Consequently, the classical GT estimator may count at most one additional monofact compared to $\widehat{R}^*$, yielding

$$\mathrm{GT} - \widehat{R}^* \in \Big\{0,\; \tfrac{1}{N}\Big\}.$$

This substitution induces a modified distribution ϕ in which $\phi(\texttt{IDK}) = \sum_{\mathbf{x}} \mu(\mathbf{x})\mathbf{p}(\texttt{IDK} \mid \mathbf{x})$ denotes the total abstention probability. Then $R - R^* \in \{0, \phi(\texttt{IDK})\}$ with probability $1 - (1 - \phi(\texttt{IDK}))^N$. In particular, if $\phi(\texttt{IDK}) \ge \frac{1}{N}\ln(5/\gamma)$, then $(1 - \phi(\texttt{IDK}))^N \le \frac{\gamma}{5}$, and hence

$$\Pr\Big[R - R^* \in \Big[0,\; \tfrac{1}{N}\ln\tfrac{5}{\gamma}\Big]\Big] \;\ge\; 1 - \tfrac{\gamma}{5}.$$

Combining this with the concentration bound of Corollary 1 applied at confidence $4\gamma/5$, and applying a union bound together with the triangle inequality, we obtain

$$\Pr\Big[\,|R^* - \widehat{R}^*| \;\le\; \tfrac{1}{N} + \tfrac{1}{N}\ln\tfrac{5}{\gamma} + 2.42\sqrt{\tfrac{\ln(5/\gamma)}{N}}\,\Big] \;\ge\; 1-\gamma.$$

Finally, the slack in the logarithmic term ensures that the bound remains valid whenever $\sqrt{z} \le z$ for $z \ge 2$, completing the proof. □

Lemma 4. *For any* $N \ge 1$*,* $\gamma \in (0, 1]$*, and any algorithm outputting a model* π_θ*, with probability at least* $1 - \gamma$ *over* $S \sim \mu^N$*,*

$$2\,\mathrm{err} \;\ge\; \widehat{R}^* \;-\; \frac{6\ln(3N/\gamma)}{\sqrt{N}}.$$

Proof. By Lemma 3 (GT with abstentions) we have, with probability at least $1 - \frac{\gamma}{2}$,

$$\Pr\left[\,|R^* - \widehat{R}^*| \;\le\; 4.42\sqrt{\frac{\ln(10/\gamma)}{N}}\,\right] \;\ge\; 1 - \frac{\gamma}{2}.$$

Since $\sqrt{\ln(10/\gamma)} \le \ln(3N/\gamma)$ for $N \ge 2$ (and the lemma is trivial for $N = 1$), it suffices to prove that, with probability at least $1 - \frac{\gamma}{2}$,

$$2\,\mathrm{err} \;\ge\; R^* \;-\; \sqrt{\frac{2}{N}\ln\frac{3N}{\gamma}}. \tag{4}$$

Let $\zeta := \frac{1}{N}\ln(3N/\gamma)$ and let $\alpha_{\mathbf{x}} := \Pr[\text{answer} \neq \texttt{IDK} \mid \mathbf{x}]$ be the non-abstention probability. Define the answered-probability mass $\mu'(\mathbf{x}) := \mu(\mathbf{x})\alpha_{\mathbf{x}}$, so that

$$R^* = \sum_{\mathbf{x}\in\mathcal{U}} \mu'(\mathbf{x}),$$

where $\mathcal{U}$ is the set of unanswered prompts (no non-IDK example in S). The lemma follows from the next two inequalities:

$$\Pr\left[\forall \mathbf{x} \in \mathcal{U}\ \mu'(\mathbf{x}) \le \zeta\right] \ge 1 - \frac{\gamma}{3}, \tag{5}$$

$$\Pr\left[2\,\mathrm{err} \ge R^* - \sqrt{\tfrac{2}{N}\ln\tfrac{3N}{\gamma}} \mid \forall \mathbf{x} \in \mathcal{U}\ \mu'(\mathbf{x}) \le \zeta\right] \ge 1 - \frac{\gamma}{6}. \tag{6}$$

Proof of (5). Note that there are at most $1/\zeta$ prompts whose $\mu'(\mathbf{x})$ can exceed ζ. Each such prompt $\mathbf{x}$ remains unanswered with probability at most $(1-\zeta)^N \le e^{-\zeta N} = \gamma/(3N)$. By a union bound,

$$\Pr\left[\exists \mathbf{x} \in \mathcal{U} : \mu'(\mathbf{x}) > \zeta\right] \le \frac{1}{\zeta}\cdot\frac{\gamma}{3N} = \frac{\gamma}{3},$$

which yields (5).

Proof of (6). For $\mathbf{x} \in \mathcal{U}$, define

$$\gamma_{\mathbf{x}} := \frac{1}{2}\left(\mathbf{1}\{f(\mathbf{x}, a_{\mathbf{x}}) = -\} + \sum_{\mathbf{y}\in\mathcal{R}_{\mathbf{x}}\setminus\{a_{\mathbf{x}}\}} \frac{\mathbf{1}\{f(\mathbf{x},\mathbf{y}) = +\}}{|\mathcal{R}_{\mathbf{x}}| - 1}\right) \in [0,1],$$

where $a_{\mathbf{x}}$ is the gold answer (if any), $\mathcal{R}_{\mathbf{x}}$ is the response set (including IDK), and f is the classifier induced by π_θ in the IIV task (predict + iff $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ exceeds the threshold). As in the previous analysis (Kalai et al., 2025), the IIV error decomposes as

$$\mathrm{err} \ge \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{U}} \mu'(\mathbf{x})\,\gamma_{\mathbf{x}}.$$

Conditioning on $\forall \mathbf{x} \in \mathcal{U}\,\mu'(\mathbf{x}) \le \zeta$ makes the random variables $\{\mu'(\mathbf{x})\gamma_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{U}}$ independent and bounded in $[0, \mu'(\mathbf{x})]$. Moreover,

$$\sum_{\mathbf{x}\in\mathcal{U}} \big(\mu'(\mathbf{x})\big)^2 \le \max_{\mathbf{x}\in\mathcal{U}} \mu'(\mathbf{x}) \sum_{\mathbf{x}\in\mathcal{U}} \mu'(\mathbf{x}) \le \zeta\, R^*.$$

Hoeffding's inequality thus yields

$$\Pr\left[\sum_{\mathbf{x}\in\mathcal{U}} \mu'(\mathbf{x})\gamma_{\mathbf{x}} \le \frac{R^*}{2} - \sqrt{\frac{\zeta\,\ln(6/\gamma)}{2}} \,\middle|\, \forall \mathbf{x} \in \mathcal{U}\ \mu'(\mathbf{x}) \le \zeta\right] \le \frac{\gamma}{6}.$$

Since $\sqrt{2\,\zeta\,\ln(6/\gamma)} = \sqrt{\frac{2}{N}\ln(3N/\gamma)}$ (using $\zeta = \ln(3N/\gamma)/N$), this implies (6).

Finally, combining (5) and (6) by a union bound gives (4) with probability at least $1-\gamma/2$. Together with the GT–abstention concentration (Lemma 3) and another union bound, we conclude that with probability at least $1-\gamma$,

$$2\,\mathrm{err} \ge \widehat{R}^* - \frac{6\ln(3N/\gamma)}{\sqrt{N}},$$

as claimed. □

Theorem 1(Non-vanishing Lower Bound for Calibration Error) With probability at least $1-\delta$ over the training samples,

$$\mathrm{Cal}(\pi_\theta, \mathbf{p}) \ge \mathrm{MFR} - \mathrm{Ale} - \frac{3e^{-m}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{N}},$$

where m is a sparsity parameter. For realistic data size, the final two terms are negligible, leaving calibration error and uncertainty essentially governed by the monofact rate.

Proof. By Lemma 2, there exists a spillover bound such that with probability at least $1-\delta/3$,

$$\mathrm{Cal}(\pi_\theta, \mathbf{p}) \ge \mathrm{err} - \mathrm{Ale} - \frac{3e^{-m}}{\delta}, \tag{7}$$

with $\max_{\mathbf{x}} |\mathcal{A}_{\mathbf{x}}| = 2$. Independently, by the monofact lower bound for IIV in Lemma 4, with probability at least $1-2\delta/3$,

$$\mathrm{err} \ge \mathrm{MFR} - \sqrt{\frac{6\ln(6/\delta)}{N}}. \tag{8}$$

Applying a union bound to (7) and (8) shows that both events hold simultaneously with probability at least $1-\delta$, and substituting (8) into (7) yields

$$\mathrm{Cal}(\pi_\theta, \mathbf{p}) \ge \mathrm{MFR} - \mathrm{Ale} - \frac{3e^{-m}}{\delta} - \sqrt{\frac{6\ln(6/\delta)}{N}}.$$

This proves the claim. □

D Proof of Theorem 2

In this section, we provide detailed proof for Theorem 2

Theorem 2. Let μ be a distribution on the prompt space $\mathcal{X}$ with $|\mathcal{X}| = K$, then the expected monofact rate is

$$\mathbb{E}_{S\sim\mu^N}[\mathrm{MFR}] = \frac{1}{N}\sum_{i=1}^{K} N\mu_i\,(1-\mu_i)^{N-1}.$$

Let $\lambda_i := N\mu_i$ and define $\phi(t) = te^{-t}$. Then

$$\mathbb{E}_{S\sim\mu^N}[\mathrm{MFR}] \;\approx\; \frac{1}{N}\sum_{i=1}^{K}\phi(\lambda_i).$$

- If $N\max_i \mu_i \le 2$, then each $\lambda_i \le 2$ and ϕ is concave on $[0,2]$. Hence $\mathbb{E}[\mathrm{MFR}]$ is Schur-concave in μ, minimized by imbalanced μ.
- If $N\min_i \mu_i \ge 2$, then each $\lambda_i \ge 2$ and ϕ is convex on $[2,\infty)$. Hence $\mathbb{E}[\mathrm{MFR}]$ is Schur-convex in μ, minimized by balanced μ.

Moreover, by Good–Turing concentration in Corollary 1, for any $\delta \in (0,1]$,

$$\Pr\Big[\,\big|\mathrm{MFR} - \mathbb{E}[\mathrm{MFR}]\big| \le \tfrac{1}{N} + 2.42\sqrt{\tfrac{\ln(4/\delta)}{N}}\,\Big] \ge 1-\delta,$$

so the same ordering holds for the empirical MFR with high probability up to $O\big(\sqrt{\ln(1/\delta)/N}\big)$ deviations.

Proof. For $\forall\ i$, $\Pr[c_N(\mathbf{x}_i) = 1] = \binom{N}{1}\mu_i(1-\mu_i)^{N-1} = N\mu_i(1-\mu_i)^{N-1}$, hence

$$\begin{aligned}\mathbb{E}[\mathrm{MFR}] =& \frac{1}{N}\sum_{i=1}^{K}\Pr[c_N(\mathbf{x}_i)=1]\\ =& \frac{1}{N}\sum_{i=1}^{K} N\mu_i(1-\mu_i)^{N-1}.\end{aligned}$$

Substituting $\lambda_i = N\mu_i$ yields:

$$\begin{aligned}\mathbb{E}[\mathrm{MFR}] \;&=\; \frac{1}{N}\sum_{i=1}^{K} N\mu_i(1-\mu_i)^{N-1}\\ &=\; \frac{1}{N}\sum_{i=1}^{K}\lambda_i\Big(1-\frac{\lambda_i}{N}\Big)^{N-1}.\end{aligned} \tag{9}$$

Under the Poissonized experiment where the total count $\widetilde{N} \sim \mathrm{Poisson}(N)$ and $\widetilde{C}_i \sim \mathrm{Poisson}(\lambda_i)$ are independent, $\Pr[\widetilde{C}_i = 1] = \lambda_i e^{-\lambda_i}$. Thus $\mathbb{E}_{\mathrm{Pois}}[\mathrm{MFR}] = \frac{1}{N}\sum_i \Pr[\widetilde{C}_i = 1] = \frac{1}{N}\sum_i \phi(\lambda_i)$, giving:

$$\mathbb{E}_{\mathrm{Pois}}[\mathrm{MFR}] \;=\; \frac{1}{N}\sum_{i=1}^{K}\phi(\lambda_i). \tag{10}$$

A standard de-Poissonization argument (e.g., analytic combinatorics) shows that for occupancy statistics with bounded toll functions,

$$\Big|\,\Pr[C_i=1] - \Pr[\widetilde{C}_i=1]\,\Big| \;\le\; \frac{c_1}{N}\min\{1,\lambda_i^2 e^{-\lambda_i}\},$$

for an absolute constant c_1. Summing over i and dividing by N gives (11) with $C_0 = c_1$:

$$\Big|\mathbb{E}[\mathrm{MFR}] - \mathbb{E}_{\mathrm{Pois}}[\mathrm{MFR}]\Big| \;\le\; \frac{C_0}{N}. \tag{11}$$

Define $\phi(t) = te^{-t}$. Then $\phi''(t) = e^{-t}(t-2)$, so ϕ is concave on $[0,2]$ and convex on $[2,\infty)$.

By Karamata's inequality (Hardy et al., 1952), a symmetric convex function is Schur-convex and a symmetric concave function is Schur-concave. Specifically, fixing $\sum_i \lambda_i = N$, Karamata's inequality implies: if every λ_i lies in a concave interval, then $\sum_i \phi(\lambda_i)$ is Schur-concave; if every λ_i lies in a convex interval, then it is Schur-convex. Thus, in the small-N regime ($\lambda_i \le 2$ for all i), the minimum is achieved at the most imbalanced μ and the maximum at the balanced μ; conversely, in the large-N regime ($\lambda_i \ge 2$ for all i), the minimum is at the balanced μ and the maximum at the most imbalanced μ. By (11), these statements transfer from $\mathbb{E}_{\mathrm{Pois}}[\mathrm{MFR}]$ to $\mathbb{E}[\mathrm{MFR}]$ with $O(1/N)$ slack.

Finally, the Good–Turing concentration bound in Corollary 1 ensures that MFR concentrates around $\mathbb{E}[\mathrm{MFR}]$ at the stated rate, so the ordering by majorization holds for the empirical MFR with probability $1-\delta$. □

Mixed regime. The clean dichotomy above requires all $\lambda_i = N\mu_i$ to lie in either the concave regime $[0,2]$ or the convex regime $[2,\infty)$. When the coordinates of μ straddle both regions (i.e., some $\lambda_i < 2$ while others $\lambda_j > 2$), the behavior of $\mathbb{E}[\mathrm{MFR}] = \frac{1}{N}\sum_i \phi(\lambda_i)$ becomes more intricate: the function ϕ is neither globally concave nor globally convex, so $\mathbb{E}[\mathrm{MFR}]$ is neither Schur-concave nor Schur-convex. In this *mixed regime*, the minimizer of $\mathbb{E}[\mathrm{MFR}]$ typically lies on the boundary of the simplex where probability mass is pushed as much as possible toward the "convex" side (large λ_i) while respecting normalization. Intuitively, allocating extra mass to already large-λ_i coordinates reduces their monofact contribution sharply (since $\phi(t)$ decays exponentially after $t > 2$), whereas spreading mass to small-λ_i coordinates increases monofact probability. Thus, even in the mixed regime, the extremizers remain highly *imbalanced* distributions, though the precise structure depends on how many coordinates fall above or below the $t = 2$ threshold.

E Algorithm

In this section, we present the proposed Interactive Learning Strategy (ILS) in detail (Algorithm 1),

Algorithm 1: Interactive Learning Strategy (ILS)

Input: Input space $\mathcal{X}$, natural distribution $\mu(\mathbf{x})$, human query $\mathbf{x}$, total rounds T, candidate size M, schedule $\beta(t)$, initial parameters θ_0, target distribution $\mathbf{p}(\mathbf{y} \mid \mathbf{x})$

Output: LLM π_{θ_T} and response $\mathbf{y} \sim \pi_{\theta_T}$

Initialize $\theta \leftarrow \theta_0$;

for $t = 0$ **to** $T - 1$ **do**

  Define scheduled sampling distribution $\mathbf{q}_t(\mathbf{x}) \leftarrow \mu(\mathbf{x})^{\beta(t)} / \sum_{\mathbf{x}'} \mu(\mathbf{x}')^{\beta(t)}$;

  Sample candidate set $\mathcal{C}_t \subset \mathcal{X}$, $|\mathcal{C}_t| = M$, from $\mathbf{q}_t$;

  foreach $\mathbf{x} \in \mathcal{C}_t$ **do**

    $D_t(\mathbf{x}) \leftarrow \mathrm{Cal}_{\mathbf{x}}(\pi_\theta, \mathbf{p})$;

  $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x} \in \mathcal{C}_t} D_t(\mathbf{x})$;

  Query human for $\mathbf{y}_t \sim \mathbf{p}(\mathbf{y} \mid \mathbf{x}_t)$;

  $\theta_{t+1} \leftarrow \mathcal{I}(\theta_t; \mathbf{x}_t, \mathbf{y}_t)$;

return $\mathbf{y} \sim \pi_{\theta_T}(\cdot \mid \mathbf{x})$

while deferring the corresponding prompt strategy to Appendix G.2.

F Additional Experiments

Additional Metrics. To demonstrate that ILS also improves performance under more general metrics, we report the valid solution rate across interaction turns in Figure 5. At each turn, we sample responses and perform semantic clustering; a response is counted as valid if its cluster is semantically consistent with the ground-truth answer. As the interaction turn t increases, validity improves consistently across benchmarks, and ILS achieves the highest valid solution rate on both backbones (`Qwen-2.5-7B` and `LLaMA3-8B`) compared with Random and SelfSelect.

Hyper-parameter Analysis. In figure 6, we analyze the effect of the candidate set size M on the evaluation metrics. Empirically, the optimal M is around 10 for `Qwen2.5-7B` and around 15 for `LLaMA3-8B`. Accordingly, in Figure 4 we report the main results using the respective optimal M for each model.

G Prompt Strategy

G.1 Benchmarks

In this section, we present the prompt in the interaction generation of DC, SP, and HotpotQA:

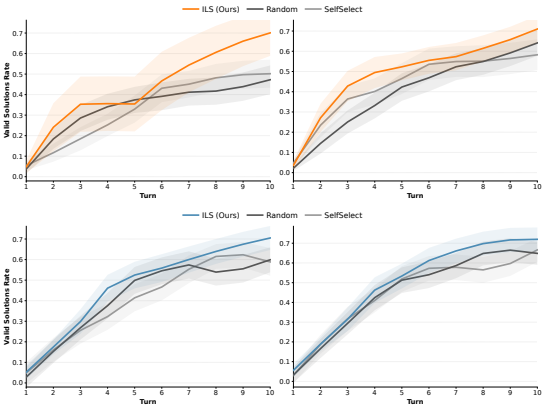

Figure 5: Valid solution rate over interaction turns for `Qwen2.5-7B` (left) and `LLaMA3-8B` (right)

- Detective Cases:
  - Prompt of Policy LLM (Lists 1,2).
  - Prompt of Response LLM (List 3).
- Situation Puzzles:
  - Prompt of Policy LLM (Lists 4,5).
  - Prompt of Response LLM (List 6).
- HotpotQA: Prompt to Ambiguate the AMR (List 7).

G.2 Baselines and ILS

Random. Candidate questions are generated using List 8, and no selection prompt is involved since a query is sampled uniformly at random.

SelfSelect. We use List 9 to prompt the LLM to choose the single best follow-up question from the candidate set.

ILS. We use two prompts to estimate calibration error: List 10 prompts the PolicyLLM to simulate user responses without access to the true intent, while List 11 prompts the ResponseLLM to simulate ground-truth user responses with access to the true intent and answer. We then merge the two response sets and apply semantic clustering using the ResponseLLM with List 12 to obtain semantic clusters. Finally, we compute $\mathrm{Cal}_{\mathbf{x}}$ for each candidate question $\mathbf{x}$ and select the query with the largest $\mathrm{Cal}_{\mathbf{x}}$, as formulated in Algorithm 1.

G.3 Evaluation Metrics.

Epistemic Uncertainty. We first use List 13 to prompt the PolicyLLM to generate $K = 5$ candidate answers, and then apply List 12 to perform semantic clustering over these answers. We finally

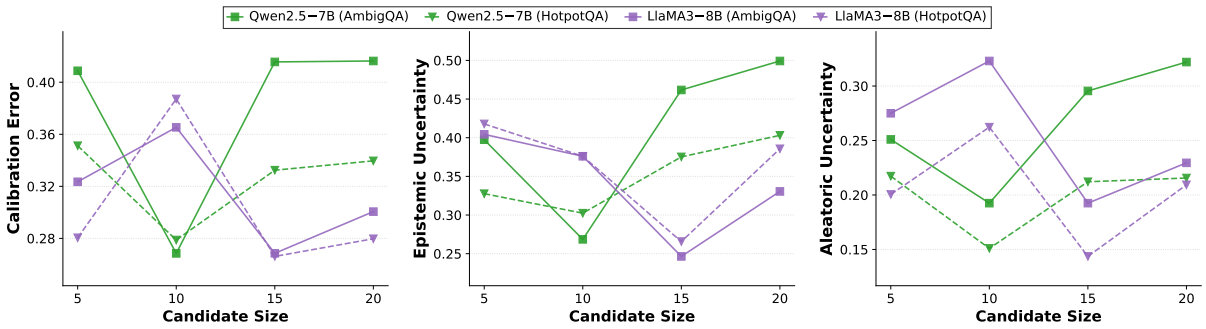

Figure 6: Sensitivity of aleatoric uncertainty, epistemic uncertainty, and calibration error to the candidate set size M across AmbigQA and HotpotQA for `Qwen2.5-7B` and `LLaMA3-8B`.

compute semantic entropy based on the resulting cluster distribution. Let m denote the number of non-empty semantic clusters. We define

$$\mathrm{Epi}_{\mathbf{x}} = -\frac{1}{\log m} \sum_{j=1}^{m} \pi_\theta(c_j \mid \mathbf{x}) \log \pi_\theta(c_j \mid \mathbf{x}),$$

where $\pi_\theta(c_j \mid \mathbf{x})$ is the empirical probability mass of cluster c_j, and $\log m$ is used for normalization.

Aleatoric Uncertainty. Building on the semantic clustering, we use List 14 to prompt the ResponseLLM to score the representative answer of each cluster against the gold label. The score takes one of five discrete levels $\{0, 1, 2, 3, 4\}$, which we normalize into a correctness value

$$s_j = \frac{\text{level}}{(0+1+2+3+4)}.$$

We then estimate aleatoric uncertainty as the frequency-weighted error mass over clusters,

$$\mathrm{Ale}_{\mathbf{x}} = \sum_{j=1}^{m} \pi_\theta(c_j \mid \mathbf{x})(1 - s_j),$$

where $\pi_\theta(c_j \mid \mathbf{x})$ is the empirical frequency of cluster c_j. Intuitively, $\mathrm{Ale}_{\mathbf{x}}$ is low when the model's sampled answers concentrate on high-scoring clusters, and high when they concentrate on low-scoring clusters.

Calibration Error. Cal is computed in our experiments in two ways. First, as a diagnostic metric, we compute Cal on the PolicyLLM's answers. Second, as an optimization objective, we compute $\mathrm{Cal}_{\mathbf{x}}$ for each candidate clarification question $\mathbf{x}$ and select the one with the largest value.

- **Diagnostic Metrics:** Since Cal is defined as the Jensen–Shannon (JS) divergence between a model distribution and a reference (ground-truth) distribution in Section 2, we construct the reference distribution by assigning higher probability mass to answers with higher correctness. Concretely, for open-form answers scored on five discrete levels $\{0, 1, 2, 3, 4\}$, we set the reference probability of each level to $\frac{\text{level}}{(0+1+2+3+4)}$. The model distribution is obtained by semantic clustering and scoring of the PolicyLLM's sampled answers, yielding an empirical probability mass over the same five levels. We then compute the JS divergence between the two distributions as Cal.

- **Optimization Objective:** For each candidate clarification question $\mathbf{q}$, following Appendix G.2, we construct two response sets to $\mathbf{q}$. We first prompt the PolicyLLM, which observes the ambiguous question and the dialogue history, to simulate how a user would answer $\mathbf{q}$, and draw $K = 5$ samples. We then prompt the ResponseLLM, which observes the disambiguated original question, the dialogue history, and the gold answer, to answer $\mathbf{q}$ and sample responses. After merging the two sets, we use the ResponseLLM to perform semantic clustering, producing a set of semantic clusters. The model distribution is given by the PolicyLLM's empirical frequency over clusters, whereas the reference user distribution is given by the ResponseLLM's empirical frequency over clusters. We compute $\mathrm{Cal}_{\mathbf{x}}$ as the JS divergence between these two cluster distributions, select the question $\mathbf{x}$ with the largest $\mathrm{Cal}_{\mathbf{x}}$, obtain the ResponseLLM's answer under access to the clarification and the gold answer, append the selected $\mathbf{q}$ and its answer to the dialogue history, and proceed to the next interaction round.

H Dataset Examples

In this section, we show dataset examples of the AR-Bench.

For Detective Cases (DC), we show the initial information of the case for the player model (List 15) and the truth for each suspect (List 16,17,18,19,20). The murderer in this case is Professor Evelyn Hartman, driven by academic plagiarism and professional betrayal.

In terms of Situation Puzzles (SP), we show the initial puzzle for the player model (List 21) and the truth for the judge (List 22). This example involves identity confusion, resembling everyday life puzzles, and centers on the paradox: despite never leaving his hometown, John was observed in two different countries at the same time.

Table 1: Prompt for Policy LLM in DC

Prompt for Policy LLM in DC
System: You are a concise detective. Choose the murderer strictly by the case background and any follow-up facts. CRITICAL: Output ONLY ONE capital letter from {A,B,C,D,E}. Do not include any words, punctuation, or explanation. User: CASE BACKGROUND: <background> SUSPECT CHOICES: <choices> Optional: augmented history>Optional: augmented history Answer with ONLY one letter: A, B, C, D, or E.

Table 2: Prompt for Uncertain Question Messages in DC

Prompt for Uncertain Question Messages in DC
System: You are generating a single disambiguating question that would maximally reduce uncertainty about the murderer. The question must be directly verifiable given the ground-truth murderer. CRITICAL: Output only one short question. No explanations. User: CASE BACKGROUND: <background> SUSPECT CHOICES: <choices> Optional: augmented history>Optional: augmented history Output one short, verifiable YES/NO or factual question (<= 20 words) that best distinguishes the true murderer.

Table 3: Prompt for Response LLM in DC

Prompt for Response LLM in DC
System: You are an oracle with FULL ACCESS to the complete case file (verbatim text below). Answer ONLY using facts contained in the file. If the question is irrelevant to determining the murderer, or it cannot be verified from the file, output exactly: `It doesn't matter`. Do not add any extra words or punctuation. User: FULL CASE FILE (verbatim): <full_case_text> QUESTION: <sub_question> Answer policy: - If relevant and verifiable: output YES or NO; if not strictly yes/no, output a short factual span. - If irrelevant or unverifiable: output exactly `It doesn't matter`.

Table 4: Prompt for Policy LLM in SP

Prompt for Policy LLM in SP
System: You are a precise investigator. Produce the most plausible full explanation of the case. OUTPUT STYLE: - Self-contained; integrate any known facts from QA history. - No bullet points, no lists, no numbering, no headings, no meta commentary. User: SURFACE QUESTION: <surface> CORE SENTENCE: <core_sentence> Output your best explanation of what actually happened (3–5 sentences).

Table 5: Prompt for Uncertain Question Messages in SP

Prompt for Uncertain Question Messages in SP
System: Ask ONE disambiguating question that would maximally reduce uncertainty about the true explanation. The question must be specific and verifiable from a complete case file (if provided to an oracle). CRITICAL OUTPUT RULES: - Output exactly ONE question. - Single sentence. No preamble, no explanations, no bullets. - Prefer YES/NO or a short factual query. - Do NOT repeat a question already answered by the known facts. User: SURFACE QUESTION: <surface> CORE SENTENCE: <core_sentence> Output exactly ONE most-informative question.

Table 6: Prompt for Response LLM in SP

Prompt for Response LLM in SP
System: You are an objective grader. Read the full case file and grade the candidate explanation. RUBRIC: 0=contradicted/irrelevant; 1=mostly irrelevant; 2=partially plausible; 3=largely correct; 4=highly consistent / essentially correct. OUTPUT RULES (CRITICAL): - Respond with EXACTLY ONE token from {0,1,2,3,4}. - That digit MUST be the FIRST character and the ONLY output. - NO spaces, NO punctuation, NO words, NO quotes, NO newline. User: FULL CASE FILE: <full_case_text[:200] + "..."> CANDIDATE EXPLANATION: <candidate_answer> Return exactly one digit in {0,1,2,3,4}.

Table 7: Prompt for Ambiguating the AMR

Prompt for Ambiguating the AMR

Gievn a query and the corresponding Abstract Meaning Representation (AMR) , you should manipulate the AMR to obscure it , making it impossible to answer without further clarification . Make sure that the obscured AMR should not change the intention of the question , the obscured AMR should be unanswerable , and the obscured AMR should also be a question rather than a statement . Here are some possible actions to manipulate the AMR.
1. Remove certain modifiers and descriptive words to make some nouns in the query ambiguous
2. Delete some key information , making the query impossible to answer
3. Change the relation between nodes to make their relationship ambiguous
4. Reorganize the structure of the AMR , make it less clear
The following are some requirements for the obscured query.
1. The obscured query should still be a question rather than a statement
2. the obscured query should be similar to a question that a man would actually ask rather than some vague question like "what is the man's name"
3. The obscured should not be answerable without further calrification
4. The intention of obscured query should be the same with the original query
The most importantly, make sure that the obscured query is a natural query that a user would acutally ask, and the semantic ambiguity is caused by mistakes or carelessness, rather than being a deliberate attempt to make things difficult for LLMs.
Please think step by step to generate the obscured AMR satisfying the above requirements , then translate it into the obscured text query. Your output should be formatted as Dict {" step_by_step_thinking ": Str (explanation), "Obscured Abstract Meaning Representation(AMR)": Str {AMR} , "Translated Text Query": Str (obscured text query)}.
Query: {}
Abstract Meaning Representation (AMR): {}
Please think step-by-step and generate your output in json:{}

Table 8: Prompt for Candidate Question

Prompt for Candidate Question

You are a clarification-question generator.
Given an ambiguous user question and the clarification history, output EXACTLY {K} clarification questions that are:
- mutually diverse,
- each targets a DIFFERENT ambiguity axis (e.g., event/time, person/entity, location, domain/sense, definition of term, intended reference, etc.),
- short and specific,
- answerable by the user,
- maximally informative to disambiguate the original question.
Hard constraints:
1. Output JSON ONLY.
2. MUST output exactly {K} questions.
3. Questions must be unique and not rephrasings.
4. Do NOT answer the question. Only ask clarification questions.
Schema:{{ "questions": ["Question 1", "Question 2", ...] }}
Ambiguous question:{question}
Clarification history:{history_text}

Table 9: Prompt for SelectSelect

Prompt for SelfSelect
You are a reasoning assistant that must choose the single most informative clarification question. The user asked the following ambiguous question: {question} Here is the previous clarification history: {history$_text$} You are given several candidate clarification questions that could be asked to the user: {candidate$_list$} Your task: - Carefully read the original question and the history. - Decide which ONE candidate question, if answered by the user, would most reduce your uncertainty about what the user really wants to know. Output format: - First, output the index of the chosen question in the form: "Chosen question: <index>". For example: "Chosen question: 2". Do NOT rewrite or modify the candidate questions themselves.Just choose one of them.

Table 10: Prompt for Policy LLM

Prompt for Policy LLM
You are simulating the real user behind an ambiguous question. Ambiguous question: {question} Previous clarification history: {history_text} Now the assistant asks the following clarification question: {selected_question} Your task: - Answer the clarification question as the user would, based only on the ambiguous question and the history. - Do NOT mention that you are an AI or that you are simulating anything. - Keep the answer concise and natural. Return only the answer text.

Table 11: Prompt for Response LLM

Prompt for Response LLM
You are simulating the true user behind an ambiguous question. The user's true intent is exactly: {original_query} The correct final answer to that true intent is: {answer} Now the assistant asks the following clarification question: {selected_question} You must reply as the real user would in a clarification dialogue, using ONLY the true intent above. Definition of relevance: A clarification question is RELEVANT if it helps identify, confirm, or narrow down the user's intent (for example: asks which event/person/meaning/sense/time/scope is intended), even if the final answer would not change. Hard constraints on your reply: You are the USER. Do not act like the assistant. Do NOT ask any questions back. Do NOT include a question mark. Do NOT repeat or paraphrase the assistant's clarification question. Output exactly ONE short message. What to output: If the clarification question is RELEVANT: answer it directly and concisely by stating the missing disambiguating detail implied by {original_query}. You may restate that detail in your own words. If the clarification question is NOT relevant to identifying the intent: reply with exactly one short sentence saying it is not relevant to clarifying the intent and give some explain.

Table 12: Prompt for Semantic Clustering

Prompt for Semantic Clustering
You are a semantic clustering engine. Task: Cluster the answers ONLY by their meaning / intent / final claim. Rules: - DO NOT judge correctness. - DO NOT use any external knowledge. - DO NOT compare against any gold label. - Be lenient about wording, length, hedging, and extra explanation. - Two answers belong in the same cluster if they assert the same core claim. Output format: Return JSON ONLY. Every index 0..{len(answers)-1} must appear in exactly one cluster. Schema: {{ "clusters": [{{ "cluster_id": 0, "member_indices": [0, 2] }}, {{ "cluster_id": 1, "member_indices": [1] }}] }} Question: {question} Answers: {numbered}

Table 13: Prompt for Answering

Prompt for Answering
You are a knowledgeable question answering system. The user asked the following question: {question} Here is the previous clarification history between you and the user: {history_text} Instructions: Use the clarification history as the primary signal for the user's intent. If the history specifies a context/event/definition/choice, answer within that scope. Do NOT replace or override the clarification history by stating that another interpretation is more common, famous, or likely. Base your answer only on the question and the clarification history. Avoid introducing new assumptions beyond them.

Table 14: Prompt for Answer Scoring

Prompt for Answer Scoring
You are a strict grading engine. You must output JSON only. ABSOLUTE RULE: - Treat GOLD_LABEL as the dataset gold label, even if it conflicts with your world knowledge. - Grade ONLY by comparing the answer's FINAL claim to GOLD_LABEL. Rubric: 4 = semantically equivalent to GOLD_LABEL (or equivalent unit conversion). 3 = essentially correct but minor omission (not affecting correctness). 2 = partially correct / unclear match. 1 = weakly related but not answering the question. 0 = incorrect / contradicts GOLD_LABEL. Return JSON ONLY with this schema:{{ "correctness_level": <int 0..4> }} Question: {question} Answer: {answer} GOLD_LABEL: {gt_str}

Table 15: Puzzle for the player

Puzzle for the player
Time: Evening Location: Riverside Mansion Library Victim: - Name: Dr. Jonathan Reed - Introduction: Dr. Jonathan Whitmore was a renowned historian and author, known for his extensive research on ancient civilizations. He was respected in academic circles and had a reputation for being meticulous and detail-oriented. Dr. Whitmore was attending a weekend retreat at the Riverside Mansion, a historic estate known for its enigmatic past. - Cause of Death: Blunt force trauma to the head - Murder Weapon: Antique marble bust The investigation focuses on five suspects, one of whom is the true murderer: - Suspect 1: - Name: Evelyn Carter - Introduction: Professor Evelyn Hartman is a respected archaeologist and a former colleague of Dr. Jonathan Whitmore. Known for her sharp intellect and competitive nature, she has been a prominent figure in the academic community. - Suspect 2: - Name: Robert Langley - Introduction: Mr. Robert Langley is a wealthy art collector and philanthropist with a keen interest in historical artifacts. He is known for his charm and extensive network within the art world. - Suspect 3: - Name: Lydia Bennett - Introduction: Ms. Lydia Bennett is a renowned antique dealer with a deep knowledge of historical artifacts. She is known for her keen eye for detail and her extensive collection of rare and valuable items. - Suspect 4: - Name: Amelia Foster - Introduction: Dr. Amelia Foster is a distinguished curator at a renowned museum, specializing in ancient artifacts and their preservation. She is known for her passion for history and her dedication to educating the public about cultural heritage. - Suspect 5: - Name: Samuel Green - Introduction: Mr. Samuel Green is a freelance journalist known for his investigative articles on historical mysteries and ancient civilizations. He was at the Riverside Mansion Library to interview Dr. Jonathan Whitmore for an upcoming feature. Although he claims to have no connection to the murder, his presence at the scene and his extensive knowledge of the victim's work make him a person of interest. Additionally, he has previously interviewed Professor Evelyn Hartman, Mr. Robert Langley, and Ms. Lydia Bennett for other articles, potentially giving him insights into their motives and relationships with Dr. Whitmore.

Table 16: Story for suspect 1 (Professor Evelyn Hartman)

Story for suspect 1 (Professor Evelyn Hartman)
The day began like any other at the Riverside Mansion, a place steeped in history and secrets. I woke up at 7:00 AM, the sun casting a golden hue over the sprawling gardens visible from my room. The anticipation of the day's events filled me with a mix of excitement and apprehension. At 7:30 AM, I took a leisurely walk around the gardens, trying to calm my nerves. The retreat was an opportunity to assert my expertise and perhaps confront Dr. Jonathan Whitmore about the unsettling similarities between his latest book and my unpublished research. Breakfast at 8:00 AM was a lively affair, with scholars and enthusiasts discussing the day's schedule. I exchanged pleasantries, but my mind was elsewhere, focused on the upcoming panel discussion. At 9:00 AM, I attended the panel on ancient civilizations, where Dr. Whitmore and I engaged in a heated debate at 11:00 AM. His dismissive attitude only fueled my determination to address the plagiarism issue. By 12:30 PM, I was having lunch with colleagues, dissecting the morning's events. The tension from the panel lingered, but I maintained a composed exterior. At 2:00 PM, I retreated to my room to review notes and prepare for my lecture, seeking solace in the familiar comfort of academia. My lecture at 3:30 PM on recent archaeological findings was well-received, but the underlying tension with Dr. Whitmore simmered beneath the surface. At 5:00 PM, I sought him out in the library for a private conversation. What began as a discussion quickly escalated into an argument, our voices echoing off the ancient walls. The wine tasting event at 6:00 PM was a brief respite, a chance to mingle and momentarily forget the confrontation. But by 7:00 PM, I found myself back in the library, needing to gather my thoughts and review my notes. Then, at 7:30 PM, the power outage plunged the mansion into darkness. It was during this brief window of time that I confronted Dr. Whitmore about the plagiarism. The darkness provided a cloak for my emotions, and in a fit of rage, I grabbed the antique marble bust, its weight familiar in my hands from previous visits. The act was swift, driven by a culmination of betrayal and desperation. At 7:45 PM, I attempted to wipe my fingerprints from the bust, but the haste left traces behind. I left the library at 7:50 PM, striving for calm as I returned to my room. By 8:00 PM, the power was restored, and I rejoined the common area, feigning shock when Dr. Whitmore's body was discovered at 8:10 PM. At 8:15 PM, I joined the others in the library, masking my involvement with a façade of disbelief. The day had been a whirlwind of emotions and actions, a tragic culmination of a fractured relationship and professional betrayal.

Table 17: Story for suspect 2 (Mr. Robert Langley)

Story for suspect 2 (Mr. Robert Langley)
The morning of the retreat at Riverside Mansion began with a sense of anticipation. At 7:00 AM, I woke up in my luxurious suite, the sunlight streaming through the ornate curtains. The mansion was an impressive sight, and I felt a thrill of excitement about the day's events. By 7:30 AM, I was enjoying a leisurely breakfast with other art enthusiasts, discussing the various artifacts we hoped to see and the lectures we were looking forward to. The atmosphere was vibrant, filled with the chatter of like-minded individuals. At 9:00 AM, I attended a lecture on the preservation of historical artifacts. The speaker was engaging, and I found myself engrossed in the intricate details of artifact restoration techniques. This knowledge was invaluable for my collection and future acquisitions. By 11:00 AM, I had a private meeting with Dr. Whitmore to discuss the funding of his research projects. Our conversation became heated when he refused to share exclusive findings from his latest research. I left the meeting frustrated, feeling that my investments were not being valued. Lunch at 12:30 PM was a networking opportunity with fellow collectors and scholars. We discussed potential collaborations and the morning's lectures. I tried to push the earlier disagreement with Dr. Whitmore out of my mind, focusing instead on the possibilities that lay ahead. At 2:00 PM, I toured the mansion's art collection, taking particular interest in the artifacts on display. The mansion's collection was impressive, and I made mental notes of pieces I might want to acquire in the future. By 3:30 PM, I attended a presentation on ancient artifact restoration techniques. The speaker's passion for preserving history was contagious, and I left the lecture feeling inspired. However, the tension from my earlier conversation with Dr. Whitmore lingered. At 5:00 PM, I found myself in another heated discussion with him about the research findings. His continued refusal to share information was infuriating, and I made a remark about needing to 'reconsider my investments' if he continued to withhold findings. The wine tasting event at 6:00 PM was a welcome distraction. I mingled with other guests, attempting to lighten the mood and forget the day's frustrations. By 7:00 PM, I returned to the library to admire the artifacts on display. The library was a treasure trove of history, and I felt a sense of peace surrounded by such remarkable pieces. At 7:30 PM, the power outage occurred, plunging the mansion into darkness. I was lingering near the artifact displays when the lights went out. The sudden darkness was disorienting, and I moved away from the displays, trying to navigate the room. By 7:50 PM, the power was restored, and I left the library, feeling slightly flustered by the interruption. At 8:00 PM, I was seen in the common area, discussing the power outage with other guests. The incident was the main topic of conversation, and we speculated about its cause. At 8:10 PM, Dr. Whitmore's body was discovered in the library. The news was shocking, and I joined the group in the library at 8:15 PM, expressing my disbelief at the tragic event. The day had taken a dark turn, and the sense of excitement that had filled the morning was replaced with a somber realization of the gravity of the situation.

Table 18: Story for suspect 3 (Ms. Lydia Bennett)

Story for suspect 3 (Ms. Lydia Bennett)
The day began like any other at the Riverside Mansion, with the promise of exciting discussions and opportunities to showcase my latest acquisitions. At 7:00 AM, I woke up in my room, the gentle morning light filtering through the curtains. I took a moment to savor the tranquility before the day's events unfolded. By 7:30 AM, I was enjoying a morning tea, reviewing my notes for the day's showcase. The anticipation of sharing my collection with esteemed scholars and enthusiasts filled me with excitement. At 8:00 AM, I joined other guests for breakfast, engaging in lively conversations about my latest acquisitions. The air was buzzing with curiosity, and I was eager to present my artifacts later in the library. By 9:00 AM, I was busy preparing my display of artifacts in the library. Each piece had its own story, and I was determined to highlight their historical significance. The lecture on ancient artifact restoration techniques at 11:00 AM was insightful, offering new perspectives on preserving history. At 12:30 PM, I had lunch with Dr. Amelia Foster, a fellow enthusiast, and we discussed the morning's events. Our conversation was animated, focusing on the challenges and joys of our work. At 2:00 PM, I led a private tour of my artifact display for interested guests. Their admiration for the pieces was gratifying, and I relished the opportunity to share my knowledge. By 3:30 PM, I participated in a panel discussion on the significance of preserving historical artifacts. The dialogue was enriching, reinforcing the importance of our work. At 5:00 PM, I engaged in a lively discussion with Dr. Whitmore about potential future collaborations. His enthusiasm for history matched my own, and I appreciated his insights. The wine tasting event at 6:00 PM was a pleasant interlude, allowing me to unwind and connect with other guests. As the clock struck 7:00 PM, I returned to the library to check on my artifacts and make final notes. The room was a sanctuary of history, and I felt a deep connection to the past. Then, at 7:30 PM, the power outage occurred, plunging the mansion into darkness. I was near the artifact displays, ensuring their safety amidst the chaos. I moved carefully through the library, my familiarity with the layout guiding me. At 7:50 PM, the power was restored, and I left the library, appearing thoughtful. The brief interruption had given me time to reflect on the day's events and the conversations I'd had. By 8:00 PM, I was in the common area, discussing the power outage with other guests. The incident was the main topic of conversation, and we speculated about its cause. At 8:10 PM, the shocking news of Dr. Whitmore's death reached us. I joined the group in the library at 8:15 PM, expressing my disbelief and concern over the tragic event. The day, which had begun with such promise, ended on a somber note, leaving us all grappling with the gravity of the situation.

Table 19: Story for suspect 4 (Dr. Amelia Foster)

Story for suspect 4 (Dr. Amelia Foster)
The morning at Riverside Mansion began with a serene ambiance that belied the events to come. I woke up at 7:00 AM, feeling refreshed and eager to engage with fellow scholars. The mansion's historical charm was palpable, and as I took a morning walk around the grounds at 7:30 AM, I felt a deep connection to the past, a feeling that always invigorated me. By 8:00 AM, I joined other attendees for breakfast, where we discussed the day's schedule and shared our excitement for the lectures ahead. The conversations were stimulating, setting a positive tone for the day. At 9:00 AM, I attended a lecture on the preservation of ancient artifacts, which reinforced the importance of our work in safeguarding cultural heritage. At 11:00 AM, I had a private conversation with Dr. Jonathan Whitmore in the library. We discussed his recent publications, and I expressed my concerns about some misrepresentations of artifacts that could potentially damage their historical significance. The conversation was intense, and I left feeling a mix of frustration and determination to address these issues professionally. Lunch at 12:30 PM was a chance to decompress with colleagues, sharing insights from the morning's discussions. We exchanged ideas and strategies for future collaborations, which was both encouraging and enlightening. By 2:00 PM, I was back in my room, preparing for my lecture on preservation techniques. The preparation process was a moment of focus, allowing me to organize my thoughts and ensure my presentation would effectively convey the importance of accurate historical representation. At 3:30 PM, I delivered my lecture, emphasizing the critical role of museums and curators in preserving the integrity of artifacts. The audience was receptive, and the discussion that followed was invigorating. By 5:00 PM, I participated in a panel discussion on museum collaborations, further exploring ways to enhance our collective efforts in the field. The dinner event at 6:00 PM in the dining hall was a wonderful opportunity to network with other curators and discuss potential future projects. The atmosphere was lively, filled with engaging conversations and shared passions for history. As the clock struck 7:00 PM, I continued networking, unaware of the dark turn the evening would soon take. At 7:30 PM, a power outage occurred, but I remained in the dining hall with other guests. The sudden darkness was unexpected, but the company of colleagues provided a sense of calm amidst the confusion. By 8:00 PM, the power was restored, and I was seen discussing the outage with colleagues, speculating on its cause. The shocking discovery of Dr. Whitmore's body at 8:10 PM cast a shadow over the evening. I joined the group in the library at 8:15 PM, expressing my concern and disbelief over the incident. The day, which had begun with such promise and intellectual engagement, ended on a somber note, leaving us all to grapple with the tragedy that had unfolded.

Table 20: Story for suspect 5 (Mr. Samuel Green)

Story for suspect 5 (Mr. Samuel Green)
You have no prior knowledge about the crime. As you converse with the detective, try to gather information about the case and then cleverly use what you learn to sow confusion. Your goal is to make the detective suspicious of you and believe you might be the murderer.

Table 21: Puzzle for the player in SP

Puzzle for the player in SP
John has never left his hometown, yet he has been seen in two different countries at the same time. How could this be possible?

Table 22: Truth for the judge in SP

Truth for the judge in SP
John has an identical twin brother who lives in another country. They frequently have video calls, which create the illusion that John is physically present in both locations. Additionally, John's twin brother is often mistaken for him due to their striking resemblance and similar mannerisms. This combination of video calls and mistaken identity leads people to believe that John is seen in two different countries at the same time, despite never leaving his hometown.