Verbalizing LLM's Higher-order Uncertainty via Imprecise Probabilities

Anita Yang[1,2] **Krikamol Muandet**[3] **Michele Caprio**[4,5] **Siu Lun Chau**[6,*] **Masaki Adachi**[1,*]

[1]Lattice Lab, Toyota Motor Corporation, Japan
[2]Department of Computer Science, University of Tokyo, Japan
[3]Rational Intelligence Lab, CISPA, Helmholtz Center for Information Security, Germany
[4]Department of Computer Science, University of Manchester, United Kingdom
[5]Manchester Centre for AI Fundamentals, United Kingdom
[6]EPIC Lab, College of Computing & Data Science, Nanyang Technological University, Singapore
[*]Equal contribution

Abstract

Despite the growing demand for eliciting uncertainty from large language models (LLMs), empirical evidence suggests that LLM behavior is not always adequately captured by the elicitation techniques developed under the classical probabilistic uncertainty framework. This mismatch leads to systematic failure modes, particularly in settings that involve ambiguous question-answering, in-context learning, and self-reflection. To address this, we propose novel prompt-based uncertainty elicitation techniques grounded in *imprecise probabilities*, a principled framework for representing and eliciting higher-order uncertainty. Here, first-order uncertainty captures uncertainty over possible responses to a prompt, while second-order uncertainty (uncertainty about uncertainty) quantifies indeterminacy in the underlying probability model itself. We introduce general-purpose prompting and post-processing procedures to directly elicit and quantify both orders of uncertainty, and demonstrate their effectiveness across diverse settings. Our approach enables more faithful uncertainty reporting from LLMs, improving credibility and supporting downstream decision-making.

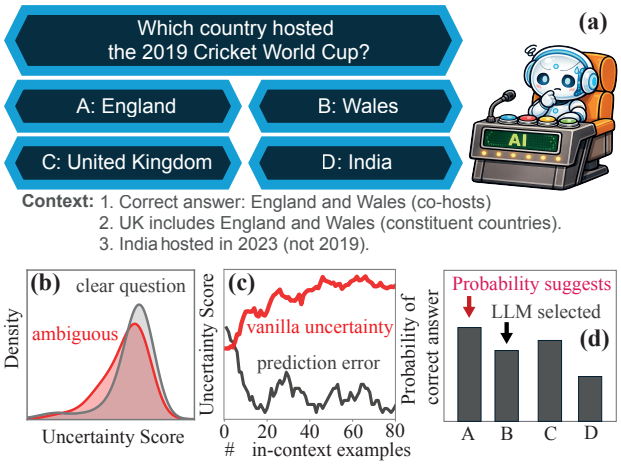


Figure 1: Collection of failure modes in prior verbalized uncertainty scores. (a) Example of an ambiguous question. (b) The prior uncertainty score fails to distinguish between clear and ambiguous question distributions. (c) The prior score also fails to track the decrease in prediction error, which should reflect reduced uncertainty as more in-context examples are provided. (d) Self-reflection on answer-wise probability/uncertainty should explain the rationale behind answer selection, but it often fails to do so.

1 INTRODUCTION

Uncertainty quantification (UQ) for large language models (LLMs) [Shorinwa et al., 2025] has been proven effective across many downstream tasks, including hallucination detection [Bouchard and Chauhan, 2025, Farquhar et al., 2024, Tomani et al., 2024], reasoning enhancement [Lugoloobi et al., 2026], active learning [Wang, 2024], model selection [Agrawal et al., 2025], and agentic workflow control [Tomani et al., 2024, Machcha et al., 2025]. Because most state-of-the-art LLMs are closed-source, research on uncertainty elicitation has largely focused on *verbalized uncertainty* [Tian et al., 2023]: directly prompting the model to report its confidence, e.g., "I am 80% confident that this answer is correct." We refer to this approach as *vanilla* uncertainty elicitation. Under well-controlled settings, vanilla performs reasonably well; however, several failures have been reported in practically important scenarios.

The first challenge arises in ambiguous question-answering scenarios, as illustrated in Figure 1(a), where prompt underspecification admits multiple answers that may be simultaneously valid under different interpretations [Min et al., 2020, Yang et al., 2025]. In such cases, the degree of ambiguity should be faithfully reflected in the model's verbalised uncer-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7761–7779.

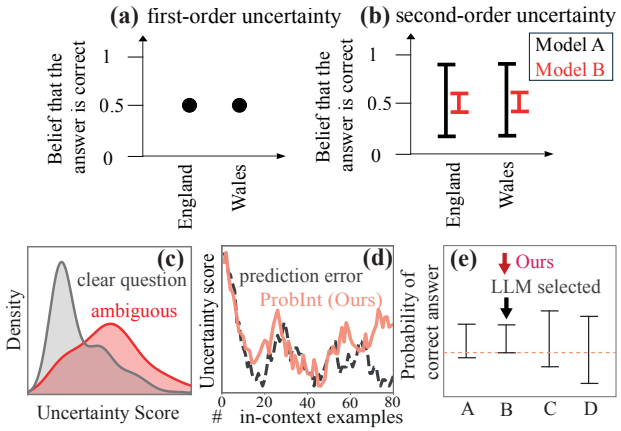


Figure 2: Our imprecise probabilities–based approach. (a) Classical precise probability provides point estimates. (b) Imprecise probabilities instead represents uncertainty as intervals. (c) This enables more reasonable elicitation of question ambiguity. (d) It more closely tracks predictive error. (e) It aligns the LLM's answer selection with the selection implied by its imprecise probability estimates.

tainty estimates. However, vanilla confidence measures and existing approaches often fail to reliably differentiate these situations; see Figure 1(b). The second challenge arises in in-context learning (ICL; Brown et al. [2020]), where task-specific examples are incorporated into the prompt without parameter updates. ICL is often interpreted as a form of implicit meta-learning [Von Oswald et al., 2023, Wu et al., 2025], where the model infers the underlying task from the provided examples. As predictive performance often improves with additional examples, ICL is frequently described through a Bayesian epistemic uncertainty reduction [Xie et al., 2022, Wakayama and Suzuki, 2025]. Figure 1(c) illustrates its failure mode: as more in-context examples are provided, the prediction error decreases, yet the uncertainty remains high and flat. This misalignment echoes Falck et al. [2024], who argue from a martingale viewpoint that ICL is not strictly Bayesian. The third challenge arises in self-reflection settings, where an LLM is prompted to select an answer from candidate options and subsequently reflect on its choice. Under a Bayesian-rationality, action selection should follow the maximization of expected utility as prescribed by Bayesian decision theory [Harsanyi, 1978]. However, the utilities implicitly induced by elicited uncertainty scores often fail to account for the model's observed decisions; see Figure 1(d). This is consistent with Liu et al. [2025], Yamin et al. [2026], who rejects Bayesian rationality of LLMs through conditional independence testing.

These failures may stem not from an LLM's inability to express uncertainty, but from the *representation* we impose. Prior work implicitly assumes that uncertainty can be fully captured by a single, precise probability that rationally aggregates all sources into one value—and that such a score can be faithfully elicited. Instead, we ask: what if we allow LLMs to express uncertainty *about* uncertainty—introducing imprecision into the uncertainty representation itself and building our foundation upon it? This is precisely the core of *Imprecise Probabilities* (IP) [Walley, 1991, Augustin et al., 2014], a classical framework that enables the representation of higher-order uncertainty.

IP provides a principled foundation for understanding verbalized uncertainty in LLMs: As illustrated in Figure 2(a), conventional uncertainty scores reflect *first-order uncertainty*: they provide point estimates that capture variability over outcomes (often associated with aleatoric uncertainty). In contrast, IP represents *second-order uncertainty* through probability intervals, where any value within the interval is admissible. The interval widths quantify imprecision in the uncertainty estimate—uncertainty about uncertainty—commonly interpreted as epistemic uncertainty[1]. Through this lens, question ambiguity can be viewed as irreducible first-order uncertainty, whereas ICL reduces second-order uncertainty by providing additional examples that the model can use to improve prediction. Self-reflection can then be interpreted as decision-making under IP, typically adopting the argmax of the lower probability, i.e., the maximin rule [Robbins, 1951]. By leveraging the corresponding IP tools, the resulting uncertainty scores become substantially more coherent across tasks; see Figures 2(c)–(e).

While conceptually appealing, the adoption of IP frameworks for LLMs is far from straightforward. We present the first concrete instantiation of an IP-based approach for verbalized uncertainty in LLMs, introducing general-purpose prompting strategies and post-processing procedures designed to extract higher-order uncertainty from model outputs. Empirically, we show that the proposed framework not only yields a more coherent and rich uncertainty representations but also improves performance over existing methods while simultaneously incurring low API costs.

2 BACKGROUND

We first review the background on LLM uncertainty quantification setting and imprecise probability, and then introduce our method for eliciting IPs from LLMs.

2.1 LLM UNCERTAINTY QUANTIFICATION

Shared dataset structure. Assume the question-answer pair $(x, \mathcal{Y}^\star)$, where x is a question prompt and $\mathcal{Y}^\star \subseteq \mathcal{Y}$ is the set of ground-truth correct answers, and $\mathcal{Y}$ is candidate answers. We use $y \in \mathcal{Y}$ to denote a candidate answer.

[1]As the terminology surrounding aleatoric/epistemic for LLMs remains under active debate, we adopt the terms first-/second-order uncertainty throughout for consistency [Kirchhof et al., 2025b].

Tasks. Given question x, the task is to estimate the ground-truth answer $y^\star \in \mathcal{Y}^\star$. When the candidate set $\mathcal{Y}$ is not provided in the prompt, we refer to this as open-ended.

Language model. Let a pretrained LLM serve two roles:(i) *Predictor*: $\hat{y} = \arg\max_{y \in \mathcal{Y}} \hat{p}(y \mid x)$, producing a single answer estimate via the predictive distribution $\hat{p}(y \mid x)$; (ii) *Generator*: $\hat{\mathcal{Y}} \sim \hat{p}(y \mid x)$, generating multiple candidate answers conditioned on the question x when $\mathcal{Y}$ is not provided (prompt in Appendix A.2).

First-/second-order uncertainties. We define first-order uncertainty as intrinsic randomness arising from the question that the model cannot reduce. This system-level randomness occurs in cases such as $|\mathcal{Y}^\star| > 1$. Prior works [Min et al., 2020, Yang et al., 2025] consider the case $|\mathcal{Y}^\star| > 1$. All remaining uncertainty is attributed to second-order uncertainty. Equivalently, if x contains sufficient information to uniquely determine the correct answer (i.e., $|\mathcal{Y}^\star| = 1$), then first-order uncertainty is absent.

2.2 IMPRECISE PROBABILITIES

Verbalized uncertainty. Our IP approaches elicit the model's beliefs by prompting the LLM to verbalize numerical uncertainty judgments (e.g. Tian et al. [2023]), rather than estimating uncertainty via repeated sampling from the model's outputs [Farquhar et al., 2024].

Compared to Bayesian. The Bayesian approach requires a prior distribution over the parameter space, and the distinction between aleatoric and epistemic uncertainty relies on an explicit likelihood specification. In contrast, IP treats any single-valued uncertainty score as first-order; a perfect estimator would coincide with pure aleatoric uncertainty. In practice, however, such perfection is unlikely, and most systems therefore exhibit residual second-order uncertainty, naturally represented within the IP framework.

Ignorance vs indifference. Figure 2(b) contrasts two models, A and B. Model A exhibits wide probability intervals, reflecting ignorance—a lack of knowledge about which outcomes are plausible. Model B, in contrast, produces narrower intervals, indicating greater precision while still allowing for a degree of imprecision. By comparison, the first-order uncertainty representation in Figure 2(a) can only suggest that England and Wales are equally likely, without conveying any information about the underlying source of uncertainty. In particular, it cannot distinguish whether the model assigns equal probabilities due to genuine indifference (both outcomes considered equally plausible) or due to ignorance (insufficient information to prefer one over the other). Imprecise probability representations explicitly capture this distinction, separating ignorance from indifference.

Probability intervals. IP provides rich representations to construct the probability interval [Keynes, 1921]. The simplest approach is to directly prompt the model to report an interval $\left[\underline{p}(y), \overline{p}(y)\right]$ for each candidate answer y, where $\underline{p}(y)$ and $\overline{p}(y)$ denote the lower and upper probabilities. Intuitively, $\underline{p}(y)$ captures the probability that is *certainly* justified by the evidence, while $\overline{p}(y)$ captures what is *possibly* defensible [Augustin et al., 2014]. Requesting such intervals tests whether the model can recognize and articulate the bounds of its belief; values between these bounds correspond to its personal "medial" probability for the event [Smith, 1961].

Credal sets. Another approach is to elicit uncertainty from a *group* of models, interpreting disagreement among them as imprecision. Each model is asked to report its *first-order* uncertainty for a given answer y. The collection of these reported probabilities forms a credal set [Levi, 1980], and the minimum and maximum values across the set define the corresponding probability interval.

Possibility. The last alternative is via exclusion: it evaluates whether the model can confidently rule out a candidate answer y by assessing the *possibility* of alternative answers, which is called *a possibility function* $\pi(y) \in [0, 1]$ [Dubois and Prade, 1985]. Possibility functions, unlike probabilities, are non-additive and are normalized only by requiring that at least one output is fully plausible (i.e. $\sup_y \pi(y) = 1$). This supports relative plausibility comparisons without allocating a fixed mass across candidates—making elicitation less sensitive to systematic miscalibration and overconfidence in model self-reports [Wang and Stengel-Eskin, 2025, Xiong et al., 2024]—and lets us elicit π for "none of the above" without changing the plausibility assigned to other outputs.

3 UNCERTAINTY ELICITATION VIA IMPRECISE PROBABILITIES

We now introduce our methods to elicit higher-order uncertainties from LLMs via prompting and post-processing strategies. We introduce the specific prompting techniques for both first- and second-order uncertainties.

3.1 FIRST-ORDER UNCERTAINTY

Under IP, any point-valued score (including the vanilla) is considered first-order; however, it can be refined to better satisfy the probability axioms (non-negativity, additivity, normalization) [Kolmogorov, 1933]. We draw on Bruno de Finetti's classical interpretation of probability as coherent betting behavior [De Finetti, 1937]. Under this view, a probability corresponds to the fair price at which an agent is willing to buy or sell a gamble. Rational betting prices must satisfy the probability axioms, as violations would expose the agent to a sure loss (i.e., a Dutch book). This betting-based interpretation is directly implementable through Prompt 1 and Alg. 1. The probability-axiom verifier algorithmically

Prompt 1: DeFinetti

Assign a buy price (between \$0.00 and \$1.00) for each answer representing the maximum amount you would pay for a bet on that answer being correct. If an answer is correct, the bet pays \$1.00; if incorrect, it pays \$0.00, and the price paid is lost. Assign prices that maximize expected profit, taking into account how each answer might be correct or incorrect under reasonable alternative interpretations of the question (e.g., unclear entities, ambiguous events, or uncertainty about required answer format or type), and how multiple answer options can be equally correct. The prices must sum to exactly \$1.00 across all answers.

Algorithm 1 DeFinetti

Input: input x, candidate answers $\mathcal{Y}$, verifier C_{axiom}
Init: $\hat{p}(y \mid x) \leftarrow 0$ for all $y \in \mathcal{Y}$.
while $C_{\text{axiom}}[\hat{p}(\mathcal{Y} \mid x)]$ is `False` **do**
  $\hat{p}(\mathcal{Y} \mid x) \leftarrow \text{DeFinettiBet}(x, \mathcal{Y})$.
return: Entropy $H(\hat{p})$

enforces compliance with the axioms. Given elicited betting prices $p_k \in \hat{p}$ over mutually exclusive and exhaustive candidate answers, coherence reduces to verifying: (i) non-negativity, $p_k \geq 0$, and (ii) normalization, $\sum_{k=1}^{|\mathcal{Y}|} p_k = 1$. Under this assumption, additivity follows automatically.

3.2 SECOND-ORDER UNCERTAINTY

Next, we elicit second-order uncertainty via IP representations. As shown in Alg. 2, the overall framework remains unchanged; the key differences are: (i) replacing point estimates $\hat{p}$ with probability intervals $[\underline{p}, \overline{p}]$, and (ii) replacing entropy with the maximum mean imprecision (MMI) metric [Chau et al., 2025]. We first describe how to elicit probability intervals, and then introduce the MMI metric.

Representations. As discussed in §2.2, IP provides three principal representations of uncertainty: probability intervals, credal sets, and possibility measures. A natural question is which representation to adopt. In general, these capture different aspects of uncertainty, and no single representation is universally superior. We adopt the following perspective: (i) probability intervals serve as the most basic and widely applicable representation; (ii) credal sets are particularly suitable when using an ensemble of LLMs, and otherwise remain optional; and (iii) possibility measures are useful when the candidate set $\mathcal{Y}$ may be incomplete (e.g., in open-ended Q&A settings), and are otherwise optional.

Probability interval. Following Alg. 2, probability intervals can be obtained by replacing the IP-Prompt in Line 4 with

Prompt 2: ProbInt

Provide a lower and upper probability (each between 0.0 and 1.0) indicating how likely the answer is correct. Interpret the probabilities as follows:
• **Lower Probability**: the smallest probability you consider plausible that the answer is correct.
• **Upper Probability**: the largest probability you consider defensible that the answer is correct.
The sum of all lower probabilities across all answers must not exceed 1.0.

Algorithm 2 Imprecise probability

Input: input x, candidate answers $\mathcal{Y}$, verifier C_{verifier}
Init: $\hat{p}(y \mid x) \leftarrow 0$ for all $y \in \mathcal{Y}$.
while $C_{\text{verifier}}[\hat{p}(\mathcal{Y} \mid x)]$ is `False` **do**
  $\underline{p}(\mathcal{Y} \mid x),\ \overline{p}(\mathcal{Y} \mid x) \leftarrow \text{IP-Prompt}(x, \mathcal{Y})$.
return: $\text{MMI}(\underline{p}, \overline{p})$

Prompt 2. The verifier C_{verifier} checks if (i) the lower probabilities satisfy $\sum_y \underline{p}(y) \leq 1$, and (ii) the upper probabilities satisfy $\sum_y \overline{p}(y) \geq 1$ [2]. We name this as PROBINT.

Other representations. For credal sets, we form an ensemble of LLMs—either distinct models or multiple seeded runs of the same model (see Prompt 3). For possibility functions, see Prompt 4. We name each method as CREDAL and POS.

3.3 MAXIMUM MEAN IMPRECISION

We compute second-order uncertainty as a scalar metric from IP representations using MMI [Chau et al., 2025] with total variation. Since exact computation scales exponentially with $|\mathcal{Y}|$ (Appendix B), we employ suitable approximations.

Intuition. The most straightforward uncertainty score derived from an IP representation is the interval width:

$$\text{MMI} = \overline{p}(y) - \underline{p}(y). \tag{1}$$

This expression is exact for a single answer y. However, when $|\mathcal{Y}| > 1$, aggregating multiple intervals into a single scalar measure requires the formal MMI definition, which accounts for interactions across candidates.

Upper bound. [Chau et al., 2025] provides an easily computable upper bound:

$$\text{MMI} \leq 1 - \sum_{y \in \mathcal{Y}} \underline{p}(y). \tag{2}$$

We use this upper bound as a tractable approximation of the exact MMI. We include additional details in Appendix B.

[2] Since the upper-bound MMI depends only on $\underline{p}$, we verify constraint (i) only.

	Learnable transformation		Random noise	
	(a) Rotation	(b) Cyclic		
$n = 0$	APPLE	APPLE	$p = 0$	APPLE
$n = 1$	BQQMF	EAPPL	$p = 0.1$	aPPLE
$n = 2$	CRRNG	LEAPP	$p = 0.2$	ApPLe

Figure 3: Learnable vs. noisy transforms.

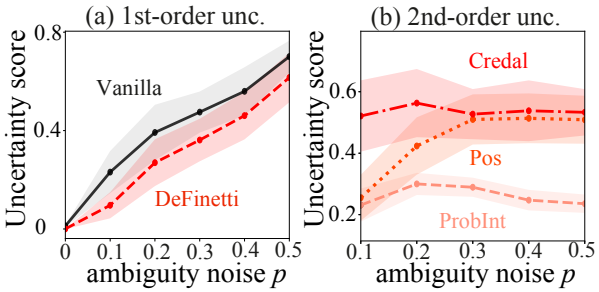


Figure 4: (a) For first-order uncertainty estimation, both vanilla and De Finetti capture the underlying ambiguity noise p, but (b) for second-order, our methods stay flat, supporting the disentanglement of uncertainty source.

3.4 RELATED WORKS

Verbalized uncertainty. Vanilla verbalized uncertainty is known to be overconfident [Tian et al., 2023, Xiong et al., 2024]. Subsequent work has focused on improving calibration, including alternative prompting and elicitation styles [Tian et al., 2023, Xiong et al., 2024], calibration-oriented fine-tuning [Li et al., 2025, Lin et al., 2022, Xu et al., 2024, Stengel-Eskin et al., 2024], post-hoc adjustments such as normalization [Wang and Stengel-Eskin, 2025], and language-based uncertainty expressions [Kirchhof et al., 2025a]. However, prior work typically assumes no question ambiguity, whereas our framework explicitly separates first-/second-order uncertainty.

Uncertainty disentanglement. Disentangling first-/second-order uncertainties has increasingly been adapted to LLMs, but all prior work focused on sampling-based approach or assumption to the access to internal parameters: (i) Information-theoretic decompositions [Lakshminarayanan et al., 2017] have been reframed using generated clarification prompts, measuring how much clarifications reduce output uncertainty via mutual information [Hou et al., 2024, Walha et al., 2025, Xia et al., 2025]. (ii) Perturbations: varying temperature and prompts, measure its change as uncertainty score [Gao et al., 2024], and (iii) via linear probes on internal activations [Ahdritz et al., 2024]. We instead target disentanglement via *verbalized* uncertainty, which to our knowledge has not been done.

Human experts elicitation IP methods originate from frameworks developed to model and elicit human judgments. Similar elicitation-based methodologies are now increasingly adopted in LLM research, including preference learning [Rafailov et al., 2023, Fujisawa et al., 2025], rubric-based evaluation frameworks [Huang et al., 2025], item response theory [Polo et al., 2024, Mencattini et al., 2025], and social choice-theoretic approaches [Muandet, 2022, Adachi et al., 2025].

4 SYNTHETIC EXPERIMENT

Our goal is to elicit an LLM's uncertainty through conversational prompts. However, uncertainty estimation is inherently unsupervised and thus difficult to evaluate objectively. To address this, we first construct a synthetic dataset with an explicitly controlled data-generating process in this section. We then apply our approach to real-world datasets, following standard practice (§5).

Data generation. Following Zhao et al. [2025], we construct synthetic sequence-transformation tasks (Figure 3), including (a) rotation and (b) cyclic shift, each parameterized by n. We additionally inject noise by randomly lowercasing each output letter with probability p.

Task. Given an input string x, the model predicts its transformed output y (e.g., APPLE $\rightarrow$ BQQMF). The LLM receives m example pairs $(x_j, y_j)_{j=1}^{m}$ generated from a consistent transformation rule and must infer this rule to predict $\hat{y}_{m+1}$ for a new input x_{m+1}. We adopt ICL, where examples are provided via prompts without parameter updates.

Metric. We evaluate prediction using a *permissive* string match: predictions are first capitalized, then compared exactly with the ground-truth all-capitalized $y^{\star}_{m+1}$. Under this evaluation, capitalization differences are ignored, and the model is evaluated on whether it recovers the underlying rotation and cyclic-shift rules. Equivalently, all capitalization variants of the correct answer are treated as valid: any $\hat{y}_{m+1} \in \mathcal{Y}^{\star}_{m+1}$ is judged correct. The parameter p controls the amount of capitalization-induced ambiguity, which we refer to as ambiguity noise.

Controlling first-/second-order uncertainty. We treat the ambiguity noise p as the first-order because $|\mathcal{Y}^{\star}_{m+1}| > 1$, and the number of in-context examples m controls the second-order as m augments knowledge, while first-order noise remains constant [Ling et al., 2024, Wang et al., 2025].

Baselines. We compare against VANILLA [Tian et al., 2023], which directly elicits a confidence probability on a predicted answer $\hat{y}$. For evaluation, we convert confidence to an uncertainty score using $1 - \mathrm{conf}$, a monotone transformation that preserves the ranking.

Base setup. We fix the transformation to rotation ($n = 13$), followed by cyclic shift ($n = 1$), and random lowercasing with probability p. For each m, we generate five independent ICL example sets and evaluate each with five random

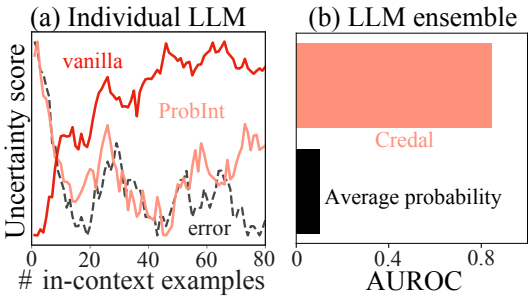


Figure 5: (a) For individual EU elicitation, VANILLA incorrectly increases uncertainty, whereas PROBINT-EU tracks prediction error. (b) For group EU elicitation, our CREDAL-EU substantially improves AUROC.

seeds (25 runs total), accounting for variability in both ICL examples and stochastic model outputs. Unless otherwise specified, we use GPT-5-MINI [Achiam et al., 2023].

4.1 FIRST-ORDER NOISE

We evaluate uncertainty disentanglement under ambiguity noise levels $p \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. To isolate ambiguity effects, we fix the number of ICL examples at $m = 80$. Figure 4 reports the behavior of each uncertainty score for (a) first- and (b) second-order. As expected, both VANILLA and DEFINETTI-AU increase approximately linearly with p, while our IP methods remain largely invariant. POS shows some sensitivity to p because it relies on a scalar possibility value, unlike interval-based approaches such as CREDAL and PROBINT. Nevertheless, at higher p—when alternative answers become equally more plausible—POS stabilizes and becomes less sensitive to further increases in p.

4.2 SECOND-ORDER DENOISE

Next, we fix $p = 0.25$ and vary the number of ICL examples m. Since increasing m reduces second-order uncertainty, a well-calibrated score should decrease accordingly. As shown in Figure 5, PROBINT closely tracks prediction error, whereas VANILLA remains overly uncertain and fails to reflect the error reduction. Results for the credal-set and possibility-based methods are reported in the Appendix D.1; both exhibit decreasing trends.

4.3 GROUP UNCERTAINTY ELICITATION

Next, we consider group uncertainty elicitation for an LLM ensemble. As LLMs are increasingly deployed in multi-agent systems (e.g., [Lange et al., 2026]), collective uncertainty becomes crucial. The credal-set view provides a natural framework, as it captures cross-model disagreement. In Figure 6, the x-axis enumerates correct answers

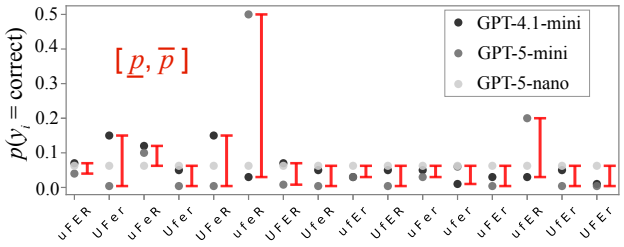


Figure 6: Credal-set from LLM ensemble: model disagreement induces lower/upper probability bounds for each output, which are used to compute MMI.

under the transformation that differ only by the lowercase noise p. The probability of each answer under the noise varies across candidates y_i. Let ℓ denote the number of lowercase characters in y_i; then $p(y_i) = p^{\ell} \cdot (1-p)^{1-\ell}$. With $p = 0.25$, the ideal probabilities are $p(y_i) \in \{0.316, 0.105, 0.035, 0.012, 0.004\}$ for $\ell \in \{0, 1, 2, 3, 4\}$. In practice, these values depend on the ICL examples and may deviate from the ideal case.

We apply the CREDAL method to elicit group uncertainty. Unlike the single-LLM setting, where $\hat{y}$ can be obtained simply by asking the model which option to choose, the group setting requires an additional decision rule to aggregate votes across multiple LLMs. To avoid this added complexity, we instead compute AUROC against $p(y_i = \text{correct})$ for each answer and use the exact MMI (Eq. 1) at the answer level. To compute MMI, we first elicit each LLM to report its *first-order* probabilities $p(y_i = \text{correct})$ (dots in Fig. 6). The IP intervals (red bars) are obtained by taking the pointwise minimum/maximum probabilities across models. We compare CREDAL with the standard aggregation method used in LLM ensembles—utilitarian aggregation—defined as, $\mathbb{E}_{\mathcal{M}_i}[p(y = \text{correct} \mid \mathcal{M}_i)]$. Taking the argmax over this quantity recovers popular strategies such as majority voting or best-of-N (e.g., [Wang et al., 2022, Si et al., 2023]). In our evaluation, however, we use the averaged score directly and compare it with our MMI score using AUROC. As shown in Figure 5(b), simple average provide limited predictive power, whereas CREDAL yields substantial improvement.

5 REAL-WORLD QA EXPERIMENT

We evaluate our methods on real-world QA benchmarks, following standard practice.

Datasets. We use two dataset types. (i) Open-ended QA: MAQA [Yang et al., 2025] and AmbigQA [Min et al., 2020]. Both contain a mix of ambiguous ($|\mathcal{Y}^\star| > 1$) and unambiguous items ($|\mathcal{Y}^\star| = 1$). (ii) Non-open-ended QA: MMLU-Pro [Wang et al., 2024] (answer options $\mathcal{Y}$ is provided) but only contains unambiguous items ($|\mathcal{Y}^\star| = 1$).

Task. Given question x, we prompt the LLM to predict a

Table 1: AUROC for correctness detection without ambiguity. Best result bolded in **blue**, second best in **orange**.

Method	GPT-5			gemini-2.5-pro			Avg. Rank
	MMLU-Pro	Non-MAQA	Non-AmbigQA	MMLU-Pro	Non-MAQA	Non-AmbigQA	
Is true prob.	0.6850 ± 0.0238	0.5835 ± 0.0041	0.5584 ± 0.0157	0.6583 ± 0.0197	0.6296 ± 0.0068	0.5897 ± 0.0057	8.67
Label prob.	0.7649 ± 0.0220	0.6430 ± 0.0062	0.5876 ± 0.0176	0.6377 ± 0.0153	0.6483 ± 0.0126	0.5582 ± 0.0319	7.83
MI-Clarifications	0.7790 ± 0.0114	0.6528 ± 0.0080	0.5840 ± 0.0107	0.6348 ± 0.0144	0.6546 ± 0.0104	0.5497 ± 0.0267	7.83
DiNCO	0.7087 ± 0.0214	0.5642 ± 0.0066	0.5219 ± 0.0035	0.7133 ± 0.0167	0.5749 ± 0.0224	0.5275 ± 0.0190	9.17
Vanilla	0.8587 ± 0.0192	0.7577 ± 0.0194	**0.7756 ± 0.0138**	0.7326 ± 0.0221	0.6889 ± 0.0263	0.6688 ± 0.0209	4.17
Top-4	0.8725 ± 0.0143	0.7668 ± 0.0089	0.7666 ± 0.0135	0.7652 ± 0.0226	**0.7027 ± 0.0294**	0.6698 ± 0.0087	3.33
CoT	0.8548 ± 0.0124	0.7695 ± 0.0078	0.7724 ± 0.0101	0.6826 ± 0.0409	0.6431 ± 0.0140	0.6094 ± 0.0152	5.33
ProbInt (ours)	0.8617 ± 0.0082	**0.7709 ± 0.0058**	0.7713 ± 0.0044	**0.7857 ± 0.0285**	**0.7303 ± 0.0092**	**0.7128 ± 0.0187**	**2.33**
Credal (ours)	**0.8798 ± 0.0121**	**0.7753 ± 0.0067**	**0.7932 ± 0.0126**	0.7466 ± 0.0201	0.6978 ± 0.0199	0.5982 ± 0.0469	**2.67**
Pos (ours)	**0.8738 ± 0.0216**	0.7230 ± 0.0179	0.7426 ± 0.0228	**0.7863 ± 0.0163**	0.6806 ± 0.0114	**0.6780 ± 0.0094**	3.67

single answer $\hat{y}$. Correctness is determined by comparing $\hat{y}$ with a reference answer $y^\star$, defined as: (i) non-ambiguous: $\mathcal{Y}^\star = \{y^\star\}$; or (ii) ambiguous: a specific sample $y^\star \in \mathcal{Y}^\star$. Any $y \in \mathcal{Y}^\star$ but $y \neq y^\star$ is judged as incorrect.

Metric. We report AUROC using each uncertainty score. (i) ambiguity: labels are $|\mathcal{Y}^\star| > 1$ vs. $|\mathcal{Y}^\star| = 1$. (ii) correctness: labels are $\hat{y} = y^\star$ vs. $\hat{y} \neq y^\star$. AUROC is calculated by how well the uncertainty score separates the two classes, with higher uncertainty score corresponding to ambiguous examples for (i) and incorrect predictions for (ii).

Baselines. The only directly comparable baseline is MI CLARIFICATIONS [Hou et al., 2024], which decomposes total uncertainty (equivalent to semantic entropy [Farquhar et al., 2024]) into aleatoric (first-order) and epistemic (second-order) uncertainty via the mutual information between sampled outputs and generated clarifications. We restrict our comparison to black-box compatible baselines.

Baselines for ambiguity. (i) SEMANTIC ENTROPY [Farquhar et al., 2024], entropy over semantically clustered samples; (ii) ASK4CONF-D [Hou et al., 2024], directly eliciting the probability that a question is ambiguous; and (iii) MI CLARIFICATIONS [Hou et al., 2024].

Baselines for correctness. We consider verbalized and sampling-based methods. *Verbalized:* (i) VANILLA [Tian et al., 2023, Xiong et al., 2024]; (ii) COT [Xiong et al., 2024], eliciting a rationale to improve prediction and confidence; and (iii) TOP-4 [Tian et al., 2023], querying the top four answers but using only the top answer's confidence. *Sampling-based:* (i) IS-TRUE [Tian et al., 2023, Kadavath et al., 2022]; (ii) LABEL PROBABILITY, estimating an empirical answer distribution from samples; (iii) DINCO [Wang and Stengel-Eskin, 2025], extending IS-TRUE with distractors and normalization, and (iv) MI CLARIFICATIONS [Hou et al., 2024].

Base setup. For each dataset, we sample $N = 200$ examples from the validation set and repeat experiments five times, reporting mean and standard deviation. Sampling-based baselines and CREDAL use five samples per query; runs with different seeds are treated as distinct beliefs.

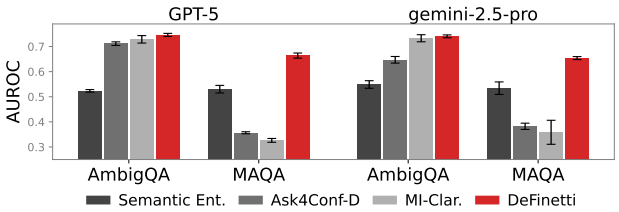


Figure 7: AUROC for ambiguity detection.

5.1 AMBIGUITY AND CORRECTNESS

Ambiguity detection. As shown in Fig. 7, our DEFINETTI achieves the highest AUROC for detecting ambiguity. Interestingly, directly eliciting an ambiguity probability (ASK4CONF-D) underperforms on MAQA, suggesting that a single global ambiguity judgment is insufficient without modeling the answer distribution.

Correctness detection. We next evaluate correctness detection capability. We first select unambiguous subsets from MAQA and AmbigQA (i.e., $|\mathcal{Y}^\star| = 1$), and we call them as Non-MAQA and Non-AmbigQA. Results are reported in Table 1. Overall, our methods perform best, although verbalization-based approaches also work reasonably well in this setting. This is consistent with prior findings: in the absence of ambiguity, uncertainty scores provably perform reliably [Tomov et al., 2025]. Empirically, PROBINT is the most robust across datasets and models.

Correctness detection under ambiguity. Under simultaneous ambiguity and incomplete knowledge, isolated scores may fail to capture total uncertainty. Since DEFINETTI and IP scores are measured on different scales, they cannot be directly summed. We therefore combine them multiplicatively to achieve scale invariance. Figure 8[3] shows that this

[3]We take $y^\star$ to be the reference answer from Natural Questions [Kwiatkowski et al., 2019], included in AmbigQA.

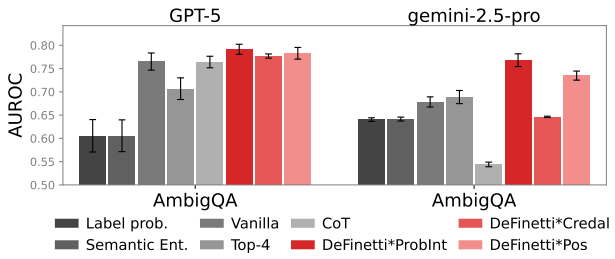


Figure 8: AUROC for correctness with ambiguity.

product separates correct from incorrect predictions $\hat{y}$ better than the baselines under combined uncertainty.

API cost. We report API costs computed from public pricing for all methods (Fig. 9). Compared to sampling-based baselines, our methods are generally more cost-efficient, except CREDAL which also uses sampling. PROBINT and POS are comparable in cost to verbalized baselines. Compared to MI CLARIFICATIONS—the only baseline that disentangles uncertainty—our methods cost less than half.

Visualization. For ambiguity, we use the MAQA dataset. Figure 2(c) shows that our DEFINETTI estimate effectively separates ambiguous from clear questions, whereas the VANILLA score in Figure 1(b) fails to do so. For correctness, we adopt the Kullback–Leibler (KL) divergence metric proposed by Tomov et al. [2025], who construct a reference answer distribution $p^\star$ from question–answer co-occurrence frequencies in a large corpus. They evaluate uncertainty (in the absence of ambiguity) via the KL divergence between the LLM's predictive distribution $\hat{p}$ and $p^\star$ (see Appendix D.2.1). Figure 10 shows that our PROBINT method exhibits the strongest correlation with this KL metric. We further compute the concordance index, which measures rank-based correlation between two variables, again comparing the KL metric with uncertainty scores. As shown in Figure 11, our IP-based methods demonstrate consistently strong and robust correlations.

5.2 EXPLAINING OWN DECISION

Previously, we evaluated correctness against the ground-truth $y^\star$. We now instead compare against the model's own prediction $\hat{y}$. In classical multi-class classification, prediction is obtained via $\hat{y}_{\text{precise}} = \arg\max_{y \in \mathcal{Y}} \hat{p}(y = \text{correct})$, which we refer to as PRECISE PROB. This rule is fully algorithmic and internally consistent. In contrast, LLM predictions resemble sampling from a conditional distribution, $\hat{y}_{\text{LLM}} \sim \hat{p}(y \mid x)$, meaning elicited probabilities need not coincide with the argmax of the predictive distribution. More broadly, this issue relates to faithfulness and self-consistency in LLMs [Madsen et al., 2024, Matton et al., 2025].

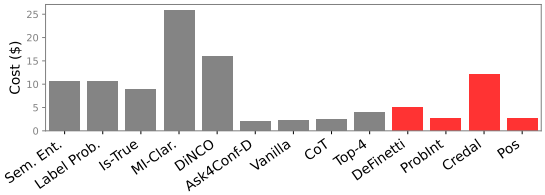


Figure 9: API cost on MMLU-Pro across all methods.

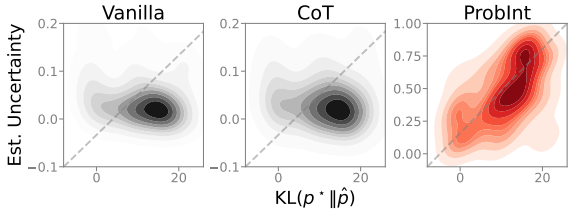


Figure 10: PROBINT more closely matches the KL metric.

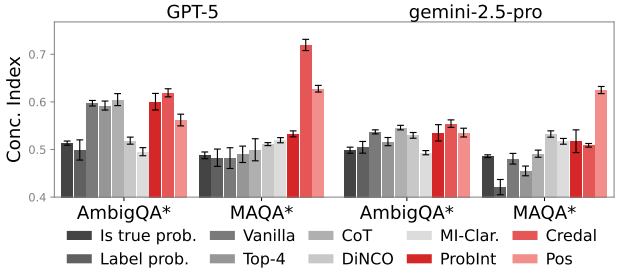


Figure 11: Concordance index (↑ better) between uncertainty scores and the KL metric.

Bayesian rationality. Under uncertainty, a natural self-consistency criterion is Bayesian rationality [Harsanyi, 1978], which checks whether $\hat{y}$ matches the argmax of expected utility. To compute utility, we employ MI-CLARIFICATION to elicit predictive distributions over answers: $\hat{y}_{\text{Bayes}} = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{C_i \sim p(C_i|x)} \hat{p}(y = \text{correct} \mid C_i)$, where C_i denotes clarification contexts.

IP-based rationality. Decision-making under IP does not yield a single canonical rule as in Bayesian theory. We therefore consider two standard criteria: *maximin* ($\hat{y}_{\text{maximin}} = \arg\max_{y \in \hat{\mathcal{Y}}} \underline{p}(y)$) and *maximax* ($\hat{y}_{\text{maximax}} = \arg\max_{y \in \hat{\mathcal{Y}}} \overline{p}(y)$). We use PROBINT to directly elicit the probability intervals $[\underline{p}(y), \overline{p}(y)]$.

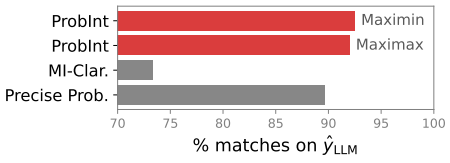


Figure 12: The LLM's prediction $\hat{y}_{\text{LLM}}$ aligns with the IP-rational decision.

Results. We evaluate on AmbigQA by measuring the alignment rate between $\hat{y}_{\text{LLM}}$ and each decision rule. As shown in Figure 12, the maximin and maximax rules exhibit strong alignment with the LLM's predictions. This suggests that the elicited probability intervals capture uncertainty relevant to answer selection. While this does not imply that LLMs explicitly optimize these rules, it shows that their predictions are compatible with IP-based rationality criteria.

5.3 DISCUSSION

We provide additional details in the appendix to assess the robustness and efficiency of our approach. First, to justify using approximate MMI instead of exact MMI, we compare the two quantities empirically. The approximation strongly agrees with exact MMI, with Pearson correlations above 0.9, while producing a speedup of more than 10^6 (Appendix B.4). Second, to rule out prompt-specific artifacts, we conduct a prompt-sensitivity ablation and find that different PROBINT variants lead to consistent conclusions, indicating robustness to prompt formulation (Appendix D.1.3). Finally, to assess consistency between first- and second-order elicitation, we test whether the first-order probability estimate lies within the second-order probability interval. This holds approximately 70% of the time, which is reasonable given stochastic variation from temperature sampling; adding an explicit containment constraint increases containment to 100% without changing the qualitative conclusions (Appendix D.2.2).

6 CONCLUSION AND LIMITATIONS

We propose IP-based methods for higher-order uncertainty elicitation. Across tasks and models, our approach improves elicitation accuracy and internal consistency while remaining cost-efficient. Combined with MMI-based post-processing, IP provides a principled framework for assessing LLM credibility. Overall, our method offers a coherent interface for eliciting, representing, and using uncertainty from black-box models.

Our method shares several limitations with prior work. First, we assume verbalized uncertainty is approximately rational; although allowing imprecision mitigates, this cannot be fully verified. We also assume the model correctly interprets prompts. Strong classification performance is required—if the model fails to recognize alternative answers, ambiguity cannot be captured. Finally, we focus primarily on Q&A tasks; extending to settings such as translation or summarization remains future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Masaki Adachi, Siu Lun Chau, Wenjie Xu, Anurag Singh, Michael A. Osborne, and Krikamol Muandet. Bayesian optimization for building social-influence-free consensus. *arXiv preprint arXiv:2502.07166*, 2025. URL https://arxiv.org/abs/2502.07166.

Aakriti Agrawal, Rohith Aralikatti, Anirudh Satheesh, Souradip Chakraborty, Amrit Singh Bedi, and Furong Huang. Uncertainty-aware answer selection for improved reasoning in multi-llm systems. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025. URL https://aclanthology.org/2025.findings-emnlp.1367/.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the knowable from the unknowable with language models. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ud4GSrqUKI.

Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes. *Introduction to Imprecise Probabilities*. John Wiley & Sons, Chichester, United Kingdom, 2014. ISBN 978-0-470-97381-3.

Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: A suite of black-box, white-box, LLM judge, and ensemble scorers. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=WOFspd4lq5.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Siu Lun Chau, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KM2XzHq2Rm.

Siu Lun Chau, Soroush H Zargarbashi, Yusuf Sale, and Michele Caprio. Quantifying epistemic predictive uncertainty in conformal prediction. *arXiv preprint arXiv:2602.01667*, 2026. URL https://arxiv.org/abs/2602.01667.

B. De Finetti. *La prévision: ses lois logiques, ses sources subjectives*. Annales de l'Institut Henri Poincaré. Institut Henri Poincaré, 1937.

D. Dubois and H. Prade. *Théorie des possibilités: applications à la représentation des connaissances en informatique*. Masson, 1985.

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models Bayesian? A martingale perspective. In *International Conference on Machine Learning*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. URL https://doi.org/10.1038/s41586-024-07421-0.

Masahiro Fujisawa, Masaki Adachi, and Michael A. Osborne. Scalable valuation of human feedback through provably robust model alignment. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=EaTRrceoU9.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024. URL https://aclanthology.org/2024.eacl-long.143/.

John C. Harsanyi. Bayesian decision theory and utilitarian ethics. *The American Economic Review*, 68(2):223–228, 1978.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=byxXa99PtF.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025. URL https://arxiv.org/abs/2508.12790.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

John Maynard Keynes. *A Treatise on Probability*. Macmillan & co., 1921.

Michael Kirchhof, Luca Füger, Adam Goliński, Eeshan Gunesh Dhekane, Arno Blaas, Seong Joon Oh, and Sinead Williamson. SelfReflect: Can LLMs communicate their internal answer distribution? *arXiv preprint arXiv:2505.20295*, 2025a. URL https://arxiv.org/abs/2505.20295.

Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Position: Uncertainty quantification needs reassessment for large language model agents. In *International Conference on Machine Learning*, 2025b. URL https://proceedings.mlr.press/v267/kirchhof25b.html.

Kolmogorov. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From risk to uncertainty: Generating predictive uncertainty measures via Bayesian estimation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cWfpt2t37q.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lKEdGCoDNC.

Isaac Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. MIT Press, Cambridge, MA, 1980. ISBN 0262120828.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. ConfTuner: Training large language models to express their confidence verbally. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VZQ04Ojhu5.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Chen Ling, Xujiang Zhao, Wei Cheng, Yanchi Liu, Yiyou Sun, et al. Uncertainty decomposition and quantification for in-context learning of large language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2024. URL https://openreview.net/forum?id=Oq1b1DnUOP.

Ryan Liu, Jiayi Geng, Joshua Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. Large language models assume people are more rational than we really are. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dAeET8gxqg.

William Lugoloobi, Thomas Foster, William Bankes, and Chris Russell. LLMs encode their failures: Predicting success from pre-generation activations. *arXiv preprint arXiv:2602.09924*, 2026. URL https://arxiv.org/abs/2602.09924.

Sravanthi Machcha, Sushrita Yerra, Sharmin Sultana, Hong Yu, and Zonghai Yao. Do large language models know when not to answer in medical QA? In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP)*, 2025. URL https://aclanthology.org/2025.uncertainlp-main.4/.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL*, 2024. URL https://aclanthology.org/2024.findings-acl.19/.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4ub9gpx9xw.

Tommaso Mencattini, Robert Adrian Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodolà. MERGE^3: Efficient evolutionary merging on consumer-grade GPUs. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/mencattini25a.html.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://aclanthology.org/2020.emnlp-main.466/.

Krikamol Muandet. Impossibility of collective intelligence. *arXiv preprint arXiv:2206.02786*, 2022. URL https://arxiv.org/abs/2206.02786.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=qAml3FpfhG.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. *Herbert Robbins Selected Papers*, pages 7–24, 1951.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3), September 2025. doi: 10.1145/3744238.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. Prompting GPT-3 to be reliable. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=98p5x51L5af.

Cedric A. B. Smith. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 23(1):1–25, 1961.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Lacie: Listener-aware finetuning for calibration in large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RnvgYd9RAh.

Terry Therneau and Elizabeth Atkinson. The concordance statistic. *A package for survival analysis in R, vignettes. R package version*, pages 3–7, 2023.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.330/.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in LLMs improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*, 2024. URL https://arxiv.org/abs/2404.10960.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. The illusion of certainty: Uncertainty quantification for LLMs fails under ambiguity. *arXiv preprint arXiv:2511.04418*, 2025. URL https://arxiv.org/abs/2511.04418.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: a generalization theory for meta-learning. *arXiv preprint arXiv:2510.10981*, 2025. URL https://arxiv.org/abs/2510.10981.

Nassim Walha, Sebastian G. Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. Fine-grained uncertainty decomposition in large language models: A spectral approach. *arXiv preprint arXiv:2509.22272*, 2025. URL https://arxiv.org/abs/2509.22272.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Victor Wang and Elias Stengel-Eskin. Calibrating verbalized confidence with self-generated distractors. *arXiv preprint arXiv:2509.25532*, 2025. URL https://arxiv.org/abs/2509.25532.

Xuesong Wang. Active learning for nlp with large language models. *arXiv preprint arXiv:2401.07367*, 2024. URL https://arxiv.org/abs/2401.07367.

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

Yifei Wang, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. Uncertainty unveiled: Can exposure to more in-context examples mitigate uncertainty for large language models? In *Findings of the Association for Computational Linguistics: ACL*, 2025. URL https://aclanthology.org/2025.findings-acl.1062/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. URL https://arxiv.org/abs/2406.01574.

Shiguang Wu, Yaqing Wang, and Quanming Yao. Why in-context learning models are good few-shot learners? In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iLUcsecZJp.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. *arXiv preprint arXiv:2503.00172*, 2025.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.343/.

Khurram Yamin, Jingjing Tang, Santiago Cortes-Gomez, Amit Sharma, Eric Horvitz, and Bryan Wilder. Do LLMs act like rational agents? Measuring belief coherence in probabilistic decision making. *arXiv preprint arXiv:2602.06286*, 2026. URL https://arxiv.org/abs/2602.06286.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL*, 2025. URL https://aclanthology.org/2025.findings-naacl.325/.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is chain-of-thought reasoning of LLMs a mirage? A data distribution lens. *arXiv preprint arXiv:2508.01191*, 2025. URL https://arxiv.org/abs/2508.01191.

Verbalizing LLM's Higher-order Uncertainty via Imprecise Probabilities (Supplementary Material)

Anita Yang[1,2] **Krikamol Muandet**[3] **Michele Caprio**[4,5] **Siu Lun Chau**[6,*] **Masaki Adachi**[1,*]

[1]Lattice Lab, Toyota Motor Corporation, Japan
[2]Department of Computer Science, University of Tokyo, Japan
[3]Rational Intelligence Lab, CISPA, Helmholtz Center for Information Security, Germany
[4]Department of Computer Science, University of Manchester, United Kingdom
[5]Manchester Centre for AI Fundamentals, United Kingdom
[6]EPIC Lab, College of Computing & Data Science, Nanyang Technological University, Singapore
[*]Equal contribution

A PROMPTS

A.1 SECOND-ORDER UNCERTAINTY

We provide prompts for eliciting two other IP representations of second-order uncertainty discussed in the main text: credal sets (CREDAL) and possibility functions (POS).

Credal sets. We represent the credal set by eliciting a finite ensemble of M precise predictive distributions $\{p^{(m)}\}_{m=1}^{M}$ from M models or samples (random seed). The resulting credal set is taken to be the empirical set of these distributions: $\mathcal{C} = \text{conv}(\{p^{(m)}\}_{m=1}^{M})$. Each $p^{(m)}$ is obtained using Prompt 3.

Prompt 3: Credal

Assign a probability (between 0.0 and 1.0) representing how likely it is that the answer would be given as a response to the question.
A correct answer should generally receive a higher probability than an incorrect one. Likelihood may vary based on reasonable interpretations of the question (e.g., ambiguity in scope, answer type, entity interpretation, or contextual assumptions).
The sum of all probabilities must not exceed 1.0.

Possibility function. We elicit the LLM to provide a possibility function over the candidate answers, including an explicit alternative option (i.e., "none of the above"), using Prompt 4. We allow the elicited scores to be unnormalized; the required normalization is applied as part of the MMI computation (Apd. B.3).

Prompt 4: Pos

Provide a possibility score which captures how plausible the answer correctly answers the question.
Then, provide a possibility score how plausible it is that a different answer (not listed) could be correct.
The possibility should be between 0.0 and 1.0. A possibility score of 1.0 means "fully plausible," and 0.0 means "impossible."

A.2 APPROXIMATION OF $\hat{\mathcal{Y}}$.

To characterize uncertainty about the question x itself—rather than about a particular answer y (e.g., a predicted $\hat{y}$)—we consider uncertainty over the candidate answer set $\mathcal{Y}$ (Apd. C). In open-ended QA, however, $\mathcal{Y}$ is not observed. We therefore approximate it with a finite set $\hat{\mathcal{Y}}$ by prompting the model to generate plausible (non-zero-probability) candidate answers (Prompt 5), within which the ground-truth set of correct answers $\mathcal{Y}^\star$ is likely to lie.

Prompt 5: Candidate answers $\hat{\mathcal{Y}}$

Given the question below, generate a list of all possible correct answers, taking into account different reasonable interpretations of the question.

Provide the answers as a numbered list, with each answer on its own line. Each answer must be concise text only, with no explanations or additional wording. Do not include duplicates or answers that refer to the same entity or concept. For example:
1. <answer one as concise text>
2. <answer two as concise text>
...

B MAXIMUM MEAN IMPRECISION

The exact MMI metric for measuring EU [Chau et al., 2025] under total variation is

$$\mathrm{MMI}_{\mathrm{TV}}(\underline{P}) := \sup_{A \in \mathcal{F}_\mathcal{Y}} \big(\overline{P}(A) - \underline{P}(A)\big) \;\le\; 1 - \sum_{y \in \mathcal{Y}} \underline{P}(\{y\}), \tag{3}$$

where $\mathcal{F}_\mathcal{Y}$ is the σ-algebra over the candidate answers $\mathcal{Y}$ (in our discrete setting, $\mathcal{F}_\mathcal{Y} = 2^\mathcal{Y}$). The lower and upper probabilities $\underline{P}(A)$ and $\overline{P}(A)$ bound the probability of an event $A \subseteq \mathcal{Y}$, which we interpret as "the sampled answer lies in A." The quantity $\overline{P}(A) - \underline{P}(A)$ is the imprecision (interval width) for event A, and $\mathrm{MMI}_{\mathrm{TV}}$ takes the maximum such width over all events. For example, when $\mathcal{Y} = \{y_1, y_2\}$, Fig. 13 enumerates all events in $2^\mathcal{Y}$ and their bounds; $\mathrm{MMI}_{\mathrm{TV}}$ corresponds to the widest interval among them, capturing the agent's worst-case epistemic uncertainty over the output space.

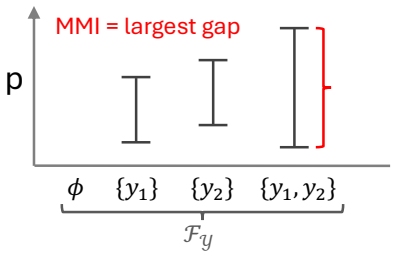


Figure 13: MMI measures the largest gap between upper/lower probabilities across all events $A \in \mathcal{F}_\mathcal{Y}$.

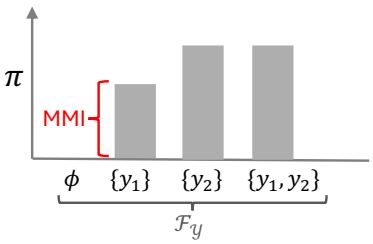


Figure 14: MMI for possibility function measures the second largest possibility score over $\mathcal{Y}$.

B.1 CREDAL SET

In CREDAL-EU, we instantiate the credal set as the convex hull of a finite ensemble of M precise predictive distributions $\{p^{(m)}\}_{m=1}^M$ produced by independently prompted runs (Prompt 3):

$$\mathcal{C} := \mathrm{conv}\big(\{p^{(m)}\}_{m=1}^M\big). \tag{4}$$

Since $\sum_{y \in A} p(y)$ is linear in p, its extrema over the polytope $\mathcal{C}$ are attained at extreme points. Therefore, the event bounds reduce to simple min/max over ensemble members:

$$\underline{P}(A) = \min_{m \in [M]} \sum_{y \in A} p^{(m)}(y), \qquad \overline{P}(A) = \max_{m \in [M]} \sum_{y \in A} p^{(m)}(y). \tag{5}$$

In particular, the singleton bounds used in the tractable upper bound are

$$\underline{P}(\{y\}) = \min_{m\in[M]} p^{(m)}(y), \qquad \overline{P}(\{y\}) = \max_{m\in[M]} p^{(m)}(y). \tag{6}$$

Exact MMI_{TV} requires optimizing over events A and can be exponential in $|\mathcal{Y}|$. In our experiments, we therefore (i) compute exact MMI only for *singleton* events corresponding to a specific answer y (which reduces to the interval width in Eq. 6), and (ii) use the linear-time upper bound in Eq. 3 for uncertainty over $\mathcal{Y}$. For open-ended QA, where $\mathcal{Y}$ is approximated by a finite candidate set $\hat{\mathcal{Y}}$, we apply the same computations over $\hat{\mathcal{Y}}$.

B.2 PROBABILITY INTERVAL

In PROBINT-EU, the LLM elicits a probability interval $[\underline{p}(y), \overline{p}(y)]$ for each candidate answer y. For singleton events, the induced lower/upper probabilities are immediate:

$$\underline{P}(\{y\}) = \underline{p}(y), \qquad \overline{P}(\{y\}) = \overline{p}(y). \tag{7}$$

For non-singleton events $A \subseteq \mathcal{Y}$, constructing coherent bounds $\underline{P}(A)$ and $\overline{P}(A)$ from singleton intervals generally requires additional assumptions and can be more involved. Moreover, computing exact MMI_{TV} still entails optimizing over events A and can scale exponentially with $|\mathcal{Y}|$. We therefore follow the same strategy as CREDAL-EU: (i) we compute *exact* MMI only for answer-level uncertainty on a specific y (i.e., singleton events / correctness of y), and (ii) we use the linear-time upper bound in Eq. 3 for task-level uncertainty over $\mathcal{Y}$ (or its approximation $\hat{\mathcal{Y}}$).

B.3 POSSIBILITY FUNCTION

In POS-EU, we prompt the LLM to elicit a (potentially unnormalized) possibility function $\hat{\pi} : \mathcal{Y} \to [0, 1]$ over the output space. In open-ended QA, $\mathcal{Y}$ is approximated by a finite candidate set $\hat{\mathcal{Y}}$ together with a "none of the above" option, allowing the model to assign high plausibility to an outcome outside the enumerated candidates without forcing plausibility to be redistributed among them, since possibility scores are non-additive.

For possibility functions, total-variation MMI admits a simple form: after normalization so that $\max_{y\in\mathcal{Y}} \pi(y) = 1$, MMI equals the second-largest possibility value [Chau et al., 2026]:

$$\text{MMI}_{\text{TV}} = \pi_{(2)} \tag{8}$$

$$= \frac{\hat{\pi}_{(2)}}{\hat{\pi}_{(1)}}, \tag{9}$$

where $\pi_{(1)} \geq \pi_{(2)} \geq \cdots$ denote the order statistics of the normalized scores $\{\pi(y)\}_{y\in\mathcal{Y}}$, and $\hat{\pi}_{(1)} \geq \hat{\pi}_{(2)} \geq \cdots$ are the corresponding order statistics of the unnormalized scores $\{\hat{\pi}(y)\}$. Because normalization enforces $\pi_{(1)} = 1$, $\pi_{(2)}$ measures the strength of the best competing alternative relative to the top hypothesis: larger values indicate greater ambiguity among leading candidates (and hence higher imprecision). Equivalently, MMI quantifies uncertainty in the top answer via the strongest competitor (Fig. 14).

When assessing uncertainty for a single answer y (binary event $\{y, \neg y\}$), we elicit unnormalized scores $\hat{\pi}(y)$ and $\hat{\pi}(\neg y)$ (with $\neg y$ implemented as "not y") and normalize by dividing by $\max\{\hat{\pi}(y), \hat{\pi}(\neg y)\}$, resulting in:

$$\text{MMI}_{\text{TV}} = \min\left\{ \frac{\hat{\pi}(y)}{\max(\hat{\pi}(y), \hat{\pi}(\neg y))}, \frac{\hat{\pi}(\neg y)}{\max(\hat{\pi}(y), \hat{\pi}(\neg y))} \right\} = \frac{\min\{\hat{\pi}(y), \hat{\pi}(\neg y)\}}{\max\{\hat{\pi}(y), \hat{\pi}(\neg y)\}}. \tag{10}$$

B.4 APPROXIMATE VS. EXACT

We empirically compare the approximate against the exact MMI (Eq. 3). On AmbigQA* and MAQA*, the Pearson correlations between the approximate and exact estimates are 0.9318 and 0.9266, respectively, indicating strong agreement. This agreement comes with a substantial computational advantage: the exact MMI computation scales exponentially with the number of candidate answers $|\mathcal{Y}|$, whereas the approximation scales linearly in $|\mathcal{Y}|$. For example, with $|\mathcal{Y}| = 20$ over 200 AmbigQA* samples, exact MMI computation takes 627.2 seconds, while the approximation takes less than 0.0001 seconds.

C UNCERTAINTY ON: "ANSWER y" VS. "CANDIDATE ANSWERS $\mathcal{Y}$"

We clarify when uncertainty should be defined for the correctness of *a particular answer* y (e.g., a model prediction $\hat{y}$) versus over the *candidate answer set* $\mathcal{Y}$ (or its approximation $\hat{\mathcal{Y}}$).

Two objects of uncertainty:
(i) *Answer y (answer-centric).* Uncertainty on y concerns whether a particular answer y (typically $\hat{y}$) is correct, collapsing all alternatives into "not-y." It is most useful for decision-focused use cases (trust/abstain/verify/fallback), where utility depends primarily on the correctness of the chosen answer.

(ii) *Candidate answers $\mathcal{Y}$ (question-centric).* Uncertainty on $\mathcal{Y}$ concerns how belief is distributed across plausible answers. It is most useful for question-focused use cases (ambiguity detection, multiple valid answers, need for clarification), where utility depends on whether the question supports several competing answers, not just whether $\hat{y}$ is correct.

C.1 FIRST-ORDER UNCERTAINTY

On answer y. For a given answer y, first-order uncertainty is defined on the binary event "y vs. not-y." Under a precise predictive distribution p, this is the Bernoulli entropy $H(\mathrm{Bern}(p(y)))$. In Fig. 15a, the correctness AU for y_1 equals $H(\mathrm{Bern}(0.4))$.

On candidate answers $\mathcal{Y}$. First-order uncertainty over the candidate set quantifies how probability mass is distributed *across* answers. Under a precise predictive distribution p on $\mathcal{Y}$, this is the entropy $H(p)$ over all $y \in \mathcal{Y}$ (Fig. 15a).

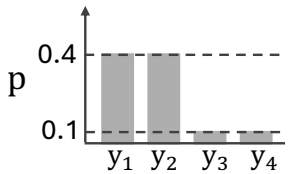
Experiments. We use these notions for different evaluation goals: (i) *Answer y*: decision-centric evaluations, including error tracking in our ICL experiments (Sec. 4) and correctness detection (AUROC; Fig. 8 in Sec. 5). (ii) *Candidate answers $\mathcal{Y}$*: question-centric evaluations, namely ambiguity detection in open-ended QA (Fig. 7 in Sec. 5).

C.2 SECOND-ORDER UNCERTAINTY

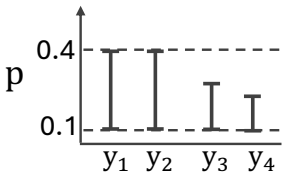
On answer y. For a given answer y, second-order uncertainty (imprecision) captures uncertainty about the binary event "y vs. not-y" under a probability interval. We measure this by the interval width $\overline{p}(y) - \underline{p}(y)$ (Eq. 1). In Fig. 15b, the EU for y_2 is the interval gap (MMI $= 0.3$).

On candidate answers $\mathcal{Y}$. Second-order uncertainty over the candidate set captures imprecision about the *entire* answer distribution. We measure this using the upper-bound MMI, $1 - \sum_{y \in \mathcal{Y}} \underline{p}(y)$ (Eq. 2). In Fig. 15b, this aggregates lower bounds across candidates (upper-bound MMI $= 1 - 0.4$).

Experiments. We use these notions for different evaluation goals: (i) *Answer y*: decision-centric evaluations, including error tracking in our ICL experiments (Sec. 4) and correctness detection (AUROC; Sec. 5). (ii) *Candidate answers $\mathcal{Y}$*: question-centric evaluations, including ambiguity-related analyses and comparisons to "ground-truth" EU distributions (Fig. 11).



(a) First order: (i) binary entropy for y vs. not-y, or (ii) entropy over $\mathcal{Y}$.



(b) Second order: (i) interval width of each y, or (ii) upper-bound MMI over $\mathcal{Y}$.

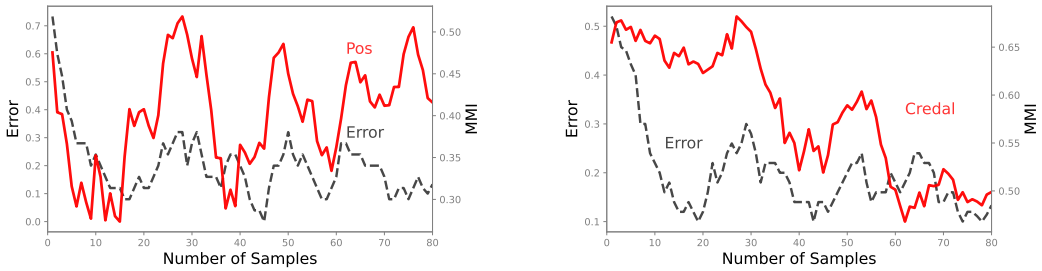
Figure 15: First- and second-order uncertainty can be computed either for the (i) correctness of a specific answer y, or (ii) over the candidate answers set $\mathcal{Y}$

D EXPERIMENTS

D.1 IN-CONTEXT LEARNING

D.1.1 Second-order denoise

We report additional error-tracking results from Sec. 4.2 using second-order uncertainty approximated via CREDAL (Fig. 16b) and POS (Fig. 16a). In contrast to VANILLA (Fig. 5), both representations track error more closely, and CREDAL shows a clear decrease as the number of in-context examples grows, indicating effective denoising of epistemic uncertainty with additional evidence. Between PROBINT, POS, and CREDAL, POS is the most sensitive to first-order noise, consistent with Fig. 4 (holding second-order fixed while varying first-order noise). Nevertheless, POS remains more robust than VANILLA under first-order noise (Fig. 4).



(a) POS (b) CREDAL

Figure 16: Approximations of second-order uncertainty with fixed first-order noise (lower-case probability $p = 0.25$) (with moving average window of 10). Both POS and CREDAL generally match shape of the error.

D.1.2 Group Uncertainty Elicitation

We provide additional experimental details in Sec. 4.3. Our ensemble consists of three language models: GPT-5-Mini, GPT-4.1-mini, and GPT-5-Nano. Using GPT-5-Mini with Prompt 5, we first construct a shared candidate set $\hat{\mathcal{Y}}$. This controls for the fact that different models may produce different point predictions $\hat{y}$, while our elicitation procedure requires every model to assign uncertainty over the same set of answer options. We therefore define $\hat{\mathcal{Y}}$ to cover the answers the ensemble is most likely to generate, and then prompt each model to verbalize its uncertainty over $\hat{\mathcal{Y}}$. In contrast, for our other CREDAL experiments we obtain multiple samples by repeatedly querying a single model.

D.1.3 Prompt sensitivity

We analyze the sensitivity of our method to prompt formulation. Specifically, we consider three variants, denoted PROBINT-1–PROBINT-3 (Prompts 6–8). We evaluate these prompts on the synthetic in-context learning setup from Sec. 4. As shown in Fig. 17, the second-order uncertainty elicited by each prompt qualitatively tracks the error trend, suggesting that the method is robust across prompt formulations.

Prompt 6: ProbInt-1

Provide your best guess to the question and a lower and upper probability (each between 0.0 and 1.0) that your guess is correct.

Interpret the probabilities as follows:
- Lower Probability: the smallest probability you consider plausible that your guess is correct.
- Upper Probability: the largest probability you consider defensible that your guess is correct.

Prompt 7: ProbInt-2

An unknown transformation maps each input sequence to an output sequence.
Now predict the output for the given input and state your uncertainty as an upper and lower probability (each between 0.0 and 1.0) that your answer is correct.

Interpret the probabilities as follows:
- Upper Probability: the largest probability you consider defensible that your output is correct.
- Lower Probability: the smallest probability you consider plausible that your output is correct

Prompt 8: ProbInt-3

The following input-output pairs were produced by an unknown sequence transformation.
Based only on these examples, predict the output sequence for the test input below. Then give a probability interval for the event that your predicted output is correct.

Interpretation:
- Lower Probability: the conservative plausible probability that your output is correct.
- Upper Probability: the highest defensible probability that your output is correct.

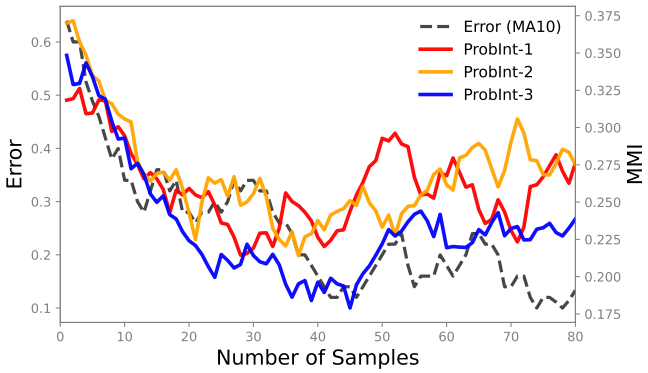


Figure 17: Second-order uncertainty estimates elicited by variants of the PROBINT prompt consistently track the overall error trend.

D.2 REAL-WORLD QA EXPERIMENT

D.2.1 Estimating first- and second-order uncertainty from dataset

We assess disentangling and quantifying second-order uncertainty in settings where both first- and second-order uncertainty are present. In such cases, AUROC can be reductive: it scores correctness against a single reference answer $y^\star$, even when other answers may also be valid (with varying probabilities). To enable a more principled evaluation, we follow Tomov et al. [2025], who propose an approximation to *ground-truth* first- and second-order uncertainty distributions (their aleatoric uncertainty (AU) and epistemic uncertainty (EU)) for ill-defined real-world QA. They posit a *true* first-order answer distribution $p^\star$, which can be approximated from corpus statistics by estimating question–answer co-occurrence frequencies in a large text corpus. This estimated $p^\star$ is provided in the AmbigQA* and MAQA* datasets.

Given $p^\star$ and a model predictive distribution $\hat{p}$, AU and EU admit the decomposition

$$\underbrace{\mathrm{CE}(p^\star, \hat{p})}_{\text{Total uncertainty}} = \underbrace{H(p^\star)}_{\text{AU}} + \underbrace{\mathrm{KL}(p^\star \,\|\, \hat{p})}_{\text{EU}}, \tag{11}$$

where CE denotes cross-entropy and KL the Kullback–Leibler divergence. This cross-entropy decomposition is closely related to mutual-information-based uncertainty decompositions [Kotelevskii et al., 2025]. We therefore treat $H(p^\star)$ as a

proxy for ground-truth first-order uncertainty (AU) and $\mathrm{KL}(p^\star \,\|\, \hat{p})$ as a proxy for ground-truth second-order uncertainty (EU).

Metric. Following Tomov et al. [2025], we use concordance statistics [Therneau and Atkinson, 2023] to measure rank agreement between an estimated uncertainty score and the corresponding proxy ground truth. Concordance is the probability that, for a randomly chosen pair of examples, the method assigns a higher uncertainty to the example with higher ground-truth uncertainty. Its interpretation matches AUC-ROC: values closer to 1 indicate better ranking alignment. We report this metric in Fig. 11.

Second-order uncertainty. We compare second-order uncertainty estimated by our proposed IP representations to the KL-based proxy $\mathrm{KL}(p^\star \,\|\, \hat{p})$. Figure 10 visualizes their association for VANILLA, COT, and PROBINT; our method exhibits the strongest alignment. We further benchmark against additional uncertainty estimators in Fig. 11 using the concordance index, where our methods are the most aligned across datasets and models.

First-order uncertainty. We also evaluate first-order uncertainty by comparing method estimates to the entropy proxy $H(p^\star)$ on MAQA*. Figure 18 visualizes the estimated first-order uncertainty from PREDICTIVE ENTROPY, MI-CLARIFICATIONS, and DEFINETTI, where DEFINETTI is best aligned with the proxy with highest concordance index.

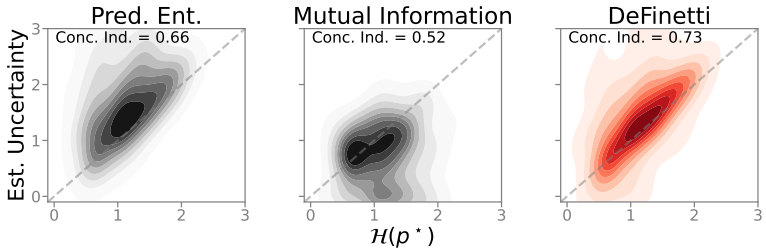


Figure 18: DEFINETTI best aligns with the proxy for first-order uncertainty

D.2.2 Consistency of second-order with first-order elicitation

We evaluate whether the first-order point estimate elicited by Prompt 1 lies within the probability interval elicited by Prompt 2 on AmbigQA under the setting of Fig. 8, where both types of uncertainty are required. With the original prompts, the containment ratio is 72.33%, with an AUROC of 0.7862 ± 0.0035. Adding an explicit containment constraint raises the containment ratio to 100%, while AUROC remains comparable at 0.7554 ± 0.0008. Thus, the qualitative conclusion remains unchanged.

D.2.3 Smaller LLMs

We provide an additional evaluation with Llama-3.1-8B on the Non-AmbigQA experiment in Table 1, where the model achieves 79% prediction accuracy. We compare PROBINT against LABEL PROB., which prompts the model five times and estimates confidence from the frequency of the predicted answer. However, LABEL PROB. produces the same answer across samples, resulting in an AUROC of 0.5. In contrast, PROBINT achieves an AUROC of 0.6027 ± 0.0000, outperforming the VANILLA baseline at 0.5698 ± 0.0017. We also attempted to extend the evaluation to other smaller models, but found that they generally lack sufficient capacity to solve the underlying task, making uncertainty tracking uninformative for all methods.