

Credal Ensembles for Uncertainty Quantification in LLMs

Vani Kanjirangat

SUPSI, IDSIA

Lugano, Switzerland

vani.kanjirangat@supsi.ch

Oscar Lithgow-Serrano

SUPSI, IDSIA

Lugano, Switzerland

oscarwilliam.lithgow@supsi.ch

Alessandro Antonucci

SUPSI, IDSIA

Lugano, Switzerland

alessandro.antonucci@supsi.ch

Abstract

We consider the problem of uncertainty quantification in LLMs from a black-box perspective. To effectively disentangle aleatoric and epistemic uncertainty, we use an ensemble method that produces responses paired with so-called *credal* sets of distributions. We combine different ensembling techniques, based on temperature variation, multiple models, paraphrasing, and permutations, with semantic-similarity and verbalised elicitation methods, resulting in new, computationally cheap measures of uncertainty. A benchmark on closed-ended (BoolQ and CommonsenseQA) and open-ended (AmbigQA) questions shows good correlation with predictive accuracy. Notably, when considering ambiguous tasks, our analysis reveals a tendency of LLMs to focus on a single interpretation, an important flaw in current LLM-based solutions.

1 Introduction

In settings governed by noisy generative processes, *uncertainty quantification* (UQ, [Der Kiureghian and Ditlevsen, 2009](#)) conventionally distinguishes between two sources: *aleatoric* uncertainty (AU) as the irreducible randomness inherent to the data-generating process itself, and *epistemic* uncertainty (EU) reflecting the lack of knowledge due to limited data or model misspecification. Whereas AU represents an intrinsic limit, EU is in principle reducible through additional data or improved modelling. UQ is particularly consequential in high-stakes applications: high EU flags cases where the model itself is untrustworthy and should prompt caution, further data collection, or human oversight, while AU calls for decision-making that accounts for a fundamental predictability limit.

Several techniques have been successfully developed for UQ in deep learning models, mostly Bayesian approaches placing and propagating distributions over model parameters to obtain predictive uncertainty ([Abdar et al., 2021](#)). Extending

these methods to LLMs, however, is challenging primarily for computational reasons: LLMs comprise billions of parameters, making it prohibitively expensive to maintain or sample full posterior distributions over weights. Moreover, their autoregressive nature compounds uncertainty across long output sequences in ways that classical Bayesian machinery is not designed to handle ([Malinin and Gales, 2021](#)). This also precludes the use of deep *ensembles*, trained with different initializations.

Since LLMs are inherently probabilistic models, using their raw output probabilities directly for UQ would be an appealing alternative to more expensive approaches. Yet existing work has uncovered calibration issues that undermine this straightforward strategy ([Kadavath et al., 2022](#)). Following [Hofman et al. \(2024\)](#), we instead approach UQ through *credal* ensembles inducing sets of distributions (called credal sets, CS, [Walley, 1991](#)) jointly regarded as generalised uncertainty models. We instantiate this idea by exploring different ensemble techniques for LLMs that require neither re-training nor low-level interventions, and evaluate the corresponding UQ estimates on QA tasks.

With closed-ended benchmarks (*BoolQ* and *CommonsenseQA* in our tests), we can easily obtain CSs based on the raw probabilities by means of question paraphrasing and permutation strategies. Tests on these benchmarks allows us to verify the suitability of our *credal-ensemble* EU and AU descriptors and their calibration. With open-ended tasks (*AmbigQA*), we propose instead a semantic clustering of sampled free-form responses. Such a procedure is iterated to obtain CSs by considering different models or temperatures.

Our empirical analysis based on those benchmarks address the following research questions.

1. **RQ1** When applied to LLMs, does credal-ensemble UQ produce calibrated measures that correlate well with predictive accuracy?

2. **RQ2** Do LLMs adequately represent AU when facing ambiguous questions?
3. **RQ3** Which variation strategy is better suited to capture EU in LLMs?

To properly address and answer the questions, we first review the relevant literature in Section 2. The basics of credal-ensemble UQ are reviewed in Section 3 and applied to LLMs in Section 4. The experimental setup is discussed in Section 5 and the results presented in Section 6. This allows to answer the research questions in Section 7. Conclusions are in Section 8. Limitations are finally discussed in Section 9

2 Related Works

The recent survey by Shorinwa et al. (2025) categorises UQ methods for LLMs into four families: token-level, mechanistic interpretability, self-verbalised, and semantic-similarity approaches. The first two are incompatible with our *black-box* focus, i.e., methods that do not rely on access to the model’s intermediate outputs or architecture. See Gao et al. (2024) for a promising, but computationally expensive example of white-box approach. Our method builds on the semantic-similarity approach of Kuhn et al. (2023) and Farquhar et al. (2024), which clusters sampled generations by meaning, before computing a measure. We extend this idea by using variations that yield an *ensemble* of measures rather than a single one, this being key to disentangle EU from AU.

Ensembles are, of course, a standard way to estimate EU in deep learning (Lakshminarayanan et al., 2017), where disagreement across independently trained models captures reducible uncertainty; however, training such ensembles directly for LLMs is computationally prohibitive, and in any case incompatible with our black-box setting. To generate an ensemble more cheaply, temperature scaling within a single model has been proposed as a proxy (Guo et al., 2017), although recent work questions whether temperature affects prediction confidence independently of the model’s underlying knowledge (Desai and Durrett, 2020). This suggests that we should consider a multi-model approach as a good alternative to temperature variation from the broader perspective of UQ rather than calibration alone.

Regarding self-verbalised methods, Xiong et al. (2024) evaluate whether LLMs can express cal-

ibrated confidence through verbalised probability estimates and find systematic overconfidence across model families. Similar miscalibration surfaces in structured reasoning tasks as reported in Lithgow-Serrano et al. (2025). The paper shows that instruction-tuned models express near-certain verbalised confidence in causal relation classification despite near-random accuracy, and that output entropy over the classification options remains near its theoretical maximum—indicating the failure reflects missing causal representations rather than miscalibration alone. We also elicit verbalised probabilities, but focusing specifically on the distribution over interpretations rather than answer confidence.

Finally, ours is not the first paper to relate CSs to LLMs. Prior work in this direction has focused on LLM training (Mubashar et al., 2025), while the closest to our proposal is the recent work of Yang et al. (2026), which focuses on elicitation without specific attention to UQ.

3 Bayesian and Credal UQ

Notation and basics. Given a variable V taking values in a finite set $\mathcal{V}$, we denote a *probability mass function* (PMF) over V as P , with $P(v)$ the probability assigned to state $v \in \mathcal{V}$ and write $m := |\mathcal{V}|$ for the number of possible states. The *entropy* of P is defined as $H[P] := -\sum_{v \in \mathcal{V}} P(v) \log_m P(v)$, where the sum is restricted over the states with strictly positive probability and the use of base- m logarithms normalises H to the range $[0, 1]$. The entropy achieves its maximum, equal to one, with a uniform PMF assigning the same probability to all the states, while the minimum equal to zero corresponds to degenerate (i.e., zero/one) PMFs. A CS, i.e., a set of PMFs is denoted instead as $\mathcal{P}$; here we consider only CSs consisting of finitely many PMFs. To characterize the dissimilarity between two PMFs, we use the Kullback-Leibler (KL) pseudo-distance:

$$\delta[P, P'] := \sum_{v \in \mathcal{V}} P(v) \log_m \frac{P(v)}{P'(v)}, \quad (1)$$

where the sum is over the states receiving strictly positive probability from both P and P' . Since δ is not symmetric, we symmetrise it.

UQ of CSs. Following Hofman et al. (2024), a conservative, genuinely *credal*, UQ for the CS $\mathcal{P}$ can be obtained by considering the maximal entropy attained within $\mathcal{P}$ and the maximum (KL)

dissimilarity between its elements as, respectively, AU and EU measures, i.e.,

$$\text{AU}_C := \max_{P \in \mathcal{P}} H[P], \quad (2)$$

$$\text{EU}_C := \max_{P, P' \in \mathcal{P}} \delta[P, P']. \quad (3)$$

When a second-order distribution over the elements of $\mathcal{P}$ is available, we can consider as an alternative a *Bayesian* UQ. This is based on the CS *centre of mass*, obtained by averaging the elements of $\mathcal{P}$ with the second-order probabilities as weights, and similarly averaging the entropies. The average entropy is then taken as an AU measure, while the EU is identified with the difference between the entropy of the centre of mass and the AU (the latter difference being a natural measure of the additional uncertainty introduced by not knowing which element of $\mathcal{P}$ is the “true” one). If there is no reason to consider one element of the CS more likely than another, we can simply assume a uniform distribution over $\mathcal{P}$ and obtain:

$$\text{AU}_B := |\mathcal{P}|^{-1} \sum_{P \in \mathcal{P}} H[P], \quad (4)$$

$$\text{EU}_B := H[|\mathcal{P}|^{-1} \sum_{P \in \mathcal{P}} P] - \text{AU}_B. \quad (5)$$

As both our AU definitions in Eqs. (2) and (4) are based on entropies, their values are also in $[0, 1]$. The maximum AU, i.e., one, is achieved, respectively, when the CS includes at least a uniform PMF, and when the centre of mass is such a PMF. The minimum AU, i.e., zero, is instead achieved by both measures when the CS is a single degenerate PMF.

AU gap. We call *ground-truth* AU (GT-AU) the actual level of aleatoric noise in the data-generating process, that is the entropy of the ground-truth PMF. For single “factual” predictions, this should be the entropy of a degenerate PMF giving GT-AU=0. For *ambiguous* predictive tasks where different states, say n out of the total number of states m , should be regarded as equivalent and then equally likely, we have instead:

$$\text{GT-AU} = n \cdot \left(-\frac{1}{n} \log_m \frac{1}{n} \right) = \log_m n. \quad (6)$$

Having a uniform PMF over the valid options is an assumption that may not hold in general, thus GT-AU should be better intended as an upper bound. Nevertheless, we regard the *AU gap*,

defined as the difference between AU and GT-AU, as a calibration measure. A negative gap represents an *interpretation collapse*, i.e., the model over-commits to one interpretation/option among the valid ones. Vice versa, a positive gap can be achieved only if the model assigns probability mass to a number of states n exceeding the total number m . This means that the model spreads probability beyond what the interpretations say, thus indicating inconsistency or over-splitting.

4 Retrieving PMFs and CSs from LLMs

PMFs by direct retrieval. LLMs treat *tokens*, i.e., chunks of text not necessarily corresponding to words, as the discrete units of both input and output. Tokenisation splits text into a fixed vocabulary of such pieces, typically on the order of 10^4 or 10^5 entries depending on the model. Given a context, an LLM computes a PMF over the next token, and text is generated autoregressively, one token at a time. Because most LLMs expose their (log-)probabilities via API, each output can in principle be paired with a PMF over the vocabulary for single-token outputs, or over the space of possible token sequences for multi-token outputs. Here we focus on QA tasks, where the output is a response. In closed-ended settings, semantically equivalent answers can be identified using simple rule-based techniques, and their probabilities summed accordingly. This procedure reduces the raw, token-level PMF to a low-cardinality one whose states correspond to the possible answers, plus one residual state collecting all remaining alternatives (which typically carries negligible probability mass). In open-ended settings, however, obtaining such a low-cardinality PMF is less straightforward.

PMFs by response clustering. For open-ended QA, say that each question is asked multiple times and the different answers are collected, without extracting the raw probabilities. A *semantic* clustering technique can be used to merge similar answers into k clusters, and then use the cardinalities of each cluster to induce a PMF over the k states of a cluster membership variable. We achieve that by two different clustering techniques. In the first, responses are grouped by bidirectional NLI entailment using a cross-encoder. Two responses r and r' are merged into the same cluster if both probabilities of entailment, say $P(r|r')$ and $P(r'|r)$ exceed a certain threshold τ . As no ground-truth answers are used for this approach, we call it *free* clustering.

Alternatively, we can match each response against the ground-truth valid answers using substring inclusion (score = 1.0) with the cosine-similarity fallback with threshold θ . This leads to a cluster of *correct* responses together with an additional set of incorrect clusters. We call this approach *constrained* clustering. The PMF we induce from both clustering techniques has probabilities proportional to the cardinalities of each cluster.

PMFs by verbalised elicitation. As a third elicitation strategy, rather than sampling (and clustering) free-form responses, we might prompt the model to first enumerate plausible interpretations of the question and assign each a probability before giving a final answer (Figure 1). Parsing the model’s enumerated pairs of interpretation and probability and mapping each interpretation to the closest ground-truth answers using substring matching with sentence-embedding cosine similarity fallback yields a PMF over interpretations directly from the model’s own stated probabilities.

Prompt: *This question may have multiple valid interpretations. First list each plausible interpretation and assign a probability between zero and one reflecting how likely it is the intended meaning (probabilities should sum to roughly 1). Then give your final answer.*

Figure 1: A prompt for verbalised PMF elicitation.

From PMFs to CSs. The techniques described above for retrieving PMFs from LLM responses can be readily iterated, yielding CSs instead of single PMFs. We consider four such methods. Question paraphrases are obtained by asking the model to generate a paraphrase of the original question. For closed-ended questions, we discard paraphrases that lead to a different answer than the one obtained with the original question. A simpler alternative is option permutation, applicable to multiple-choice questions, where the order of the options is shuffled in the prompt; as with paraphrasing, we discard permutations that lead to a different answer. Note that, unlike paraphrasing, this method bounds the maximum number of responses/PMFs by the number of possible permutations (i.e., the factorial of the number of options). For open-ended questions, where the PMF is instead obtained by clustering the sampled responses, different PMFs can be obtained by varying either the sampling temperature of a given model or the model itself. We

call these four variation techniques *paraphrasing*, *permutation*, *temperature* and *model*.

5 Experimental Setup

Datasets. As a benchmark for closed-ended polar questions, we use the development set (3,270 questions) of the popular *BoolQ* dataset (Clark et al., 2019). To also account for multiple-choice questions, we use the validation set (1,221 questions) of *CommonsenseQA* (Talmor et al., 2019), in which each question offers five answer choices, only one of which is correct. For open-ended QA, we instead turn to *AmbigQA* (Min et al., 2020), which provides two question types. *Factual* questions admit exactly one correct answer, expressed through multiple surface-form paraphrases (Figure 2, top); model responses are matched against all such paraphrases. *Ambiguous* questions instead admit several genuinely distinct interpretations, each associated with its own disambiguating constraint and gold answer (Figure 2, bottom). We sample 400 questions of each type; the number of valid interpretations of ambiguous questions ranges from two to four.

Factual question: What is found at the centre of an atom?
Valid answer 1: nucleus
Valid answer 2: nuclei
Valid answer 3: atomic nucleus

Ambiguous question: When did Ohio become part of the United States?
Interpretation 1: When was Ohio admitted as a state?
Valid answer: March 1, 1803
Interpretation 2: When was Ohio’s statehood retroactively recognized?
Valid answer: August 7, 1953

Figure 2: AmbigQA questions.

Models. We consider the four instruction-tuned models listed in Table 1. For conciseness, and without risk of ambiguity, we denote them as *Llama*, *Qwen*, *Gemma2*, and *Gemma3*. For *BoolQ*, we compute $n = 30$ question paraphrases, while for *CommonsenseQA* we collect the responses to all the $5! = 120$ permutations of the answer choices. In both cases, we retain only the responses consistent with the response to the original question. With *AmbigQA*, in the multi-model variation experiments, we generate six responses per question using fixed random seeds at the temperatures in

Table 1. In the temperature-variation experiments, each model generates five responses per question at each of four temperatures $T \in \{0.3, 0.7, 1.0, 1.3\}$, with $\text{top-}p=0.95$ and $\text{max_new_tokens}=64$. As cross-encoder for the free clustering we use instead `nli-deberta-v3-small`. Finally, for verbalised elicitation, we use the model `all-MiniLM-L6-v2` with the same parameters and matching procedure as for constrained clustering.

Model	Temperature	top- p
llama-3.1-8B-instruct	0.80	0.95
qwen2.5-14B-instruct	0.70	0.90
gemma-2-9B-it	0.70	0.95
gemma-3-12B-it	0.70	0.95

Table 1: Models and decoding hyper-parameters.

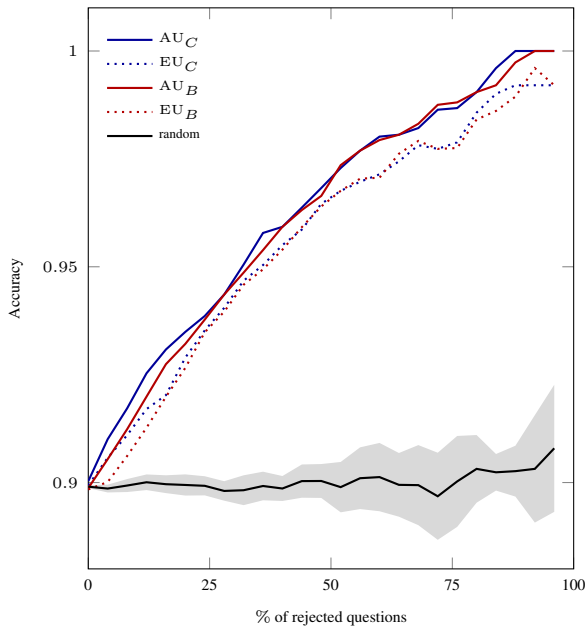


Figure 3: BoolQ accuracy-rejection plot.

6 Results

The code used to generate the empirical results presented in this section is available at github.com/idsia-papers/2026-uncertainNLP.

Accuracy-rejection plots. Let us first evaluate the AU and EU in both their *credal* (Equations (2) and (3)) and *Bayesian* (Equations (4) and (5)) versions. To assess whether our confidence measures are informative in terms of predictive correctness, we consider the *rejection-accuracy* plots for BoolQ and CommonsenseQA. Predictions are ranked by their associated confidence and progressively rejected in order of increasing uncertainty; for each

rejection rate r , we report accuracy on the retained $1-r$ fraction of the test instances. A well-calibrated confidence measure should yield a monotonically increasing curve, reflecting higher accuracy as low-confidence (and presumably less reliable) predictions are excluded. As a baseline we consider a *random* ranking, which is iterated five times, thus also allowing us to display the deviations from the mean accuracies. The results for Gemma2 are depicted in Figures 3 and 4. Experiments with the other models in Table 1 show similar patterns. The monotonic behaviour reflects a good discriminative power, slightly more prominent for AU measures. This is expected as AU captures irreducible ambiguity in the input itself, i.e., precisely the inputs where the model is most likely to be wrong. Notably, both the Bayesian and the credal approaches appear equally effective. For the sake of conciseness, in the rest of this section we only present the results for the Bayesian measures.

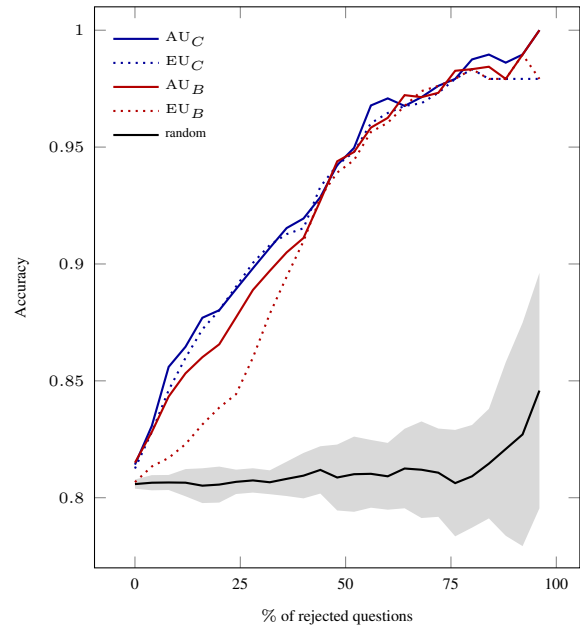


Figure 4: CommonsenseQA accuracy-rejection plot.

Parameter tuning for clustering. As a threshold on the cosine similarity for the constrained clustering used in AmbigQA, we choose the value $\theta=0.65$ after a sensitivity analysis across $\theta \in \{0.45, 0.50, \dots, 0.75\}$ with Llama. As shown in Figure 5, the Bayesian AU on both question types varies by less than 0.03 across the full range, with no elbow points. The threshold of free clustering to decide entailments is instead set to $\tau = 0.5$.

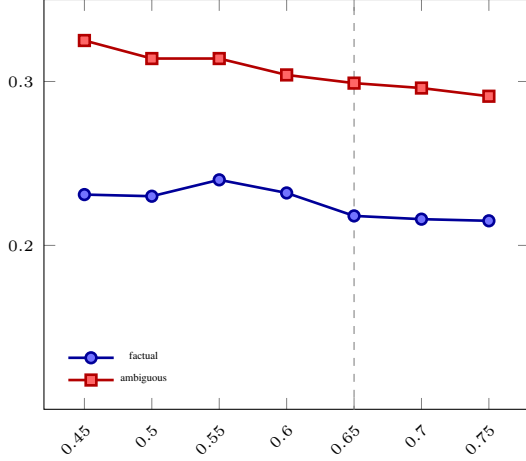


Figure 5: AU sensitivity to similarity threshold θ .

Ambiguous questions. Let us first consider the experimental results for the ambiguous questions of AmbigQA, which are depicted in Table 2.

Table 2: AmbigQA ambiguous question results (GT-AU mean = 1.40 bits).

Var.	Model	Clust.	AU	EU	AU gap
Temperature	Llama	constr.	0.298	0.122	-1.101
	Qwen	constr.	0.197	0.070	-1.203
	Gemma2	constr.	0.203	0.087	-1.196
	Gemma3	constr.	0.103	0.035	-1.297
	Llama	free	1.768	1.322	+0.369
	Qwen	free	1.276	0.879	-0.124
	Gemma2	free	0.961	0.621	-0.439
	Gemma3	free	0.587	0.352	-0.813
Model	All	constr.	0.192	0.368	-1.208
	All	free	1.231	1.663	-0.168

The negative AU gaps for most of the experiments denote a generalised interpretation collapse, i.e., the models seem to be unable to express the intrinsic ambiguity of the questions and concentrate the probability mass on single interpretations. We also note a higher AU with the free clustering with respect to the constrained one. This models a possible over-fragmentation of the invalid interpretations induced by the lack of ground-truth labels. This is also true for the EU. The dominant values of the EU with respect to the AU for model variations and not for temperature variation are also highlighted in Figure 6. Different models, trained on different data and objectives, genuinely disagree on which interpretation should receive higher probability. Each model individually collapses onto *one* interpretation (low AU), but different models collapse onto *different* interpretations (high EU).

For a deeper validation of the results in Table 2,

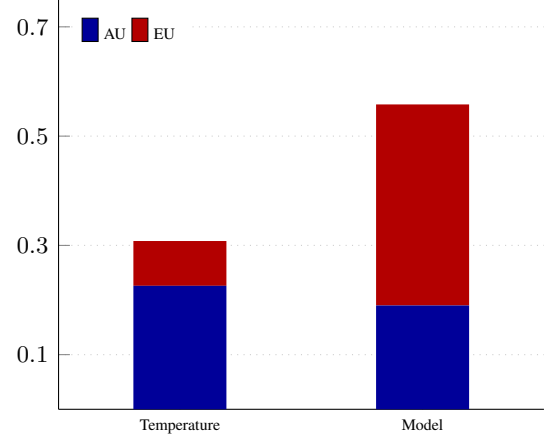


Figure 6: AU/EU decomposition for AmbigQA ambiguous questions for temperature and model variations.

we replace the clustering techniques with the verbalised elicitation on a sample of 50 ambiguous questions for two models. The AU results depicted in Table 3 are notably very similar, and this shows a good robustness of our measure with respect to the elicitation procedure. Not surprisingly, the EU results are less robust, which reflects the fact that a different *epistemic* background is considered for elicitation.

Table 3: Verbalised vs. constrained clustering for a sample of 50 ambiguous questions of AmbigQA.

Model	Elicitation	AU	EU	AU gap
Llama	clustering	0.271	0.130	-1.181
	verbalised	0.215	0.331	-1.237
Qwen	clustering	0.159	0.055	-1.293
	verbalised	0.164	0.259	-1.288

The AU_B values in Table 2 are obtained by averaging per-population entropies over the set $\mathcal{P}$ of temperature or model variations (Equation (4)). To rule out the possibility that this pooling is itself responsible for the observed interpretation collapse, we recompute the AU gap with $\mathcal{P}$ restricted to a single population, $|\mathcal{P}|=1$ (e.g. $T=1.0$, or a single model), which reduces Equation (4) exactly to the standard *semantic entropy* baseline (Kuhn et al., 2023; Farquhar et al., 2024). For constrained clustering, the $|\mathcal{P}|=1$ AU gap remains negative for every model tested, ranging from -1.03 (Llama) to -1.29 (Gemma3) bits at $T=1.0$, closely tracking the pooled values in Table 2 (-1.10 to -1.30). The finding is robust to *which* single population is chosen: across all four temperatures and all four models individually, the $|\mathcal{P}|=1$ gap ranges

from -0.99 to -1.35 bits, never approaching zero. Under free clustering the gap is markedly attenuated or reverses sign for some models (e.g., Qwen: $+0.02$; Llama: $+0.52$), consistent with the over-fragmentation effect noted above. Interpretation collapse is therefore not an artifact of averaging over multiple sampling populations. Note that we measured differences in entropy-based measures in bits by converting the base of the logarithm to two.

Factual questions. The results for factual questions are in Table 5. AU gaps are not reported as, for these questions, GT-AU is zero by construction and AU directly measures model inaccuracy or inconsistency. Gemma3 is the most accurate model (smallest AU). The EU with model variations again substantially exceeds the EU for temperature variations, confirming that inter-model disagreement is a stronger epistemic signal than intra-model temperature variation regardless of question type.

Robustness of the collapse effect. Two additional checks confirm the AU gap is not a measurement artifact. First, as the PMFs we obtain for AmbigQA are based on frequentist estimates from few samples, we might consider the Miller-Madow small-sample bias correction (Miller, 1955) to the entropy, i.e.,

$$\hat{H}[P] := H[P] + \frac{m-1}{2k \ln 2}, \quad (7)$$

where P is the PMF estimated from k samples and $m \leq |\mathcal{V}|$ is the number of states in $\mathcal{V}$ with strictly positive empirical count (which corrects the known downward bias of naive entropy at low N). This correction leaves the gap essentially unchanged (Table 6): at $k = 5$ responses per population, 95.9% of the naive AU gap survives correction, rising to 98.9% at $k = 50$ (Llama, constrained clustering). Second, the AU gap does not merely appear on average; it widens monotonically with the number of annotated gold interpretations (Table 7), from -0.70 bits (2 interpretations) to -2.64 bits (8 interpretations) (Pearson $r = -0.81$, Spearman $r = -0.79$, both $p < 10^{-6}$). Models do not scale their expressed uncertainty with the true ambiguity of the question; instead, the gap between what they express and what the annotation space warrants grows as ambiguity increases.

7 Findings

The experiments detailed in the previous section allow us to answer the research questions formulated

in the introduction.

Regarding RQ1, on the suitability of the ensemble-based UQ measures, our tests on closed-ended QA show good correlation with predictive accuracy, indicating strong potential for the credal AU and EU measures. This aligns with earlier results in purely machine-learning tasks (Hüllermeier and Waegeman, 2021) and, to the best of our knowledge, represents the first successful application of these techniques to LLM tasks.

Regarding RQ2, when answering ambiguous questions such as those in AmbigQA, we would expect LLMs to distribute probability across valid interpretations in proportion to the ambiguity of the annotated answer space. Instead, even after performing UQ and removing EU, we observe a systematic *interpretation collapse* at the AU level as well: all models assign disproportionate probability to a single interpretation, yielding strongly negative AU gaps. This result is robust to the elicitation method, as verbalised techniques yield equally collapsed distributions, confirming that the effect is not an artifact of free-form decoding specifically. This negative finding calls for novel, dedicated approaches to properly address ambiguity in LLM tasks.

Regarding RQ3, we expect AU to dominate for factual questions and EU to dominate for ambiguous ones. This pattern does not emerge under temperature variation, which is commonly used to induce output diversity and is often treated as a proxy for EU. It does emerge, however, under model variation, where EU is two to four times higher than under temperature variation. In practice, temperature variation appears to capture decoding noise rather than knowledge uncertainty, while model diversity is necessary to capture genuine EU. This is in line with the findings of Desai and Durrett (2020), though, to the best of our knowledge, this is the first application of this insight to LLMs in a UQ context.

8 Conclusions

We presented a credal framework for quantifying aleatoric and epistemic uncertainty in LLM responses, based on ensembles obtained through different variation techniques. Our measures successfully capture LLM uncertainty, and the black-box nature of our approach makes it computationally cheap and easy to implement, requiring no access to model internals. Applying these measures to

Table 4: AU gap in bits on the *ambiguous* questions of AmbigQA under constrained clustering, comparing the averaged framework value (Equation (4), $|\mathcal{P}|=4$) against the single-population ($|\mathcal{P}|=1$) baseline at each individual temperature.

Model	Pooled ($ \mathcal{P} =4$)	$T=0.3$	$T=0.7$	$T=1.0$	$T=1.3$
Llama	-1.101	-1.238	-1.140	-1.034	-0.993
Qwen	-1.203	-1.277	-1.208	-1.180	-1.147
Gemma2	-1.196	-1.307	-1.229	-1.150	-1.100
Gemma3	-1.297	-1.346	-1.301	-1.286	-1.254

Table 5: AmbigQA single-factual results.

Variation	Model	AU	EU
Temperature	Llama	0.217	0.075
	Qwen	0.127	0.044
	Gemma2	0.139	0.049
	Gemma3	0.070	0.022
Model	All	0.139	0.233

Table 6: Effects of Miller-Madow correction (Llama3, constrained clustering, subset of 50 ambiguous questions).

m	AU (H)	AU ($\hat{H}$)	Survival [%]
5	-1.180	-1.132	95.9
10	-1.138	-1.104	97.1
20	-1.107	-1.085	98.1
50	-1.069	-1.058	98.9

ambiguous tasks revealed a systematic tendency of LLMs to concentrate probability mass on a single interpretation, a form of interpretation collapse that persists even after accounting for epistemic uncertainty and is robust across elicitation strategies. Finally, our experiments highlighted the importance of multi-model ensembles for properly capturing epistemic diversity, as temperature-based variation alone fails to recover comparable levels of EU.

9 Limitations

Although promising, this is only a preliminary study on the potential of ensemble uncertainty quantification in LLMs, and several limitations remain to be addressed in future work. We list the most pressing ones below.

Further empirical investigation is needed to determine which approach among the credal and the Bayesian is best suited to uncertainty quantification from ensembles of output distributions in practice. The Bayesian approach relies on additional assumptions that may not generally hold, whereas the credal approach may prove overly conservative in open-ended tasks.

Table 7: AU gap by number of annotated gold interpretations (Llama3, constrained clustering, ambiguous questions, $N=400$). Mean/Median columns report the AU gap in bits.

n_{interp}	# Questions	Mean	Median
2	219	-0.703	-0.820
3	108	-1.321	-1.585
4	31	-1.723	-1.820
5	16	-1.963	-2.020
6	15	-2.144	-2.062
7	7	-2.348	-2.627
8	4	-2.638	-2.698

The above discussed interpretation *collapse* represents a serious limitation to the reliability of LLMs when faced with ambiguous tasks. Our measures might provide a sound tool to evaluate this adverse effect, but dedicated mitigation strategies should be devised.

Moreover, our concept of *gap* in the aleatoric uncertainty assumes a uniform prior over the annotated interpretations. In real cases, some interpretations may be more probable than others. For example, for the Ohio question in Figure 2, the 1803 statehood date is the more familiar reading than the 1953 retroactive congressional declaration. If the true prior is skewed, we overstate genuine ambiguity, and the gaps are conservative upper bounds on the collapse effect. A rigorous treatment would require human-annotated interpretation priors.

Finally, our constrained clustering assigns responses to ground-truth interpretation clusters or to an *invalid* bucket when no cluster matches. This bucket conflates two qualitatively different cases. Consider again the Ohio question: a response like “*Ohio was colonized in 1788*” is invalid because factually wrong. But a response like “*The Northwest Ordinance of 1787 first incorporated Ohio’s territory*” is also marked invalid, despite representing a historically defensible alternative reading of the question that was simply not annotated in AmbigQA. Both cases make the aleatoric uncertainty lower, but only the first is a genuine model error.

References

- Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, and 1 others. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.
- Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302. Association for Computational Linguistics.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346. Association for Computational Linguistics.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. 2024. Quantifying aleatoric and epistemic uncertainty: A credal approach. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling (SPIGM)*, Vienna, Austria.
- Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.
- 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](#). Preprint, arXiv:2207.05221.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pages 6402–6413. Curran Associates, Inc.
- Oscar William Lithgow-Serrano, Vani Kanjirang, and Alessandro Antonucci. 2025. Causal understanding by llms: The role of uncertainty. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 208–228.
- Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria*.
- George Miller. 1955. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Muhammad Mubashar, Shireen Kudukil Manchingal, and Fabio Cuzzolin. 2025. [Random-set large language models](#). Preprint, arXiv:2504.18085.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peter Walley. 1991. *Statistical reasoning with imprecise probabilities*. Chapman and Hall.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, volume 2024, pages 23650–23678.

Anita Yang, Krikamol Muandet, Michele Caprio, Siu Lun Chau, and Masaki Adachi. 2026. [Verbalizing LLM’s higher-order uncertainty via imprecise probabilities](#). *Preprint*, arXiv:2603.10396.