Evidential Semantic Entropy for LLM Uncertainty Quantification

Lucie Kunitomo-Jacquin[1], Edison Marrese-Taylor[1,2], Ken Fukuda[1], Masahiro Hamasaki[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
[2]Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
kunitomo-jacquin.lucie@aist.go.jp edison.marrese@aist.go.jp
emarrese@weblab.t.u-tokyo.ac.jp ken.fukuda@aist.go.jp masahiro.hamasaki@aist.go.jp

Abstract

Quantifying uncertainty in large language models (LLMs) is crucial for applications where safety is a concern, as it helps identify factually incorrect LLM answers, commonly referred to as hallucinations. Recently, advancements have been made in quantifying uncertainty, specifically by incorporating the semantics of sampled answers to estimate entropy. These methods typically rely on a normalized probability that is calculated using a limited number of sampled answers. However, we note these estimation methods fail to account for the effects of the semantics that are possible to be obtained as answers, but are not observed in the sample. This is a significant oversight, since a heavier tail of unobserved answer probabilities indicates a higher level of overall uncertainty. To alleviate this issue, we propose Evidential Semantic Entropy (EVSE), which leverages evidence theory to represent both total ignorance arising from unobserved answers and partial ignorance stemming from the semantic relationships among the observed answers. Experiments show that EVSE significantly improves uncertainty quantification performance. Our code is available at: https://github.com/lucieK-J/EvidentialSemanticEntropy.git.

1 Introduction

The advent of large language models (LLMs) has revolutionized numerous fields by demonstrating remarkable capabilities across a diverse array of tasks. However, despite their impressive performance, these models often struggle with reliability issues, particularly due to factual inaccuracies in their outputs. In this context, quantifying their confidence and adjusting them for various tasks can reduce risks and enhance the quality of outputs from LLMs.

However, the uncertainty quantification of LLMs remains challenging since the output possibilities for these models are substantially greater than those of discriminative models. As the generation length increases, the number of potential outcomes grows exponentially, making it unfeasible to evaluate all possible answers (Geng et al., 2024).

We can distinguish two types of uncertainty in LLMs: aleatory uncertainty, stemming from inherent randomness, and epistemic uncertainty, resulting from a lack of information. Following previous works (Nikitin et al., 2024), we aim to quantify a measure of total uncertainty, i.e., aleatory and/or epistemic, as both types of uncertainty contribute to model errors. Among the methods of uncertainty quantification for LLMs, we identify black-box methods, which assume access only to the generations, and white-box methods, which also utilize internal states of the LLM or token-level probabilities. In this paper, we focus on white-box methods utilizing token-level probabilities, based on multiple sampling. Specifically, we explore methods that estimate uncertainty through entropy calculation from a sample of answers generated by the LLM. In this line of research, recent advancements have been made by integrating answer semantics into entropy calculations. Specifically, the semantic entropy (SE) method (Kuhn et al., 2023) takes into account that not all answers convey the same meaning and dramatically enhances the uncertainty quantification performance of the predictive entropy. Indeed, variability in answers with identical meanings does not necessarily indicate high uncertainty. Using their example, when asked, *What is the capital of France?*, if a language model hesitates between "Paris" and "Rome", it indicates a higher level of uncertainty than if it oscillates between "Paris" and "It's Paris".

In this paper, we highlight that multiple sampling-based methods employed for quantifying uncertainty in language models are themselves subject to epistemic uncertainty, as they rely on only a glimpse of the probability distribution of possible answers for practical considerations. This

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7107–7122

epistemic uncertainty is not accounted for in the calculation of SE, significantly impacting the quantification of total uncertainty, particularly when the generated sample size is small. In the SE method, the likelihoods of sampled answers sharing equivalent meaning, i.e., grouped in the same cluster by SE method, are aggregated (semantic likelihood). The issue arises when this semantic likelihood is normalized to sum to 1 for entropy calculation. It is precisely with this normalization step that we lose the crucial information related to epistemic uncertainty arising from the answers that are not observed in the sample. To illustrate this point, we present the running example of this paper in Example 1. To make this example interesting, we intentionally chose probabilities that, when normalized, reduce to the same as in the first scenario.

Example 1 (Running Example) *For the question:* "Sir William Walton's 'Crown Imperial Coronation March' was written for whose coronation?", *we sampled five answers, "george vi", "george VI", "queen elizabeth ii", "queen elizabeth" and "edward viii". Then we consider two scenarios: in the first, we have a lighter tail of unobserved answers, and in the second scenario this tail mass is more consequent* (see Table 1).

This example demonstrates that the SE method can quantify the uncertainty in a strictly identical degree for two cases of very different uncertainty situations.

Our contributions can be summarized as follows.

- We introduce Evidential Semantic Entropy (EVSE) which mobilizes Evidence Theory for modeling two kinds of ignorance that have not yet been considered in entropy-based UQ methods for LLMs: **total ignorance** arising from the unobserved answers, and **partial ignorance** arising from the vagueness of answers that can vary in precision (e.g. "queen elizabeth" versus "queen elizabeth ii".)
- To account for partial ignorance, we introduce a novel clustering method aimed not only at grouping answers with the same meaning but also at identifying relationships between the resulting clusters.
- We show empirical evidence demonstrating the significant superiority of our proposed method for LLMs UQ in the short sequence generation setting, outperforming seven baseline methods. This short sequence generation setting is important for real-time dialogue applications, especially for LLMs deployed on mobile devices, which require a quick and efficient generation approach.

2 Background

Let us denote by x the object about which we quantify uncertainty; in our case study, x refers to the input given to the LLM.

2.1 Semantic Entropy

Let $\mathcal{S}$ represent the set of all possible answers which might be generated by the LLM. Each possible answer in $\mathcal{S}$ consists of a sequence of tokens drawn from a vocabulary set $\mathcal{T}$. The probability of an N-tokens long sequence $s = (t_1, \ldots, t_N)$ is obtained by the product of conditional probabilities of new tokens given past tokens:

$$p(s|x) = \prod_i p(t_i|t_{<i}, x). \tag{1}$$

Instead of basing UQ on the probability of possible LLM answer defined above, Semantic entropy is based on the Semantic likelihood, that is the probability over meanings (Kuhn et al., 2023; Farquhar et al., 2024). These meanings are captured by groups of answers sharing equivalent meaning and called semantic clusters. Formally, the set of semantic clusters $\mathcal{C}$ is a sub-σ-algebra of the event-space of all possible answers $\mathcal{S}$ where the probability of a semantic cluster $c \in \mathcal{C}$ is defined by

$$p(c|x) \quad = \quad \sum_{s \in c} p(s|x). \tag{2}$$

Semantic entropy is defined as the entropy over the probability of semantic clusters:

$$SE(x) = -\sum_{c \in \mathcal{C}} p(c|x) \log p(c|x). \tag{3}$$

As it is not realistic to compute the probability of all semantic clusters in $\mathcal{C}$, semantic entropy is estimated using M answers, denoted $s_1, \ldots, s_M$ sampled from the LLM for a fixed input x. Let C_{obs} denote the set of observed clusters, that is to say clusters of sampled answers. Semantic entropy (3) is calculated using a Rao–Blackwellized Monte Carlo estimator (Farquhar et al., 2024):

$$SE(x) \approx -\sum_{c \in C_{\text{obs}}} p'(c|x) \log p'(c|x)), \tag{4}$$

where the probability of each observed cluster c is normalized as $p'(c|x) = {}^{p(c|x)}/_{\sum_{c \in C_{\text{obs}}} p(c|x)}$ to ensure $\sum_{c \in C_{\text{obs}}} p'(c|x) = 1$ (Farquhar et al., 2024).

Answers s	Likelihood $p(s\|x)$	Semantic likelihood $p(c\|x) = \sum_{s \in c} p(s\|x)$	Normalized Semantic likelihood $p(c\|x)/\sum_{c \in C_{\text{obs}}} p(c\|x)$	Semantic entropy $SE(x)$
"george vi"	0.4	0.5	0.625	2
"george VI"	0.1			
"queen elizabeth ii"	0.1	0.1	0.125	
"queen elizabeth"	0.1	0.1	0.125	
"edward viii"	0.1	0.1	0.125	
$\sum$	**0.8**	**0.8**	1	

(a) Scenario 1 : Answer distribution with lighter tail

Answers s	Likelihood $p(s\|x)$	Semantic likelihood $p(c\|x) = \sum_{s \in c} p(s\|x)$	Normalized Semantic likelihood $p(c\|x)/\sum_{c \in c_{\text{obs}}} p(c\|x)$	Semantic entropy $SE(x)$
"george vi"	0.2	0.25	0.625	2
"george VI"	0.05			
"queen elizabeth ii"	0.05	0.05	0.125	
"queen elizabeth"	0.05	0.05	0.125	
"edward viii"	0.05	0.05	0.125	
$\sum$	**0.4**	**0.4**	1	

(b) Scenario 2 : Answer distribution with heavier tail

Table 1: Two scenarios with equal semantic entropy despite different probabilities of unobserved answers. In scenario 2a, this probability is $1 - \sum_s p(s|x) = 0.2$ (lighter tail). In contrast, scenario 2b features a heavier tail of probability 0.6. The notation descriptions are provided in §A.

2.2 Evidence Theory

The Evidence theory, a.k.a, Dempster-Shafer theory or Belief Functions theory (Shafer, 1976; Smets and Kennes, 1994), is an extension of probability theory, allowing masses of evidence to be allocated to subsets of a frame of discernment Θ for representing both epistemic and aleatoric uncertainties. In the Evidence theory, the uncertain information is represented by a *mass function*, also called *basic belief assignment* (bba), $m : 2^{\Theta} \rightarrow [0, 1]$ satisfying

$$\sum_{A \subseteq \Theta} m(A|x) = 1. \tag{5}$$

For a set $A \subseteq \Theta$, the quantity $m(A)$ is defined as the mass of belief allocated exactly to the set A and not to more specific subsets of A. The quantity, $m(A)$ can be interpreted as the probability that we know nothing more than "$\theta \in A$" (Shafer, 1976) where θ denotes the true element in Θ. We call focal element a subset of Θ which is assigned a strictly positive mass and denote the set of focal elements by $\mathbb{F}_m = \{B \subseteq \Theta | m(B) > 0\}$.

3 Method

Evidential Semantic Entropy (EVSE) consists of five steps. First, a set of answers is sampled from the LLM for the input x. These answers are then clustered, with relationships between clusters also identified. Next, an interpretation within the evidential framework is conducted, where both total and partial ignorance are captured through the construction of a frame of discernment and a mass assignment. This process enables the quantification of the LLM's uncertainty in the evidential framework regarding the input x. A flow chart illustrating EVSE is provided in Figure 1.

Step 1: Answers sampling We begin by sampling M answers from the same model for which we aim to quantify uncertainty regarding the input x, and denote $S_{\text{obs}} \subseteq \mathcal{S}$ the set of sampled answers. Note that $|S_{\text{obs}}| \leq M$ because identical answers might be sampled multiple times. Each sampled answer $s \in S_{\text{obs}}$ has an associated probability $p(s|x)$ which is computed as in Equation 1 without any normalization and accounting for the end-

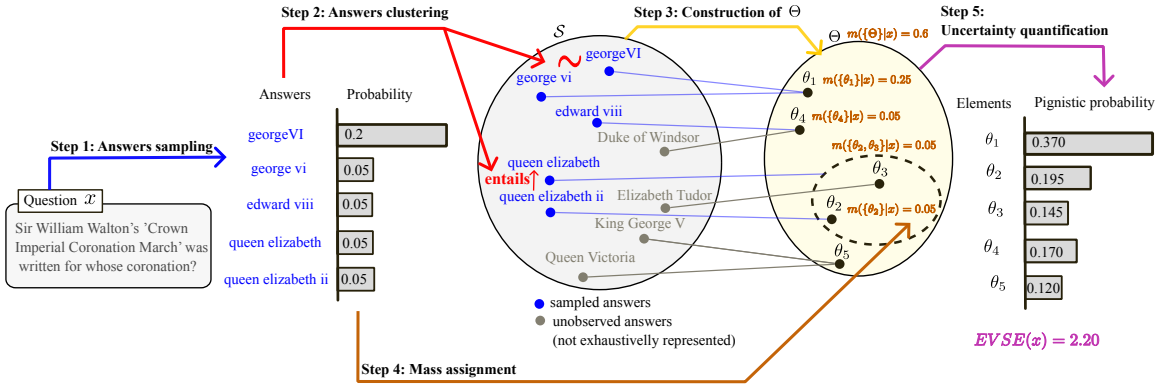


Figure 1: Illustration of EVSE method considering the second scenario of our running example.

of-sequence-token as recommended in Kunitomo-Jacquin et al. (2025) to ensure a proper calculation of the unobserved answers probability.

Step 2: Answers clustering Similarly to Kuhn et al. (2023), our clustering method is based on the entailment relation, denoted $entails$, between answers predicted by a Natural Language Inference (NLI) model. Two answers, s_i and s_j, are considered equivalent, and denoted $s_i \sim s_j$, if both s_i $entails$ s_j and s_j $entails$ s_i are predicted by the NLI model. However, due to language ambiguity and potential NLI model errors, the entailment relation is not perfectly transitive. As a result, clusters obtained in the manner of Kuhn et al. (2023), by successively clustering sequences upon finding bilateral entailment, may exhibit intra-cluster inconsistencies, as acknowledged by the authors.

In our work, we aim not only to cluster answers but also to identify inter-cluster relations to account for the partial ignorance conveyed by an answer with vague meaning, such as "queen elizabeth", from our running example. We define these inter-cluster relations as the extension of the concept of entailment to semantic clusters: c_1 $entails$ c_2 indicates that the meaning conveyed by answers in cluster c_2 is entailed by the meaning conveyed by answers in cluster c_1.

Furthermore, we need to eliminate all cycles, i.e., paths that start and end at the same cluster through entailment relations, to interpret the clusters as distinct subsets of possible answers in the next steps of our method. Therefore, we perform clustering and identification of inter-cluster relations by successively eliminating cycles and prioritizing intra-cluster consistency when merging and updating the inter-cluster relations. Example 2 illustrates this step via our running example.

Example 2 (Running Example) *Let us illustrate our clustering on the running example, where 5 answers are sampled:* s_1 = *"george vi",* s_2 = *"george VI",* s_3 = *"queen elizabeth ii",* s_4 = *"queen elizabeth" and* s_5 = *"edward viii". We assume entailment results are as follows:* s_1 $entails$ s_2, s_2 $entails$ s_1 *and* s_3 $entails$ s_4. *In the first iteration,* s_1 *and* s_2 *are detected to be in a cycle, so they are merged into one cluster. Then, the procedure is complete since there are no more cycles. Finally we obtain the set of clusters* $C_{obs} = \{\{s_1, s_2\}, \{s_3\}, \{s_4\}, \{s_5\}\}$ *and one inter-cluster entailment relation:* $\{s_3\}$ $entails$ $\{s_4\}$.

The detailed procedure and associated pseudo-code for our clustering is provided for reference in the §B and empirical comparison with SE clustering can be found in §C, in our supplementary material.

Step 3: Construction of the frame of discernment We construct our frame of discernment, denoted as $\Theta = \{\theta_1, \ldots, \theta_{|\Theta|}\}$, as the coarsest partition of the set of all semantically mutually exclusive possible meanings such that every observed cluster, i.e., element of C_{obs}, can be represented by a subset $c \subseteq \Theta$ and such that relationships between two clusters c_1 $entails$ c_2, are expressed as strict inclusions $c_1 \subset c_2$. This structure may require the introduction of an element $\theta \in \Theta$ such that $\theta \in c_2$ but $\theta \notin c_1$ to account for answers that entail answers in c_2 but are not semantically equivalent to those in c_1. Furthermore, all other possible answers not sharing meaning with any clusters are represented by one additional element in Θ. Example 3 shows the results of applying this step to our running example.

Example 3 (Running Example) *Figure 2 shows the constructed frame of discernment for the run-*

ning example. Notice that element θ_3 allows us to differentiate between answers meaning "queen Elizabeth" from answers meaning "queen Elizabeth II", and element θ_5 allows us to account for the unobserved, yet possible answers meanings.

Step 4: Mass assignment To represent the uncertain information available for input x, we define a mass function within the evidential framework $m(.|x) : 2^{\Theta} \rightarrow [0, 1]$ as

$$m(c|x) = \sum_{s \in c} p(s|x), \forall c \in C_{\text{obs}}, \quad (6)$$

$$m(\Theta|x) = 1 - \sum_{s \in S_{\text{obs}}} p(s|x). \quad (7)$$

In Equation 6, we allocate the probabilities of the observed answer meanings to their corresponding semantic clusters and in Equation 7, the remaining mass, which corresponds to unobserved answer meanings, is assigned to the total set Θ. This is because the unobserved answers may include answer sharing meaning with any previously observed answers, as well as any other potential answers with different meanings within the set $\mathcal{S}$.

Step 5: Uncertainty quantification We quantify the total uncertainty about input x by means of the entropy proposed by Jousselme et al. (2006), referred as the ambiguity measure of m which satisfies a set of reasonable properties (Jousselme et al., 2006; Urbani et al., 2023):

$$EVSE(x) = -\sum_{\theta \in \Theta} BetP_m(\theta|x) \log BetP_m(\theta|x), \quad (8)$$

where $BetP_m$ is the Pignistic probability distribution (Dubois and Prade, 1982; Smets et al., 1989) defined as follows for $\theta \in \Theta$:

$$BetP_m(\theta|x) = \sum_{A|\theta \in A} \frac{m(A|x)}{|A|}, \quad (9)$$

and where $|A|$ denotes the number of elements in A. Let us note that in the case of null probability of the unobserved answers, that is $m(\Theta|x) = 0$, and if no inter-cluster relations exist, $EVSE(x)$ reduces to $SE(x)$, letting aside the consideration for the practical calculation of sequence probability of Equation 1. Indeed, following recommendations from previous works, we chose not to perform any normalization of the sequence probability and to include the end-of-sequence (EOS) token probability to ensure a properly defined probability distribution on $\mathcal{S}$ (Kunitomo-Jacquin et al., 2025). Example 4 shows the results of applying the entire method to our running example.

Element θ	Pignistic probability $BetP_m(\theta\|x)$	Evidential semantic entropy $EVSE(x)$
θ_1	0.540	1.83
θ_2	0.190	
θ_3	0.090	
θ_4	0.140	
θ_5	0.040	

(a) Scenario 1 : Answer distribution with lighter tail

Element θ	Pignistic probability $BetP_m(\theta\|x)$	Evidential semantic entropy $EVSE(x)$
θ_1	0.370	2.20
θ_2	0.195	
θ_3	0.145	
θ_4	0.170	
θ_5	0.120	

(b) Scenario 2 : Answer distribution with heavier tail

Table 2: Pignistic probability distributions and evidential semantic entropy for the two scenarios of Table 1. In scenario 1 (2a), $m(\{\theta_1\}|x) = 0.5, m(\{\theta_2\}|x) = m(\{\theta_2, \theta_3\}|x) = m(\{\theta_4\}|x) = 0.1$ and $m(\{\Theta\}|x) = 0.2$ (lighter tail). In scenario 2 (2b), $m(\{\theta_1\}|x) = 0.25, m(\{\theta_2\}|x) = m(\{\theta_2, \theta_3\}|x) = m(\{\theta_4\}|x) = 0.05$ and $m(\{\Theta\}|x) = 0.6$ (heavier tail).

Example 4 (Running Example) *Table 2 shows that our representation effectively distinguishes between scenarios that represent different levels of uncertainty regarding the input x: when the distribution has a heavier tail of unobserved answers, the evidential semantic entropy is higher.*

3.1 Complexity Analysis

The most computationally intensive step is the answers clustering step, which involves answer clustering and peaks at $O(M^3)$. This complexity arises from the cycle detection operations, which can reach $O(M^2)$ and may be repeated up to M times if all answers need to be merged.

In contrast, the complexity of the uncertainty quantification step (Equations 8-9) within the evidence framework is negligible, requiring only $O(|\Theta| \times |\mathbb{F}_m|)$ operations, where $\mathbb{F}_m$ is the set

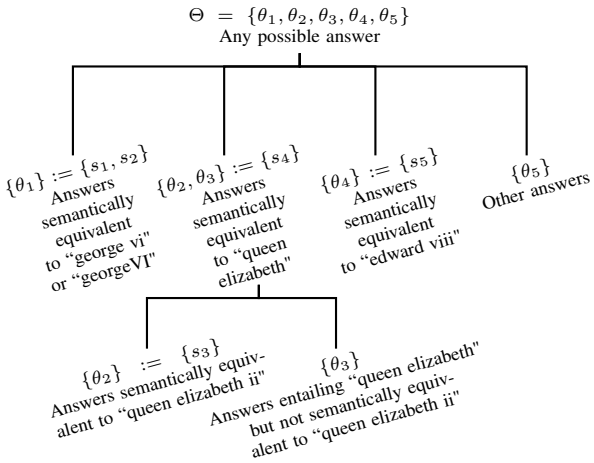


Figure 2: Interpretation of different kinds of answers through subsets of Θ.

of focal elements defined in section 2.2. This complexity is upper bounded by the worst-case scenario of $O(M^2)$ operations, which occurs when all clusters are nested and consist of single answers (in this extreme case $|\Theta| = 2M$ and $|\mathbb{F}_m| = M + 1$).

For comparison, previous methods such as Semantic Entropy (SE) complexity is $O(M^2)$, while Kernel Language Entropy (KLE)'s complexity is $O(M^3)$.

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments were performed on four datasets. Two are general knowledge datasets, TriviaQA (Joshi et al., 2017) and SQuAD (Rajpurkar et al., 2018), one contain general domain questions from Google search, Natural Questions (NQ) (Kwiatkowski et al., 2019), and one dataset contains natural language math problems (SVAMP) (Patel et al., 2021).

Models. Our experiments focus on the uncertainty quantification of 5 families of LLMs. We select 2 model sizes for each family where available, and test both the base and instruction-tuned versions. Concretely, we work with LLama 2 models (Touvron et al., 2023) with 7B and 13B parameters (*Llama-2-7b*, *Llama-2-7b-chat*; *Llama-2-13b*, *Llama-2-13b-chat*), the Falcon models (Almazrouei et al., 2023) with 7B and 40N parameters (*falcon-7b*, *falcon-7b-instr*; *falcon-40b*, *Falcon-40b-Instr*, and Mistral models (Jiang et al., 2023) with 7B parameters (*Mistral-7B-v0.1*, *Mistral-7B-Instr*). In terms of NLI models, we used DeBERTa-Large-MNLI (He et al., 2020) which offers an efficient and accurate alternative.

Sampling We conducted our sampling using a prompting method to obtain short sequence type answers, as shown in Figure 4.

Sampling prompt

Answer the following question as briefly as possible.
{Question}

Figure 4: Prompt for our LLMs, requesting short sequence type answers, where {Question} is a placeholder for the content of the actual question.

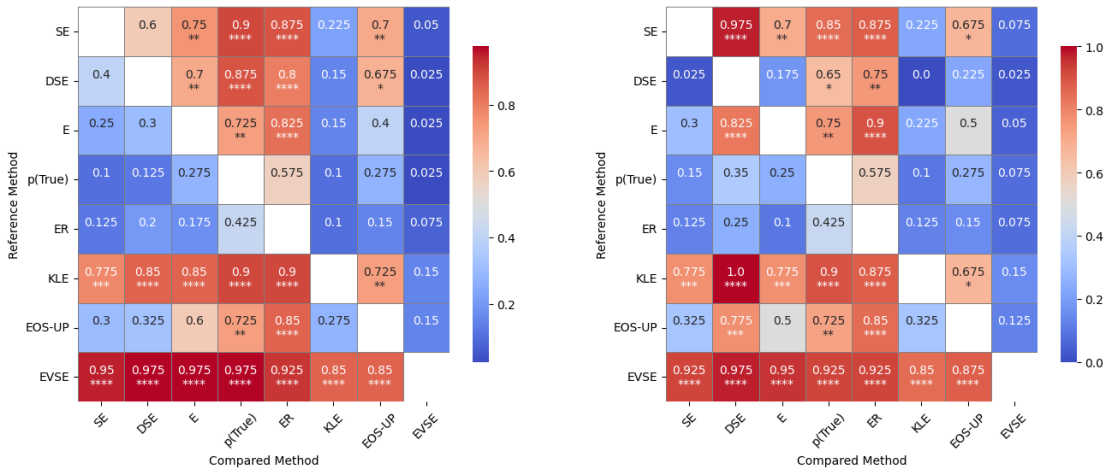
Unless specified, we used $M = 5$ number of samples. Following the methodologies of previous studies (Farquhar et al., 2024; Nikitin et al., 2024), we employed top-K sampling with $K = 50$ and nucleus sampling with $p = 0.9$ at a temperature of $T = 1$.

In line with previous works, we evaluated the model's accuracy by sampling an additional answer at a lower temperature ($T = 0.1$). To prevent getting empty answer, we provided 5-few shots to the model. We used another large language model, gpt4o-mini, to compare this answer at a lower temperature with the ground truth answers from the datasets. The prompts for checking answers correctness are provided in §D.

Metrics We evaluate uncertainty quantification methods by measuring their ability in predicting model output accuracy using the Area under the Receiver Operating Curve (AUROC). In line with previous studies, we also evaluate the uncertainty quantification methods with the Area Under the Accuracy-Rejection Curve (AUARC) which computes the area under the rejection accuracy curve for all possible thresholds (Nadeem et al., 2009). All reported scores in terms of AUROC or AUARC in this paper are calculated using 500 pairs of input questions and answers from a given dataset.

Baselines We compare our method against a wide variety of baselines selected from previous work, which we briefly introduce below.

- **Semantic entropy (SE)** (Kuhn et al., 2023) as presented in Equation 4.
- **Discrete Semantic Entropy (DSE)** (Kuhn et al., 2023; Farquhar et al., 2024), which



(a) AUROC (b) AUARC

Figure 3: Summary of the 40 experiments. Each cell contains the fraction of experiments where a reference method (in rows) outperforms a comparative method (in columns). When applicable, the significance level of the outperformance assessed with a binomial statistical significance test is indicated: * for p-value < 0.05, ** for p-value < 0.01, and *** for p-value < 0.001.

consists in a variant of SE where cluster probabilities are approximated by $p(c|x) \approx |\{s, s \in C\}|/M$.

- **Token predictive Entropy (E)** (Malinin and Gales, 2020), the predictive entropy of the output distribution $H(x) = -1/M \sum_{s=1}^{M} \log p(s_m|x)$.
- **p(true)** (Kadavath et al., 2022), a method sampling M answers to use them as brainstormed answers in a prompt asking if the highest probability answer is true (option "a") or false (option "b"). The probability that the LLM answers "a" to the multiple-choice question is used as a measure of uncertainty.
- **Embedding Regression (ER)**, previous used as a strong baseline by Farquhar et al. (2024) and originally inspired by Kadavath et al. (2022). It consists of training a logistic regression classifier to assess if the model provided the correct answer, relying on the final embedding of the LLM.
- **Kernel Language Entropy (KLE)** (Nikitin et al., 2024), which encodes semantic similarity of LLM outputs using positive semidefinite kernels and quantifies uncertainty with von Neumann entropy, considering pairwise semantic dependencies for finer uncertainty estimates. We use the best instantiation of their methods namely K_{HEAT} with default choices parameters.
- **EOS-Inclusive Unobserved answers Probability (EOS-UP)** consists of quantifying the LLM uncertainty using the probability of unobserved answers $\mathbb{P}(\overline{S_{\text{obs}}}|x)$ calculated as recommended in Kunitomo-Jacquin et al. (2025).

Further experiments details are provided in §D.

4.2 Results

Overall Performance The main experiments consist in evaluating our method, EVSE, and our selected baselines via AUROC and AUARC over the 4 datasets with all our considered models. The overall results, presented in Figure 3, show that our methods clearly outperforms all the other baselines in terms of win rate. Furthermore, these results are statistically significant with a p-value < 0.01 according to a binomial statistical test considering win rates over 40 experiments (4 datasets $\times$ 10 models). Detailed experiments for two selected models are provided in Table 3. The same table for the remaining 8 models is provided in §E, in our supplementary material where we also provide SE and EVSE values on four concrete examples. 6

Influence of instruction-tuning We conducted the same analysis of win rates shown in Figure 3 separately for sets of base and instruction-tuned models and reported the full results in §E. These results shows that the better performance of EVSE is preserved even if we restrict the analysis to instruction-tuned models exept for KLE were outperformance is not significant. For base models,

		NQ		SQuAD		SVAMP		TriviaQA	
		AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Llama-2-7b	SE	75.5 ±2.5	48.3 ±3.2	**77.6** ±2.9	33.8 ±3.1	86.0 ±2.3	62.9 ±3.0	81.7 ±2.4	79.4 ±2.0
	DSE	75.5 ±2.5	45.0 ±3.0	75.9 ±3.0	29.6 ±2.9	84.5 ±2.3	57.2 ±3.1	81.1 ±2.3	77.9 ±1.9
	E	76.7 ±2.6	49.1 ±3.1	67.3 ±3.7	29.5 ±3.1	86.8 ±2.2	63.3 ±3.2	79.9 ±2.4	79.0 ±1.9
	p(True)	63.9 ±3.1	40.9 ±3.2	58.6 ±3.4	23.9 ±2.7	76.4 ±2.9	56.0 ±3.3	72.2 ±2.7	75.2 ±2.3
	ER	57.2 ±3.0	35.4 ±3.1	52.1 ±3.6	21.2 ±2.5	79.4 ±2.7	58.4 ±3.3	65.0 ±3.0	70.8 ±2.6
	KLE	76.3 ±2.6	48.6 ±3.1	77.2 ±2.9	33.7 ±3.2	86.4 ±2.2	63.1 ±3.1	81.5 ±2.4	79.3 ±2.0
	EOS-UP	**78.2** ±2.4	49.0 ±3.0	71.2 ±3.2	30.9 ±3.2	**89.2** ±1.9	**64.6** ±3.2	80.0 ±2.4	78.7 ±2.0
	EVSE	77.5 ±2.5	**49.1** ±3.0	77.3 ±3.0	**33.8** ±3.2	88.6 ±2.1	64.4 ±3.1	**83.2** ±2.2	**80.4** ±1.9
falcon-40b-instr	SE	72.7 ±2.6	56.3 ±2.9	77.4 ±2.8	37.6 ±3.2	84.0 ±2.2	75.0 ±2.4	78.6 ±3.0	83.4 ±1.8
	DSE	72.3 ±2.6	53.2 ±2.9	76.5 ±2.8	34.0 ±3.0	82.5 ±2.3	71.3 ±2.6	79.5 ±2.9	83.3 ±1.7
	E	71.8 ±2.6	55.3 ±3.1	72.3 ±3.1	35.5 ±3.2	82.5 ±2.4	74.8 ±2.5	75.0 ±2.8	82.8 ±1.7
	p(True)	69.6 ±2.6	53.1 ±2.9	64.5 ±3.3	30.9 ±3.0	74.8 ±2.8	70.8 ±2.8	77.1 ±2.8	84.1 ±1.5
	ER	63.4 ±2.9	48.8 ±3.1	63.8 ±3.2	29.4 ±3.0	82.7 ±2.4	75.0 ±2.4	68.0 ±3.1	80.4 ±2.0
	KLE	74.0 ±2.6	56.7 ±2.9	77.3 ±2.9	37.6 ±3.2	84.2 ±2.3	75.3 ±2.5	79.1 ±3.1	83.5 ±1.7
	EOS-UP	67.3 ±2.8	52.9 ±3.1	75.7 ±3.0	36.7 ±3.2	84.5 ±2.3	75.7 ±2.4	72.5 ±3.0	82.0 ±1.9
	EVSE	**74.0** ±2.4	**56.9** ±2.9	**78.6** ±2.8	**38.5** ±3.3	**85.0** ±2.2	**75.8** ±2.4	**80.4** ±2.9	**84.1** ±1.8

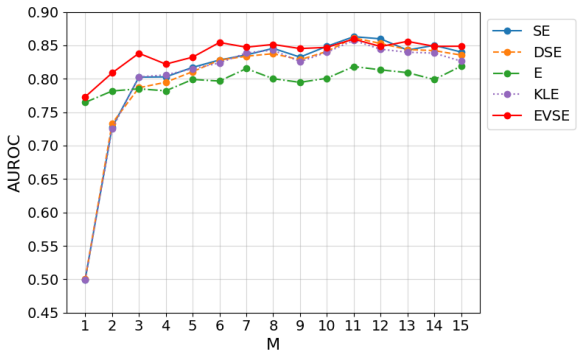
Table 3: Detailed experimental results of the seven baseline methods and ours (EVSE) on four datasets for the uncertainty quantification of two LLMs. AUROC and AUARC scores (×100) are reported to one decimal place, with the best performance highlighted in bold, and second best performance underlined. Displayed confidence intervals are from 1,000 bootstrap resamples.

our method remains superior for all compared methods.

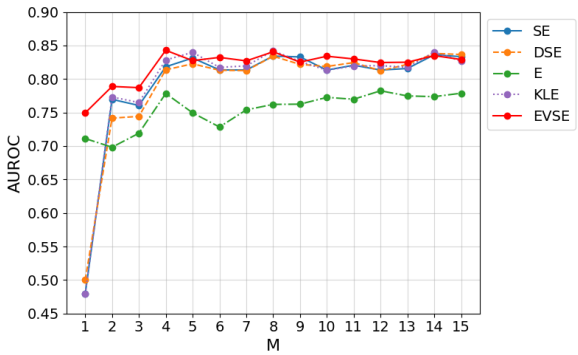
Influence of sample size Our proposed method shows improved robustness when sample size is limited, as depicted in Figure 5. This significant advantage of our approach is obtained by our expressive representation of the LLM uncertainty. By effectively capturing the total ignorance arising from the probability of unobserved answers, our method accounts for the greater uncertainty typically found in scenarios with less samples available. In other words, in contrast to other entropy-based methods, we model epistemic uncertainty, i.e., arising from a lack of information, that is inherent to these methods that have access to only a portion of the probability distributions of all possible answers in $\mathcal{S}$.

Ablation study To justify the choice of the Pignistic probability for uncertainty estimation (Equation 8) in the way of Jousselme (Jousselme et al., 2006), we conducted an ablation study comparing our approach with four alternative measures in the Evidence theory. These methods were selected based on their relevance and reasonable computational complexity:

- **EVSE$_{\text{Höhle}}$**, EVSE with Entropy of Höhle (Höhle, 1982) (instead of Jousselme's), $EVSE_{Höhle}(x) = -\sum_{A \in \mathbb{F}_m} m(A|x) \log Bel(A|x)$ where $Bel(A|x)$ is belief of A, $Bel(A|x) =$



(a)



(b)

Figure 5: Influence of the sample size for uncertainty quantification in terms of AUROC of entropy-based methods on TriviaQA dataset for model Llama-2-7b (5a) and model falcon-40b-Instr (5b).

$\sum_{B \subseteq A} m(B|x)$.

- **EVSE$_{\text{Yager}}$**, EVSE with Entropy of Yager (Yager, 1983), $EVSE_{Yager}(x) = -\sum_{A \in \mathbb{F}_m} m(A|x) \log Pl(A|x)$, where $Pl(A|x)$ is the plausibility of A, $Pl(A) = \sum_{B \cap A \neq \emptyset} m(B|x)$.

- **EVSE$_{\text{Nguyen}}$**, EVSE with Entropy of Nguyen (Nguyen, 1987), $EVSE_{Nguyen}(x) = -\sum_{A \in \mathbb{F}_m} m(A|x) \log m(A|x)$.

- **EVSE$_{\text{Dubois \& Prade}}$**, EVSE with Entropy of Dubois and Prade (Dubois and Prade, 1987), $EVSE_{Dubois\ \&\ Prade}(x) = -\sum_{A \in \mathbb{F}_m} m(A|x) \log |A|$.

We compared these alternatives with our method in the same setting as we did for the overall results and reported the win rates in Figure 6. This shows that the outperformance of the proposed EVSE method is significant over each of the alternatives.

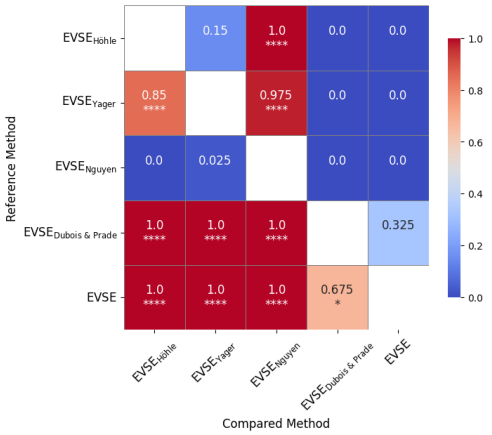
Our interpretation of why taking the entropy of the Pignistic probability performs better is as follows. In standard question-answering tasks, correctness is a binary outcome (0 or 1), influenced by both epistemic and aleatoric uncertainties. A measure of total uncertainty, such as the entropy of the pignistic probability, is therefore particularly effective. In contrast, Höhle, Yager, and Nguyen primarily capture discord which is a source of aleatoric uncertainty, while Dubois & Prade focus on non-specificity which is associated with epistemic uncertainty (Urbani et al., 2023).

5 Conclusion

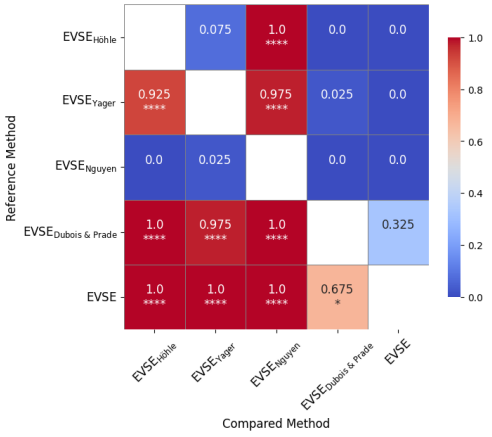
We have introduced a novel method for quantifying the uncertainty of language models, which falls within the category of white-box approaches based on a sample of answers from the LLM.

We showed that the representation of total and partial ignorance arising from the unobserved answers and the answers vagueness, can be integrated in an entropy-based UQ method. To achieve this integration, we mobilized Evidence Theory, which allows capturing the two aforementioned types of uncertainty. Our experiments show that our proposed model can significantly outperform previous work in the short sequence generation setting considered in this paper.

In our future work, we consider studying the influence of LLM calibration and longer sequence on uncertainty quantification.



(a) AUROC



(b) AUARC

Figure 6: Summary of the 40 experiments for the ablation study. Results are presented in the same way as in Figure 3.

Acknowledgements

This paper is based on results obtained from a project, JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Limitations

Restriction to short sequence generation Our experiments demonstrate that EVSE can outperform previous works in the case of short sequences, as guided by our considered sampling prompt. However, since longer sequences tend to exhibit lower probabilities, a long sequence setting leads to an overestimation of the total uncertainty captured by $m(|\Theta|)$. To obtain the same result in this setting, it would be necessary to mitigate the long-sequence probability bias. Approximations, such as sequence length normalization (Malinin and Gales, 2020), are not applicable because they do not preserve the sum of all probability values to be 1.

Restriction to white-box setting Unlike other white-box methods, such as SE of KLE, which can also work in a black-box setting, EVSE requires the probability of the sequences to deduce the probability of the unobserved, which, by definition, are not sampled. For this reason, EVSE could not be adapted to function in the black-box setting which access model generations only.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. arXiv preprint arXiv:2311.16867.

Didier Dubois and Henri Prade. 1982. On several representations of an uncertain body of evidence. In IFAC Symposium on Theory and Application of Digital Control (IFAC 1982), pages 167–181. North Holland.

Didier Dubois and Henri Prade. 1987. Properties of measures of information in evidence and possibility theories. Fuzzy sets and systems, 24(2):161–182.

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

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Ulrich Höhle. 1982. Entropy with respect to plausibility measures. In Proc. of 12th IEEE Int. Symp. on Multiple Valued Logic, Paris, 1982.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

A-L Jousselme, Chunsheng Liu, Dominic Grenier, and Éloi Bossé. 2006. Measuring ambiguity in the evidence theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 36(5):890–903.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664.

Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, and Ken Fukuda. 2025. On the role of unobserved sequences on sample-based uncertainty quantification for llms. Preprint, arXiv:2510.04439.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:452–466.

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

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with

a reject option. In Machine Learning in Systems Biology, pages 65–81. PMLR.

Hung T Nguyen. 1987. On entropy of random sets and possibility distributions. The Analysis of Fuzzy Information, 1:145–156.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. Advances in Neural Information Processing Systems, 37:8901–8929.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2080–2094, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Glenn Shafer. 1976. A mathematical theory of evidence, volume 42. Princeton university press.

Philippe Smets and Robert Kennes. 1994. The transferable belief model. Artificial intelligence, 66(2):191–234.

Philippe Smets and 1 others. 1989. Constructing the pignistic probability function in a context of uncertainty. In UAI, volume 89, pages 29–40.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Michele Urbani, Gaia Gasparini, and Matteo Brunelli. 2023. A numerical comparative study of uncertainty measures in the dempster–shafer evidence theory. Information Sciences, 639:119027.

Ronald R Yager. 1983. Entropy and specificity in a mathematical theory of evidence. International Journal of General System, 9(4):249–260.

A Glossary

The main notations used in this paper are described in Table 4.

Notation	Description
$\mathcal{S}$	The set of all possible answers.
S_{obs}	The set of sampled answers.
$\mathcal{C}$	The set of semantic clusters.
C_{obs}	A set of observed clusters.
s	A sequence.
c	A cluster.
x	The input given to the LLM.
$\mathcal{T}$	The vocabulary set.
N	A sequence length.
M	The number of sampled answers.
Θ	A frame of discernment.
m	A mass function.
$\mathbb{F}_m$	The set of focal elements
$\mathbf{E}$	entailment matrix
$\mathbf{R}$	set of inter-cluster relations

Table 4: Notation glossary.

B Detailed clustering procedure

Our clustering method is described below and the associated pseudo-code is provided in Algorithm 1.

- First, we predict and store entailment for each pair of answers in an entailment matrix denoted $\mathbf{E} = (e_{ij})_{1 \leq i,j \leq |S_{\text{obs}}|}$ where $e_{ij} = 1$ if $s_i \; entails \; s_j$ and 0 otherwise.
- Then, each answer is treated as a separate cluster for initialization. Using depth-first search, cycles of entailment relation between clusters are successively detected in the entailment matrix.
- At each iteration, two clusters detected in a cycle are merged, and the entailment matrix is updated via the definition of a new matrix $\mathbf{E}'$ in which only the shared entailment relationships are preserved in order to enhance intra-cluster consistency. As clusters are merged, the entailment matrix is dynamically updated to reflect these changes. Redundant rows and columns, corresponding to obsolete clusters, are removed.
- Once the process of merging clusters is complete and no further cycles of entailment exist, we proceed to identify and store inter-cluster entailment relations.

C SE Clustering vs. EVSE Clustering

Both SE and EVSE clustering methods represent compromises to address inconsistencies arising from the imperfect transitivity of the entailment relation. To examine the differences in behavior

Algorithm 1 Sampled answers clustering and identification of inter-cluster relationships

Input: $\mathbf{E} = (e_{ij})_{0 \leq i,j \leq |S_{\text{obs}}|}$ // Entailment matrix of dimension $|S_{\text{obs}}| \times |S_{\text{obs}}|$
Outputs: $\mathbf{c} = (c_0, c_1, \ldots, c_{|S_{\text{obs}}|-1})$ and $\mathbf{R}$ // List of cluster assignments for each answer and list of inter-cluster relations

// Initialization: each answer is treated as a separate cluster
$l \leftarrow |S_{\text{obs}}|$
$\mathbf{c} \leftarrow (0, \ldots, l-1)$
$\mathbf{E}' \leftarrow \mathbf{E}$
$l' \leftarrow l$

// Successively merge two clusters when they are detected in a cycle
while there exists a cycle in $\mathbf{E}'$ **do**
  $(i, j) \leftarrow$ find a cycle in $\mathbf{E}'$ using depth-first search
  for $k = 0$ to $l - 1$ **do**
    if $c_k = j$ **then**
      $c_k \leftarrow i$ // Merge clusters j into i
    end if
  end for

  // Keep only the entailment relationships shared by both clusters i and j
  for $k = 0$ to $l' - 1$ **do**
    if $e'_{i,k} \neq e'_{j,k}$ **then**
      $e'_{i,k} \leftarrow 0$
    end if
    if $e'_{k,i} \neq e'_{k,j}$ **then**
      $e'_{k,i} \leftarrow 0$
    end if
  end for

  // Update $\mathbf{E}'$ and $\mathbf{c}$ by removing information related to obsolete cluster j
  Remove row j from $\mathbf{E}'$
  Remove column j from $\mathbf{E}'$
  $l' \leftarrow l' - 1$
  for $k = 0$ to $l - 1$ **do**
    if $c_k > j$ **then**
      $c_k \leftarrow c_k - 1$ // Adjust cluster indexes to fit $\mathbf{E}'$
    end if
  end for
end while

// Retrieve inter-cluster relationships
$\mathbf{R} \leftarrow$ empty list
for $i = 0$ to $l' - 1$ **do**
  for $j = 0$ to $l' - 1$ **do**
    if $i \neq j$ **and** $e'_{i,j} = 1$ **then**
      Add $c_i \; entails \; c_j$ to $\mathbf{R}$
    end if
  end for
end for

between our clustering and SE clustering with respect to these inconsistencies, we define two types of inconsistency. First intra-cluster inconsistency (intraCI) counts the number of pairs of sequences belonging to the same cluster but which do not entails each other bilaterally:

$$\text{intraCI} = |\{(s_i, s_j) \in C \times C, C \in \mathcal{C} \mid s_i \not\sim s_j\}|, \tag{10}$$

where $s_i \not\sim s_j$ means that at least one of the two entailments $s_i \; entails \; s_j$ or $s_j \; entails \; s_i$ is false. Second, we define the inter-cluster inconsistency (interCI) which counts the pairs of sequences in different clusters despite they entail each other:

$$\text{interCI} = |\{(s_i, s_j) \in C_i \times C_j, C_i \neq C_j, C_i, C_j \in \mathcal{C} | s_i \sim s_j\}|. \tag{11}$$

The comparison between our clustering method and the SE method clustering is presented in Table 5. We observe that for both considered models, we effectively prioritize the minimization of intra-cluster inconsistencies. However, the trade-off is a higher presence of inter-cluster inconsistencies compared to SE clustering.

	Llama-2-7b		falcon-40b-instr	
	Avg. intraCI	Avg. interCI	Avg. intraCI	Avg. interCI
SE	0.006	0.024	0.012	0.028
EVSE	0.002	0.030	0.004	0.028

Table 5: Average intra-cluster inconsistency (Avg. intraCI) and average inter-cluster inconsistency (Avg. interCI) for SE clustering and EVSE clustering (ours) for model falcon-40b-instruct and TriviaQA dataset.

D Additional Experiments Details

Hardware All experiments were run on NVIDIA A100 80GB PCIe GPUs. For reference, the experiment on all eight methods using the TriviaQA dataset took approximately 3 hours, utilizing the *falcon-40b-Instruct* model.

Licenses The datasets used in this paper are released under Apache 2.0 (TriviaQA), CC BY-SA 4.0 (SQuAD), MIT (SVAMP), and CC BY-SA 3.0 (NQ).

Prompts To check the correctness of the answers, we used the same prompts as previous studies presented in Figure 7.

Prompt for checking answers correctness (single answer)

*We are assessing the quality of answers to the following question: **{question}** \n The expected answer is: **{correct_answer}**. \n The proposed answer is: **{predicted_answer}** \n Within the context of the question, does the proposed answer mean the same as the expected answer? \n Respond only with yes or no.\n Response:*

Prompt for checking answers correctness (multiple answers)

*We are assessing the quality of answers to the following question: **{question}** \n The following are expected answers to this question: **{correct_answers}**. \n The proposed answer is: **{predicted_answer}** \n Within the context of the question, does the proposed answer mean the same as any of the expected answers? \n Respond only with yes or no.\n Response:*

Figure 7: Prompts fed to the model in our experiments when providing a single (top) and many correct answers (bottom), where **placeholders** are denoted in bold.

E Additional results

Detailed experiments for the remaining 8 models is provided in Table 7. In Table 6, SE and EVSE values on four concrete examples are presented.

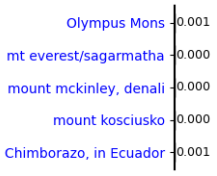
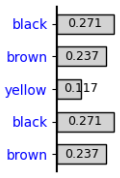
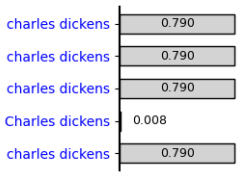
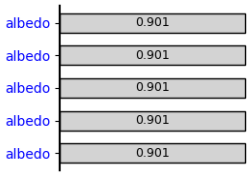
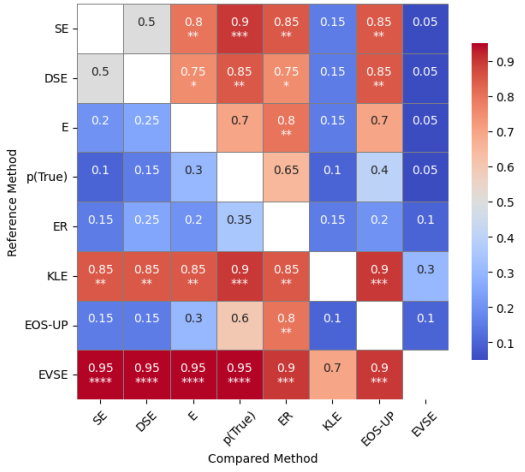
Question ($= x$)	True answer	Model prediction	Sampled answers	$m(\Theta\|x)$	SE	EVSE
What mountain's peak is the farthest point from Earth's centre/center?	chimborazo	mount everest		0.998	1.46	2.58
On a map of the London Underground, what colour is the Northern Line?	Black	brown		0.374	0.95	1.85
Which author created the fictional character Sergeant Cuff?	wilkie collins	charles dickens		0.202	0	0.47
What is the astronomical term for the measure of the reflective ability of a heavenly body?	albedo	albedo		0.099	0	0.28

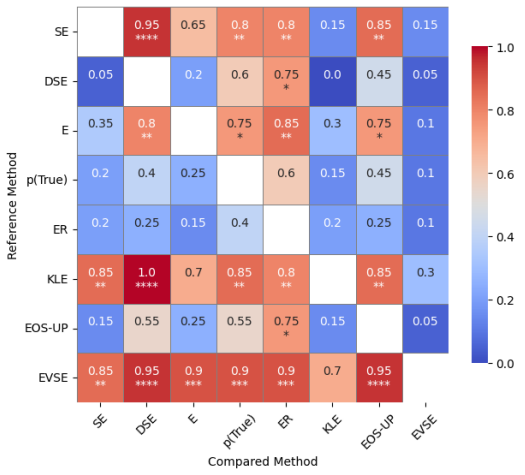
Table 6: Uncertainty quantification values for SE and EVSE methods on four examples from the TriviaQA dataset, using the falcon-40b-instr model.

		NQ		SQuAD		SVAMP		TriviaQA	
		AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Llama-2-13b-chat	SE	71.5 ±2.5	51.9 ±3.1	76.3 ±2.8	33.3 ±3.2	83.9 ±2.3	66.3 ±2.9	74.6 ±2.8	77.9 ±2.0
	DSE	71.6 ±2.5	50.3 ±2.9	77.3 ±2.6	32.0 ±2.9	83.6 ±2.2	63.7 ±3.0	75.1 ±2.7	77.9 ±2.1
	E	71.4 ±2.6	**52.8** ±3.0	75.1 ±2.9	32.9 ±3.2	83.7 ±2.4	67.1 ±2.9	73.8 ±3.0	78.3 ±2.1
	p(True)	69.6 ±2.7	51.9 ±3.0	78.0 ±2.7	34.3 ±3.2	81.5 ±2.3	66.0 ±3.0	**82.1** ±2.3	**83.0** ±1.6
	ER	62.5 ±2.7	47.0 ±3.0	62.6 ±3.4	27.2 ±2.8	**88.7** ±2.0	**69.4** ±2.8	65.3 ±3.1	74.8 ±2.5
	KLE	**72.5** ±2.4	52.5 ±3.0	77.7 ±2.5	33.8 ±2.9	84.8 ±2.3	66.9 ±2.9	75.0 ±2.8	78.0 ±2.0
	EOS-UP	62.7 ±2.9	46.7 ±3.1	76.2 ±2.9	33.9 ±2.9	77.9 ±2.7	63.3 ±3.1	68.9 ±2.9	77.0 ±2.3
	EVSE	72.2 ±2.6	51.5 ±3.0	**79.1** ±2.6	**35.4** ±3.2	85.6 ±2.3	67.3 ±3.1	77.4 ±2.8	79.8 ±2.1
Llama-2-13b	SE	74.5 ±2.5	59.2 ±2.8	74.1 ±3.1	34.7 ±3.3	86.8 ±2.0	71.5 ±2.5	82.2 ±2.7	83.5 ±1.7
	DSE	74.9 ±2.5	55.7 ±2.8	74.2 ±3.1	31.0 ±2.9	85.9 ±2.0	67.5 ±2.8	82.9 ±2.5	83.0 ±1.7
	E	73.7 ±2.4	57.3 ±3.0	71.5 ±3.3	33.7 ±3.2	87.0 ±1.9	71.4 ±2.6	82.4 ±2.5	84.0 ±1.6
	p(True)	73.4 ±2.5	57.4 ±2.9	65.1 ±3.3	29.8 ±3.0	68.8 ±3.0	59.9 ±3.4	84.1 ±2.2	**85.0** ±1.4
	ER	61.9 ±2.8	49.4 ±3.0	63.7 ±3.2	27.3 ±2.7	83.2 ±2.3	68.8 ±2.9	65.0 ±3.1	77.2 ±2.2
	KLE	76.0 ±2.3	59.8 ±2.8	73.7 ±3.1	34.4 ±3.1	87.5 ±2.0	71.8 ±2.6	82.5 ±2.7	83.6 ±1.7
	EOS-UP	74.3 ±2.5	58.1 ±2.9	75.9 ±3.1	35.8 ±3.1	88.0 ±1.9	71.8 ±2.6	83.6 ±2.3	84.4 ±1.6
	EVSE	**77.4** ±2.3	**60.4** ±2.8	**75.9** ±3.1	**35.9** ±3.3	**88.4** ±1.9	**72.3** ±2.7	**85.4** ±2.3	85.0 ±1.6
Llama-2-7b-chat	SE	71.7 ±2.7	40.5 ±3.2	72.1 ±3.4	29.8 ±3.1	83.2 ±2.2	58.5 ±3.1	79.8 ±2.3	**72.7** ±2.2
	DSE	72.1 ±2.7	38.7 ±2.9	71.4 ±3.2	27.7 ±2.9	82.3 ±2.3	56.3 ±3.1	**80.2** ±2.1	72.3 ±2.2
	E	74.4 ±2.6	42.5 ±3.2	69.9 ±3.2	28.6 ±3.0	84.4 ±2.3	**60.4** ±3.2	74.9 ±2.6	70.3 ±2.5
	p(True)	65.2 ±3.1	37.1 ±3.1	64.3 ±3.5	25.7 ±2.8	62.7 ±3.4	47.9 ±3.5	75.3 ±2.5	70.8 ±2.5
	ER	56.4 ±3.1	31.6 ±3.0	51.6 ±3.4	19.1 ±2.4	82.8 ±2.4	59.6 ±3.3	71.1 ±2.7	69.4 ±2.4
	KLE	73.1 ±2.6	40.9 ±3.1	72.5 ±3.3	30.0 ±3.0	83.3 ±2.2	58.6 ±3.2	79.7 ±2.3	72.6 ±2.4
	EOS-UP	70.4 ±2.8	40.3 ±3.1	68.2 ±3.5	27.5 ±2.9	75.2 ±2.8	54.9 ±3.4	70.3 ±2.7	68.5 ±2.7
	EVSE	**74.5** ±2.6	**42.7** ±3.2	**73.8** ±3.2	**30.9** ±3.2	**85.4** ±2.2	60.2 ±3.4	79.0 ±2.4	72.2 ±2.6
Mistral-7b-instr	SE	72.6 ±3.0	33.0 ±3.2	71.9 ±3.5	28.6 ±3.1	88.3 ±1.8	70.6 ±2.6	84.3 ±1.9	74.2 ±2.1
	DSE	74.3 ±3.1	30.6 ±2.8	72.4 ±3.3	25.4 ±2.6	86.7 ±2.0	65.9 ±3.0	83.7 ±2.0	71.9 ±2.2
	E	73.9 ±3.0	33.8 ±3.1	69.5 ±3.3	26.4 ±2.9	83.6 ±2.4	68.8 ±2.8	76.5 ±2.4	70.3 ±2.5
	p(True)	64.5 ±3.3	28.7 ±2.9	66.3 ±3.5	25.7 ±3.0	74.2 ±2.9	61.2 ±3.2	82.6 ±2.1	73.6 ±2.3
	ER	64.5 ±3.2	27.9 ±2.9	61.3 ±3.6	23.1 ±2.7	**89.9** ±1.8	**71.4** ±2.7	71.7 ±2.6	66.8 ±2.7
	KLE	75.4 ±2.9	**34.4** ±3.1	**74.7** ±3.1	**29.2** ±2.9	88.3 ±1.9	70.4 ±2.6	**84.5** ±2.0	74.4 ±2.2
	EOS-UP	70.8 ±3.1	32.2 ±3.1	67.9 ±3.4	25.7 ±2.9	85.6 ±2.1	68.6 ±2.7	75.0 ±2.4	69.6 ±2.5
	EVSE	**75.6** ±2.9	34.3 ±3.1	72.8 ±3.4	28.6 ±3.1	89.4 ±1.8	71.2 ±2.6	84.3 ±1.9	**74.6** ±2.2
Mistral-7b	SE	73.2 ±2.6	53.1 ±3.1	68.9 ±3.3	33.7 ±3.2	86.8 ±2.1	76.7 ±2.3	84.8 ±2.5	84.6 ±1.6
	DSE	73.5 ±2.4	50.1 ±2.9	70.8 ±3.2	31.1 ±2.8	85.7 ±2.1	73.8 ±2.5	84.0 ±2.5	83.4 ±1.6
	E	69.0 ±2.7	48.8 ±3.2	71.6 ±3.0	34.0 ±3.3	83.4 ±2.4	76.0 ±2.4	81.9 ±2.5	84.4 ±1.6
	p(True)	72.8 ±2.7	52.6 ±2.8	72.3 ±3.0	34.0 ±3.1	77.6 ±2.7	72.7 ±2.6	83.8 ±2.6	84.4 ±1.7
	ER	60.1 ±2.9	43.9 ±2.9	62.9 ±3.1	28.4 ±3.0	**89.3** ±1.9	**78.3** ±2.3	67.0 ±3.1	78.2 ±2.1
	KLE	74.8 ±2.6	53.8 ±2.9	69.9 ±3.3	34.1 ±3.1	86.7 ±2.2	76.6 ±2.4	85.2 ±2.5	84.6 ±1.6
	EOS-UP	73.2 ±2.6	52.8 ±3.1	72.2 ±3.2	35.8 ±3.2	80.9 ±2.4	74.4 ±2.5	82.4 ±2.5	84.2 ±1.6
	EVSE	**75.9** ±2.4	**54.5** ±3.0	**74.0** ±3.0	**36.0** ±3.2	88.0 ±2.0	77.2 ±2.4	**85.4** ±2.5	**84.9** ±1.7
falcon-40b	SE	72.0 ±2.5	57.8 ±3.0	75.2 ±2.8	36.8 ±3.2	84.0 ±2.2	74.3 ±2.4	82.2 ±2.8	85.5 ±1.5
	DSE	71.7 ±2.5	54.4 ±2.8	74.5 ±3.0	33.3 ±2.9	82.3 ±2.3	69.4 ±2.7	82.2 ±2.8	85.0 ±1.5
	E	72.4 ±2.5	57.6 ±2.9	69.8 ±3.3	34.4 ±3.3	83.2 ±2.3	74.4 ±2.5	74.5 ±2.9	83.5 ±1.7
	p(True)	63.4 ±2.8	51.7 ±3.0	53.5 ±3.6	25.9 ±2.8	75.6 ±2.8	69.0 ±2.9	76.2 ±3.1	84.1 ±1.7
	ER	64.0 ±2.8	50.8 ±3.1	60.1 ±3.2	27.5 ±2.8	84.0 ±2.3	73.5 ±2.6	73.9 ±3.2	83.2 ±1.7
	KLE	72.5 ±2.6	58.4 ±2.8	76.1 ±2.8	36.8 ±3.2	84.3 ±2.3	74.6 ±2.5	82.5 ±2.9	85.6 ±1.6
	EOS-UP	71.3 ±2.7	57.5 ±3.0	71.4 ±3.2	34.9 ±3.1	83.4 ±2.3	74.9 ±2.4	79.7 ±2.7	85.4 ±1.6
	EVSE	**73.0** ±2.5	**58.4** ±2.8	**77.9** ±2.7	**37.5** ±3.2	**85.8** ±2.1	**75.2** ±2.4	**84.2** ±2.8	**86.2** ±1.6
falcon-7b-instruct	SE	68.4 ±3.6	28.5 ±3.2	74.0 ±4.0	18.4 ±2.7	65.4 ±4.2	24.1 ±3.2	82.9 ±2.1	59.4 ±2.9
	DSE	68.5 ±3.6	25.7 ±2.8	72.8 ±4.0	15.5 ±2.3	60.6 ±3.9	19.8 ±2.5	83.2 ±1.9	54.7 ±3.0
	E	**70.1** ±3.4	28.3 ±3.1	61.5 ±4.7	14.6 ±2.4	64.5 ±4.5	25.3 ±3.4	80.6 ±2.3	57.4 ±2.9
	p(True)	42.6 ±3.5	15.7 ±2.1	59.2 ±4.1	13.1 ±2.2	64.9 ±4.1	24.4 ±3.4	72.1 ±2.5	49.6 ±3.1
	ER	60.1 ±3.2	22.3 ±2.7	57.7 ±4.6	13.3 ±2.3	65.9 ±4.1	24.6 ±3.3	73.4 ±2.5	50.8 ±3.1
	KLE	69.9 ±3.5	**29.0** ±3.1	**75.8** ±3.9	**18.6** ±2.7	65.1 ±4.1	24.5 ±3.3	84.7 ±2.0	60.2 ±2.9
	EOS-UP	69.8 ±3.5	28.0 ±3.0	72.0 ±4.1	17.4 ±2.5	**71.5** ±3.8	**27.1** ±3.5	80.1 ±2.2	56.1 ±3.1
	EVSE	69.0 ±3.7	28.7 ±3.1	75.6 ±3.7	18.4 ±2.7	69.6 ±3.9	26.2 ±3.4	**85.7** ±1.8	**60.4** ±2.8
falcon-7b	SE	69.3 ±2.7	42.1 ±3.1	73.9 ±3.4	25.7 ±2.9	73.0 ±3.3	37.1 ±3.6	83.0 ±2.1	71.4 ±2.3
	DSE	68.3 ±2.9	37.9 ±2.8	70.9 ±3.5	20.7 ±2.5	63.5 ±3.4	28.2 ±3.0	81.7 ±2.2	68.2 ±2.4
	E	65.8 ±3.0	40.6 ±3.1	68.0 ±3.9	23.2 ±2.8	69.5 ±3.6	35.9 ±3.4	76.6 ±2.4	68.5 ±2.6
	p(True)	54.5 ±3.0	32.2 ±2.9	43.9 ±3.8	13.3 ±2.0	61.9 ±3.7	31.0 ±3.4	60.3 ±2.8	56.3 ±3.0
	ER	61.3 ±2.9	36.3 ±3.1	52.9 ±3.9	17.2 ±2.5	72.0 ±3.3	35.6 ±3.6	71.9 ±2.5	64.5 ±2.8
	KLE	70.3 ±2.7	42.7 ±3.2	72.3 ±3.6	25.2 ±3.1	72.5 ±3.4	36.4 ±3.7	83.6 ±2.0	71.8 ±2.3
	EOS-UP	70.8 ±2.9	**43.6** ±3.2	**77.3** ±3.1	**26.9** ±3.0	**76.3** ±3.2	**38.9** ±3.7	81.0 ±2.1	71.0 ±2.4
	EVSE	**71.3** ±2.7	43.6 ±3.2	75.6 ±3.2	26.2 ±3.0	75.4 ±3.3	37.8 ±3.6	**84.2** ±2.0	**72.2** ±2.3

Table 7: Same as Table 3 but for the remaining 8 models.

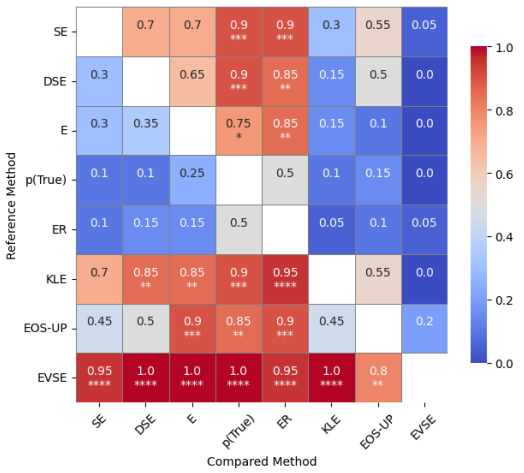


(a) AUROC

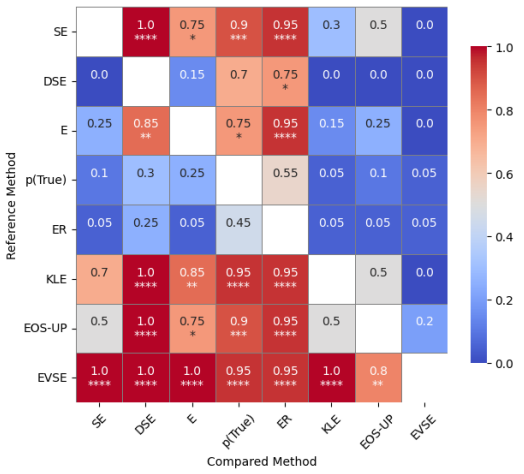


(b) AUARC

Figure 8: Summary of the 20 experiments for **instruction-tuned** models. The remaining description is same as Figure 3.



(a) AUROC



(b) AUARC

Figure 9: Summary of the 20 experiments for **base** models. The remaining description is same as Figure 3.