EVIDENTIAL SEMANTIC UNCERTAINTY FOR AUTOREGRESSIVE LLM GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty quantification (UQ) is essential for improving the reliability of autoregressive generation in large language models (LLMs). In open-ended generation, uncertainty can arise at lexical, syntactic, and semantic levels, among which semantic uncertainty is most indicative of answer correctness. A widely used approach, semantic entropy, estimates uncertainty by sampling multiple answers, clustering them into semantically consistent groups, and computing the entropy over group probabilities. However, this formulation assumes disjoint answer groups and fails to capture partial semantic overlap. Recent methods address this limitation by leveraging spectral clustering and kernel-based representations to model dispersion in semantic space. Despite these advances, an important failure mode remains unaddressed: cases where the model produces consistently incorrect yet semantically coherent answers. In this work, we observe that such incorrect generations are often associated with low evidential support, providing a complementary signal for uncertainty estimation. Building on this insight, we propose an evidential-theoretic framework that models answer groups using multinomial opinion and hyper-opinion. The proposed framework enables a principled decomposition of semantic uncertainty into: (i) lack of supporting evidence and (ii) conflict among competing answer groups. Our formulation explicitly captures inter-group semantic overlap and yields more expressive uncertainty measures aligned with the underlying semantic structure. Experiments across multiple question-answering benchmarks and LLMs demonstrate that our method consistently improves uncertainty estimation and enhances the detection of incorrect answers.

1 INTRODUCTION

With the rapid adoption of large language models (LLMs) across diverse applications, ensuring the reliability and trustworthiness of their outputs has become increasingly important. Because LLMs can produce fluent yet incorrect responses, providing an uncertainty, or confidence estimate alongside each generation is crucial for helping users decide when to trust the model and when to seek additional verification. Uncertainty quantification (UQ) (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Sensoy et al., 2018; Liu et al., 2020; Vaze et al., 2021) has long been studied as a means of improving the reliability of machine learning predictions, particularly in high-stakes settings.

However, UQ for open-ended autoregressive generation differs fundamentally from standard classification. Generation unfolds over multiple tokens, with each decoding step corresponding to a classification problem over the vocabulary. Consequently, sequence-level uncertainty is commonly estimated by aggregating token-level scores, such as entropy or negative log-likelihood (Ren et al., 2022; Malinin & Gales, 2020). Nevertheless, in many applications, the most consequential uncertainty is semantic: the model may be uncertain among meaningfully distinct answers, rather than merely among alternative wordings or syntactic realizations of the same answer. Token-level measures do not explicitly separate semantic uncertainty from lexical and syntactic variability, and may assign high uncertainty to harmless surface-form variation. This limits their ability to faithfully indicate when a generated response is semantically unreliable.

To address this limitation, semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) estimates uncertainty at the level of meaning rather than surface form. Given a question, it samples multiple responses from an LLM and clusters them into groups corresponding to semantically distinct answers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

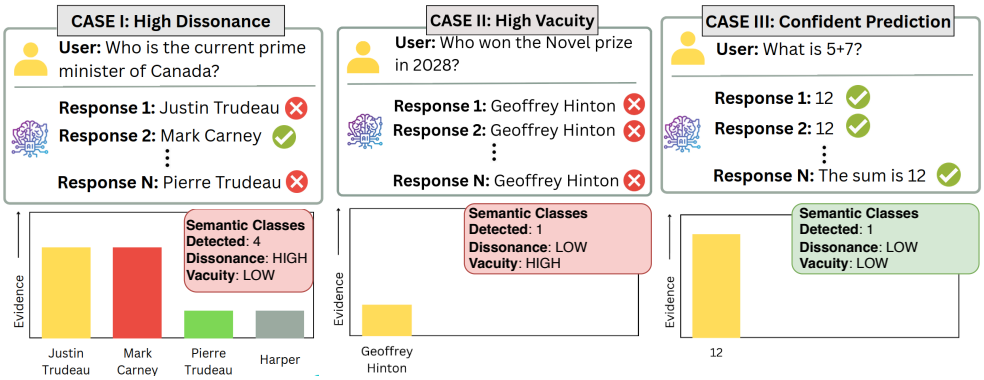


Figure 1: Overview of three cases of output from LLM. Case I: Model is incorrect in its prediction due to confusion between answers. Case II: Model is incorrect in its prediction due to lack of evidence. Case III: Model is correct in its prediction and fine-grained uncertainties are absent.

Two responses are assigned to the same group if they are judged to be semantically equivalent. The entropy of the resulting group distribution is then used as a semantic-level uncertainty score for detecting incorrect responses.

However, recent work (Lin et al., 2023; Nikitin et al., 2024) has shown that semantic-equivalence grouping induces a hard partition of the answer space, implicitly assuming that semantic clusters are disjoint. This assumption can be overly restrictive, as plausible answers may partially overlap in meaning rather than fall into mutually exclusive groups. Capturing such overlap is important: when answer groups are highly similar, the effective semantic disagreement should be lower, and the resulting uncertainty score should decrease accordingly. To address this hard-partitioning limitation, Lin et al. (Lin et al., 2023) construct an answer–answer similarity graph and apply spectral clustering, allowing semantic groups to be informed by graded similarity rather than strict equivalence alone. Nikitin et al. (Nikitin et al., 2024) define a semantic kernel over generated answers and use von Neumann entropy to quantify the dispersion of generations in the induced semantic space.

Yet an important failure mode remains: an LLM may produce answers that are semantically consistent but incorrect. As illustrated in Figure 1(Case II), the sampled responses may collapse to a single wrong semantic mode. Existing approaches, including semantic entropy, spectral-clustering-based uncertainty, and kernel entropy, can therefore assign low uncertainty simply because the generations exhibit little semantic dispersion. This is problematic: *semantic concentration is not equivalent to reliability when the model consistently converges to an incorrect answer*.

To address uncertainty quantification for open-ended language generation, we propose a novel framework grounded in evidence theory (Jøsang, 2016; Josang et al., 2018). Starting from semantic answer groups (e.g., obtained through semantic-equivalence clustering), we map the group structure into evidential opinions, including both multinomial opinions and hyper-opinion representations. Figure 2 provides an overview of the proposed formulation. Beyond sampled answers, we further extract token- and sequence-level evidence from the model and aggregate it within each semantic group, producing both group-specific evidence and total evidence for the overall opinion. This evidential representation enables a fine-grained decomposition of semantic uncertainty into two complementary components: (i) conflict uncertainty, which captures disagreement or competition among semantic answer groups, and (ii) vacuity, or lack-of-evidence uncertainty, which captures insufficient total support for the model's prediction. As illustrated in Figure 1, conflict uncertainty reflects semantic ambiguity among plausible answers, while vacuity is critical for identifying unreliable generations even when the samples are narrowly concentrated, such as when the model produces semantically consistent but incorrect answers.

The two evidential opinions (multinomial opinion and hyper-opinion) differ in how they model semantic answer groups. Multinomial opinions assume that groups are mutually exclusive, whereas hyper-opinions relax this assumption by explicitly allowing non-exclusive, overlapping groups. By incorporating semantic overlap into the uncertainty computation, hyper-opinions can more faithfully capture dispersion in semantic space than multinomial opinions, leading to improved incorrect-answer detection when semantic clusters are not well separated. Our contributions are threefold:

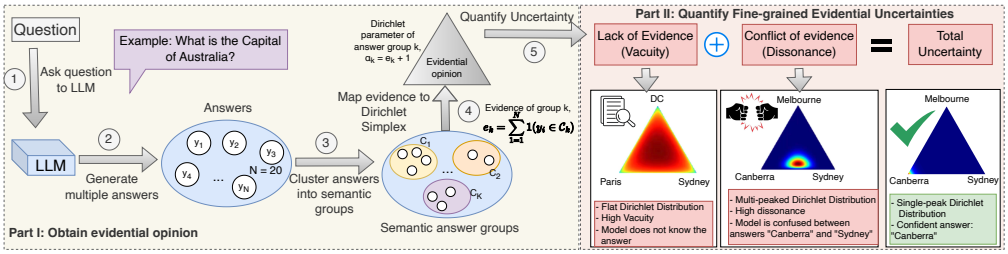


Figure 2: Overview of the proposed method. Part I: Obtain evidential opinions from semantic clusters, Part II: Quantify fine-grained evidential uncertainties.

- We propose an evidential framework for uncertainty quantification in open-ended LLM generation. The framework decomposes uncertainty into *conflict* across semantic answer groups and *vacuity*, i.e., lack of total evidence, enabling the detection of unreliable outputs even when generated answers are semantically consistent.
- We provide a theoretical analysis showing how fine-grained evidential uncertainty components can be combined to detect incorrect generations.
- We introduce a hyper-opinion formulation to model partial semantic overlap among answer groups, yielding more faithful uncertainty estimates when semantic clusters are non-exclusive and addressing a key limitation of semantic entropy.

We evaluate the proposed formulation on four open-ended generation datasets and three LLMs, demonstrating consistent improvements over prior uncertainty quantification baselines.

2 Related Work

2.1 Uncertainty for non-autoregressive outputs

Uncertainty quantification (UQ) in machine learning has a long history for non-autoregressive prediction settings. Bayesian deep learning provides a principled framework for characterizing aleatoric and epistemic uncertainty (Kendall & Gal, 2017). Monte Carlo dropout offers a computationally efficient approximation for uncertainty estimation (Brach et al., 2020), while deep ensembles are another widely used approach to quantify predictive uncertainty (Lakshminarayanan et al., 2017; Wilson & Izmailov, 2020). Simpler confidence proxies, such as the maximum softmax probability or maximum logit score, or logit margin between top-2 predictions, have also been used—particularly in open-set and out-of-distribution detection settings (Vaze et al., 2021).

More recently, evidential learning methods have shown strong performance for both in-distribution and out-of-distribution uncertainty estimation by explicitly modeling fine-grained uncertainty components (Sensoy et al., 2018; Shi et al., 2020). Hyper-opinion representations have further been explored to capture non-exclusive class structure and improve uncertainty estimates in image classification (Qu et al., 2024). In this work, we extend hyper-opinion–based modeling to uncertainty quantification for autoregressive language generation.

2.2 Uncertainty for autoregressive outputs

UQ for autoregressive generation is challenging because uncertainty must be assessed over sequences rather than single predictions. A common approach is to aggregate token-level uncertainty using perplexity (Ren et al., 2022; Bakman et al., 2024), which summarizes the negative log-likelihood across generated tokens. Other methods aim to expose uncertainty from internal model signals: for instance, training an auxiliary classifier on the final/intermediate layer representations to predict whether an output is incorrect (Feldman et al., 2023; Azaria & Mitchell, 2023; Liu et al., 2024). Furthermore, the LLM can be prompted to assess the truthfulness of its own generations, using the predicted probability of "True" as a confidence score (Kadavath et al., 2022; Yang et al., 2024).

A complementary line of work relies on sample-based uncertainty: multiple answers are generated for the same prompt, and their inconsistency is used as an uncertainty signal (Kuhn et al., 2023; Farquhar et al., 2024; Fomicheva et al., 2020; Lin et al., 2023; Nikitin et al., 2024; Qiu & Miikkulainen, 2024; Vashurin et al.; Duan et al., 2024). Early approaches used lexical diversity among

samples (Fomicheva et al., 2020), while semantic entropy (Kuhn et al., 2023) demonstrated that semantic variability, captured by clustering answers into distinct meaning groups, yields a stronger indicator of incorrectness. An extension to semantic entropy assigns importance to tokens according to their relevance while computing confidence for answer clusters (Duan et al., 2024). However, semantic-entropy-style grouping typically assumes disjoint semantic clusters and does not account for overlap between groups. To relax this assumption, graph-based formulations (Lin et al., 2023) and kernel-based similarity measures over the answer set (Nikitin et al., 2024) have been proposed to better capture the geometry of the semantic space. Leveraging the semantic similarities with other responses, the sample-level confidence score is also quantified (Qiu & Miikkulainen, 2024).

In this work, we explicitly model semantic overlap among answer groups using an evidential-theoretic formulation, and we further introduce an additional uncertainty source that captures lack of evidence from the model, complementing dispersion/conflict-based signals.

3 METHODOLOGY

Given an input query x_{query} to an LLM, the model generates a token sequence $\boldsymbol{y} = [y_1, \ldots, y_t, \ldots, y_T]$, where T denotes the output length and y_t is the token produced at generation step t. The goal is to assign an uncertainty score that is *high* when the generated response $\boldsymbol{y}$ is incorrect and *low* when it is correct. The proposed methodology, as illustrated in Figure 2, consists of two parts: (i) obtain evidential opinion for x_{query}, and (ii) quantify fine-grained uncertainties given evidential opinion.

3.1 PART I: FROM SEMANTIC CLUSTERS TO EVIDENTIAL OPINIONS

Given an input query x_{query}, we sample a set of N responses $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ from the model through decoding. Next, we group these answers into K semantic clusters $\mathcal{Y}^{sem} = \{C_1, \ldots, C_K\}$, where K is determined automatically by induced clustering. Two responses $\boldsymbol{y}_n$ and $\boldsymbol{y}_m$ are assigned to same cluster if they are semantically equivalent (determined via an NLI model (Kuhn et al., 2023)).

To assign evidential support to these semantic clusters, we derive evidence directly from the model's token logits. Although cross-entropy training aligns the model's logits with token-level confidence (Vaze et al., 2021; Kadavath et al., 2022), raw token logits in autoregressive generation fail to distinguish true semantic uncertainty from lexical variability across different wordings. Conversely, relying solely on sample frequencies ignores the model's internal confidence, making it difficult to detect when an LLM consistently generates wrong answers with low confidence. To bridge this gap, we aggregate token logits across generated sequences within each semantic cluster C_k. For a response $\boldsymbol{y}^{ik}$ (i^{th} answer in cluster k), we apply a non-negative transformation $\mathcal{E}$ (e.g., ReLU) to its token logits $f_\theta(y_t^{ik} \mid x_{\text{query}}, \boldsymbol{y}_{<t}^{ik})$ and average over sequence length T^{ik}. The total evidence e_k for semantic cluster C_k is the sum across all its constituent responses:

$$e(\boldsymbol{y}^{ik}) = \frac{1}{T^{ik}} \sum_{t=1}^{T^{ik}} \mathcal{E}\left(f_\theta(y_t^{ik} \mid x_{\text{query}}, \boldsymbol{y}_{<t}^{ik})\right), \quad e_k = \sum_{\boldsymbol{y}^{ik} \in C_k} e(\boldsymbol{y}^{ik}) \tag{1}$$

Following the subjective logic (Jøsang, 2016), a Dirichlet PDF is defined over K classes, parameterized by evidence $\alpha_k = e_k + a_k W$, where $\boldsymbol{a}$ is the base rate vector (usually set to uniform vector $a_k = 1/K$); W (usually set to $W = K$) is the non-informative prior weight. The Dirichlet strength is evaluated as $S = \sum_{k=1}^{K} e_k + W$. We define an evidential multinomial opinion $w = (\boldsymbol{b}, u, \boldsymbol{a})$ over K classes, where: b_k is the belief mass assigned to cluster C_k and u is the uncertainty mass or lack of evidence, referred to as "vacuity". The mapping of the Dirichlet PDF and the evidential opinion w is given by the following relation:

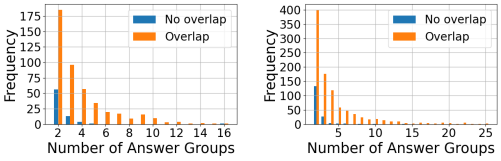
$$b_k = \frac{e_k}{S}, \qquad u = \frac{W}{S}, \qquad \sum_{k=1}^{K} b_k + u = 1 \tag{2}$$

Hyper-opinion Representation: While standard multinomial opinion assumes classes are mutually exclusive, semantic clusters may contain partial overlap with each other. We empirically demonstrate the prevalence of overlap between answer groups in Figure 3 on the TriviaQA test set using two LLMs. We first compute the number of semantic groups K for each test sample. Then, for each

sample, we check whether any pair of groups exhibits partial overlap, defined as an entailment relation between the groups. Specifically, letting $\mathcal{Q}$ denote the set of unique group pairs, we declare a sample to have overlap if

$$\exists (i,j) \in \mathcal{Q} : \ \mathcal{C}_i \to \mathcal{C}_j \ \vee \ \mathcal{C}_j \to \mathcal{C}_i,$$

where $\mathcal{C}_i \to \mathcal{C}_j$ indicates that class i entails group j. We treat group-level entailment as holding when at least one response in $\mathcal{C}_i$ entails a response in $\mathcal{C}_j$. For each distinct value of K, we plot the fraction of samples with *no-overlap* versus *overlap*. Here, *overlap* means that at least two answer groups in the sample are non-exclusive, while *no-overlap* means that none of the groups are non-exclusive. As shown in Figure 3, overlap becomes increasingly common as K grows: questions that admit many candidate answer groups are very likely to produce groups that entail one another, and the proportion of no-overlap cases becomes small. This empirical trend motivates the need to model inter-group overlapping in uncertainty quantification.



(a) Gemma Model (b) Llama Model

Figure 3: Visualization of overlap of answer groups for TriviaQA dataset (a) Gemma-3-4b-it and (b) Llama-2-7b-chat-hf

To capture this non-exclusive structure, we extend our framework to an evidential hyper-opinion, $w^H = (\boldsymbol{b}^H, u^H, \boldsymbol{a}^H, \tilde{\boldsymbol{a}}^H)$ (Josang et al., 2018). Similar to multinomial opinion, $\boldsymbol{b}^H$, u^H, and $\boldsymbol{a}^H$ denote belief masses, uncertainty mass, and standard base rate for each class. $\tilde{\boldsymbol{a}}^H$ is an inter-group base rate defined for two classes and represents the fraction of the semantic space between two classes that does not overlap. Further, we define the non-informative prior weight W^H as the effective number of semantic sets in $\mathcal{Y}^{sem}$, accounting for partial overlap between them. By leveraging the inter-group base rate $\tilde{\boldsymbol{a}}^H$ and W^H, the hyper-opinion effectively adjusts the uncertainty score when an LLM's response falls into the gray area between two semantically similar clusters. We discuss the exact formulation of $\tilde{\boldsymbol{a}}^H$ and W^H in Section 3.2 . The relation of belief and uncertainty masses differs from multinomial opinion only in the definition of W^H:

$$S^H = \sum_{k=1}^{K} e_k + W^H, \quad b_k^H = \frac{e_k}{S^H}, \quad u^H = \frac{W^H}{S^H}, \quad \sum_{k=1}^{K} b_k^H + u^H = 1 \tag{3}$$

3.2 Part II: Quantify Fine-grained Evidential Uncertainties

First, we introduce the preliminaries of quantifying fine-grained evidential uncertainties. Next, we present the methodology to quantify semantic uncertainties for multinomial and hyper opinions.

Fine-grained Evidential Uncertainties: Following the principles of subjective logic and theory of evidence (Jøsang, 2016), a multinomial opinion over K classes is derived from evidence $\mathbf{e} = [e_1, \ldots, e_K]$. This allows us to express uncertainties when the model "doesn't know" (high vacuity) and when the model is "conflicted" (high dissonance) between classes, respectively. The definition of vacuity(u) is given in equation 2. We define dissonance(`diss`) for multinomial opinion follows:

$$\texttt{diss} = \sum_{i=1}^{K} \left(b_i \frac{\sum_{j \neq i} b_j Bal(b_i, b_j)}{\sum_{j \neq i} b_j} \right), \quad Bal(b_i, b_j) = \begin{cases} 2\frac{\min(b_i, b_j)}{b_i + b_j}, & \text{if } b_i b_j > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Multinomial Semantic Uncertainties: For multinomial opinion, we quantify fine-grained uncertainties following Equations 2 and 4. Vacuity is a function of the total evidence $\sum_k^K e_k$ and increases as the sum of the evidence decreases, thereby revealing a lack of evidence in the model. In Figure 4, we plot the distribution of total evidence for correct and incorrect samples for the setting of TriviaQA, Gemma model, empirically demonstrating its confidence characteristics. In addition to the evidence sum, vacuity is also a function of the non-informative prior weight W, which is typically a fixed, predefined constant for standard subjective logic (Jøsang, 2016; Sensoy et al., 2018). In the proposed multinomial formulation, we define W as a dynamic parameter, $W = K = |\mathcal{Y}^{sem}|$, where K denotes the cardinality of the semantic classes discovered during the

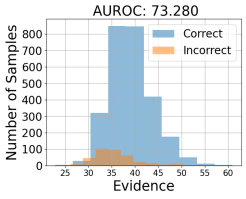


Figure 4: Evidence distribution

clustering phase. An increase in the value of K (distinct semantic classes) requires a higher sum of evidence to lower the vacuity. In this way, u leverages the power of both signals (evidence and number of semantic classes) to estimate the vacuity from the model. For dissonance (uncertainty of conflicts between beliefs), we define $Bal(b_i, b_j)$, a conflict between a pair of answer groups (with indices i, j). Next, we aggregate the conflict across all the pairs, defined by dissonance ($\texttt{diss}$). Belief mass is used as a weight for the conflict $Bal(b_i, b_j)$ when aggregating in Equation 4, thereby assigning greater importance to conflicts between higher belief masses.

The representative examples of high vacuity, high dissonance, and confident prediction are presented in Figure 2 with corresponding Dirichlet PDFs. To unify these scores, we characterise the interplay between vacuity and dissonance by deriving a bound for b_{cor}, the belief mass of the correct semantic label in Proposition 1.

Proposition 1. *Consider an LLM that outputs a semantic set of labels $\mathcal{Y}^{sem}$ for an input x_{query}. If b_{cor} is the belief mass of the correct label, $\texttt{diss}$ is the dissonance, and u is vacuity, then the correct belief is bounded by the uncertainties u and $\texttt{diss}$ according to:*

$$b_{cor} \leq 1 - (u + \frac{\texttt{diss}}{2}) \quad \textit{where} \quad 0 \leq b_{cor} \leq 1, \quad 0 \leq \texttt{diss} \leq 1 \quad \& \quad 0 \leq u \leq 1 \tag{5}$$

The inequality provides the theoretical upper bound of the belief mass of the correct semantic label. b_{cor} is penalized by the presence of fine-grained evidential uncertainties: u and $\texttt{diss}$.

Please refer to Appendix C for the proof. Following the proposition, we aggregate fine-grained uncertainties to a unified score: $U^{\texttt{multi}} = u + \frac{\texttt{diss}}{2}$. The formulation being directly related to b_{cor}, justifies the goal of uncertainty in LLM to identify incorrectness. Contrary to a standard classification problem, where vacuity is leveraged to detect out-of-distribution samples and dissonance is leveraged to detect misclassifications for in-distribution samples, the unified score helps in uncovering both sources of uncertainty that lead to incorrectness for LLM generation.

Hyper-opinion Semantic Uncertainties: We present the fine-grained uncertainties: hyper-vacuity and hyper-dissonance for handling non-exclusive groups. For hyper-vacuity u^H, we adjust the prior weight W^H to represent the effective number of semantic sets ($K_{\text{effective}}$) via diversity of the answer space, explicitly accounting for overlap between answer clusters. For the same amount of total evidence, dynamically reducing W^H for overlapping answer groups helps reduce the vacuity, thereby modeling the effect of semantic overlap.

To measure the true semantic diversity of the answer space, we utilize the Normalized Graph Laplacian, $L = I - D^{-1/2}AD^{-1/2}$, where A is the semantic similarity matrix (A_{ij} denotes the similarity between clusters i and j), and D is the diagonal degree matrix. Following spectral graph theory (Von Luxburg, 2007; Lin et al., 2023), the eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K \leq 2$ of L reveal the underlying structure of the semantic graph. The algebraic multiplicity of the eigenvalue $\lambda = 0$ corresponds exactly to the count of fully disjoint connected components (Von Luxburg, 2007). However, semantic similarity matrices typically yield a single connected component with non-zero continuous eigenvalues. To capture true semantic diversity under this continuous setting, we explore two complementary spectral strategies to compute $W^H = K_{\text{effective}}$: (i) **Thresholded Spectral Aggregation**: We treat the concentration of small eigenvalues near zero as a soft indicator of cluster count. We estimate the effective cluster count by summing small eigenvalues ($\lambda_k < \tau$), effectively giving maximum weight to independent clusters ($\lambda_k \approx 0$), resulting in hyper-vacuity:

$$u^H(\text{Sum}) = \frac{W^H(\text{Sum})}{S^H}, \quad W^H(\text{Sum}) = \sum_{k=1}^{K} \max(0, \tau - \lambda_k) \tag{6}$$

where $\tau > 0$ acts as a spectral threshold hyperparameter. This formulation relaxes hard zero-multiplicity counting: eigenvalues near zero contribute near-maximal weight toward $K_{\text{effective}}$, whereas larger eigenvalues ($\lambda_k \geq \tau$) signal weak partition boundaries or noisy connections and are filtered out. Adjusting τ allows us to control how strictly an eigenvalue must approach zero to contribute to the effective set size. (ii) **Participation Ratio**: Inspired by effective dimensionality estimation in high-dimensional representations (Chun et al., 2026), we adapt the Participation Ratio (PR) to measure the effective semantic rank of the answer space continuously. Instead of counting discrete clusters with a fixed cutoff threshold τ, PR measures the spectral variance of the Laplacian, thereby reflecting the true semantic diversity of the answer space. Intuitively, when the semantic answers

Table 1: Performance comparison (AUROC and AUARC) across datasets and models.

Method	TriviaQA AUROC	TriviaQA AUARC	SQuAD AUROC	SQuAD AUARC	NQ AUROC	NQ AUARC	BioASQ AUROC	BioASQ AUARC
Model: gemma-3-4b-it								
NLL	62.02	86.39	63.46	81.59	64.09	35.34	62.30	56.28
SE	69.44	87.40	72.20	83.65	70.60	39.26	68.44	60.43
DSE	69.20	87.36	72.54	83.70	72.02	40.09	66.43	59.16
PE	66.01	86.98	61.44	81.28	62.50	34.70	57.94	55.02
P(True)	59.62	85.30	50.05	76.76	59.56	31.93	59.54	54.68
Eigen Value	72.57	88.07	74.20	84.07	76.07	41.73	73.79	62.78
Degree	71.95	87.94	74.09	84.02	74.45	41.00	71.62	61.67
KLE	71.87	87.92	73.71	84.14	72.43	40.26	74.81	63.09
Max Prob	69.15	87.34	72.44	83.69	71.45	39.43	66.42	59.05
CoCoA	70.91	87.92	73.05	83.81	71.72	39.92	67.83	59.63
Semantic Density	69.17	81.26	71.77	83.23	73.59	40.68	69.94	60.85
EOS-UP	68.66	82.59	74.75	84.54	69.44	38.65	68.89	60.03
EVSE	68.15	82.72	74.92	84.68	69.12	38.56	70.63	60.65
U^{multi}	76.64	89.42	77.85	86.31	70.95	39.66	66.77	59.58
U^{hyper} (PR)	77.52	89.52	78.54	86.42	76.42	42.05	75.78	63.90
U^{hyper} (Sum)	**77.68**	**89.54**	**78.71**	**86.45**	**77.00**	**42.28**	**76.48**	**64.05**
Model: Llama-2-7b-chat-hf								
NLL	64.89	88.78	61.55	86.18	64.15	48.43	62.56	61.45
SE	73.65	90.03	68.65	87.62	66.07	49.94	60.31	60.64
DSE	75.36	90.21	65.87	87.23	66.34	50.31	66.12	63.71
PE	69.09	88.85	63.18	86.51	57.00	45.19	52.52	57.20
P(True)	66.57	89.15	53.10	84.19	57.30	44.44	64.20	63.46
Eigen Value	77.57	90.43	70.90	87.87	71.29	52.62	73.17	67.18
Degree	76.73	90.35	70.17	87.79	69.68	51.98	71.38	66.27
KLE	77.66	90.44	69.29	87.75	68.22	51.39	69.77	65.53
Max Prob	75.24	90.11	68.21	87.52	65.86	49.88	66.11	63.54
CoCoA	76.28	90.50	69.21	87.64	69.63	51.70	67.91	64.58
Semantic Density	78.34	90.57	71.15	87.95	70.26	52.07	70.44	65.78
EOS-UP	76.94	89.95	66.46	87.18	66.04	49.83	60.92	59.60
EVSE	76.12	89.94	67.36	87.40	66.67	50.24	59.44	59.19
U^{multi}	76.79	90.68	69.39	87.99	66.42	50.50	67.17	64.35
U^{hyper} (PR)	**79.36**	**90.91**	**72.22**	**88.40**	71.99	53.29	74.91	67.99
U^{hyper} (Sum)	78.69	90.86	72.20	88.37	**72.59**	**53.45**	**75.44**	**68.13**
Model: Qwen3-VL-8B-Instruct								
NLL	65.62	90.42	62.57	87.54	63.00	44.21	66.07	66.12
SE	74.46	91.40	62.81	87.62	63.16	44.53	64.22	65.15
DSE	75.80	91.49	64.52	87.94	71.36	48.86	72.22	69.32
PE	59.74	89.70	61.97	87.43	55.81	39.50	59.71	61.43
P(True)	68.06	90.35	61.31	87.48	61.74	41.21	70.05	68.31
Eigen Value	80.19	91.92	62.40	87.47	73.48	50.18	77.03	72.73
Degree	79.69	91.86	62.60	87.48	72.37	49.66	77.04	72.69
KLE	79.35	91.84	63.57	87.66	73.58	50.75	76.87	72.69
Max Prob	75.51	91.42	65.18	88.00	70.06	47.42	71.71	68.96
CoCoA	76.69	91.67	65.11	88.02	67.51	46.93	68.92	68.05
Semantic Density	77.83	91.68	66.71	88.50	72.65	50.29	75.75	71.72
EOS-UP	73.93	91.04	66.02	87.89	68.75	47.13	72.99	69.78
EVSE	73.70	90.97	65.94	88.02	70.03	47.82	74.33	70.08
U^{multi}	76.44	91.58	65.13	87.97	70.16	48.34	71.82	69.53
U^{hyper} (PR)	80.14	91.87	67.08	88.42	73.80	**51.11**	77.86	73.04
U^{hyper} (Sum)	**81.63**	**92.14**	**68.25**	**88.70**	**73.67**	50.84	**78.69**	**73.62**

are fully diverse and distinct, $W^H(\text{PR}) \to K$, reflecting K unique answer groups. Conversely, as answers become highly overlapping and redundant, $W^H(\text{PR}) \to 1$, effectively reducing the counted weight to represent a single consolidated semantic concept.

$$u^H(\text{PR}) = \frac{W^H(\text{PR})}{S^H}, \quad W^H(\text{PR}) = \frac{\left(\sum_{k=1}^{K} \lambda_k\right)^2}{\sum_{k=1}^{K} \lambda_k^2} \tag{7}$$

Next, for measuring the actual conflict between answer groups, we leverage inter-group base rate $\tilde{\boldsymbol{a}}^H$, which is defined for group i and j such that $\tilde{\boldsymbol{a}}^H(i,j)Bal(b_i^H, b_j^H)$ measures conflict between beliefs of two regions of the classes that are semantically non-overlapping. Hyper-dissonance aggregates all the effective conflicts between answer groups:

$$\texttt{diss}^H = \sum_{i=1}^{K} \left(b_i^H \frac{\sum_{j \neq i} b_j^H \times \tilde{\boldsymbol{a}}^H(i,j) \times Bal(b_i^H, b_j^H)}{\sum_{j \neq i} \tilde{\boldsymbol{a}}^H(i,j) b_j^H} \right) \tag{8}$$

Remark: Multinomial opinion is a special form of hyper-opinion when $\tilde{\boldsymbol{a}}^H(i,j) = 1$ and $W^H = |\mathcal{Y}^{sem}|$, *i.e.,* groups i and j are disjoint from each other. The role of inter-group base rates and W^H for three cases of overlaps is illustrated in Figure 8. We leverage NLI model classification output for obtaining the inter-group base rates and similarity matrix such that $\tilde{\boldsymbol{a}}^H(i,j) = \tilde{\boldsymbol{a}}^H(j,i) = \frac{p(\text{contradiction}|C_i,C_j)+p(\text{contradiction}|C_j,C_i)}{2}$ and $A_{ij} = A_{ji} = \frac{p(\text{entailment}|C_i,C_j)+p(\text{entailment}|C_j,C_i)}{2}$. Finally, the total uncertainty for hyper-opinion is defined as $U^{\texttt{hyper}}(\text{Sum}) = u^H(\text{Sum}) + \frac{\texttt{diss}^H}{2}$ and $U^{\texttt{hyper}}(\text{PR}) = u^H(\text{PR}) + \frac{\texttt{diss}^H}{2}$. By defining the total uncertainty as $u^H + \frac{diss^H}{2}$, we are measuring the theoretical upper bound penalty on the correct belief mass as established in Equation 5.

4 EXPERIMENTS

In this section, we present the experiments conducted to verify the effectiveness of the proposed method. First, we discuss the datasets and models used in the experiments. Next, we discuss the compared baselines. We present a comparison of the proposed methods with baselines and study the impact of the proposed components.

Datasets, Models and Evaluation Metrics: We use three large language models: gemma-3-4b-it (Google, 2024), Llama-2-7b-chat-hf (Touvron et al., 2023) and Qwen3-VL-8B-Instruct (Bai et al., 2025). We conduct experiments with four open-ended generation datasets: (i) TriviaQA (Joshi et al., 2017), (ii) SquadQA (Rajpurkar et al., 2018) (reading comprehension and Question Answering datasets), (iii) Natural Questions (NQ) (Kwiatkowski et al., 2019), and (iv) Biomedical question answering (BioASQ) (Krithara et al., 2023). We use AUROC, and AUARC as evaluation metrics following the experiments of KLE (Nikitin et al., 2024). Detailed discussions on evaluation metrics are presented in Appendix D.2.

Comparison Baselines: We include following baselines in our experiments (i) Negative log likelihood (NLL), (ii) Semantic Entropy (SE), (iii) Discrete Semantic Entropy (DSE) (Kuhn et al., 2023), (iv) Predictive Entropy (PE) (Malinin & Gales, 2020) (iv) p(True) (Kadavath et al., 2022), (Feldman et al., 2023), (v) EigenValue(EigV) (vi) Degree (Lin et al., 2023),(vii) Kernel Language Entropy (Nikitin et al., 2024), (viii) Max Prob, (ix) CoCoA (Vashurin et al.), (x) Semantic Density (Qiu & Miikkulainen, 2024), (xi) Evidential Semantic Entropy (EVSE) (Kunitomo-Jacquin et al., 2026) and (xii) EOS-Inclusive UP (EOS-UP) (Kunitomo-Jacquin et al., 2025). Detailed discussions on baselines are presented in Appendix D.2.

Performance Comparison: The comparison of the proposed method, along with baselines, is presented in Table 1 for four datasets and three models. Our method includes total multinomial uncertainty: U^{multi} and two variants of hyper-opinion-based uncertainty: U^{hyper}(sum) and U^{hyper}(pr). For hyper-vacuity in U^{hyper}(sum), we set $\tau = 1.0$. The comparison reveals that either of the variants of our primary approach U^{hyper} (participation ratio or sum) achieves superior results across all settings, consistently yielding the highest AUROC and AUARC scores. This study confirms the robustness of the proposed formulation as a reliable uncertainty estimator. The strength of the proposed hyper-opinion-based formulation lies in (i) quantifying both sources of uncertainty: confusion in answers and lack of evidence for generated answers through hyper dissonance and hyper vacuity, and (ii) capturing partial semantic overlap between answers. Total multinomial uncertainty U^{multi}, quantifying both sources of uncertainty, also demonstrates competitive performance across multiple settings. However, since it does not capture partial semantic overlap between generated answers, the performance is suboptimal in comparison to the hyper-opinion-based formulation. This can be further demonstrated in the impact of each component of the multinomial and hyper-opinion-based formulation in the next section.

Hybrid approaches that include both internal model confidence alongside response consistency, such as CoCoA and EVSE, remain limited by their reliance on point-estimate probabilities compared to the proposed evidential framework. Further, baselines that consider only internal signals (NLL, PE, EOS-UP) or only semantic inconsistency among answers (SE, DSE, KLE, Eigenvalue, Degree, Max Prob and Semantic Density) also demonstrate suboptimal performance. Among them, the most competitive baselines are the ones that consider the partial semantic relationship between answers (KLE, eigenvalue, degree). By explicitly modeling partial overlaps among answer groups alongside fine-grained uncertainties, U^{hyper} demonstrates superior performance across all evaluated settings. This validates the importance of treating semantic sets as non-exclusive.

Impact of uncertainty components: In this section, we study the impact of each component of the proposed multinomial and hyper-opinion (sum variant) for two datasets and the gemma-3-4b-it model. The results are presented in Table 2. First, hyper-vacuity outperforms standard vacuity and hyper-dissonance outperforms standard dissonance across both datasets. This confirms the benefit of accounting for partial semantic overlap among generated answers. Next, combining hyper-dissonance and hyper-vacuity into total uncertainty (U^{hyper}) yields superior performance compared to using either component in isolation. Among components, hyper-vacuity proves more effective uncertainty estimator than hyper-dissonance. Further, comparing across datasets reveals that integrating hyper-dissonance into total uncertainty provides a larger performance boost on NQ than on TriviaQA.

Table 2: Impact of individual uncertainties

Setting	TriviaQA		NQ	
	AUROC	AUARC	AUROC	AUARC
Hyper-opinion				
u^H	77.51	89.53	76.34	41.92
$\texttt{diss}^H$	69.78	87.43	74.92	41.34
U^{hyper}	**77.68**	**89.54**	**77.00**	**42.28**
Multinomial opinion				
u	76.67	89.41	72.35	40.22
$\texttt{diss}$	68.30	87.26	68.54	38.82
U^{multi}	76.64	89.42	70.95	39.66

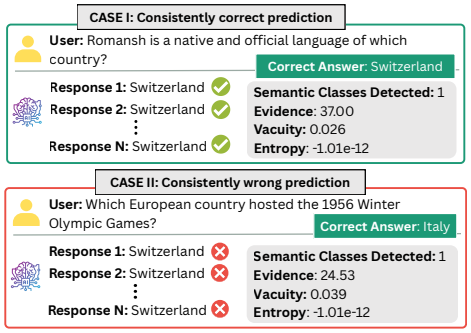
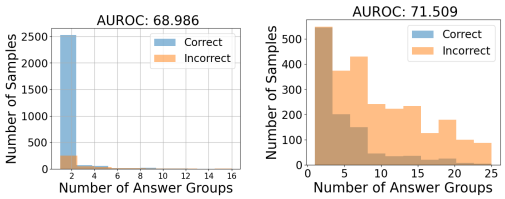


Figure 5: Qualitative analysis



(a) TriviaQA dataset (b) NQ dataset

Figure 6: Distribution of K

To understand why hyper-dissonance provides a significantly larger uplift on NQ than on TriviaQA, we analyze the distribution of the number of answer groups (K) across both datasets. For TriviaQA (Figure 6a), the majority of samples, both correct and incorrect, concentrate at very low values of K. In this low-K setting, separating correct from incorrect predictions relies primarily on the evidence rather than inter-answer-group conflict. Consequently, hyper-vacuity alone already captures most of the uncertainty signal on TriviaQA, leaving little room for hyper-dissonance to add further value. In contrast, Natural Questions (NQ) exhibits much higher values of K for both correct and incorrect samples (Figure 6b). Standard multinomial dissonance struggles here because treating these many answer groups as mutually exclusive falsely penalizes correct samples that contain semantically overlapping clusters, assigning them high uncertainty. Hyper-dissonance resolves this by accounting for partial semantic overlap when measuring conflict, effectively lowering the uncertainty for correct, overlapping answers. Because NQ suffers far more from large values of semantic groups (K), resolving this overlap via hyper-dissonance yields a much stronger boost to total uncertainty (U^{hyper}) than it does on TriviaQA.

Qualitative Analysis: We conduct a qualitative study of the effectiveness of the proposed method in uncertainty quantification by demonstrating 2 different scenarios. We present the study in Figure 5. Both of the questions are from TriviaQA and gemma-3-4b-it model. In the first case, the model consistently outputs the correct answer, resulting in a low value of semantic entropy. Due to a high amount of evidence while generating the answer, vacuity is also low. When the model consistently outputs the wrong answer in the case II, the semantic entropy remains the same as in case I. Due to a low value of evidence sum, vacuity increases in comparison to case I. Hence, even in the extreme case of a consistent wrong answer ($K = 1$), vacuity is able to differentiate between two cases, resulting in accurate detection of incorrect generation. In Appendix D.16, we present a qualitative analysis of how the hyper-opinion handles partial semantic overlap between answer groups.

5 Conclusion

In this work, we propose an evidential framework for uncertainty quantification in open-ended language generation. Our approach combines fine-grained uncertainty components that capture (i) *conflict* across semantic answer groups and (ii) *vacuity* due to insufficient evidence from the model. Moreover, by explicitly modeling overlap among answer groups, we obtain uncertainty estimates that better reflect the dispersion of the semantic answer space. We validate the proposed method through experiments on multiple open-ended generation datasets using three LLMs, and conduct ablations to quantify the contribution of each component. Further, the impact study of total evidence, number of answer groups K, total answers generated N, alternative belief quantification strategy, impact of $\tilde{\boldsymbol{a}}^H$, NLI model, accuracy estimation, and sampling strategies is presented in the Appendix. This extensive study further validates the superior performance of the proposed formulation.

AI USAGE

In this work, the use of generative AI tools for synthetic data generation, cleaning and reformatting dataset, qualitative and thematic data analysis are not applicable. LLM tools have been utilized to polish writing, not for generating ideas, or methods or proposition. Tools were used to help debug Python errors while implementing methodolology, and latex errors while creating the plots, and tables for the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our experimental results, we provide details across the paper and Appendix. Experimental configurations, hyperparameter settings, decoding parameters, and hardware specifications across evaluated language models are described in Section 4 and Appendix. Descriptions of the benchmark datasets, prompt templates, clustering protocols, and evaluation metrics used in our experiments are provided in Section 4 and Appendix. Furthermore, source code containing our evidence extraction pipeline, opinion mapping algorithms, and evaluation suite is linked anonymously in Appendix.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7752–7767, 2024.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.

Kai Brach, Beate Sick, and Oliver Dürr. Single shot mc dropout approximation. *arXiv preprint arXiv:2007.03293*, 2020.

Chanwoo Chun, Abdulkadir Canatar, SueYeon Chung, and Daniel Lee. Estimating dimensionality of neural representations from finite samples. In *International Conference on Learning Representations*, volume 2026, pp. 93949–93979, 2026.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063, 2024.

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

Philip Feldman, James R Foulds, and Shimei Pan. Trapping llm hallucinations using tagged context prompts. *arXiv preprint arXiv:2306.06085*, 2023.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. Translategemma technical report. *arXiv preprint arXiv:2601.09012*, 2026.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Google. Gemma: Introducing new state-of-the-art open models. *https://blog.google/technology/developers/gemma-open-models/*, 2024.

Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.

Audun Josang, Jin-Hee Cho, and Feng Chen. Uncertainty characteristics of subjective opinions. In *2018 21st International Conference on Information Fusion (FUSION)*, pp. 1998–2005. IEEE, 2018.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

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

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.

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

Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, and Ken Fukuda. On the role of unobserved sequences on sample-based uncertainty quantification for llms. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 179–183, 2025.

Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, Ken Fukuda, and Masahiro Hamasaki. Evidential semantic entropy for llm uncertainty quantification. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7107–7122, 2026.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Xin Liu, Farima Fatahi Bayat, and Lu Wang. Enhancing language model factuality via activation-based confidence calibration and guided decoding, 2024. URL `https://arxiv.org/abs/2406.13230`.

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

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

Deep S Pandey, Spandan Pyakurel, and Qi Yu. Be confident in what you know: Bayesian parameter efficient fine-tuning of vision foundation models. *Advances in Neural Information Processing Systems*, 37:44814–44844, 2024.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533, 2024.

Jingen Qu, Yufei Chen, Xiaodong Yue, Wei Fu, and Qiguang Huang. Hyper-opinion evidential deep learning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37: 84645–84668, 2024.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2022.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Weishi Shi, Xujiang Zhao, Feng Chen, and Qi Yu. Multifaceted uncertainty estimation for label-efficient deep learning. *Advances in neural information processing systems*, 33:17247–17257, 2020.

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

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms, 2024. URL https://arxiv.org/abs/2412.14737.

Evidential Semantic Uncertainty for Autoregressive LLM Generation

Supplementary Material

TABLE OF CONTENTS

A Organization **14**

B Summary of Notations **14**

C Mathematical Proof **14**

D Details of Experiments and Additional Results **15**

D.1 Implementation Details 15

D.2 Details of Baselines and Evaluation Metrics 15

D.3 Impact of N 16

D.4 Impact of sampling strategies 16

D.5 Performance of K and Total Evidence 16

D.6 Performance on machine translation task 17

D.7 Impact of belief quantification 18

D.8 Study of $\tilde{\boldsymbol{a}}^H$ and W^H 18

D.9 Impact of NLI model 19

D.10 Impact of Accuracy Estimation 19

D.11 Impact of τ 20

D.12 Performance of Internal Signal-based Scores 20

D.13 Calibration performance 20

D.14 Handling consistently wrong output 21

D.15 Limitation of semantic entropy in capturing overlaps 22

D.16 Qualitative Analysis 23

E Limitations and Future Work **23**

F Impact Statement **24**

G Source Code **24**

A ORGANIZATION

First, we present a summary table of notations in Appendix B. Next, we present proof of proposition 1 in Appendix C. Next, we present details of experiments and additional results in Appendix D and discuss limitations and potential future work in E. We discuss impact statement in F. We provide link to the source code in Appendix G.

B SUMMARY OF NOTATIONS

Notation	Description
U	Uncertainty score
$\mathbf{b}$	Belief masses of classes
w	Multinomial opinion
$\boldsymbol{\alpha}$	Dirichlet PDF parameters
$\mathbf{a}$	Base rate of K classes
u	Vacuity (uncertainty mass) of multinomial opinion
$\texttt{diss}$	Dissonance of multinomial opinion
$U^{\texttt{multi}}$	Total uncertainty of multinomial opinion
f_θ	Neural network with parameter θ
K	Number of classes
e_k	Evidence of class k
$\mathcal{E}$	Activation function
W	Non-informative prior weight
S	Strength of an opinion
w^H	Hyper-opinion
S^H	Strength of Hyper-opinion
$\mathbf{b}^H$	Hyper belief masses of classes
u^H	Hyper-vacuity (uncertainty mass) of hyper-opinion
$U^{\texttt{hyper}}$	Total uncertainty of hyper-opinion
$\texttt{diss}^H$	Hyper-Dissonance of hyper-opinion
$\mathbf{a}^H$	Base rate of classes in hyper-opinion
W^H	Non-informative prior weight of hyper-opinion
L	Normalized graph Laplacian
A	Similarity matrix between clusters
λ_k	k^{th} eigen value of L
$\tilde{\boldsymbol{a}}^H$	Inter-group base rate of hyper-opinion
x_{query}	Question asked to LLM
$\mathbf{y}$	Autoregressive output for input x_{query}
t	t^{th} token
T	Total number of generated tokens in $\mathbf{y}$
N	Total answers generated for x_{query}
$\mathcal{Y}$	Set of total answers, $\lvert\mathcal{Y}\rvert = N$
$\mathcal{Y}^{sem}$	Set of semantic classes, , $\lvert\mathcal{Y}^{sem}\rvert = K$
$\mathcal{C}_k$	k^{th} answer group
$\mathcal{Q}$	Set of all unique pairs of answer groups
$Bal(b_i, b_j)$	Belief balance between two answer groups
b_{cor}	Belief mass of correct label
ib	Sum of beliefs of incorrect labels

C MATHEMATICAL PROOF

In this section, we provide the proof for the proposition 1. First, we divide belief masses into correct and incorrect classes. b_{cor} denotes the belief mass of the correct semantic label and ib denotes the

sum of beliefs of incorrect labels. From (Pandey et al., 2024), we leverage the relationship between ib and *diss*:

$$
\begin{aligned}
ib &\geq \frac{diss}{2} \\
u + ib &\geq u + \frac{diss}{2} \quad \text{[Adding } u \text{ to both side]} \\
-(u + ib) &\leq -(u + \frac{diss}{2}) \quad \text{[Multiplying both side by } -1\text{]} \\
1 - (u + ib) &\leq 1 - (u + \frac{diss}{2}) \quad \text{[Adding 1 to both side]} \\
b_{cor} &\leq 1 - (u + \frac{diss}{2}) \quad \text{[Using relation from Equation 2, } u + b_{cor} + ib = 1\text{]}
\end{aligned}
$$

This completes the proof of Proposition 1.

D Details of Experiments and Additional Results

D.1 Implementation Details

We conduct our experiments on an NVIDIA RTX A6000 GPU with 48 GB of VRAM. The proposed method and baselines are implemented in Python using PyTorch and the Hugging Face transformers library. We evaluate open-weights models from distinct families: google/gemma-3-4b-it, meta-llama/Llama-2-7b-chat-hf, and Qwen/Qwen3-VL-8B-Instruct, testing across diverse text backbones, as well as a Vision-Language Model (VLM) evaluated in text-only mode, to assess cross-architecture generalization.

For inference, we use the value of $N = 25$ in our experiments. To obtain N answers, we use nucleus sampling with a value of top p as 0.9 and temperature as 1.0. We leverage deberta_xlarge as an NLI model. For obtaining the accuracy of a sample, we use the Meta-Llama-3-8B-Instruct model as a judge to decide if the predicted answer is correct, given the question and the reference answer. The specific prompt is as follows:

Evaluation Prompt Template

We are assessing the quality of answers to the following question: <*question*>
The expected answer is: <*correct answer*>.
The proposed answer is: <*predicted answer*>.
Within the context of the question, does the proposed answer mean the same as the expected answer?
Respond only with yes or no.
Response:

The specific instructions/code to load the models and obtain the results using Python are provided in the source code: G.

D.2 Details of Baselines and Evaluation Metrics

In this section, we provide description of the baselines we experimented in the paper and evaluation metrics we used. The baselines are:

- **NLL (Negative Log Likelihood)**: It is quantified as the average of the negative log likelihood across the tokens of the generated content.
- **PE (Predictive Entropy)** (Malinin & Gales, 2020): It is quantified as the average of standard entropy across the tokens.
- **Discrete Semantic Entropy(DSE)** (Kuhn et al., 2023): First, N samples are generated for a single query, and clusters are obtained based on bidirectional entailment. Each cluster

obtains probability as the frequency of the samples belonging to the cluster. Finally, entropy is calculated over these clusters.

- **Semantic Entropy(SE)** (Kuhn et al., 2023): First, clusters are obtained similar to the process in Discrete Semantic Entropy. Then, the probability of each cluster is assigned as the sum of the probability of responses belonging to the cluster and normalized across clusters. The probability of a response is the product of conditional probabilities across tokens. Finally, entropy is calculated over these clusters.
- **P(True)**(Kadavath et al., 2022): The evaluated model is asked to assess the correctness of the answer among "True" or "False". The probability of the "True" token is used as a confidence score.
- **Degree / EigenValue (EigV)** (Lin et al., 2023): It defines a graph laplacian over the similarity matrix of the responses. The sum of eigenvalues and degree of the Laplacian is leveraged to define the semantic dispersion of generated answers.
- **Kernel Language Entropy (KLE)** (Nikitin et al., 2024): It generalizes SE by using positive semidefinite kernels to capture fine-grained semantic similarities rather than "hard" clusters.
- **Max Prob**: The maximum of the probability among clusters from the semantic entropy baseline is leveraged as a confidence score.
- **CoCoA** (Vashurin et al.): A hybrid framework that integrates internal confidence (conditional probabilities) with external consistency (similarity between multiple generated samples).
- **Semantic Density** (Qiu & Miikkulainen, 2024): Measures uncertainty by analyzing the concentration of generated responses in a continuous semantic space.
- **EVSE (Evidential Semantic Entropy)** (Kunitomo-Jacquin et al., 2026): Leverages Dempster-Shafer evidence theory to model uncertainty and derives basic belief assignments (mass functions) from standard normalized sequence/token probabilities.
- **EOS-UP (EOS-Inclusive Unobserved Probability)** (Kunitomo-Jacquin et al., 2025): Quantifies uncertainty using the total probability mass of unobserved response sequences while strictly including the End-of-Sequence (EOS) token.

For evaluation, we leverage metric of binary classification (AUROC). The classification problem in this case is to detect incorrect samples (positive label) using the uncertainty score. Further, we report the Area Under the Accuracy-Rejection Curve (AUARC) to evaluate the practical utility of uncertainty for selective prediction. This metric measures the model's accuracy improvement as a function of the percentage of samples rejected (abstention) based on their assigned uncertainty.

D.3 Impact of N

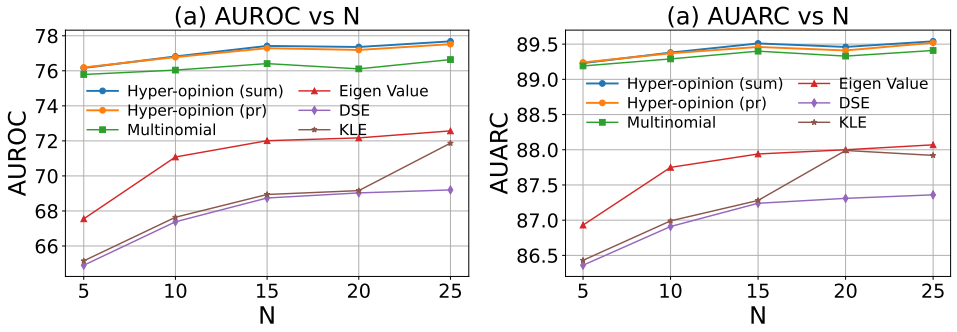
In this section, we present a comparative study of N's impact on the proposed method, along with competitive baselines. The comparison of both metrics (AUROC and AUARC) is presented in Figures 7a and 7b for the TriviaQA dataset using the Gemma model. The study shows that the performance improves as we increase the value of N for all the compared methods. Further, the proposed formulation (both hyper-opinion variants and multinomial) remains competitive for all values of N. It is also observed that with a lower value of N, the performance of strong baselines: KLE and eigenvalue, is significantly affected compared to multinomial and hyper-opinion methods.

D.4 Impact of sampling strategies

For Table 1, we obtained N answers using nucleus sampling. In this section, we study the impact of alternative sampling methods: top-k, and temperature sampling on the proposed method, along with competitive baselines. The results on TriviaQA, gemma-3-4b-it model are presented in Table 3. The superior performance either of the variants of the proposed U^{hyper} is observed across all sampling strategies. This validates the robust performance of the proposed method.

D.5 Performance of K and Total Evidence

As a motivation to evidential formulation, we demonstrated the confidence characteristics of total evidence in Figure 4. We also plotted the distribution of K for correct and incorrect samples in Figure



(a) AUROC comparison (b) AUARC comparison

Figure 7: (a) AUROC and (b) AUARC for proposed method and competitive baselines for different values of N

Table 3: Impact of sampling strategies

Decoding	DSE	$U^{\texttt{multi}}$	KLE	Eigen Value	$U^{\texttt{hyper}}$ (Sum)	$U^{\texttt{hyper}}$ (PR)
			AUROC			
Nucleus	69.20	76.64	71.87	72.57	**77.68**	77.52
Top-k	71.09	75.04	70.61	71.27	77.67	**77.87**
Temperature	66.28	75.57	66.48	69.64	75.94	**76.60**
			AUARC			
Nucleus	87.36	89.42	87.92	88.07	**89.54**	89.52
Top-k	87.73	89.31	86.93	87.78	89.62	**89.63**
Temperature	86.76	89.20	86.80	87.04	89.25	**89.36**

Table 4: AUROC comparison of K and Evidence

Model	Method	TriviaQA	Squad	NQ	BioASQ
Gemma	K	68.98	72.80	71.50	65.83
	Evidence	73.28	74.52	71.30	66.94
	$U^{\texttt{multi}}$	76.64	77.85	70.95	66.77
	$U^{\texttt{hyper}}$ **(PR)**	77.52	78.54	76.42	75.78
	$U^{\texttt{hyper}}$ **(Sum)**	**77.68**	**78.71**	**77.00**	**76.48**
LLaMA	K	73.67	68.61	66.31	65.98
	Evidence	71.43	67.69	66.19	70.39
	$U^{\texttt{multi}}$	76.79	69.39	66.42	67.17
	$U^{\texttt{hyper}}$ **(PR)**	**79.36**	**72.22**	71.99	74.91
	$U^{\texttt{hyper}}$ **(Sum)**	78.69	72.20	**72.59**	**75.44**

6a to reveal its uncertainty characteristics. In this section, we present the performance comparison of these metrics with the proposed method. The results across 4 datasets are presented in Table 4. The performance of both K and evidence is observed to be competitive in comparison to the proposed metrics, strengthening the motivation behind the proposed evidential formulation.

D.6 Performance on machine translation task

In the main Table 1, we present results on QA datasets. In this study, we experiment with a machine translation dataset: WMT14 (Bojar et al., 2014) using model: translategemma-4b-it (Finkelstein et al., 2026). The AUROC/AUARC results of the proposed methods are presented in Table 5, along with baselines. The proposed formulations outperform the baselines, demonstrating the robustness of uncertainty metrics in tasks beyond Question Answering.

Table 5: Performance on WMT14

Method	AUROC	AUARC
NLL	54.67	58.04
SE	52.43	57.01
DSE	52.41	57.02
PE	56.72	59.25
Eigen Value	55.78	58.76
KLE	55.81	58.69
Max Prob	52.84	57.20
CoCoA	55.03	58.32
Semantic Density	54.62	58.06
$U^{\texttt{multi}}$	64.54	65.61
$U^{\texttt{hyper}}$ (PR)	<u>65.33</u>	<u>65.94</u>
$U^{\texttt{hyper}}$ (Sum)	**65.40**	**66.09**

D.7 Impact of belief quantification

Table 6: Impact of belief score

Setting	AUROC	AUARC
Min Belief	74.86	88.75
Prod Belief	74.52	88.69
$U^{\texttt{hyper}}$ (pr)	77.52	89.52
$U^{\texttt{hyper}}$ (sum)	**77.68**	**89.54**

In this section, we present the performance comparison of alternative methods of calculating the belief of answer groups. In the proposed method (both variants of $U^{\texttt{hyper}}$) , we first quantify the average evidence of each answer and aggregate them for obtaining evidence of the class. In this study, we study how performance changes when we quantify the belief at token level and aggregate them across multiple tokens to obtain belief of an answer, and consequently obtain belief of each class. We consider two alternative designs for token belief aggregation to obtain belief of an answer: (i) by getting the minimum among token beliefs, and (ii) getting the product of belief tokens as the answer's belief score. The belief of each token is obtained as:

$$b_t = \frac{e_t}{W}, \quad e_t = \mathcal{E}(f_\theta(y_t^{ik} \mid x_{\text{query}}, \boldsymbol{y}_{<t}^{ik}) \tag{9}$$

IWhile quantifying the belief of each token, following the evidential theory, the value of W becomes the vocabulary size of the model, making belief masses insignificant. We reduce the size of W to the total number of tokens with positive logits. The results are presented for the TriviaQA dataset and the gemma-3-4b-it model in Table 6. While the other alternatives also show competitive performance for detecting incorrect samples, the proposed variants of hyper-opinion based uncertainty are seen to be the superior design choice for belief quantification.

D.8 Study of $\tilde{\boldsymbol{a}}^H$ and W^H

Table 7: Impact of $\tilde{\boldsymbol{a}}^H$ and W^H

Setting	AUROC	AUARC
Setting 1 (pr)	77.16	89.47
Setting 1 (sum)	77.13	89.47
Setting 2 (pr)	76.51	89.31
Setting 2 (sum)	76.53	89.29
$U^{\texttt{hyper}}$ (pr)	77.52	89.52
$U^{\texttt{hyper}}$ (sum)	77.68	89.54

The proposed formulation uses the base rate $\tilde{\boldsymbol{a}}^H$ and non-informative prior weight W^H in hyper-opinion-based fine-grained uncertainties. The illustration of $\tilde{\boldsymbol{a}}^H$ and W^H is presented in Figure 8 for the cases of no-overlap, partial overlap, and complete overlap. The design of the base rate should be such that it falls in the range of $[0, 1]$, becomes maximum when there is no overlap, becomes minimum when there is complete overlap, and increases as it starts to contradict with each other. Similarly, the design of W^H depends on similarity matrix A. With this requirement, we leverage contradiction scores from an NLI model for quantifying $\tilde{\boldsymbol{a}}^H$and entailment score for quantifying A. Along with contradiction and entailment scores, we also receive neutral and entailment scores from NLI model. The neutral score refers to two answers that neither contradict nor entail each other. In this section, we study the impact of using alternative ways for quantifying $\tilde{\boldsymbol{a}}^H$ and A. The results are presented in

Table 8: Impact of NLI model

Method	microsoft/deberta-v2-xlarge-mnli		microsoft/deberta-large-mnli	
	AUROC	AUARC	AUROC	AUARC
DSE	69.20	87.36	69.43	87.41
Eigen Value	72.57	88.07	73.77	88.40
KLE	71.87	87.92	69.53	87.41
U^{multi}	76.64	89.42	76.53	89.41
U^{hyper} (PR)	77.52	89.52	**77.45**	**89.50**
U^{hyper} (Sum)	**77.68**	**89.54**	77.29	89.49

Table 7. In setting 1, we leverage both the neutral score and the contradiction score for $\tilde{\boldsymbol{a}}^H$. In setting 2, we leverage both the neutral score and entailment score for A. The comparison reveals a decrease in performance due to the other choices, justifying the design of the proposed formulation. Further, the addition of neutral score is more harmful to setting 2 (*i.e.,* for obtaining similarity matrix A).

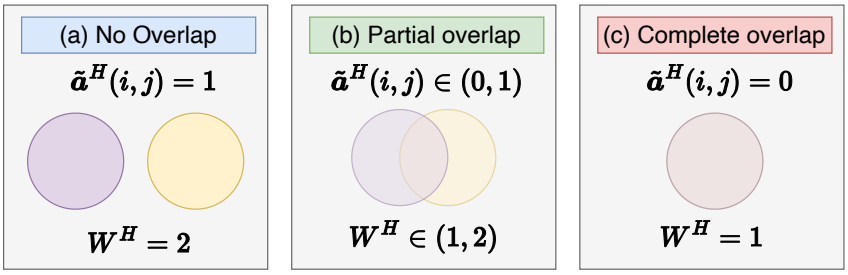


Figure 8: Visualization of inter-group base rates and effective number of semantic sets for three cases of overlaps in hyper-opinion representation

D.9 Impact of NLI model

In our experiments, we leverage microsoft/deberta-xlarge-mnli as an NLI model. In this section, we study the impact of an alternative NLI model: microsoft/deberta-large-mnli. The results of the proposed method and competitive methods are presented in Table 8. Using both of the NLI models, the proposed method consistently outperforms the baselines, confirming the robust effectiveness.

Table 9: Impact of accuracy estimation

Method	LLM		ROUGE-L	
	AUROC	AUARC	AUROC	AUARC
DSE	69.20	87.36	68.36	80.96
Eigen Value	72.57	88.07	72.05	82.31
KLE	71.87	87.92	69.68	81.39
U^{multi}	76.64	89.42	75.50	84.17
U^{hyper} (pr)	77.52	89.52	**75.86**	**84.28**
U^{hyper} (sum)	**77.68**	**89.54**	75.62	84.24

D.10 Impact of Accuracy Estimation

We calculated whether the answer is correct or not using LLM estimation for our experiments. In this section, we use an alternative way to quantify the accuracy and compare the performance of the proposed method with competitive baselines. The results of TriviaQA dataset and gemma-3-4b-it model are presented in Table 9. The accuracy from LLM is 87.99 and Rouge-L score is 79.11. While changing the evaluation metric from LLM estimation to ROUGE-L lowers absolute scores across

all methods, both proposed variants ($U_{\text{hyper(PR)}}$ and $U_{\text{hyper(sum)}}$) perform almost identically (75.86 and 75.62 AUROC) and consistently outperform all baselines.

D.11 Impact of τ

In Equation 6, we introduce a hyperparameter τ, which acts as a filter for eigenvalues to be used in the effective number of cluster counts ($u^{\texttt{hyper}}$ (sum)). In the main experiments, we use $\tau = 1.0$. In this section, we study the impact of hyperparameter τ. We also evaluate a discrete counting strategy for estimating $K_{\text{effective}}$ as an alternative to continuous spectral aggregation. Here, we directly count the number of near-zero eigenvalues below a fixed cutoff δ, given by $\sum_{k=1}^{K} \mathbb{I}(\lambda_k < \delta)$. The result is presented for the gemma-3-4b-it model and the TriviaQA dataset in Table 10. The result demonstrates that $\tau = 1.0$ yields the best result, justifying the choice of value used in the main experiment. Further, the count-based formulation instead of sum (as proposed in the paper) demonstrates suboptimal performance, justifying the choice of sum-based aggregation instead of count in the proposed method.

Table 10: Impact of τ and δ

τ	AUROC	AUARC
1.2	77.09	89.42
1.1	77.30	89.47
1.0	**77.68**	**89.54**
0.9	77.19	89.45
0.8	77.08	89.43
0.6	77.13	89.43
0.4	77.21	89.44
0.2	77.37	89.50
0.1	77.50	89.53
δ	AUROC	AUARC
0.2	77.00	89.42
0.1	77.01	89.42

D.12 Performance of Internal Signal-based Scores

In this section, we compare the performance of single-generation internal signals as confidence and uncertainty estimators against our proposed framework, as summarized in Table 11 for gemma-3-4b-it model and TriviaQA dataset. Conventional probability-based metrics such as Negative Log-Likelihood and Predictive Entropy show the weakest performance, primarily because softmax normalization flattens confidence disparities and induces overconfidence that masks model errors. Signals like Energy Score (Liu et al., 2020) and Logit Margin (difference in logits between top-2 tokens) offer slight improvements, but unnormalized signals, specifically Evidence and Logit, perform significantly better by preserving raw magnitude information prior to probability distortion. However, single-generation internal signals in isolation remain fundamentally limited because they cannot capture semantic ambiguities that emerge across alternative output spaces. Our method overcomes this limitation, outperforming all single-generation baselines. Rather than relying solely on isolated internal magnitudes, our approach provides a principled mechanism that directly ties internal signals to semantic inconsistency across multiple generations by decomposing uncertainty into fine-grained evidential components: hyper-vacuity (total lack of underlying evidence) and hyper-dissonance (conflicting evidence across semantic clusters).

D.13 Calibration performance

In this section, we study the calibration behavior of the proposed confidence metrics with baselines. The hyper-belief is the maximum belief masses presented in Methodology section. Inspired by the effectiveness of hyper-vacuity and hyper-dissonance, we also propose hyper-prob measure for this study to represent the confidence of predicted answer as:

Table 11: Performance evaluation of internal signals.

Signal	AUROC	AUARC
NLL	62.02	86.39
Predictive Entropy	66.01	86.98
Energy Score	66.04	86.76
Logit margin	65.08	86.86
Evidence (single generation)	73.04	88.65
Logit	73.04	88.65
Ours	**77.68**	**89.54**

$$p^H = \max_{k=1}^{K} \frac{e_k + \sum_{k'=1, k' \neq k}^{K} A_{kk'} e_{k'} + \mathbf{a}^H W^H}{S^H} \tag{10}$$

The evidence of class k' is shared with class k in probability calculation according to the similarity in semantic space through similarity score $A_{kk'}$. Baseline: Prod prob is obtained by multiplying conditional probabilities of tokens across a response generation. Similarly, baseline: max prob is maximum among the probabilities of semantic clusters. As a confidence score, we also define hyper-belief as the maximum among the belief scores: $b^H_{\max} = \max_{k=1}^{K} b^H_k$.

To evaluate the calibration performance, the standard calibration metric, Expected Calibration Error (ECE) is used. The results for TriviaQA and NQ datasets using the gemma-3-4b-it model are presented in Table 12. The proposed confidence metrics demonstrate competitive results in comparison to baselines.

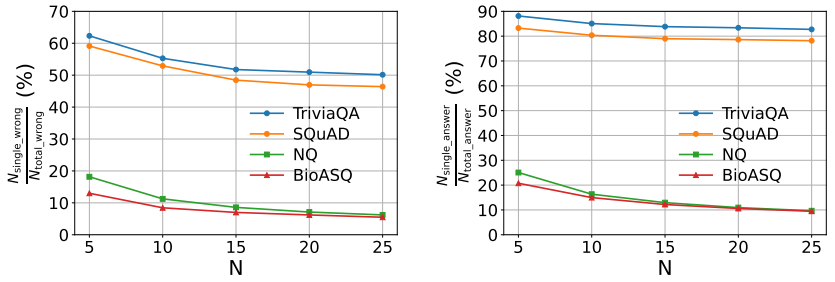
Table 12: Calibration performance comparison (ECE ↓)

Method	TriviaQA	NQ
Prod Prob	0.261	<u>0.146</u>
Max Prob	**0.081**	0.280
Semantic Density	0.109	0.556
Hyper-Belief	**0.081**	0.285
Hyper-prob	<u>0.097</u>	**0.122**

D.14 Handling consistently wrong output

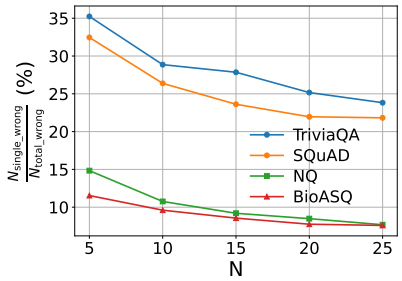
Competitive baselines (Nikitin et al., 2024; Lin et al., 2023) account for semantic overlap beyond hard equivalence classes. However, these methods still fail in the setting where the model is *consistently wrong* (eg: CASE II in Figure 1). When the sampled answers collapse to a single semantic mode, dispersion-based uncertainty metrics assign low uncertainty despite incorrectness.

To empirically demonstrate how often this extreme failure mode occurs, we evaluate `gemma-3-4b-it` and `Llama-2-7b-chat-hf` across four benchmark datasets as the sample size N increases (Figure 9). Across both architectures, a substantial fraction of errors stem from this "consistently wrong" phenomenon, where the model persistently collapses to a single, identical wrong answer ($\frac{N_{\text{single_wrong}}}{N_{\text{total_wrong}}}$). This failure mode is particularly pronounced on TriviaQA and SQuAD, reaching over 60% for `gemma-3-4b-it` and over 32% for `Llama-2-7b-chat-hf` at $N = 5$, confirming across model families that model uncertainty cannot be reliably measured by answer inconsistency alone. To explain the disparity across datasets, Figures 9b and 9d show the overall response consistency across all queries ($\frac{N_{\text{single_answer}}}{N_{\text{total_answer}}}$). The variation across datasets fundamentally comes down to how confidently the models respond overall. On TriviaQA and SQuAD, both models are generating highly consistent answers, producing the exact same answer over 78% of the time for `gemma-3-4b-it` and over 46% for `Llama-2-7b-chat-hf`. Because the models rarely vary their outputs on these tasks, their mistakes naturally collapse into a single, confident wrong answer.

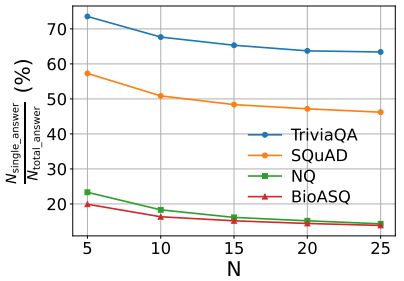


(a) Single Wrong Answer (gemma-3-4b-it)

(b) Overall Single Answer(gemma-3-4b-it)

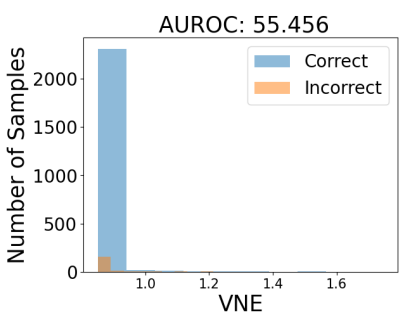


(c) Overall Single Answer(Llama-2-7b-chat)

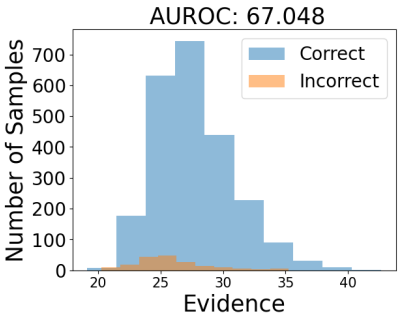


(d) Overall Single Answer(Llama-2-7b-chat)

Figure 9: Visualization of the distribution of single wrong answer ratio and overall single answer ratio across sample sizes N and models



(a) Von-neumann entropy (K=1)



(b) Evidence (K=1)

Figure 10: Visualization of distribution of (a) von-neuman entropy (K=1) and (b) evidence (K=1)

In contrast, on NQ and BioASQ, overall response consistency remains low ($< 25\%$) for both models, indicating naturally diverse outputs. As a result, when either model makes a mistake on these datasets, it spreads those errors across multiple distinct wrong answers rather than settling on just one.

To demonstrate how the competitive method handles this failure mode, we plot in Figure 10a the distributions of von Neumann entropy (VNE) for correct versus incorrect samples conditioned on $K = 1$, i.e., when all sampled answers form a single semantic group. The near-overlap of these distributions, and the corresponding AUROC, further confirms the limitation of this competitive baseline: it does not assign high uncertainty to incorrect samples when semantic spread is low. In contrast, Figure 10b shows the distribution of total evidence for the same $K = 1$ subset, indicating that evidence provides additional signal to separate correct from incorrect cases even when semantic dispersion is absent. This study further justifies the formulation in handling the challenging scenarios.

D.15 Limitation of semantic entropy in capturing overlaps

In this section, we discuss the limitation of semantic entropy. As shown in Figure 11, the same query can yield two scenarios (Cases I and II). In both cases, bi-directional entailment induces $K = 3$ semantic answer groups, and we assume the groups contain the same number of sampled responses.

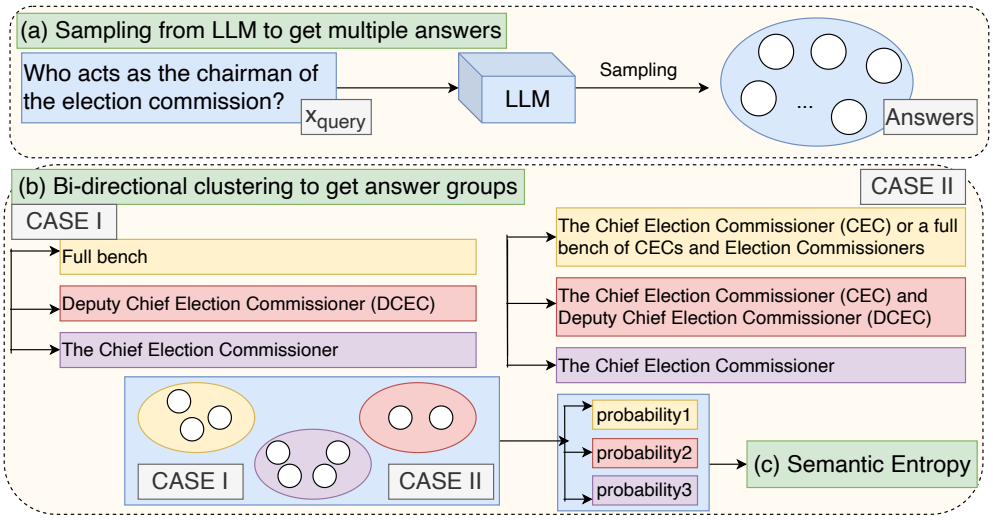


Figure 11: Overview of the bi-directional clustering to obtain answer groups (a) Sampling from LLM to get N answers for a question. (b) Cluster the answers into groups using bi-directional clustering. CASE I and CASE II are different possible instances of answer groups. Both instances have 3 groups shown in yellow, red, and green colors. The groups contain 3, 2, and 4 answers respectively. Probabilities are calculated from the answer groups to obtain semantic entropy.

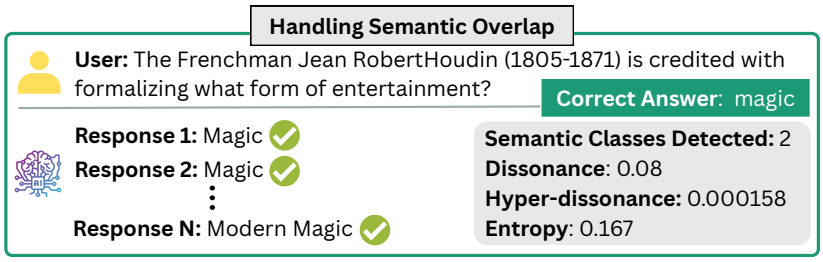


Figure 12: Qualitative analysis of semantic overlap

Consequently, semantic entropy assigns the *same* uncertainty score to both cases, since it depends only on the group counts. This is undesirable. In Case I, the groups are largely disjoint and represent genuinely distinct meanings, whereas in Case II, the groups exhibit substantial semantic similarity (i.e., overlap) with one another. Intuitively, Case II reflects lower effective semantic disagreement and should therefore receive a *lower* uncertainty score than Case I. A related limitation arises for multinomial opinions, which also treat groups as mutually exclusive singletons. This exclusivity assumption overestimates uncertainty when clusters are not well separated.

D.16 Qualitative Analysis

In this section, we provide the case of handling semantic overlap using hyper-opinion representation, which is illustrated in Figure 12. The number of distinct semantic classes detected is 2, but there exists a partial overlap between them. Semantic entropy generates a high value. Similarly, dissonance value is also higher in comparison to hyper-dissonance. By leveraging the inter-group base rate between the detected semantic classes, we are able to lower uncertainty value from hyper-opinion representation.

E Limitations and Future Work

In this work, we propose an evidential formulation to quantify fine-grained uncertainties for detecting incorrect outputs from a large language model. The uncertainty quantification is based on answer

groups obtained from clustering N generated answers. The limitation of the proposed work and recent state-of-the-arts like semantic entropy, KLE, and EigenValue is that it requires N answers for accurate uncertainty estimates, which is computationally expensive. From the ablation study, it is demonstrated that our method maintains competitive performance in comparison to baselines even when we generate less number of total answers N. However, the general trend demonstrates that performance increases with an increase in N. To make uncertainty quantification computationally efficient and competitive with baselines that generate N answers, we leave the uncertainty quantification using just 1 sample as a future work.

F Impact Statement

With the increase in usage of LLMs in the majority of tasks, it is important to ensure that the models are reliable. In this work, we enhance the reliability of language outputs from foundation models through uncertainty quantification. The proposed uncertainty quantification method acts as a useful tool to ensure trust in the predicted answer from the model.

G Source Code

The implementation code of the proposed method is provided in this link.