Uncertainty Quantification for Black-box LLMs via Star Graphs Connectivity: Exploring Alternatives for Semantic Density

Zhaoye Li[1], Huan Chen[1], Huibin Tan[1], Long Lan[1], Yize Sui[1], and Jing Ren[1] (✉)

College of Computer Science and Technology,
National University of Defense Technology, Changsha 410073, China
{lizhaoye23,chenhuan14,tanhb_,long.lan,suiyize18,renjing}@nudt.edu.cn

Abstract. Large language models (LLMs) excel in natural language processing but are prone to generating hallucinations. One approach to detecting hallucinations in LLM outputs is uncertainty quantification. These methods assign relative scores to generated responses, indicating their likelihood of being correct or hallucinatory. A well-known technique is Semantic Density, which uses the "density" of a target response in the semantic space as a proxy for its confidence. This approach addresses two limitations of Semantic Entropy: its uncertainty score is prompt-wise, and it only checks for binary semantic equivalence rather than capturing nuanced differences between two responses. Despite the success of Semantic Density, it relies on token-level probabilities, which are inaccessible in black-box LLMs, limiting its broader applicability. In this paper, we propose alternatives to Semantic Density by reconstructing uncertainty indicators from Semantic Entropy. We introduce a weighted star graph centered on the target response, reflecting the fine-grained semantic relationships between the target and other semantics within the output space. We propose using the connectivity of this star graph as a proxy for the confidence of the target response. Specifically, we present three methods based on graph density, the spectral radius of the adjacency matrix, and the spectral radius of the graph Laplacian. Our analysis shows that our approaches have a comparable computational cost to Semantic Density but outperform it in terms of both applicability and performance, making them robust alternatives.

Keywords: Uncertainty Quantification · Large Language Models · Trustworthy AI.

1 Introduction

Large language models (LLMs) excel in natural language processing, dialogue generation, and text summarization [1, 27]. However, they often produce content that sounds plausible but is factually incorrect, a phenomenon known as "hallucination" [10]. This issue is especially concerning in safety-critical fields

like healthcare, where misinformation can have severe consequences. One way to assess the reliability of LLM outputs is through uncertainty quantification (UQ), which assigns uncertainty/confidence scores to responses, thereby highlighting those that are likely accurate and those that are more prone to hallucination.

The key principle of UQ is that higher divergence (lower consistency) among multiple responses to the same input suggests a higher risk of hallucination [6,17]. One well-known method for measuring the divergence of LLM output distributions is Semantic Entropy [6, 13]. This method evaluates the degree of *semantic* divergence among multiple responses sampled from the model by calculating the predictive entropy over the predicted *meaning* distribution. Despite the success of Semantic Entropy in capturing semantic uncertainties, it has the following limitations: First, the returned uncertainty score is assigned to the prompt (i.e., prompt-wise) rather than to the individual responses being evaluated (i.e., not response-wise) [6]. Given that LLMs can generate diverse responses to the same prompt, applying the same uncertainty score to multiple potentially distinct responses is problematic [22]. Second, when comparing two responses, Semantic Entropy merely assesses semantic equivalence—treating responses with only subtle differences the same as those with major differences—and thus ignores the fine-grained distinctions that could improve the precision of uncertainty quantification (UQ) [22]. To address the two issues of Semantic Entropy, researchers have proposed Semantic Density [22]. Semantic Density is a response-wise uncertainty indicator that quantifies the confidence of LLM responses in semantic space. In this process, additional reference responses are sampled, and their fine-grained semantic differences to the target response (i.e., the response being evaluated for its reliability) are calculated. Finally, the "density" of the target response is estimated and serves as a proxy for the confidence of the target response. Although Semantic Density has made significant progress in addressing the aforementioned issues of Semantic Entropy and more accurately quantifying uncertainty in LLM responses, its applicability remains limited. This is because calculating Semantic Density requires obtaining probability information for each token generated by the LLM. However, in many cases, LLMs operate as black boxes via APIs, where users only have access to the final response text and cannot obtain token-level probability data.

This paper aims to reconstruct uncertainty/confidence indicators based on Semantic Entropy, which are suitable for black-box LLMs and serve as alternatives to Semantic Density. To measure the uncertainty/confidence level of a given target response y, we use an edge-weighted star graph to capture the fine-grained semantic relationships between y and the reference responses. By assessing the connectivity of this graph, we can evaluate the confidence of y. Specifically, we use y as the central node and other sampled reference responses as leaf nodes to form a star graph. The edge weights correspond to the semantic similarity—a continuous value between 0 and 1—that reflects degrees of semantic relatedness, rather than a binary equivalence, between y and these reference responses. Under this design, there is a positive correlation between the connectivity of the star graph and the semantic consistency between the reference responses

and the target response y. The greater the connectivity, the closer the semantic alignment between the reference responses and the target response. This suggests that the target response resides within a confident region of the output semantic space, thereby reducing the likelihood of it being a hallucination, as a key feature of hallucinations is semantic divergence [17]. Comparatively speaking, this approach contrasts with Semantic Entropy by focusing on evaluating the confidence of individual responses (by linking star graph connectivity to response confidence), rather than assessing the divergence of responses related to the prompt. This directly addresses the first limitation of Semantic Entropy. Additionally, by setting the edge weights of the star graph to reflect fine-grained semantic similarity, we capture subtle semantic differences rather than simply determining semantic equivalence, thus resolving the second limitation. Unlike Semantic Density, which requires token-level probability data to model response confidence, our star graph-based approach relies solely on the text output from LLMs, overcoming the practical limitations of Semantic Density.

We propose three simple and effective methods to measure the connectivity of the star graph in order to assess the confidence of LLM responses: graph density, the spectral radius of the adjacency matrix, and the spectral radius of the graph Laplacian. We evaluated these methods on four question-answer datasets that are widely used in current UQ literature. The experimental results show that all three methods we propose outperform baseline approaches, including Semantic Entropy and Semantic Density, achieving a new state-of-the-art (SOTA). We further validate the superiority of our methods when handling varying numbers of reference responses, and target responses with different degrees of diversity. We conclude that our methods serve as effective alternatives to Semantic Density. This is because our methods not only exhibit superior performance compared to both Semantic Density and Semantic Entropy but also offer broader applicability, requiring only the text output from LLMs rather than token probabilities. Our contributions are summarized as follows:

- We propose a new perspective that addresses two limitations of Semantic Entropy by employing the connectivity of a star graph—centered on the target response and reflecting fine-grained semantic relations—as a proxy for that response's uncertainty/confidence.
- We propose three response-wise uncertainty/confidence indicators by calculating the graph density, the spectral radius of the adjacency matrix, and the spectral radius of the graph Laplacian. Additionally, we derive possible simplified expressions and theoretical upper and lower bounds.
- Analysis and experimental results demonstrate that the three proposed methods can serve as viable alternatives to Semantic Density for the following reasons: (1) our methods have broader applicability as they only require access to the LLM's text output, without needing token probabilities; (2) the computational cost of our methods is comparable to that of Semantic Density; (3) all three methods outperform baseline approaches, including Semantic Entropy and Semantic Density, achieving a new SOTA performance.

2 Related work

In the literature on LLMs, the terms "uncertainty" and "confidence" are often used interchangeably, viewed as two aspects of the same principle—like two sides of the same coin [7, 9, 15, 22]: high confidence typically corresponds to lower uncertainty (or higher certainty). However, some studies emphasize distinguishing between the two concepts [17], with uncertainty being considered a characteristic of the predicted distribution. Despite these differences, all approaches share a common goal: deriving a score that reflects the trustworthiness of LLM responses. Higher uncertainty or lower confidence often signals potential hallucinations. In contrast, lower uncertainty or higher confidence generally indicates greater accuracy.

Recently, a variety of UQ methods [3, 6, 17, 19, 21, 22] have emerged. These methods differ in how they model uncertainty and the types of information they utilize, including the LLM's output text, token-level probabilities, internal embeddings, and model weights. UQ methods can be categorized into the two main types: white-box and black-box [7, 11, 17]. Black-box methods have access only to the LLM's output text, while white-box methods can also access the model's internal mechanisms and numerical outputs. Among these methods, Semantic Entropy [6] stands out as one of the most historically significant and is also a prominent white-box approach. Early methods, such as Predictive Entropy [18], combined lexical and semantic uncertainty, ignoring the fact that different lexical expressions can convey the same meaning. Semantic Entropy addresses this limitation by eliminating lexical uncertainty through semantic clustering (where semantically equivalent responses are grouped into clusters), marking a significant advancement in UQ. Since then, almost all UQ methods have incorporated semantics into their frameworks. Other representative methods include Deg [17], Ecc [17], and EigV [17], which exemplify black-box approaches. These methods utilize a weighted complete graph to represent the relationships between different semantics within the LLM output space, aiming to quantify semantic divergence. EigV [17] estimates the number of connected components in the graph by analyzing the eigenvalues of the graph Laplacian. In contrast, Deg [17] and Ecc [17] measure output diversity using the graph's degree matrix and the spectral embedding of its nodes, respectively. In addition to these key methods, several other approaches have been proposed. Discrete Semantic Entropy [6] serves as a black-box approximation of Semantic Entropy. Kernel Language Entropy [21] extends Semantic Entropy by incorporating fine-grained semantic relations beyond equivalence. Lastly, DUE [3] captures asymmetric logical relationships among reference responses.

3 Methodology

3.1 Task Formalization

Given a black-box LLM (where only the output text is available and the internal workings as well as numeric outputs, such as token logits, are not accessible), an

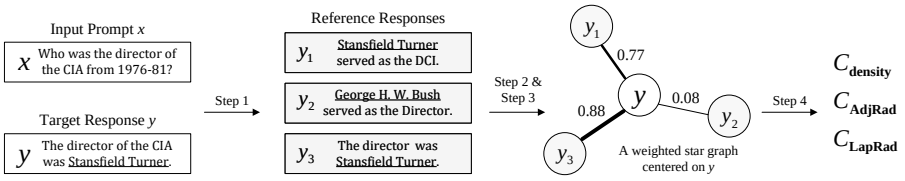

Fig. 1. In Step 1, we sample M additional responses, denoted as $y_1, y_2, \ldots, y_M$, which serve as reference responses to evaluate the reliability of y. In Step 2, we measure the semantic similarity between y and each of $y_1, y_2, \ldots, y_M$. In Step 3, we construct a star graph with y at the center and $y_1, y_2, \ldots, y_M$ as leaf nodes. Finally, in Step 4, the connectivity of the star graph is calculated, which serves as a proxy for the confidence of the target response.

input prompt x (e.g., an input question), and a target model-generated response y, the goal is to design a response-wise uncertainty/confidence indicator $C(y; x)$. It is important to note that: (1) Similar to [22], this paper specifically focuses on short-form responses, which are defined as single-proposition statements[1] [6]. (2) The objective is to derive a *relative* confidence score for ranking responses, *distinguishing* between correct and incorrect ones[2], rather than calculating the exact probability of response correctness. High confidence generally indicates correctness, whereas low confidence may indicate a potential hallucination. (3) As in [22], the methods proposed here act as confidence indicators, with higher output scores indicating a greater likelihood of correctness. If the output scores are negated, these indicators represent uncertainty indicators instead.

3.2 Step 1: Sampling Reference Responses

In Step 1, we sample M additional responses, denoted as $y_1, y_2, \ldots, y_M$, which serve as *reference responses* to evaluate the reliability of y. Following [22], we employ Diverse Beam Search [25] for sampling, since it tends to generate diverse and highly probable responses, thus providing good coverage of the LLM's semantic output space.

[1] These short-form responses are typically brief, consisting of only a few words or, at most, a single sentence, in contrast to longer paragraphs.

[2] For example, for two LLM input-output pairs $(x1, o1)$ and $(x2, o2)$, if the relative confidence score for $(x1, o1)$ is higher than that for $(x2, o2)$, then the probability of the event "$o1$ being correct for $x1$" is higher than that of the event "$o2$ being correct for $x2$". Consider the methods we propose, where (x_1, o_1) and (x_2, o_2) correspond to star graphs G_1 and G_2, respectively. If G_1 exhibits higher connectivity than G_2, it indicates that the reference responses for o_1 provide stronger support for x_1 than the reference responses for o_2 provide for x_2. Hence, the probability that "o_1 is correct for x_1" is higher than the probability that "o_2 is correct for x_2.".

3.3 Step 2: Measuring Response Similarities

In Step 2, we present a method for calculating the semantic similarity (a value between 0 and 1, not limited to binary equivalence) between the target response y and each of the reference responses $y_1, y_2, \ldots, y_M$.

The output logits of natural language inference (NLI) models have been shown to effectively measure the semantic similarity between two sentences within a given context [17][3]. Following the best practices outlined in [17], we assess the semantic similarity, denoted as s_i, between y and y_i within the context of the input prompt x. We concatenate[4] x with y and y_i to form $x \oplus y$ and $x \oplus y_i$. These concatenated strings are then fed into the NLI model twice. In the first pass, $x \oplus y$ is treated as the premise and $x \oplus y_i$ as the hypothesis. In the second pass, $x \oplus y_i$ serves as the premise, while $x \oplus y$ is the hypothesis. The softmax function is applied to the predicted logits from the NLI model, and the similarity score is computed as the average of the entailment logits from both passes, as shown in the following equation:

$$s_i = \frac{1}{2}\left(\hat{p}_{entail}(x \oplus y, x \oplus y_i) + \hat{p}_{entail}(x \oplus y_i, x \oplus y)\right) \tag{1}$$

Since $\hat{p}_{entail}$ is constrained within the interval $[0, 1]$, it follows that the similarity score s_i is also constrained within this range.

3.4 Step 3: Constructing a Star Graph

In Step 3, we construct a graph to capture the fine-grained semantic relationships between the target response y and each of the reference responses $y_1, y_2, \ldots, y_M$. This many-to-one relationship naturally forms a star graph. A star graph consists of a central node, which is connected to several peripheral nodes, known as leaf nodes. The central node, often referred to as the "hub," serves as the primary connector, while the leaf nodes are connected only to the central node and have no edges between each other.

In this study, we construct a weighted, undirected star graph consisting of $M+1$ nodes. The central node represents y, the target response that needs to be evaluated for credibility, while the M leaf nodes are represented by $y_1, y_2, \ldots, y_M$. The weight of the edge (y, y_i) is assigned the value s_i, which denotes the semantic similarity between y and y_i, as calculated in Step 2. Intuitively, in the star graph, the greater the edge weight, the higher the reachability from the leaf nodes to the central node (implying the greater importance of the central node), and the tighter the connections between all nodes in the graph. In other words, a higher semantic similarity between the reference and target responses indicates that the

[3] A more common approach to measuring semantic similarity is to compute the cosine similarity between SBERT sentence embeddings [23]. We present detailed experimental analyses and discussion in the supplementary materials (Section 5), which indicate that SBERT-based similarity is unsuitable for quantifying uncertainty.

[4] The input template used to obtain the NLI model's output logits is described in the supplementary materials (Section 10).

target response (or target semantics) lies within a more confident region of the LLM output semantic space. We quantify the "proximity of leaf nodes to the central node" as a measure of the connectivity of the star graph. The detailed methodology will be provided in the following section. Before that, we introduce several key symbols and definitions. Specifically, the adjacency matrix of the graph is given by:

$$W = \begin{pmatrix} 0 & s_1 & s_2 & \cdots & s_M \\ s_1 & 0 & 0 & \cdots & 0 \\ s_2 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_M & 0 & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} 0 & \boldsymbol{s}^T \\ \boldsymbol{s} & \mathbf{0} \end{pmatrix}. \tag{2}$$

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_M)^T$, which is an M-dimensional column vector, and $\boldsymbol{s}^T$ is its transpose, which is a row vector. In addition, the degree matrix is denoted as D, which is a diagonal matrix where each diagonal element represents the degree of a node. D is calculated as follows:

$$D = \operatorname{diag}\left(\sum_{i=1}^{M} s_i, s_1, s_2, \ldots, s_M\right). \tag{3}$$

3.5 Step 4: Calculating the Uncertainty/Confidence

In this section, we present three methods for measuring the connectivity of the star graph, which serve as a proxy for the confidence of the target response.

Graph Density For an undirected simple *binary*[5] graph, graph density [5] is a measure of how full a graph is, reflecting the ratio between the actual number of edges present (current capacity) and the maximum possible number of edges (total possible capacity). Since the graph in our paper has a fixed number of M edges and the edge weights (defined by similarities) are non-negative and bounded within $[0, 1]$, we extend the definition of graph density. We do so by calculating the sum of all edge weights (current capacity) divided by the total sum of the maximum possible weights of all edges (total possible capacity).

$$C_{\text{Density}}(y; x) = \frac{\sum_{i=1}^{M}\sum_{j=i+1}^{M} w_{ij}}{M \cdot \sup_{1\le i<j\le M}\{w_{ij}\}} = \frac{\sum_{i=1}^{M}\sum_{j=i+1}^{M} w_{ij}}{M \cdot 1} = \frac{1}{M}\sum_{i=1}^{M} s_i \tag{4}$$

C_{Density} can be interpreted as the average similarity between the reference responses $y_1, y_2, ..., y_M$ and the target response y. Since s_i is bounded within the interval $[0, 1]$, it follows that C_{Density} is also constrained within this range.

[5] In this case, the term "binary" could refer to a graph where each edge either exists or does not exist (i.e., each edge is either 0 or 1, with no other possibilities).

The Spectral Radius of the Adjacency Matrix The spectral radius, defined as the largest absolute value of the eigenvalues of a matrix, plays a pivotal role in graph theory. For a weighted (with non-negative weights) undirected graph (such as the star graph introduced in Step 3), the spectral radius of the adjacency matrix serves as a key indicator of the graph's connectivity [2]. A larger radius typically signifies stronger interactions between vertices, thereby facilitating more efficient propagation of information or flow across the graph [2].

We propose using the spectral radius of the adjacency matrix W as a proxy for the confidence of the target response. Through simplified analysis, we find that the spectral radius is essentially the ℓ_2-norm of $\boldsymbol{s}$.

$$C_{\text{AdjRad}}(y;x) = \sqrt{\sum_{i=1}^{M} s_i^2} = \|\boldsymbol{s}\|_2 \quad . \tag{5}$$

The proof of this relationship is presented as follows. Consider the eigenvalue equation:

$$W \begin{pmatrix} p \\ \boldsymbol{q} \end{pmatrix} = \lambda(W) \begin{pmatrix} p \\ \boldsymbol{q} \end{pmatrix}, \tag{6}$$

where $\lambda(W)$ is the eigenvalue and $\begin{pmatrix} p \\ \boldsymbol{q} \end{pmatrix}$ is the corresponding eigenvector. Here, p is a scalar and $\boldsymbol{q} = (q_1, q_2, \ldots, q_M)^T \in \mathbb{R}^M$ is an M-dimensional vector. Expanding the matrix multiplication in Equation 6, we get the following system of equations: (i) The first equation from the top row of W is:

$$\boldsymbol{s}^T \boldsymbol{q} = \lambda(W) p. \tag{7}$$

(ii) For the remaining M rows in W, we get $s_i p = \lambda(W) q_i$ $(1 \le i \le M)$, which implies

$$q_i = \frac{s_i}{\lambda(W)} p \quad (1 \le i \le M). \tag{8}$$

Now, substitute Equation 8 for all i into Equation 7. This gives:

$$\boldsymbol{s}^T \boldsymbol{q} = \sum_{i=1}^{M} s_i \left(\frac{s_i}{\lambda(W)} p \right) = \frac{p}{\lambda(W)} \sum_{i=1}^{M} s_i^2. \tag{9}$$

Equating Equation 9 with $\lambda(W)p$ (The right-hand side of Equation 7), we obtain $\frac{p}{\lambda(W)} \sum_{i=1}^{M} s_i^2 = \lambda(W)p$. Assuming $p \neq 0$[6], further simplification yields: $\lambda(W) = \pm\sqrt{\sum_{i=1}^{M} s_i^2}$. Since W is a real symmetric matrix, its eigenvalues are real. Therefore, the largest absolute value of the eigenvalues is:

$$C_{\text{AdjRad}} = \sqrt{\sum_{i=1}^{M} s_i^2} = ||\boldsymbol{s}||_2. \tag{10}$$

[6] This assumption does not affect the computation of the spectral radius, as explained in the supplementary materials (Section 7).

Since s_i is bounded within the interval $[0, 1]$, it follows that C_{AdjRad} is constrained within the range $[0, \sqrt{M}]$.

The Spectral Radius of the Graph Laplacian The spectral radius of the Laplacian matrix $L = D - W$ can also give indirect insights into the graph's connectivity [2]. If the radius is large, it suggests that the graph might contain nodes with very high degrees, which could indicate potential clusters [2]. We propose using the spectral radius of the graph Laplacian as a proxy for the confidence of the target response. It is well-known that the eigenvalues of the Laplacian matrix are always real (since it is a real symmetric matrix) and non-negative (since it is positive semi-definite). Therefore, the spectral radius is essentially the largest eigenvalue of the Laplacian matrix. Formally, we define:

$$C_{\text{LapRad}}(y; x) = \lambda_{\max}(L) = \lambda_{\max}(D - W). \tag{11}$$

It is evident that both C_{Density} and C_{AdjRad} (as shown in Equations 4 and 5) are tightly bounded within a specific interval. In fact, a similar conclusion holds for C_{LapRad}, despite the difficulty in deriving an explicit expression for it. In the supplementary materials (Section 6), we rigorously prove that C_{LapRad} is bounded within the interval $[0, M + 1]$. The established boundary allows practitioners to define a fixed threshold within this range for filtering out unreliable responses, thereby improving the reliability of the remaining responses (which will subsequently be evaluated using the AUARC metric).

Theorem 1 (Boundary Properties of C_{LapRad}). *For any $0 \leq s_i \leq 1$ where $1 \leq i \leq M$, C_{LapRad} is bounded within the interval $[0, M + 1]$. Specifically, $C_{LapRad} = 0$ if and only if $s_i = 0$ for all $1 \leq i \leq M$, and $C_{LapRad} = M + 1$ if and only if $s_i = 1$ for all $1 \leq i \leq M$.*

3.6 Comparative Analysis with Existing Approaches

- **Comparison with Semantic Entropy** Compared to Semantic Entropy, our methods are specifically designed to evaluate the confidence of a given response, rather than the divergence of reference responses associated with the prompt. This addresses the first issue identified with Semantic Entropy. Furthermore, by assigning edge weights to reflect fine-grained semantic similarity, we effectively capture subtle semantic differences among responses rather than simply determining whether they are semantically equivalent. This approach addresses the second limitation. *The ablation experiment demonstrates that the two aforementioned points are effective, with detailed setup and results provided in the supplementary materials (Section 14).*
- **Comparison with Semantic Density** Our methods only require access to the LLM's output text, without relying on token-level probability data. This overcomes the practical limitations associated with Semantic Density. In terms of computational cost, our proposed methods and Semantic Density have similar overhead. Both require sampling M reference responses initially.

Our methods run the NLI model $2M$ times in the second step, while Semantic Density requires *up to* $2M$ runs. Given that the NLI model consumes far fewer resources than LLMs, the computational overhead remains comparable. Further explanations are provided in the supplementary materials (Section 3).

– **Comparison with Graph-Based Methods** This part of the content is provided in the supplementary materials (Section 2).

4 Experiments and Result Analysis

4.1 Experimental Setups

Datasets and Models Currently, the evaluation of UQ in LLMs primarily focuses on question-answer datasets [3, 6, 13, 17, 21, 22]. We assess performance across a diverse range of question-answering domains, including biomedical science (BioASQ [24], 2,814 questions), trivia knowledge (TriviaQA [12], 9,960 questions), scientific knowledge (SciQ [26], 1,000 questions), and natural questions (NQ [14], 3,610 questions derived from real-world Google Search data). Detailed information regarding the datasets, their splits, and example questions for each dataset can be found in the supplementary materials (Sections 8 and 9). We utilize five well-known LLMs for evaluation, with model sizes ranging from 1B to 32B parameters. These models include Llama-3.2-1B[7], Llama-3.2-3B[8], Gemma2-2B[9], Mistral-7B-v0.3[10] and QWen1.5-32B[11]. For the NLI model used to calculate response similarities, we employ DeBERTa-Large-MNLI [8].

Evaluation Metrics Evaluation metrics include AUROC and AUARC [20], the primary measures in current uncertainty quantification literature [3, 13, 17, 21]. AUROC (Area Under the Receiver Operating Characteristic Curve) measures how well confidence scores distinguish between correct and incorrect responses. An AUROC of 0.5 indicates random guessing, while an AUROC of 1 signifies perfect discrimination, where all correct responses have higher confidence scores than all incorrect ones. Additionally, QA accuracy can be improved by abstaining from (or rejecting) low-confidence responses. This improvement is quantified using AUARC (Area Under the Accuracy-Rejection Curve) [20], which measures the area under the accuracy-rejection curve at various thresholds. The rejection accuracy at a given threshold is determined by the accuracy of the remaining responses after rejecting those with confidence scores below this threshold.

[7] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[8] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[9] https://huggingface.co/google/gemma-2-2b-it
[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[11] https://huggingface.co/Qwen/Qwen1.5-32B

Baseline Methods We included 10 baseline methods for comparison, consisting of five white-box and five black-box approaches. Although this paper primarily focuses on black-box scenarios, we also integrated state-of-the-art white-box methods to highlight the performance advantages of the proposed approaches. The white-box methods have complete access to the LLMs, including their internal mechanisms, numerical outputs, and generated text, whereas the black-box methods are limited to the LLM's output text only. White-box baselines include Predictive Entropy (PE) [18], Length-Normalized Likelihood (LNL) [19], Semantic Entropy (SE) [6], Shifting Attention to Relevance (SAR) [4], and Semantic Density (SD) [22]. Black-box baselines include Discrete Semantic Entropy (DSE) [6], Kernel Language Entropy (KLE) [21], EigV [17], Deg [17], DUE [3]. We provide a brief introduction to the baselines and their implementation details in the supplementary materials (Section 15).

Response Generation All responses to the questions were generated in free-form text. The prompts used for generating these responses are provided in the supplementary materials (Section 9). Following [22], we used Diverse Beam Search [25] to sample 10 responses for each question by configuring 10 groups, with each group containing one beam. In the main experiment, we used the responses generated by the first group (generated through greedy search) as the target responses, and the responses generated by the remaining groups as the reference responses.

Correctness Metrics (Metrics for Assessing the Accuracy of Target Response) Following [6], we prompted GPT-4-0613 to verify whether the target response aligned with any ground truth answers provided by the datasets[12]. In [22], a target response is considered correct if its Rouge-L score [16] with respect to any ground truth answer exceeds 0.3. The results of using Rouge-L for correctness judgment are included in the supplementary materials (Section 13).

4.2 Main Results

The evaluation results for Llama3.2-3B and QWen1.5-32B are presented in Table 1, while those for Llama3.2-1B, Gemma2-2B, and Mistral-7B are included in the supplementary materials (Section 12). Based on the results from *five* LLMs, *four* datasets, *two* correctness metrics (GPT-4 judge and Rouge-L judge), and *two* evaluation metrics (AUROC and AUARC), totaling 80 experimental combinations (5×4×2×2=80), we present the following findings:

– **Each of the three proposed methods consistently outperforms the baseline methods across all four datasets and five LLMs.** For example, when evaluated using Llama3.2-1B, C_{AdjRad} achieves up to a 10.48%

[12] The prompt used for auto-generated correctness judgment, together with the performance evaluation of the LLM's correctness judgments, is provided in the supplementary materials (Section 11).

Table 1. Evaluation results on Llama3.2-3B and QWen1.5-32B. Results from our methods that surpass all baselines are in **bold**. The best baseline results are highlighted in green. The optimal outcomes from the three proposed methods are in blue. The correctness of a target response is determined by GPT-4-0613 based on whether it matches any of the ground truth answers. All results are presented as percentages.

Method	BioASQ		NQ		SciQ		TriviaQA	
	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Llama3.2-3B								
PE	53.55	27.51	63.75	23.47	55.51	41.50	61.54	38.78
LNL	64.59	33.97	61.42	20.89	61.14	52.59	60.18	36.57
SE	74.26	38.84	73.10	27.38	64.53	47.53	72.01	45.96
SAR	74.32	39.34	73.59	28.16	65.26	48.39	72.31	46.62
SD	75.33	40.80	72.15	27.70	69.94	51.50	74.08	47.41
DSE	74.18	38.58	72.92	27.21	64.62	47.24	71.78	45.95
KLE	73.05	38.80	70.80	26.48	63.30	47.47	71.63	45.84
EigV	70.77	36.12	70.52	25.79	58.39	43.99	69.65	44.16
Deg	74.88	40.02	74.18	28.48	65.85	49.52	72.65	46.89
DUE	73.16	38.62	72.30	27.30	62.84	47.17	70.90	45.71
C_{Density}	**78.32**	**42.21**	**77.81**	**30.50**	**76.61**	**56.49**	**77.04**	**49.26**
C_{AdjRad}	**78.18**	**42.05**	**78.23**	**31.07**	**78.14**	**57.85**	**77.72**	**49.67**
C_{LapRad}	**78.28**	**42.21**	**78.00**	**30.65**	**77.09**	**56.84**	**77.21**	**49.34**
QWen1.5-32B								
PE	43.62	20.08	55.66	22.60	41.11	66.33	60.56	70.30
LNL	66.01	30.67	63.93	23.53	60.10	72.78	55.38	59.99
SE	64.93	28.76	72.85	30.26	74.25	82.21	81.28	81.84
SAR	67.42	30.49	73.45	32.25	77.39	84.53	81.64	82.20
SD	71.57	32.71	74.83	32.49	79.17	85.85	87.65	86.04
DSE	64.98	28.69	72.42	30.05	74.15	82.64	81.06	81.77
KLE	66.31	29.50	72.38	30.77	75.00	84.15	82.72	82.97
EigV	64.04	29.29	70.23	28.48	72.26	81.61	79.50	81.00
Deg	66.88	29.90	74.34	31.97	77.22	84.85	83.26	83.40
DUE	66.37	29.82	72.37	30.65	76.08	84.39	81.54	82.70
C_{Density}	**72.85**	**32.73**	**78.88**	**34.91**	**82.58**	**87.08**	**90.17**	**87.18**
C_{AdjRad}	**73.17**	**32.76**	**79.19**	**35.11**	**82.44**	**87.01**	**90.06**	**87.15**
C_{LapRad}	**73.00**	**32.75**	**79.03**	**34.98**	**82.63**	**87.10**	**90.19**	**87.19**

higher AUROC on the SciQ dataset and up to a 5.97% improvement on the TriviaQA dataset compared to the best baseline results. To confirm that the observed performance differences are statistically significant, we conducted pairwise significance tests (as detailed in the supplementary materials (Section 18)). The results show that all p-values are significantly less than 0.05, thereby confirming the consistent performance improvement of each proposed method.

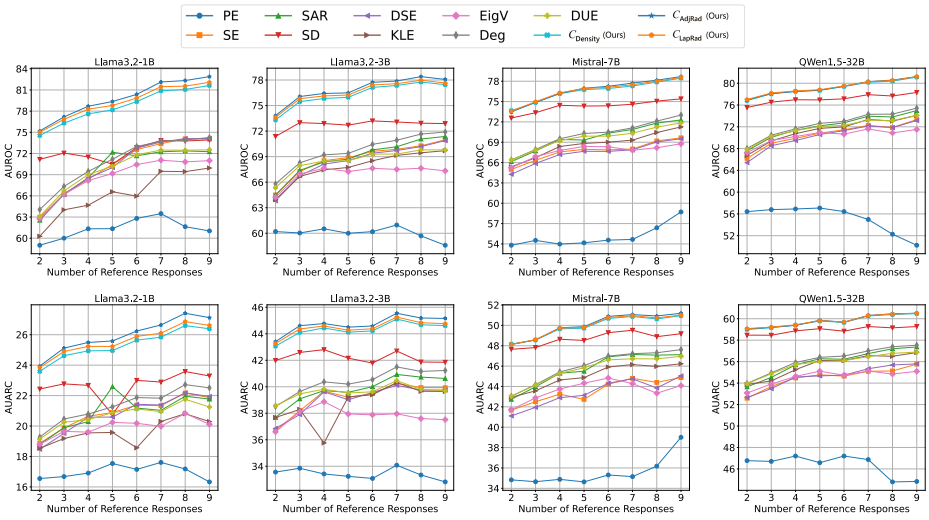

Fig. 2. The performance across varying numbers of reference responses is evaluated. Each point on the curve represents the average result across four datasets. We include all baselines for comparison, except for Length-Normalized Likelihood (LNL), as LNL operates without the need for reference responses. Each of our methods consistently outperforms all baseline methods across all numbers of reference responses.

- **Compared to Semantic Density, each of the three proposed methods shows a considerably greater improvement over Semantic Entropy.** Both Semantic Density and the three proposed methods are derived from improvements addressing two limitations of Semantic Entropy. Experimental results demonstrate that each of our proposed methods significantly outperforms Semantic Density, even without accessing the token logits (token probability) of the LLM output, highlighting both the superior performance and broader applicability of UQ modeling based on star graph connectivity over the density-based approach.
- **Compared to Semantic Density, our methods exhibit superior compatibility across different LLM sizes.** In experiments with Llama3.2-1B, Gemma2-2B, and Llama3.2-3B, Semantic Density occasionally underperforms relative to Deg. In contrast, our proposed methods consistently outperform baseline approaches across all LLM sizes, from smaller models like Llama3.2-1B to larger models like QWen1.5-32B.

4.3 Robustness of Our Proposed Methods

We adhere to the experimental setup outlined in [22] to further validate the robustness of our methods. Two experiments were conducted:

- **Performance across Varying Numbers of Reference Responses** This experiment employs the same setup as the main experiment, with the sole

distinction being the variation in the number of reference responses. Experimental results, as shown in Fig. 2, demonstrate that: (1) the performance of our methods generally improves as the number of reference responses increases. (2) Under varying numbers of reference responses, our methods consistently outperform baseline methods. (3) Compared with the baseline methods, our approaches demonstrate significantly higher generation efficiency. Specifically, our methods achieve comparable AUROC or AUARC scores while requiring fewer reference responses. Notably, in the Llama3.2-1B experiments, our approaches attain superior performance using only 2 reference responses, whereas the baseline methods require 9 reference responses to achieve similar results.

– **Performance on Target Responses with Varying Degrees of Diversity** In practical applications, users may have differing preferences for response generation strategies. Some users may favor a greedy sampling strategy, which yields more certain and consistent responses, while others may require a broader range of diverse responses. Given this consideration, we conducted this experiment. The diversity of the responses generated by diverse beam search varies across different beam groups (the first group performs a greedy beam search, while the subsequent groups encourage more diverse responses). Therefore, we use the responses from groups 2, 4, 6, 8, and 10 as target responses representing higher diversity, with responses from the other groups serving as reference responses. This setup follows the methodology outlined in [22]. Experimental results, as presented in Fig. 3, demonstrate that all of our proposed methods consistently outperform baseline approaches.

5 Future Work: Extending Our Methods for Detecting Token-Level Hallucination

Compared to Semantic Density, our methods show stronger advantages in addressing the two limitations of Semantic Entropy. However, similar to Semantic Entropy and Semantic Density, our methods can only assess the overall relative correctness of the entire target response, but cannot evaluate specific tokens (words or word pieces). To address this issue, we propose the following solution. First, the entire response (which may be a lengthy paragraph) is decomposed into multiple question-answer pairs, with each pair corresponding to a specific text snippet in the original content. Subsequently, the answer within each pair is treated as the target response, and additional short-form reference responses are generated based on the associated question. Confidence scores are then calculated for each question-answer pair. Given that each question-answer pair uniquely corresponds to a specific text snippet in the original content, the confidence score can be interpreted as the confidence level for the respective text snippet. This approach enables more precise identification of hallucinations at the token level.

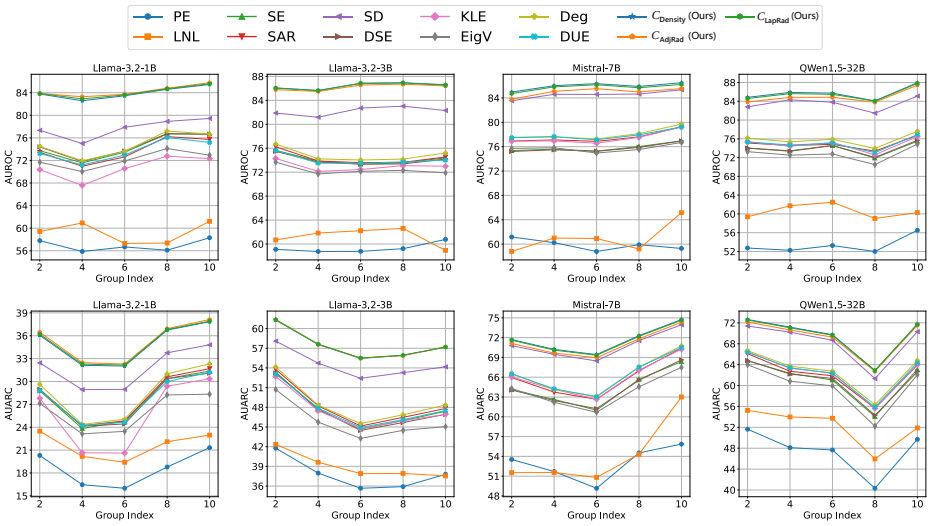

Fig. 3. Performance on target responses with varying degrees of diversity. We use responses from groups 2, 4, 6, 8, and 10 as target responses with varying diversity levels, while other groups provide reference responses. Each curve point shows the average result across four datasets. All baselines are included for comparison.

6 Conclusion

In this paper, we propose simple yet effective methods for uncertainty quantification as alternatives to Semantic Density. Specifically, we provide a new perspective that addresses two limitations of Semantic Entropy by using the connectivity of a specially tailored star graph as a proxy for the confidence of the target response. We propose using the graph density, the spectral radius of the adjacency matrix, and the spectral radius of the graph Laplacian as proxies for confidence. Analysis and experimental results demonstrate that the three proposed methods can serve as viable alternatives to Semantic Density.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No. 62376282). Furthermore, we would like to express our gratitude to the area chair and reviewers for their insightful and constructive feedback, which greatly enhanced the clarity and overall quality of this paper.

References

1. Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., Han, B.: Unveiling causal reasoning in large language models: Reality or mirage? Advances in Neural Information Processing Systems **37**, 96640–96670 (2024)
2. Chung, F.R.: Spectral graph theory, vol. 92. American Mathematical Soc. (1997)

3. Da, L., Chen, T., Cheng, L., Wei, H.: LLM uncertainty quantification through directional entailment graph and claim level response augmentation. arXiv preprint arXiv:2407.00994 (2024)
4. Duan, J., Cheng, H., Wang, S., Zavalny, A., Wang, C., Xu, R., Kailkhura, B., Xu, K.: 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)
5. ERDdS, P., R&wi, A.: On random graphs I. Publ. math. debrecen **6**(290-297), 18 (1959)
6. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
7. Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., Gurevych, I.: 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). pp. 6577–6595 (2024)
8. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021)
9. Hu, M., Zhang, Z., Zhao, S., Huang, M., Wu, B.: Uncertainty in natural language processing: Sources, quantification, and applications. arXiv preprint arXiv:2306.04459 (2023)
10. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems (2023)
11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM computing surveys **55**(12), 1–38 (2023)
12. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: 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). pp. 1601–1611 (2017)
13. Kuhn, L., Gal, Y., Farquhar, S.: Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In: The Eleventh International Conference on Learning Representations (2023)
14. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al.: Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019)
15. Liang, X., Song, S., Zheng, Z., Wang, H., Yu, Q., Li, X., Li, R.H., Wang, Y., Wang, Z., Xiong, F., et al.: Internal consistency and self-feedback in large language models: A survey. arXiv preprint arXiv:2407.14507 (2024)
16. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04). pp. 605–612 (2004)
17. Lin, Z., Trivedi, S., Sun, J.: Generating with confidence: Uncertainty quantification for black-box large language models. Transactions on Machine Learning Research (2024)

18. Lindley, D.V.: On a measure of the information provided by an experiment. The Annals of Mathematical Statistics **27**(4), 986–1005 (1956)
19. Murray, K., Chiang, D.: Correcting length bias in neural machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 212–223 (2018)
20. Nadeem, M.S.A., Zucker, J.D., Hanczar, B.: Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In: Machine Learning in Systems Biology. pp. 65–81. PMLR (2009)
21. Nikitin, A., Kossen, J., Gal, Y., Marttinen, P.: Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
22. Qiu, X., Miikkulainen, R.: Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
23. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
24. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics **16**, 1–28 (2015)
25. Vijayakumar, A., Cogswell, M., Selvaraju, R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse beam search for improved description of complex scenes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
26. Welbl, J., Liu, N.F., Gardner, M.: Crowdsourcing multiple choice science questions. In: Proceedings of the 3rd Workshop on Noisy User-generated Text. pp. 94–106 (2017)
27. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)