Lowest Span Confidence: A Zero-Shot Metric for Efficient and Black-Box Hallucination Detection in LLMs

Anonymous ACL submission

Abstract

Hallucinations in Large Language Models (LLMs), *i.e.*, the tendency to generate plausible but non-factual content, pose a significant challenge for their reliable deployment in high-stakes environments. However, existing hallucination detection methods generally operate under unrealistic assumptions, *i.e.*, either requiring expensive intensive sampling strategies for consistency checks or white-box LLM states, which are unavailable or inefficient in common API-based scenarios. To this end, we propose a novel efficient zero-shot metric called **Lowest Span Confidence (LSC)** for hallucination detection under minimal resource assumptions, **only requiring a single forward with output probabilities**. Concretely, LSC evaluates the joint likelihood of semantically coherent spans via a sliding window mechanism. By identifying regions of lowest marginal confidence across variable-length n-grams, LSC could well capture local uncertainty patterns strongly correlated with factual inconsistency. Importantly, LSC can mitigate the dilution effect of perplexity and the noise sensitivity of minimum token probability, offering a more robust estimate of factual uncertainty. Extensive experiments across multiple state-of-the-art (SOTA) LLMs and diverse benchmarks show that LSC consistently outperforms existing zero-shot baselines, delivering strong detection performance even under resource-constrained conditions.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across diverse natural language processing tasks, enabling their deployment in high-stakes domains such as healthcare (Wang et al., 2025), finance (Dong et al., 2025), and autonomous agents (Wang et al., 2024). However, their tendency to generate fluent yet factually incorrect content—commonly known as hallucination—remains a critical barrier to reliable usage (He et al., 2024; Manakul et al., 2023; Lin et al., 2022). In contexts where factual accuracy is paramount, even a single hallucinated claim can erode user trust and lead to harmful outcomes, underscoring the need for effective and deployable hallucination detection mechanisms.

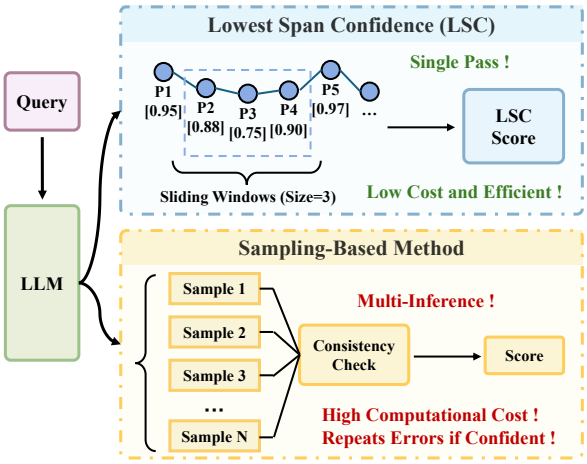


Figure 1: Comparison between LSC and Sampling-based Methods.

Hallucination detection methods can be broadly categorized into two paradigms: response consistency analysis and internal state inspection.

Δ Response consistency analysis relies on generating multiple outputs for the same input and measuring semantic or lexical agreement among them (Manakul et al., 2023; Zhang et al., 2023; Chen et al., 2024). Although these methods can identify inconsistent generations, they incur substantial computational overhead due to repeated sampling, which is often prohibitive in latency-sensitive applications. More critically, they are prone to mode collapse: *when an LLM is overconfident in an erroneous fact, it may reproduce the same hallucination across all samples, rendering consistency-based signals unreliable.*

Δ Internal state inspection usually leverages white-box model information, such as attention weights or hidden states (He et al., 2024; Chuang

et al., 2024). Though capable of detecting uncertainty without external retrieval, these methods require access to high-dimensional intermediate representations, while most commercial LLM APIs only expose generated tokens and optionally output log-probabilities. As a result, both major lines of work fail to satisfy the dual requirements of efficiency and black-box compatibility, limiting their practical utility.

In this work, we propose Lowest Span Confidence (LSC), a simple yet effective zero-shot metric for hallucination detection that operates under minimal resource assumptions. Unlike global metrics such as perplexity, whose signal can be diluted by long, high-confidence contexts, or single-token approaches like minimum token probability (Min-P), which are highly sensitive to local noise, LSC leverages a sliding window over the model's output log-probabilities to evaluate the joint confidence of semantically coherent spans. By identifying the span with the lowest aggregated confidence across variable-length n-grams, LSC effectively captures localized uncertainty patterns that strongly correlate with factual inconsistencies. Critically, LSC requires only a single forward pass and access to output token probabilities, rendering it highly efficient and fully compatible with black-box, API-based deployments of large language models.

The main contributions of this work are summarized as follows:

- We rethink the potential assumptions of existing hallucination detection methods, *i.e.*, either requiring expensive intensive sampling strategies for consistency checks or white-box LLM states. To address these issues, we propose **Lowest Span Confidence (LSC)**, a novel zero-shot metric using only a single forward pass and output token probabilities.

- We design a sliding window mechanism that strikes a balance between global context awareness and local sensitivity, where LSC can effectively identify continuous hallucinated segments while avoiding the noise and instability inherent in token-level metrics.

- We conduct extensive experiments across multiple LLMs and diverse hallucination benchmarks. Results show that LSC consistently outperforms existing zero-shot baselines, demonstrating that high-quality hallucination detection can be achieved efficiently under resource-constrained conditions.

2 Related Work

Hallucination poses a significant challenge to the reliable development of trustworthy LLMs (Li et al., 2024; Zhang et al., 2025). Here, we review two main paradigms for hallucination detection methods: response consistency analysis and internal state inspection.

Response Consistency Analysis. A predominant stream of research detects hallucinations based on the inconsistency of generated content. SelfCheckGPT (Manakul et al., 2023) samples multiple stochastic responses to verify if they support the original answer. SAC3 (Zhang et al., 2023) evaluates consistency across different LLMs or rephrased queries. In the embedding space, Eigenscore (Chen et al., 2024) attempts to quantify semantic inconsistency. Similarly, Lexical Similarity (Lin et al., 2022) utilizes the average similarity among responses as a consistency metric. Additionally, AGSER (Liu et al., 2025) leverages attentive and non-attentive queries to construct a hallucination estimator. However, these methods incur high computational costs due to repeated LLM inference. Furthermore, they rely heavily on randomness; if an LLM is overly confident in an incorrect answer, the resampling process may consistently reproduce the same error, rendering consistency checks ineffective (Zhang et al., 2023; Chen et al., 2024).

Internal State Inspection. The internal states of LLMs offer signals for hallucination detection (Azaria and Mitchell, 2023; Zhong et al., 2025). Classifiers can be trained using hidden states (He et al., 2024) or attention values (Chuang et al., 2024). Alternatively, some approaches incorporate external tools to assist in detection (Cheng et al., 2024; Yin et al., 2024). Research has also focused on parameter refinement to enhance factuality, employing techniques such as alignment (Zhang et al., 2024b), truthful space editing (Zhang et al., 2024a), over-trust penalties (Leng et al., 2024), and confidence calibration (Liu et al., 2024). However, these training-based approaches require annotated datasets and often suffer from poor generalization across different models and domains (Orgad et al., 2024). Furthermore, accessing white-box parameters (e.g., hidden states) is often impractical in real-world deployment scenarios.

3 Methodology

In this section, we formally define the problem of hallucination detection and introduce our proposed metric, *Lowest Span Confidence* (LSC). As illustrated in Figure 1, unlike computationally intensive consistency-based methods that require multiple generations to verify facts, LSC operates efficiently in a single inference pass.

3.1 Problem Formulation

Let x denote a user query and $\mathcal{M}$ a large language model. The model generates a response sequence $y = \{t_1, t_2, \ldots, t_T\}$, where t_i is the i-th token and T is the sequence length. Each token is sampled from the conditional probability distribution $P(t_i \mid x, y_{<i})$, with $y_{<i} = \{t_1, \ldots, t_{i-1}\}$. The objective of hallucination detection is to design a scoring function $S(y)$ such that lower scores indicate a higher likelihood of non-factual content.

A standard baseline for uncertainty estimation is **perplexity (PPL)**, defined as the exponential of the average negative log-likelihood over the generated sequence:

$$\mathrm{PPL}(y) = \exp\left(-\frac{1}{T}\sum_{i=1}^{T}\log P(t_i \mid x, y_{<i})\right). \tag{1}$$

However, perplexity suffers from the *dilution effect*: a short hallucinated span can be masked by a long stretch of high-confidence, factual tokens, yielding deceptively low global perplexity scores.

To mitigate this issue, prior work such as SelfCheckGPT (Manakul et al., 2023) has proposed using the minimum token probability,

$$\text{Min-P} = \min_i P(t_i \mid x, y_{<i}), \tag{2}$$

which is highly sensitive to low-probability tokens. However, Min-P is prone to false positives—a single low-probability token (e.g., a rare but factually correct proper noun) does not necessarily indicate a hallucination.

3.2 Lowest Span Confidence

Our core hypothesis is that hallucinations typically manifest not as isolated tokens, but as *coherent semantic spans*, such as incorrect entities, dates, or relational phrases. Consequently, assessing the joint confidence of consecutive tokens provides a more reliable signal of factuality than evaluating tokens in isolation.

Given the generated response $y = \{t_1, \ldots, t_T\}$, we define the token-wise probability sequence $P = \{p_1, p_2, \ldots, p_T\}$, where $p_i = P(t_i \mid x, y_{<i})$. To capture localized uncertainty while suppressing token-level noise, we apply a sliding window of fixed size w. For each valid starting position j (where $1 \le j \le T - w + 1$), the corresponding window is $W_j = \{p_j, p_{j+1}, \ldots, p_{j+w-1}\}$. The confidence of this span is computed as the arithmetic mean of its constituent probabilities:

$$C_j^{\text{mean}} = \frac{1}{w}\sum_{k=0}^{w-1} p_{j+k}. \tag{3}$$

We then define the **Local Span Confidence** (LSC) score of the full response y as the minimum span confidence across all windows:

$$\mathrm{LSC}(y) = \min_j C_j^{\text{mean}}. \tag{4}$$

By aggregating token probabilities over contiguous spans, LSC effectively mitigates the influence of spurious low-probability tokens (e.g., rare but correct words) while remaining sensitive to extended low-confidence regions that often signal hallucinatory content.

4 Experiments

4.1 Experimental Setup

Datasets. By (Chen et al., 2024), we conduct comprehensive evaluations on four widely adopted Question Answering (QA) benchmarks:

(1) Natural Questions (NQ) (Kwiatkowski et al., 2019): We utilize the validation split, comprising 3,610 QA pairs.

(2) TriviaQA (Joshi et al., 2017): We employ the validation set of the *rc.nocontext* subset, which contains 9,960 deduplicated examples.

(3) SQuAD 2.0 (Rajpurkar et al., 2016): Using the development split, we filter out unanswerable queries (where *is_impossible* is True), resulting in a subset of 5,928 samples.

(4) CoQA (Reddy et al., 2019): We use the development split consisting of 7,983 conversational QA pairs. These datasets cover a diverse range of sequence lengths and question types, providing a robust testbed for hallucination detection.

Models. To verify the effectiveness and scalability of our method across different architectures and model sizes, we select representative models from two prominent open-source families:

Table 1: Main results of hallucination detection on four QA datasets. AUC_s and AUC_r denote AUROC calculated based on semantic similarity and ROUGE-L correctness, respectively. **Bold** indicates the best performance, and underlined indicates the second best. All numbers are reported in percentages.

Models	Datasets	NQ			TriviaQA			SQuAD			CoQA		
	Methods	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC
LLaMA-13B	Perplexity	69.7	69.6	31.8	81.1	81.3	50.4	39.7	44.8	-6.6	51.6	59.1	4.7
	Energy	62.5	63.4	24.4	69.8	70.8	37.0	35.4	38.5	-21.1	43.7	50.5	-9.7
	LN-Entropy	68.1	67.7	28.0	77.3	77.1	42.5	53.5	56.4	9.8	58.0	63.8	16.5
	Lexical Similarity	71.1	70.9	33.3	70.9	71.5	42.9	62.0	64.3	23.4	68.7	72.9	36.1
	EigenScore	73.0	71.7	36.2	71.4	71.4	44.1	66.1	66.4	32.7	**71.4**	73.4	**40.5**
	AGSER	66.8	66.4	25.9	68.3	68.9	32.7	50.8	51.9	-1.5	65.7	66.0	25.5
	LSC	**73.8**	**72.8**	**37.0**	**81.6**	**81.7**	**54.4**	**69.1**	**69.7**	**34.8**	69.5	**73.9**	37.8
LLaMA-7B	Perplexity	71.7	72.8	30.7	82.5	82.9	50.3	40.6	46.2	-5.5	47.9	56.8	0.1
	Energy	62.8	64.6	20.5	74.6	75.6	43.2	32.2	37.4	-18.9	44.3	52.3	-7.1
	LN-Entropy	71.4	72.0	28.6	78.8	79.0	43.4	53.5	56.5	7.2	55.3	62.1	8.4
	Lexical Similarity	69.3	70.7	29.3	72.5	73.4	44.9	64.2	65.4	25.2	68.2	71.7	29.9
	EigenScore	71.4	71.7	32.8	73.3	73.7	46.2	67.5	67.8	**34.2**	**71.3**	**72.6**	**36.3**
	AGSER	64.4	64.5	25.1	70.1	70.9	37.6	43.2	45.5	-5.3	62.2	63.0	20.0
	LSC	**74.8**	**75.3**	**36.3**	**83.3**	**83.5**	**56.4**	**69.9**	**69.4**	32.7	67.7	72.1	31.3
Qwen-7B	Perplexity	76.4	75.4	15.3	83.6	83.7	48.8	38.3	42.0	-14.0	60.2	61.9	11.8
	Energy	58.1	57.4	-5.7	74.0	74.4	38.3	28.4	30.8	-34.5	36.8	39.8	-17.0
	LN-Entropy	77.2	75.7	28.2	80.2	80.3	47.0	48.9	51.6	-1.2	69.4	69.2	27.9
	Lexical Similarity	77.0	76.7	32.5	74.2	75.1	46.9	57.2	59.8	12.9	**71.6**	**71.8**	34.6
	EigenScore	78.9	78.0	**44.3**	74.5	75.0	48.5	**58.5**	**60.0**	**14.2**	71.5	71.0	**38.7**
	AGSER	54.9	57.0	-1.0	66.6	67.9	31.5	45.9	48.8	-7.6	69.1	69.1	27.4
	LSC	**81.6**	**81.0**	39.4	**84.6**	**84.6**	**57.2**	53.8	56.0	7.3	71.5	70.9	34.7
Qwen-3B	Perplexity	71.6	71.0	3.4	84.4	84.5	50.7	45.8	51.4	3.9	55.0	58.5	8.3
	Energy	47.2	47.0	-28.8	74.2	74.3	38.7	38.9	42.4	-7.0	41.1	44.2	-10.1
	LN-Entropy	72.7	72.4	15.0	81.0	81.2	43.0	49.5	53.6	4.8	65.1	65.7	19.6
	Lexical Similarity	72.0	73.6	21.1	76.6	77.5	49.8	55.1	58.6	12.0	67.2	67.9	29.1
	EigenScore	78.4	78.2	**42.3**	77.7	78.0	53.7	57.9	59.5	15.3	68.4	68.0	**35.0**
	AGSER	59.7	59.8	8.1	66.0	67.1	31.9	54.2	57.3	12.2	67.2	66.7	25.7
	LSC	**82.3**	**82.0**	41.6	**85.4**	**85.6**	**61.9**	**61.3**	**62.1**	**19.7**	**69.3**	**69.3**	31.4

`LLaMA-2` (Touvron et al., 2023) and `Qwen2.5` (Yang et al., 2025). For LLaMA-2, we evaluate the **7B** (Llama-2-7b-chat-hf) and **13B** (Llama-2-13b-chat-hf). For Qwen2.5, we conduct a more granular scaling analysis by including the **0.5B** (Qwen2.5-0.5B-Instruct), **3B** (Qwen2.5-3B-Instruct), **7B** (Qwen2.5-7B-Instruct), and **32B** (Qwen2.5-32B-Instruct). All models are loaded using the official pre-trained weights from Hugging Face. In the following sections, we refer to these models as LLaMA and Qwen for simplicity.

Baselines. We compare LSC with a diverse set of established hallucination detection methods. For uncertainty-based metrics derived directly from model outputs, we include **Perplexity** (Ren et al., 2022), the **Energy** score (Liu et al., 2020), and Length-normalized Entropy (**LN-Entropy**) (Malinin and Gales, 2020). Regarding consistency-based approaches that rely on multiple stochastic samples, we compare with **Lexical Similarity** (Lin et al., 2022), **EigenScore** (Chen et al., 2024), and the attention-guided method **AGSER** (Liu et al., 2025). More details are provided in Appendix B.

Evaluation Metrics. Consistent with established protocols for uncertainty estimation (Ren et al., 2022; Chen et al., 2024), we assess the performance of hallucination detection methods using two primary metrics. First, we report the Area Under the ROC Curve (AUROC), which quantifies the detector's capability to distinguish between factual and non-factual generations. A higher AUROC indicates superior binary classification per-

formance. Second, we employ the Pearson Correlation Coefficient (PCC) to evaluate the linear correlation between the proposed detection scores and the ground-truth correctness of the responses.

Correctness Measure. To obtain binary labels for AUROC calculation (*i.e.*, identifying whether a response is a hallucination), we adopt two complementary criteria following (Chen et al., 2024):

(1) Lexical Overlap: We use ROUGE-L (Lin, 2004) (F1-score) to measure surface-level similarity. A response is labeled as correct if the score exceeds a threshold of 0.5.

(2) Semantic Equivalence: To capture correctness beyond exact wording, we utilize the cosine similarity of sentence embeddings extracted by *nli-roberta-large* (Reimers and Gurevych, 2019). Responses with a similarity score above 0.9 are considered correct. The sensitivity analysis regarding these thresholds is detailed in Section 4.3.

Implementation Details. Our experiments are implemented using PyTorch and the Hugging Face Transformers library. For the primary response generation, we employ greedy decoding. For consistency-based baselines that require stochastic resampling, we strictly adhere to the default hyperparameter configurations in prior work (Chen et al., 2024) to ensure faithful reproduction. Specifically, the decoding parameters are set to a temperature of 0.5, top-p of 0.99, and top-k of 10. To balance estimation accuracy with computational efficiency, the number of sampled generations is limited to $K = 5$. Regarding our proposed LSC, the sliding window size is set to $w = 3$ by default. A sensitivity analysis of the window size is presented in Section 4.3. All evaluations were conducted on 8 NVIDIA H200 GPUs.

4.2 Main Results

Overall Performance. As presented in Table 1, our proposed LSC demonstrates robust superiority across diverse benchmarks, effectively bridging the gap between detection accuracy and computational efficiency. In comparison with standard uncertainty-based baselines (e.g., Perplexity, Energy, and LN-Entropy), LSC consistently achieves significant margins, as exemplified by a 0.8% improvement in AUC_s and a 6.1% gain in PCC over Perplexity on TriviaQA with LLaMA-7B. These results thereby validate the sliding window's ability to mitigate both the dilution effect of global metrics and the noise susceptibility of token-level indicators. Furthermore, relative to computationally intensive consistency-based methods like EigenScore, which necessitate multiple stochastic samplings, LSC operates in a single inference pass yet secures the top performance in the majority of settings, particularly dominating the NQ and TriviaQA datasets across all tested models. Even in scenarios where EigenScore excels (e.g., CoQA with LLaMA-13B), LSC remains a highly competitive second-best, proving that span-based confidence aggregation offers a compelling trade-off between state-of-the-art detection accuracy and practical deployment latency.

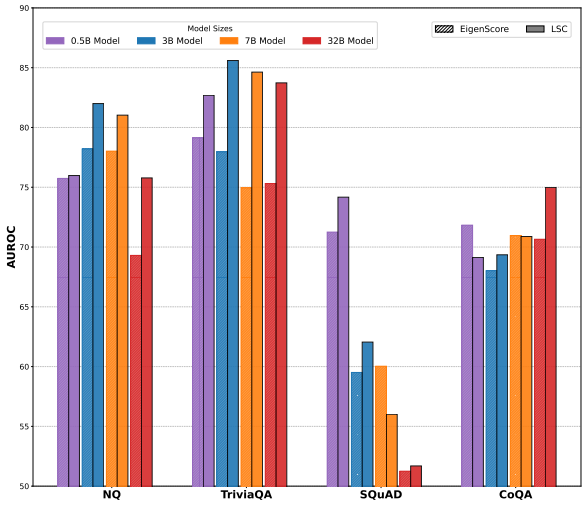


Figure 2: Scalability analysis across model sizes.

Impact of Model Size. We further investigate the robustness of LSC across varying model scales using the Qwen2.5 family (0.5B to 32B), as shown in Figure 2. Overall, LSC demonstrates remarkable scalability and consistency, proving effective across the entire parameter spectrum. This trend suggests that as models become more capable and fluent, LSC's span-based confidence aggregation remains a robust indicator of factuality, whereas consistency-based metrics may struggle. Detailed numerical results are provided in Appendix A.

4.3 Ablation Study

Sensitivity to Sliding Window Size. We investigate the sensitivity of LSC to the sliding window size w, which governs the trade-off between token-level granularity and contextual smoothing, by varying w from 1 to 8 across four datasets as illustrated in Figure 3. The results reveal a consistent trend where performance initially improves as w increases from 1 and typically peaks at $w = 2$

Table 2: Ablation results on correctness measure thresholds. Impact of varying ROUGE-L and Sentence Similarity thresholds on hallucination detection performance (AUROC) using LLaMA-7B and Qwen-7B on the NQ dataset.

Method	LLaMA-7B						Qwen-7B					
	ROUGE-L			Sentence Similarity			ROUGE-L			Sentence Similarity		
Threshold	0.3	0.5	0.7	0.7	0.8	0.9	0.3	0.5	0.7	0.7	0.8	0.9
Perplexity	**68.5**	72.8	74.0	65.5	69.4	71.7	71.2	75.4	76.9	62.3	69.2	76.4
Energy	65.8	64.6	64.2	59.0	62.0	62.8	57.4	57.4	57.4	47.7	53.9	58.1
LN-Entropy	65.5	72.0	73.8	64.0	67.9	71.4	71.2	75.7	77.5	65.6	70.8	77.2
Lexical Similarity	67.2	70.7	71.3	62.8	66.8	69.3	73.8	76.7	77.2	65.6	70.9	77.0
EigenScore	65.0	71.7	73.0	65.5	68.6	71.4	74.2	78.0	79.4	**71.0**	74.1	78.9
AGSER	61.7	64.5	65.9	61.4	62.3	64.4	55.9	57.0	56.8	49.3	52.4	54.9
LSC	67.4	**75.3**	**77.5**	**66.1**	**70.6**	**74.8**	**76.2**	**81.0**	**82.4**	69.1	**74.3**	**81.6**

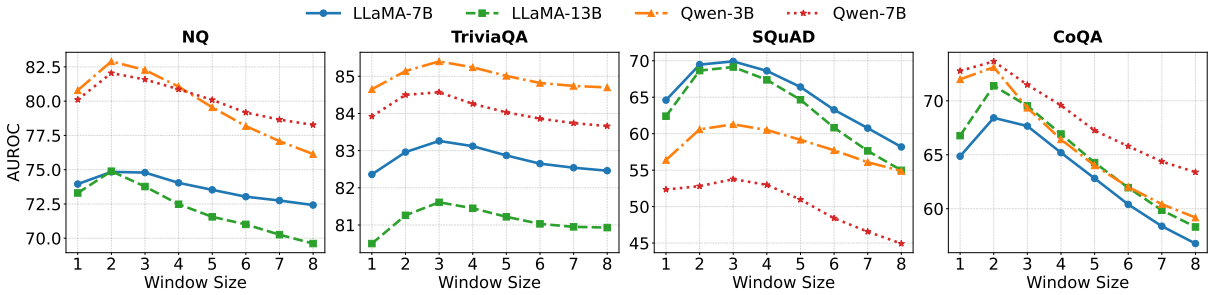


Figure 3: Ablation results of sliding window size w. The curves illustrate the impact of varying window size (from 1 to 8) on LSC detection performance across four datasets and four LLMs.

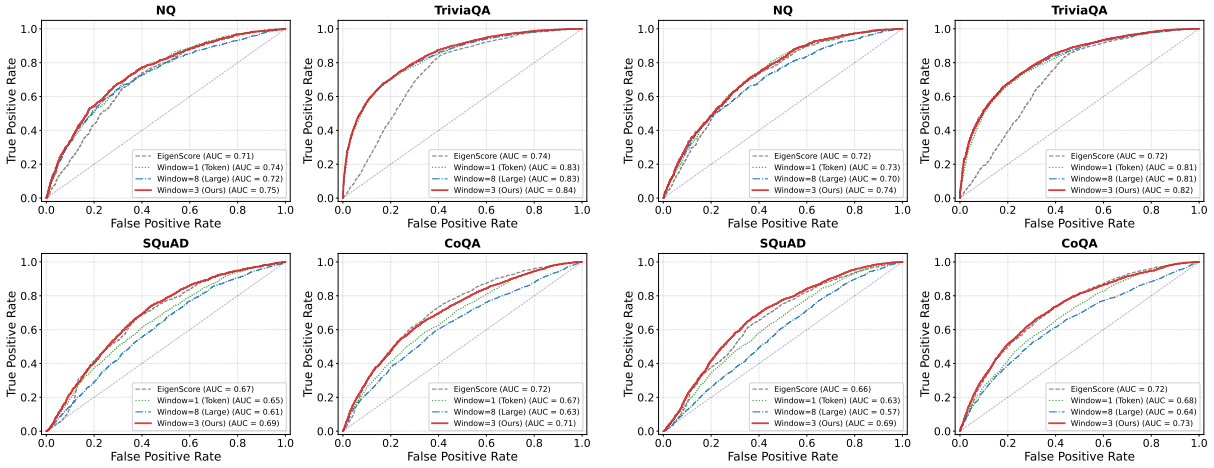
or $w = 3$, supporting our hypothesis that hallucinations often manifest as short semantic units rather than isolated tokens. Conversely, extending w beyond 4 leads to a discernible decline in AUROC, as excessively large windows dilute the signal of local uncertainty peaks by incorporating surrounding high-confidence tokens. Consequently, setting $w = 3$ serves as a robust configuration that yields near-optimal performance across diverse models and tasks.

Sensitivity to Correctness Thresholds. We evaluate the robustness of detection metrics against varying evaluation strictness by adjusting thresholds for both ROUGE-L and Sentence Similarity on the NQ dataset. As reported in Table 2, LSC demonstrates consistent superiority across nearly all configurations. Notably, the performance advantage of LSC over strong baselines like EigenScore becomes more pronounced under stricter correctness criteria, such as when the Sentence Similarity threshold exceeds 0.9, where LSC achieves an AUROC of 74.8% and 81.6% for LLaMA-7B and Qwen-7B respectively. This evidence indicates that LSC is particularly effective at identifying subtle hallucinations that require high-precision verification while maintaining stability even when the definition of correctness is relaxed.

4.4 More Analysis

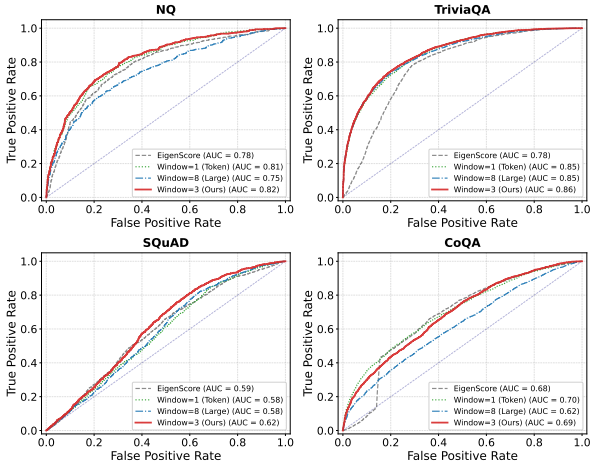
We investigate the intrinsic mechanism of LSC through a multi-granular analysis, combining macroscopic quantitative evaluations of discriminative capability across diverse models with microscopic qualitative insights into confidence trajectories and specific failure cases.

Macroscopic Analysis. We further quantify the global performance of LSC by analyzing Receiver Operating Characteristic (ROC) curves across four LLMs, as shown in Figure 4. We select EigenScore as the representative state-of-the-art baseline, alongside token-level ($k = 1$) and large-window ($k = 8$) variants to validate our windowing hypothesis. The results consistently demonstrate that the LSC curve strictly dominates the baselines across all tested models. This superiority proves robust across different model families and sizes. The comparison highlights the impact of window granularity: token-level is susceptible to local noise, result-

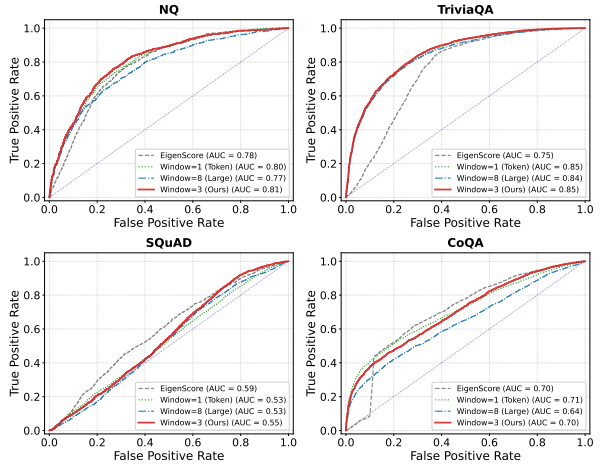


(a) LLaMA-7B

(b) LLaMA-13B



(c) Qwen-3B



(d) Qwen-7B

Figure 4: **Macroscopic performance evaluation via ROC curves across four representative LLMs.** Comparison of LSC against the SOTA baseline EigenScore and window size variations ($k = 1, 8$). Across different model families (LLaMA, Qwen) and scales (3B to 13B), LSC (solid red curve) consistently encloses the baselines. This demonstrates that LSC achieves the highest AUROC by maintaining a superior True Positive Rate while effectively suppressing False Positives.

ing in high False Positive Rates at strict thresholds; conversely, the large-window tends to over-smooth the probability distribution, which dilutes error signals and hampers the recall of subtle hallucinations. LSC effectively balances this trade-off, maintaining high sensitivity without sacrificing precision.

Microscopic Analysis. As shown in Figure 6, we visualize the confidence trajectories of generated responses to elucidate the critical trade-off governed by window size. The top panel depicts a non-hallucinated sample. Here, the token-level probability ($k = 1$) exhibits a sharp drop at the initial token ("Article"), likely due to stochastic token-level noise rather than factual error. LSC effectively mitigates this noise by aggregating local tokens, maintaining a high confidence score indicative of factuality. Conversely, the bottom panel displays a hallucinated response. The large window ($k = 8$) excessively smooths the probability distribution, diluting the error signal with surrounding high-confidence tokens and resulting in a False Negative (FN). LSC, however, remains sensitive to the span-level semantic unit.

Case Study. To qualitatively demonstrate the advantages of LSC, we analyze specific failure cases of baseline metrics as shown in Figure 5. In the first example, the model generates an incomplete response that repeats the question context but fails to provide the key entity ("Tepuis"), a hallucination that deceives global metrics like Perplexity due to

Case 1: Consistent Hallucination (Baseline False Negative)

Question: What are the table-top mountains found in the Guiana Highlands of South America (especially Venezuela) that inspired Arthur Conan Doyle's The Lost World and also appear prominently in the landscapes of the Pixar hit Up?
GT Answer: Tepuis
LLM Response: The table-top mountains of the Guiana Highlands of South America,

---

Detection Metrics (✓: Success, ✗: Failure):
Perplexity: 0.630 ✗ **Energy**: -19.918 ✗
LN-Entropy: 0.186 ✓ **Lexical Similarity**: 0.527 ✗
EigenScore: -1.201 ✗ **AGSER**: 0.600 ✗
LSC (Ours): **0.397** ✓

Sampled Generations (for Consistency Baselines):
['The table-top mountains of the Guiana Highlands in South America,', 'The table-top mountains of the Guiana Highlands in South America,', 'The Andes mountains of South America.', ...]

Case 2: Factual Response (Baseline False Positive)

Question: What were the kid's names?
GT Answer: Greta and Tony
LLM Response: Greta and Tony.

Detection Metrics (✓: Success, ✗: Failure):
Perplexity: 0.323 ✗ **Energy**: -19.823 ✓
LN-Entropy: 0.156 ✗ **Lexical Similarity**: 0.600 ✗
EigenScore: -1.831 ✗ **AGSER**: 0.333 ✗
LSC (Ours): **0.803** ✓

Sampled Generations (for Consistency Baselines):
['Tony and Greta.', 'Greta and Tony.', 'Tony and Greta.', 'Greta and Tony.', 'Greta and Tony.']

Figure 5: **Qualitative comparison of specific cases. Left (Hallucination Sample):** Selected to demonstrate how LSC correctly identifies errors where baselines suffer from false negatives. **Right (Non-hallucination Sample):** Selected to show LSC's robustness in verifying factual content where consistency-based baselines trigger false positives due to benign phrasing variations.

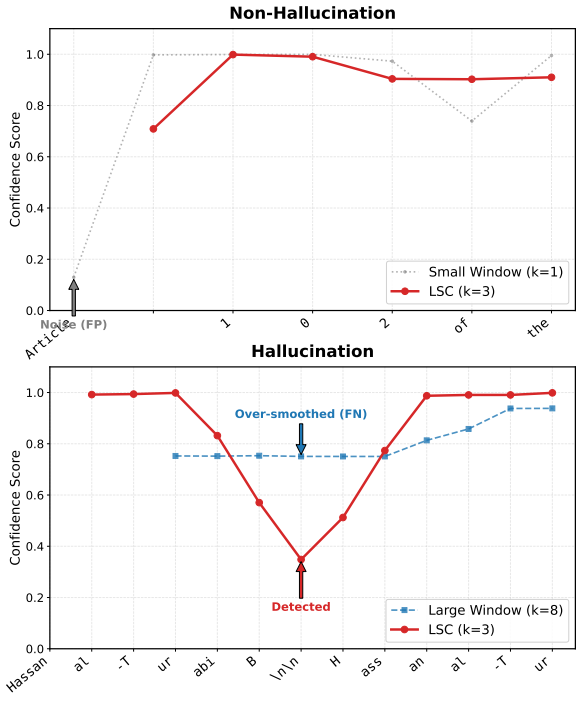


Figure 6: Microscopic analysis of trajectories.

the high-probability prefix. Similarly, consistency-based methods such as EigenScore fail to flag this error because the model exhibits mode collapse, consistently generating the same non-factual completion across multiple samples. In contrast, LSC successfully detects the hallucination by identifying the sharp drop in confidence at the span level, regardless of the high global likelihood or repeated outputs. This qualitative evidence confirms that LSC provides a more fine-grained and robust estimation of uncertainty compared to methods that rely on global averaging or stochastic sampling. Additional case studies are provided in Appendix C.

5 Conclusion

In this work, we introduced **Lowest Span Confidence (LSC)**, a novel zero-shot metric designed to bridge the gap between detection accuracy and computational efficiency in identifying LLM hallucinations. By leveraging a sliding window mechanism to evaluate span-level confidence, LSC effectively captures local uncertainty patterns, overcoming the dilution effect of global perplexity and the noise sensitivity of single-token metrics. Crucially, our approach operates under minimal resource assumptions, requiring only a single forward pass and output log-probabilities, thereby rendering it highly practical for black-box API scenarios where white-box states are unavailable. Extensive evaluations across diverse benchmarks confirm that LSC consistently outperforms existing zero-shot baselines, offering a robust and scalable solution for trustworthy LLM deployment in real-world applications.

Limitations

While LSC establishes a robust and efficient metric for identifying hallucinations via span-level uncertainty, our current study is limited to post-hoc detection without integrating active mitigation strategies to correct the generated errors. The potential of leveraging LSC as a granular feedback signal remains underexplored, specifically regarding how these localized uncertainty patterns can guide model refinement techniques such as preference optimization or reinforcement learning to penalize non-factual generation during training. Future work should investigate transforming LSC from a passive evaluation metric into an active objective for alignment, thereby closing the loop between efficient detection and the intrinsic reduction of hallucinations in foundation models.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14600–14615.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436.

Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihan Chen, and Zhi-Qi Cheng. 2025. Large language model agents in finance: A survey bridging research, practice, and real-world deployment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17889–17907.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10218–10230.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4160–4173.

Qiang Liu, Xinlong Chen, Yue Ding, Bowen Song, Weiqiang Wang, Shu Wu, and Liang Wang. 2025. Attention-guided self-reflection for zero-shot hallucination detection in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21016–21032.

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

Xin Liu, Farima Fatahi Bayat, and Lu Wang. 2024. Enhancing language model factuality via activation-based confidence calibration and guided decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10436–10448.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In

Proceedings of the 2023 conference on empirical methods in natural language processing, pages 9004–9017.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

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

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Yinuo Wang, Baiyang Wang, Robert Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. 2025. Trustworthy medical question answering: An evaluation-centric survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27477–27490.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. Sac3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024a. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Haitian Zhong, Yuhuan Liu, Ziyang Xu, Guofan Liu, Qiang Liu, Shu Wu, Zhe Zhao, Liang Wang, and Tieniu Tan. 2025. React: Representation extraction and controllable tuning to overcome overfitting in llm knowledge editing. *arXiv preprint arXiv:2505.18933*.

A Evaluation on More LLMs

To further assess the scalability and robustness of our proposed method, we extend our evaluation to extreme model sizes within the Qwen family, specifically Qwen-0.5B and Qwen-32B. The results are presented in Table 3. On the small-scale **Qwen-0.5B**, LSC remains highly competitive, achieving the best performance on TriviaQA and SQuAD. More notably, on the large-scale **Qwen-32B**, LSC demonstrates dominant performance, outperforming the strong baseline EigenScore across most datasets, particularly on **CoQA**, where LSC surpasses EigenScore by a significant margin (e.g., 77.5% vs. 71.2% in AUC_s). This suggests that as LLMs become more capable and their outputs more consistent, as an internal probability-based metric, our proposed LSC becomes increasingly critical for distinguishing factual confidence from mere repetition.

B More Baseline Introduction

We compare LSC with a diverse set of established hallucination detection methods. The implementation details of these baselines are introduced as follows:

Uncertainty-based Metrics. These methods derive hallucination scores directly from the model's output distribution without requiring external retrieval or multiple generations.

- **Perplexity** (Ren et al., 2022): A standard metric for evaluating the uncertainty of language models. It is calculated as the exponential of the negative average log-likelihood of the generated sequence. Lower perplexity indicates higher model confidence and generally correlates with higher factual correctness.

- **Energy** (Liu et al., 2020): Originally proposed for out-of-distribution detection, this metric calculates the energy score based on the log-sum-exp of the logits at each token step. We average the energy scores over the generated sequence. Higher energy values typically suggest that the input/output pattern is out of the model's knowledge distribution, indicating potential hallucinations.

- **LN-Entropy** (Malinin and Gales, 2020): Length-normalized Entropy measures the average information entropy of the predictive distribution across the generated sequence. Unlike raw entropy, it normalizes the cumulative entropy by the sequence length to prevent bias towards shorter sentences, serving as a robust indicator of the model's predictive uncertainty.

Consistency-based Metrics. These methods rely on the premise that LLMs are likely to generate diverse answers when hallucinating, either through stochastic sampling or input perturbations. For all stochastic sampling baselines (Lin et al., 2022) (Chen et al., 2024) , we sample $K = 5$ responses for each query. For AGSER, we follow the official implementation settings.

- **Lexical Similarity** (Lin et al., 2022): This metric measures the surface-level consistency among the stochastically sampled responses. Following the implementation in (Lin et al., 2022), we calculate the average pair-wise ROUGE-L F1 score between all generated samples. A lower lexical overlap suggests high divergence and a higher likelihood of hallucination.

- **EigenScore** (Chen et al., 2024): A state-of-the-art method that evaluates consistency in the semantic space. It computes the covariance matrix of the sentence embeddings of the sampled responses and applies Singular Value Decomposition (SVD). The metric utilizes the logarithm of the singular values to measure the effective dimensionality (divergence) of the semantic space. Higher dimensionality indicates inconsistent meanings and thus hallucination.

- **AGSER** (Liu et al., 2025): Attention-Guided Self-Reflection is a zero-shot approach that utilizes the model's internal attention mechanisms. It identifies "attentive" and "non-attentive" parts of the input query to construct counterfactual inputs. The hallucination score is derived by measuring the consistency difference between responses generated from these processed queries and the original response.

C More Hallucination Detection Cases

We provide additional case studies to qualitatively compare LSC with baseline methods, illustrating distinct failure modes of existing metrics. A critical challenge in hallucination detection is *mode collapse*, where the LLM consistently generates the same non-factual answer. As shown in the first case (regarding the location of a bread plate), the model

Table 3: Supplementary hallucination detection results for Qwen-0.5B and Qwen-32B. **Bold** indicates best performance; underlined indicates second best.

Models	Datasets	NQ			TriviaQA			SQuAD			CoQA		
	Methods	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC	AUC_s	AUC_r	PCC
Qwen-0.5B	Perplexity	73.0	74.6	19.5	78.3	79.1	28.7	53.8	56.9	15.1	48.9	54.9	4.2
	Energy	62.6	60.9	7.3	69.0	68.5	16.6	47.8	49.3	-2.8	43.1	47.2	-8.2
	LN-Entropy	70.3	73.0	9.1	73.1	74.2	19.3	58.9	60.9	15.5	56.0	58.6	9.7
	Lexical Similarity	68.2	74.7	14.8	75.3	76.7	31.4	67.0	69.8	31.7	67.2	69.0	31.7
	EigenScore	**74.9**	75.7	30.4	79.3	79.4	45.5	69.8	71.2	36.6	**72.0**	**71.8**	**38.4**
	AGSER	63.9	61.9	15.8	55.6	56.5	7.5	63.1	64.7	22.0	67.0	66.1	25.7
	LSC	73.2	**76.0**	**34.0**	**82.1**	**82.7**	**53.2**	**72.1**	**74.2**	**41.1**	66.4	69.1	32.8
Qwen-32B	Perplexity	72.0	72.6	24.5	82.4	82.5	51.4	34.5	44.3	-6.3	66.6	68.3	19.9
	Energy	59.0	60.0	6.6	66.2	67.4	28.0	27.0	34.2	-17.0	26.5	36.7	-20.3
	LN-Entropy	70.2	69.8	25.4	80.3	80.3	46.3	44.8	49.7	2.2	75.3	73.0	30.1
	Lexical Similarity	67.3	69.7	26.2	74.4	75.4	51.3	47.1	50.4	4.7	73.0	71.9	34.0
	EigenScore	68.1	69.3	31.9	75.1	75.3	53.5	**49.1**	51.2	7.6	71.2	70.7	41.1
	AGSER	67.3	68.8	26.8	69.9	71.0	37.5	42.5	49.2	2.7	74.6	71.2	30.1
	LSC	**75.2**	**75.8**	**38.7**	**83.6**	**83.7**	**60.3**	48.3	**51.7**	**8.1**	**77.5**	**75.0**	**41.2**

repeatedly generates the incorrect answer "to the left," deceiving consistency-based metrics like Lexical Similarity and EigenScore into assigning high factual scores (False Negatives). In contrast, **LSC** successfully flags this hallucination (marked with ✓) by detecting the underlying low confidence of the generated span, regardless of its repetition. The subsequent cases further demonstrate LSC's precision in correctly validating factual responses where other baselines produce false alarms.

Question: where is the bread plate located in a formal setting
GTAns: at each place
LLMAns: to the left of the dinner plate
Perplexity: 0.134 ✗
Energy: -22.274 ✗
LN-Entropy: 0.029 ✗
Lexical Similarity: 0.943 ✗
EigenScore: -1.778 ✗
AGSER: 0.623 ✓
LSC: 0.770 ✓
BatchGenerations: ['to the left of the dinner plate', 'to the right of the dinner plate', 'to the left of the dinner plate', 'to the left of the dinner plate', 'to the left of the dinner plate']

Question: What industry was centered in Bedfordshire?
GTAns: lace
LLMAns: The British lace industry.
Perplexity: 0.132 ✗
Energy: -24.715 ✓
LN-Entropy: 0.065 ✗
Lexical Similarity: 0.743 ✗
EigenScore: -1.765 ✗
AGSER: 0.375 ✗
LSC: 0.801 ✓
BatchGenerations: ['The British lace industry.', 'One of the main centers of the British lace industry.', 'One of the main centers of the British lace industry.', 'The British lace industry.', 'One of the main centers of the British lace industry.']

Question: Has he taken legal action over his dismissal?
GTAns: he plans to
LLMAns: Yes.
Perplexity: 0.365 ✗
Energy: -17.177 ✗
LN-Entropy: 0.267 ✗
Lexical Similarity: 0.650 ✗
EigenScore: -1.693 ✗
AGSER: 1.000 ✓
LSC: 0.698 ✓
BatchGenerations: ['Yes.', 'Yes.', 'Yes.', 'Yes, Bao has hired a lawyer and plans to file a lawsuit against the county.', 'Yes.']

Question: If two thirds of the Rhine flows through the Maas, where doe the other one third flow through?
GTAns: Pannerdens Kanaal
LLMAns: To answer this question,
Perplexity: 0.322 ✓
Energy: -27.622 ✗
LN-Entropy: 0.247 ✓
Lexical Similarity: 0.600 ✓
EigenScore: -1.708 ✓
AGSER: 0.000 ✓
LSC: 0.642 ✓
BatchGenerations: ['If two-thirds of the Rhine flows through the Maas,', 'To answer this question,', 'To answer this question,', 'To answer this question,', 'To answer this question,']

Question: What is the most common form of oxygen on the planet?
GTAns: dioxygen
LLMAns: O2.
Perplexity: 0.638 ✓
Energy: -19.346 ✓
LN-Entropy: 0.098 ✓
Lexical Similarity: 0.400 ✓
EigenScore: -1.670 ✓
AGSER: 0.000 ✓
LSC: 0.573 ✓
BatchGenerations: ['The most common form of oxygen on the planet is dioxygen,', 'The most common form of oxygen on the planet is dioxygen,', 'O2.', 'O2.', 'O2.']

Question: What US TV sitcom that ran for 7 years to 19 March 1977 was the first to have an independent career woman, Mary Richards, as the central character?
GTAns: The Mary Tyler Moore Show
LLMAns: Mary Tyler Moore Show
Perplexity: 0.140 ✓
Energy: -20.661 ✓
LN-Entropy: 0.034 ✗
Lexical Similarity: 0.956 ✓
EigenScore: -1.910 ✓
AGSER: 0.889 ✓
LSC: 0.863 ✓
BatchGenerations: ['The Mary Tyler Moore Show', 'Mary Tyler Moore Show', 'Mary Tyler Moore Show', 'Mary Tyler Moore Show', 'Mary Tyler Moore Show']

Question: Who wrote the play The Three Sisters?
GTAns: Anton Chekhov
LLMAns: William Shakespeare
Perplexity: 0.587 ✗
Energy: -16.962 ✓
LN-Entropy: 0.106 ✗
Lexical Similarity: 0.630 ✗
EigenScore: -0.911 ✗
AGSER: 0.400 ✗
LSC: 0.591 ✗
BatchGenerations: ['William Shakespeare wrote The Three Sisters.', 'William Shakespeare wrote the play.', 'William Shakespeare', 'William Shakespeare wrote the play The Three Sisters.', 'William Shakespeare']

Question: Pattern recognition receptors recognize components present in broad groups of what?
GTAns: microorganisms
LLMAns: microorganisms.
Perplexity: 0.629 ✗
Energy: -22.171 ✗
LN-Entropy: 0.274 ✗
Lexical Similarity: 0.400 ✗
EigenScore: -1.805 ✗
AGSER: 1.000 ✓
LSC: 0.592 ✗
BatchGenerations: ['microbes.', 'microbes.', 'Microorganisms.', 'microorganisms.', 'microbes.']