Novelty Assessment Report

Paper: Semantic Uncertainty Quantification of Hallucinations in LLMs: A Quantum Tensor Network Based Method
PDF URL: https://openreview.net/pdf?id=11kPIEkj75
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Large language models (LLMs) exhibit strong generative capabilities but remain vulnerable to confabulations, fluent yet unreliable outputs that vary arbitrarily even under identical prompts. Leveraging a quantum tensor network–based pipeline, we propose a quantum physics-inspired uncertainty quantification framework that accounts for the aleatoric uncertainty in token sequence probability for semantic equivalence-based clustering of LLM generations. In turn, this offers a principled and interpretable scheme for hallucination detection. We further introduce an entropy-maximization strategy that prioritizes high-certainty, semantically coherent outputs and highlights entropy regions where LLM decisions are likely to be unreliable, offering practical guidelines for when human oversight is warranted. We evaluate the robustness of our scheme under different generation lengths and quantization levels, dimensions overlooked in prior studies, demonstrating that our approach remains reliable even in resource-constrained deployments. A total of 116 experiments on TriviaQA, NQ, SVAMP, and SQuAD across multiple architectures (Mistral-7B, Mistral-7B-instruct, Falcon-rw-1b, LLaMA-3.2-1b, LLaMA-2-13b-chat, LLaMA-2-7b-chat, LLaMA-2-13b and LLaMA-2-7b) show consistent improvements in AUROC and AURAC over state-of-the-art baselines.

Core Task Landscape

This paper addresses: **Uncertainty Quantification for Hallucination Detection in Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Uncertainty Quantification Methodologies and Frameworks**
- **Hallucination Detection and Characterization**
- **Application-Specific Detection and Evaluation**
- **Mitigation and Calibration Strategies**
- **Evaluation Frameworks and Benchmarking**
- **Uncertainty in Conditional Generation and Predictive Tasks**
- **Robustness and Reliability Analysis**

Complete Taxonomy Tree

- Uncertainty Quantification for Hallucination Detection in Large Language Models Survey Taxonomy
- Uncertainty Quantification Methodologies and Frameworks
  - Comprehensive Survey and Taxonomy Studies (5 papers)
  - [1] Uncertainty quantification and confidence calibration in large language models: A survey (Liu Xiao-ou, 2025) View paper
  - [2] A survey of confidence estimation and calibration in large language models (Cai, 2024) View paper
  - [4] A survey of uncertainty estimation methods on large language models (Xia Zhi-qiu, 2025) View paper
  - [5] A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions (Ola Shorinwa, 2025) View paper
  - [17] Uncertainty Quantification for Hallucination Detection in Large Language Models: Foundations, Methodology, and Future Directions (Kang, 2025) View paper
  - Semantic-Based Uncertainty Estimation ★ (5 papers)
  - [0] Semantic Uncertainty Quantification of Hallucinations in LLMs: A Quantum Tensor Network Based Method (Anon et al., 2026) View paper
  - [13] Detecting hallucinations in large language models using semantic entropy (Sebastian Farquhar, 2024) View paper
  - [15] Semantic entropy probes: Robust and cheap hallucination detection in llms (Kossen, 2024) View paper
  - [26] Semantic energy: Detecting llm hallucination beyond entropy (Ma Huan, 2025) View paper
  - [44] Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space (Risto Miikkulainen, 2024) View paper
  - Token-Level and Probability-Based Methods (4 papers)
  - [27] Can LLMs Detect Their Confabulations? Estimating Reliability in Uncertainty-Aware Language Models (Zhou, 2025) View paper
  - [32] Geometric Uncertainty for Detecting and Correcting Hallucinations in LLMs (Phillips Edward, 2025) View paper
  - [36] Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs (Artem Vazhentsev, 2025) View paper
  - [49] Fact-checking the output of large language models via token-level uncertainty quantification (Baldwin, 2024) View paper
  - Black-Box and Ensemble-Based Approaches (3 papers)
  - [3] Uncertainty-aware fusion: An ensemble framework for mitigating hallucinations in large language models (Dey Prasenjit, 2025) View paper

- [11] Uncertainty Quantification for Language Models: A Suite of Black-Box, White-Box, LLM Judge, and Ensemble Scorers (Dylan Bouchard, 2025) View paper
- [29] Quantifying uncertainty in answers from any language model and enhancing their trustworthiness (Jiuhai Chen, 2024) View paper
- Specialized Uncertainty Quantification Techniques (4 papers)
- [12] Luq: Long-text uncertainty quantification for llms (Basaldella, 2024) View paper
- [22] LLM Uncertainty Quantification through Directional Entailment Graph and Claim Level Response Augmentation (Da, 2024) View paper
- [35] To Believe or Not to Believe Your LLM (Kuzborskij, 2024) View paper
- [46] Clue: Concept-level uncertainty estimation for large language models (Wang, 2024) View paper

- Hallucination Detection and Characterization
  - Theoretical Foundations and Taxonomy (2 papers)
  - [6] Theoretical Foundations and Mitigation of Hallucination in Large Language Models (Gumaan, 2025) View paper
  - [21] A Survey on Hallucination in Large Language and Foundation Models (Pegah Ahadian, 2025) View paper
  - Reference-Free Detection Methods (2 papers)
  - [9] Enhancing uncertainty-based hallucination detection with stronger focus (Deng Cheng, 2023) View paper
  - [10] Reference-free Hallucination Detection for Large Vision-Language Models (Geng, 2024) View paper
  - Reasoning and Chain-of-Thought Hallucinations (1 papers)
  - [14] Auditing Meta-Cognitive Hallucinations in Reasoning Large Language Models (Liu Yilian, 2025) View paper
  - Span-Level and Fine-Grained Detection (2 papers)
  - [47] Fact-level confidence calibration and self-correction (Yuan, 2024) View paper
  - [50] When Models Lie, We Learn: Multilingual Span-Level Hallucination Detection with PsiloQA (Elisei Rykov, 2025) View paper
- Application-Specific Detection and Evaluation
  - Vision-Language Model Hallucination Detection (2 papers)
  - [23] Visual Perception Uncertainty Learning for Hallucination Detection in Large Vision-Language Models (Runze Zhao, 2025) View paper
  - [38] Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation (Zhang, 2024) View paper
  - Domain-Specific Applications (4 papers)
  - [25] Natural Language Inference as a Judge: Detecting Factuality and Causality Issues in Language Model Self-Reasoning for Financial Analysis (Yilin Wu, 2025) View paper
  - [40] Hallucinations vs. Predictions: Reframing Uncertainty in LLM-Generated Medical Responses (S Afroogh, 2025) View paper
  - [42] Uncertainty estimation of large language models in medical question answering (Wu Jiaxin, 2024) View paper
  - [45] LBAP: Improved Uncertainty Alignment of LLM Planners using Bayesian Inference (Manocha, 2024) View paper
  - Practical Deployment and Real-World Evaluation (3 papers)
  - [8] Look before you leap: An exploratory study of uncertainty measurement for large language models (Huang Yuheng, 2023) View paper
  - [30] Reconsidering LLM Uncertainty Estimation Methods in the Wild (Avestimehr, 2025) View paper
  - [34] Look Before You Leap: An Exploratory Study of Uncertainty Analysis for Large Language Models (Yuheng Huang, 2025) View paper
- Mitigation and Calibration Strategies
  - Confidence Calibration Techniques (2 papers)
  - [20] Comparing Uncertainty Measurement and Mitigation Methods for Large Language Models: A Systematic Review (Toyoda Kentaroh, 2025) View paper
  - [37] Confidence is key: Uncertainty estimation in large language models and vision language models (Groot, 2024) View paper
  - Hallucination Mitigation and Self-Correction (2 papers)
  - [39] Enhancing Multi-Agent Consensus Through Third-Party LLM Integration: Analyzing Uncertainty and Mitigating Hallucinations in Large Language Models (Zhihua Duan, 2024) View paper
  - [48] Enhancing reliability in large language models: Self-detection of hallucinations with spontaneous self-checks (Steven Behore, 2024) View paper
  - Adaptive Risk Management and Abstention (1 papers)
  - [19] Learning Conformal Abstention Policies for Adaptive Risk Management in Large Language and Vision-Language Models (Tayebati, 2025) View paper
- Evaluation Frameworks and Benchmarking
  - Software Tools and Packages (1 papers)
  - [7] UQLM: A Python Package for Uncertainty Quantification in Large Language Models (Ho-Kyeong, 2025) View paper
  - Evaluation Methodology and Pitfalls (2 papers)
  - [16] Robust Uncertainty Quantification for Factual Generation of Large Language Models (Yuhao Zhang, 2025) View paper
  - [28] Addressing Pitfalls in the Evaluation of Uncertainty Estimation Methods for Natural Language Generation (Ielanskyi, 2025) View paper
  - Shared Tasks and Competitive Benchmarks (1 papers)
  - [33] Halu-NLP at SemEval-2024 Task 6: MetaCheckGPT - A Multi-task Hallucination Detection using LLM uncertainty and meta-models (Hoblitzell Andrew, 2024) View paper
  - Hybrid and Integrated Detection Pipelines (1 papers)
  - [41] Integrating Token-Level Uncertainty, Bidirectional NLI, and Semantic Entropy for Robust Hallucination Detection in Large Language Models (Shalini Raghuvanshi, 2025) View paper
- Uncertainty in Conditional Generation and Predictive Tasks
  - Image Captioning and Data-to-Text Generation (1 papers)
  - [43] On Hallucination and Predictive Uncertainty in Conditional Language Generation (Yijun Xiao, 2021) View paper
  - In-Context Learning Uncertainty (1 papers)
  - [24] Uncertainty Quantification for In-Context Learning of Large Language Models (Bai, 2024) View paper
- Robustness and Reliability Analysis
  - Error Detection and Label Quality (1 papers)
  - [18] Are llms better than reported? detecting label errors and mitigating their effect on model performance (Omer Nahum, 2025) View paper

- Uncertainty in Evolving Knowledge Domains (1 papers)
- [31] Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI (G Chrysos, 2025) View paper

Narrative

Core task: uncertainty quantification for hallucination detection in large language models. The field has organized itself around several complementary perspectives. At the broadest level, researchers distinguish between foundational uncertainty quantification methodologies—ranging from token-level confidence scores to semantic-based measures—and the specific problem of hallucination detection and characterization, which seeks to identify when models generate content unsupported by their training data or input context. Application-specific branches address domains such as medical question answering or visual perception, while mitigation and calibration strategies explore how to reduce or correct unreliable outputs. Evaluation frameworks and benchmarking efforts provide standardized testbeds, and robustness analyses examine model behavior under distribution shift or adversarial conditions. Representative works like Semantic Entropy Detection[13] and Semantic Entropy Probes[15] illustrate how semantic clustering of model outputs can reveal uncertainty, while surveys such as Uncertainty Calibration Survey[1] and Confidence Calibration Survey[2] synthesize broader methodological trends.

Within the semantic-based uncertainty estimation cluster, a central theme is moving beyond simple token probabilities to capture meaning-level variability. Semantic Entropy Detection[13] pioneered clustering semantically equivalent generations to estimate epistemic uncertainty, and Semantic Entropy Probes[15] extended this by training lightweight classifiers on hidden states. Quantum Tensor Uncertainty[0] contributes to this line by proposing tensor-based representations that encode richer structural information about semantic uncertainty, positioning itself alongside Semantic Energy[26] and Semantic Density[44], which also explore geometric or energy-based formulations. These approaches contrast with token-level methods like Token-Level NLI Entropy[41] or fact-level calibration schemes such as Fact-level Calibration[47], highlighting an ongoing trade-off between computational efficiency and the granularity of uncertainty estimates. The original work's emphasis on quantum-inspired tensor decompositions offers a novel mathematical lens within this active subfield, complementing the probabilistic clustering strategies of its immediate neighbors.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Detecting hallucinations in large language models using semantic entropy

Authors: Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, Yarin Gal | **Year/Venue**: 2024 • Nature | **URL**: View paper

Abstract

Large language model (LLM) systems, such as ChatGPT1 or Gemini2, can show impressive reasoning and question-answering capabilities but often â□□hallucinateâ□□ false outputs and unsubstantiated answers3,4. Answering unreliably or without the necessary information prevents adoption in diverse fields, with problems including fabrication of legal precedents5 or untrue facts in news articles6 and even posing a risk to human life in medical domains such as radiology7. Encouraging truthfulness through ...

Relationship Analysis

Both papers belong to the Semantic-Based Uncertainty Estimation category, focusing on quantifying uncertainty through semantic equivalence analysis of LLM generations to detect hallucinations. They overlap in using semantic clustering (bidirectional entailment via DeBERTa) and entropy-based measures to assess confabulation risk in question-answering tasks. The original paper differs by introducing a quantum tensor network framework for uncertainty quantification of token sequence probabilities, incorporating perturbation theory and entropy maximization to calibrate probabilities, whereas the candidate paper uses standard semantic entropy computed directly over meaning clusters without additional UQ layers or probability adjustment mechanisms.

---

2. Semantic entropy probes: Robust and cheap hallucination detection in llms

Authors: Kossen, Jannik, Jannik Kossen, Han Jiatong, Jiatong Han, et al. (16 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We propose semantic entropy probes (SEPs), a cheap and reliable method for uncertainty quantification in Large Language Models (LLMs). Hallucinations, which are plausible-sounding but factually incorrect and arbitrary model generations, present a major challenge to the practical adoption of LLMs. Recent work by Farquhar et al. (2024) proposes semantic entropy (SE), which can detect hallucinations by estimating uncertainty in the space semantic meaning for a set of model generations. However, the...

Relationship Analysis

Both papers belong to the semantic-based uncertainty estimation category, using semantic equivalence across model generations to quantify uncertainty for hallucination detection. They overlap in employing semantic entropy computed from clustered LLM outputs based on bidirectional entailment (via DeBERTa) to detect confabulations. The key difference is that the original paper introduces a quantum tensor network framework to quantify uncertainty in token sequence probabilities and calibrate them via entropy maximization, while the candidate paper proposes semantic entropy probes (SEPs) that train linear classifiers on LLM hidden states to predict semantic entropy from a single generation, eliminating the need for multiple samples at test time.

---

3. Semantic energy: Detecting llm hallucination beyond entropy

Authors: Ma Huan, Pan Jiadong, Huan Ma, Liu Jing, Jiadong Pan, et al. (21 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Models (LLMs) are being increasingly deployed in real-world applications, but they remain susceptible to hallucinations, which produce fluent yet incorrect responses and lead to erroneous decision-making. Uncertainty estimation is a feasible approach to detect such hallucinations. For example, semantic entropy estimates uncertainty by considering the semantic diversity across multiple sampled responses, thus identifying hallucinations. However, semantic entropy relies on post-soft...

Relationship Analysis

Both papers belong to the Semantic-Based Uncertainty Estimation category, focusing on quantifying uncertainty through semantic analysis of LLM generations to detect hallucinations. They share the approach of clustering semantically equivalent outputs and computing uncertainty measures beyond naive token-level entropy. The key difference is that the original paper employs quantum tensor network perturbation theory to quantify aleatoric uncertainty in token sequence probabilities and uses entropy maximization for calibration, while the candidate paper introduces an energy-based formulation using logits (rather than probabilities) to capture epistemic uncertainty that semantic entropy fails to detect, particularly when models produce identical but incorrect responses.

---

4. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space

Authors: Risto Miikkulainen, Xin QIU | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ With the widespread application of Large Language Models (â¦ reliable uncertainty measurement in detecting hallucinations â¦ simple baselines on uncertainty measurement for LLMs, â¦

Relationship Analysis

Both papers belong to the Semantic-Based Uncertainty Estimation category, focusing on quantifying uncertainty through semantic analysis of LLM generations rather than purely lexical measures. They overlap in using semantic clustering and entropy-based metrics to detect hallucinations, both addressing the limitation that lexically different outputs may be semantically equivalent. The key difference is that the original paper employs quantum tensor network perturbation theory to quantify uncertainty in token sequence probabilities and uses entropy maximization for calibration, while the candidate paper proposes semantic density as a response-wise confidence metric based on kernel density estimation in semantic space weighted by generation probabilities.

Contributions Analysis

Overall novelty summary. The paper proposes a quantum tensor network-based framework for uncertainty quantification in LLM hallucination detection, focusing on semantic equivalence-based clustering of token sequence probabilities. It resides in the 'Semantic-Based Uncertainty Estimation' leaf, which contains five papers including the original work. This leaf sits within the broader 'Uncertainty Quantification Methodologies and Frameworks' branch, one of seven major branches in a taxonomy covering fifty papers. The semantic-based cluster represents a moderately populated research direction, with sibling works like Semantic Entropy Detection and Semantic Entropy Probes establishing the core paradigm of clustering semantically equivalent generations to estimate epistemic uncertainty.

The taxonomy reveals neighboring leaves addressing token-level probability methods, black-box ensemble approaches, and specialized techniques for long-text or concept-level estimation. The original work's quantum tensor formulation diverges from the probabilistic clustering strategies dominant in its immediate leaf, instead offering a geometric or structural perspective akin to Semantic Energy and Semantic Density in related branches. The scope note for this leaf explicitly excludes token-level methods, positioning the work within a meaning-space analysis paradigm. Nearby branches on hallucination detection and mitigation strategies suggest the field balances foundational uncertainty estimation with practical deployment concerns.

Across three contributions, the analysis examined eighteen candidate papers with no clear refutations identified. The quantum tensor network framework examined three candidates with zero refutable overlaps, suggesting limited prior work on tensor-based semantic uncertainty in the search scope. The entropy maximization strategy examined ten candidates with no refutations, indicating potential novelty in calibration approaches within the semantic clustering paradigm. Robustness evaluation across quantization and generation lengths examined five candidates with no refutations, highlighting that these dimensions may be underexplored in prior semantic-based methods. The limited search scope means these findings reflect top-eighteen semantic matches rather than exhaustive coverage.

Given the moderate density of the semantic-based uncertainty leaf and the absence of refutations among eighteen examined candidates, the work appears to introduce a distinct mathematical formalism within an established research direction. The quantum tensor approach and robustness dimensions may offer incremental advances over probabilistic clustering baselines, though the limited search scope precludes definitive claims about field-wide novelty. The analysis captures top semantic neighbors but does not cover the full fifty-paper taxonomy or broader literature on quantum-inspired machine learning methods.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Quantum tensor network-based uncertainty quantification framework for token sequence probabilities

Description: The authors introduce a novel framework that leverages quantum tensor networks and perturbation theory to quantify uncertainty in token sequence probabilities. This physics-inspired approach provides a deterministic, one-shot method for assessing local sensitivity of sequence probabilities to model perturbations, addressing a gap in prior hallucination detection methods.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Neurosymbolic Visual Transform Based on Logic Tensor Network for Defect Detection

URL: View paper

Brief Assessment

Logic Tensor Network[53] focuses on neurosymbolic visual transforms for defect detection in images, treating image patches as tokens. This is fundamentally different from the original paper's quantum tensor network approach for uncertainty quantification in language model token sequence probabilities for hallucination detection.

---

2. Uncertainty Quantification of Large Language Models through Multi-Dimensional Responses

URL: View paper

Brief Assessment

Multi-Dimensional Responses[52] focuses on multi-dimensional uncertainty quantification through semantic and knowledge similarity matrices using tensor decomposition, not quantum tensor networks or perturbation theory for token sequence probabilities.

---

3. Sequential uncertainty quantification with contextual tensors for social targeting

URL: View paper

Brief Assessment

Sequential Contextual Tensors[51] focuses on social network targeting with probabilistic tensor regression for user-product heterogeneity, not on uncertainty quantification for token sequence probabilities in language models. The domains and technical approaches are fundamentally different.

Contribution 2: Entropy maximization strategy for calibrating token sequence probabilities

Description: The authors propose a principled method that adjusts token sequence probabilities by maximizing Rényi entropy while penalizing deviations weighted by uncertainty. This enables selection of more reliable outputs and identifies regions requiring human oversight, going beyond simple entropy thresholding used in prior work.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. On the Entropy Calibration of Language Models

URL: View paper

Brief Assessment

Entropy Calibration[61] focuses on calibrating a model's total entropy over entire generations to match log loss on human text, addressing whether generation entropy matches expected loss. The original paper addresses calibrating token sequence probabilities by

maximizing Rényi entropy while penalizing deviations weighted by uncertainty for hallucination detection. These are distinct calibration objectives with different technical approaches and applications.

---

2. Regularizing Neural Networks by Penalizing Confident Output Distributions

URL: View paper

Brief Assessment

Penalizing Confident Outputs[59] focuses on penalizing low entropy output distributions as a regularization technique in supervised learning for neural networks. The original paper proposes a quantum tensor network-based uncertainty quantification framework that adjusts token sequence probabilities by maximizing Rényi entropy while penalizing deviations weighted by uncertainty for hallucination detection in LLMs. These are fundamentally different applications and methodological approaches.

---

3. Semantic uncertainty in advanced decoding methods for LLM generation

URL: View paper

Brief Assessment

Advanced Decoding Uncertainty[65] focuses on semantic uncertainty across different decoding methods (speculative sampling, chain-of-thought) for LLM generation tasks, not on calibrating token sequence probabilities through entropy maximization with uncertainty-weighted penalties as proposed in the original paper.

---

4. Entropy-based adaptive weighting for self-training

URL: View paper

Brief Assessment

Entropy Adaptive Weighting[60] focuses on entropy-based weighting for self-training in mathematical reasoning tasks, not on calibrating token sequence probabilities for hallucination detection in LLMs. The candidate uses entropy to measure model uncertainty across multiple generated answers to prioritize training data, while the original paper uses entropy maximization to adjust token probabilities for semantic clustering and hallucination detection.

---

5. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space

URL: View paper

Brief Assessment

Semantic Density[44] focuses on kernel density estimation in semantic space for response-wise confidence, not on entropy maximization with uncertainty-weighted penalties for calibrating token sequence probabilities as proposed in the original paper.

---

6. Distinguishing the knowable from the unknowable with language models

URL: View paper

Brief Assessment

Knowable from Unknowable[62] focuses on distinguishing epistemic from aleatoric uncertainty using supervised probes and in-context learning tests, not on entropy maximization for calibrating token probabilities weighted by uncertainty estimates.

---

7. Revisiting Entropy in Reinforcement Learning for Large Reasoning Models

URL: View paper

Brief Assessment

Entropy Reasoning Models[66] focuses on entropy collapse during RLVR training for reasoning models, not on calibrating token sequence probabilities for hallucination detection through entropy maximization with uncertainty weighting.

---

8. Test-Time Distillation for Continual Model Adaptation

URL: View paper

Brief Assessment

Test-Time Distillation[63] addresses entropy bias in model fusion for continual test-time adaptation, not token sequence probability calibration in LLMs. The candidate focuses on vision models and distribution shifts, while the original work targets hallucination detection in language models through quantum-inspired uncertainty quantification.

---

9. CATfOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration

URL: View paper

Brief Assessment

CATfOOD[64] focuses on calibrating model predictions for out-of-domain QA tasks using counterfactual data augmentation, not on entropy maximization for calibrating token sequence probabilities in the context of hallucination detection.

---

10. Enhancing In-context Learning via Linear Probe Calibration

URL: View paper

Brief Assessment

Linear Probe Calibration[67] focuses on calibrating output probabilities in in-context learning using linear probes and Shannon entropy for reliability assessment, not on entropy maximization with Rényi entropy and uncertainty-weighted adjustments for hallucination detection.

Contribution 3: Robustness evaluation across quantization levels and generation lengths

Description: The authors systematically assess their hallucination detection framework across multiple quantization settings (16-bit, 8-bit, 4-bit) and varying generation lengths. This evaluation addresses practical deployment scenarios that prior hallucination detection studies have not examined, demonstrating the method's applicability to real-world resource-constrained environments.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Team Cantharellus at SemEval-2025 task 3: Hallucination span detection with fine tuning on weakly supervised synthetic data

URL: View paper

Brief Assessment

Weakly Supervised Detection[57] focuses on hallucination span detection in multilingual text using fine-tuned models on synthetic data. It does not evaluate robustness across quantization levels or varying generation lengths, which are the core dimensions of the original contribution.

---

2. Automated Topic Page Generation Using Multi-Agent LLMs

URL: View paper

Brief Assessment

Multi-Agent Topic Generation[58] focuses on automated scientific topic page generation using multi-agent LLMs, not hallucination detection frameworks. The candidate does not address quantization-level robustness or generation length variations in hallucination detection contexts.

---

3. An Empirical Study on Prompt Compression for Large Language Models

URL: View paper

Brief Assessment

Prompt Compression Study[54] focuses on prompt compression methods for LLMs across different tasks (summarization, QA, VQA), not on hallucination detection robustness across quantization levels and generation lengths. The technical focus and evaluation dimensions are fundamentally different.

---

4. Large language models with adaptive token fusion: A novel approach to reducing hallucinations and improving inference efficiency

URL: View paper

Brief Assessment

Adaptive Token Fusion[55] focuses on efficient model compression and reducing hallucinations through token fusion mechanisms, not on systematic evaluation of hallucination detection frameworks across quantization settings and generation lengths as the original paper does.

---

5. Stochastic lexical dissonance injection for self-consistent reasoning in large language models: A quantitative investigation

URL: View paper

Brief Assessment

Lexical Dissonance Injection[56] does not address quantization levels or generation length variations in hallucination detection. The candidate focuses on lexical perturbation methods for self-consistency, which is a different technical approach from the systematic robustness evaluation across deployment constraints presented in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 21 papers and found 4 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Semantic entropy probes: Robust and cheap hallucination detection in llms

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. Detecting hallucinations in large language models using semantic entropy

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Semantic Uncertainty Quantification of Hallucinations in LLMs: A Quantum Tensor Network Based Method View paper
- [1] Uncertainty quantification and confidence calibration in large language models: A survey View paper
- [2] A survey of confidence estimation and calibration in large language models View paper
- [3] Uncertainty-aware fusion: An ensemble framework for mitigating hallucinations in large language models View paper
- [4] A survey of uncertainty estimation methods on large language models View paper
- [5] A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions View paper
- [6] Theoretical Foundations and Mitigation of Hallucination in Large Language Models View paper
- [7] UQLM: A Python Package for Uncertainty Quantification in Large Language Models View paper
- [8] Look before you leap: An exploratory study of uncertainty measurement for large language models View paper
- [9] Enhancing uncertainty-based hallucination detection with stronger focus View paper
- [10] Reference-free Hallucination Detection for Large Vision-Language Models View paper
- [11] Uncertainty Quantification for Language Models: A Suite of Black-Box, White-Box, LLM Judge, and Ensemble Scorers View paper
- [12] Luq: Long-text uncertainty quantification for llms View paper
- [13] Detecting hallucinations in large language models using semantic entropy View paper
- [14] Auditing Meta-Cognitive Hallucinations in Reasoning Large Language Models View paper
- [15] Semantic entropy probes: Robust and cheap hallucination detection in llms View paper
- [16] Robust Uncertainty Quantification for Factual Generation of Large Language Models View paper
- [17] Uncertainty Quantification for Hallucination Detection in Large Language Models: Foundations, Methodology, and Future Directions View paper
- [18] Are llms better than reported? detecting label errors and mitigating their effect on model performance View paper
- [19] Learning Conformal Abstention Policies for Adaptive Risk Management in Large Language and Vision-Language Models View paper
- [20] Comparing Uncertainty Measurement and Mitigation Methods for Large Language Models: A Systematic Review View paper

- [21] A Survey on Hallucination in Large Language and Foundation Models View paper
- [22] LLM Uncertainty Quantification through Directional Entailment Graph and Claim Level Response Augmentation View paper
- [23] Visual Perception Uncertainty Learning for Hallucination Detection in Large Vision-Language Models View paper
- [24] Uncertainty Quantification for In-Context Learning of Large Language Models View paper
- [25] Natural Language Inference as a Judge: Detecting Factuality and Causality Issues in Language Model Self-Reasoning for Financial Analysis View paper
- [26] Semantic energy: Detecting llm hallucination beyond entropy View paper
- [27] Can LLMs Detect Their Confabulations? Estimating Reliability in Uncertainty-Aware Language Models View paper
- [28] Addressing Pitfalls in the Evaluation of Uncertainty Estimation Methods for Natural Language Generation View paper
- [29] Quantifying uncertainty in answers from any language model and enhancing their trustworthiness View paper
- [30] Reconsidering LLM Uncertainty Estimation Methods in the Wild View paper
- [31] Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI View paper
- [32] Geometric Uncertainty for Detecting and Correcting Hallucinations in LLMs View paper
- [33] Halu-NLP at SemEval-2024 Task 6: MetaCheckGPT - A Multi-task Hallucination Detection using LLM uncertainty and meta-models View paper
- [34] Look Before You Leap: An Exploratory Study of Uncertainty Analysis for Large Language Models View paper
- [35] To Believe or Not to Believe Your LLM View paper
- [36] Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs View paper
- [37] Confidence is key: Uncertainty estimation in large language models and vision language models View paper
- [38] Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation View paper
- [39] Enhancing Multi-Agent Consensus Through Third-Party LLM Integration: Analyzing Uncertainty and Mitigating Hallucinations in Large Language Models View paper
- [40] Hallucinations vs. Predictions: Reframing Uncertainty in LLM-Generated Medical Responses View paper
- [41] Integrating Token-Level Uncertainty, Bidirectional NLI, and Semantic Entropy for Robust Hallucination Detection in Large Language Models View paper
- [42] Uncertainty estimation of large language models in medical question answering View paper
- [43] On Hallucination and Predictive Uncertainty in Conditional Language Generation View paper
- [44] Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space View paper
- [45] LBAP: Improved Uncertainty Alignment of LLM Planners using Bayesian Inference View paper
- [46] Clue: Concept-level uncertainty estimation for large language models View paper
- [47] Fact-level confidence calibration and self-correction View paper
- [48] Enhancing reliability in large language models: Self-detection of hallucinations with spontaneous self-checks View paper
- [49] Fact-checking the output of large language models via token-level uncertainty quantification View paper
- [50] When Models Lie, We Learn: Multilingual Span-Level Hallucination Detection with PsiloQA View paper
- [51] Sequential uncertainty quantification with contextual tensors for social targeting View paper
- [52] Uncertainty Quantification of Large Language Models through Multi-Dimensional Responses View paper
- [53] Neurosymbolic Visual Transform Based on Logic Tensor Network for Defect Detection View paper
- [54] An Empirical Study on Prompt Compression for Large Language Models View paper
- [55] Large language models with adaptive token fusion: A novel approach to reducing hallucinations and improving inference efficiency View paper
- [56] Stochastic lexical dissonance injection for self-consistent reasoning in large language models: A quantitative investigation View paper
- [57] Team Cantharellus at SemEval-2025 task 3: Hallucination span detection with fine tuning on weakly supervised synthetic data View paper
- [58] Automated Topic Page Generation Using Multi-Agent LLMs View paper
- [59] Regularizing Neural Networks by Penalizing Confident Output Distributions View paper
- [60] Entropy-based adaptive weighting for self-training View paper
- [61] On the Entropy Calibration of Language Models View paper
- [62] Distinguishing the knowable from the unknowable with language models View paper
- [63] Test-Time Distillation for Continual Model Adaptation View paper
- [64] CATfOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration View paper
- [65] Semantic uncertainty in advanced decoding methods for LLM generation View paper
- [66] Revisiting Entropy in Reinforcement Learning for Large Reasoning Models View paper
- [67] Enhancing In-context Learning via Linear Probe Calibration View paper