Novelty Assessment Report

Paper: Revisiting Hallucination Detection Through The Lens Of Effective Rank-based Uncertainty
PDF URL: https://openreview.net/pdf?id=0O6Xj6ljIN
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Detecting hallucinations in large language models (LLMs) remains a fundamental challenge for their trustworthy deployment. Going beyond basic uncertainty-driven hallucination detection frameworks, we propose a simple yet powerful method that quantifies uncertainty by measuring the effective rank of hidden states derived from multiple model outputs and different layers. Grounded in the spectral analysis of representations, our approach provides interpretable insights into the model's internal reasoning process through semantic variations, while requiring no extra knowledge or additional modules, thus offering a combination of theoretical elegance and practical efficiency. Meanwhile, we theoretically demonstrate the necessity of quantifying uncertainty both internally (representations of a single response) and externally (different responses), providing a justification for using representations among different layers and responses from LLMs to detect hallucinations. Extensive experiments demonstrate that our method effectively detects hallucinations and generalizes robustly across various scenarios, contributing to a new paradigm of hallucination detection for LLM truthfulness.

Core Task Landscape

This paper addresses: **Hallucination Detection in Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

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

- **Detection Methods Based on Model Internal States**
- **Detection Methods Based on Output Analysis**
- **Detection Methods Using External Knowledge and Retrieval**
- **Specialized Detection Frameworks and Benchmarks**
- **Testing and Validation Methodologies**
- **Domain-Specific Hallucination Detection**
- **Theoretical Foundations and Feasibility Analysis**
- **Prompt Engineering and Diversion-Based Detection**
- **Multimodal Hallucination Detection**
- **Comprehensive Surveys and Reviews**
- ... and 3 more categories

Complete Taxonomy Tree

- Hallucination Detection in Large Language Models Survey Taxonomy
- Detection Methods Based on Model Internal States
  - Representation-Based Uncertainty Quantification ★ (3 papers)
  - [0] Revisiting Hallucination Detection Through The Lens Of Effective Rank-based Uncertainty (Anon et al., 2026) View paper
  - [7] Unsupervised real-time hallucination detection based on the internal states of large language models (Weihang Su, 2024) View paper
  - [50] INSIDE: LLMs' internal states retain the power of hallucination detection (Chen Chao, 2024) View paper
  - Neural Probe and Layer-Specific Detection (1 papers)
  - [38] Neural Probe-Based Hallucination Detection for Large Language Models (Shize Liang, 2025) View paper
- Detection Methods Based on Output Analysis
  - Sampling-Based Consistency Detection (2 papers)
  - [6] Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models (Gales, 2023) View paper
  - [31] InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers (Barkan, 2024) View paper
  - Probability and Uncertainty Estimation (2 papers)
  - [12] Detecting hallucinations in large language models using semantic entropy (Sebastian Farquhar, 2024) View paper
  - [49] Detecting llm hallucinations using monte carlo simulations on token probabilities (Grant Ledger, 2024) View paper
  - Self-Evaluation and Familiarity Assessment (1 papers)
  - [5] Zero-resource hallucination prevention for large language models (Luo Jun-Yu, 2024) View paper
- Detection Methods Using External Knowledge and Retrieval
  - Retrieval-Augmented Verification (2 papers)
  - [14] Mitigating entity-level hallucination in large language models (Weihang Su, 2024) View paper
  - [42] Automated Hallucination Detection and Mitigation in Large Language Model (S. Srinivasan, 2025) View paper
  - Context-Based and NLI Detection (1 papers)
  - [44] Developing a reliable, fast, general-purpose hallucination detection and mitigation service (Wang, 2025) View paper

- Specialized Detection Frameworks and Benchmarks
  - Multi-Component Detection Systems (1 papers)
  - [3] Hademif: Hallucination detection and mitigation in large language models (X Zhou, 2025) View paper
  - Evaluation Benchmarks and Taxonomies (4 papers)
  - [4] Fine-grained hallucination detection and editing for language models (Asai, 2024) View paper
  - [18] Diahalu: A dialogue-level hallucination evaluation benchmark for large language models (Chen Qin, 2024) View paper
  - [45] Felm: Benchmarking factuality evaluation of large language models (Chen Shi-qi, 2023) View paper
  - [46] Benchmarking Hallucination in Large Language Models Based on Unanswerable Math Word Problem (Sun Yuhong, 2024) View paper
  - Error Detection in LLM Responses (2 papers)
  - [28] Evaluating LLMs at detecting errors in LLM responses (Kamoi, 2024) View paper
  - [33] Hallucination Detection and Evaluation of Large Language Model (Chenggong Zhang, 2025) View paper
- Testing and Validation Methodologies
  - Metamorphic Testing Approaches (2 papers)
  - [16] Drowzee: Metamorphic testing for fact-conflicting hallucination detection in large language models (Ningke Li, 2024) View paper
  - [39] Hallucination Detection in Large Language Models with Metamorphic Relations (YANG Borui, 2025) View paper
  - Automated Testing Frameworks (1 papers)
  - [40] The earth is flat? unveiling factual errors in large language models (Wang, 2024) View paper
- Domain-Specific Hallucination Detection
  - Code Generation Hallucination Detection (1 papers)
  - [32] Exploring and evaluating hallucinations in llm-powered code generation (Liu Fang, 2024) View paper
  - Product and E-Commerce Hallucination Detection (1 papers)
  - [20] Hallucination detection in LLM-enriched product listings (L Jiang, 2024) View paper
- Theoretical Foundations and Feasibility Analysis (1 papers)
  - [43] (Im)possibility of Automated Hallucination Detection in Large Language Models (Karbasi, 2025) View paper
- Prompt Engineering and Diversion-Based Detection (2 papers)
  - [27] GPT Hallucination Detection Through Prompt Engineering (Marco Siino, 2024) View paper
  - [37] Hallucination Detection in Large Language Models Using Diversion Decoding (Basel Abdeen, 2025) View paper
- Multimodal Hallucination Detection (1 papers)
  - [1] Hallucination of multimodal large language models: A survey (Zechen Bai, 2024) View paper
- Comprehensive Surveys and Reviews (5 papers)
  - [2] A comprehensive survey of hallucination mitigation techniques in large language models (Tonmoy Islam, 2024) View paper
  - [10] A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions (Lei Huang, 2025) View paper
  - [17] The dawn after the dark: An empirical study on factuality hallucination in large language models (Junyi Li, 2024) View paper
  - [24] Tutorial Proposal: Hallucination in Large Language Models (Vipula Rawte, 2024) View paper
  - [26] Hallucination detection and hallucination mitigation: An investigation (Luo, 2024) View paper
- Misinformation and Fake News Detection Using LLMs
  - LLM-Based Fake News Detection (3 papers)
  - [19] Bad actor, good advisor: Exploring the role of large language models in fake news detection (Cao Juan, 2024) View paper
  - [34] Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data (Raza, 2025) View paper
  - [35] FakeGPT: fake news generation, explanation and detection of large language models (Huang Yue, 2023) View paper
  - Misinformation Detection in Specialized Domains (4 papers)
  - [8] Fmdllama: Financial misinformation detection based on large language models (Zhiwei Liu, 2025) View paper
  - [29] Detecting Health Misinformation by Leveraging LLM Models and Debunk List (M Rostami, 2025) View paper
  - [30] Analysis of disinformation and fake news detection using fine-tuned large language model (Pavlyshenko, 2023) View paper
  - [47] An Information Reliability Framework for Detecting Misinformation based on Large Language Models (Venkata Sai Prathyush Turaga, 2024) View paper
  - Comment and Feedback Verification (2 papers)
  - [22] Let silence speak: Enhancing fake news detection with generated comments from large language models (Qiong Nan, 2024) View paper
  - [36] Comments and feedback verification system using large language model (Akshay Kumar Kushwaha, 2024) View paper
- LLM-Generated Misinformation and Disinformation
  - Disinformation Generation Capabilities (1 papers)
  - [11] Disinformation capabilities of large language models (Bielikova Maria, 2024) View paper
  - Evolving LLM-Generated Disinformation Detection (1 papers)
  - [15] Catching chameleons: Detecting evolving disinformation generated using large language models (Bohan Jiang, 2024) View paper
  - Detector Bias and Adaptation Challenges (2 papers)
  - [13] Adapting fake news detection to the era of large language models (Jinyan Su, 2024) View paper
  - [21] Fake news detectors are biased against texts generated by large language models (Su, 2023) View paper
  - Dual Roles of LLMs in Fake News Ecosystems (3 papers)
  - [23] Combating misinformation in the age of llms: Opportunities and challenges (Canyu Chen, 2024) View paper
  - [25] Factuality challenges in the era of large language models and opportunities for fact-checking (Isabelle Augenstein, 2024) View paper
  - [41] From deception to detection: The dual roles of large language models in fake news (Dorsaf Sallami, 2024) View paper
- Detection of LLM-Generated Text
  - General LLM Output Detection (1 papers)
  - [9] Llm-check: Investigating detection of hallucinations in large language models (Siddhant Bharti, 2024) View paper
  - Theoretical Limits of LLM Text Detection (1 papers)
  - [48] Limits of detecting text generated by large-scale language models (Lav R. Varshney, 2020) View paper

Narrative

Core task: hallucination detection in large language models. The field has organized itself around several complementary perspectives. Detection Methods Based on Model Internal States exploit hidden representations and uncertainty signals within the model itself, while Detection Methods Based on Output Analysis examine generated text for consistency or semantic coherence without requiring internal access. Detection Methods Using External Knowledge and Retrieval verify claims against trusted sources, and Specialized Detection Frameworks and Benchmarks provide standardized evaluation environments. Additional branches address Testing and Validation Methodologies, Domain-Specific Hallucination Detection (e.g., code generation, product listings), Theoretical Foundations exploring feasibility limits, Prompt Engineering and Diversion-Based Detection that manipulate inputs to reveal inconsistencies, Multimodal Hallucination Detection extending beyond text, and Comprehensive Surveys and Reviews synthesizing progress. Parallel branches on Misinformation and Fake News Detection Using LLMs, LLM-Generated Misinformation and Disinformation, and Detection of LLM-Generated Text reflect concerns about adversarial uses and content provenance.

Within the internal-state branch, representation-based uncertainty quantification has attracted considerable attention, with methods like Semantic Entropy Detection[12] and INSIDE[50] leveraging latent features to estimate confidence. Effective Rank Uncertainty[0] contributes to this cluster by proposing a novel uncertainty measure derived from representation geometry, positioning itself alongside Unsupervised Hallucination Detection[7] which also avoids labeled data. These approaches contrast with output-analysis techniques such as SelfCheckGPT[6] that rely on sampling consistency, and with external-knowledge methods like Hademif[3] that ground outputs in retrieval. A recurring theme across branches is the trade-off between requiring model access versus operating in black-box settings, and between general-purpose detectors and domain-tailored solutions. The original work's focus on effective rank places it squarely in the representation-based uncertainty camp, offering a geometric lens that complements entropy-based and probe-based neighbors while remaining agnostic to specific task domains.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Unsupervised real-time hallucination detection based on the internal states of large language models

Authors: Weihang Su, Changyue Wang, Qingyao Ai, Hu Yiran, Yiran Hu, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ Further, to facilitate future research and to improve the reproducibility of this paper, we introduce a new benchmark for LLM hallucination detection named HELM: Hallucination â¦

Relationship Analysis

Both papers belong to the Representation-Based Uncertainty Quantification category, using internal model states to detect hallucinations. They overlap in leveraging hidden layer representations to measure uncertainty without external knowledge, but differ fundamentally in their approach: the original paper uses effective rank of singular values from multiple layers and responses to quantify semantic dispersion, while the candidate paper (MIND) trains a supervised classifier on hidden states using an unsupervised framework based on entity mention correctness in Wikipedia continuations.

---

2. INSIDE: LLMs' internal states retain the power of hallucination detection

Authors: Chen Chao, Liu Kai, Chao Chen, Chen Ze, Kai Liu, et al. (16 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Knowledge hallucination have raised widespread concerns for the security and reliability of deployed LLMs. Previous efforts in detecting hallucinations have been employed at logit-level uncertainty estimation or language-level self-consistency evaluation, where the semantic information is inevitably lost during the token-decoding procedure. Thus, we propose to explore the dense semantic information retained within LLMs' \textbf{IN}ternal \textbf{S}tates for halluc\textbf{I}nation \textbf{DE}tect...

Relationship Analysis

Both papers belong to the Representation-Based Uncertainty Quantification category, using spectral analysis of hidden state representations to detect hallucinations in LLMs. They overlap in their core approach of analyzing internal model states through eigenvalue decomposition of embedding covariance matrices to measure uncertainty. The key difference is that the original paper (Effective Rank) uses entropy of normalized singular values to compute an effective rank metric across multiple layers and responses, while the candidate paper (INSIDE/Eigenscore) uses the log-determinant of the covariance matrix as a differential entropy approximation, primarily from a single middle layer, and additionally introduces a test-time feature clipping technique to handle overconfident hallucinations.

Contributions Analysis

Overall novelty summary. The paper proposes using effective rank of hidden states across multiple outputs and layers to quantify uncertainty for hallucination detection. It resides in the 'Representation-Based Uncertainty Quantification' leaf, which contains only three papers including this one. This is a relatively sparse research direction within the broader taxonomy of 50 papers across 36 topics, suggesting the specific approach of spectral analysis on representations is not yet heavily explored. The sibling papers in this leaf include semantic entropy methods and unsupervised detection frameworks, indicating a small but active cluster focused on internal-state uncertainty without external resources.

The taxonomy reveals a well-populated neighboring branch on 'Sampling-Based Consistency Detection' and 'Probability and Uncertainty Estimation' under output analysis, which examines generated text rather than internal representations. The 'Neural Probe and Layer-Specific Detection' leaf sits adjacent within the same parent branch, training classifiers on activations rather than computing geometric properties. The scope note for the original leaf explicitly excludes output probability methods, clarifying that effective rank operates on hidden states rather than token distributions. This positioning suggests the work bridges representation geometry with uncertainty quantification, a niche distinct from both probe-based and sampling-based neighbors.

Among 26 candidates examined, the contribution-level analysis reveals mixed novelty signals. The effective rank-based uncertainty contribution examined 6 candidates with 1 refutable match, while the theoretical justification for multi-response uncertainty examined 10 candidates with 3 refutable matches, and the training-free framework examined 10 candidates with 1 refutable match. These statistics indicate that within the limited search scope, some prior work addresses overlapping ideas—particularly around combining internal and external uncertainty signals. However, the majority of examined candidates (21 out of 26 across all contributions) did not clearly refute the claims, suggesting the specific combination of effective rank, multi-layer analysis, and theoretical grounding may offer distinguishing elements despite conceptual overlap with existing representation-based methods.

Based on the top-26 semantic matches and citation expansion, the work appears to occupy a moderately novel position within a sparse but growing research direction. The limited search scope means exhaustive prior art may exist beyond these candidates, particularly in adjacent fields like representation learning or spectral methods in deep learning. The taxonomy context shows the field has many alternative detection paradigms (output consistency, external retrieval, probes), but fewer works specifically applying spectral geometry

to hidden states for uncertainty quantification, lending some distinctiveness to the approach despite partial overlaps identified in the analysis.

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

Contribution 1: Effective Rank-based Uncertainty for Hallucination Detection

Description: The authors introduce a novel uncertainty quantification method that computes the effective rank of embedding matrices constructed from LLM hidden states across multiple responses and layers. This spectral analysis approach provides an interpretable measure of uncertainty corresponding to the effective number of distinct semantic categories, requiring no additional training or external knowledge.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Revisiting Hallucination Detection with Effective Rank-based Uncertainty

URL: View paper

Brief Assessment

Effective Rank Revisited[53] is the same paper as the original submission. The candidate paper title explicitly references this work as '[53]', indicating it is citing itself rather than presenting prior work that would refute novelty claims.

---

2. Learning Probabilistic Box Embeddings for Effective and Efficient Ranking

URL: View paper

Brief Assessment

Probabilistic Box Embeddings[54] focuses on ranking tasks in information retrieval using box embeddings to model diversity and uncertainty in queries/items, not on uncertainty quantification for hallucination detection in language models using effective rank of hidden states.

---

3. On-Device Large Language Models: A Survey of Model Compression and System Optimization

URL: View paper

Brief Assessment

On-Device LLM Survey[52] is a comprehensive survey on model compression and system optimization for edge deployment. It does not present novel uncertainty quantification methods or hallucination detection techniques using effective rank of hidden states.

---

4. UNComp: Uncertainty-Aware Long-Context Compressor for Efficient Large Language Model Inference

URL: View paper

Brief Assessment

UNComp[51] applies effective rank to KV cache compression in LLMs for efficiency optimization, not hallucination detection. The candidate focuses on identifying sparsity patterns for compression during inference, while the original paper uses effective rank to quantify uncertainty across multiple model responses for detecting hallucinations.

---

5. Uncertainty Quantification with Generative-Semantic Entropy Estimation for Large Language Models

URL: View paper

Prior Art Analysis

Generative-Semantic Entropy[56] demonstrates that prior work exists using spectral analysis of embedding matrices for uncertainty quantification in LLMs. Both papers compute uncertainty by analyzing the eigenvalue/singular value spectrum of covariance matrices constructed from hidden state embeddings across multiple model responses. The candidate paper explicitly calculates spectral-entropy of covariance matrices from latent embeddings and interprets this as the 'effective dimension' or 'effective subspace rank' of the semantic manifold - the same conceptual framework as effective rank. The candidate's approach of extracting embeddings from multiple generations, computing their covariance, and using spectral properties to measure uncertainty predates the original paper's claimed novelty.

Evidence

Evidence 1 - **Rationale**: Both papers cite Roy & Vetterli (2007) and use the same mathematical concept of effective rank/effective subspace rank to measure semantic diversity in embeddings. The candidate's spectral-entropy calculation is mathematically equivalent to effective rank computation. - **Original**: we leverage the notion ofeffective rank(roy & vetterli, 2007) of the matrix of embedding vectors as a measure of uncertainty. effective rank serves as a smooth measure of the divergence among vectors within a matrix, and is employed in our method to detect semantic variations in different embedding ... - **Candidate**: we compute the entropy of the spectrum of σ, which is to say we are approximating the "information content", viz., the effective subspace rank of the semantic manifold (roy & vetterli, 2007) encapsulated by the set: {z(i)}i=m i=1 }. simply put, the more "semantically diverse" the set of outputs for ...

Evidence 2 - **Rationale**: Both methods follow the same core approach: generate multiple responses, extract embeddings from these responses, and use spectral analysis to quantify uncertainty through semantic diversity. - **Original**: for each query q, we first construct the representation matrix extracted from different layers and responses to calculate the effective rank. To this end, we samplem1 responses and extract embeddings from m2 layers per response, resulting in m=m 1 xm 2 embedding vectors - **Candidate**: To estimate predictive uq, we generate multiple language outputs for the input data and then approximate the semantic diversity of the low-dimensional manifold encapsulating these generated outputs. A large semantic diversity indicates high model uncertainty. gsee proceeds in four total steps: (1) g...

---

6. Boosting Accuracy & Efficiency: Teaching LLMs to

URL: View paper

Brief Assessment

Boosting Accuracy Efficiency[55] provides insufficient context to assess novelty claims. The candidate's full text contains only fragmentary mentions of effective rank and uncertainty quantification without detailed methodology or application to hallucination detection in LLMs.

Contribution 2: Theoretical Justification for Multi-Response Uncertainty Quantification

Description: The authors provide theoretical analysis showing that aleatoric uncertainty dominates and obscures epistemic uncertainty within single forward passes of LLMs. This theoretical framework justifies why multiple sampled responses are necessary to effectively detect hallucinations by externalizing the model's internal probability distribution as semantic divergence.

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. Semantically diverse language generation for uncertainty estimation in language models

URL: View paper

Prior Art Analysis

Diverse Language Generation[66] demonstrates that the theoretical framework for distinguishing aleatoric and epistemic uncertainty in language models, and the necessity of multiple sampled responses for uncertainty quantification, was established prior to the original paper. The candidate paper provides a formal theoretical foundation showing that aleatoric semantic uncertainty (semantic entropy) can be decomposed from total predictive uncertainty, and explicitly argues that sampling multiple output sequences is necessary to capture semantic diversity that reflects the model's internal probability distribution. This directly addresses the same theoretical justification that the original paper claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers provide theoretical decomposition of uncertainty into aleatoric and epistemic components. The candidate establishes this framework earlier, showing the theoretical foundation was already present. - **Original**: we theoretically demonstrate the necessity of quantifying uncertainty both internally (representations of a single response) and externally (different responses), providing a justification for using representations among different layers and responses from llms to detect hallucinations - **Candidate**: The total uncertainty, given by the posterior expectation of the crossentropyce(·;·), is additively decomposed into aleatoric and epistemic uncertainty. The aleatoric uncertainty is the shannon entropyh(·)of the predictive distribution under the given model. The epistemic uncertainty is the posterio...

---

2. A survey on uncertainty quantification methods for deep learning

URL: View paper

Brief Assessment

Uncertainty Quantification Survey[65] provides a general taxonomy of UQ methods for DNNs based on data versus model uncertainty sources, but does not address the specific theoretical analysis of aleatoric versus epistemic uncertainty dynamics within LLM forward passes or justify multi-response sampling for hallucination detection.

---

3. Uncertainty quantification for in-context learning of large language models

URL: View paper

Prior Art Analysis

In-context Uncertainty[68] provides a theoretical framework that decomposes predictive uncertainty into aleatoric and epistemic components, explicitly demonstrating that aleatoric uncertainty (from demonstrations) and epistemic uncertainty (from model configurations) are distinct sources requiring separate quantification. The paper formulates this decomposition from a Bayesian perspective and shows that both types of uncertainty must be measured to effectively assess LLM reliability. This directly addresses the same theoretical question as the original paper—why multiple responses are necessary—but from a different angle (demonstration variability vs. internal representation analysis).

Evidence

Evidence 1 - **Rationale**: Both papers provide theoretical justification for why uncertainty must be quantified from multiple sources. The candidate explicitly decomposes uncertainty into aleatoric (demonstrations) and epistemic (model configurations) components, while the original argues for internal vs. external quantification. Both justify multi-response sampling, though from different theoretical perspectives. - **Original**: we theoretically demonstrate the necessity of quantifying uncertainty both internally (representations of a single response) and externally (different responses), providing a justification for using representations among different layers and responses from llms to detect hallucinations. - **Candidate**: we delve into the predictive uncertainty of llms associated with in-context learning, highlighting that such uncertainties may stem from both the provided demonstrations (aleatoric uncertainty) and ambiguities tied to the model's configurations (epistemic uncertainty). we propose a novel formulation...

Evidence 2 - **Rationale**: The candidate provides a formal mathematical framework showing that total uncertainty entangles both aleatoric and epistemic components, requiring decomposition. This theoretical formulation directly supports the necessity of measuring uncertainty from multiple sources (demonstrations and model parameters). - **Original**: meanwhile, we theoretically demonstrate the necessity of quantifying uncertainty both internally (representations of a single response) and externally (different responses), providing a justification for using representations among different layers and responses from llms to detect hallucinations. - **Candidate**: epistemic uncertainty (eu). let h(·) be the differential entropy of a probability distribution, the total uncertainty in eq. (1) can be quantified as h (yt | x1:t), which entangles both aleatoric (i.e., demonstration sampling) and epistemic (i.e., model parameter θ) uncertainties.

Evidence 3 - **Rationale**: Both papers theoretically analyze how aleatoric uncertainty relates to the generation process. The candidate explicitly defines aleatoric uncertainty as arising from demonstration variability and provides a mutual information framework for quantification, while the original argues it obscures epistemic uncertainty in single-pass generation. - **Original**: we demonstrate how aleatoric uncertainty (the inherent stochasticity of llms) progressively amplifies and obscures epistemic uncertainty (uncertainty in the knowledge and capabilities encoded by the model's parameters) within the model's internal representations during a single sequence generation w... - **Candidate**: aleatoric uncertainty (au). in terms of au, the randomness comes from different sets of demonstration x1:t-1 and their corresponding latent concept z. to estimate au, we can quantify the mutual information between yt and latent concept z, which can often be leveraged as an evaluation metric of au

Evidence 4 - **Rationale**: The candidate provides a theoretical framework showing that epistemic uncertainty requires conditioning on different model parameters and demonstration sets, which necessitates multiple samples. This directly parallels the original's justification for multi-response sampling, though the candidate focuses on demonstration variability rather than internal representation divergence. - **Original**: this finding provides both the theoretical justification for the necessity of semantic sampling through multiple responses and the foundational rationale for its effectiveness. - **Candidate**: to estimate the eu, we condition eq. (1) on a specific realization of the model parameter θ, yielding p(yt |x1:t , θ) =r p(yt |x1:t , z,θ)p(z|x1:t)dz with an associated entropy h(yt |x1:t , z,θ). the expected value of this entropy under different demonstration sets can be expressed as ez [h(yt |x1:...

---

4. The role of predictive uncertainty and diversity in embodied ai and robot learning

URL: View paper

Brief Assessment

Predictive Uncertainty Diversity[69] focuses on aleatoric vs. epistemic uncertainty in robotics and embodied AI contexts, not specifically on LLM hallucination detection or the necessity of multiple sampled responses for externalizing probability distributions as semantic divergence.

---

5. SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models

URL: View paper

Brief Assessment

SPUQ[67] focuses on perturbation-based methods to address both aleatoric and epistemic uncertainty through input perturbations, not on theoretical analysis of how aleatoric uncertainty dominates epistemic uncertainty within single forward passes or why multiple responses are necessary to externalize probability distributions as semantic divergence.

---

6. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty

URL: View paper

Prior Art Analysis

Epistemic Uncertainty Prompting[71] provides a theoretical framework demonstrating that aleatoric uncertainty dominates epistemic uncertainty in single forward passes, necessitating multiple sampled responses. This directly addresses the same theoretical justification claimed by the original paper. Both papers analyze the decomposition of uncertainty into aleatoric and epistemic components, and both conclude that multiple responses are necessary to detect epistemic uncertainty effectively. The candidate paper presents formal mathematical analysis showing how aleatoric uncertainty obscures epistemic uncertainty within single responses, which is the core theoretical claim of the original contribution.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly distinguish between epistemic and aleatoric uncertainty as fundamental concepts for their theoretical frameworks. - **Original**: we theoretically demonstrate the necessity of quantifying uncertainty both internally (representations of a single response) and externally (different responses), providing a justification for using representations among different layers and responses from llms to detect hallucinations. - **Candidate**: we distinguish between two sources of uncertainty: epistemic and aleatoric (wen et al., 2022; osband et al., 2023; johnson et al., 2024). epistemic uncertainty arises from the lack of knowledge about the ground truth (e.g., facts or grammar in the language), stemming from various reasons such as ins...

Evidence 2 - **Rationale**: The candidate demonstrates a theoretical framework for why multiple responses are necessary, establishing prior work on this justification. - **Original**: meanwhile, we theoretically demonstrate the necessity of quantifying uncertainty both internally (representations of a single response) and externally (different responses), providing a justification for using representations among different layers and responses from llms to detect hallucinations. - **Candidate**: if multiple responses are obtained to the same query from the ground truth (the language), they should be independent from each other, that is, in probabilistic interpretation, the joint distribution of these multiple responses, for a fixed query, must be a product distribution. this observation can ...

Evidence 3 - **Rationale**: Both papers provide formal theoretical frameworks for quantifying epistemic uncertainty and explain why single responses are insufficient. - **Original**: we subsequently argue, however, that within any single sampled response, the information signal pertaining to epistemic uncertainty is largely masked by aleatoric uncertainty. furthermore, this aleatoric uncertainty can propagate and become amplified through the model's reasoning process. - **Candidate**: to measure epistemic uncertainty, we need to quantify how far the estimated pseudo joint distribution ˜qis from the ground truth ˜p. one natural choice is the following definition: definition 5.4 (epistemic uncertainty metric). given an input x ∈x , we say that the epistemic uncertainty of ˜qis quanti...

Evidence 4 - **Rationale**: Both papers cite the same theoretical foundation (decomposition of uncertainty) and apply it to justify multi-response sampling for detecting epistemic uncertainty. - **Original**: according to the perspective of depeweg et al. (2018), the uncertainty in the predictions of deep neural networks can be decomposed into two primary types:aleatoric uncertainty, which is inherent in the data and model architecture due to its ambiguity or inherent stochasticity, andepistemic uncertai... - **Candidate**: our uncertainty metric, similarly to the ones considered by, e.g., wen et al. (2022); osband et al. (2023), is based on analyzing the joint distribution of responses: if multiple responses are sampled jointly according to the ground-truth distribution, they should be independent (as one instantiatio...

---

7. Quantifying Uncertainties in Natural Language Processing Tasks

URL: View paper

Brief Assessment

Quantifying NLP Uncertainties[70] focuses on decomposing aleatoric vs. epistemic uncertainty using Bayesian neural networks for NLP tasks, but does not provide theoretical analysis about why aleatoric uncertainty dominates within single forward passes or justify multi-response sampling for hallucination detection in LLMs.

---

8. TAE: Topic-aware encoder for large-scale multi-label text classification

URL: View paper

Brief Assessment

Topic-aware Encoder[73] focuses on multi-label text classification using topic modeling, not uncertainty quantification in language models or hallucination detection through aleatoric/epistemic uncertainty decomposition.

---

9. The Geometry of Creative Variability: How Credal Sets Expose Calibration Gaps in Language Models

URL: View paper

Brief Assessment

Creative Variability Geometry[74] focuses on calibrating model uncertainty against human creative variation using credal sets in open-ended generation tasks. The original paper addresses aleatoric vs. epistemic uncertainty decomposition within LLM internal representations for hallucination detection, which is a fundamentally different problem domain and technical approach.

---

10. Uncertainty in natural language generation: From theory to applications

URL: View paper

Brief Assessment

Uncertainty NLG[72] provides a comprehensive theoretical framework for uncertainty in NLG but does not specifically address the dominance of aleatoric over epistemic uncertainty within single forward passes of LLMs, nor does it provide the mathematical justification for why multiple sampled responses are necessary to detect hallucinations through semantic divergence.

Contribution 3: Training-Free Hallucination Detection Framework

Description: The authors develop a lightweight, efficient hallucination detection approach that operates directly on pre-trained LLMs without requiring retrieval systems, auxiliary models, or fine-tuning. The method achieves competitive or superior performance compared to existing baselines while maintaining computational efficiency comparable to standard generation.

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. Counterfactual probing for hallucination detection and mitigation in large language models

URL: View paper

Brief Assessment

Counterfactual Probing[57] focuses on generating counterfactual statements and analyzing model sensitivity to perturbations, requiring multiple API calls for probe generation. The original paper uses effective rank of hidden states from internal representations without external probes or perturbations, representing a fundamentally different technical approach to training-free detection.

---

2. SelfCheckAgent: Zero-Resource Hallucination Detection in Generative Large Language Models

URL: View paper

Brief Assessment

SelfCheckAgent[61] focuses on black-box hallucination detection using multiple sampled responses and semantic consistency checking, while the original paper proposes a white-box method using effective rank of internal hidden states. These are fundamentally different approaches (black-box vs. white-box access).

---

3. Self-introspective decoding: Alleviating hallucinations for large vision-language models

URL: View paper

Brief Assessment

Self-introspective Decoding[60] focuses on alleviating hallucinations during text generation through contrastive decoding strategies, not on detecting hallucinations after generation. The original paper develops a detection framework, while this candidate develops a generation-time mitigation strategy.

---

4. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph

URL: View paper

Brief Assessment

Think-on-Graph[62] focuses on integrating LLMs with knowledge graphs for reasoning tasks, not on hallucination detection. It addresses hallucination by retrieving external knowledge during reasoning, which is fundamentally different from the original paper's training-free uncertainty-based detection approach that operates purely on internal model representations.

---

5. Zero-resource hallucination prevention for large language models

URL: View paper

Brief Assessment

Zero-resource Hallucination Prevention[5] focuses on pre-detection and prevention of hallucinations by evaluating model familiarity with input concepts before generation, rather than post-generation detection. The original paper proposes a post-generation detection method using effective rank of hidden states from multiple model outputs.

---

6. Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention

URL: View paper

Brief Assessment

Activation Steering Decoding[63] focuses on hallucination mitigation in vision-language models through activation intervention, not hallucination detection in text-only LLMs. The original paper develops a detection framework using effective rank of hidden states, while the candidate proposes a decoding-time intervention method for LVLMs.

---

7. Mitigating Hallucination in Large Vision-Language Models via Adaptive Attention Calibration

URL: View paper

Brief Assessment

Adaptive Attention Calibration[64] addresses hallucination in vision-language models through attention mechanism calibration, while the original paper focuses on uncertainty quantification in text-only LLMs using effective rank of hidden states. These are fundamentally different approaches to different problems.

---

8. Long-form hallucination detection with self-elicitation

URL: View paper

Brief Assessment

Long-form Self-elicitation[58] focuses on long-form content hallucination detection using self-generated thoughts and knowledge hypergraphs, not on training-free uncertainty-based detection for general LLM outputs without external knowledge.

---

9. Do language models know when they're hallucinating references?

URL: View paper

Brief Assessment

Hallucinating References[59] focuses specifically on detecting hallucinated references (books/articles) through consistency checks about reference metadata, not general hallucination detection across diverse tasks. The original paper addresses broader hallucination detection using effective rank-based uncertainty quantification across multiple benchmarks.

---

10. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models

URL: View paper

Prior Art Analysis

SelfCheckGPT[6] demonstrates prior work on training-free hallucination detection that operates without retrieval systems, auxiliary models, or fine-tuning. The candidate paper explicitly states it is a 'zero-resource black-box hallucination detection' method that requires 'no external database' and works by sampling multiple responses from the LLM itself. This directly refutes the novelty claim of the original paper's training-free framework, as SelfCheckGPT[6] was published in 2023 (EMNLP), predating the original submission to ICLR 2026.

Evidence

Evidence 1 - **Rationale**: SelfCheckGPT[6] explicitly claims to be the 'first zero-resource hallucination detection solution' that requires no additional tools or external knowledge, directly challenging the original paper's claim to novelty in developing a training-free detector. - **Original**: a lightweight, training-free detector. we introduce a rapid hallucination detection method that does not require additional tools, fine-tuning, or external knowledge. - **Candidate**: to the best of our knowledge, selfcheckgpt is the first work to analyze model hallucination of general llm responses, and is the first zero-resource hallucination detection solution that can be applied to black-box systems.

Appendix: Text Similarity Detection

Textual similarity detection checked 28 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. Revisiting Hallucination Detection with Effective Rank-based Uncertainty

Detected in: Contribution: contribution_1

⚠ **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. Unsupervised real-time hallucination detection based on the internal states of large language models

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] Revisiting Hallucination Detection Through The Lens Of Effective Rank-based Uncertainty View paper
- [1] Hallucination of multimodal large language models: A survey View paper
- [2] A comprehensive survey of hallucination mitigation techniques in large language models View paper
- [3] Hademif: Hallucination detection and mitigation in large language models View paper
- [4] Fine-grained hallucination detection and editing for language models View paper
- [5] Zero-resource hallucination prevention for large language models View paper
- [6] Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models View paper
- [7] Unsupervised real-time hallucination detection based on the internal states of large language models View paper
- [8] Fmdllama: Financial misinformation detection based on large language models View paper
- [9] Llm-check: Investigating detection of hallucinations in large language models View paper
- [10] A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions View paper
- [11] Disinformation capabilities of large language models View paper
- [12] Detecting hallucinations in large language models using semantic entropy View paper
- [13] Adapting fake news detection to the era of large language models View paper
- [14] Mitigating entity-level hallucination in large language models View paper
- [15] Catching chameleons: Detecting evolving disinformation generated using large language models View paper
- [16] Drowzee: Metamorphic testing for fact-conflicting hallucination detection in large language models View paper
- [17] The dawn after the dark: An empirical study on factuality hallucination in large language models View paper
- [18] Diahalu: A dialogue-level hallucination evaluation benchmark for large language models View paper
- [19] Bad actor, good advisor: Exploring the role of large language models in fake news detection View paper
- [20] Hallucination detection in LLM-enriched product listings View paper
- [21] Fake news detectors are biased against texts generated by large language models View paper
- [22] Let silence speak: Enhancing fake news detection with generated comments from large language models View paper
- [23] Combating misinformation in the age of llms: Opportunities and challenges View paper
- [24] Tutorial Proposal: Hallucination in Large Language Models View paper
- [25] Factuality challenges in the era of large language models and opportunities for fact-checking View paper
- [26] Hallucination detection and hallucination mitigation: An investigation View paper
- [27] GPT Hallucination Detection Through Prompt Engineering View paper
- [28] Evaluating LLMs at detecting errors in LLM responses View paper
- [29] Detecting Health Misinformation by Leveraging LLM Models and Debunk List View paper
- [30] Analysis of disinformation and fake news detection using fine-tuned large language model View paper
- [31] InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers View paper
- [32] Exploring and evaluating hallucinations in llm-powered code generation View paper
- [33] Hallucination Detection and Evaluation of Large Language Model View paper
- [34] Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data View paper
- [35] FakeGPT: fake news generation, explanation and detection of large language models View paper
- [36] Comments and feedback verification system using large language model View paper
- [37] Hallucination Detection in Large Language Models Using Diversion Decoding View paper
- [38] Neural Probe-Based Hallucination Detection for Large Language Models View paper
- [39] Hallucination Detection in Large Language Models with Metamorphic Relations View paper
- [40] The earth is flat? unveiling factual errors in large language models View paper
- [41] From deception to detection: The dual roles of large language models in fake news View paper
- [42] Automated Hallucination Detection and Mitigation in Large Language Model View paper
- [43] (Im)possibility of Automated Hallucination Detection in Large Language Models View paper
- [44] Developing a reliable, fast, general-purpose hallucination detection and mitigation service View paper
- [45] Felm: Benchmarking factuality evaluation of large language models View paper
- [46] Benchmarking Hallucination in Large Language Models Based on Unanswerable Math Word Problem View paper
- [47] An Information Reliability Framework for Detecting Misinformation based on Large Language Models View paper
- [48] Limits of detecting text generated by large-scale language models View paper
- [49] Detecting llm hallucinations using monte carlo simulations on token probabilities View paper
- [50] INSIDE: LLMs' internal states retain the power of hallucination detection View paper
- [51] UNComp: Uncertainty-Aware Long-Context Compressor for Efficient Large Language Model Inference View paper
- [52] On-Device Large Language Models: A Survey of Model Compression and System Optimization View paper

- [53] Revisiting Hallucination Detection with Effective Rank-based Uncertainty View paper
- [54] Learning Probabilistic Box Embeddings for Effective and Efficient Ranking View paper
- [55] Boosting Accuracy & Efficiency: Teaching LLMs to View paper
- [56] Uncertainty Quantification with Generative-Semantic Entropy Estimation for Large Language Models View paper
- [57] Counterfactual probing for hallucination detection and mitigation in large language models View paper
- [58] Long-form hallucination detection with self-elicitation View paper
- [59] Do language models know when they're hallucinating references? View paper
- [60] Self-introspective decoding: Alleviating hallucinations for large vision-language models View paper
- [61] SelfCheckAgent: Zero-Resource Hallucination Detection in Generative Large Language Models View paper
- [62] Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph View paper
- [63] Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention View paper
- [64] Mitigating Hallucination in Large Vision-Language Models via Adaptive Attention Calibration View paper
- [65] A survey on uncertainty quantification methods for deep learning View paper
- [66] Semantically diverse language generation for uncertainty estimation in language models View paper
- [67] SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models View paper
- [68] Uncertainty quantification for in-context learning of large language models View paper
- [69] The role of predictive uncertainty and diversity in embodied ai and robot learning View paper
- [70] Quantifying Uncertainties in Natural Language Processing Tasks View paper
- [71] To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty View paper
- [72] Uncertainty in natural language generation: From theory to applications View paper
- [73] TAE: Topic-aware encoder for large-scale multi-label text classification View paper
- [74] The Geometry of Creative Variability: How Credal Sets Expose Calibration Gaps in Language Models View paper