White-Box Feature Mining for Hallucination Detection in Large Language Models

Haibin Lai*
Southern University of Science and Technology
CS306 Data Mining Course Report
Shenzhen, China
12211612@mail.sustech.edu.cn

Abstract

Large language models (LLMs) have shown strong generation ability in question answering, summarization, code generation, and multi-step reasoning. However, they may produce fluent but factually incorrect, contextually inconsistent, or logically unsupported content, a phenomenon commonly known as hallucination. Because hallucinations are often hidden inside well-formed and confident natural language, keyword matching and hand-written rules are unreliable. This report studies the scientific problem of how data mining techniques can transform a model's internal generation traces into compact features for hallucination detection. We focus on white-box detection, in which the detector accesses internal signals produced during generation, including attention maps, hidden activations, logits, and predictive entropy. We argue that attention weights are a particularly informative signal: the ratio of attention placed on the input context versus the model's own generated tokens reflects whether a claim is grounded in evidence. We analyze the main challenges of white-box mining, namely high-dimensional and weakly interpretable internal tensors, the gap between token-level uncertainty and semantic hallucination, fine-grained localization in long-form text, and cross-model transferability. We then review representative attention-based, hidden-state, and logit-entropy methods, and show how they can be unified through an anomaly-detection view that localizes hallucinations to specific sentences or tokens. Finally, we propose a white-box framework that mines multi-layer attention and internal-state features for hallucination risk scoring and evidence attribution.

Keywords

large language models, hallucination detection, white-box detection, attention weights, internal states, anomaly detection

1 Introduction

Language models are powerful and widely used [23, 24]. But even the best LLMs often generate text that is factually wrong or nonsensical [7, 13, 26]. This problem is called hallucination [11]. In LLMs, a hallucination is text that reads fluently but does not match the intended meaning, the facts, the logic, or the context [17]. Such errors are especially harmful in high-stakes domains such as medicine, law, and finance, where one fabricated fact can mislead users [18]. What makes the problem hard is precisely this surface plausibility: a hallucinated claim is usually a small, locally incorrect span embedded in an otherwise correct and grammatical passage [10], so it cannot be caught by keyword matching or simple lexical rules.

*Student IDs: Haibin Lai 12211612

Table 1: Taxonomy of hallucination detection methods by the information available to the detector.

Type	Detection idea	Observed signal	Work
White-box	Attention mining	Attention maps	[5, 27, 29]
	Hidden states	Activations	[4, 6, 25]
	Logits / entropy	Output distribution	[30]
Black-box (resource-free)	Self-consistency	Multiple samples	[9, 20]
	Cross-examination	Generated answers	[7]
Black-box (resource)	Knowledge / retrieval	Atomic facts	[14, 22]

A growing body of work tries to detect hallucinations automatically. A useful way to organize these methods is by how fine-grained the unit of judgment is and what information the detector observes. Rather than validating a whole sentence at once, recent approaches first identify the key concepts, such as named entities and keyphrases, that carry the factual content of a response, and treat them as the candidates most likely to be hallucinated [30]. They then estimate how uncertain the model is about each candidate. When the model has access to logit outputs, the token-level probabilities of a concept give a direct signal: a low probability on even a single token of an entity (for example, an unfamiliar surname) is strong evidence that the model is guessing, whereas averaging the probabilities tends to wash out this signal [30]. Such uncertainty scores act as cheap, generation-time indicators that can flag suspicious spans before an expensive external check is run.

Existing hallucination detection methods can be broadly organized by the information available to the detector, as summarized in Table 1. *White-box* methods access internal generation signals—attention maps, hidden activations, logits, and layer-wise representations—and mine them into features that indicate whether a generated claim is grounded. *Black-box* methods observe only inputs and outputs: zero-resource variants exploit repeated sampling, self-consistency, and LLM-as-a-judge evaluation, whereas non-zero-resource variants compare generated claims against retrieved documents, external databases, or knowledge graphs [20, 22]. Black-box detectors are flexible and apply even to closed commercial models, but they act only after generation and cannot observe the model's internal uncertainty; white-box detectors are finer-grained and need only a single forward pass, at the cost of requiring access to model weights. Some recent benchmarks further split off a *gray-box* category for methods that read only output probabilities or

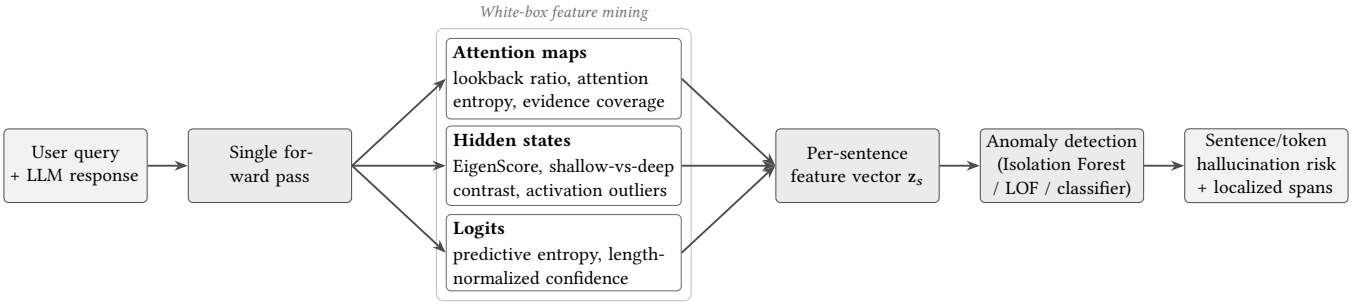

Figure 1: A white-box, attention-centered feature mining workflow for LLM hallucination detection, integrating attention-, hidden-state-, and logit-based signals from prior work [4, 5, 30]. A single forward pass yields attention, hidden-state, and logit signals; per-sentence features are fused and scored by anomaly detection.

logits without internal states [16]; for clarity we keep these under white-box, since they still rely on model-internal signals.

This report therefore focuses on white-box hallucination detection. In white-box settings, the detector can access internal model signals produced during generation, including hidden activations, logits, predictive entropy, attention maps, gradients, and layer-wise representations. These signals provide a richer view of how the model selects tokens, stores factual knowledge, attends to input context, and balances parametric knowledge against external evidence. From a data mining perspective, the key challenge is to transform these high-dimensional internal traces into compact, interpretable, and transferable features for hallucination detection.

2 Core Scientific Question

The core scientific question studied in this report is:

> In open-domain scenarios where absolute ground truth is unavailable or expensive to annotate, how can we mine inconsistencies among LLM outputs, generation distributions, internal states, and external knowledge in order to automatically identify, localize, and explain semantic hallucinations?

This question has two levels. The first is detection: deciding whether a generated passage contains hallucinated content. The second is attribution: identifying which sentence, entity relation, or reasoning step is suspicious, and returning evidence that can support human review. Detection determines whether a system can issue timely warnings, while attribution determines whether the system can be trusted and improved in real applications.

3 Research Challenges

We focus on the challenges of mining a model's internal generation traces, rather than on external evaluation. Four difficulties stand out for white-box hallucination detection.

3.1 High-Dimensional and Weakly Interpretable Internal Signals

During generation, an LLM produces logits, attention weights, hidden states, and token probabilities at every layer and every head. These signals form high-dimensional tensors with complex layer-wise dependencies, and they cannot be read directly as human-understandable hallucination indicators. For example, a single response of about 512 tokens from Qwen3.6-27B (27B parameters, 64 layers, of which 16 use gated full attention with 24 query heads, hidden size 5120, vocabulary 248320) already induces on the order of $16 \cdot 24 \cdot T^2 \approx 1.0 \times 10^8$ attention weights and $L\,T\,d \approx 1.7 \times 10^8$ hidden-state values. INSIDE shows that internal states still retain useful information for hallucination detection [4], but the challenge, illustrated in Figure 2, is to compress these hundreds of millions of raw scalars into a handful of stable, low-dimensional, and interpretable mining features.

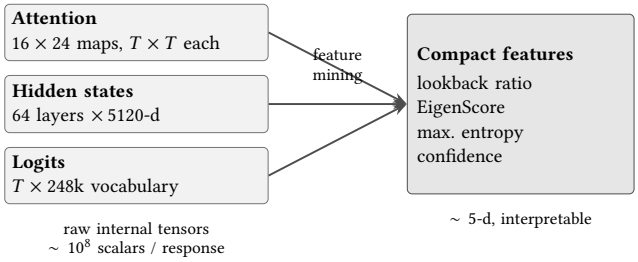

Figure 2: Challenge of high-dimensional internal signals: one forward pass of Qwen3.6-27B yields on the order of 10^8 attention and hidden-state scalars, which must be compressed into a few interpretable features.

3.2 Token-Level Uncertainty Is Not Hallucination

A natural idea is to treat high predictive entropy or low token probability as a hallucination signal. However, open-ended language admits many valid expressions, so a high-entropy token is often only stylistic variation rather than a factual error. Conversely, a model can be confidently wrong and assign high probability to a fabricated entity. The challenge is to design features that separate harmless surface uncertainty from genuine factual or contextual hallucination, for example by contrasting attention to evidence against attention to self-generated tokens.

Table 2: Comparison of White-Box Feature Families

Feature family	Internal signal	Strength	Limitation
Attention mining	Per-head lookback ratio, attention map [5, 27, 29]	Single forward pass; localizes ungrounded spans; transfers across model sizes	Needs open weights; head selection matters
Hidden states	EigenScore, layer contrast, probes, subspaces, layer dynamics [1, 3, 4, 6, 8, 12, 25, 28, 31, 32]	Dense semantics; supports activation-level anomaly localization	High memory cost; architecture-specific
Logits / entropy	Predictive entropy, confidence [30]	Cheap; available even with limited access	Token uncertainty ≠ hallucination

3.3 Error Dilution and Fine-Grained Localization

LLM hallucination is usually not random noise. It is often a local error embedded in an otherwise fluent paragraph, such as a single wrong birth year, affiliation, or publication title in an otherwise correct biography. Passage-level scores are diluted by the overall fluency of the text. Effective white-box detection therefore requires mapping internal features back to specific sentences, entities, or tokens, and then applying local anomaly discovery to flag the exact suspicious span.

3.4 Cross-Model Transfer and Access Constraints

White-box features depend on architecture-specific quantities, so a detector trained on one model may transfer poorly to another model family with different depth, head count, or attention behavior. Moreover, attention and activation signals are only available for open-weight models; closed commercial APIs expose neither, which limits applicability. A practical white-box detector must therefore aim for features that are as model-agnostic as possible and degrade gracefully when only partial internal access is available.

4 White-Box Technical Directions

A survey of hallucination detection groups white-box features into five types: hidden-layer activations, logits, entropy, attention weights, and gradients [17]. We organize them into three feature families and then describe how anomaly detection turns these features into a localized hallucination decision.

4.1 Attention-Based Mining

Attention weights directly record where the model looks when generating each token. The central hypothesis, formalized by Lookback Lens, is that contextual hallucination is related to how much an LLM attends to the provided context versus its own previously generated tokens [5]. Let $a_{t,i}^{(l,h)}$ be the attention weight from the token generated at step t to position i in attention head h of layer l. According to Lookback Lens [5], splitting the positions into the context span C and the already-generated span G, the lookback ratio of one head is defined as

$$\mathrm{LR}_t^{(l,h)} = \frac{\sum_{i \in C} a_{t,i}^{(l,h)}}{\sum_{i \in C} a_{t,i}^{(l,h)} + \sum_{j \in G} a_{t,j}^{(l,h)}}. \tag{1}$$

Stacking $\mathrm{LR}_t^{(l,h)}$ over all layers and heads gives a feature vector for token t; averaging over a span gives a sentence-level feature. A low ratio means the model is "talking to itself" instead of grounding on evidence, which signals higher hallucination risk. Notably, a simple linear classifier on lookback-ratio features matches a much richer hidden-state detector and even transfers from a 7B model to a 13B model without retraining [5].

Beyond a single ratio, the shape of the attention map carries information. LLM-Check analyzes the attention map of a single response to design lightweight detection scores that need no repeated sampling [27]. For retrieval-augmented generation, ReDeEP uses mechanistic interpretability to decouple the contribution of external context from parametric knowledge, combining attention weights with logits to detect when the model ignores retrieved evidence [29]. It frames detection causally, treating parametric knowledge and external context as mutually confounding factors and explicitly separating their contributions instead of mixing them as plain token-probability uncertainty does. Useful attention-derived statistics include attention entropy (how diffuse the attention is), the attention mass placed on entity tokens, and the evidence coverage (whether some context spans receive almost no attention from any head). AggTruth aggregates per-head attention scores over the context to obtain a single contextual-faithfulness statistic that is cheap to compute at generation time [21], while spectral analysis of the attention maps treats their eigenvalue spectrum as a feature, capturing global structure that a single ratio misses [2]. By contrasting attention to evidence against attention to self-generated tokens, this family directly addresses Challenge 3.2: it separates harmless surface uncertainty from genuinely ungrounded content. The transferability of the lookback ratio across model sizes also mitigates Challenge 3.4.

4.2 Hidden States and Activations

Hidden activations encode the dense semantics that logits lose after token decoding. According to INSIDE [4], this can be exploited by computing the EigenScore, which uses the eigenvalues of the covariance matrix of several responses' embeddings to measure semantic self-consistency in the dense representation space. DoLa observes that factual knowledge is more concentrated in deeper layers, and contrasts the projections of shallow and deep layers into the vocabulary space to expose content that deviates from the model's internal knowledge [6]. Rateike et al. take an explicit anomaly view: they run statistical tests on the activation distributions of hallucinated versus faithful text and locate the anomalous activation units, together with the input features that trigger them [25].

A second line treats the hidden states as a feature space to be probed or clustered. Supervised probes such as SAPLMA train a classifier on input–output hidden states to predict whether a statement is true [1], while MIND removes the labeling cost by learning the probe from generation-time states in an unsupervised, real-time setting [28]. Subspace methods instead look for a truthfulness direction: CCS finds a contrast-consistent direction from paired hidden states without supervision [3], and HaloScope estimates a hallucination-related subspace from unlabeled generations and fits a classifier on the induced membership [8]. SEP further shows that semantic entropy can be approximated by a probe on the hidden states of a single generation, avoiding repeated sampling [12]. Finally, dynamics-guided features track how representations evolve: PRISM uses prompt-guided hidden states to improve cross-domain transfer [31], and ICR Probe builds scores from cross-layer hidden-state updates together with attention maps [32], linking this family back to the attention signals of Section 4.1. By compressing high-dimensional hidden states into a single scalar (EigenScore) or a small set of anomalous units, this family responds to Challenge 3.1, turning raw internal tensors into compact and interpretable features.

4.3 Logits and Predictive Entropy

The most accessible white-box signal is the output distribution. Following standard uncertainty estimation for language generation [13], a simple token-level measure is predictive entropy

$$H_t = -\sum_{v \in V} p(v \mid x, y_{<t}) \log p(v \mid x, y_{<t}), \tag{2}$$

where V is the vocabulary. Rising entropy near important entity tokens suggests the model is uncertain about the corresponding fact. Varshney et al. use such low-confidence signals to detect and then mitigate hallucinations during generation, validating low-probability spans before the error propagates [30]. Because long sequences accumulate more low-probability terms, length-normalized log-probability is commonly used so that comparisons across spans of different length remain fair. Recent work refines this raw uncertainty: Semantic Energy aggregates the energy of the output distribution over semantically equivalent answers, separating confident errors from harmless lexical variation better than plain entropy [19], and a Fast Fourier Transform of the per-layer probability signals along the token axis exposes periodic patterns that distinguish faithful from hallucinated spans [15].

4.4 Anomaly Detection for Localization

The three feature families above can be unified through a common data-mining pattern. Each sentence (or atomic span) is represented by a multi-level feature vector that concatenates attention statistics, hidden-state summaries, and entropy or confidence values:

$$\mathbf{z}_s = \left[\, \overline{\mathrm{LR}}_s,\ H_s^{\max},\ \mathrm{EigenScore}_s,\ \mathrm{conf}_s^{\mathrm{ent}},\ \dots \,\right]. \tag{3}$$

Treating most generated sentences as "normal," hallucinated spans appear as outliers in this feature space. Unsupervised detectors such as isolation forest, local outlier factor, or density-based clustering can then score and rank spans, while a lightweight supervised classifier can be trained when labels are available. This view directly answers the localization requirement of Challenge 3.3: instead of one passage-level score, the detector returns the precise sentence or token where the internal signals look abnormal, so a local error is no longer diluted by the overall fluency of the passage. Compared with purely black-box methods, white-box mining offers finer localization and needs only a single forward pass, at the cost of requiring model access and careful feature selection.

5 Proposed White-Box Method

Based on the above analysis, this report proposes an attention-centered white-box framework, as shown in Figure 1. With a single forward pass, the first stage segments the response into sentences and entity spans and records the attention maps, hidden states, and logits aligned to each span. The second stage mines features per span: per-head lookback ratios and attention entropy from the attention branch, EigenScore and layer-contrast scores from the hidden-state branch, and predictive entropy and length-normalized confidence from the logit branch. The third stage feeds the concatenated feature vector $\mathbf{z}_s$ into an anomaly detector (isolation forest or LOF when labels are scarce, or a lightweight classifier when labels exist) to produce a sentence- or token-level hallucination risk.

The goal is not only higher accuracy, but also interpretable localization. For every high-risk span, the system reports the dominant internal signal behind the alert, for example a collapsed lookback ratio, an outlier activation unit, or abnormally high entropy on an entity token. This design keeps the detector itself transparent rather than turning it into another opaque black box.

Experiments can be conducted on open-domain question answering, biography generation, and summarization. Evaluation data can be drawn from TruthfulQA, FActScore, or manually constructed fact-checking datasets. Detection performance can be measured by accuracy, recall, F1, and AUC. Localization performance can be measured by sentence-level or atomic-fact-level F1. Explanation quality can be evaluated by evidence coverage and agreement with human reviewers.

False positives and false negatives should be evaluated separately. In high-risk domains such as medicine and law, recall should be prioritized because severe hallucinations should not be missed. In everyday writing assistance, precision may matter more because frequent false alarms can interrupt the user experience.

6 Conclusion

LLM hallucination detection is fundamentally a data mining problem over a model's internal generation traces. Surface text similarity alone cannot reliably find factual errors hidden inside fluent language. This report focused on white-box mining, and argued that attention weights, especially the ratio of attention to context versus self-generated tokens, are a compact and transferable signal for whether a claim is grounded. By combining attention, hidden-state, and logit features under a shared anomaly-detection view, the detector can localize hallucinations at the sentence or token level and report the internal signal behind each alert. Future work should further address cross-model transfer of attention features, low-cost long-form detection, and the reliability of the detector itself, so that LLMs can be used more safely in high-stakes applications.

References

[1] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 967–976.

[2] Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. 2025. Hallucination Detection in LLMs Using Spectral Features of Attention Maps. *arXiv preprint arXiv:2502.17598* (2025).

[3] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations*.

[4] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *The Twelfth International Conference on Learning Representations*.

[5] Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1419–1436.

[6] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[7] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. *arXiv preprint arXiv:2305.13281* (2023).

[8] Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. In *Advances in Neural Information Processing Systems*, Vol. 37.

[9] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630 (2024), 625–630. doi:10.1038/s41586-024-07421-0

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55. doi:10.1145/3703155

[11] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38. doi:10.1145/3571730

[12] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. *arXiv preprint arXiv:2406.15927* (2024).

[13] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. 9459–9474.

[15] Jinxin Li, Gang Tu, ShengYu Cheng, Junjie Hu, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. 2025. LLM Hallucination Detection: A Fast Fourier Transform Method Based on Hidden Layer Temporal Signals. *arXiv preprint arXiv:2509.13154* (2025).

[16] Xinyi Li, Zhen Fang, Yongxin Deng, Jinyuan Luo, Hongnan Ma, Changdae Oh, Zijing Shi, Shanshan Ye, Hanchen Wang, Shu-Lin Chen, Yadan Luo, Mengyue Yang, Sean Du, Sharon Li, and Ling Chen. 2026. OpenHalDet: A Unified Benchmark for Hallucination Detection across Diverse Generation Scenarios. *arXiv preprint arXiv:2606.06959* (2026).

[17] Zituo Li, Jianbin Sun, Guangzhou Chen, Xinyue Fang, Ruijing Cui, Zhiliang Tian, Zhen Huang, and Kewei Yang. 2026. Survey of Hallucination Detection Methods for Large Language Models. *Journal of Computer Research and Development* 63, 1 (2026). doi:10.7544/issn1000-1239.202550069 In Chinese.

[18] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3214–3252. doi:10.18653/v1/2022.acl-long.229

[19] Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. 2025. Semantic Energy: Detecting LLM Hallucination Beyond Entropy. *arXiv preprint arXiv:2508.14496* (2025).

[20] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 9004–9017. doi:10.18653/v1/2023.emnlp-main.557

[21] Piotr Matys, Jan Eliasz, Konrad Kiełczyński, Mikołaj Langner, Teddy Ferdinan, Jan Kocoń, and Przemysław Kazienko. 2025. AggTruth: Contextual Hallucination Detection using Aggregated Attention Scores in LLMs. *arXiv preprint arXiv:2506.18628* (2025).

[22] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741

[23] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. 27730–27744.

[25] Miriam Rateike, Celia Cintas, John Wamburu, Tanya Akumu, and Skyler Speakman. 2023. Weakly Supervised Detection of Hallucinations in LLM Activations. In *Advances in Neural Information Processing Systems*, Vol. 36.

[26] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. Out-of-Distribution Detection and Selective Generation for Conditional Language Models. In *The Eleventh International Conference on Learning Representations*.

[27] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 37. 34188–34216.

[28] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

[29] Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2024. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability. *arXiv preprint arXiv:2410.11414* (2024).

[30] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. *arXiv preprint arXiv:2307.03987*.

[31] Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. 2024. Prompt-Guided Internal States for Hallucination Detection of Large Language Models. *arXiv preprint arXiv:2411.04847* (2024).

[32] Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025. ICR Probe: Tracking Hidden State Dynamics for Reliable Hallucination Detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.