HDIR: TRAINING-FREE HALLUCINATION DETECTION VIA DISCRIMINATIVE INTERNAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models have achieved remarkable success in natural language processing, but they remain prone to hallucinations, generating plausible yet factually incorrect content. Effective hallucination detection is therefore crucial for assessing the factuality of generated content. Existing methods mainly rely on output signals or internal representations: the former are efficient and training-free but overlook richer semantic information encoded in model states, while the latter can better exploit such information but often require additional training. To address these limitations, we propose **H**allucination Detection based on **D**iscriminative **I**nternal **R**epresentations (HDIR), a training-free framework that exploits internal representations for hallucination detection. Our analysis shows that the ℓ_2 norm of attention-weighted internal representations provides a strong discriminative signal for hallucination detection. We define this measure as Semantic Intensity and further propose a Cohen's d-based selection strategy to identify the most discriminative attention heads and dimensions, thereby strengthening the discriminative signal of Semantic Intensity. Extensive experiments across multiple benchmarks and LLMs demonstrate that HDIR achieves strong overall hallucination detection performance, generalizes well across datasets, and complements output-level uncertainty signals, with their combination yielding further performance gains.

1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has brought profound transformation to the field of natural language processing. Owing to their powerful capabilities in language understanding and generation, LLMs have demonstrated broad application value across a variety of downstream tasks, including knowledge-based question answering and domain-specific analysis (Wang et al., 2024; Niu & Tian, 2025; Su et al., 2025). However, LLMs are widely known to suffer from the hallucination problem during generation (Huang et al., 2025; Alansari & Luqman, 2026; Lu et al., 2026), wherein the model produces content that appears plausible but is factually incorrect. This issue severely constrains the practical deployment of LLMs in high-reliability scenarios. Notably, even with the introduction of techniques such as retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023) and supervised fine-tuning (SFT) (Hu et al., 2022), hallucinations cannot be fully eliminated. This suggests that LLM hallucination may be rooted in deeper internal mechanisms of the model, rather than arising solely from external engineering limitations (Yao et al., 2023). Consequently, how to effectively detect hallucinations in content generated by LLMs has emerged as an important research topic that has attracted increasing attention from the academic community.

Existing hallucination detection methods can be broadly categorized by the source of information they utilize and whether they require training. Output-based methods estimate hallucination risk using external output signals, such as consistency across multiple sampled responses (Manakul et al., 2023; Abdaljalil et al., 2026) or output uncertainty (Malinin & Gales, 2020; Ma et al., 2025). These methods are generally training-free and easy to deploy, but rely solely on the model's final outputs, overlooking the rich discriminative signals encoded in internal representations. Moreover, consistency-based methods additionally require sampling multiple generations, incurring substantial inference overhead. Internal-representation-based methods, in contrast, leverage hidden-layer representations from the generation process to capture signals indicative of hallucination (Azaria & Mitchell, 2023; Su et al., 2024b). However, most existing approaches directly train classifiers or probes on internal representations, primarily treating them as features for hallucination detection

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rather than explicitly identifying components with stronger hallucination-discriminative signals. As a result, the discriminative contributions of individual components remain less explicit.

To address these limitations, we propose HDIR, a training-free hallucination detection method that directly exploits discriminative information encoded in internal representations, without requiring repeated sampling. Specifically, HDIR constructs an attention-weighted representation from the hidden states of an intermediate layer and defines its ℓ_2 norm as **Semantic Intensity**. Our analysis shows that factual generations consistently exhibit higher Semantic Intensity than hallucinated ones, making it an effective signal for hallucination detection. To further enhance its discriminative capability, we introduce a selection strategy based on Cohen's d effect size (Cohen, 2013) that leverages an analysis set to identify the most discriminative attention heads and representation dimensions, thereby improving detection performance. Furthermore, we show that HDIR is complementary to output-based uncertainty estimation methods, and that their fusion further improves hallucination detection performance. The key contributions of our work are summarized as follows:

- We propose HDIR, a hallucination detection method that requires neither repeated sampling nor model training. HDIR quantifies Semantic Intensity from attention-weighted internal representations to distinguish factual from hallucinated generations.
- We propose a Cohen's d-based selection strategy to identify the most discriminative attention heads and representation dimensions, showing that a small subset of internal features can capture robust, dataset-agnostic hallucination-discriminative signals.
- Extensive experiments show that HDIR achieves superior overall hallucination detection performance across models and benchmarks, while complementing output-based uncertainty signals and further improving performance through score fusion.

2 Related Work

2.1 Output-based Hallucination Detection

Output-based methods assess the reliability of generated content using signals derived from model outputs, including confidence and consistency across multiple generations.

Confidence-based methods rely on signals from a single generation pass, such as predictive entropy (Malinin & Gales, 2020) and sequence-level log-probability metrics (Ren et al., 2022). In contrast, sampling-based methods estimate reliability from the consistency of multiple generations. Lexical Similarity (Lin et al., 2022) measures lexical overlap among sampled responses, while SelfCheckGPT (Manakul et al., 2023) evaluates consistency by comparing independently generated responses. PTrue (Kadavath et al., 2022) prompts the model to self-assess the truthfulness of its generated answer, whereas Semantic Entropy (Kuhn et al., 2023) and Semantic Embedding Uncertainty (Grewal et al., 2024) quantify semantic uncertainty based on the distribution and embedding-space variation of multiple generations, respectively.

Recent studies further explore combining complementary signals, such as sampling consistency with internal representations (Chen et al., 2024) or output confidence with self-evaluation (Kumar et al., 2024). Although these methods are generally training-free, sampling-based components require multiple generations, while output-level signals are derived from final model outputs and may not capture the finer-grained signals encoded in internal representations.

2.2 Representation-based Hallucination Detection

Internal representations encode rich semantic information, motivating their use for hallucination detection. CCS (Burns et al., 2022) and RepE (Zou et al., 2023) demonstrate that truth-related knowledge, along with other high-level concepts, is encoded in the hidden states of LLMs. Existing representation-based methods exploit such internal information in different ways. SAPLMA (Azaria & Mitchell, 2023) trains classifiers on hidden-layer activations, while MIND (Su et al., 2024b) constructs an unsupervised detector from last-layer token representations. Recent studies further investigate the localization and extraction of discriminative internal features, including key tokens (Orgad et al., 2025), selected layers and neurons (Zhang et al., 2025b), prompt-guided representations (Zhang et al., 2025a), and hallucination subspaces (Du et al., 2024). However, these approaches

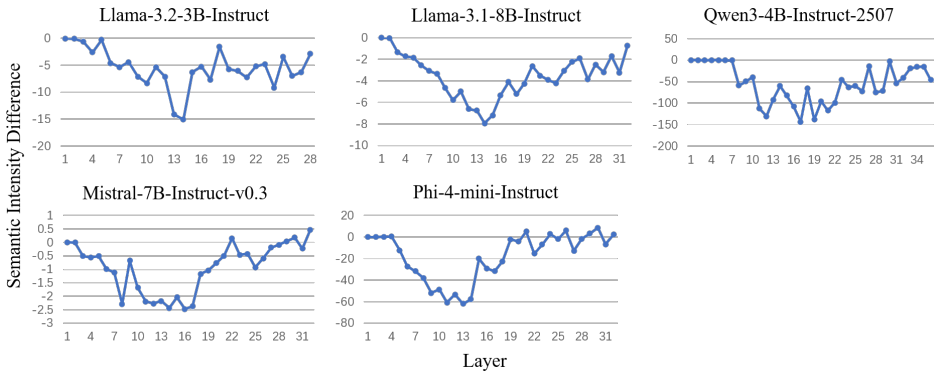


Figure 1: Layer-wise difference in mean Semantic Intensity (hallucinated - factual) across five LLMs.

primarily rely on trained classifiers or probes to identify features predictive of hallucinations, with less emphasis on directly quantifying the hallucination-discriminative signals encoded in individual internal components.

In contrast, HDIR directly analyzes representations from factual and hallucinated responses to identify discriminative internal components and then quantifies these signals, without training an auxiliary detector.

3 Motivation

Previous studies have shown that hidden-layer representations of LLMs encode discriminative information related to factuality and semantics (Azaria & Mitchell, 2023; Su et al., 2024b; Niu et al., 2026). We hypothesize that appropriately aggregating these representations and measuring the magnitude of the resulting representation yields a Semantic Intensity that is strongly correlated with hallucination. To validate this hypothesis, we conduct a preliminary empirical analysis on a separately constructed analysis set (Section 5.1.1) across multiple LLMs. Specifically, we compute attention-weighted internal representations for each Transformer layer using attention weights averaged over all heads, measure their Semantic Intensity by the ℓ_2 norm, and compute the layer-wise difference in mean Semantic Intensity between hallucinated and factual responses.

As shown in Figure 1, hallucinated samples consistently exhibit lower Semantic Intensity than factual ones across all evaluated LLMs. The discrepancy is most pronounced in the intermediate Transformer layers and gradually diminishes toward both shallow and deeper layers. This pattern likely reflects the functional specialization of Transformer layers, with intermediate layers encoding richer semantic and factual information (Azaria & Mitchell, 2023). Hallucinated generations may therefore exhibit weaker internal semantic signals in these layers, leading to lower Semantic Intensity. We also examine alternative representation construction approaches, with the corresponding results reported in Appendix A.

These observations suggest that Semantic Intensity derived from intermediate-layer attention-weighted representations provides a promising signal for hallucination detection. The same pattern is also observed in code generation, with results reported in Appendix B.

4 Method

Motivated by the empirical findings in Section 3, we propose HDIR, whose overall pipeline is illustrated in Figure 2. HDIR consists of three main components. First, it constructs an attention-weighted internal representation that emphasizes semantic information associated with highly attended tokens. Second, it quantifies its Semantic Intensity as a discriminative signal for hallucination detection. Finally, it uses a Cohen's d-based selection strategy to identify discriminative attention heads and dimensions for computing the final hallucination detection score.

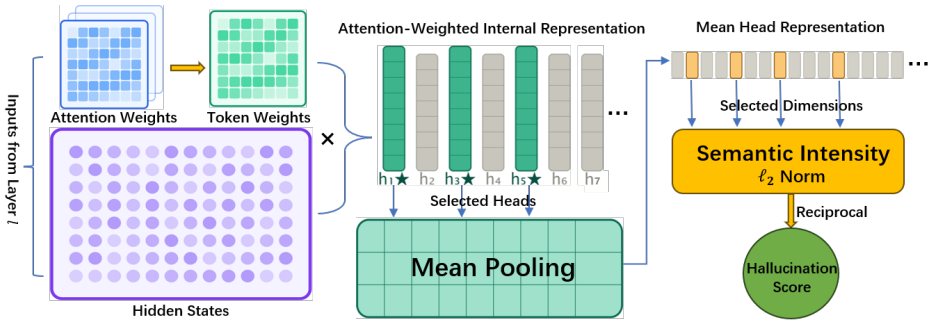


Figure 2: The framework of HDIR.

4.1 Attention-Weighted Internal Representation

During LLM generation, different tokens contribute unequally to the final output, with key tokens playing a more important role in determining factuality (Lee et al., 2024; Orgad et al., 2025; Zhang et al., 2025b). The attention mechanism (Vaswani et al., 2017) assigns an attention weight to each token at every generation step, reflecting the extent to which the token is attended to when producing subsequent tokens. Aggregating these weights across generation steps yields a token-level score that serves as a proxy for the token's overall contribution to the generated response (Zhang et al., 2023; Su et al., 2024a). We therefore use these aggregated attention weights to perform weighted aggregation over the hidden-layer representations, yielding a sequence-level semantic representation.

Specifically, given the hidden states $\mathbf{H}^l \in \mathbb{R}^{T\times d}$ and the attention weight matrix $\mathbf{A}^l \in \mathbb{R}^{K\times T\times T}$ at layer l, where T denotes the sequence length, d denotes the hidden dimension, and K denotes the number of attention heads, we first sum the attention weights over the query dimension and normalize by T, obtaining the token weights for the k-th attention head:

$$w^l_{k,t} = \frac{1}{T}\sum_{i=1}^{T} \mathbf{A}^l_{k,i,t}, \quad t = 1, \ldots, T. \tag{1}$$

The resulting token weights are then used to compute a weighted sum of the hidden states, yielding a head-specific internal representation:

$$\tilde{\mathbf{h}}^l_k = \sum_{t=1}^{T} w^l_{k,t} \cdot \mathbf{H}^l_t, \quad \tilde{\mathbf{h}}^l_k \in \mathbb{R}^d \tag{2}$$

Stacking these head-specific representations across all attention heads yields the attention-weighted internal representation at layer l.

$$\tilde{\mathbf{H}}^l = \left[\tilde{\mathbf{h}}^l_1; \tilde{\mathbf{h}}^l_2; \cdots; \tilde{\mathbf{h}}^l_K\right] \in \mathbb{R}^{K\times d} \tag{3}$$

While the formulation applies to any Transformer layer, our empirical analysis shows that hallucination-discriminative signals are generally stronger in the middle layers. We therefore use the middle layer, $l = \lfloor L/2 \rfloor$, as a representative layer, providing a simple yet effective choice without requiring multi-layer aggregation. For simplicity, we omit the layer superscript hereafter and denote the resulting representation as $\tilde{\mathbf{H}} \in \mathbb{R}^{K\times d}$, which is used in all subsequent formulations.

4.2 Semantic Intensity for Hallucination Detection

Given the attention-weighted internal representation, we define its ℓ_2 norm as Semantic Intensity, which quantifies the overall magnitude of the resulting semantic representation. Since hallucinated and factual generations exhibit a clear difference in Semantic Intensity, Semantic Intensity can serve as a discriminative signal for hallucination detection.

Specifically, based on the middle-layer multi-head representation $\tilde{\mathbf{H}} \in \mathbb{R}^{K\times d}$ obtained above, we perform mean pooling over all K attention heads while retaining all representation dimensions to obtain an aggregated semantic representation:

$$\bar{\mathbf{h}} = \frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{h}}_k, \quad \bar{\mathbf{h}} \in \mathbb{R}^d \tag{4}$$

We then compute the ℓ_2 norm of this aggregated representation as the Semantic Intensity.

$$\|\bar{\mathbf{h}}\|_2 = \sqrt{\sum_{j=1}^{d}(\bar{h}_j)^2} \tag{5}$$

As hallucinated content exhibits systematically lower Semantic Intensity than factual generations, its reciprocal can thus serve as a preliminary hallucination score, with higher values indicating a greater likelihood of hallucination:

$$s_{\text{pre}} = \frac{1}{\|\bar{\mathbf{h}}\|_2 + \epsilon} \tag{6}$$

where ϵ is a small smoothing constant used to prevent numerical instability and is fixed throughout the paper. A higher value of s_{pre} indicates a lower Semantic Intensity of the model's internal representation, suggesting that the corresponding content is more likely to be hallucinated.

However, not all attention heads and representation dimensions contribute equally to the discriminative signal of Semantic Intensity. We introduce a selection strategy to identify the most hallucination-discriminative heads and dimensions, reducing redundant information and strengthening the effectiveness of Semantic Intensity for hallucination detection.

4.3 Discriminative Head and Dimension Selection

We propose a two-stage feature selection strategy based on Cohen's d effect size (Cohen, 2013) to identify the most discriminative attention heads and representation dimensions. Cohen's d quantifies the standardized difference between the means of two groups, providing a scale-invariant measure of the difference between factual and hallucinated generations. Feature selection is performed on the analysis set, and the selected attention heads and representation dimensions are subsequently fixed for hallucination detection.

Stage 1: Attention Head Selection. For each attention head k, we compute the absolute Cohen's d effect size based on the Semantic Intensity of factual and hallucinated samples:

$$\text{score}(k) = \frac{|\mu_k^+ - \mu_k^-|}{\sqrt{\frac{(n^+-1)(\sigma_k^+)^2+(n^--1)(\sigma_k^-)^2}{n^++n^--2}} + \epsilon}, \tag{7}$$

where μ_k^+ and μ_k^- denote the mean Semantic Intensity of factual and hallucinated samples, respectively, for the k-th attention head. σ_k^+ and σ_k^- denote the corresponding sample standard deviations, while n^+ and n^- denote the numbers of factual and hallucinated samples, respectively.

The score standardizes the difference between the class means using the pooled standard deviation, thereby accounting for both between-class separability and within-class dispersion. A higher score indicates a stronger capability of the corresponding head to discriminate between hallucinated and factual generations. After ranking all K attention heads in descending order of their effect sizes, we progressively construct candidate subsets by adding heads in this order. For each subset, we average the representations of the selected attention heads and compute the resulting Semantic Intensity for each sample, which is then evaluated using AUROC on the analysis set. The subset achieving the highest AUROC is selected as the final attention head set, denoted by $\mathcal{H}^* \subseteq \{1, \cdots, K\}$.

Stage 2: Dimension Selection Once the attention head set $\mathcal{H}^*$ has been determined, the attention-weighted representations produced by the selected heads are mean-pooled to obtain a unified sequence-level representation. For each representation dimension, we compute the absolute Cohen's d effect size based on its absolute activation values, which correspond to the ℓ_2 norm of the one-dimensional representation. The dimensions are then ranked in descending order of their effect sizes. Starting from the highest-ranked dimension, additional dimensions are progressively incorporated using the same greedy accumulation strategy as in head selection. In practice, the optimal performance is typically achieved with only a small number of top-ranked dimensions. The subset achieving the highest AUROC is selected as the final dimension set, denoted by $\mathcal{D}^* \subseteq \{1, \cdots, d\}$.

After the above two-stage selection process, the final hallucination detection score is defined as:

$$s_{\text{final}} = \frac{1}{\|\bar{\mathbf{h}}_{\mathcal{H}^*,\mathcal{D}^*}\|_2 + \epsilon} \tag{8}$$

where $\bar{\mathbf{h}}_{\mathcal{H}^*,\mathcal{D}^*}$ denotes the final internal representation obtained from the attention-weighted multi-head representation at the middle layer, by retaining only the heads corresponding to the head set $\mathcal{H}^*$, performing mean pooling over their attention-weighted representations, and then selecting the dimension set $\mathcal{D}^*$ from the pooled representation.

Compared with the preliminary score s_{pre}, s_{final} focuses on the most discriminative heads and dimensions, effectively filtering out redundant information and thereby enhancing the discriminative power of the detection signal.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets And Metrics

We conduct experiments on five open-domain QA benchmark datasets, spanning both single-hop and multi-hop reasoning tasks. NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023) are single-hop QA datasets, with PopQA focusing on long-tail entity knowledge, whereas HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020) require multi-hop reasoning. We use the development splits for NQ, TriviaQA, HotpotQA, and 2WikiMultiHopQA, and the full dataset for PopQA, which lacks an official split.

Test Set. For each dataset, the first 500 questions serve as the test set. The target LLM generates a response to each question, from which internal representations are extracted to compute the hallucination score. A judge model assesses response correctness as ground truth for evaluation.

Analysis Set. For each dataset, the target LLM generates responses to the questions following the test set, which are judged for correctness by the same judge model. As the resulting label proportions vary across datasets, we sequentially collect responses until 100 factual and 100 hallucinated samples are obtained per dataset, yielding a model-specific analysis set of 1,000 samples (500 factual, 500 hallucinated) across the five datasets. This set is used both for the preliminary analysis in Section 3 and for selecting discriminative attention heads and dimensions in our method, with no overlap with the test set.

We adopt AUROC as the evaluation metric, as it measures classification performance across varying decision thresholds and is widely used in hallucination detection.

5.1.2 Models

We evaluate all methods on five open-source LLMs: Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-4B-Instruct-2507 (Yang et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Phi-4-mini-Instruct (Abouelenin et al., 2025), covering multiple model families and 3B–8B parameters. We refer to them as Llama3.2-3B, Llama3.1-8B, Qwen3-4B, Mistral-7B, and Phi-4-mini, respectively. We further use Qwen2.5-72B-Instruct (Qwen et al., 2025) as the judge model to determine response correctness for both analysis-set construction and test evaluation, providing the ground-truth labels.

Table 1: AUROC comparison across LLMs and QA datasets. The left block reports results averaged over datasets for each model; the right block reports results averaged over models for each dataset.

Method	Per-Model Average					Per-Dataset Average				
	Llama3.2-3B	Llama3.1-8B	Qwen3-4B	Mistral-7B	Phi-4-mini	2WikiMultiHopQA	HotpotQA	PopQA	TriviaQA	NQ
LNPE	0.6196	**0.7551**	0.7024	0.6502	0.6789	0.6708	0.6798	0.6511	0.6972	0.7073
SCG	0.6013	0.6001	0.5843	0.5831	0.5449	0.6139	0.4969	0.6083	0.6039	0.5907
SE	0.5423	0.4794	0.5462	0.4693	0.4940	0.5695	0.4969	0.4164	0.5284	0.5202
SEU	0.5709	0.6457	0.6886	0.6554	0.6621	0.6737	0.6499	0.4897	0.7116	0.6979
PTrue	0.6414	0.4125	0.5505	0.6819	0.5495	0.6194	0.5280	0.5302	0.5693	0.5891
MIND	0.6249	0.6832	0.5895	0.5041	0.5306	0.5402	0.6078	0.6329	0.5824	0.5690
SAPLMA	0.6220	0.6218	0.5511	0.5495	0.5251	0.4956	0.5789	0.6623	0.5563	0.5762
INSIDE	0.5529	0.5896	0.5635	0.5423	0.5569	0.6113	0.5516	0.4723	0.5928	0.5772
HDIR	**0.6998**	0.7346	**0.7976**	**0.7235**	**0.6935**	**0.7065**	**0.6922**	**0.7501**	**0.7721**	**0.7283**

5.1.3 BASELINES

For output-level methods, we compare against LNPE (Malinin & Gales, 2020) as a representative confidence-based approach, along with several multi-sample consistency methods: the N-gram variant of SelfCheckGPT (SCG) (Manakul et al., 2023), Semantic Entropy (SE) (Kuhn et al., 2023), Semantic Embedding Uncertainty (SEU) (Grewal et al., 2024), and PTrue (Kadavath et al., 2022). For representation-based methods, we compare against SAPLMA (Azaria & Mitchell, 2023) and MIND (Su et al., 2024b). We further include INSIDE (Chen et al., 2024), which compares internal representations across multiple sampled generations.

5.1.4 IMPLEMENTATION DETAILS

For HDIR and other non-sampling-based baselines, we use deterministic generation with `do_sample=False`, while sampling-based baselines use a temperature of 0.8, top-p of 0.9, and 5 sampled responses. Training-based baselines are implemented following their original protocols. The random seed is fixed at 42, and the maximum number of newly generated tokens is set to 64. All experiments are conducted on NVIDIA A100 GPUs with 80GB of memory. The discriminative attention heads and dimensions selected by HDIR are provided in Appendix C.

5.2 MAIN RESULTS

Table 1 reports the average AUROC across five QA datasets for each LLM, and across five LLMs for each QA dataset, providing a per-model and per-dataset view of the compared methods' performance, respectively.

Among output-based methods, LNPE achieves the strongest overall performance, demonstrating that output-level confidence remains a highly effective signal for hallucination detection. SEU consistently outperforms SE across all five models and five datasets, suggesting that embedding-based uncertainty provides more discriminative signals than semantic entropy alone. However, both SEU and SE perform worst on PopQA, suggesting that semantic-similarity-based uncertainty estimation is less effective under long-tail knowledge distributions. SCG exhibits relatively stable overall performance among output-based baselines, but its detection performance remains moderate. PTrue exhibits a similarly narrow spread across datasets but much larger variation across models, suggesting that its performance is sensitive to model architecture.

Among representation-based methods, SAPLMA and MIND perform well on Llama-family models and achieve their strongest results on PopQA, but show greater variation across model architectures and datasets, suggesting less consistent transfer across the evaluated settings. INSIDE exhibits moderate overall performance but is more sensitive to dataset differences than to model architecture, potentially due to the additional variability introduced by multi-sampling.

In contrast, HDIR achieves strong and competitive performance across all models and datasets. These results show that Semantic Intensity, combined with the discriminative head and dimension selection strategy, consistently captures hallucination-discriminative features across diverse model architectures and QA domains, using a single set of selected heads and dimensions per model without dataset-specific adaptation or additional training. We additionally evaluate the robustness of HDIR to the choice of judge LLM, with results reported in Appendix D.

Table 2: Ablation study of head and dimension selection strategies with an alternative ranking criterion, reported as AUROC averaged over datasets.

Method	Llama3.2-3B	Llama3.1-8B	Qwen3-4B	Mistral-7B	Phi-4-mini
w/o Selection (All Heads + All Dims)	0.6496	0.6335	0.6051	0.6274	0.5323
w/ Head Selection (All Dims)	0.6837	0.7124	0.7368	0.7132	0.6983
w/ Head + Dim Selection (Normalized Gap)	0.6956	0.7210	**0.8044**	0.4291	**0.7408**
HDIR	**0.6998**	**0.7346**	0.7976	**0.7235**	0.6935

Table 3: Comparison of AUROC across head and dimension selection strategies under different analysis set compositions, averaged over datasets.

Method	#Samples	Llama3.2-3B	Llama3.1-8B	Qwen3-4B	Mistral-7B	Phi-4-mini
Same-Dataset	200	0.5305	0.5387	0.7821	0.7065	**0.7070**
Cross-Dataset	200	0.6441	0.6971	0.7163	0.6603	0.6561
Mixed-Dataset	200	0.6488	0.7316	0.7736	0.6676	0.6920
HDIR	1000	**0.6998**	**0.7346**	**0.7976**	**0.7235**	0.6935

5.3 Ablation Study

We conduct an ablation study to evaluate the proposed head and dimension selection strategy under four settings: (1) no selection, (2) head selection only, (3) head and dimension selection with Normalized Gap, and (4) HDIR. Here, Normalized Gap replaces Cohen's d effect size with the ratio of the inter-class mean difference to the average of the two classes' mean feature magnitudes. We further analyze the impact of layer-selection configurations and the number of selected attention heads and representation dimensions in Appendices E and F, respectively.

As shown in Table 2, head selection alone provides substantial improvements, with Phi-4-mini achieving a 0.166 AUROC gain over the no-selection setting, indicating that the selected attention heads contain more discriminative information for hallucination detection. Dimension selection further improves performance on most models. Compared with Normalized Gap, HDIR performs better on most models, with comparable results on the rest. Notably, Normalized Gap's performance drops sharply on Mistral-7B, whereas HDIR remains stable, suggesting greater robustness of our strategy.

5.4 Cross-Dataset Generalization of Selected Heads and Dimensions

This subsection investigates the cross-dataset generalization of the selected heads and dimensions, evaluating whether the selection strategy remains effective under different analysis-set construction strategies. Specifically, we compare four strategies for constructing the analysis set as follows: (1) Same-Dataset, using 200 samples from the target dataset; (2) Cross-Dataset, using 200 samples collected from the four non-target datasets (50 samples per dataset); (3) Mixed-Dataset, using 200 samples evenly sampled from all five datasets (40 samples per dataset, including the target dataset); and (4) HDIR, using the full 1,000-sample analysis set adopted in the main experiments. All analysis sets contain equal numbers of factual and hallucinated samples.

As shown in Table 3, Same-Dataset is highly model-dependent: although it has access to samples from the target dataset, its performance degrades substantially across the Llama family. In contrast, Cross-Dataset achieves consistent performance across models without using target-dataset samples, indicating that features selected from non-target datasets transfer effectively across datasets and architectures. Mixed-Dataset further benefits from the increased diversity of the analysis data, while HDIR uses the larger, full analysis set and achieves the best overall performance. These results suggest that the discriminative head and dimension selection strategy identifies robust, dataset-agnostic hallucination-discriminative features rather than dataset-specific patterns. We also examine the cross-model transferability of the selected heads and dimensions in Appendix G.

5.5 Combining with LNPE

The HDIR score quantifies Semantic Intensity computed from internal representations, while LNPE reflects output-level uncertainty. These signals provide complementary information, motivating

Table 4: Comparison of HDIR, LNPE, and different fusion strategies in terms of AUROC, averaged over datasets.

Method	Llama3.2-3B	Llama3.1-8B	Qwen3-4B	Mistral-7B	Phi-4-mini
LNPE	0.6196	0.7551	0.7024	0.6502	0.6789
HDIR	0.6998	0.7346	0.7976	0.7235	0.6935
Product Fusion	0.6419	0.7885	0.7659	0.6967	0.7315
Dynamic Weighted Fusion	0.7010	0.7664	0.8228	**0.7345**	0.7451
Complementary Product Fusion	**0.7180**	**0.7889**	**0.8305**	0.7293	**0.7474**

their fusion for improved hallucination detection. Let s_{HDIR} and s_{LNPE} denote the scores produced by HDIR and LNPE, respectively. We consider three fusion strategies: Product Fusion, Dynamic Weighted Fusion, and Complementary Product Fusion.

Product Fusion. The two scores are directly multiplied:

$$s_{\text{fusion}} = s_{\text{HDIR}} \cdot s_{\text{LNPE}} \tag{9}$$

Dynamic Weighted Fusion. We first record the minimum and maximum values of s_{HDIR} and s_{LNPE} on the analysis set and use them to normalize both scores to $[0, 1]$ during inference, yielding $\hat{s}_{\text{HDIR}}$ and $\hat{s}_{\text{LNPE}}$, respectively. Values outside the observed range are clipped to $[0, 1]$. Since values closer to 0.5 indicate lower confidence, we define the confidence of each signal as its distance from 0.5: $c_{\text{HDIR}} = |\hat{s}_{\text{HDIR}} - 0.5|$ and $c_{\text{LNPE}} = |\hat{s}_{\text{LNPE}} - 0.5|$. These confidence measures are normalized into fusion weights:

$$w_{\text{HDIR}} = c_{\text{HDIR}}/(c_{\text{HDIR}} + c_{\text{LNPE}} + \epsilon), \quad w_{\text{LNPE}} = c_{\text{LNPE}}/(c_{\text{HDIR}} + c_{\text{LNPE}} + \epsilon) \tag{10}$$

The final fusion score is the confidence-weighted sum of the normalized scores:

$$s_{\text{fusion}} = w_{\text{HDIR}} \cdot \hat{s}_{\text{HDIR}} + w_{\text{LNPE}} \cdot \hat{s}_{\text{LNPE}} \tag{11}$$

Complementary Product Fusion. Using the same normalized scores, Complementary Product Fusion combines the factual and hallucination supports of the two signals through complementary products, followed by normalization:

$$s_{\text{fusion}} = \frac{\hat{s}_{\text{HDIR}} \cdot \hat{s}_{\text{LNPE}}}{\hat{s}_{\text{HDIR}} \cdot \hat{s}_{\text{LNPE}} + (1 - \hat{s}_{\text{HDIR}})(1 - \hat{s}_{\text{LNPE}}) + \epsilon} \tag{12}$$

Table 4 compares the three fusion strategies against HDIR and LNPE alone. As shown in the table, Product Fusion consistently improves over LNPE but shows inconsistent performance relative to HDIR across models, suggesting that a simple product of the two scores cannot fully exploit the complementary information they carry. In contrast, Dynamic Weighted Fusion consistently outperforms both individual scores across all models by adaptively weighting the two signals. Complementary Product Fusion achieves the best overall performance, further confirming the complementarity between HDIR and LNPE. These results confirm that HDIR captures complementary information for output-level uncertainty estimation, making it effective not only as a standalone detector but also as a complementary component for further improving hallucination detection.

6 Conclusion

In this paper, we propose HDIR, a training-free hallucination detection method that exploits discriminative information in LLM internal representations. HDIR introduces Semantic Intensity to quantify differences in attention-weighted representations, revealing an effective signal for distinguishing factual and hallucinated generations. It further identifies discriminative attention heads and dimensions via Cohen's d-based selection, focusing Semantic Intensity on informative components. Extensive experiments show that HDIR achieves the best or highly competitive performance across most settings and generalizes well across datasets. Moreover, HDIR provides complementary information to the output-based uncertainty signal of LNPE, with their combination further improving detection performance. Overall, our findings highlight the value of analyzing LLM internal representations before using them for hallucination detection. Such analysis reveals that hallucination-discriminative information is concentrated in a small subset of internal components, which can be efficiently identified and leveraged for training-free detection.

7 AI USE STATEMENT

In this work, we have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing key elements for proving mathematical claims, assisting in writing proofs, proposing or refining hypotheses, designing research methods or experiments, providing feedback on research methods or experiments, implementing methods, assisting in translation, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and these tasks were not applicable to this work.

Additionally, we used generative AI tools for editing and polishing the writing of this manuscript to improve its clarity, readability, and language. We have reviewed all AI-assisted work and carefully checked the edited text for accuracy and to ensure that it faithfully reflects our intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Samir Abdaljalil, Parichit Sharma, Erchin Serpedin, Rachad Atat, and Hasan Kurban. Sindex: Semantic inconsistency index for hallucination detection in llms. *IEEE Open Journal of the Computer Society*, 2026.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61:100970, 2026.

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

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

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

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yashvir S Grewal, Edwin V Bonilla, and Thang D Bui. Improving uncertainty quantification in large language models via semantic embeddings. *arXiv preprint arXiv:2410.22685*, 2024.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yifan Wu, YK Li, et al. Deepseek-coder: when the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 315–334, 2024.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Seongmin Lee, Hsiang Hsu, and Chun-Fu Chen. Llm hallucination reasoning with zero-shot knowledge test. In *Workshop on Socially Responsible Language Modelling Research*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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

Fang Liu, Yang Liu, Lin Shi, Zhen Yang, Li Zhang, Xiaoli Lian, Zhongqi Li, and Yuchi Ma. Beyond functional correctness: Exploring hallucinations in llm-generated code. *IEEE Transactions on Software Engineering*, 2026.

Ming Lu, Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yi Hua, Yushan Cai, Yuanbo Xu, et al. Loki's dance of illusions: A comprehensive survey of hallucination in large language models. *IEEE Transactions on Computational Social Systems*, 2026.

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. Semantic energy: Detecting llm hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*, 2025.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. 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*, pp. 9004–9017, 2023.

Muyuan Niu and Cong Tian. Coopkg: An academic knowledge graph for question answering systems. In *International Conference on Database Systems for Advanced Applications*, pp. 436–445. Springer, 2025.

Muyuan Niu, Jie Su, and Cong Tian. Ua-rag: Uncertainty-aware dynamic retrieval-augmented generation. *Neurocomputing*, pp. 134357, 2026.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

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

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12991–13013, 2024a.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, 2024b.

Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. Judge: Benchmarking judgment document generation for chinese legal system. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3573–3583, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Changyue Wang, Weihang Su, Hu Yiran, Qingyao Ai, Yueyue Wu, Cheng Luo, Yiqun Liu, Min Zhang, and Shaoping Ma. Lekube: A legal knowledge update benchmark. *arXiv preprint arXiv:2407.14192*, 2024.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21806–21818, 2025a.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. Detecting hallucination in large language models through deep internal representation analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 8357–8365, 2025b.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 915–932, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

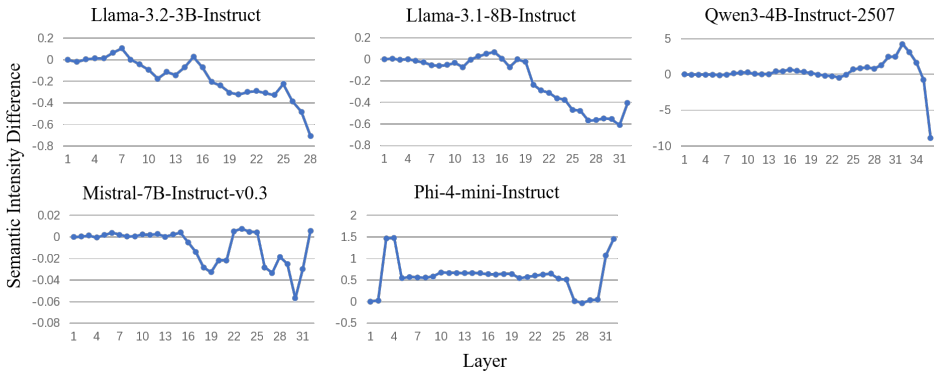


Figure 3: Layer-wise difference in mean Semantic Intensity (hallucinated - factual) under **Last-Token Extraction** across five LLMs.

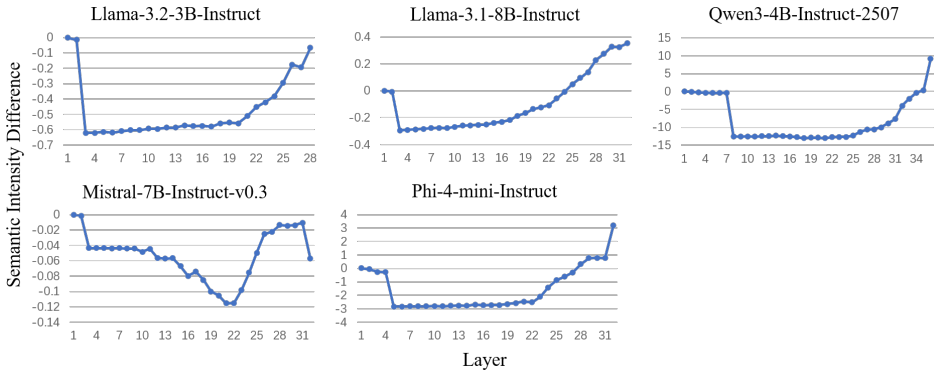


Figure 4: Layer-wise difference in mean Semantic Intensity (hallucinated - factual) under **Mean Pooling Aggregation** across five LLMs.

A Alternative Representation Construction

To further support our choice of attention-weighted aggregation, we examine two alternative representation construction approaches commonly used in prior work: (1) **Last-Token Extraction**, which takes the hidden state corresponding to the last token position in the sequence as the sequence-level representation, i.e., $\mathbf{h} = \mathbf{H}_{-1}$; and (2) **Mean Pooling Aggregation**, which averages the hidden states of all tokens along the sequence dimension, i.e., $\mathbf{h} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_t$, where T denotes the sequence length. For both approaches, we compute the layer-wise difference in mean Semantic Intensity between hallucinated and factual samples on the analysis set. As shown in Figure 3 and Figure 4, neither Last-Token Extraction nor Mean Pooling Aggregation exhibits a consistent and sufficiently discriminative pattern between hallucinated and factual samples across layers and models.

Last-Token Extraction shows that, for Llama-family models and Qwen3-4B, the Semantic Intensity of hallucinated samples becomes progressively lower than that of factual samples as the layer index increases. Nevertheless, the Semantic Intensity values of the two groups substantially overlap across many layers and models, limiting their separability. In addition, Mistral-7B and Phi-4-mini exhibit trends inconsistent with those of the other models, making it difficult to identify a consistent hallucination-related pattern.

Consistent with the trend observed in our main motivation analysis, Mean Pooling Aggregation generally exhibits a decreasing-then-increasing pattern in the Semantic Intensity gap. This observation further supports the presence of hallucination-discriminative signals in internal representations. However, the gap between the Semantic Intensity of factual and hallucinated samples remains relatively small despite the similar layer-wise trend, motivating the use of attention weighting to better

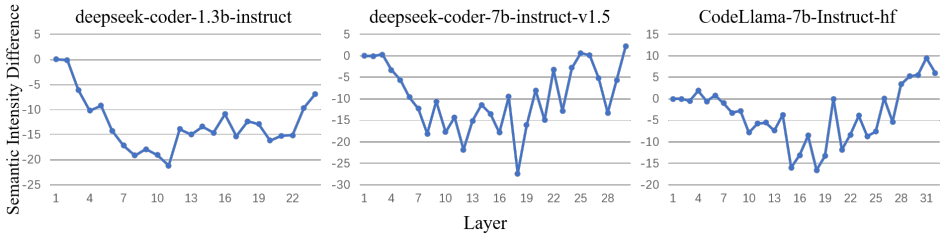


Figure 5: Layer-wise difference in mean Semantic Intensity (hallucinated - factual) across code LLMs.

Table 5: Selected attention heads and dimensions (0-indexed), ranked by Cohen's d, for each backbone LLM.

Method	Head Indices	Dimension Indices
Llama3.2-3B	3, 23, 14, 18	70, 953, 2607
Llama3.1-8B	26, 4	3947
Qwen3-4B	10, 6, 5, 30	1029, 122, 2238, 979, 2047
Mistral-7B	1, 24, 0	1504, 3287, 746
Phi-4-mini	23, 6, 20, 3	1699, 1582

highlight such signals. Moreover, the layer at which the Semantic Intensity gap reaches its minimum varies across models: it tends to occur in earlier middle layers for most models, but later for Mistral-7B.

In contrast, our attention-weighted aggregation produces a much clearer and more consistent separation, suggesting that it yields a more discriminative semantic representation space for distinguishing hallucinated from factual samples.

B Semantic Intensity for Code Hallucination

To examine whether the observed Semantic Intensity signal extends beyond open-domain QA, we conduct an additional analysis on the code hallucination dataset introduced by Liu et al. (Liu et al., 2026). The dataset builds on CoderEval and HumanEval and contains model-generated code responses together with human-annotated hallucination labels. Specifically, it includes 460 CoderEval problems (230 Java and 230 Python) and 164 HumanEval Python problems, with generations from multiple models. We focus on the three open-source models with accessible internal representations: DeepSeek-Coder-1.3B-Instruct (Guo et al., 2024), DeepSeek-Coder-7B-Instruct-v1.5, and CodeLlama-7B-Instruct-hf (Roziere et al., 2023).

For each model, we use its own generated responses to these problems and select 100 hallucinated and 100 factual samples according to the original human annotations. We then compute Semantic Intensity at each Transformer layer and calculate the difference in mean Semantic Intensity between hallucinated and factual generations. The resulting layer-wise patterns are shown in Figure 5.

Consistent with our observations in open-domain question answering, hallucinated code generations exhibit lower Semantic Intensity than non-hallucinated generations, with the difference being more pronounced in intermediate Transformer layers. The consistent patterns observed across open-domain QA and code generation further support the effectiveness of our aggregation approach and Semantic Intensity in capturing internal signals associated with hallucination.

C Feature Selection

For each backbone LLM, the discriminative attention heads and dimensions are independently identified using its corresponding analysis set. Attention heads are selected via greedy search over the full head space, while dimension selection is restricted to the top-200 candidates ranked by Cohen's

Table 6: AUROC comparison across LLMs and QA datasets using Llama-3.1-70B-Instruct as the judge model. The left block reports results averaged over datasets for each model, while the right block reports results averaged over models for each dataset.

Method	Per-Model Average					Per-Dataset Average				
	Llama3.2-3B	Llama3.1-8B	Qwen3-4B	Mistral-7B	Phi-4-mini	2WikiMultiHopQA	HotpotQA	PopQA	TriviaQA	NQ
LNPE	0.5470	**0.7321**	0.6749	0.6499	0.6592	0.6497	0.6695	0.5547	0.7004	0.6888
SCG	0.5078	0.5363	0.5744	0.5705	0.5385	0.6174	0.4654	0.5233	0.5810	0.5405
SE	0.5204	0.4192	0.5827	0.4619	0.5001	0.5616	0.4897	0.4236	0.5074	0.5019
SEU	0.5461	0.6640	0.6742	0.6826	0.6892	0.6639	0.6350	0.5297	0.7184	**0.7091**
PTrue	0.6754	0.3671	0.5594	0.7178	0.5657	0.6392	0.5220	0.5721	0.5655	0.5865
MIND	0.7320	0.7210	0.6197	0.4624	0.5066	0.5275	0.6176	0.7126	0.6248	0.5592
SAPLMA	0.6654	0.6635	0.5717	0.5524	0.5420	0.5034	0.5927	0.7091	0.6024	0.5873
INSIDE	0.4737	0.5365	0.5724	0.5390	0.5761	0.6115	0.5282	0.4342	0.5755	0.5482
HDIR	**0.7701**	0.7222	**0.7701**	**0.7434**	**0.7609**	**0.7276**	**0.6971**	**0.8658**	**0.7719**	0.7044

Table 7: Selected attention heads and dimensions (0-indexed), ranked by Cohen's d, for each backbone LLM using Llama-3.1-70B-Instruct as the judge LLM.

Method	Head Indices	Dimension Indices
Llama3.2-3B	14, 23	70, 953, 1973, 860, 430, 1149, 2234, 1532, 1486, 699, 125
Llama3.1-8B	26	3524, 4081
Qwen3-4B	10, 30, 28, 6, 5, 11, 29	1029, 122, 1909
Mistral-7B	24, 1, 11	1504, 3287, 2669
Phi-4-mini	23	1010, 2349, 409, 896, 848

d effect size. The selected features are then fixed and used for evaluation across all test datasets. The selected head and dimension indices for each backbone are reported in Table 5.

As shown in the table, only a small number of heads and dimensions are needed to achieve optimal performance on the analysis set, indicating that hallucination-discriminative signals are highly concentrated rather than dispersed across the entire representation space.

D Robustness to an Alternative Judge LLM

As shown in Table 6, replacing the judge LLM leads to variation in AUROC values relative to those reported in Table 1, as the judge LLM affects both analysis set construction and evaluation label assignment. Nevertheless, the performance patterns of different methods remain consistent across the two settings, with HDIR achieving the best overall performance. These results show that HDIR is robust to the choice of judge LLM.

We additionally report the selected heads and dimensions for each model using the same analysis set, but with its labels generated by Llama-3.1-70B-Instruct as the judge LLM, as shown in Table 7. Comparing the results with those in Table 5, we observe a consistent pattern: models with fewer selected heads tend to have more selected dimensions, and vice versa. Moreover, the smaller selected head sets are largely contained within the corresponding larger sets, while the selected dimensions exhibit substantial overlap across the two judges without forming a strict subset relationship.

These findings suggest that hallucination-discriminative information occupies a relatively constrained representation space, where attention heads and representation dimensions can provide complementary ways of capturing this information. Moreover, the selected heads and dimensions capture a stable set of discriminative components intrinsic to each target model, largely independent of the specific judge LLM used to construct the analysis set, further supporting the robustness of the proposed strategy.

E Analysis of Layer Configuration

To compare the impact of different layer selection strategies on detection performance, we design the following set of controlled experiments. For multi-layer representations, Semantic Intensity is computed for each layer and then averaged.

Table 8: AUROC comparison under different layer selections with and without feature selection, averaged over datasets.

Method	Llama3.2-3B	Llama3.1-8B	Qwen3-4B	Mistral-7B	Phi-4-mini	AVG ± STD
Pre-layer-all	0.6373	0.6770	0.6672	0.6131	0.5784	0.6346±**0.0360**
Pre-layer-2/3	0.6812	0.6771	0.6582	0.6248	0.5497	**0.6382**±0.0485
Pre-layer-1/3	0.6696	0.6680	0.6458	0.6271	0.5700	0.6361±0.0366
Pre-layer-middle	0.6496	0.6335	0.6051	0.6274	0.5323	0.6096±0.0412
HDIR-layer-all	0.6744	0.7662	0.7931	0.6867	0.6871	0.7215±0.0485
HDIR-layer-2/3	**0.7210**	**0.7675**	0.7932	0.7030	0.6489	0.7267±0.0505
HDIR-layer-1/3	0.7012	0.7657	**0.8126**	0.7018	0.6729	**0.7308**±0.0509
HDIR	0.6998	0.7346	0.7976	**0.7235**	**0.6935**	0.7298±**0.0371**

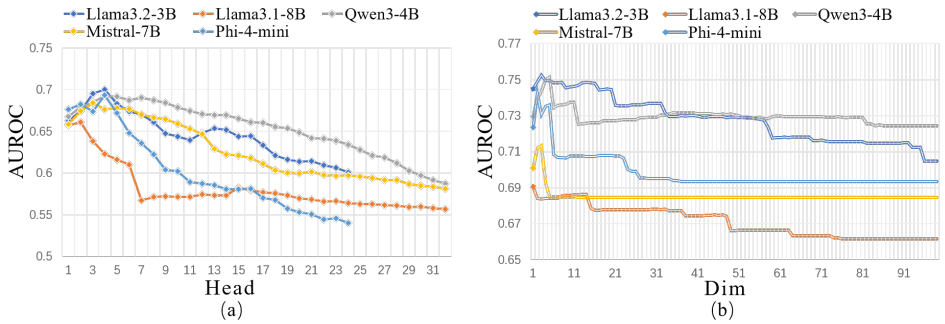


Figure 6: AUROC on the analysis set as a function of the number of selected (a) attention heads and (b) hidden dimensions across five LLMs.

The four variants prefixed with "Pre" correspond to settings without attention head and dimension selection, listed from top to bottom as follows: all layers, the middle two-thirds of layers, the middle one-third of layers, and the middle layer only. The four variants prefixed with "HDIR" adopt the same set of layer selection strategies, but further incorporate discriminative head and dimension selection on top of each respective layer configuration.

As shown in Table 8, among the four "Pre" variants, multi-layer aggregation generally outperforms single-layer representations for the preliminary score, indicating that multi-layer averaging can mitigate the influence of layer-specific noise on the resulting representation. With head and dimension selection, the single middle-layer strategy matches multi-layer aggregation with lower computational cost and greater stability. We attribute this to the mixing of representations across layers in multi-layer averaging. In contrast, single-layer representations preserve layer-specific information, allowing the selected features to be clearly traced to a specific layer. This facilitates stable and discriminative feature selection, while achieving performance comparable to that of multi-layer aggregation.

This suggests that the proposed selection strategy can effectively identify hallucination-discriminative features from a single intermediate layer, offering a simpler and more interpretable alternative to multi-layer aggregation.

F Analysis of Head and Dimension Selection

This section analyzes how detection performance varies with the number of selected heads and dimensions on the analysis set through greedy accumulation curves.

Figure 6(a) shows how AUROC varies across different models as attention heads are progressively added in descending order of their Cohen's d effect size scores. AUROC rises sharply and reaches its peak with only a small number of selected heads, confirming that effect-size-based ranking effectively identifies the most informative heads for hallucination detection. However, once the number of selected heads exceeds a certain threshold, AUROC begins to decline, indicating that not all heads

Table 9: Average AUROC across the datasets after replacing the selected attention heads and dimensions with the corresponding components selected from other models. Model-specific results are included for comparison.

Model	Source	AUROC
Llama3.2-3B	Llama3.2-3B	0.6998
	Phi-4-mini	0.4117
Phi-4-mini	Phi-4-mini	0.6935
	Llama3.2-3B	0.5334
Llama3.1-8B	Llama3.1-8B	0.7346
	Mistral-7B	0.6116
Mistral-7B	Mistral-7B	0.7235
	Llama3.1-8B	0.5043

contribute equally to distinguishing hallucinated from factual outputs. Heads with lower effect-size scores tend to introduce redundant or task-irrelevant information, which in turn degrades the model's overall discriminative capability.

Based on the optimal number of attention heads identified for each model, Figure 6(b) presents the greedy accumulation curves for representation dimensions, ranked by their Cohen's d effect size scores. Similar to head selection, the best performance is achieved with only a small subset of dimensions. Llama3.1-8B reaches its peak AUROC with a single dimension, while other models achieve optimal performance with only a few dimensions, followed by a gradual decline as more dimensions are added. These results show that hallucination-discriminative signals are concentrated in a small number of representation dimensions even within the selected attention heads.

G Cross-Model Transferability of Selected Heads and Dimensions

This experiment examines whether the identified discriminative features are transferable across different models.

Among the LLMs evaluated in this study, Llama-3.2-3B and Phi-4-mini share the same number of attention heads and representation dimensions, as do Llama-3.1-8B and Mistral-7B. We exchange the selected heads and dimensions between each pair and evaluate their hallucination detection performance, comparing the results with those obtained using model-specific selections.

As shown in Table 9, the *Source* column specifies the model from which the selected attention heads and representation dimensions are obtained, and each reported score represents the average AUROC across the five test datasets. Compared with the model-specific selections, using heads and dimensions selected from another model leads to a substantial performance degradation. This indicates that hallucination-discriminative heads and dimensions are largely model-specific and are not directly transferable across LLMs, even when they share the same number of heads and dimensions. These results further support our model-specific component selection strategy, showing that effective hallucination-discriminative heads and dimensions are model-specific, while the selection strategy itself generalizes across diverse LLM architectures.

H Limitations and Future Work

While HDIR demonstrates strong performance across diverse models and datasets, several limitations remain. First, our method relies on a small labeled analysis set to determine the number of discriminative attention heads and dimensions to retain. Although this process is far less costly than training a classifier, it still requires a limited amount of labeled data and a judge model for annotation. Second, although HDIR can be applied across different LLM architectures, the resulting hallucination-discriminative heads and representation dimensions are model-specific and cannot be directly transferred across models. When applying HDIR to a new model, these discriminative

features therefore need to be re-identified, although the process remains lightweight and requires no additional training.

For future work, we plan to explore unsupervised strategies for head and dimension selection, potentially reducing reliance on labeled analysis data and external judge models. We also aim to investigate more efficient feature selection across LLM architectures, thereby reducing the adaptation cost of applying HDIR to new models. Motivated by the consistent Semantic Intensity patterns observed in our code hallucination analysis, we plan to investigate the applicability of HDIR to code hallucination detection and other generation tasks.