

Do LLMs Really Know What They Don't Know? Internal States Mainly Reflect Knowledge Recall Rather Than Truthfulness

Chi Seng Cheang¹ Hou Pong Chan² Wenxuan Zhang³ Yang Deng^{1*}

¹Singapore Management University ²DAMO Academy, Alibaba Group


³Singapore University of Technology and Design

cs.cheang.2025@phdcs.smu.edu.sg, houpong.chan@alibaba-inc.com

wxzhang@sutd.edu.sg, ydeng@smu.edu.sg

Abstract

Recent work suggests that LLMs “know what they don’t know”, positing that hallucinated and factually correct outputs arise from distinct internal processes and can therefore be distinguished using internal signals. However, hallucinations have multifaceted causes: beyond simple knowledge gaps, they can emerge from training incentives that encourage models to exploit statistical shortcuts or spurious associations learned during pretraining. In this paper, we argue that when LLMs rely on such learned associations to produce hallucinations, their internal processes are mechanistically similar to those of factual recall, as both stem from strong statistical correlations encoded in the model’s parameters. To verify this, we propose a novel taxonomy categorizing hallucinations into Unassociated Hallucinations (UHs), where outputs lack parametric grounding, and Associated Hallucinations (AHs), which are driven by spurious associations. Through mechanistic analysis, we compare their computational processes and hidden-state geometries with factually correct outputs. Our results show that hidden states primarily reflect whether the model is *recalling parametric knowledge* rather than the truthfulness of the output itself. Consequently, AHs exhibit hidden-state geometries that largely overlap with factual outputs, rendering standard detection methods ineffective. In contrast, UHs exhibit distinctive, clustered representations that facilitate reliable detection.

 <https://github.com/AndyCheang/knowledge-recall-vs-truthfulness>

1 Introduction

Large language models (LLMs) are notorious for producing hallucinations (Zhang et al., 2023b; Huang et al., 2025), where generated outputs appear plausible but are factually incorrect. Recent

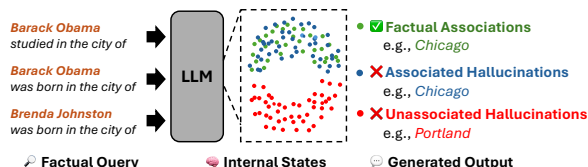


Figure 1: Illustration of three categories of knowledge. Associated hallucinations follow similar internal knowledge recall processes with factual associations, while unassociated hallucinations arise when the model’s output is detached from the input.

research suggests that LLMs’ internal states contain signals correlated with factual correctness, enabling hallucination detection using internal representations, such as residual streams (Azaria and Mitchell, 2023; Gottesman and Geva, 2024), attention weights (Yüksekgönül et al., 2024), and output token logits (Orgad et al., 2025; Varshney et al., 2023a). However, since LLMs are not explicitly trained to represent truthfulness, it remains unclear whether these signals genuinely reflect truthfulness or instead capture other confounding factors. Understanding what these signals actually encode is critical for reliably deploying LLMs in real-world applications.

In this work, we argue that such internal signals primarily reflect the model’s internal process of recalling parametric knowledge, rather than truthfulness itself. Consequently, these signals can reliably detect hallucinations only when hallucinated and factually correct outputs are produced by distinct internal mechanisms. For example, as shown in Figure 1, given the prompt “Brenda Johnston was born in the city of”, a model that lacks the relevant factual knowledge about the subject (“Brenda Johnston”) may hallucinate a completion such as “Portland”. In contrast, given the prompt “Barack Obama studied in the city of”, the model can leverage knowledge encoded about the subject (“Barack Obama”) to produce the factually correct output “Chicago”. These two cases are likely supported

*Corresponding author.

by different internal mechanisms: the former lacks knowledge about the subject entity, while the latter relies on encoded knowledge relevant to the queried subject. As a result, internal representations can reflect this difference in how the model processes the subject entity, enabling these cases to be distinguished.

However, hallucinations do not always arise from missing knowledge. When a model exploits learned statistical shortcuts or spurious correlations (Lin et al., 2022b; Kang and Choi, 2023; Cheang et al., 2023), the resulting hallucinations may be produced through mechanisms similar to those underlying factual recall. For instance, “*Barack Obama*” frequently co-occurs with “*Chicago*” in the model’s pre-training corpora. The model can leverage this statistical association to produce a factually correct output (e.g., “*Barack Obama studied in the city of Chicago.*”), but it may also leverage the same association to produce a hallucinated response (e.g., “*Barack Obama was born in the city of Chicago.*”). In both cases, the model relies on the same encoded statistical association involving the subject entity. Consequently, the resulting internal representations may not provide reliable signals to distinguish hallucinated outputs from factual ones, limiting the effectiveness of existing representation-based hallucination detection methods.

Based on this observation, we hypothesize that the effectiveness of representation-based hallucination detection depends on how the model leverages its parametric knowledge when producing a response, particularly whether the generated output is driven by learned associations involving the subject entity. To investigate this hypothesis, we go beyond labeling outputs solely by factual correctness and instead categorize them according to their relationship with the subject entity through a causal intervention. Specifically, we label factually correct outputs as **Factual Associations (FAs)**. For outputs that are factually incorrect, we further classify them as **Unassociated Hallucinations (UHs)**, whose outputs lack strong learned associations with the subject entity, and **Associated Hallucinations (AHs)**, which are driven by strong but spurious associations involving the subject entity.

Using this taxonomy, we conduct mechanistic and empirical analyses of these knowledge categories, yielding three key observations: First, **AHs and FAs share highly similar internal processes and representational geometries**. Building on the analytical framework of Geva et al. (2023),

we examine the internal mechanisms underlying model predictions by tracing how information propagates across layers and token positions during inference. We observe that because AHs and FAs are both driven by learned associations with the subject, their hidden state representations overlap in the hidden space. In contrast, UHs lack this reliance on subject associations and are instead generated through a different internal process, allowing them to remain more separable in the representation space.

Second, **existing hallucination detection methods struggle to distinguish AHs from FAs**. Since these methods rely on internal states that primarily reflect the process of knowledge recall rather than truthfulness, their performance degrades significantly for AH samples (AUROC \approx 0.48–0.69 for LLaMA). However, UHs are more reliably detected (AUROC \approx 0.86–0.93) due to their more separable representational geometry.

Third, **representational overlap constrains the effectiveness of refusal tuning for AHs**. We compare tuning effectiveness under two settings: (i) training the model to refuse AHs, and (ii) training the model to refuse UHs. In both settings, the model is trained to maintain its original factual responses for FAs. Because UH representations are more separable from FAs, the model can successfully learn distinct generative behaviors, achieving an 82% refusal rate on UH samples. Conversely, because AH representations overlap substantially with FAs, the model struggles to differentiate them to learn refusal behaviors, resulting in a refusal rate of only 33% for AH samples.

2 Related Work

Existing hallucination detection methods can be broadly categorized into two types: *representation-based* and *confidence-based*. **Representation-based methods** assume that an LLM’s internal hidden states can reflect the correctness of its generated responses. These approaches train a classifier (often a linear probe) using the hidden states from a set of labeled correct/incorrect responses to predict whether a new response is hallucinatory (Li et al., 2023; Azaria and Mitchell, 2023; Su et al., 2024; Ji et al., 2024; Chen et al., 2024; Ni et al., 2025; Xiao et al., 2025). **Confidence-based methods**, in contrast, assume that a lower confidence during the generation led to a higher probability of hallucination. These methods quantify uncer-

tainty through various signals, including: (i) token-level output probabilities (Guerreiro et al., 2023; Varshney et al., 2023a; Orgad et al., 2025); (ii) directly querying the LLM to verbalize its own confidence (Lin et al., 2022a; Tian et al., 2023; Xiong et al., 2024; Yang et al., 2024b; Ni et al., 2024; Zhao et al., 2024); or (iii) measuring the semantic consistency across multiple outputs sampled from the same prompt (Manakul et al., 2023; Kuhn et al., 2023; Zhang et al., 2023a; Ding et al., 2024). A response is typically flagged as a hallucination if its associated confidence metric falls below a pre-determined threshold.

However, a growing body of work reveals a critical limitation: even state-of-the-art LLMs are poorly calibrated, meaning their expressed confidence often fails to align with the factual accuracy of their generations (Kapoor et al., 2024; Xiong et al., 2024; Tian et al., 2023). This miscalibration limits the effectiveness of confidence-based detectors and raises a fundamental question about the extent of LLMs’ self-awareness of their knowledge boundary, *i.e.*, whether they can reliably “know what they don’t know” (Yin et al., 2023; Li et al., 2025). Despite recognizing this problem, prior work does not provide a mechanistic explanation for its occurrence. To this end, our work addresses this explanatory gap by employing mechanistic interpretability techniques to trace the internal computations underlying knowledge recall within LLMs.

3 Dataset Construction

In this section, we outline our dataset construction for mechanistic and empirical analyses under two conditions: hallucinations produced with and without leveraging the learned associations related to the subject entity. Given an input query q , the ground-truth answer y , and the model’s response \hat{y} , standard evaluation of hallucination detection methods typically assigns a factual correctness label by comparing \hat{y} with y . To study hallucinations produced through different internal mechanisms, we go beyond factual correctness: for each hallucinatory sample, we perform a causal intervention to estimate its reliance on learned subject associations and categorize it accordingly.

3.1 Data Collection

Factual Query Prompt Creation We focus on a knowledge-based question answering setting,

where each example corresponds to a knowledge triple (subject, relation, object) (s, r, o) . To construct factual query prompts, we first collect knowledge triples from Wikidata (Vrandečić and Krötzsch, 2014). Each (s, r) pair is then converted into a cloze-style factual query q using a hand-crafted prompt template for each relation r . The corresponding object o is treated as the ground-truth answer y . To ensure a well-defined evaluation setting, we follow Gekhman et al. (2025) and select only relations for which the correct object is objectively verifiable. Details on relation selection and prompt templates are provided in Appendix A.1.

Generating Model Responses For each query, we prompt LLMs to generate a response \hat{y} using greedy decoding. We conduct experiments on two widely adopted open-source LLMs: LLaMA-3 (Dubey et al., 2024) and Mistral-v0.3 (Jiang et al., 2023). Due to space constraints, full implementation details are provided in Appendix A.2.

3.2 Categorization of Knowledge

We categorize each response based on two criteria: (1) *factual correctness* and (2) *reliance on subject representations*. Each sample is then categorized into one of the following categories:

- **Factual Associations (FA)** refer to factual knowledge that is reliably stored in the parameters or internal states of an LLM and can be recalled to produce correct, verifiable outputs.
- **Associated Hallucinations (AH)** refer to non-factual content produced when an LLM relies on input-triggered parametric associations.
- **Unassociated Hallucinations (UH)** refer to non-factual content produced without reliance on parametric associations to the input.

3.3 Labeling Procedure

We detail the labeling procedure as follows:

Factual Correctness We assess the factual correctness of model responses following the framework proposed by Gekhman et al. (2025) (details in Appendix A.3). If a model’s response \hat{y} matches the ground-truth answer y , it is labeled as a **Factual Association (FA)**. Otherwise, the prediction is considered a hallucination and is further categorized based on its reliance on subject representations.

Subject Representation Reliance Geva et al. (2023) analyzes LLMs’ internal activations to characterize mechanisms underlying factual genera-

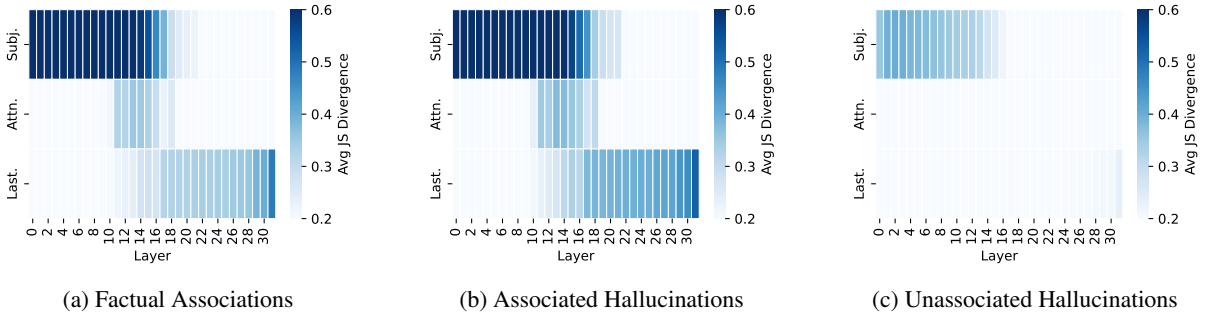


Figure 2: Effect of interventions across layers of LLaMA-3-8B. The heatmap shows JS divergence between the output distribution before and after intervention. Darker color indicates that the intervened hidden states are more causally influential on the model’s predictions. Top row: patching representations of subject tokens. Middle row: blocking attention flow from subject to the last token. Bottom row: patching representations of the last token.

	LLaMA-3-8B	Mistral-7B-v0.3
Factual Association	3,506	3,354
Associated Hallucination	1,406	1,284
Unassociated Hallucination	7,381	7,655
Total	12,293	12,293

Table 1: Dataset statistics across categories.

tion. Given a factual query prompt (e.g., “Barack Obama was born in the city of”), a model typically follows three steps to produce the correct completion (e.g., “Honolulu”). First, early layers of LLMs construct enriched subject representations encoding attributes specific to the subject (e.g., “Barack Obama”). Second, attention modules in middle layers propagate this subject information to the prediction position (i.e., the last token). Finally, upper layers extract the relevant attribute (e.g., “Honolulu”) from the propagated information to produce the next token.

While this mechanism has been established for factual generation, the internal processes underlying hallucinations remain less understood. We hypothesize that hidden states primarily reflect *how* the model leverages its parametric knowledge, rather than whether the output is factually correct. Consequently, hallucinations produced through statistical shortcuts or spurious correlations may follow internal processes similar to factual recall.

To test this hypothesis, we categorize hallucinatory outputs based on their reliance on learned associations with the subject entity. Specifically, we perform a causal intervention by blocking attention from the subject tokens and measuring the resulting shift in the output distribution. Because attention is the primary mechanism through which subject representations interact with subsequent tokens, a large distributional shift under masking

indicates a strong reliance on the subject representation. We quantify this shift using the Jensen–Shannon (JS) divergence between the original and masked output distributions. We set the threshold to the average JS divergence observed for all correct answers (FAs): hallucinations are categorized as **Associated Hallucinations (AH)** if the JS divergence exceeds this threshold; otherwise, they are categorized as **Unassociated Hallucinations (UH)**. Table 1 summarizes the final data statistics.

4 Analysis of Internal States in LLMs

As established by Geva et al. (2023), factual generation relies on three components: (i) the enrichment of subject representations in the early layers, (ii) an attention flow from the subject to the last token in the middle layers, and (iii) the decoding of the last token representation in the upper layers. In this section, we first visualize this overall information flow across all categories (§4.1), and analyze each component in §4.2 through §4.4. Parallel experimental results on Mistral are summarized in Appendix C.

4.1 Visualization of Information Flow

In §3.3, we categorized hallucinations by intervening on the attention flow from the subject tokens. To provide a more granular understanding of the internal mechanisms established by Geva et al. (2023), we perform fine-grained causal interventions on all three components across each knowledge category. This allows us to examine whether the enrichment and decoding phases remain consistent between Associated Hallucinations (AHs) and Factual Associations (FAs).

Experimental Setup We measure the importance of each component by observing the resulting shift in the output distribution after intervention; a

large shift indicates the component is critical for the prediction, while a small shift suggests a limited role. We quantify each shift using Jensen-Shannon (JS) divergence between the original and intervened output distributions. Specifically, for subject-token and last-token representations, we corrupt the corresponding hidden states at each layer ℓ . For the attention mechanism, we follow Geva et al. (2023) to mask the attention flow between the subject and last tokens at layer ℓ , using a window size of 5 layers.

Experimental Results As shown in Figure 2, FAs and AHs share highly similar information flows. The components that are crucial for producing factually correct outputs are also crucial for producing AHs. This pattern suggests that, despite being factually incorrect, AHs are generated through the same underlying internal mechanisms as factual recall. In contrast, UHs display substantially weaker information flow, reflecting their limited reliance on the subject entity during generation.

These findings indicate that hidden states primarily capture **whether the model is recalling parametric knowledge**, rather than **whether the generated output is factually correct**. These results challenge the prevailing assumption in prior works (Azaria and Mitchell, 2023; Gottesman and Geva, 2024; Yüsekçönül et al., 2024; Orgad et al., 2025; Varshney et al., 2023b) that internal representations inherently encode factual correctness. Instead, our results show that factual outputs and association-driven hallucinations can share nearly identical internal computational pathways.

It is worth noting that the strong causal effect of attention-flow interventions for FAs and AHs is expected, since our labeling procedure already uses attention masking to distinguish AHs from UHs. However, interventions on subject-token representations and last-token hidden states provide an independent validation of our hypothesis. These components remain highly influential for both FAs and AHs, confirming that the shared mechanism extends beyond the attention pathways used during labeling.

4.2 Analysis of Subject Representations

Having established the overall information flow, we next analyze the subject representations in the early layers. Our goal is to understand why the internal behavior at subject positions is highly similar for

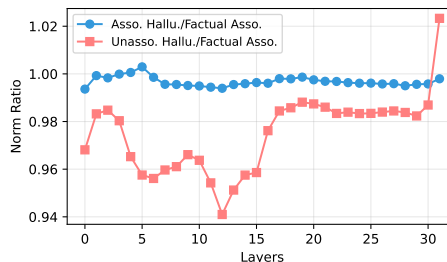


Figure 3: Norm ratio curves of subject representations in LLaMA-3-8B, comparing AHs and UHs against FAs as the baseline.

AHs and FAs, yet differs substantially for UHs. To this end, we first examine the norm of these representations (§4.2.1), and then analyze how this behavior correlates with subject prevalence in the pre-training data (§4.2.2).

4.2.1 Norm of Subject Representations

We measure the average L_2 norm of subject representations for each knowledge category across layers (details in Appendix B). To facilitate direct comparison, we report the L_2 norms of both hallucination types (AHs and UHs) as ratios relative to the norm of factually correct predictions (FAs). A ratio close to 1 indicates norms comparable to FAs, while ratios below 1 indicate smaller norms.

As shown in Figure 3, for LLaMA-3-8B, the norm of AH samples is highly comparable to that of FAs, with ratios remaining close to 1 across layers. In contrast, UH samples consistently exhibit smaller norms than FAs in nearly all layers, with the ratio dropping to approximately 0.95 in the middle layers (layers 10–15). This suggests that **AHs and FAs activate subject representations with similar strength, whereas UHs involve weaker subject encoding**. We further analyze the subspace overlap between subject hidden states and MLP weight matrices in Appendix D.1, which provides additional evidence that AHs and FAs occupy highly similar representational regions.

4.2.2 Correlation with Subject Popularity

To explain why AHs exhibit subject representation norms comparable to FAs while UHs remain weaker, we hypothesize that this difference arises from subject popularity in the pre-training data. Prior work (Kang et al., 2024) shows that entities appearing more frequently during pre-training tend to produce hidden representations with larger norms. Accordingly, we expect that FAs and AHs predominantly involve high-popularity entities,

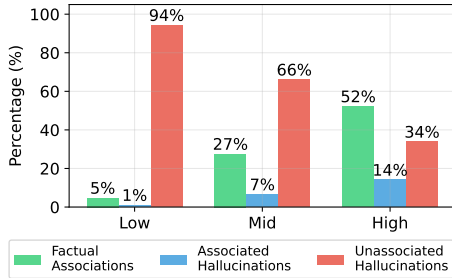


Figure 4: Sample distribution across different subject popularity (low, mid, high) in LLaMA-3-8B, measured by monthly Wikipedia page views.

while UHs are more common for low-popularity entities.

Following [Mallen et al. \(2023a\)](#), we use average monthly Wikipedia page views as a proxy for subject frequency during pre-training. We partition subjects into three popularity bins and analyze the distribution of FAs, AHs, and UHs across them. The distribution in Figure 4 confirms that the prevalence of both FAs and AHs is strongly associated with subject popularity. In the lowest popularity bin, both FAs and AHs are rarely observed (5% for FAs and 1% for AHs). In contrast, their prevalence significantly increases in the highest popularity bin (52% for FAs and 14% for AHs). However, the prevalence of UHs shows an inverse trend: while the lowest popularity bin is dominated by UHs (94%), their prevalence drops to 34% in the highest popularity bin.

This observation highlights a critical implication: if high-popularity subjects account for a substantial portion of real-world language model usage, AHs may constitute a substantial portion of hallucinations encountered in practical deployments. Meanwhile, as AHs arise mainly on popular subjects, they are often indistinguishable from FAs by popularity-based heuristics, contradicting prior work ([Mallen et al., 2023a](#)) that links popularity to hallucinations.

4.3 Visualization of Attention Flow

Having examined how the model forms subject representations, we next visualize how this information is propagated to the last token of the input where the model generates the object of a knowledge tuple. While the strong reliance of AHs on this attention flow is an expected consequence of our causal intervention labeling, visualizing this mechanism explicitly illustrates the divergence between AHs/FAs and UHs, providing a more com-

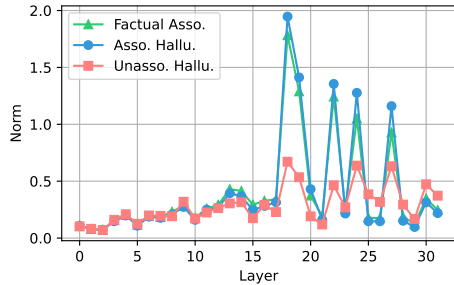


Figure 5: Subject-to-last attention contribution norms across layers in LLaMA-3-8B. Values show the norm of the attention contribution from subject tokens to the last token at each layer.

prehensive view of the internal mechanism behind how the model produces outputs for each category.

To quantify the specific contribution from subject tokens (s_1, \dots, s_n) to the last token, we compute the attention contribution from subject tokens to the last position:

$$\mathbf{a}_{\text{last}}^\ell = \sum_k \sum_h A_{\text{last}, s_k}^{\ell, h} (\mathbf{h}_{s_k}^{\ell-1} W_V^{\ell, h}) W_O^{\ell, h}. \quad (1)$$

where $A_{i,j}^{\ell, h}$ denotes the attention weight assigned by the h -th head in the layer ℓ from the last position i to subject token j . Here, $\mathbf{a}_{\text{last}}^\ell$ represents the subject-to-last attention contribution at layer ℓ . Intuitively, if subject information is critical for prediction, this contribution should have a large norm; otherwise, the norm should be small.

Figure 5 shows that in LLaMA-3-8B, both AHs and FAs exhibit large attention-contribution norms in mid-layers, indicating a strong information flow from subject tokens to the target token. In contrast, UHs show consistently lower norms, implying that their predictions rely far less on subject information. [Yüksekgönül et al. \(2024\)](#) previously argued that high attention flow from subject tokens signals factuality and proposed using attention-based hidden states to detect hallucinations. Our results challenge this view: **the model propagates subject information just as strongly when generating AHs as when producing correct facts.**

4.4 Analysis of Last Token Representations

Since both FAs and AHs involve strong information propagation from subject representations, whereas this propagation is weaker for UHs, we hypothesize that this disparity in information flow also leads to distinct geometric properties in the last token representations. Specifically, the strong subject propagation for FAs and AHs suggests that their last-token representations receive significant subject-specific updates, causing them to *diverge* in

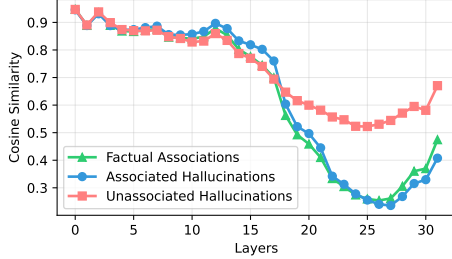


Figure 6: Cosine similarity of target-token hidden states across layers in LLaMA-3-8B.

representation space. In contrast, because subject information is weakly propagated for UH samples, their last-token representations receive smaller updates and therefore are expected to remain *tightly clustered*, dominated by the shared features of the prompt template.

To verify this, we compute the layer-wise cosine similarity among last-token representations \mathbf{h}_T^ℓ within each knowledge category (Figure 6). The results indicate that the strength of propagated subject information significantly affects the geometric properties. In early layers, cosine similarity is high for all categories (≈ 0.9). From the mid-layers onward, where subject representations propagate for FAs and AHs, their last-token representations diverge significantly. By layer 25, the cosine similarity for both FAs and AHs drops to ≈ 0.2 . In contrast, UHs remain moderately clustered, with cosine similarity only declining to ≈ 0.5 .

To further illustrate this, we present a t-SNE visualization of the last token’s representations at layer 25 of LLaMA-3-8B in Figure 7. The visualization aligns with our cosine similarity analysis: hidden representations of UHs form a distinct cluster separated from FAs, whereas AHs substantially overlap with FAs. We quantitatively validate this observation using two cluster separability metrics (the Silhouette coefficient and Davies-Bouldin Index) in Appendix D.2, which show significantly higher separability for UHs compared to AHs. Additional visualizations can be found in Appendix D.3.

Additionally, we observe that the strength of subject representation propagation also affects the entropy of the output distribution (Figure 8). A strong subject-specific update concentrates the output distribution on specific subject-related entities, resulting in low-entropy outputs for FAs and AHs. Conversely, a weak subject-specific update leads to a wider spread of probability mass covering many possible answers (e.g., assigning probability to multiple possible names for the prompt “The name of

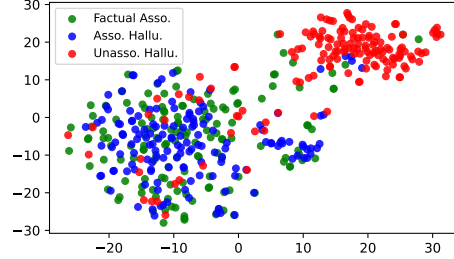


Figure 7: t-SNE visualization of last token’s representations at layer 25 of LLaMA-3-8B.

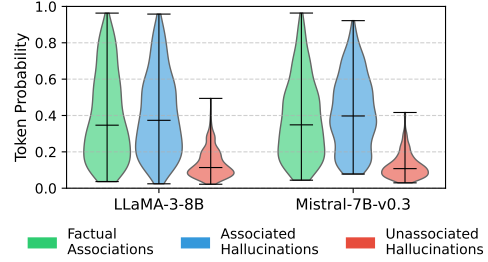


Figure 8: Distribution of last token probabilities.

the father of <subject> is”), resulting in significantly higher output entropy.

5 Revisiting Hallucination Detection

The mechanistic analysis in §4 reveals that *internal states of LLMs primarily capture how the model utilizes its parametric knowledge, rather than whether the output is truthful*. This suggests that representation-based signals struggle to distinguish associated hallucinations (AHs) from factual associations (FAs) but can more effectively separate unassociated hallucinations (UHs) from FAs.

In this section, we quantify the extent to which existing hallucination detection methods can separate AHs from FAs and compare this performance to detecting UHs.

Experimental Setups We revisit the effectiveness of widely-adopted white-box hallucination detection approaches that use internal state probing as well as black-box approaches that rely on scalar features. We evaluate on three settings: 1) **AH Only** (1,000 FAs and 1,000 AHs for training; 200 of each for testing), 2) **UH Only** (1,000 FAs and 1,000 UHs for training; 200 of each for testing), and 3) **Full** (1,000 FAs and 1,000 hallucination samples mixed of AHs and UHs for training; 200 of each for testing). For each setting, we repeat data splitting with five random seeds and report mean AUROC and standard deviation.

White-box methods: We extract and normalize

Methods	LLaMA		Mistral	
	AH Only	UH Only	AH Only	UH Only
Subject	0.65 ± 0.02	0.91 ± 0.01	0.57 ± 0.02	0.81 ± 0.02
Attention	0.58 ± 0.04	0.92 ± 0.02	0.58 ± 0.07	0.87 ± 0.01
Last Token	0.69 ± 0.03	0.93 ± 0.01	0.63 ± 0.02	0.92 ± 0.01
Probability	0.49 ± 0.01	0.86 ± 0.01	0.46 ± 0.00	0.89 ± 0.00
Subject Pop.	0.48 ± 0.01	0.87 ± 0.01	0.52 ± 0.01	0.84 ± 0.01

Table 2: Hallucination detection performance on **AH Only** and **UH Only** settings. Detailed statistical significance analysis is provided in Appendix E.1.

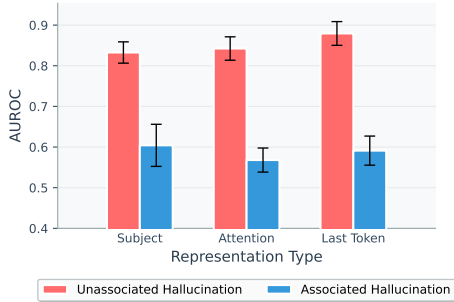


Figure 9: Hallucination detection performance on the **Full** setting (LLaMA-3-8B).

internal features and then train a probe.

- **Subject representations:** last subject token hidden state from three consecutive layers (Gottesman and Geva, 2024).
- **Attention flow:** attention weights from the last token to subject tokens across all layers (Yükselgönül et al., 2024).
- **Last-token representations:** final token hidden state from the last layer (Orgad et al., 2025).

Black-box methods: We test two commonly used scalar features, including **answer token probability** (Orgad et al., 2025) and **subject popularity** (average monthly Wikipedia page views) (Mallen et al., 2023a). As discussed in §4.2.2 and §4.4, these features are also related to whether the model relies on encoded knowledge to produce outputs rather than with truthfulness itself.

Experimental Results Table 2 shows that hallucination detection methods perform differently in **AH Only** and **UH Only** settings. For white-box probes, all approaches effectively distinguish UHs from FAs (AUROC \approx 0.91–0.93). In contrast, performance drops sharply on the **AH Only** setting (AUROC \approx 0.58–0.69). Black-box methods follow the same pattern (AUROC \approx 0.86–0.87 on **UH Only** and \approx 0.48–0.49 on **AH Only**). Notably, the performance of black-box methods on **AH Only** is close to random guessing (AUROC \approx 0.48–0.49), indicating that token uncertainty and subject popularity provide little signal for separating AHs from

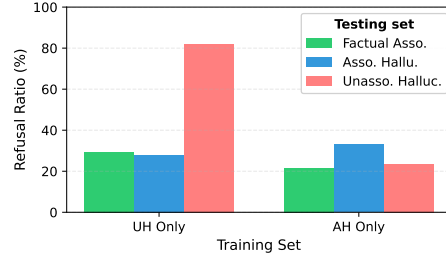


Figure 10: Refusal tuning performance across three types of samples (LLaMA-3-8B).

FAs. Although white-box representations show slightly better separability than black-box methods, such performance is far from reliable for deployment (Hosmer Jr et al., 2013). Figure 9 further highlights this disparity under the **Full** setting: detection is consistently stronger on UH samples than on AH samples, and adding AHs to the training set significantly dilutes performance on UHs (AUROC \approx 0.9 on **UH Only** vs. \approx 0.8 on **Full**).

These results confirm that both internal probes and black-box methods capture whether a model draws on parametric knowledge, not whether its outputs are factually correct. Our further analysis in Appendix E.2 reveals that consistency-based detection methods also struggle to distinguish AHs from FAs. UHs are easier to detect because they bypass this knowledge, while associated hallucinations are produced through the same recall process as FAs, leaving limited internal cues to distinguish them. An analysis of sensitivity-specificity trade-offs in Appendix E.3 further confirms this limitation, showing that while UHs can be detected with high sensitivity and specificity simultaneously, AH detection requires a significant trade-off between them. As a result, **LLMs appear to have limited intrinsic awareness of their own truthfulness, and detection methods relying on these signals risk misclassifying AHs as correct**, fostering harmful overconfidence in model outputs.

6 Challenges of Refusal Tuning

A common strategy to mitigate potential hallucinations in the model’s responses is to fine-tune LLMs to refuse answering when they cannot provide a factual response, *e.g.*, **Refusal Tuning** (Zhang et al., 2024). For such refusal capability to generalize, the training data must contain a shared feature pattern across hallucinated outputs, allowing the model to learn and apply it to unseen cases.

Our analysis in the previous sections shows that

this prerequisite is not met. Because AH representations are geometrically diverse, it is difficult for models to capture a consistent feature pattern among AH samples, thereby hindering the generalization of refusal responses to unseen AH samples. In contrast, UH representations tend to form distinct cluster, presenting consistent feature patterns that should facilitate easier generalization of the refusal capability.

Experimental Setups To verify the hypothesis, we conduct refusal tuning on LLMs under two settings: 1) **UH Only**, where 1,000 UH samples are paired with 10 refusal templates, and 1,000 FA samples are preserved with their original answers. 2) **AH Only**, where 1,000 AH samples are paired with refusal templates, with 1,000 FA samples again remain unchanged. We then evaluate both models on 200 samples each of FAs, UHs, and AHs. A response matching any refusal template is counted as a refusal, and we report the **Refusal Ratio** as the proportion of samples eliciting refusals. This measures not only whether the model refuses appropriately on UHs and AHs, but also whether it “over-refuses” on FA samples.

Experimental Results Figure 10 shows that training with UHs leads to strong generalization across UHs, with refusal ratios of 82% for LLaMA. However, this effect does not transfer to AHs, where refusal ratios fall to 28%, respectively. Moreover, some FA cases are mistakenly refused (29.5%). These results confirm that UHs share a common activation subspace, supporting generalization within the category, while AHs and FAs lie outside this space. By contrast, training with AHs produces poor generalization. On AH test samples, refusal ratio is only 33%, validating that their subject-specific hidden states prevent consistent refusal learning. Generalization to UHs is also weak (23.5%), again reflecting the divergence between AH and UH activation spaces.

Overall, these findings indicate that **the generalizability of refusal tuning is constrained by the heterogeneous nature of hallucinations**. While UH representations are internally consistent enough to support learnable refusal generalization, AH representations are too geometrically diverse for the model to learn a consistent and robust refusal pattern.

7 Conclusions and Future Work

In this work, we revisit the widely accepted claim that hallucinations can be detected from a model’s internal states. Our mechanistic analysis reveals that **hidden states encode whether models rely on their parametric knowledge rather than truthfulness**. As a result, detection methods succeed most reliably when outputs are detached from the input but struggle when hallucinations arise from the same knowledge-recall process as correct answers.

These findings lead to three key implications. First, future evaluations should **report detection performance separately** for Associated Hallucinations (AHs) and Unassociated Hallucinations (UHs), as they stem from fundamentally different internal processes and require distinct detection strategies. Second, relying solely on hidden states is insufficient for reliable hallucination detection. Future research should integrate LLMs with **external feedback mechanisms**, such as fact-checking modules or retrieval-based verifiers, to assess factuality more robustly. Third, future studies should **prioritize improving AH detection**. Because AHs occur more frequently in widely known or highly popular topics (§4.2), their undetected errors pose greater risks to user trust and the practical reliability of LLMs.

Limitations

We identify several limitations of our work.

Focus on Factual Knowledge While our analysis identifies challenging cases of hallucination detection methods, our study is primarily limited to factual completion prompts. It does not extend to long-form or open-ended text generation tasks (Wei et al., 2024; Min et al., 2023; Huang and Chen, 2024). Future work should broaden this investigation to these tasks in order to draw more comprehensive conclusions.

Lack of Analysis on Prompt-based Hallucination Detection Approaches Our analysis focuses on white-box hallucination detection methods based on internal states and two black-box approaches based on external features. We do not include verbalization-based strategies (Lin et al., 2022a; Tian et al., 2023; Xiong et al., 2024; Yang et al., 2024b; Ni et al., 2024; Zhao et al., 2024), such as prompting the model to report or justify

its confidence explicitly, which constitute a different line of approach. Exploring such approaches may offer complementary insights into how models internally represent and express uncertainty.

Applicability to Black-box LLMs or Large Reasoning Models Our study is limited to open-source LLMs. Conducting mechanistic analyses on commercial black-box LLMs is not permitted due to access restrictions. Future work could explore alternative evaluation protocols or collaboration frameworks that enable partial interpretability analyses on such systems. In addition, recent studies (Mei et al., 2025; Zhang et al., 2025) have begun examining the internal states of large reasoning models for hallucination detection, suggesting a promising direction for extending our methodology to models with multi-step reasoning capabilities.

Ethical Considerations

This work analyzes the internal mechanisms of large language models using data constructed from Wikidata (Vrandečić and Kröttsch, 2014), which is released under the Creative Commons CC0 1.0 Universal license, allowing unrestricted use and redistribution of its data. All data are derived from publicly available resources, and no private or sensitive information about individuals is included. We used LLM tools for language polishing.

Acknowledgments

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 24-SIS-SMU-002) and the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016).

References

- Amos Azaria and Tom M. Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.
- Chi Seng Cheang, Hou Pong Chan, Derek F. Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia S. Chao. 2023. [Can lms generalize to future data? an empirical analysis on text summarization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16205–16217. Association for Computational Linguistics.
- 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Michael Han Daniel Han and Unsloth team. 2023. [Unsloth](#).
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. [Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models](#). *CoRR*, abs/2402.10612.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. [Inside-out: Hidden factual knowledge in llms](#). *CoRR*, abs/2503.15299.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting recall of factual associations in auto-regressive language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12216–12235. Association for Computational Linguistics.
- Daniela Gottesman and Mor Geva. 2024. [Estimating knowledge in large language models without generating a single token](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 3994–4019.
- Nuno Miguel Guerreiro, Elena Voita, and André F. T. Martins. 2023. [Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1059–1075. Association for Computational Linguistics.
- David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

- Chao-Wei Huang and Yun-Nung Chen. 2024. [Factalign: Long-form factuality alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16363–16375.
- 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 Trans. Inf. Syst.*, 43(2):42:1–42:55.
- Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. [LLM internal states reveal hallucination risk faced with a query](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Cheongwoong Kang and Jaesik Choi. 2023. [Impact of co-occurrence on factual knowledge of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735.
- Katie Kang, Amrith Setlur, Claire J. Tomlin, and Sergey Levine. 2024. [Deep neural networks tend to extrapolate predictably](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024*.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. [Large language models must be taught to know what they don’t know](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- 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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Kenneth Li, Oam Patel, Fernanda Vi  gas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). *Advances in Neural Information Processing Systems*, 36:41451–41530.
- Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. [Knowledge boundary of large language models: A survey](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025*, pages 5131–5157.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. [Teaching models to express their uncertainty in words](#). *Trans. Mach. Learn. Res.*, 2022.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. [Truthfulqa: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 3214–3252.
- Alex Mullen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. [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), ACL 2023*, pages 9802–9822.
- Alex Mullen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. [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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.
- 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.
- Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. 2025. [Reasoning about uncertainty: Do reasoning models know when they don’t know?](#) *CoRR*, abs/2506.18183.
- 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, EMNLP 2023*, pages 12076–12100.
- Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. [When do llms need retrieval augmentation? mitigating llms’ overconfidence helps retrieval augmentation](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages

- 11375–11388. Association for Computational Linguistics.
- Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. 2025. [Towards fully exploiting LLM internal states to enhance knowledge boundary perception](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24315–24329. Association for Computational Linguistics.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szepktor, Hadas Kotek, and Yonatan Belinkov. 2025. [LLMs know more than they show: On the intrinsic representation of LLM hallucinations](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.
- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. [Simple entity-centric questions challenge dense retrievers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6138–6148. Association for Computational Linguistics.
- Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. [Un-supervised real-time hallucination detection based on the internal states of large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14379–14391. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.
- Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jian-shu Chen, and Dong Yu. 2023a. [A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation](#). *CoRR*, abs/2307.03987.
- Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jian-shu Chen, and Dong Yu. 2023b. [A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation](#). *arXiv preprint arXiv:2307.03987*.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. [Long-form factuality in large language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#). *CoRR*, abs/1910.03771.
- Chenghao Xiao, Hou Pong Chan, Hao Zhang, Mahani Aljunied, Lidong Bing, Noura Al Moubayed, and Yu Rong. 2025. [Analyzing llms’ knowledge boundary cognition across languages through the lens of internal representations](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24099–24115. Association for Computational Linguistics.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- 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, and 22 others. 2024a. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. [Alignment for honesty](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. [Do large language models know what they don’t know?](#) In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.
- Gal Yona, Roei Aharoni, and Mor Geva. 2024. [Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6737–6751. Association for Computational Linguistics.
- Mert Yüsekönül, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. [Attention satisfies: A constraint-satisfaction lens on factual errors of](#)

language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. *R-tuning: Instructing large language models to say ‘i don’t know’*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024*, pages 7113–7139.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Kumar Sricharan. 2023a. *Sac³: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency*. *CoRR*, abs/2311.01740.

Qingjie Zhang, Yujia Fu, Yang Wang, Liu Yan, Tao Wei, Ke Xu, Minlie Huang, and Han Qiu. 2025. *On the self-awareness of large reasoning models’ capability boundaries*. *Preprint*, arXiv:2509.24711.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. *Siren’s song in the AI ocean: A survey on hallucination in large language models*. *CoRR*, abs/2309.01219.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. *Knowing what llms DO NOT know: A simple yet effective self-detection method*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024*, pages 7051–7063.

Appendix

A Datasets and Implementations

A.1 Selected Relations and Prompt Templates

We employed a set of criteria to select relations from Wikidata in order to construct our dataset. Our criteria largely follow the framework proposed by Gekhman et al. (2025). Specifically, we require that each factual query in the dataset be unambiguous: given a subject–relation pair, the object should be unique and easy verifiable. The criteria are as follows:

- **Avoid granularity ambiguity.** We exclude relations whose answers can vary in their level of detail. For example, in *location* queries, the response could be expressed as a city, state, or country, making it ill-defined (Yona et al., 2024).

Relation	Prompt Template
father	The name of the father of [subject] is
mother	The name of the mother of [subject] is
spouse	The name of the spouse of [subject] is
date of birth	The birth date of [subject] is

Table 3: Relations and prompt templates for querying factual knowledge of models. [subject] is a placeholder replaced with subject entities.

- **Avoid surface-level guessing.** We exclude relations whose correct answers can often be inferred from shallow patterns. For instance, *country of citizenship* can frequently be guessed from shallow lexical patterns, rather than reflecting actual memorization (Mallen et al., 2023b).

Following these criteria, Gekhman et al. (2025) narrowed the 24 relations introduced by Sciavolino et al. (2021) down to four. However, we observe that their filtering primarily addresses ambiguity at the relation and object levels, but does not consider ambiguity at the subject level. In practice, some relations involve subjects that are inherently ambiguous. For example, the relation *record label* can be problematic because many songs share identical names, leading to unclear subject–object mappings.

To mitigate such cases, we apply an additional subject-level filtering step and restrict our dataset to relations where the subject is a person, thereby reducing ambiguity. In addition, we manually include certain relations to strengthen the dataset. Concretely, we use the following four relations: P22 (*father*), P25 (*mother*), P26 (*spouse*), and P569 (*date of birth*). We show the list of the templates used to create our dataset in Table 3.

A.2 Implementation Details

Checkpoints and GPU resources. All the checkpoints used in our experiments are provided by the Hugging Face Transformers library (Wolf et al., 2019). Specifically, we use the checkpoint “meta-llama/Meta-Llama-3-8B”¹ and “mistralai/Mistral-7B-v0.3”² for the experiments of response generation (§3), hidden-state analysis (§4) and accessing the performance of hallucination detection methods (§5). For refusal tuning (§6), we use checkpoints

¹<https://huggingface.co/meta-llama/Meta-Llama-3-8B>

²<https://huggingface.co/mistralai/Mistral-7B-v0.3>

LLM Judge Prompt
<p>I will give you a factual query (e.g., “The name of the father of <subj>”), a gold answer to the factual query, and a proposed answer generated by an LLM. You need to compare the proposed answer to the gold answer and assign it one of the possible grades using the steps below.</p> <p>Possible grades are:</p> <p>A: CORRECT B: INCORRECT C: WRONG GOLD D: ERROR</p> <p>Spelling errors, synonyms, abbreviations, or hedging expressions (e.g., “it is possible that”) should not alter the grade if the person referred to in the proposed answer matches the gold answer.</p> <p>Steps:</p> <p>Step 1: If the gold answer does not correspond to an answer for the question, output “C” and finish. Otherwise, proceed to Step 2.</p> <p>Step 2: Extract all predicted entities from the proposed answer. Proceed to Step 3.</p> <p>Step 3: If each predicted entity refers to the answer mentioned in the gold answer, output “A” and finish. Otherwise, proceed to Step 4.</p> <p>Step 4: If the predicted entity does not refer to the gold answer, output “B” and finish. Otherwise, proceed to Step 5.</p> <p>Step 5: Double-check whether the proposed answer refers to a different answer from the gold answer. If it does, output “B.” Otherwise, output “D” and finish.</p> <p>Input format:</p> <p>Question: {question} Gold answer: {gold_answer} Proposed answer: {proposed_answer}</p> <p>Instruction: Output your reasoning steps. After that, conclude your response with “Output:” followed by the letter (A, B, C, or D). Do not provide any further explanation.</p>

Figure 11: LLM Judge prompt used for evaluation.

provided by the Unsloth framework (Daniel Han and team, 2023), namely “unsloth/llama-3-8b”³ and “unsloth/mistral-7b-v0.3”⁴, which enable more efficient fine-tuning. All experiments are conducted on 4 NVIDIA L40S GPUs.

Decoding algorithm. We employ greedy decoding (temperature = 0) for response generation, with models run in BF16 precision.

PEFT settings for refusal tuning. For refusal tuning, we fine-tune with both models using QLoRA (Dettmers et al., 2023), implemented with the Unsloth framework (Daniel Han and team, 2023), with rank $r = 8$, and $\alpha = 8$. QLoRA adapters are applied to all attention and MLP modules, and each model is fine-tuned for one epoch.

³<https://huggingface.co/unsloth/llama-3-8b>

⁴<https://huggingface.co/unsloth/mistral-7b-v0.3>

A.3 Labeling Factual Correctness

Here, we outline the process of determine Factual Correctness of a model prediction. We construct correctness labels through a two-stage process. First, we use spaCy⁵ Named Entity Recognizer to extract the target entity from the model’s output. If it matches the ground truth, the answer is marked correct. Otherwise, or if extraction fails, we rely on Qwen2.5-14B-Instruct (Yang et al., 2024a) as a judge to compare the predicted answer with the ground truth. Following Gekhman et al. (2025), we design the evaluation prompt, which is shown in Figure 11.

B Calculation of Subject Representation Norms

To investigate the geometric properties of subject representations, we compute the average L_2 norm

⁵<https://spacy.io/>

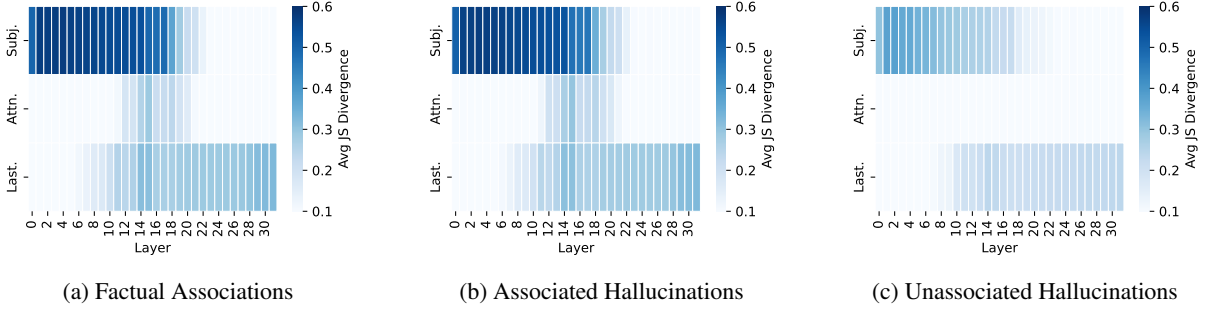


Figure 12: Effect of interventions across layers of Mistral-7B-v0.3. The heatmap shows JS divergence between the output distribution before and after intervention. Darker color indicates that the intervened hidden states are more causally influential on the model’s predictions. Top row: patching representations of subject tokens. Middle row: blocking attention flow from subject to the last token. Bottom row: patching representations of the last token.

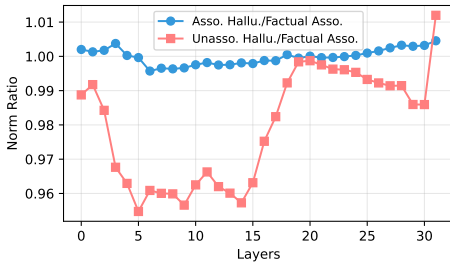


Figure 13: Norm ratio curves of subject representations in Mistral-7B-v0.3, comparing AHs and UHs against FAs as the baseline. At earlier layers, the norm of UH samples is significantly lower than that of AH samples.

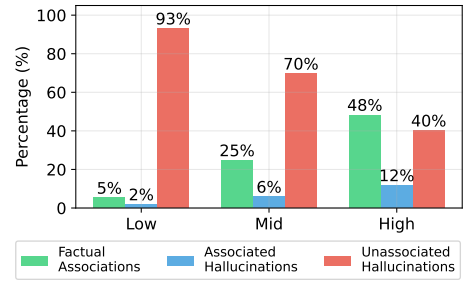


Figure 15: Sample distribution across different subject popularity (low, mid, high) in Mistral-7B-v0.3, measured by monthly Wikipedia page views.

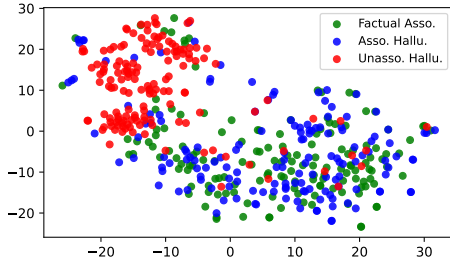


Figure 14: t-SNE visualization of last token’s representations at layer 25 of Mistral-7B-v0.3.

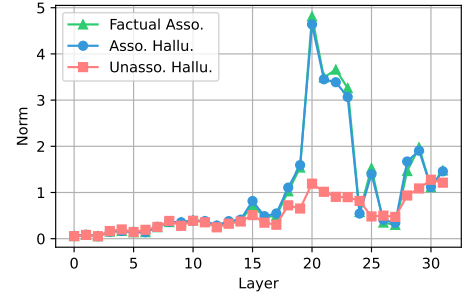


Figure 16: Subject-to-last attention contribution norms across layers in Mistral-7B-v0.3. Values show the norm of the attention contribution from subject tokens to the last token at each layer.

of the hidden states across different knowledge categories. For a single sample with subject tokens s_1, \dots, s_n , we first compute the average norm across all subject tokens at layer ℓ :

$$\bar{N}_{\text{sample}}^{(\ell)} = \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{h}_{s_i}^{(\ell)} \right\|_2 \quad (2)$$

Then, for a given category (e.g., FAs, UHs, or AHs) containing K samples, we aggregate these values to obtain the per-category average norm:

$$\bar{N}_{\text{category}}^{(\ell)} = \frac{1}{K} \sum_{j=1}^K \bar{N}_{\text{sample},j}^{(\ell)} \quad (3)$$

where $\bar{N}_{\text{sample},j}^{(\ell)}$ is the average norm for the j -th sample in the category. By repeating this process for each layer $\ell \in \{1, \dots, L\}$, we obtain a layer-wise norm profile. Finally, to facilitate direct comparison, we report the norm ratio of hallucinations relative to factual associations:

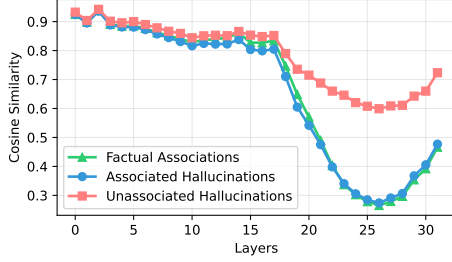


Figure 17: Cosine similarity of target-token hidden states across layers in Mistral-7B-v0.3. From mid-layers onward, FAs and AHs diverge sharply as subject information propagates, while UHs remain more clustered, confirming weaker subject-dependent updates.

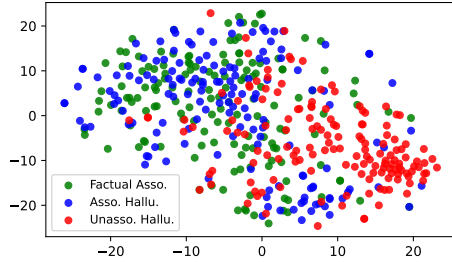


Figure 18: t-SNE visualization of subject tokens' representations at layer 11 of LLaMA-3-8B.

$$\text{Ratio}_{\text{AH/FA}}^{(\ell)} = \frac{\bar{N}_{\text{AH}}^{(\ell)}}{\bar{N}_{\text{FA}}^{(\ell)}} \quad (4)$$

$$\text{Ratio}_{\text{UH/FA}}^{(\ell)} = \frac{\bar{N}_{\text{UH}}^{(\ell)}}{\bar{N}_{\text{FA}}^{(\ell)}} \quad (5)$$

C Parallel Experiments on Mistral

This section is for documenting parallel experiments conducted on the Mistral-7B-v0.3 model under the same settings as described in the main text (Figures 12–21).

The results from Mistral exhibit similar patterns to those observed in LLaMA, as described before. Specifically, we find consistent patterns in the model's internal computations, hidden-state behaviors, and the performance of hallucination detection and refusal tuning experiments.

D Further Analysis on Internal States

D.1 Analysis of Parametric Subspace Overlap for Subject Representation

We investigate why early layers encode subject representations differently across knowledge types by examining how inputs interact with the parametric

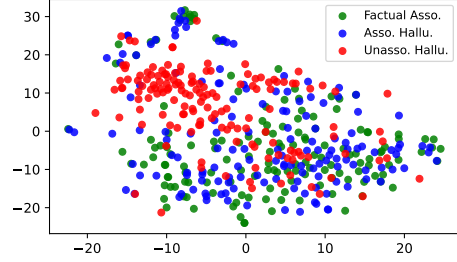


Figure 19: t-SNE visualization of subject tokens' representations at layer 11 of Mistral-7B-v0.3.

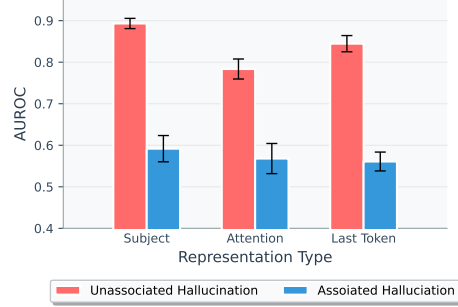


Figure 20: Hallucination detection performance on the Full setting (Mistral-7B-v0.3).

knowledge stored in MLP modules. Inspired by Kang et al. (2024), the output norm of an MLP layer depends on how well its input aligns with the subspace spanned by the weight matrix: *poorly aligned inputs yield smaller output norms*.

For each MLP layer ℓ , we analyze the down-projection weight matrix W_{down}^{ℓ} and its input x^{ℓ} . Given the input x_s^{ℓ} corresponding to the subject tokens, we compute its overlap ratio with the top singular subspace V_{top} of W_{down}^{ℓ} :

$$r(x_s^{\ell}) = \frac{\|x_s^{\ell \top} V_{\text{top}} V_{\text{top}}^{\top}\|^2}{\|x_s^{\ell}\|^2}. \quad (6)$$

A higher overlap ratio $r(x_s^{\ell})$ indicates stronger alignment to the subspace spanned by W_{down}^{ℓ} , leading to larger output norms.

To highlight relative deviations from the factual baseline (FA), we report the relative ratios between AH/FA and UH/FA. Focusing on the layer with the largest UH norm shift, Figure 22 shows that UHs have significantly lower $r(x_s^{\ell})$ than AHs in both LLaMA and Mistral. This reveals that early-layer parametric weights are more aligned with FA and AH subject representations than with UH subjects, producing higher norms for the former ones. These results also suggest that the model has sufficiently learned representations for FA and AH subjects during pretraining but not for UH subjects.

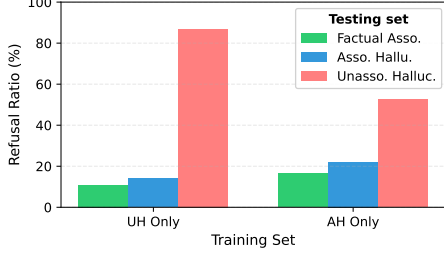


Figure 21: Refusal tuning performance across three types of samples (Mistral-7B-v0.3).

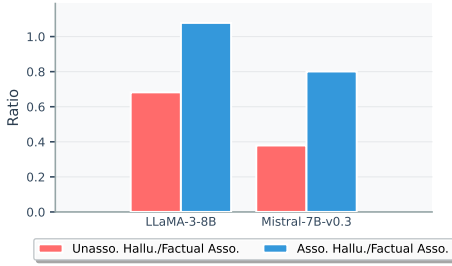


Figure 22: Comparison of subspace overlap ratios.

D.2 Quantitative Analysis of Cluster Separability

To quantitatively validate the visual separability observed in the t-SNE visualizations (Figures 7 and 14), we compute two cluster separability metrics on the hidden representations: the Silhouette coefficient and the Davies-Bouldin Index (DBI). The Silhouette coefficient ranges from -1 to 1 , where higher values indicate better-defined clusters. The DBI measures the average similarity ratio of each cluster with its most similar cluster, where lower values indicate better separation.

The results shown in Table 4 quantitatively confirm our observations from the t-SNE visualizations. When comparing UHs and FAs, both models exhibit high Silhouette coefficients (0.33–0.40) and low DBI values (0.93–1.13), indicating clear cluster separation between UHs and FAs. In contrast, when comparing AHs and FAs, the Silhouette coefficients are near zero (0.0022 and -0.0011) and DBI values are substantially higher (14.22 and 25.25), indicating strong overlap between AHs and FAs. These metrics support our claim that FAs and UHs form well-separated clusters in the representation space, while AHs strongly overlap with FAs, consistent with our hypothesis that AHs rely on similar knowledge-recall processes as FAs.

Model	UH Only		AH Only	
	Sil. \uparrow	DBI \downarrow	Sil. \uparrow	DBI \downarrow
LLaMA	0.397	0.932	0.002	14.22
Mistral	0.327	1.134	-0.001	25.25

Table 4: Quantitative cluster separability metrics. Higher Silhouette (Sil., \uparrow) and lower Davies–Bouldin Index (DBI, \downarrow) indicate better separation.

Model	Method	UH Only	AH Only
Mistral	Prob.	.894 (.889–.899)	.457 (.453–.461)
	Subj. Pop.	.812 (.787–.839)	.570 (.551–.590)
	Subject	.893 (.878–.909)	.592 (.552–.631)
	Attention	.872 (.864–.881)	.584 (.500–.668)
	Last Token	.923 (.913–.934)	.629 (.610–.648)
LLaMA	Prob.	.860 (.853–.867)	.488 (.480–.495)
	Subj. Pop.	.869 (.856–.883)	.479 (.467–.491)
	Subject	.913 (.897–.930)	.648 (.626–.669)
	Attention	.920 (.889–.950)	.579 (.525–.634)
	Last Token	.935 (.923–.947)	.691 (.660–.722)

Table 5: Hallucination detection performance with 95% confidence intervals across five random seeds.

D.3 More Visualization on Hidden States

In this section, we provide t-SNE visualization of subject tokens’ hidden states in Figure 18 and Figure 19.

Compared to the last-token representations, the t-SNE visualization of subject-token hidden states shows that unassociated hallucinations (UHs) are moderately separated from factual and associated samples, they exhibit marginally lower separability than last-token representations. This observation aligns with the results in §5, where the hallucination detection performance using last-token hidden states outperforms that based on subject-token representations.

E Further Analysis on Hallucination Detection

E.1 Statistical Significance Analysis

We report the hallucination detection performance (§5) with 95% confidence intervals across five random seeds in Table 5. Across all settings relevant to our claims, AUROC on UHs is significantly higher than on AHs (all $p < 0.01$), supporting our finding that current detection methods are substantially more effective for UHs than AH.

E.2 Semantic Consistency-Based Detection

We sample 200 instances per category and generate 5 responses per prompt with temperature = 0.7. We

Model	UH Only	AH Only
LLaMA	0.8374	0.6183
Mistral	0.8437	0.6159

Table 6: Semantic consistency-based hallucination detection performance.

Threshold	UH Only		AH Only	
	Sensitivity	Specificity	Sensitivity	Specificity
0.1	0.900	0.833	0.697	0.583
0.3	0.879	0.888	0.644	0.633
0.5	0.867	0.925	0.601	0.679
0.7	0.867	0.942	0.534	0.725
0.9	0.838	0.963	0.433	0.771

Table 7: Sensitivity-specificity trade-offs at different thresholds.

compute a consistency score for each sample and evaluate hallucination detection performance using these scores, measured by AUROC. Table 6 shows the results.

The results show that the performance of this consistency-based method aligns with our findings for both the black-box and white-box methods discussed in §5: it can effectively distinguish UHs from FAs (AUROC \approx 0.84) but struggles to separate AHs from FAs (AUROC \approx 0.62). This behavioral evidence supports our mechanistic analysis: since AHs rely on similar knowledge-recall processes as FAs, they tend to produce similarly consistent outputs across runs, whereas UHs exhibit higher variance due to weaker reliance on subject representations.

E.3 Sensitivity-Specificity Trade-offs

To further examine the detection performance beyond AUROC, we analyze the trade-offs between sensitivity and specificity at various classification thresholds. We report results for detectors trained on LLaMA’s last-token representations. Tables 7 present the results.

For UH Only, we observe that detection performance is remarkably robust across the threshold range. Even as the threshold increases from 0.1 to 0.9, the detector maintains a high balance of metrics, with sensitivity remaining above 0.83 and specificity exceeding 0.96 at the highest setting. This stability indicates a clear and consistent separation in the representation space. In contrast, for AH Only, the performance is highly sensitive to the specific threshold. As the threshold moves from 0.1 to 0.9, sensitivity drops significantly from 0.697 to

0.433, while specificity only improves from 0.583 to 0.771. This trade-off suggests that establishing a stable decision boundary between AHs and FAs is challenging, confirming that their internal representations strongly overlap.