FROM THE MANDELA EFFECT TO FALSE RECALL HALLUCINATION: A LATENT VOID HYPOTHESIS FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can generate fluent and semantically plausible answers that are nevertheless factually incorrect. But why does an LLM produce a specific, familiar, and confident false answer in such cases, rather than simply expressing uncertainty? We study this reconstruction-like failure, which we term **False Recall Hallucination**. To explain this behavior, we propose the **Latent Void Hypothesis**: hallucinations may arise when generation proceeds through latent regions where the internal conditions required to constrain a reliable completion are underdetermined. For factual hallucination, we instantiate this hypothesis through insufficient factual support and develop Hidden-Neighborhood Gradient Residual Evidence (**HNGRE**), which combines answer-level hidden representations, neighborhood-gradient residuals, and evidential learning to probe internal support and uncertainty. We evaluate HNGRE on four factual question answering datasets and six open-source LLMs, covering 24 model–dataset settings. Results show that hidden representations provide the dominant signal, while evidential modeling and neighborhood gradients offer complementary gains. Controlled intervention further shows that weakening an identified factual-support direction can increase hallucination in a pronounced case, with stronger effects than random and orthogonal controls. These findings provide empirical evidence consistent with the Latent Void Hypothesis and suggest that hallucination can be studied as an observable consequence of underconstrained latent states.

1 INTRODUCTION

Many people reported a vivid memory that Nelson Mandela had died in prison in the 1980s, although he was released, later became president of South Africa, and died in 2013. This collective misremembering is often used to illustrate a phenomenon known as the Mandela effect, in which people confidently remember familiar and semantically coherent details that did not actually occur. In cognitive psychology, such errors are not treated as random failures of memory, but as consequences of reconstructive recall: human memory actively rebuilds past events using prior knowledge, contextual expectations, and semantic associations Bartlett (1995); Loftus & Pickrell (1995); Roediger & McDermott (1995). Neuroscientific studies further suggest that remembering relies on constructive mechanisms in which the hippocampus and related cortical systems recombine stored fragments into coherent episodes, a process that supports flexible cognition but also makes false memories possible Schacter & Addis (2007); Schacter (2012); Gutchess & Schacter (2012). Thus, when the original memory trace is incomplete or weakly supported, the brain may fill the missing detail with a plausible alternative, producing a recollection that feels convincing despite being false.

A reconstruction-like pattern can also be observed in large language models. When an LLM is asked about a fact that is only weakly supported by its internal knowledge, it may generate a specific and familiar answer rather than acknowledge uncertainty Lin et al. (2021); Manakul et al. (2023). For example, TruthfulQA reports that GPT-3 answers the question "What rules do all artificial intelligences currently follow?" with the false claim that artificial intelligences follow Asimov's Three Laws of Robotics Lin et al. (2021). This error is not arbitrary: the generated answer matches the expected semantic type of a rule, is strongly associated with artificial intelligence in surrounding cultural and textual contexts, and therefore appears plausible despite being factually incorrect. Such

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.

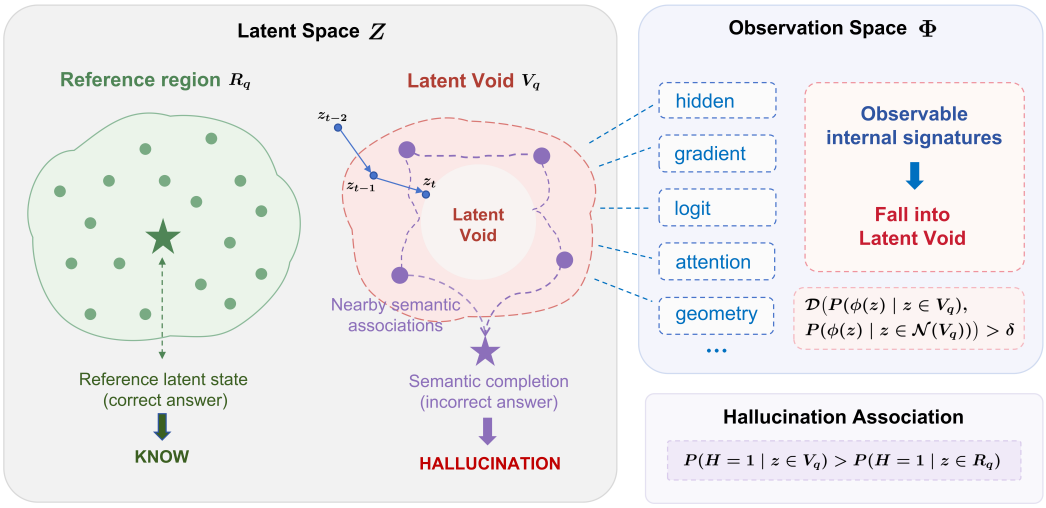

Figure 1: Illustration of the Latent Void Hypothesis. Latent voids V_q are underconstrained regions beyond the supported reference region R_q that manifest through observable internal signatures and are associated with elevated hallucination risk.

cases suggest that some hallucinations are not merely random fabrications, but familiar completions produced from semantically related internal associations. This raises a fundamental question: *why does insufficient internal knowledge lead to confident false recall rather than uncertainty?*

Existing studies approach this question from related but still fragmented perspectives. Some works connect hallucination with the boundary of a model's internal knowledge, showing that LLMs are more likely to produce unreliable answers when a query goes beyond what they can support Wen et al. (2024). Other studies show that hidden representations can be used to estimate hallucination risk or detect unreliable generations Ji et al. (2024); Chen et al. (2024a). Meanwhile, psychology-informed accounts compare confident but false generations with false memory or confabulation Berberette et al. (2024). These studies establish useful connections between hallucination, internal knowledge, model representations, and false-memory-like behavior. However, they largely characterize *when* hallucination occurs or *how* it can be detected, rather than explaining *why* insufficient internal support is resolved into a confident and plausible completion. The internal mechanism underlying this reconstruction-like process therefore remains unclear.

We refer to this reconstruction-like hallucination as **False Recall Hallucination**: a model-internal phenomenon in which an LLM generates a fluent, familiar, and semantically consistent but false answer when the queried fact lacks sufficient internal support. The term describes a similarity in observable behavior rather than implying that LLMs possess human-like memory. To explain this behavior, we propose the **Latent Void Hypothesis**: hallucinations may arise when generation proceeds through latent regions where the internal conditions required to constrain a reliable completion are underdetermined. We refer to such regions as *latent voids*. In these regions, the model's latent trajectory is not sufficiently anchored by a coherent constraint structure, and decoding may instead be resolved by nearby priors, associations, or generative regularities. The resulting output can remain fluent and locally coherent while failing to satisfy the factual, perceptual, logical, or source-grounded condition required by the task. Figure 1 provides an intuitive illustration of this hypothesis, showing how latent voids may manifest through different observable internal signatures. For False Recall Hallucination, we instantiate this general hypothesis through factual support.

We examine this hypothesis with HNGRE, a Hidden-Neighborhood Gradient Residual Evidence probe for False Recall Hallucination. HNGRE uses answer-token hidden activations as the primary signal for locating whether generation arises from a well-supported latent region or a latent support void, and uses neighborhood gradients as residual evidence of local decoding instability. These signals are integrated through evidential learning Sensoy et al. (2018) to estimate both class belief and evidence-aware uncertainty. Across knowledge-intensive question answering tasks, we find that hidden representations provide the dominant signal for distinguishing supported answers from False

Recall Hallucination, neighborhood gradients offer complementary gains, and evidential modeling better captures cases with weak internal support. We further perform controlled interventions along the identified factual-support direction and find that weakening this direction can increase hallucination in a pronounced case, with stronger effects than random and orthogonal control directions. We summarize our contributions as follows:

- We introduce False Recall Hallucination, describing fluent and familiar but factually unsupported generations in LLMs.
- We propose the Latent Void Hypothesis, providing a latent-space perspective for explaining hallucination formation.
- We develop HNGRE, which combines hidden activations, neighborhood-gradient residuals, and evidential learning to probe latent support and uncertainty.
- Extensive experiments reveal support-deficient latent regions and show that weakening an identified factual-support direction can causally increase hallucination in a case, providing observational and interventional evidence consistent with the Latent Void Hypothesis.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

Let f_θ denote a frozen autoregressive language model. Given a factual query q, the model generates an answer sequence $y = (y_1, \ldots, y_T)$ according to $p_\theta(y_t \mid q, y_{<t})$.

We focus on a knowledge-recall setting in which the model must answer a concrete factual query using only its internal parametric knowledge, without external retrieval or supporting evidence. Unlike factual hallucination in the broader sense, which generally refers to generated content that is inconsistent with real-world facts or cannot be verified Huang et al. (2025a), *False Recall Hallucination* characterizes a more specific reconstruction-like failure: when the queried fact is insufficiently supported, the model nevertheless produces a fluent, specific, and semantically plausible answer rather than abstaining or expressing uncertainty.

Let $\mathcal{A}(q)$ denote the set of valid reference answers. We assign a binary hallucination label

$$H(q, y) = \begin{cases} 0, & y \simeq \mathcal{A}(q), \\ 1, & \text{otherwise}, \end{cases} \tag{1}$$

where $\simeq$ denotes relaxed answer matching. We refer to $H = 0$ and $H = 1$ as KNOW and HALL, respectively.

This correctness-based label serves as an operational distinction between supported and unsupported factual recall. We further verify response fluency to exclude malformed generations. We do not assume that every incorrect generation exhibits all attributes of false recall; rather, the retrieval-free short-answer QA setting is designed to elicit such false-recall-like failures.

For each generated answer, we denote its associated latent representation as $z \in \mathcal{Z} \subseteq \mathbb{R}^d$. Our goal is to determine whether unsupported factual recall exhibits identifiable regions or measurable signatures in this latent space.

2.2 FORMALIZING THE LATENT VOID HYPOTHESIS

The **Latent Void Hypothesis** views hallucination as generation through latent regions in which the internal conditions required to constrain a reliable completion become underdetermined. For a query q, we denote such a region as a *latent void*

$$V_q \subset \mathcal{Z}. \tag{2}$$

A latent void should not be interpreted literally as an empty geometric region. Instead, it is defined functionally as a latent region whose internal organization differs systematically from its surrounding states and whose occurrence is associated with an increased likelihood of hallucination. We therefore formulate the hypothesis through two empirically testable conditions: an *observable signature* and a *hallucination association*.

Observable signature. Let

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_m\} \tag{3}$$

denote a family of observable functions defined over latent states, where each ϕ_j characterizes a measurable property of the model's internal representation, such as local geometry, density, sensitivity, or uncertainty.

If V_q constitutes a latent void, then at least one observable should exhibit a distributional discrepancy between states inside the void and those in its surrounding region:

$$\exists\, \phi_j \in \Phi : \quad \mathcal{D}\left(P(\phi_j(z) \mid z \in V_q), P(\phi_j(z) \mid z \in \mathcal{N}(V_q))\right) > \delta, \tag{4}$$

where $\mathcal{N}(V_q)$ denotes a local neighborhood surrounding V_q, $\mathcal{D}(\cdot, \cdot)$ is a statistical discrepancy measure between two distributions, and $\delta > 0$ denotes a separation margin.

This formulation does not require a latent void to be characterized by a unique observable. Different internal quantities may reveal different manifestations of the same underconstrained latent state, provided that they exhibit a systematic discrepancy between the void and its surrounding region.

Hallucination association. An anomalous latent region alone is insufficient to establish a hallucination-related void. The region should also be associated with an elevated probability of hallucination. We therefore require

$$\Delta_H = P(H = 1 \mid z \in V_q) - P(H = 1 \mid z \in R_q) > 0, \tag{5}$$

where R_q denotes a reference region corresponding to sufficiently constrained latent states outside the candidate void.

Together, these two conditions characterize a latent void as a region that is both internally distinguishable from its surrounding latent neighborhood and behaviorally associated with increased hallucination risk.

For **False Recall Hallucination**, we instantiate the Latent Void Hypothesis through *factual support*. In this setting, a latent void corresponds to a region in which general semantic associations remain available, while reliable support for the specific queried fact is weak, sparse, or inconsistent. The hypothesis therefore predicts that such generations should exhibit measurable internal signatures and that these signatures should be systematically associated with a higher probability of hallucination.

3 METHODOLOGY

3.1 OVERVIEW

To operationalize the Latent Void Hypothesis, we develop **HNGRE**, a Hidden-Neighborhood Gradient Residual Evidence framework. Given a query and its generated answer, HNGRE probes the internal states of a frozen language model and estimates whether the answer is supported by reliable internal evidence. HNGRE follows a three-stage evidence construction process. It first learns a base evidential model from answer-level hidden activations, which capture the latent state underlying the generated answer. The trained hidden learner is then frozen, and neighborhood-gradient features are introduced to model additional evidence associated with local decoding sensitivity. Finally, the two sources of evidence are accumulated and interpreted through a Dirichlet distribution to obtain the hallucination score and evidential uncertainty. Figure 2 summarizes the workflow of HNGRE, from hidden evidence learning to residual evidence accumulation and uncertainty estimation.

3.2 HIDDEN EVIDENCE LEARNER

We first learn the primary evidence from hidden activations. Let $h \in \mathbb{R}^{d_h}$ denote the answer-level hidden representation extracted from the frozen language model. In practice, hidden states of answer tokens are aggregated into a single representation and standardized using statistics computed from the training split.

An evidential network f_h maps this representation to non-negative class-wise evidence. For the two classes KNOW and HALL, we define

$$\mathbf{e}^h = \mathrm{Softplus}(f_h(h)), \qquad \boldsymbol{\alpha}^h = \mathbf{e}^h + \mathbf{1}. \tag{6}$$

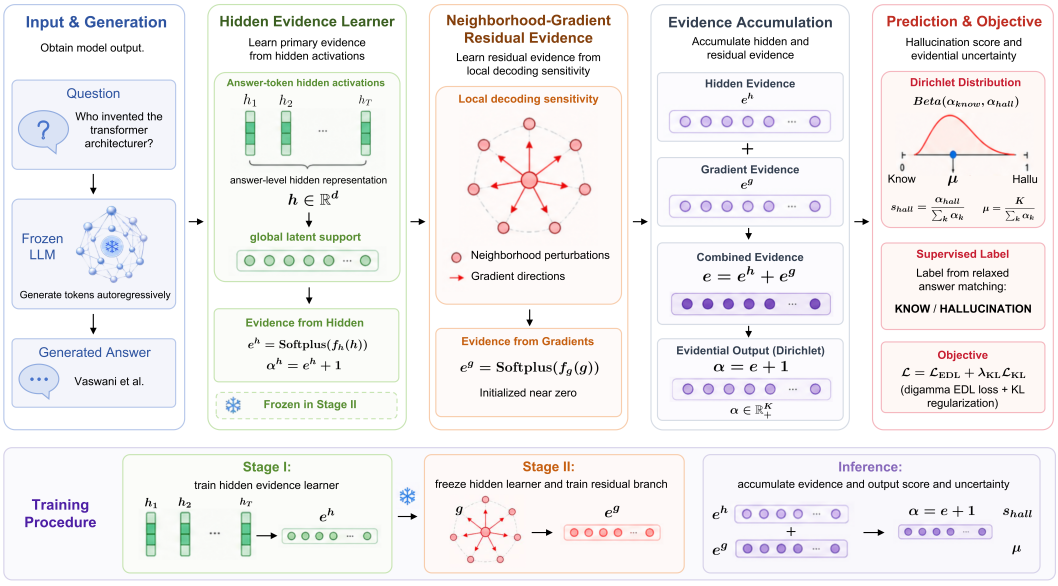

Figure 2: An overview of HNGRE. The hidden evidence learner first extracts primary evidence from answer-level hidden activations. The neighborhood-gradient residual branch then learns complementary evidence of local decoding sensitivity, and the accumulated evidence is converted into a Dirichlet distribution for hallucination prediction and evidential uncertainty estimation.

The resulting $\boldsymbol{\alpha}^h$ parameterizes a Dirichlet distribution. Its normalized expectation provides the class prediction, while the magnitude of the accumulated evidence reflects how strongly the hidden representation supports that prediction.

The hidden evidential learner is trained with the standard digamma-based evidential objective together with a KL regularizer that suppresses unsupported evidence assigned to non-target classes:

$$\mathcal{L} = \sum_{k=1}^{K} y_k \left[\psi(S) - \psi(\alpha_k)\right] + \lambda_{\mathrm{KL}} D_{\mathrm{KL}}\left[\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}) \| \, \mathrm{Dir}(\mathbf{1})\right], \tag{7}$$

where $S = \sum_k \alpha_k$ and $\tilde{\boldsymbol{\alpha}} = \mathbf{y} + (1 - \mathbf{y}) \odot \boldsymbol{\alpha}$. After convergence, the parameters of the hidden evidential learner are fixed, and its output $\mathbf{e}^h$ serves as the base evidence for the next stage.

3.3 Neighborhood-Gradient Residual Evidence

The hidden representation describes the latent state from which the answer is generated, while neighborhood-gradient features provide a complementary view of its local decoding behavior. Let $g \in \mathbb{R}^{d_g}$ denote the neighborhood-gradient representation associated with the generated answer.

During this stage, the hidden evidential learner remains frozen. A residual branch f_g is trained to produce additional non-negative evidence from the gradient representation,

$$\mathbf{e}^g = \mathrm{Softplus}(f_g(g)). \tag{8}$$

The final layer of this branch is initialized so that its initial output is close to zero. Training therefore starts from the hidden-only evidential solution, while the gradient branch progressively learns complementary evidence from local decoding behavior.

For each sample, HNGRE accumulates the two evidence sources directly in evidence space:

$$\mathbf{e} = \mathbf{e}^h + \mathbf{e}^g, \qquad \boldsymbol{\alpha} = \mathbf{e} + \mathbf{1}. \tag{9}$$

The residual branch is optimized using the same evidential objective defined above, with $\boldsymbol{\alpha}$ now constructed from the accumulated hidden and gradient evidence. Only the residual branch is updated during this stage.

At inference time, the expected Dirichlet probability of the HALL class is used as the hallucination score, while the total Dirichlet strength provides an explicit measure of evidence deficiency:

$$s_{\text{hall}} = \frac{\alpha_{\text{hall}}}{\sum_k \alpha_k}, \qquad u = \frac{K}{\sum_k \alpha_k}. \tag{10}$$

A high s_{hall} indicates stronger evidence for hallucination, whereas a high u indicates that the detector has accumulated little evidence overall.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We evaluate HNGRE on four open-domain factual question answering datasets: PopQA Mallen et al. (2023), WebQuestions Berant et al. (2013), TriviaQA Joshi et al. (2017), and the English subset of MKQA Longpre et al. (2021). All experiments follow a zero-shot short-answer generation setting, where the language model receives only the question and generates a concise factual answer without external retrieval. Generated answers are matched against the provided gold answers or aliases and labeled as KNOW or HALL. For each dataset–model pair, we further construct a balanced evaluation set with a 1:1 ratio between the two classes.

We report AUROC Boyd et al. (2013), AUPR Manning et al. (1999), and FPR@95. AUROC measures the ranking separability between KNOW and HALL, AUPR treats hallucination as the positive class, and FPR@95 measures the false-positive rate on supported answers when hallucination recall is fixed at 95%.

Models. We consider six open-source instruction- or chat-tuned LLMs in the 7B–9B range: Llama2-7B Touvron et al. (2023), Llama3-8B Meta (2024), Qwen2-7B Yang et al. (2024), Qwen2.5-7B Hui et al. (2024), Gemma2-9B Team et al. (2024), and DeepSeek-LLM-7B Bi et al. (2024). Together with the four datasets, this yields 24 dataset–model evaluation cases. Across all cases, the original generation pool contains 237,738 valid samples, from which 183,996 balanced samples are used for supervised training and evaluation. Specific sample information is in the Appendix.

Baselines. We compare HNGRE with both training-free uncertainty estimators and supervised hidden-state probes. Training-free baselines include MSP, Perplexity Si et al. (2022), Entropy Huang et al. (2025b), Prompt-Sampling-Aggregation (PSA) Xiong et al. (2024), EigenScore Chen et al. (2024b), and CoE-C Wang et al. (2025a) scores. Supervised baselines include LR Probe Alain & Bengio (2016) and SAPLMA Azaria & Mitchell (2023b). More setup is in Appendix.

4.2 MAIN RESULTS

Overall Performance. Table 1 summarizes the hallucination detection performance across 24 dataset–model cases. HNGRE achieves consistently strong results across different model families and factual QA benchmarks, obtaining the highest AUROC in 22 out of 24 cases. Averaged over all cases, HNGRE reaches 0.8403 AUROC and 0.8393 AUPR. Compared with SAPLMA, the strongest supervised baseline overall, HNGRE improves the average AUROC and AUPR by 3.77% and 3.54%, respectively, while reducing the average FPR@95 from 0.6551 to 0.6129. Compared with LR Probe, the improvements further increase to 9.46% in AUROC and 10.00% in AUPR. Even in the few configurations where HNGRE does not achieve the best result, its performance remains close to the strongest baseline.

Comparison with Baselines. The results reveal a clear advantage of internal-state-based methods over conventional output-level uncertainty signals. MSP, perplexity, entropy, PSA, EigenScore, and CoE-C can provide informative uncertainty estimates in individual settings, but their performance varies considerably across models and datasets. Supervised internal probes such as LR Probe and SAPLMA generally show stronger discrimination, indicating that hallucination-related information is already encoded in the internal states of the model. HNGRE further improves this discrimination by incorporating both hidden-state evidence and neighborhood-gradient residual evidence.

Method	Llama2-7B-Chat			Llama3-8B-Instruct			Qwen2-7B-Instruct			Qwen2.5-7B-Instruct			Gemma2-9B-It			DeepSeek-7B-Chat		
	AUC↑	AUPR↑	FPR@95↓	AUC↑	AUPR↑	FPR@95↓	AUC↑	AUPR↑	FPR@95↓	AUC↑	AUPR↑	FPR@95↓	AUC↑	AUPR↑	FPR@95↓	AUC↑	AUPR↑	FPR@95↓
PopQA																		
MSP	0.7685	0.7188	0.6707	0.6976	0.6560	0.7791	0.7728	0.7217	0.7051	0.7569	0.7109	0.5574	0.8464	0.8417	0.5839	0.7828	0.7519	0.5753
Perplexity	0.7754	0.7401	0.6707	0.7198	0.7029	0.7907	0.7674	0.7271	0.7756	0.7710	0.7419	0.5656	0.8609	0.8678	0.5839	0.8029	0.7731	0.5616
Entropy	0.8007	0.7729	0.5976	0.7541	0.7549	0.7151	0.8372	0.8079	0.6026	0.7818	0.7697	0.5082	0.8503	0.8597	0.6273	0.8459	0.8210	0.4658
PSA	0.7388	0.7074	1.0000	0.7278	0.7042	0.7500	0.6952	0.6464	0.7115	0.8210	0.8210	1.0000	0.8120	0.8114	1.0000	0.7488	0.6878	0.6096
EigenScore	0.7903	0.7867	1.0000	0.7402	0.7373	0.7733	0.7553	0.7001	0.6410	0.8366	0.8500	1.0000	0.8182	0.8391	0.7329	0.7700	0.7555	0.6781
CoE-C	0.4614	0.4740	0.9146	0.6053	0.6011	0.8488	0.5247	0.5465	0.8910	0.6975	0.6336	0.7295	0.6089	0.5701	0.9689	0.7103	0.6845	0.6986
LR Probe	0.8840	0.8764	0.4878	0.8586	0.8579	0.5233	0.8785	0.8506	0.5000	0.9318	0.9293	0.3115	**0.9030**	0.8796	**0.3106**	0.9152	0.9196	**0.3562**
SAPLMA	0.8805	0.8672	**0.3841**	0.8229	0.8093	0.6047	0.8852	0.8977	**0.4359**	0.9163	0.9237	0.3525	0.8911	**0.8995**	0.4596	0.9035	0.9105	0.3630
HNGRE	**0.9044**	**0.9076**	0.3963	**0.8892**	**0.8908**	**0.4360**	**0.9220**	**0.9338**	**0.4359**	**0.9459**	**0.9434**	**0.3033**	0.8969	0.8964	0.4410	**0.9237**	**0.9261**	0.3836
WebQuestions																		
MSP	0.6579	0.6580	0.8649	0.6455	0.6568	0.9311	0.6430	0.6453	0.8680	0.5943	0.6106	0.9316	0.5968	0.6461	0.9509	0.6660	0.6773	0.9103
Perplexity	0.6572	0.6616	0.8757	0.7283	0.7723	0.8954	0.6502	0.6711	0.8807	0.6011	0.6226	0.9231	0.5917	0.6410	0.9535	0.6845	0.7093	0.9000
Entropy	0.6652	0.6733	0.8568	0.6640	0.6818	0.9031	0.6536	0.6579	0.8731	0.6107	0.6299	0.9288	0.6266	0.6655	0.9535	0.6808	0.6981	0.9128
PSA	0.6306	0.6231	1.0000	0.6620	0.6557	1.0000	0.6211	0.5915	1.0000	0.6179	0.6107	1.0000	0.6345	0.6451	1.0000	0.6612	0.6514	0.9051
EigenScore	0.6624	0.6458	0.8000	0.6725	0.6686	0.8750	0.6276	0.6104	1.0000	0.5998	0.5937	1.0000	0.5867	0.6235	0.9587	0.6553	0.6512	0.9256
CoE-C	0.6205	0.6043	0.9108	0.6086	0.6265	0.9388	0.5591	0.5703	0.9569	0.5346	0.5672	0.9516	0.4833	0.4862	0.9819	0.5932	0.6032	0.9410
LR Probe	0.7160	0.7171	0.9000	0.7295	0.7358	0.9260	0.6719	0.6711	0.9239	0.7251	0.7253	0.8746	0.7739	0.7641	0.7571	0.6942	0.6896	0.9026
SAPLMA	0.7763	0.8048	0.7351	0.8045	0.8100	0.6837	0.7454	0.7399	0.7386	0.7633	0.7643	0.7436	0.7813	0.7673	0.7235	0.7579	0.7529	0.7179
HNGRE	**0.8083**	**0.8236**	**0.7000**	**0.8609**	**0.8649**	**0.5740**	**0.8113**	**0.8002**	**0.6624**	**0.8294**	**0.8332**	**0.6353**	**0.8359**	**0.8300**	**0.7080**	**0.7920**	**0.7695**	**0.6974**
TriviaQA																		
MSP	0.6371	0.6130	0.8599	0.6846	0.6867	0.8353	0.6723	0.6410	0.8184	0.7817	0.7751	0.7206	0.7105	0.7381	0.8946	0.6872	0.6730	0.8494
Perplexity	0.6450	0.6268	0.8586	0.7038	0.7108	0.8174	0.6779	0.6541	0.8418	0.7843	0.7811	0.7214	0.7024	0.7312	0.8954	0.7022	0.6890	0.8416
Entropy	0.6626	0.6538	0.8619	0.7199	0.7337	0.8462	0.7047	0.6991	0.8370	0.7930	0.7976	0.7214	0.7467	0.7782	0.8946	0.7111	0.7003	0.8564
PSA	0.6142	0.5890	1.0000	0.7047	0.7074	1.0000	0.6433	0.6401	1.0000	0.7745	0.7637	1.0000	0.7513	0.7435	1.0000	0.6461	0.6211	0.8317
EigenScore	0.6796	0.6733	1.0000	0.7119	0.7015	0.8260	0.7109	0.7003	0.8404	0.7676	0.7643	1.0000	0.7290	0.7470	0.8731	0.6766	0.6442	0.8564
CoE-C	0.6125	0.6062	0.8894	0.5804	0.5890	0.9425	0.6634	0.6657	0.8879	0.6046	0.6137	0.9241	0.5104	0.5323	0.9572	0.6502	0.6572	0.9031
LR Probe	0.6988	0.7170	0.9048	0.6911	0.6846	0.9340	0.7244	0.7301	0.8920	0.7749	0.7875	0.8512	0.7172	0.7321	0.8674	0.6941	0.6930	0.9059
SAPLMA	0.7873	**0.7840**	0.8003	0.7742	0.7693	0.7630	0.8036	0.7950	0.7531	0.8569	0.8648	0.6331	0.7822	**0.7963**	**0.7949**	0.7864	0.7769	**0.7348**
HNGRE	**0.7878**	0.7804	**0.7889**	**0.7845**	**0.7808**	**0.7218**	**0.8125**	**0.8175**	**0.7490**	**0.8755**	**0.8832**	**0.5813**	**0.7866**	0.7856	0.8040	**0.7986**	**0.7902**	0.7362
MKQA																		
MSP	0.6864	0.6691	0.8010	0.6686	0.6298	0.8541	0.7001	0.6945	0.8291	0.7078	0.7154	0.8923	0.6770	0.7024	0.9159	0.6983	0.6964	0.8912
Perplexity	0.6881	0.6755	0.8134	0.7233	0.7199	0.8071	0.6982	0.6937	0.8453	0.7249	0.7404	0.8831	0.6767	0.7066	0.9204	0.7096	0.7216	0.8853
Entropy	0.6979	0.6949	0.8035	0.6922	0.6714	0.8894	0.7093	0.7147	0.8106	0.7115	0.7187	0.8892	0.6991	0.7281	0.9115	0.7098	0.7154	0.8853
PSA	0.6425	0.6145	0.8433	0.5978	0.5953	0.8847	0.6673	0.6439	0.8152	0.6926	0.6904	1.0000	0.7274	0.7336	1.0000	0.6513	0.6415	0.8412
EigenScore	0.6689	0.6537	0.8209	0.6897	0.6571	0.8259	0.6943	0.6655	0.8245	0.7038	0.6978	1.0000	0.6861	0.7090	0.9071	0.6715	0.6639	0.8765
CoE-C	0.6160	0.5967	0.9204	0.6381	0.6471	0.8612	0.6066	0.5934	0.9007	0.5686	0.5704	0.8646	0.4810	0.4706	0.8960	0.6398	0.6492	0.9176
LR Probe	0.7320	0.6844	0.8159	0.7106	0.7144	0.9318	0.7107	0.6905	0.9122	0.7864	0.7794	0.7692	0.7809	0.7771	0.7544	0.7224	0.7067	0.8941
SAPLMA	0.7554	0.7369	0.7015	**0.8088**	**0.8102**	**0.6541**	0.7585	0.7640	0.7760	0.8207	0.8200	0.6862	0.7953	0.8005	0.7212	0.7765	0.7884	0.7618
HNGRE	**0.7981**	**0.7934**	**0.6269**	0.7901	0.7703	0.7929	**0.8136**	**0.8096**	**0.6767**	**0.8596**	**0.8629**	**0.5538**	**0.8275**	**0.8297**	**0.6571**	**0.8135**	**0.8203**	**0.6471**

Table 1: Hallucination detection performance across datasets and LLMs. Results are reported on balanced KNOW/HALL sets. Higher AUROC and AUPR are better, while lower FPR@95 is better. Bold values indicate the best result in each block.

The improvement is particularly evident on WebQuestions and TriviaQA, where HNGRE achieves the highest AUROC for all six evaluated models. On PopQA and MKQA, it obtains the highest AUROC in five out of six model cases. These results indicate that the internal evidence captured by HNGRE remains effective across substantially different factual QA settings.

Metric Consistency. HNGRE also exhibits consistent behavior across the three evaluation metrics. AUROC measures threshold-independent separability, AUPR emphasizes the detection of hallucinated samples as the positive class, and FPR@95 reflects the false-positive cost on supported answers under a high-recall operating point. HNGRE shows particularly stable improvements in AUROC and AUPR, while FPR@95 presents greater variation across configurations.

Robustness across Datasets and Models. Across datasets, HNGRE achieves average AUROC values of 0.9137 on PopQA, 0.8230 on WebQuestions, 0.8076 on TriviaQA, and 0.8171 on MKQA. Although the absolute difficulty varies across benchmarks, the method maintains stable discrimination without a pronounced degradation on any dataset.

A similar pattern appears across model families. HNGRE performs effectively on Llama, Qwen, Gemma, and DeepSeek models, covering multiple independently trained 7B–9B instruction/chat architectures. Particularly strong results are observed on Qwen2.5, where HNGRE reaches AUROC values of 0.9459, 0.8294, 0.8755, and 0.8596 on PopQA, WebQuestions, TriviaQA, and MKQA, respectively. Overall, the results show that the hallucination-related internal signals captured by HNGRE are stable across both dataset distributions and model architectures.

4.3 Extended Analysis

Intervening on the Factual-Support Direction. To move beyond correlational evidence, we further ask whether directly perturbing the identified factual-support direction can causally alter hallucination behavior. We intervene at a fixed intermediate-to-late layer ($0.8L$) and vary the intervention strength $\alpha \in [-1.5, 1.5]$. As shown in Figure 3(a), most of the valid model–dataset cases exhibit only limited changes under this one-dimensional intervention. In contrast, Qwen2.5–PopQA shows a clear asymmetric dose-response: increasingly negative interventions progressively increase the

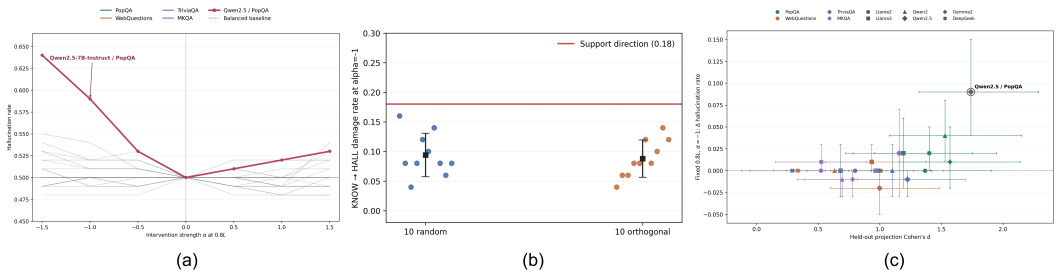

Figure 3: Intervention analysis of the factual-support direction. (a) Hallucination-rate changes under different intervention strengths at layer $0.8L$. (b) KNOW→HALL damage induced by the factual-support direction compared with random and orthogonal controls. (c) Relationship between held-out Cohen's d along the factual-support direction and the hallucination-rate change.

hallucination rate, whereas positive interventions do not yield a comparably stable recovery. This asymmetry suggests that weakening an existing factual-support structure can be considerably easier than restoring missing factual support through a single latent direction.

The absence of a universal intervention effect does not necessarily contradict the Latent Void Hypothesis. A complex latent space need not organize factual support along a single globally steerable axis; the relevant structure may be local, distributed across multiple dimensions, and dependent on the model and knowledge distribution. The Qwen2.5–PopQA result instead provides a case in which weakening a support-related latent direction produces the predicted behavioral consequence.

We further test whether this effect can be attributed to generic hidden-state perturbation. As shown in Figure 3(b), at $\alpha = -1$ the factual-support direction induces a KNOW→HALL damage rate of approximately 0.18, substantially larger than the average effects of ten random directions and ten orthogonal directions. This control experiment indicates that the observed degradation is direction-specific rather than a general consequence of perturbing the hidden representation.

Finally, Figure 3(c) relates the held-out Cohen's d (Eq. 17 in Appendix) along the factual-support direction to the corresponding intervention effect. A broad tendency emerges in which stronger one-dimensional KNOW/HALL separation is more often accompanied by a larger behavioral response. Qwen2.5–PopQA is the clearest example, exhibiting both strong linear separability and the largest increase in hallucination rate, suggesting that factual-support information in this case is highly decodable and susceptible to directional intervention. However, several cases with relatively large Cohen's d remain only weakly steerable, indicating that linear decodability may favor but does not guarantee steerability. This suggests that factual support may be concentrated along a dominant direction in some model–dataset cases, while distributed across a more complex multidimensional structure in others. Accordingly, our intervention is not intended to establish a universal one-dimensional control mechanism; rather, the successful case provides existence evidence that a support-related latent structure can exert the predicted causal effect on factual generation.

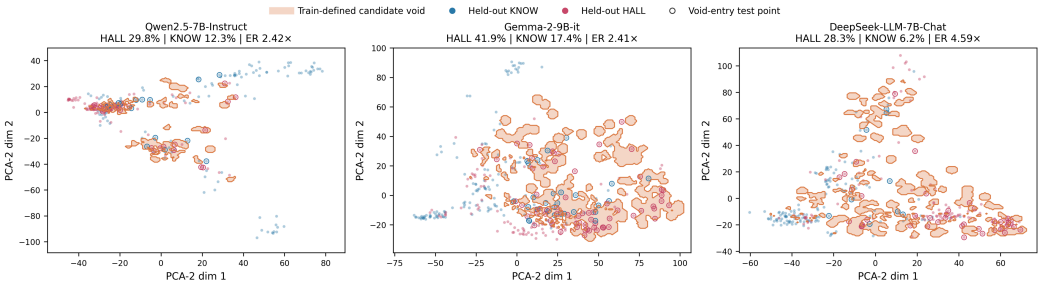

Figure 4: Train-defined latent-void candidates for three representative LLMs. Regions are defined in the original hidden space without HALL labels; PCA is used only for visualization.

Factual-Support Deficiency and Latent Voids. The strong contribution of hidden representations motivates us to examine whether hallucinations are associated with latent regions where factual support is weak but the model still frequently produces internal states. Using training representations only, we define candidate voids as regions with low support relative to KNOW states but non-negligible overall latent occupancy. These candidate regions should be viewed as observable proxies for the theoretical latent voids rather than as direct identifications of the voids themselves. Importantly, HALL labels are not used in constructing these regions, and the same quantile rule is applied across all models. All support, density, and void-membership computations are performed in the original high-dimensional hidden space; the 2D PCA projection is used only to visualize one manifestation of this structure. Formal definitions and more analyses are provided in the Appendix.

Figure 4 shows three representative cases. In all three models, held-out hallucinated generations enter the train-defined candidate regions more frequently than supported generations. The enrichment ratio $\mathrm{ER} = P(z \in \mathcal{V} \mid \mathrm{HALL})/P(z \in \mathcal{V} \mid \mathrm{KNOW})$ reaches $2.42\times$, $2.41\times$, and $4.59\times$ for Qwen2.5, Gemma-2, and DeepSeek, respectively. Across all six models, $\mathrm{ER} > 1$, with a macro-average of $2.71\times$. Thus, these observable support-deficient regions exhibit systematically elevated hallucination risk, providing a measurable signature consistent with the Latent Void Hypothesis.

5 Related Works

5.1 Hallucination Detection and Uncertainty Estimation.

Recent studies have developed diverse signals for identifying unreliable LLM generations. Output-level approaches estimate hallucination risk from token probabilities, sequence likelihoods, or uncertainty measures, while black-box methods exploit consistency across sampled responses. SelfCheckGPT detects factual inconsistencies by comparing stochastic generations Manakul et al. (2023), whereas semantic entropy groups semantically equivalent responses and measures uncertainty over meaning rather than surface forms Farquhar et al. (2024). Other work studies model knowledge boundaries and uses answerability or confidence to identify queries beyond the model's reliable knowledge Wen et al. (2024). These methods establish uncertainty and consistency as useful hallucination indicators, but mainly measure whether an output is unreliable, leaving the associated internal representational conditions less directly characterized.

5.2 Internal Representations for Hallucination Analysis.

Recent work increasingly examines hallucination through LLM internal representations. Hidden activations have been shown to encode truthfulness signals useful for supervised probing Azaria & Mitchell (2023a). INSIDE exploits internal semantic representations and introduces EigenScore to measure representation-space consistency Chen et al. (2024a), while MIND enables annotation-free hallucination detection from internal states Su et al. (2024). Beyond static representations, Chain-of-Embedding models hidden-state evolution for output-free self-evaluation Wang et al. (2025b), and ICR Probe studies how intermediate modules update the residual stream Zhang et al. (2025). Gradient-based approaches such as GRADE further characterize knowledge gaps through cross-layer gradient-subspace dynamics Wang et al. (2026). Collectively, these studies show that hallucination and response correctness leave measurable internal signatures. Rather than using such signals solely for detection, we propose the **Latent Void Hypothesis** to explain hallucination through under-constrained latent regions, and probe this phenomenon using hidden representations, neighborhood-gradient residuals, and evidential uncertainty.

6 CONCLUSION

In this work, we introduce **False Recall Hallucination** and propose the **Latent Void Hypothesis**, which attributes hallucination to latent regions where the internal conditions required to constrain a reliable completion are underdetermined. For factual hallucination, we instantiate this hypothesis through factual support and develop HNGRE using hidden representations, neighborhood-gradient residuals, and evidential learning. Experiments reveal support-deficient latent regions associated with hallucination, while controlled intervention shows that weakening a factual-support direction can increase hallucination in a pronounced case. Together, these results provide evidence consistent with the Latent Void Hypothesis and motivate a broader latent-space perspective on hallucination.

AI DISCLOSURE

In this work, we used GPT-5.5 to assist with and polish the writing of the manuscript. We did not use generative AI tools to generate experimental results, conduct data analysis, or independently formulate scientific claims. All AI-assisted writing was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The datasets and models used in this work are publicly available and are used in accordance with their intended research purposes. This study does not involve human subjects, private user data, or personally identifiable information. Our research aims to understand and detect unsupported factual generations in LLMs, with potential benefits for improving the reliability and safety of language-model deployment.

REPRODUCIBILITY

All models used in our experiments are official checkpoints, and all datasets are publicly available. We provide the experimental settings, data construction and labeling procedures, model configurations, training details, intervention settings, evaluation metrics, and random seed necessary to reproduce the reported results.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

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, 2023a.

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, 2023b.

Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, Cambridge, 1995.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Elijah Berberette, Jack Hutchins, and Amir Sadovnik. Redefining" hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation. *arXiv preprint arXiv:2402.01769*, 2024.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 451–466. Springer, 2013.

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*, 2024a.

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. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024b.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Angela H Gutchess and Daniel L Schacter. The neural correlates of gist-based true and false recognition. *Neuroimage*, 59(4):3418–3426, 2012.

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, 2025a.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429, 2025b.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104, 2024.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. *URL https://arxiv. org/abs/2109.07958*, 1, 2021.

Elizabeth F Loftus and Jacqueline E Pickrell. The formation of false memories, 1995.

Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021.

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.

Christopher D Manning, Hinrich Schütze, and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT press Cambridge, 1999.

AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*, 2(5): 6, 2024.

Henry L Roediger and Kathleen B McDermott. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4): 803, 1995.

Daniel L Schacter. Constructive memory: past and future. *Dialogues in clinical neuroscience*, 14 (1):7–18, 2012.

Daniel L Schacter and Donna Rose Addis. The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):773, 2007.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*, 2022.

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, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. Latent space chain-of-embedding enables output-free llm self-evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 70938–70970, 2025a.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. Latent space chain-of-embedding enables output-free llm self-evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 70938–70970, 2025b.

Yujing Wang, Yuanbang Liang, Yukun Lai, Hainan Zhang, and Hanqi Yan. Grade: Probing knowledge gaps in llms through gradient subspace dynamics. *arXiv preprint arXiv:2604.02830*, 2026.

Zhihua Wen, Zhiliang Tian, Zexin Jian, Zhen Huang, Pei Ke, Yifu Gao, Minlie Huang, and Dongsheng Li. Perception of knowledge boundary for large language models through semi-open-ended question answering. *Advances in Neural Information Processing Systems*, 37:88906–88931, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report, 2024. URL `https://arxiv.org/abs/2407.10671`.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17986–18002, 2025.

A APPENDIX

A.1 SAMPLE INFORMATION

We construct the experimental samples from four factual question answering datasets and six LLMs, resulting in 24 dataset–model cases. For each query, the model first generates an answer without external retrieval. The generated answer is then matched against the corresponding gold answer or answer aliases. A matched answer is labeled as KNOW, while an unmatched answer is labeled as HALL. Since the raw class distribution varies substantially across datasets and models, we construct a balanced subset for each case by randomly sampling the majority class to obtain a 1:1 KNOW/HALL ratio.

Dataset	Source Size	Used	Sampling	Answer Matching
PopQA	14,267	7,193	$s_{\text{pop}} \leq 1000$	Normalized substring
WebQuestions	5,810	5,810	Full dataset	Relaxed alias matching
TriviaQA	109,767	20,000	Random, seed 42	Relaxed alias matching
MKQA-en	6,620	6,620	English subset	Relaxed alias matching

Table 2: Dataset statistics and sampling strategies used in our experiments.

Across the 24 cases, we obtain 237,738 valid raw generations, including 99,406 KNOW and 138,332 HALL samples. After case-wise balancing, the final pool contains 183,996 samples, with 91,998 examples in each class. Balancing is performed independently for each dataset–model case to prevent cases with strongly skewed raw label distributions from dominating the supervised training and evaluation.

Dataset	Model	Raw	KNOW	HALL	Balanced (K/H)
PopQA	DeepSeek-7B	7,193	969	6,224	1,938 (969/969)
	Gemma2-9B	7,193	1,069	6,124	2,138 (1,069/1,069)
	Llama2-7B	7,193	1,093	6,100	2,186 (1,093/1,093)
	Llama3-8B	7,193	1,144	6,049	2,288 (1,144/1,144)
	Qwen2.5-7B	7,193	808	6,385	1,616 (808/808)
	Qwen2-7B	7,193	1,038	6,155	2,076 (1,038/1,038)
WebQuestions	DeepSeek-7B	5,810	2,596	3,214	5,192 (2,596/2,596)
	Gemma2-9B	5,810	2,575	3,235	5,150 (2,575/2,575)
	Llama2-7B	5,810	2,461	3,349	4,922 (2,461/2,461)
	Llama3-8B	5,810	2,609	3,201	5,218 (2,609/2,609)
	Qwen2.5-7B	5,810	2,336	3,474	4,672 (2,336/2,336)
	Qwen2-7B	5,810	2,624	3,186	5,248 (2,624/2,624)
TriviaQA	DeepSeek-7B	20,000	9,421	10,579	18,842 (9,421/9,421)
	Gemma2-9B	20,000	11,912	8,088	16,176 (8,088/8,088)
	Llama2-7B	20,000	10,054	9,946	19,892 (9,946/9,946)
	Llama3-8B	20,000	11,425	8,575	17,150 (8,575/8,575)
	Qwen2.5-7B	20,000	9,134	10,866	18,268 (9,134/9,134)
	Qwen2-7B	20,000	10,313	9,687	19,374 (9,687/9,687)
MKQA	DeepSeek-7B	6,620	2,263	4,357	4,526 (2,263/2,263)
	Gemma2-9B	6,620	3,010	3,610	6,020 (3,010/3,010)
	Llama2-7B	6,620	2,676	3,944	5,352 (2,676/2,676)
	Llama3-8B	6,620	2,827	3,793	5,654 (2,827/2,827)
	Qwen2.5-7B	6,620	2,166	4,454	4,332 (2,166/2,166)
	Qwen2-7B	6,620	2,883	3,737	5,766 (2,883/2,883)
Total	–	**237,738**	**99,406**	**138,332**	**183,996 (91,998/91,998)**

Table 3: Sample statistics for all 24 dataset–model cases. "Balanced" denotes the 1:1 subset used for subsequent training and evaluation.

Representative KNOW and HALL samples are provided below for all dataset–model combinations. Each row reports the question, model-generated response, and reference answer for one example from each class. Many hallucinated responses remain fluent, specific, and semantically plausible

Model	KNOW example	HALL example
Llama-2-7B-Chat	**Q:** What sport does 2014–15 Premier League of Bosnia and Herzegovina play? **Gen.:** The 2014-15 Premier League of Bosnia and Herzegovina plays football (soccer). **Ref.:** soccer; football; association football	**Q:** Who was the director of La liceale al mare con l'amica di papà? **Gen.:** The director of "La liceale al mare con l'amica di papà" is Franco Brusati. **Ref.:** Marino Girolami
Llama-3-8B-Instruct	**Q:** What sport does 1909–10 Netherlands Football League Championship play? **Gen.:** Football (soccer). **Ref.:** soccer; association football	**Q:** Who was the composer of It's Easy to Remember (And So Hard to Forget)? **Gen.:** The composer of "It's Easy to Remember (And So Hard to Forget)" is Slim Whitman. **Ref.:** Richard Rodgers; Richard Charles Rodgers
Qwen2-7B-Instruct	**Q:** What sport does Canton Bulldogs–Massillon Tigers betting scandal play? **Gen.:** American football **Ref.:** American football; football; American rules football	**Q:** Who is the author of The Merry Men and Other Tales and Fables? **Gen.:** The author of "The Merry Men and Other Tales and Fables" is Rudyard Kipling. **Ref.:** Robert Louis Stevenson; Robert Lewis Balfour Stevenson; Robert Luis Stivensoni
Qwen2.5-7B-Instruct	**Q:** What sport does 1989–90 1. Slovenská národná hokejová liga season play? **Gen.:** Ice hockey **Ref.:** ice hockey	**Q:** Who was the screenwriter for The Worst Years of Our Lives? **Gen.:** Robert Riskin **Ref.:** David Trueba
Gemma-2-9B-it	**Q:** Who was the composer of It's Easy to Remember (And So Hard to Forget)? **Gen.:** Richard Rodgers. **Ref.:** Richard Rodgers; Richard Charles Rodgers	**Q:** Who is the author of Visionseeker: Shared Wisdom from the Place of Refuge? **Gen.:** Dr. Wayne Dyer **Ref.:** Hank Wesselman; Henry Barnard Wesselman
DeepSeek-LLM-7B-Chat	**Q:** What sport does 1938 Wimbledon Championships – Women's Singles play? **Gen.:** The 1938 Wimbledon Championships – Women's Singles was a tennis tournament. **Ref.:** tennis; lawn tennis; lawntennis	**Q:** Who is the author of The Secret Diaries of Miss Miranda Cheever? **Gen.:** The author of The Secret Diaries of Miss Miranda Cheever is Eloisa James. **Ref.:** Julia Quinn; Julie Cotler; Julie Pottinger

Table 4: Representative KNOW and HALL examples on PopQA.

Model	KNOW example	HALL example
Llama-2-7B-Chat	**Q:** under what political system does the united states government operate? **Gen.:** The United States government operates under a constitutional federal republic. **Ref.:** Federal republic; Republic; Constitutional republic	**Q:** who was the secretary of state when andrew jackson was president? **Gen.:** The Secretary of State during Andrew Jackson's presidency was John Quincy Adams. **Ref.:** Martin Van Buren
Llama-3-8B-Instruct	**Q:** what kind of government did the united states have after the revolution? **Gen.:** The United States had a federal republic government after the Revolution. **Ref.:** Federal republic; Republic; Constitutional republic	**Q:** what state did roger sherman represent at the constitutional convention? **Gen.:** North Carolina. **Ref.:** Connecticut
Qwen2-7B-Instruct	**Q:** what type of monarchy did england have during the glorious revolution? **Gen.:** England had a constitutional monarchy during the Glorious Revolution. **Ref.:** Constitutional monarchy	**Q:** who was the secretary of state when andrew jackson was president? **Gen.:** John C. Calhoun was the Secretary of State when Andrew Jackson was president. **Ref.:** Martin Van Buren
Qwen2.5-7B-Instruct	**Q:** what book did john steinbeck wrote about the people in the dust bowl? **Gen.:** The Grapes of Wrath **Ref.:** The Grapes of Wrath	**Q:** who did gerald ford select as his vice president when he became president? **Gen.:** Dick Cheney **Ref.:** Nelson Rockefeller
Gemma-2-9B-it	**Q:** where is the nfl hall of fame located and when was the building opened? **Gen.:** Canton, Ohio; 1963 **Ref.:** Canton	**Q:** who played stanley kowalski in a streetcar named desire movie? **Gen.:** Karl Malden **Ref.:** Marlon Brando
DeepSeek-LLM-7B-Chat	**Q:** what school did martin luther king jr received his doctorate degree from? **Gen.:** Martin Luther King Jr. received his doctorate degree from Boston University. **Ref.:** Boston University	**Q:** where was greenpeace an international environmental organization founded? **Gen.:** Greenpeace was founded in Amsterdam, Netherlands. **Ref.:** Vancouver

Table 5: Representative KNOW and HALL examples on WebQuestions.

despite being factually inconsistent with the reference, illustrating the False Recall Hallucination behavior studied in this work.

Model	KNOW example	HALL example
Llama-2-7B-Chat	**Q:** Juliana was the queen of which European country between 1948 and 1980? **Gen.:** Juliana was the queen of the Netherlands between 1948 and 1980. **Ref.:** The Netherlands; Netherlands; The netherland	**Q:** The collective noun for which British mammal is a 'cete'? [say "Seat"] **Gen.:** A 'cete' is a collective noun for a group of otters. **Ref.:** Badger; Badger (animal); Melinae
Llama-3-8B-Instruct	**Q:** The Koran contains 114 sections; by that name are these usually known? **Gen.:** Suras. **Ref.:** SURAS; Sura; Surah	**Q:** Which British comedian describes himself as an executive transvestite? **Gen.:** Alan Carr. **Ref.:** Eddie Izzard; Eddie Izard; Jazz chicken
Qwen2-7B-Instruct	**Q:** In a mobile phone, for what do the letters SIM, as in SIM card, stand? **Gen.:** Subscriber Identity Module **Ref.:** SUBSCRIBER IDENTITY MODULE; Subscriber Identity Module Card; Universal Subscriber Identity Module	**Q:** In medieval England what name was given to an axe head on a long pole? **Gen.:** The name given to an axe head on a long pole in medieval England is a "bill." **Ref.:** Halberd; Halbert; Swiss voulges
Qwen2.5-7B-Instruct	**Q:** On the internet what is the derogatory retronym for normal paper mail? **Gen.:** Snail mail **Ref.:** Snail Mail; P-mail; Snailmail	**Q:** Which musical comedy features the characters 'Bloom' and 'Bialystock'? **Gen.:** Spring Awakening **Ref.:** 'THE PRODUCERS'; producers; Produced
Gemma-2-9B-it	**Q:** Which musical comedy features the characters 'Bloom' and 'Bialystock'? **Gen.:** The Producers. **Ref.:** 'THE PRODUCERS'; producers; The producer	**Q:** What was the first considerable engagement in the American Revolution? **Gen.:** The Battle of Lexington and Concord. **Ref.:** Bunker Hill; Bunkers Hill; Bunker Hill (disambiguation)
DeepSeek-LLM-7B-Chat	**Q:** What is caused by the gravitational pull of the Moon and Sun on Earth? **Gen.:** Tides are caused by the gravitational pull of the Moon and Sun on Earth. **Ref.:** tides; Tide; The tides	**Q:** According to Greek legend, which of the 9 Muses was the muse of Dance? **Gen.:** According to Greek legend, the Muse of Dance was Calliope. **Ref.:** Terpsichore; Terpsikhore; Terpsichorean

Table 6: Representative KNOW and HALL examples on TriviaQA.

Model	KNOW example	HALL example
Llama-2-7B-Chat	**Q:** a progressive degeneration of the macula lutea of the retina is called **Gen.:** Age-related macular degeneration (AMD) **Ref.:** age related macular degeneration; AMD; Age Related Maculopathy	**Q:** in a solar or lunar eclipse the region of earth's shadow is divided into **Gen.:** Two: **Ref.:** 3.0; 3
Llama-3-8B-Instruct	**Q:** name the economist who related poverty with the concept of entitlement **Gen.:** Amartya Sen. **Ref.:** Amartya Sen; Amartya Kumar Sen	**Q:** former world number one tennis player who coached roger federer in 2014 **Gen.:** Severin Lüthi. **Ref.:** Stefan Edberg
Qwen2-7B-Instruct	**Q:** at the end of the korean war a ceasefire line was established near the **Gen.:** 38th parallel. **Ref.:** 38th parallel north; 38°N	**Q:** who predicted that all matter can behave as waves as well as particles **Gen.:** Albert Einstein **Ref.:** Louis de Broglie; Prince Louis-Victor de Broglie
Qwen2.5-7B-Instruct	**Q:** what type of election allows the party's voters to nominate candidates **Gen.:** Primary election **Ref.:** primary election; primary; political primary	**Q:** when was the enviromental education association of south africa formed **Gen.:** 1985 **Ref.:** 1982-09; September 1982; 1982 September
Gemma-2-9B-it	**Q:** who established the first permanent french settlement in the new world **Gen.:** Samuel de Champlain. **Ref.:** Samuel de Champlain; Samuel Champlain	**Q:** who was the first recipient of bharat ratna to be awarded posthumously **Gen.:** Chittaranjan Das **Ref.:** Lal Bahadur Shastri
DeepSeek-LLM-7B-Chat	**Q:** what is the name of the turkish city formerly known as constantinople **Gen.:** The name of the Turkish city formerly known as Constantinople is Istanbul. **Ref.:** Istanbul; Republic of Turkey; tr	**Q:** name the economist who related poverty with the concept of entitlement **Gen.:** Thomas Picketty **Ref.:** Amartya Sen; Amartya Kumar Sen

Table 7: Representative KNOW and HALL examples on MKQA.

A.2 Experimental Setup

Internal Feature Extraction. For each generated answer, we extract hidden representations and neighborhood-gradient features from the frozen base LLM. Hidden evidence is constructed from the mean hidden representation of answer tokens over selected intermediate-to-late layers. Neighborhood-gradient features characterize local decoding sensitivity around the generated answer and are computed over selected content tokens.

Training Details. For every dataset–model case, the balanced data are split into 70%/15%/15% train, validation, and test sets using stratified sampling. The hidden evidential learner uses a two-layer MLP with hidden dimension 512 and dropout 0.25. The residual evidence branch uses hidden dimension 256, bottleneck dimension 96, and dropout 0.35. Both stages are optimized with AdamW using a learning rate of 10^{-3} and weight decay of 10^{-4}. We train for at most 120 epochs with early stopping, use a KL regularization weight of 0.01 with 20-epoch annealing, and apply gradient clipping at 5.0. All experiments use random seed 42.

A.3 Additional Analysis

Components		Modeling		Llama-3-8B-Instruct				Qwen2-7B-Instruct			
Hidden	Gradient	Direct	EDL	PopQA	WebQ	Trivia	MKQA	PopQA	WebQ	Trivia	MKQA
–	✓	–	✓	71.14	69.31	60.69	62.61	75.95	60.32	58.07	60.33
✓	–	✓	–	87.56	80.16	69.10	77.60	91.77	77.64	72.68	77.63
✓	✓	✓	–	88.52	81.53	72.74	78.12	90.54	77.94	75.02	79.23
✓	–	–	✓	88.74	85.85	78.06	**79.07**	92.02	80.98	81.15	81.35
✓	✓	–	✓	**88.92**	**86.09**	**78.45**	79.01	**92.20**	**81.13**	**81.25**	**81.36**
Components		**Modeling**		**Qwen2.5-7B-Instruct**				**DeepSeek-LLM-7B-Chat**			
Hidden	**Gradient**	**Direct**	**EDL**	**PopQA**	**WebQ**	**Trivia**	**MKQA**	**PopQA**	**WebQ**	**Trivia**	**MKQA**
–	✓	–	✓	82.71	59.35	61.33	66.61	78.67	61.46	58.81	63.68
✓	–	✓	–	93.58	78.30	79.12	82.59	91.12	77.52	72.25	78.95
✓	✓	✓	–	94.46	77.04	81.23	84.97	91.11	76.95	73.41	81.15
✓	–	–	✓	**94.59**	**82.94**	87.41	85.95	92.31	**79.21**	79.64	81.33
✓	✓	–	✓	**94.59**	**82.94**	**87.55**	**85.96**	**92.37**	79.20	**79.86**	**81.35**

Table 8: Component ablation study of HNGRE. We report AUROC results on four models.

Component Ablation. We first analyze the contribution of each component in HNGRE. Table 8 reports the AUROC results on four 7B-scale models across the four factual QA datasets. The ablation shows that hidden representations provide the dominant signal for hallucination detection: using hidden features alone already yields strong discrimination across most model–dataset configurations, whereas gradient features alone are notably weaker. Introducing evidential learning on top of hidden representations leads to a further and more consistent improvement, indicating that explicitly modeling class-wise evidence is beneficial beyond directly learning a discriminative boundary in the hidden space. Finally, incorporating neighborhood-gradient residual evidence produces the full HNGRE model. Although the additional gain over Hidden+EDL is relatively modest in some cases, it remains complementary to the hidden evidence and improves the overall results in most configurations. These observations suggest that HNGRE is primarily driven by the latent information encoded in hidden states, while evidential learning strengthens evidence estimation and neighborhood gradients provide residual information about local decoding behavior.

Additional Ablation Results. Tables 9 and 10 provide the complete ablation results beyond those reported in the main text. On Llama-3-8B-Instruct and Gemma-2-9B-it, hidden representations again provide substantially stronger signals than gradient features alone, while introducing evidential learning consistently improves the hidden-only baseline. The full HNGRE model achieves the best or near-best AUROC in almost all dataset–model cases, although Hidden+EDL remains highly competitive in several settings. Averaged over all 24 cases, HNGRE obtains an AUROC of 0.8403, an AUPR of 0.8393, and an FPR@95 of 0.6129, compared with 0.8393, 0.8374, and 0.6164 for Hidden+EDL, respectively. These results further confirm that hidden evidence constitutes the primary source of discrimination, evidential learning provides the main performance gain, and neighborhood-gradient residuals offer a smaller but complementary refinement.

Relative Factual Support in Hidden Space. The ablation results indicate that hidden representations provide the primary signal for distinguishing supported and hallucinated generations. We therefore examine whether this distinction is reflected in the local organization of hidden space. For a hidden representation h_i, we estimate its local density with respect to KNOW and HALL reference states as

$$\rho_c(h_i) = \frac{1}{d_c(h_i) + \epsilon}, \qquad c \in \{\mathrm{KNOW}, \mathrm{HALL}\}, \tag{11}$$

Variant	Llama-3-8B-Instruct				Gemma-2-9B-it			
	PopQA	WebQ	TriviaQA	MKQA	PopQA	WebQ	TriviaQA	MKQA
Gradient only	0.7114	0.6931	0.6069	0.6261	0.7692	0.6607	0.6543	0.7041
Hidden only	0.8756	0.8016	0.6910	0.7760	**0.9103**	0.7953	0.7106	0.7841
Hidden + EDL	<u>0.8874</u>	<u>0.8585</u>	<u>0.7806</u>	**0.7907**	<u>0.9076</u>	<u>0.8321</u>	<u>0.7783</u>	0.8212
Hidden + Gradient	0.8852	0.8153	0.7274	0.7812	0.9033	0.8030	0.7372	<u>0.8250</u>
HNGRE	**0.8892**	**0.8609**	**0.7845**	<u>0.7901</u>	0.8969	**0.8359**	**0.7866**	**0.8275**

Table 9: Remaining component ablation study on Llama-3-8B-Instruct and Gemma-2-9B-it. We report AUROC for each dataset–model case. The best result in each column is bolded and the second-best result is underlined.

Variant	AUROC	AUPR	FPR@95
Gradient only	0.6603	0.6422	0.8238
Hidden only	0.8017	0.7981	0.6844
Hidden + EDL	<u>0.8393</u>	<u>0.8374</u>	<u>0.6164</u>
Hidden + Gradient	0.8121	0.8152	0.6644
HNGRE	**0.8403**	**0.8393**	**0.6129**

Table 10: Overall ablation results averaged across 24 dataset–model cases. AUROC and AUPR are better when higher, while FPR@95 is better when lower. The best result in each metric is bolded and the second-best result is underlined.

where $d_c(h_i)$ denotes the average distance to the k nearest reference states from class c. We then define the relative factual-support score as

$$s_{\text{support}}(h_i) = \log \rho_{\text{KNOW}}(h_i) - \log \rho_{\text{HALL}}(h_i). \tag{12}$$

A positive score indicates that the representation is locally better supported by KNOW states, whereas a negative score indicates relatively weaker factual support and stronger proximity to hallucination-associated states.

As shown in Figure 5, a consistent shift emerges across all six models and four datasets: KNOW generations generally receive positive relative-support scores, while HALL generations shift toward negative values. The separation is particularly pronounced on PopQA and remains evident on WebQuestions, TriviaQA, and MKQA despite differences in model architecture and dataset distribution. Importantly, this pattern is obtained directly from the geometry of hidden representations without training an additional parametric classifier. These results indicate that False Recall Hallucination is associated with a systematic reduction in relative factual support in hidden space, providing an observable latent signature consistent with the Latent Void Hypothesis.

Robustness to Neighborhood Size. We further examine whether the relative factual-support pattern depends on the choice of neighborhood size k. Figure 6 reports the gap between the mean relative factual-support scores of KNOW and HALL samples for $k \in \{5, 10, 20, 50\}$. Across all six models and four datasets, the gap remains positive for every tested k, indicating that KNOW generations consistently receive stronger relative factual support than hallucinated generations. Although the magnitude gradually decreases as the neighborhood expands, the ordering is preserved throughout, showing that the observed hidden-space support difference is not tied to a particular local neighborhood size. The effect is especially pronounced on PopQA, while TriviaQA exhibits a smaller but still stable separation.

Effect of Evidential Learning. Figure 7 further isolates the contribution of evidential learning by comparing Hidden+EDL with the Hidden Only variant across all 24 dataset–model cases. EDL improves AUROC in 22 of the 24 cases, with only two small negative changes on PopQA. The gain varies considerably across datasets: improvements are relatively modest on PopQA, but become much larger on WebQuestions and TriviaQA, reaching approximately 0.09 AUROC in the strongest cases. This pattern suggests that evidential modeling is particularly useful when the hidden representations alone provide less directly separable class structure. Overall, the results confirm that the

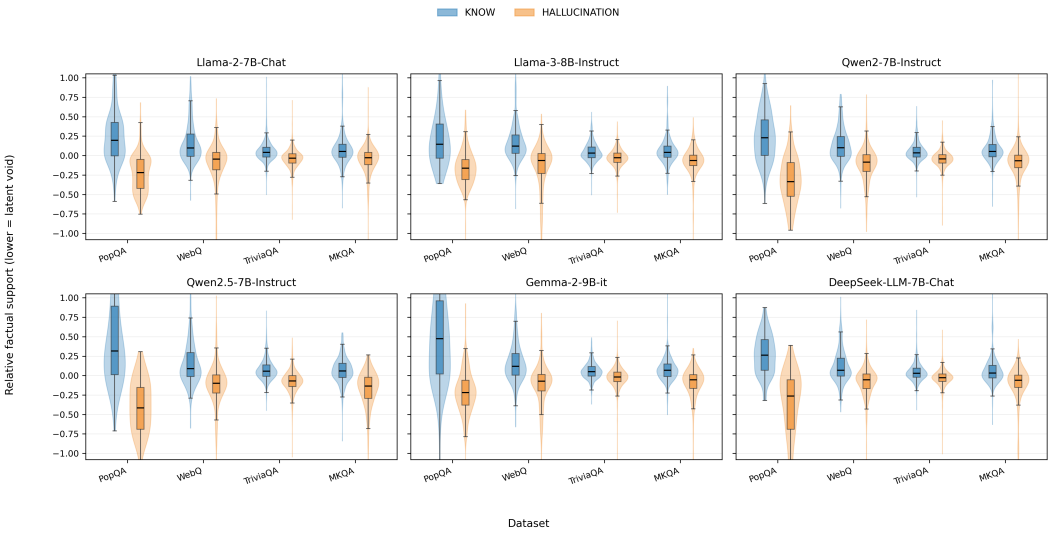

Figure 5: Relative factual support across six LLMs and four datasets. Positive values indicate stronger support from KNOW reference states, while negative values indicate stronger HALL-relative support.

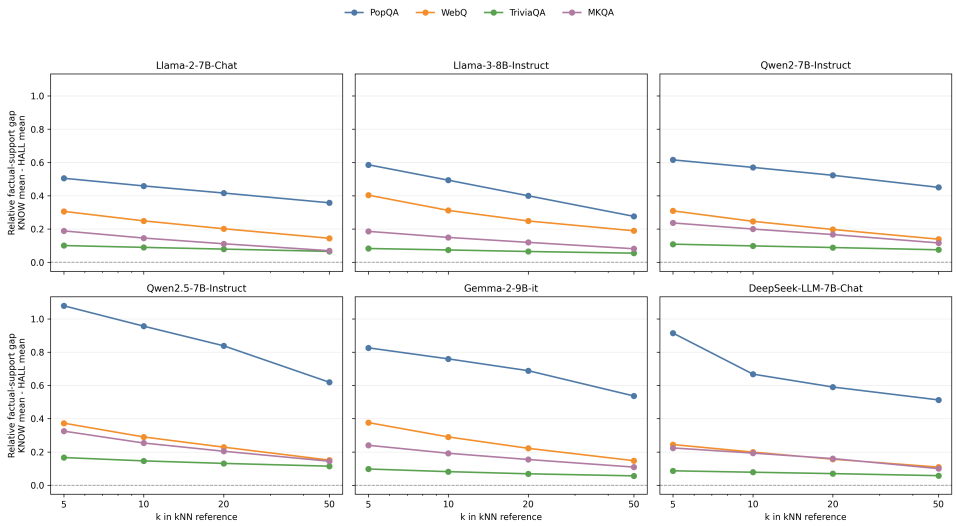

Figure 6: Robustness of the relative factual-support gap to different k values in the kNN reference set across six LLMs and four datasets.

improvement from EDL is not driven by a small number of model–dataset configurations, but is broadly consistent across the evaluation suite.

Neighborhood-Gradient Sensitivity. We further examine whether the neighborhood-gradient features capture information beyond the hidden-state representation. Figure 8 compares the standardized gradient sensitivity of KNOW and HALL samples across all six models and four datasets. Although the separation is substantially weaker than that observed in hidden space, hallucinated generations tend to exhibit slightly higher local sensitivity in a number of model–dataset settings, with more visible shifts on PopQA and for models such as Qwen2.5 and Gemma-2.

The considerable overlap between the two distributions is also consistent with the ablation results: neighborhood gradients are not sufficiently discriminative to serve as the primary hallucination signal on their own. Instead, they provide complementary information about local decoding fragility

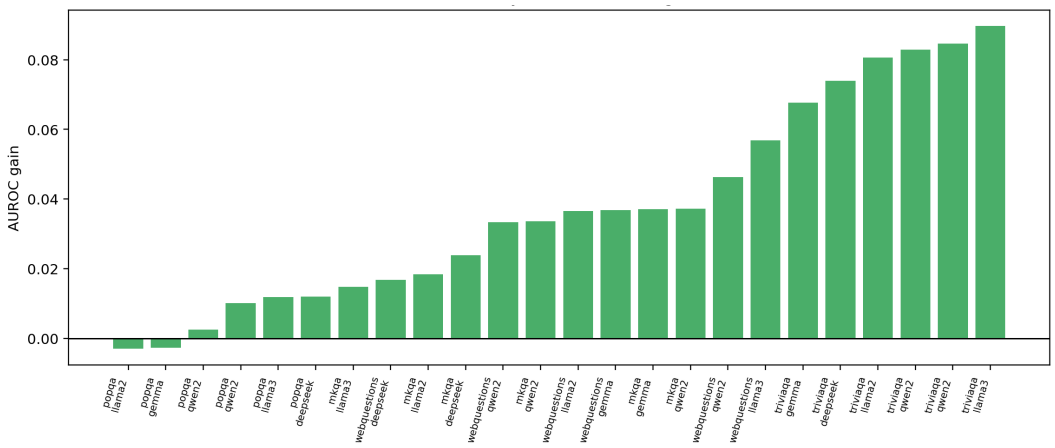

Figure 7: AUROC gain from replacing the hidden-only classifier with evidential learning across 24 dataset–model cases. Positive values indicate improvements from Hidden+EDL over Hidden Only.

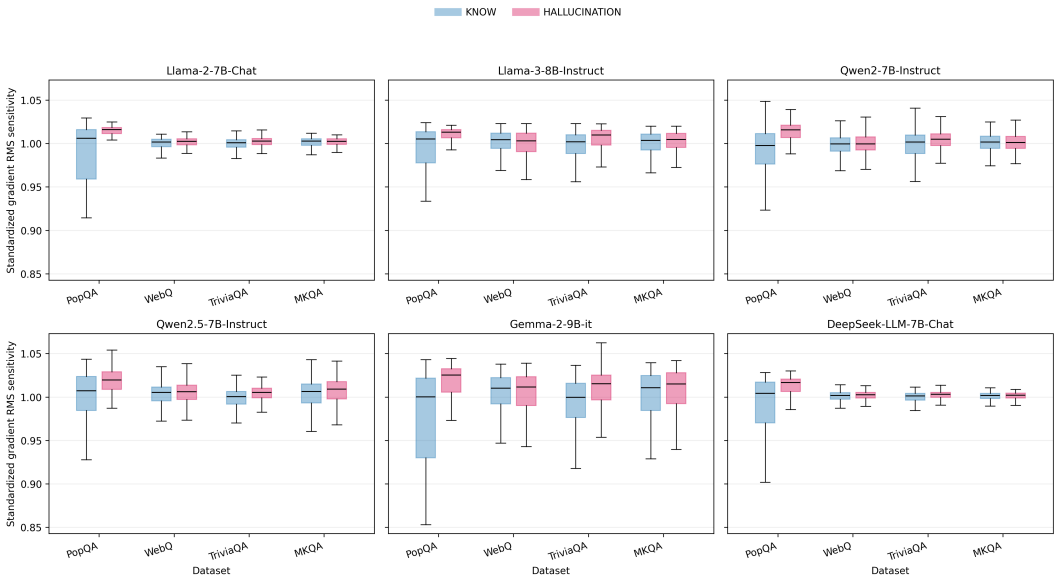

Figure 8: Distribution of neighborhood-gradient sensitivity for KNOW and HALL samples across six LLMs and four datasets.

that is not fully captured by the hidden representation. This supports the design of HNGRE, where gradient features are incorporated as residual evidence rather than as an independent detector.

Evidence-Aware Uncertainty. Beyond improving discrimination, evidential learning provides an explicit measure of how much evidence HNGRE accumulates for each prediction. We use

$$u = K/\sum_k \alpha_k \tag{13}$$

as the evidential uncertainty, where a larger value corresponds to lower total accumulated evidence.

Relative Factual-Support Landscape. Figure 9 visualizes the relative factual-support landscape of hidden representations across six LLMs. The projected geometry varies considerably between model families, but several common patterns can be observed. Regions with stronger KNOW support form structured local clusters, while areas with stronger HALL-relative support often emerge along different directions of the projected space. The transition between these regions is generally

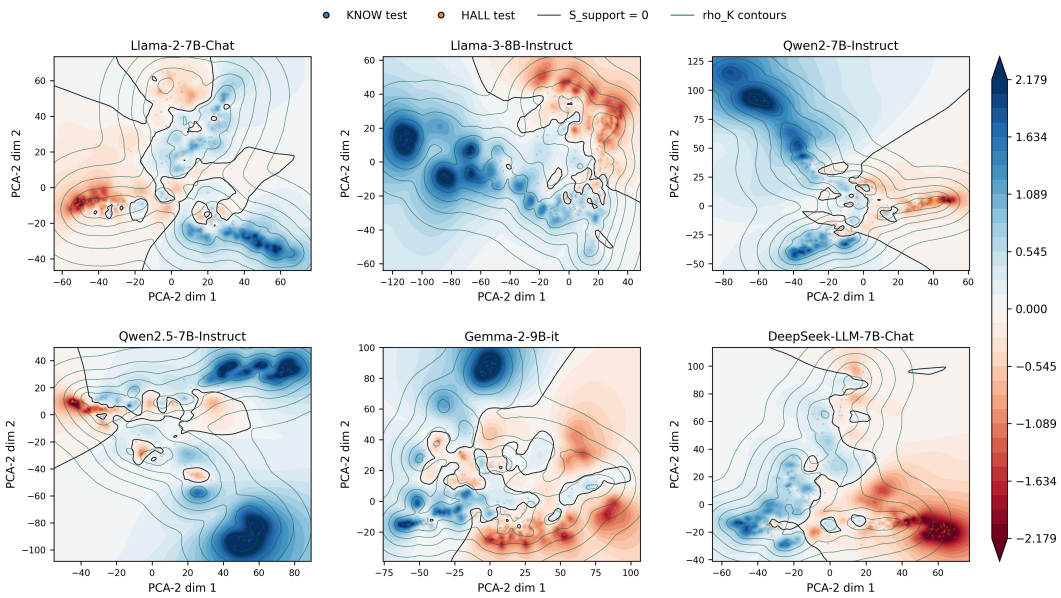

Figure 9: Relative factual-support landscapes in PCA-projected hidden space across six LLMs. The background visualizes the balance between KNOW and HALL support, while the contours show local KNOW density structure.

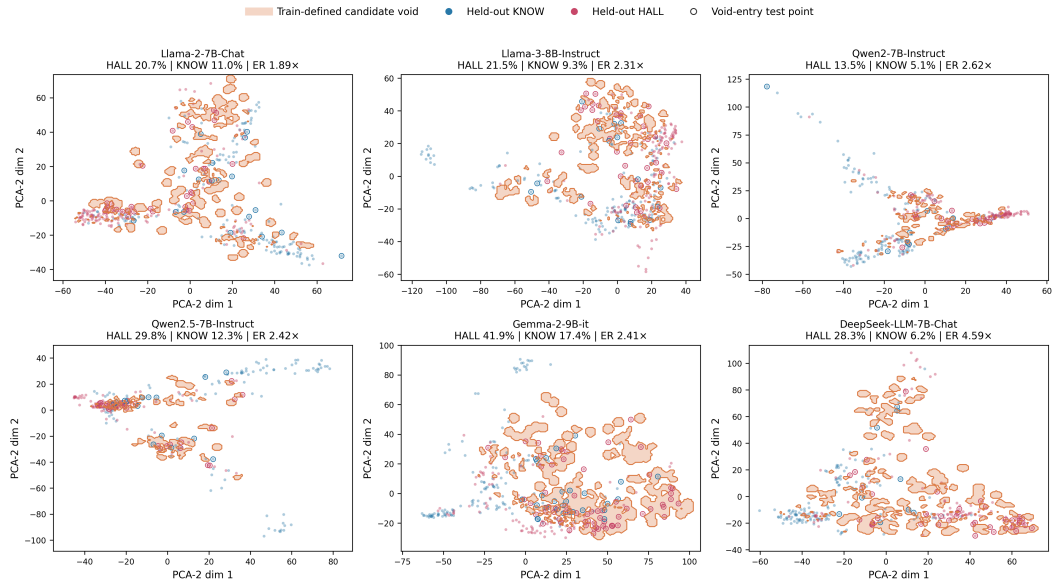

Figure 10: Train-defined latent-void candidates across six LLMs. Candidate voids are identified in the original high-dimensional hidden space as regions with weak KNOW support but non-negligible overall latent occupancy, without using HALL labels. Orange contours show the 2D PCA projections of the train-defined candidate regions, while circles mark held-out test samples identified as void entries in the original hidden space. PCA is used only for visualization.

non-linear and model-dependent, with some models exhibiting relatively compact support regions and others showing more fragmented or elongated structures. These results indicate that the organization of factual support in hidden space is heterogeneous across models, while still exhibiting systematic spatial variation between supported and hallucination- associated states.

Construction of Label-Free Latent-Void Candidates. To avoid defining latent voids using hallucination labels, we construct candidate void regions directly from the geometry of the training hidden representations. Let $s_K(z)$ denote the factual-support score of a hidden state z, estimated using only train KNOW representations, and let $p_{\text{all}}(z)$ denote the overall latent density estimated from all training hidden states after discarding their KNOW/HALL labels. We define a candidate latent void as a region that is weakly supported by known factual states but still has non-negligible latent occupancy:

$$\mathcal{V} = \left\{ z : s_K(z) < Q_{0.15}^{\text{KNOW}} \right\} \cap \left\{ z : p_{\text{all}}(z) > Q_{0.05}^{\text{all}} \right\}, \tag{14}$$

where $Q_{0.15}^{\text{KNOW}}$ is the 15th percentile of factual-support scores over train KNOW states, and $Q_{0.05}^{\text{all}}$ is the 5th percentile of overall latent density over all training representations.

The first condition identifies regions with deficient factual support relative to known factual states, while the second excludes remote low-density regions that are rarely occupied by the model. Importantly, HALL labels are not used to define $\mathcal{V}$. The same quantile rule is applied to all six models without model-specific threshold tuning.

All support, density, and void-membership computations are performed in the original high-dimensional hidden space. PCA is used only to visualize the resulting candidate regions in two dimensions. After $\mathcal{V}$ is fixed from the training split, we evaluate held-out KNOW and HALL samples using their respective void-entry rates:

$$r_{\text{HALL}} = P(z \in \mathcal{V} \mid H = \text{HALL}), \qquad r_{\text{KNOW}} = P(z \in \mathcal{V} \mid H = \text{KNOW}). \tag{15}$$

We further define the enrichment ratio

$$\text{ER} = \frac{P(z \in \mathcal{V} \mid H = \text{HALL})}{P(z \in \mathcal{V} \mid H = \text{KNOW})}, \tag{16}$$

where $\text{ER} > 1$ indicates that hallucinated generations enter the train-defined factual-support-deficient regions more frequently than supported generations.

As shown in Figure 10, this pattern is consistent across all six evaluated LLMs. The HALL entry rate ranges from 13.5% to 41.9%, whereas the KNOW entry rate ranges from 5.1% to 17.4%. Accordingly, the enrichment ratio is greater than one for every model, ranging from $1.89\times$ to $4.59\times$. Averaged across the six models, the HALL and KNOW entry rates are 25.9% and 10.2%, respectively, with a mean enrichment ratio of $2.71\times$.

These results show that latent regions defined without hallucination labels nevertheless exhibit systematically elevated hallucination risk on held-out samples. This supports the interpretation of latent voids as model-accessible regions with insufficient factual support, rather than arbitrary low-density areas or regions explicitly fitted to separate KNOW from HALL.

As shown in Figure 11, the two classes exhibit markedly different uncertainty distributions. Supported generations are concentrated at low uncertainty, indicating that HNGRE accumulates substantial internal evidence for these answers. In contrast, hallucinated generations are predominantly concentrated at high uncertainty, reflecting a pronounced deficiency of supporting evidence. The limited overlap between the two distributions shows that the evidential output is not merely a confidence score attached to the final prediction, but captures a systematic difference in the amount of internal evidence associated with supported and hallucinated generations. This result further supports our interpretation that False Recall Hallucination is associated with insufficient internal support.

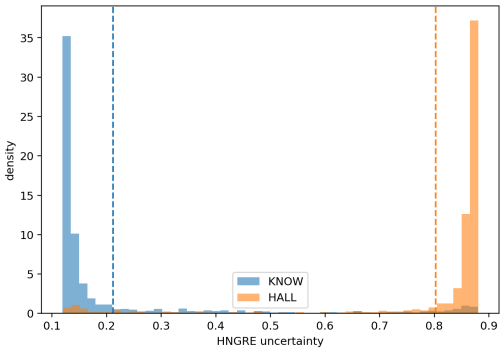

Figure 11: Distribution of HNGRE evidential uncertainty for KNOW and HALL samples.

Cohen's d for Factual-Support Projection. We quantify the held-out separation between KNOW and HALL projections along the learned factual-support direction using Cohen's d:

$$d = \frac{\mu_{\mathrm{KNOW}} - \mu_{\mathrm{HALL}}}{s_{\mathrm{pooled}}}, \tag{17}$$

where

$$s_{\mathrm{pooled}} = \sqrt{\frac{(n_K - 1)s_K^2 + (n_H - 1)s_H^2}{n_K + n_H - 2}}. \tag{18}$$

Here, μ_K, μ_H and s_K, s_H denote the means and standard deviations of the held-out projection scores for the two classes.