

Training a Generalist Hallucination Detector across Multiple Domains via Adaptive Layer Aggregation

Xinyi Li*

xinyi.li-1@student.uts.edu.au
University of Technology Sydney
Sydney, NSW, Australia

Zhen Fang†

zhen.fang@uts.edu.au
University of Technology Sydney
Sydney, NSW, Australia

Yadan Luo

lyadanluo@gmail.com
The University of Queensland
Brisbane, Queensland, Australia

Yongxin Deng*

dengyongxin1998@gmail.com
University of Technology Sydney
Sydney, NSW, Australia

Sean Du

xuefeng.du@ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Shanshan Ye

shanshan.ye@student.uts.edu.au
University of Technology Sydney
Sydney, NSW, Australia

Ling Chen†

ling.chen@uts.edu.au
University of Technology Sydney
Sydney, NSW, Australia

Abstract

Hallucination detection is essential for the reliable deployment of large language models (LLMs) in safety-critical applications, yet existing detectors often degrade under real-world domain shifts. To formalize this challenge, we introduce Multi-Domain Generalizable Hallucination Detection (MGHD), where a detector is trained on multiple labeled domains and evaluated on unseen domains without any test-domain supervision. The central challenge of MGHD stems from a previously underexplored phenomenon, termed *layer-index shift*, in which the most informative internal layer varies substantially across domains, causing detectors that rely on a fixed layer to generalize poorly. To mitigate this phenomenon, we propose a novel method **Stability-Aware Adaptive Layer Aggregation (SALA)**. SALA defines a layer-wise instability index computed from the training data, which measures how separability varies across domains. For an unseen test domain, it provides an upper bound on the possible separability drop from training to test. Guided by this bound, SALA designs a per-layer projector that balances discriminability and stability. Then using these projectors, SALA learns invariant subspaces across layers and aggregates them into a robust detector. Experiments across four widely used QA benchmarks and four LLM backbones from the LLaMA and Qwen families (7B-14B) show that SALA achieves the best average performance compared with existing hallucination detection baselines.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Artificial intelligence**; **Machine learning**.

*Both authors contributed equally to this research.

†Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
<https://doi.org/10.1145/3770855.3817792>

Keywords

LLM Hallucination Detection; Domain Generalization; Reliable Machine Learning

ACM Reference Format:

Xinyi Li, Zhen Fang, Yadan Luo, Yongxin Deng, Sean Du, Shanshan Ye, and Ling Chen. 2026. Training a Generalist Hallucination Detector across Multiple Domains via Adaptive Layer Aggregation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26)*, August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770855.3817792>

Resource Availability:

The source code of this paper is publicly available at <https://github.com/Nellie179/SALA>, with an archival snapshot available at <https://doi.org/10.5281/zenodo.20467844>.

1 Introduction

Hallucination detection [10, 13, 23, 39] is crucial for the reliable deployment of large language models (LLMs) in real-world applications, particularly in safety-critical domains such as education [16], healthcare [49], and finance [26]. This is because LLMs may produce fluent but factually incorrect or logically inconsistent outputs, commonly referred to as hallucinations [25], which can undermine user trust and cause harm in high-stakes settings. Substantial progress has been made, and many detectors achieve strong in-domain performance [3, 6, 56]. However, real deployments inevitably encounter distribution shifts, and hallucination detectors that perform well in-domain can degrade sharply when transferred across domains [27]. This persistent fragility exposes a critical generalization gap and motivates *cross-domain* hallucination detection.

To bridge this generalization gap in realistic deployment scenarios, we propose a novel, challenging, and practical problem setting, termed Multi-Domain Generalizable Hallucination Detection (MGHD). In the MGHD setting, a detector is trained on labeled data drawn from multiple training domains and then evaluated on unseen test domains, *without* access to any test-domain supervision or validation. This setup reflects practical deployments in

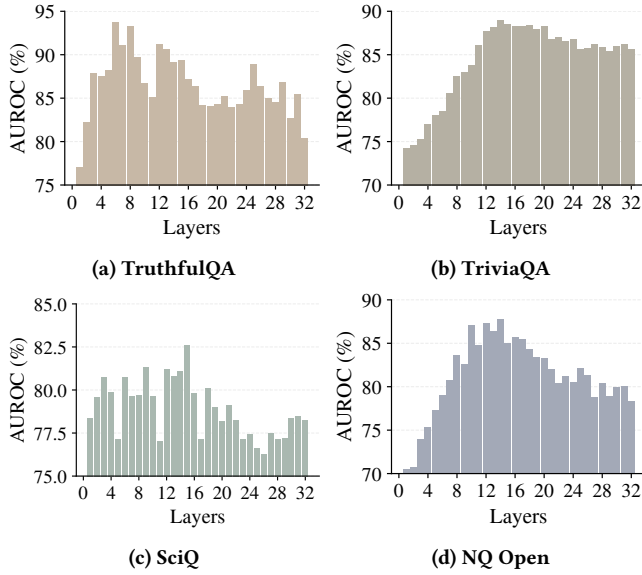


Figure 1: Layer-wise probing performance varies substantially across domains. For each dataset, we train a linear probe [2] for hallucination detection at every Transformer layer and report the resulting in-domain performance. The results reveal a clear *layer-index shift*: the layer that yields the strongest in-domain separability differs markedly across datasets, and the ranking of layers can change sharply from one domain to another. This pattern indicates that the optimal probing layer is highly domain-dependent, and there is no single layer that remains consistently optimal.

which the training and test domains are not necessarily aligned, and re-annotating data or re-tuning the detection model for each new domain is often impractical in real-world scenarios.

Mainstream hallucination detection methods rely on internal representations from a single layer, typically selected based on validation performance [3, 6, 8, 18, 57]. However, their performance often degrades in MGHD. One key reason is that the most informative layer for distinguishing truthful from hallucinated outputs varies significantly across domains (Figures 1 and 2). We refer to this phenomenon as *layer-index shift*, characterized by strong domain dependence and high variability. As a result, for MGHD, the layer chosen during training is frequently suboptimal in unseen test domains, leading to brittle generalization.

To effectively address MGHD, we begin by analyzing the underlying factors of the layer-index shift phenomenon. We provide an intuitive explanation, drawing on prior perspectives from domain generalization and invariance [1, 40, 46]. The insight is that different domains exhibit different nuisance patterns that should be ignored, while detectors trained only on training domains may fail to suppress nuisance variations that appear in unseen domains. As a result, the layer-wise separation between truthful and hallucinated samples becomes unstable. This instability implies that no single layer remains consistently informative across domains: different test domains favor different layers depending on which are least affected by domain-specific variations. This domain-dependent variability in layer informativeness gives rise to the observed layer-index shift.

Guided by this view, we propose a novel method **Stability-Aware Adaptive Layer Aggregation (SALA)**, which explicitly prioritizes components that remain stable across domains while reducing layer-level noise caused by domain-specific variations. SALA begins by defining a layer-wise instability index that measures how much a layer’s ability to distinguish between truthful and hallucinated outputs varies across training domains. Derived from training domains alone, this index proxies a layer’s expected shift-robustness at test time. Under the stated assumptions, our analysis shows that for related unseen domains, the separability drop relative to the training domains can be upper bounded in terms of this instability index, offering a principled way to allocate representation dimensions across layers. Motivated by this, SALA assigns a layer-wise dimension to each layer by balancing (i) strong separability on the training domains and (ii) low instability across domains. This limits the influence of layers whose behavior is volatile. With the selected dimensions, SALA then learns a lightweight projection for each layer to extract a compact, class-conditional subspace that emphasizes domain-stable discriminative signals. Finally, these projected features are concatenated and fed into a probe, reducing the influence of unstable layers on the final decision.

Extensive experiments are conducted on four LLM backbones (i.e., LLaMA-3.1-8B, LLaMA-3-8B-Instruct, Qwen-2.5-7B, and Qwen-2.5-14B) and four open-domain QA benchmarks treated as domains (i.e., TriviaQA [24], TruthfulQA [33], NQ Open [30], and SciQ [54]) under a leave-one-domain-out protocol. Results show that SALA achieves strong performance under domain shifts and often outperforms competitive baselines, including generalization-oriented detectors such as PRISM [56] and ICR Probe [58]. Our main contributions are summarized as follows:

- We propose Multi-Domain Generalizable Hallucination Detection (MGHD), a practically motivated setting for hallucination detection, and systematically re-assess representative baselines under a unified multi-domain protocol, revealing pronounced robustness gaps under domain shifts.
- We identify an empirical phenomenon in cross-domain hallucination detection, termed *layer-index shift*: the layer that most reliably separates truthful from hallucinated responses can differ markedly across different domains and backbones.
- We introduce a layer-wise *instability index* and establish a theoretical bound that links it to cross-domain performance degradation, showing that, under the stated assumptions, the train–test separability discrepancy for unseen domains can be explicitly upper-bounded in terms of this index in a clear, principled, and interpretable manner.
- We develop SALA, which designs per-layer projectors with adaptive projection dimensions to learn stable, transferable features and integrates complementary information across layers, yielding competitive cross-domain performance in the studied MGHD setting.

2 Related Work

Hallucination Detection. A growing literature studies how to detect hallucinations in LLM outputs using signals spanning output-space uncertainty [35], multi-sample disagreement, internal activations, and evidence-grounded verification [21, 41]. Output-based

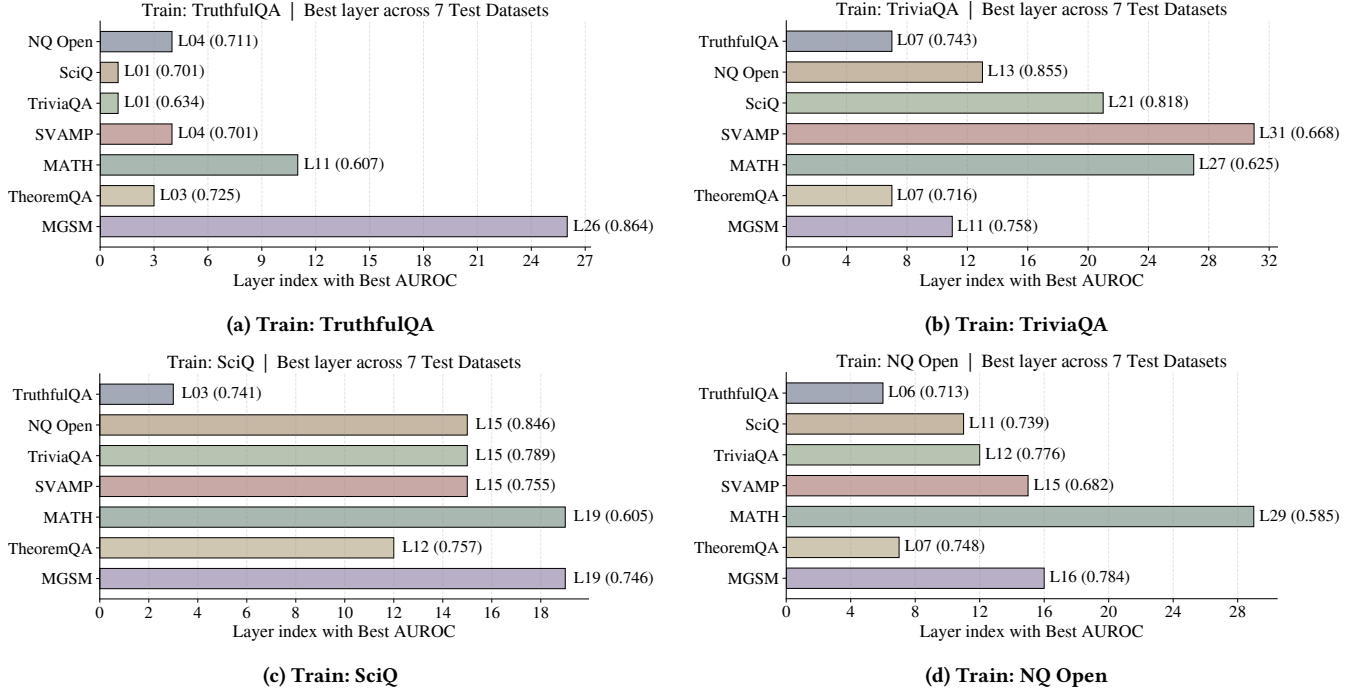


Figure 2: Cross-domain probing exhibits “layer-index shift” (no universal best layer). Each subplot trains a linear probe for hallucination detection on one training domain (title) and evaluates it on the remaining test domains (y -axis). For each test domain, we report the layer that maximizes test AUROC (bar length; value shown in parentheses). The optimal probing layer changes substantially across test domains, indicating no consensus on a single probing layer and strong domain dependence.

methods view hallucination risk as correlated with elevated uncertainty and exploit token- or sequence-level statistics such as log probabilities, perplexity, and entropy-like measures [13, 29, 38, 48]. To address the fact that the same meaning can be expressed in many surface forms, semantic-level estimators group generations by semantic equivalence and quantify uncertainty in meaning space, improving failure prediction for open-ended generation [13, 29]. Another widely used black-box family estimates reliability from sample-to-sample disagreement by generating multiple stochastic outputs for the same query and using their consistency as a proxy for factuality. SelfCheckGPT is a representative baseline in this line [39]. Complementary work elicits uncertainty in natural language by prompting models to verbalize confidence, which can be calibrated but remains sensitive to prompting and evaluation conditions [32, 55]. Beyond output signals, internal-based approaches operate in representation space, training lightweight probes or learning directions/subspaces from internal activations to predict truthfulness or semantic uncertainty in a single forward pass [3, 4, 12, 28, 51]. Recent studies further investigate robustness under domain shifts by guiding internal states with prompts or by explicitly modeling hidden-state dynamics across layers [19, 20, 56, 58]. Finally, evidence-grounded pipelines verify generations against external corpora via retrieve-and-check or post-hoc revision, improving interpretability at the cost of dependence on retrieval quality and domain coverage [5, 14, 21].

Layer Sensitivity in Internal-Space Hallucination Detection.

While internal-space detectors can be lightweight and effective, their performance is often highly sensitive to *which layer* is probed. Several works therefore conduct explicit layer-wise analyses and treat layer selection as an important design choice. For example, LLM-Check reports systematic cross-layer behaviors and includes a dedicated study on layerwise performance and efficient layer selection, highlighting that some internal signals can fluctuate substantially across layers and that mid-level layers may be close to optimal in practice [52]. Related studies observe pronounced layer-dependent separability: ICR Probe analyzes layer-wise AUROC of its internal score, identifies layers where hallucination cues peak, and motivates incorporating broader cross-layer patterns rather than relying on a single-layer statistic [58]. In long-form generation, recent token-level probing systems explicitly read hidden states from an intermediate layer to detect hallucinated entities in real time, again underscoring the practical importance of probing layer [43]. Beyond choosing a single layer, other probe-based approaches aggregate information across layers, e.g., by concatenating grouped layer representations, aiming to reduce sensitivity to any particular layer and improve robustness [15]. Overall, these findings suggest that fixed-layer probing can be brittle, especially under distribution shift, motivating adaptive mechanisms that combine layer-wise signals in a data-driven manner.

3 Preliminaries and Problem Formulation

3.1 Preliminary Analysis

We first examine whether hallucination-relevant information is consistently encoded at the same internal layer across domains. Following SAPLMA [6], we use LLaMA-3-8B-Instruct, concatenate the question-answer sequence, extract the last-token hidden state at each Transformer layer, and train one linear probe per layer.

Figure 1 reports in-domain AUROC on TruthfulQA, TriviaQA, SciQ, and NQ Open. For each dataset, training and evaluation are performed independently at each layer within the same dataset. The results show substantial layer-wise variation, indicating that the most informative layer for hallucination detection is domain-dependent. Figure 2 further reports cross-domain probing results across eight datasets. For each QA training domain, we train one probe per layer and evaluate it on the same layer of each test domain, reporting the layer that achieves the best AUROC for each test domain. The best transfer layer shifts substantially across test domains even when the training domain is fixed. This observation reveals *layer-index shift*: a layer selected from training-domain evidence may become suboptimal on unseen domains. This motivates SALA, which avoids relying on a single fixed layer and instead aggregates stable layer representations across layers.

3.2 Problem Formulation

LLMs and Internal Representations. Following Du et al. [12], Park et al. [44], we consider a base LLM as a probability distribution $\mathbb{P}_\theta(\cdot)$ over token sequences, where θ denotes the model parameters. Given an input (question) token sequence $\mathbf{Q} = [q_1, \dots, q_k]$, where q_j is the j -th token. The LLM generates an output (answer) sequence $\mathbf{A} = [q_{k+1}, \dots, q_{k+a}]$ of length a by sampling from

$$\mathbb{P}_\theta(\mathbf{A} | \mathbf{Q}) = \prod_{j=k+1}^{k+a} \mathbb{P}_\theta(q_j | q_{<j}), \quad (1)$$

where $q_{<j} = (q_1, \dots, q_{j-1})$ denotes the preceding context. A transformer based LLM with L layers produces a sequence of hidden embeddings for each input pair $(\mathbf{Q}, \mathbf{A})$. Let $\phi_{\theta,\ell}$ denote the non-linear transformation in layer ℓ . With recursive composition, the hidden state $H_{\theta,\ell}(\mathbf{Q}, \mathbf{A}) \in \mathbb{R}^{T \times d_h}$ at the ℓ -th layer is given by

$$\phi_{\theta,\ell} \circ \phi_{\theta,\ell-1} \circ \dots \circ \phi_{\theta,1}(H_{\theta,0}(\mathbf{Q}, \mathbf{A})), \quad (2)$$

where $H_{\theta,0}(\mathbf{Q}, \mathbf{A})$ is the initial input embedding, T is the length of tokens, and d_h is the hidden size. Note that most previous internal-space-based hallucination detection methods [6, 8, 18, 57] use the last-token embedding $Z_\ell(\mathbf{Q}, \mathbf{A})$ to test the factual validity of the generated sequence:

$$Z_\ell(\mathbf{Q}, \mathbf{A}) = H_{\theta,\ell}(\mathbf{Q}, \mathbf{A})[-1] \in \mathbb{R}^{d_h}. \quad (3)$$

Domains. Let $\mathcal{Q}$ and $\mathcal{A}$ be the spaces of questions and answers. We introduce two domains over $\mathcal{Q} \times \mathcal{A}$:

- The *truthful-answer domain* is a joint distribution $\mathbb{P}_{Q,T}$, where $Q \in \mathcal{Q}$ is a random variable for questions and $T \in \mathcal{A}$ is a random variable for the truthful answers.
- The *hallucinatory-answer domain* is a joint distribution $\mathbb{P}_{Q,H}$, where Q is defined as above and $H \in \mathcal{A}$ is a random variable representing the hallucinated answers.

In general hallucination detection settings, based on the dataset construction procedure [12, 44], the distribution $\mathbb{P}_{Q,H}$ is induced by the LLM $\mathbb{P}_\theta(\cdot)$ and $\mathbb{P}_Q$: for any input pair $(\mathbf{Q}, \mathbf{A}) \in \mathcal{Q} \times \mathcal{A}$,

$$\mathbb{P}_{Q,H}(\mathbf{Q}, \mathbf{A}) = \mathbb{P}_Q(\mathbf{Q}) \cdot \mathbb{P}_\theta^-(\mathbf{A} | \mathbf{Q}), \quad (4)$$

where $\mathbb{P}_\theta^-(\cdot | \mathbf{Q})$ denotes the conditional distribution induced by the LLM $\mathbb{P}_\theta(\cdot | \mathbf{Q})$ under the constraint that only hallucinated responses are retained. Since MGHD involves multiple domains, we use the index e to distinguish among them, i.e., $\mathbb{P}_{Q,T}^e$ represents the truthful-answer domain e and $\mathbb{P}_{Q,H}^e$ is the hallucinatory-answer domain e . Therefore, $(\mathbf{Q}^e, \mathbf{T}^e) \sim \mathbb{P}_{Q,T}^e$ implies that $\mathbf{T}^e$ is a correct response to the question $\mathbf{Q}^e$, and $(\mathbf{Q}^e, \mathbf{H}^e) \sim \mathbb{P}_{Q,H}^e$ implies that $\mathbf{H}^e$ is a hallucinated response to $\mathbf{Q}^e$.

Datasets. Given a question $\mathbf{Q}^e \sim \mathbb{P}_Q^e$, the LLM generates an answer $\mathbf{A}^e$ according to the conditional distribution $\mathbf{A}^e \sim \mathbb{P}_\theta(\cdot | \mathbf{Q}^e)$. Each generated answer $\mathbf{A}^e$ is then assigned a binary label $y^e \in \{0, 1\}$ based on its semantic alignment with a reference answer $\mathbf{A}_{\text{ref}}^e \sim \mathbb{P}_{T|Q}^e(\cdot | \mathbf{Q}^e)$. Specifically, if $\mathbf{A}^e$ aligns with the reference answer $\mathbf{A}_{\text{ref}}^e$, it is labeled as truthful ($y^e = 1$); otherwise, it is labeled as hallucinated ($y^e = 0$). The resulting dataset is defined as

$$\mathcal{D}_l^e = \{(\mathbf{Q}_1^e, \mathbf{A}_1^e, y_1^e), \dots, (\mathbf{Q}_{n^e}^e, \mathbf{A}_{n^e}^e, y_{n^e}^e)\}. \quad (5)$$

Note that the hallucinated samples in the dataset $\mathcal{D}_l^e$ can be regarded as being drawn from $\mathbb{P}_{Q,H}^e$, which is induced by the LLM as defined in Eq. (4).

Multi-Domain Generalizable Hallucination Detection. Here we give the formal definition of MGHD.

Problem 1 (Multi-Domain Generalizable Hallucination Detection). Given the training datasets constructed from M training (source) domains $\mathbb{P}_{Q,T}^1, \dots, \mathbb{P}_{Q,T}^M$ together with an LLM $\mathbb{P}_\theta$, i.e., for each index $e \in \{1, \dots, M\}$,

$$\mathcal{D}_l^e = \{(\mathbf{Q}_1^e, \mathbf{A}_1^e, y_1^e), \dots, (\mathbf{Q}_{n^e}^e, \mathbf{A}_{n^e}^e, y_{n^e}^e)\} \quad (6)$$

as given in Eq. (5), the aim of Multi-Domain Generalizable Hallucination Detection (MGHD) is to learn a detector G based on $\mathbb{P}_\theta(\cdot)$ and the given training datasets $\mathcal{D}_l^1, \dots, \mathcal{D}_l^M$, such that G can generalize to N unseen test domains $\mathbb{P}_{Q,T}^{M+1}, \dots, \mathbb{P}_{Q,T}^{M+N}$ well, i.e., for any $(\mathbf{Q}, \mathbf{A})$,

$$G(\mathbf{Q}, \mathbf{A}) = \begin{cases} 1, & \text{if } (\mathbf{Q}, \mathbf{A}) \sim \mathbb{P}_{Q,T}^e, \\ \forall e \in \{M+1, \dots, M+N\}, \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where 1 indicates that $\mathbf{A}$ is the truthful response and 0 indicates that $\mathbf{A}$ is the hallucinated response.

4 Methodology

Previous hallucination detection methods based on internal representations [3, 6, 8, 18, 57] typically extract features $Z_\ell(\mathbf{Q}, \mathbf{A})$ from the embeddings $H_{\theta,\ell}(\mathbf{Q}, \mathbf{A})$ at a selected layer ℓ (see Eq. (3)) and use a scoring function S for detection, i.e., given a threshold $\lambda \in \mathbb{R}$,

$$G(\mathbf{Q}, \mathbf{A}) = \begin{cases} 1, & \text{if } S(Z_\ell(\mathbf{Q}, \mathbf{A})) \geq \lambda, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where 1 indicates that $\mathbf{A}$ is the truthful response to $\mathbf{Q}$ and 0 indicates that $\mathbf{A}$ is the hallucinated response to $\mathbf{Q}$.

However, a critical yet underexplored issue lies in the choice of the layer ℓ from which $\mathbf{Z}_\ell(\mathbf{Q}, \mathbf{A})$ is extracted. As shown in Figure 2, when a detector is trained on one domain and evaluated on other unseen domains, the oracle layer that yields the best detection performance varies across domains. This observation suggests that the optimal layer for hallucination detection is highly *domain-dependent*, and that selecting a fixed layer may severely limit cross-domain generalization. In MGHD, selecting layers based on test-domain performance is infeasible due to the absence of test-domain data during training. This naturally raises the question of

how to effectively integrate embeddings across layers for MGHD without relying on explicit layer selection.

4.1 Effective Layer-wise Embedding Integration

Under domain shift, different layers encode unequal levels of generalizable information for hallucination detection, making a uniform allocation across layers suboptimal. Therefore, we suggest that effective integration of multi-layer embeddings should go beyond simple concatenation and instead be guided by a principled allocation of representational dimension, which regulates the effective information contributed by each layer during aggregation.

Given an input $(\mathbf{Q}, \mathbf{A})$, we extract layer-wise hidden representations $\{\mathbf{Z}_\ell(\mathbf{Q}, \mathbf{A})\}_{\ell=1}^L$ (see Eq. (3)), and construct domain-stable features for each layer via a corresponding projector $\mathbf{P}_{d_\ell} \in \mathbb{R}^{d_\ell \times d_h}$, i.e., for each $\ell \in \{1, \dots, L\}$,

$$\mathbf{Z}_\ell(\mathbf{Q}, \mathbf{A}; d_\ell) = \mathbf{P}_{d_\ell} \cdot \mathbf{Z}_\ell(\mathbf{Q}, \mathbf{A}), \quad (9)$$

where d_ℓ denotes the allocated layer dimension of the ℓ -th layer, corresponding to the dimension of its domain-stable projected representation. Then, we form a single fused representation by concatenating the domain-stable features from all layers, i.e.,

$$\mathbf{Z}(\mathbf{Q}, \mathbf{A}) = [\mathbf{Z}_1(\mathbf{Q}, \mathbf{A}; d_1); \dots; \mathbf{Z}_L(\mathbf{Q}, \mathbf{A}; d_L)]. \quad (10)$$

Finally, we design a learned layer-weighted scoring function for detection, i.e.,

$$\mathbf{S}_W(\mathbf{Z}(\mathbf{Q}, \mathbf{A})) = \mathbf{W}^\top \cdot \mathbf{Z}(\mathbf{Q}, \mathbf{A}), \quad (11)$$

where $\mathbf{W}$ is a d -dimensional vector, here $d = \sum_{\ell=1}^L d_\ell$. Next, we present how to learn d_ℓ , $\mathbf{P}_{d_\ell}$ and $\mathbf{W}$.

4.2 Learning the Layer-wise Projector $\mathbf{P}_{d_\ell}$

Motivated by representation learning for domain generalization [50, 59], we learn layer-wise projectors that generalize to unseen domains while preserving label-discriminative signals for hallucination detection. Each projector satisfies two conditions: (i) *Label-discriminative preservation*: it retains class-relevant information so that truthful and hallucinatory samples remain separable; (ii) *Cross-domain stability*: it reduces class-conditional domain sensitivity, yielding features that are stable under domain shifts.

Class-conditional Invariance Loss $\mathcal{L}_v(\mathbf{P}_{d_\ell})$. Following class-conditional distribution matching objectives that align feature distributions within each class across domains [37, 42], we encourage the projected representations to be domain-invariant after conditioning on the label. Given the labeled training datasets in Eq. (6), for each domain $e \in \{1, \dots, M\}$, layer $\ell \in \{1, \dots, L\}$, and class

$y \in \{0, 1\}$, we compute the empirical class-conditional mean in the projected space as

$$\mu_\ell^{e,y} = \frac{1}{|\mathcal{D}_\ell^{e,y}|} \sum_{(\mathbf{Q}^e, \mathbf{A}^e) \in \mathcal{D}_\ell^{e,y}} \mathbf{Z}_\ell(\mathbf{Q}^e, \mathbf{A}^e; d_\ell), \quad (12)$$

where $\mathcal{D}_\ell^{e,y}$ denotes the subset of $\mathcal{D}_\ell^e$ consisting of question-answer pairs with label y . To enforce class-conditional invariance, we use the following objective, which averages the pairwise discrepancies between these class-wise means:

$$\mathcal{L}_v(\mathbf{P}_{d_\ell}) = \frac{1}{M(M-1)} \sum_{y \in \{0,1\}} \sum_{1 \leq e < e' \leq M} \|\mu_\ell^{e,y} - \mu_\ell^{e',y}\|_2^2. \quad (13)$$

Separation Loss $\mathcal{L}_s(\mathbf{P}_{d_\ell})$. To retain label-discriminative information after projection, we explicitly encourage the two classes to remain well separated in the projected space. From Eq. (12), the class- y centroid averaged over the M training domains is

$$\bar{\mu}_\ell^{\cdot,y} = \frac{1}{M} \sum_{e=1}^M \mu_\ell^{e,y}. \quad (14)$$

We then promote separability by maximizing the squared distance between the two averaged centroids, equivalently by minimizing the following separation loss:

$$\mathcal{L}_s(\mathbf{P}_{d_\ell}) = -\|\bar{\mu}_\ell^{\cdot,1} - \bar{\mu}_\ell^{\cdot,0}\|_2^2. \quad (15)$$

Intuitively, $\mathcal{L}_s$ complements the invariance term by preventing the projector from collapsing the two classes while aligning domains.

Anchor Regularization $\mathcal{L}_a(\mathbf{P}_{d_\ell})$. Since the projector optimization can be sensitive to random initialization, we introduce an anchor term to stabilize training by keeping $\mathbf{P}_{d_\ell}$ close to a reference initialization $\mathbf{P}_{d_\ell,0}$:

$$\mathcal{L}_a(\mathbf{P}_{d_\ell}) = \frac{1}{d_h d_\ell} \|\mathbf{P}_{d_\ell} - \mathbf{P}_{d_\ell,0}\|_F^2, \quad (16)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathbf{P}_{d_\ell,0}$ is constructed by using *Invariant-feature Subspace Recovery* [53], which provides a data-driven initialization by recovering a low-dimensional subspace that captures invariant features.

Overall Objective of the Projectors. Combining the above components together, we learn the layer-wise projector $\mathbf{P}_{d_\ell}$ by minimizing the following loss $\mathcal{L}_{\text{pro}}(\{\mathbf{P}_{d_\ell}\}_{\ell=1}^L)$:

$$\frac{1}{L} \sum_{\ell=1}^L (\mathcal{L}_v(\mathbf{P}_{d_\ell}) + \lambda_s \mathcal{L}_s(\mathbf{P}_{d_\ell}) + \lambda_a \mathcal{L}_a(\mathbf{P}_{d_\ell})), \quad (17)$$

where $\lambda_s, \lambda_a \geq 0$ are the hyperparameters.

4.3 Learning Scoring Function $\mathbf{S}_W$

After learning the layer-wise projectors $\{\mathbf{P}_{d_\ell}\}_{\ell=1}^L$ and aggregating the projected features into $\mathbf{Z}(\mathbf{Q}, \mathbf{A})$, we train a scoring function to classify each instance as truthful or hallucinatory. Formally, we consider the scoring function given in Eq. (11), and convert it into a probabilistic prediction via a sigmoid mapping:

$$p_W(y = 1 | \mathbf{Q}, \mathbf{A}) = \sigma(\mathbf{S}_W(\mathbf{Z}(\mathbf{Q}, \mathbf{A}))), \quad (18)$$

where $\sigma(t) = 1/(1 + \exp(-t))$. Given the labeled training dataset (see Eq. (6)), we optimize the detector by minimizing the Binary Cross-Entropy (BCE) loss $\mathcal{L}_{\text{BCE}}(\mathbf{W})$:

$$-\frac{1}{\sum_{e=1}^M n^e} \sum_{e=1}^M \sum_{i=1}^{n^e} \left[y_i^e \log p_{\mathbf{W}}(y_i^e = 1 \mid \mathbf{Q}_i^e, \mathbf{A}_i^e) + (1 - y_i^e) \log(1 - p_{\mathbf{W}}(y_i^e = 1 \mid \mathbf{Q}_i^e, \mathbf{A}_i^e)) \right]. \quad (19)$$

Equivalently, by substituting Eq. (18) into Eq. (19), the optimization objective can be expressed as the standard logistic loss on the score $S_{\mathbf{W}}(\mathbf{Z}(\mathbf{Q}, \mathbf{A}))$. After training, we obtain the final detector by thresholding the score in Eq. (8).

4.4 Allocated Layer-wise Dimension d_ℓ

Here we present a training-domain-only strategy for selecting the layer-wise dimensions $\{d_\ell\}_{\ell=1}^L$, which approximates cross-domain generalization using statistics estimated from multiple training domains. The purpose of $\{d_\ell\}_{\ell=1}^L$ is to allocate effective information dimension to each layer, which preserves separability while promoting cross-domain consistency. On the one hand, this avoids excessively high feature dimensionality from naive concatenation, which would lead to computational complexity. On the other hand, adaptive allocation across layers enables the selection of features with stronger cross-domain generalization from each layer.

The challenge in designing $\{d_\ell\}_{\ell=1}^L$ lies in estimating layer-wise generalizable information using only *training-domain statistics*, in the absence of test-domain data, and allocating dimension accordingly to balance representation complexity and cross-domain generalization. To overcome this challenge, we restrict each layer-wise dimension d_ℓ to a discrete candidate set $\mathcal{K}$ of admissible projection dimensions, and design a training-only scoring criterion characterized by training-domain statistics to select layer-wise dimensions.

Layer-wise Instability β_{d_ℓ} . Here we first design an index β_{d_ℓ} that measures the cross-domain instability of layer-wise feature representations $\{\mathbf{Z}_\ell(\mathbf{Q}, \mathbf{A}; d_\ell)\}_{\ell=1}^L$ (see Eq. (9)). We begin by considering the following statistic, referred to as the *pairwise margin*:

$$D_{d_\ell}(\mathbf{Q}, \mathbf{T}, \mathbf{Q}', \mathbf{H}; \mathbf{W}) = S_{\mathbf{W}}(\mathbf{Z}_\ell(\mathbf{Q}, \mathbf{T}; d_\ell)) - S_{\mathbf{W}}(\mathbf{Z}_\ell(\mathbf{Q}', \mathbf{H}; d_\ell)). \quad (20)$$

It is clear that when $\mathbf{T}$ is the truthful response to $\mathbf{Q}$ and $\mathbf{H}$ is the hallucinated response to $\mathbf{Q}'$, the pair-wise margin directly quantifies the separability between the truthful and hallucinated answers using the scoring function $S_{\mathbf{W}}$. Then, given the e -th domain, with hallucinated-answer distribution $\mathbb{P}_{\mathbf{Q}, \mathbf{H}}^e$ and truthful-answer distribution $\mathbb{P}_{\mathbf{Q}, \mathbf{T}}^e$, introduced in Section 3.2, we compute the mean and variance of the pairwise margin:

$$\begin{aligned} M_{d_\ell}^e &= \mathbb{E}[D_{d_\ell}(\mathbf{Q}, \mathbf{T}, \mathbf{Q}', \mathbf{H}; \mathbf{W})], \\ V_{d_\ell}^e &= \text{Var}[D_{d_\ell}(\mathbf{Q}, \mathbf{T}, \mathbf{Q}', \mathbf{H}; \mathbf{W})], \end{aligned} \quad (21)$$

where $(\mathbf{Q}, \mathbf{T}) \sim \mathbb{P}_{\mathbf{Q}, \mathbf{T}}^e$ and $(\mathbf{Q}', \mathbf{H}) \sim \mathbb{P}_{\mathbf{Q}, \mathbf{H}}^e$. Next, to characterize the layer-wise cross-domain instability, we define

$$\beta_{d_\ell} = \text{Std}_{e \in \{1, \dots, M\}} R_{d_\ell}^e, \quad (22)$$

where Std is the standard deviation and

$$R_{d_\ell}^e = \frac{(M_{d_\ell}^e)^2}{(M_{d_\ell}^e)^2 + V_{d_\ell}^e}. \quad (23)$$

This ratio $R_{d_\ell}^e$ decreases as the variance $V_{d_\ell}^e$ increases or the mean $M_{d_\ell}^e$ decreases, indicating weaker separability between truthful and hallucinated responses in the learned d_ℓ -dimensional features. Conversely, a smaller variance with a larger mean drives the ratio toward 1, corresponding to stronger separability. Hence, $R_{d_\ell}^e$ serves as a measure of feature separability.

In Eq. (22), we quantify cross-domain consistency by the standard deviation of this ratio $R_{d_\ell}^e$ across domains. A larger β_{d_ℓ} implies substantial variation in separability across domains, whereas a smaller value indicates stable and consistent hallucination-truthfulness separability. Hence, β_{d_ℓ} serves as a measure of the cross-domain instability of the learned d_ℓ -dimensional representations.

Analysis of Layer-wise Instability β_{d_ℓ} . In the following, we show how β_{d_ℓ} characterizes cross-domain variation in separability.

Lemma 4.1. *Given M training domains $\mathbb{P}_{\mathbf{Q}, \mathbf{T}}^1, \dots, \mathbb{P}_{\mathbf{Q}, \mathbf{T}}^M$, and N test domains $\mathbb{P}_{\mathbf{Q}, \mathbf{T}}^{M+1}, \dots, \mathbb{P}_{\mathbf{Q}, \mathbf{T}}^{M+N}$, for any $e, e' \in \{1, \dots, M\}$ and any $t \in \{M+1, \dots, M+N\}$, we have*

$$|R_{d_\ell}^e - R_{d_\ell}^{e'}| \leq \sqrt{2M} \beta_{d_\ell}, \quad (24)$$

$$|R_{d_\ell}^t - R_{d_\ell}^e| \leq \min_{e' \in \{1, \dots, M\}} |R_{d_\ell}^t - R_{d_\ell}^{e'}| + \sqrt{2M} \beta_{d_\ell}. \quad (25)$$

Proof. See Appendix D.

Lemma 4.1 serves as an intermediate result for the subsequent analysis. Eq. (24) shows that, among the training domains, the difference in the separability measure R_{d_ℓ} between any two domains is upper bounded by $\sqrt{2M} \beta_{d_\ell}$. This implies that a small β_{d_ℓ} leads to consistent separability across training domains. Additionally, Eq. (25) extends this result in Eq. (24) to the test domains, showing that the separability gap between a test domain and any training domain is bounded by the smallest test-training discrepancy plus an additional term controlled by β_{d_ℓ} . Hence, when β_{d_ℓ} is small, the separability learned from the training domains can generalize more stably to unseen test domains. This result serves as an intermediate step for the subsequent analysis and motivates the instability-aware dimension allocation rule in Eq. (28).

Note that the upper bound in Eq. (25) contains the term

$$\min_{e' \in \{1, \dots, M\}} |R_{d_\ell}^t - R_{d_\ell}^{e'}|, \quad (26)$$

which depends on the unknown discrepancy of feature separability between the test domain and the training domains, and is therefore not directly observable in general. In Theorem 4.2, we show that under appropriate conditions, when the number of training domains M is sufficiently large, with a high probability, there exists at least one training domain whose separability measure is close to that of the test domain, rendering the above minimum term sufficiently small. This result implies that, with adequate domain coverage, the cross-domain variation of separability is predominantly governed by β_{d_ℓ} , highlighting its importance in characterizing the cross-domain instability of the learned d_ℓ -dimensional representations. The following analysis is intended to provide theoretical intuition

for instability-aware dimension allocation, rather than a universal guarantee under arbitrary domain shifts. Before presenting this theorem, we state the assumptions used in this analysis, following prior domain generalization theory [11].

Assumption 1 (Meta Assumption [11]). The distributions $\{\mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e\}_{e=1}^M$ are assumed to be drawn independently and identically distributed from a meta distribution $\mathcal{P}$ over the joint distribution space $\mathcal{P}_{Q,T,Q',H}$, where $\mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e$ is the product distribution of $\mathbb{P}_{Q,T}^e$ and $\mathbb{P}_{Q,H}^e$.

Assumption 2 (Regular Distribution [11]). $\mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e$ is a regular distribution, i.e., for any $\epsilon > 0$, $\mathcal{P}(\mathcal{N}_\epsilon^e) > 0$, where $\mathcal{P}$ is the meta distribution introduced in Assumption 1 and $\mathcal{N}_\epsilon^e = \{\mathbb{P} \in \mathcal{P}_{Q,T,Q',H} : d_{TV}(\mathbb{P}, \mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e) < \epsilon\}$, here $d_{TV}(\cdot, \cdot)$ is the total variation distance.

It is straightforward to verify that if the meta distribution $\mathcal{P}$ is discrete or continuous with a strictly positive density on its support, then Assumption 2 holds for any distribution sampled from $\mathcal{P}$.

Theorem 4.2. *Given a test domain $\mathbb{P}_{Q,T}^t$ with the corresponding hallucinated-answer distribution $\mathbb{P}_{Q,H}^t$, suppose that Assumption 1 holds and that the product distribution $\mathbb{P}_{Q,T}^t \otimes \mathbb{P}_{Q,H}^t$ is regular, i.e., Assumption 2 holds for $\mathbb{P}_{Q,T}^t \otimes \mathbb{P}_{Q,H}^t$. Then, under the stated conditions, with probability at least $1 - (1 - \mathcal{P}(\mathcal{N}_\epsilon^t))^M > 0$, the following holds: for any $e \in \{1, \dots, M\}$,*

$$|R_{d_t}^t - R_{d_t}^e| \leq C\epsilon + \sqrt{2M}\beta_{d_t}, \quad (27)$$

where $C > 0$ is a uniform constant.

Proof. See Appendix F.

Theorem 4.2 shows that the separability gap between a test domain and the training domains is jointly controlled by a neighborhood term $C\epsilon$ and the instability measure β_{d_t} . When the number of training domains M is sufficiently large and ϵ is small, then with a high probability, the bound is dominated by β_{d_t} , highlighting its importance in characterizing the cross-domain instability of the learned d_t -dimensional representations.

Accordingly, we select the layer-wise dimension by maximizing separability score penalized by the instability: given a candidate set $\mathcal{K}$ for admissible projection dimensions,

$$d_t^* = \arg \max_{d_t \in \mathcal{K}} R_{d_t} = \mathbb{E}_{e' \in \{1, \dots, M\}} [R_{d_t}^{e'}] - \lambda \cdot \beta_{d_t}, \quad (28)$$

where λ balances the separability and stability.

We use $R_{d_t}^e$ as a tractable layer-wise separability surrogate because its moment-based form makes the role of β_{d_t} explicit and enables the analysis in Theorem 4.2. In implementation, we instantiate the training-only LODO selection with AUROC for finite-sample stability, details are provided in Appendix B. And the overall framework is provided in Appendix C.

5 Experiments

5.1 Evaluation Setup

Datasets. Following the data construction and prompting setup of [44], we evaluate SALA on four generative open-domain QA

benchmarks: TriviaQA [24], TruthfulQA [33], NQ Open [30], and SciQ [54]. Following [22], we treat each dataset as a domain and study cross-domain generalization under a leave-one-domain-out (LODO) protocol: in each fold, we train on three training domains and evaluate on the held-out test domain. For all methods, we construct the training data by randomly sampling 75% of the QA pairs from each training domain; we fix this split and reuse it across baselines to ensure a fair, like-for-like comparison. The remaining 25% from each training domain is *not* used for training; it is reserved for training-only validation when needed for model selection (e.g., layer selection or hyperparameter tuning), and no test-domain data is accessed during training or selection. For evaluation, we reserve 75% of the held-out test domain as the test set and do not use the remaining 25%, yielding a lower-variance and more reliable estimate of out-of-domain performance.

Models and Baselines. We evaluate our method using two families of LLMs which provide accessible internal representations: LLaMA-3.1-8B & LLaMA-3-8B-Instruct [36] and Qwen-2.5-7B & Qwen-2.5-14B [47].

The baselines considered include: (1) training-free methods: Perplexity [31], Length Normalized Entropy (LN-entropy) [38], Semantic Entropy [29], Lexical Similarity [34], EigenScore [6], SelfCheckGPT [39], and Verbalize [34]; (2) training-based methods: SAPLMA [3], PRISM (SAPLMA instantiation) [56], ICR Probe [58], TSV (full training) [44]. In particular, PRISM and ICR Probe were originally proposed with an emphasis on generalization, and thus serve as strong and directly relevant baselines for evaluation. More details are provided in Appendix B.

Evaluation. Following [12, 29], detector performance is evaluated using the area under the receiver operating characteristic curve (AUROC), with binary labels assigned according to [44].

5.2 Experimental Results

Main Results. In Table 1, we compare SALA with three categories of baselines across all test domains and LLM backbones. Overall, SALA performs strongly, achieving the highest mean AUROC and remaining competitive across most domain-backbone settings. A consistent pattern is that several training-based baselines exhibit larger performance variance under domain shift, suggesting limited transferability when decision boundaries are fit solely from training-domain supervision. Training-free methods tend to be more stable across settings, but typically trail approaches that leverage representation learning or cross-domain objectives. By contrast, SALA improves robustness via stability-aware layer aggregation, and its advantage is most pronounced on TruthfulQA, where it achieves the largest gains over competing baselines.

For baselines that involve layer selection (e.g., SAPLMA and PRISM), we follow the best-practice selection strategies recommended in their original papers to ensure a fair comparison. In particular, for SAPLMA, we report two commonly used variants: (i) concatenating the last five layers for detection, and (ii) selecting the probing layer using a validation set constructed only from the training domains. Notably, although validation-based layer selection can be highly effective in-domain for internal-state-based detectors, our results indicate that this heuristic can break under

Table 1: AUROC (%) across four test domains and four LLM backbones. [▲]: SAPLMA using *last-5-layer concatenation* for detection. [▼]: SAPLMA using *training-validation layer selection* for cross-domain generalization. Best results are bold, second-best are underlined.

Method	TruthfulQA				TriviaQA				SciQ				NQ Open			
	LLaMA 3.1-8B	LLaMA3 3-8B-Ins	Qwen 2.5-7B	Qwen 2.5-14B	LLaMA 3.1-8B	LLaMA3 3-8B-Ins	Qwen 2.5-7B	Qwen 2.5-14B	LLaMA 3.1-8B	LLaMA3 3-8B-Ins	Qwen 2.5-7B	Qwen 2.5-14B	LLaMA 3.1-8B	LLaMA3 3-8B-Ins	Qwen 2.5-7B	Qwen 2.5-14B
Training-free methods																
Perplexity _(EMNLP2023)	71.33	64.42	70.44	73.05	50.12	67.96	55.32	68.21	57.29	53.37	56.63	55.52	60.52	58.66	71.65	72.72
LN-Entropy _(ICLR2021)	<u>71.59</u>	67.44	75.60	<u>76.14</u>	62.43	71.09	54.23	50.69	59.90	52.43	55.02	66.97	59.29	53.43	62.27	56.98
Semantic Entropy _(ICLR2023)	50.53	51.76	52.17	50.64	71.73	<u>73.51</u>	58.72	72.92	62.77	66.56	69.71	59.60	58.84	74.87	74.01	73.99
Lexical Similarity _(TMLR2024)	53.97	59.30	51.45	52.77	71.01	70.74	63.01	75.77	59.57	59.16	68.37	63.11	58.83	63.24	74.34	75.52
EigenScore _(ICLR2024)	52.72	50.67	62.18	65.12	69.52	66.73	62.23	71.08	53.71	64.90	54.17	71.95	52.30	67.54	61.71	73.54
SelfCheckGPT _(EMNLP2023)	50.80	59.12	58.15	54.96	71.98	70.13	61.98	67.12	61.14	57.50	59.19	56.33	60.01	62.06	65.84	63.88
Verbalize _(TMLR2024)	55.97	50.29	65.61	61.02	63.81	72.42	55.34	58.17	56.41	63.69	59.43	53.74	52.07	61.47	52.15	53.85
Training-based methods																
SAPLMA [▲] _(EMNLP2023)	68.72	62.97	74.43	71.47	78.20	70.60	<u>78.85</u>	<u>85.49</u>	81.02	<u>76.23</u>	<u>86.19</u>	<u>88.67</u>	64.86	74.58	<u>81.50</u>	86.84
SAPLMA [▼] _(EMNLP2023)	59.15	50.10	51.99	50.31	75.99	61.52	56.54	58.50	<u>81.87</u>	67.53	56.20	58.39	65.12	58.39	58.13	58.89
TSV (full-training) _(ICML2025)	71.52	<u>69.73</u>	<u>81.05</u>	75.33	77.00	67.74	77.62	82.78	80.33	62.46	82.68	88.07	66.96	77.80	80.10	85.05
Cross-domain specialized methods																
PRISM _(ACL2025)	70.30	62.64	79.42	72.12	70.83	68.36	71.66	84.38	75.94	72.85	87.73	88.64	61.97	<u>78.26</u>	80.75	<u>87.14</u>
ICR Probe _(ACL2025)	71.49	60.17	76.61	70.13	<u>80.78</u>	69.58	72.53	81.25	80.38	72.96	76.25	86.64	<u>66.28</u>	76.27	82.26	86.73
SALA (Ours)	79.38	71.52	86.02	76.34	81.16	74.03	81.73	86.23	82.45	77.99	86.00	88.78	65.57	80.20	80.45	87.81

Table 2: Ablation study (AUROC, %) on TruthfulQA and TriviaQA. Raw Concat tests whether gains come from using more layers, while the other rows remove or combine projector objectives ($\mathcal{L}_v$, $\mathcal{L}_s$) to isolate their contributions.

Methods	TruthfulQA		TriviaQA	
	LLaMA 3.1-8B	Qwen 2.5-7B	LLaMA 3.1-8B	Qwen 2.5-7B
Raw Concat (full layers)	66.56	80.72	80.60	80.92
Proj. only $\mathcal{L}_v$	52.43	66.56	74.25	80.33
Proj. only $\mathcal{L}_s$	74.24	82.13	76.62	79.63
Proj. only $\mathcal{L}_v + \mathcal{L}_s$	77.19	82.12	77.10	79.88
SALA (Ours)	79.38	86.02	81.16	81.73

MGHD: the layer preferred by training-domain validation is often misaligned with the optimal layer on unseen domains, resulting in clear degradation under domain shifts. This observation further reinforces the brittleness of fixed-layer probing in multi-domain generalization and motivates adaptive aggregation across layers.

This mismatch suggests that the optimal probing layer is domain-dependent, so fixed-layer detectors can be brittle under MGHD. SALA mitigates this issue by aggregating evidence across layers while prioritizing subspaces that are both stable across training domains and predictive of the labels. We also observe that gains can be smaller when the test domain is closer to the training mixture, indicating that SALA mainly improves robustness in harder shifts rather than uniformly boosting in-domain accuracy.

Ablation Study. In Table 2 and Table 3, we ablate projector-training design choices in SALA using two representative backbones across test domains. As a control, we include *Raw Concat* (all

Table 3: Ablation study (AUROC, %) on SciQ and NQ Open. Same ablations as Table 2 on the remaining domains.

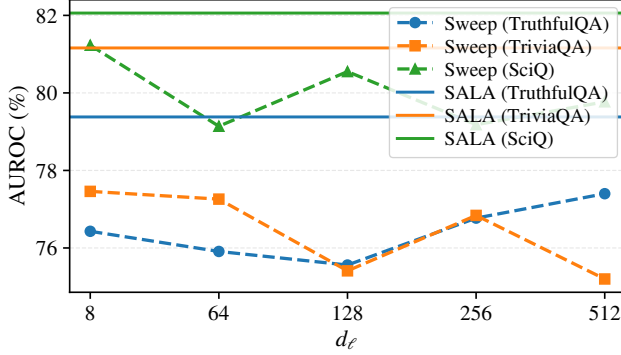
Methods	SciQ		NQ Open	
	LLaMA 3.1-8B	Qwen 2.5-7B	LLaMA 3.1-8B	Qwen 2.5-7B
Raw Concat (full layers)	81.89	84.74	65.50	80.26
Proj. only $\mathcal{L}_v$	82.76	84.68	67.31	65.69
Proj. only $\mathcal{L}_s$	80.78	80.48	63.61	64.58
Proj. only $\mathcal{L}_v + \mathcal{L}_s$	79.86	81.28	64.11	63.69
SALA (Ours)	82.45	86.00	65.57	80.45

layers) to test whether gains come simply from using more features. While naive stacking can be competitive on some tests, it degrades markedly on TruthfulQA, suggesting that indiscriminately mixing all layers introduces domain-specific noise and hurts transfer. In contrast, SALA typically improves over raw concatenation, indicating that the gains come from stability-aware feature extraction and aggregation rather than feature quantity. We further ablate the projector training objectives. Optimizing the projector with only $\mathcal{L}_v$ or only $\mathcal{L}_s$ fails to deliver consistently strong performance across domains, highlighting the importance of balancing separability and stability. All three components are essential to SALA.

Then, we compare SALA’s adaptive layer-wise dimension allocation with a fixed global-dimension baseline that applies the same projection dimension to all layers (Figure 3). Overall, per-layer allocation yields consistently stronger performance: SALA emphasizes layers with stable, label-discriminative signals and down-weights domain-specific or noisy layers, whereas a uniform dimension over-invests in less transferable layers and under-utilizes informative

Table 4: AUROC (%) across four mathematical and logical reasoning benchmarks and four LLM backbones. Best results are bold, second-best are underlined.

Method	TheoremQA				GSM8K				SVAMP				MATH			
	LLaMA 3.1-8B	LLaMA3 3.2-3B-Ins.	Qwen 3-8B	Qwen 3-14B	LLaMA 3.1-8B	LLaMA3 3.2-3B-Ins.	Qwen 3-8B	Qwen 3-14B	LLaMA 3.1-8B	LLaMA3 3.2-3B-Ins.	Qwen 3-8B	Qwen 3-14B	LLaMA 3.1-8B	LLaMA3 3.2-3B-Ins.	Qwen 3-8B	Qwen 3-14B
Training-free methods																
Perplexity	56.87	51.35	57.97	56.79	66.93	70.73	67.97	72.70	76.06	68.38	76.52	78.47	55.76	51.64	50.74	<u>72.86</u>
Self-Evaluator	52.36	52.53	51.27	51.34	51.15	50.79	51.43	52.43	56.57	54.74	54.76	59.54	52.24	50.54	<u>73.74</u>	51.66
Training-based methods																
SAPLMA	<u>73.59</u>	68.56	<u>64.14</u>	<u>65.33</u>	72.76	69.17	65.64	75.91	<u>81.19</u>	71.29	78.97	<u>90.06</u>	<u>79.64</u>	<u>74.11</u>	70.30	72.07
Cross-domain specialized methods																
PRISM	74.62	<u>69.38</u>	59.14	63.44	<u>73.68</u>	68.99	63.29	74.16	78.00	<u>73.14</u>	81.56	88.52	56.39	59.48	67.88	54.61
SALA (Ours)	73.25	69.83	65.43	66.12	74.00	<u>69.53</u>	<u>66.50</u>	<u>75.59</u>	85.68	74.01	<u>80.51</u>	91.52	81.21	75.53	75.37	73.79

**Figure 3: Sensitivity to uniform layer dimension and the benefits of adaptive allocation.** We sweep a uniform per-layer dimension $d_l \in \{8, 64, 128, 256, 512\}$ for LLaMA-3.1-8B and report AUROC on three test domains. The dashed curves show strong performance fluctuations under uniform dimension, while SALA yields consistent gains, indicated by horizontal reference lines. This benefit is most pronounced under domain shift, where the reliable layers vary across domains and layer-agnostic dimension choices become brittle.

Testing on Mathematical and Logical Reasoning. Beyond factual QA, we further assess whether our framework remains reliable on *mathematical and logical reasoning* benchmarks (Table 4). Unlike standard factual QA, these tasks demand long reasoning trajectories and suffer from severe instability. Even minor prompt variations can easily trigger complex errors, spanning from simple arithmetic slips to deep logical inconsistencies. We therefore evaluate cross-domain generalization on four reasoning benchmarks: GSM8K [9] (adopted to isolate core reasoning from multilingual variance using its larger testbed), MATH [17], SVAMP [45], and TheoremQA [7], to test whether stability-aware, layer-wise allocation generalizes in this reasoning-heavy regime.

Efficiency Analysis. Compared to the lightweight SAPLMA (Table 5), SALA achieves similar training time and peak inference VRAM. Although multi-layer aggregation adds marginal (low-millisecond) inference latency, it significantly improves cross-domain robustness without prohibitive overhead.

Hyperparameter sensitivity. We examine the sensitivity of SALA to its two regularization weights, λ_s and λ_a , in Appendix A. The results show that SALA maintains stable performance across a

Table 5: Computational cost comparison on LLaMA-3.1-8B with TruthfulQA under the same protocol.

Metric	SAPLMA	SALA
Training Time (s)	185.803	179.714
Inference Latency (ms/sample)	0.161	2.429
Inference Peak VRAM (MB)	52.27	50.86

broad range of hyperparameter configurations, with no clear dependence on a fragile optimum. In particular, the AUROC landscape remains smooth, suggesting that the proposed stability-aware objective is robust to moderate perturbations of the regularization strengths. Additional analyses and implementation details are available at <https://github.com/Nellie179/SALA>.

6 Conclusion

In this paper, we study multi-domain generalization for hallucination detection (MGHD), where detectors are trained on multiple training domains and transferred to unseen test domains. We propose SALA, a stability-aware framework that learns layer-wise stable features and aggregates them to improve cross-domain robustness. We provide theoretical analysis and extensive experiments showing that SALA mitigates the layer-index shift phenomenon and consistently outperforms strong baselines under domain shifts. We further validate its effectiveness via ablation studies and stress tests on mathematical and logical reasoning benchmarks. Overall, our results suggest that carefully extracting and combining stable internal signals can substantially strengthen the out-of-domain reliability of hallucination detectors in diverse and realistic settings.

Acknowledgments

This work is partially supported by the Australian Research Council Discovery Early Career Researcher Awards DE250100363, and the Australian Research Council Future Fellowship FT230100121. Xinyi Li and Zhen Fang are supported by DE250100363. Yongxin Deng and Ling Chen are supported by FT230100121. Yadan Luo is supported by the Australian Research Council Discovery Early Career Researcher Awards DE240100105. Sean Du is supported by the Nanyang Technological University Start-Up Grant 025730-00001 and the Ministry of Education, Singapore, AcRF Tier 1 Seed Funding Grant RS 24/25 025822-00001. We also thank Sze To Leung for assistance with code engineering and implementation organization.

References

- [1] Kartik Ahuja, Ethan Caballero, Dinghui Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. In *Proc. of NeurIPS*. 3438–3450.
- [2] Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *Proc. of ICLR*. 1–14.
- [3] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Proc. of EMNLP Findings*. 967–976.
- [4] Mohammad Beigi, Ying Shen, Runing Yang, et al. 2024. InternalInspector I^2 : Robust Confidence Estimation in LLMs through Internal States. In *Proc. of EMNLP Findings*. 12847–12865.
- [5] Tianchi Cai, Zhiwen Tan, Xierui Song, Tao Sun, Jiyang Jiang, Yunqi Xu, Yinger Zhang, and Jinjie Gu. 2024. FoRAG: Factuality-optimized Retrieval Augmented Generation for Web-enhanced Long-form Question Answering. In *Proc. of KDD*. 199–210.
- [6] 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 *Proc. of ICLR*.
- [7] Wenhui Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A Theorem-driven Question Answering Dataset. In *Proc. of EMNLP*. 7889–7901.
- [8] Yuyan Chen, Zehao Li, Shuangjie You, et al. 2025. Attributive Reasoning for Hallucination Diagnosis of Large Language Models. In *Proc. of AAAI*. 23660–23668.
- [9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. Training Verifiers to Solve Math Word Problems. *ArXiv preprint* (2021).
- [10] Yongxin Deng, Zhen Fang, Sharon Li, and Ling Chen. 2026. Beyond In-Domain Detection: SpikeScore for Cross-Domain Hallucination Detection.
- [11] Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu. 2021. Confident Anchor-Induced Multi-Source Free Domain Adaptation. In *Proc. of NeurIPS*. 2848–2860.
- [12] Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. In *Proc. of NeurIPS*.
- [13] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, et al. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* (2024), 625–630.
- [14] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proc. of ACL*. 16477–16508.
- [15] Jiatong Han, Neil Band, Muhammed Razzak, et al. 2025. Simple Factuality Probes Detect Hallucinations in Long-Form Natural Language Generation. In *Proc. of EMNLP Findings*. 16209–16226.
- [16] Emma Harvey, Allison Koenecke, and René F. Kizilcec. 2025. "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proc. of ICLR*.
- [18] John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proc. of NAACL*. 4129–4138.
- [19] Yuanwei Hu, Bo Peng, Yadan Luo, Zhen Fang, Ling Chen, and Jie Lu. 2026. Respecting Modality Gap in Post-hoc Out-of-distribution Detection with Pre-trained Vision-Language Models. *arXiv:2605.26661*
- [20] Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024. Prompt Perturbation in Retrieval-Augmented Generation based Large Language Models. In *Proc. of KDD*. 1119–1130.
- [21] Lei Huang, Weijiang Yu, Weitao Ma, et al. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* (2025), 1–55.
- [22] Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, et al. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence* (2023), 1161–1174.
- [23] Ziwei Ji, Nayeon Lee, Rita Frieske, et al. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* (2023), 1–38.
- [24] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proc. of ACL*. 1601–1611.
- [25] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, et al. 2025. Why Language Models Hallucinate. *ArXiv preprint* (2025).
- [26] Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *ArXiv preprint* (2023).
- [27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proc. of ICML*. 5637–5664.
- [28] Jannik Kossen, Jiatong Han, Muhammed Razzak, et al. 2024. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. *ArXiv preprint* (2024).
- [29] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *Proc. of ICLR*.
- [30] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *TACL* (2019), 452–466.
- [31] Yifan Li, Yifan Du, Kun Zhou, Jimpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proc. of EMNLP*. 292–305.
- [32] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. *TMLR* (2022).
- [33] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proc. of ACL*. 3214–3252.
- [34] Zhen Lin, Shubendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *TMLR* (2024).
- [35] Xiaohu Liu, Tiejun Chen, Longchao Da, et al. 2025. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. In *Proc. of KDD*. 6107–6117.
- [36] Llama Team. 2024. The Llama 3 Herd of Models. *ArXiv preprint* (2024).
- [37] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jianguang Sun, and Philip S. Yu. 2013. Transfer Feature Learning with Joint Distribution Adaptation. In *Proc. of ICCV*. 2200–2207.
- [38] Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *Proc. of ICLR*.
- [39] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proc. of EMNLP*. 9004–9017.
- [40] Mateusz Michalkiewicz, Masoud Faraki, Xiang Yu, Manmohan Chandraker, and Mahsa Baktashmotlagh. 2023. Domain Generalization Guided by Gradient Signal to Noise Ratio of Parameters. In *Proc. of ICCV*. 6154–6165.
- [41] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, et al. 2024. Large Language Models: A Survey. *ArXiv preprint* (2024).
- [42] Saeid Motiani, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. 2017. Unified Deep Supervised Domain Adaptation and Generalization. In *Proc. of ICCV*. 5716–5726.
- [43] Oscar Obeso, Andy Arditi, Javier Ferrando, et al. 2025. Real-Time Detection of Hallucinated Entities in Long-Form Generation. *ArXiv preprint* (2025).
- [44] Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, et al. 2025. Steer LLM Latents for Hallucination Detection. In *Proc. of ICML*.
- [45] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems?. In *Proc. of NAACL*. 2080–2094.
- [46] Rui Qiao and Bryan Kian Hsiang Low. 2024. Understanding Domain Generalization: A Noise Robustness Perspective. In *Proc. of ICLR*.
- [47] Qwen, An Yang, Baosong Yang, et al. 2024. Qwen2.5 Technical Report. *ArXiv preprint* (2024).
- [48] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. Out-of-Distribution Detection and Selective Generation for Conditional Language Models. In *Proc. of ICLR*.
- [49] Dimitri Roustan and François Bastardot. 2025. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interactive Journal of Medical Research* (2025), e59823.
- [50] Aadars Sahoo, Rameswar Panda, Rogerio Feris, et al. 2023. Select, Label, and Mix: Learning Discriminative Invariant Feature Representations for Partial Domain Adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4210–4219.
- [51] Ben Snyder, Marius Moisesescu, and Muhammad Bilal Zafar. 2024. On Early Detection of Hallucinations in Factual Question Answering. In *Proc. of KDD*. 2721–2732.
- [52] Gaurang Sriraman, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *Proc. of NeurIPS*.
- [53] Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. 2022. Provable Domain Generalization via Invariant-Feature Subspace Recovery. In *Proc. of ICML*. 23018–23033.
- [54] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 94–106.
- [55] 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 *Proc. of ICLR*.

- [56] Fujie Zhang, Peiqi Yu, Biao Yi, et al. 2025. Prompt-Guided Internal States for Hallucination Detection of Large Language Models. In *Proc. of ACL*. 21806–21818.
- [57] Luan Zhang, Dandan Song, Zhijing Wu, et al. 2025. Detecting Hallucination in Large Language Models Through Deep Internal Representation Analysis. In *Proc. of IJCAI*. 8357–8365.
- [58] Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, et al. 2025. ICR Probe: Tracking Hidden State Dynamics for Reliable Hallucination Detection in LLMs. In *Proc. of ACL*. 17986–18002.
- [59] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. 2019. On Learning Invariant Representations for Domain Adaptation. In *Proc. of ICML*. 7523–7532.

A Hyperparameter Sensitivity

Hyperparameter Sensitivity. Figure 4 studies the sensitivity of SALA to the two regularization weights, λ_s and λ_a , on TruthfulQA. Across a wide range of configurations, the test AUROC varies only within a narrow interval, indicating that the method is not dependent on a finely tuned hyperparameter choice. In particular, the high-performing region is not concentrated at a single isolated point: configurations around moderate stability regularization, such as $\lambda_s = 0.4$, remain competitive under different values of λ_a . This suggests that the stability-aware objective provides a consistent regularization effect rather than introducing fragile trade-offs between separability and stability. Overall, the smooth and relatively flat AUROC landscape supports the robustness of SALA and shows that its generalization performance is insensitive to moderate perturbations of λ_s and λ_a .

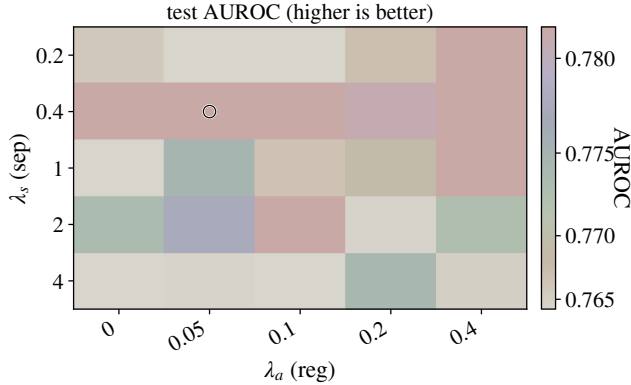


Figure 4: Hyperparameter sensitivity. Test-domain AUROC heatmap over separation (λ_s) and regularization (λ_a) weights (fixed λ_v). Optimal configuration marked.

B Details of Experiments

Practical Instantiation of Eq. (28) (Training-only). Eq. (28) is stated using the moment-based proxy $R_{d_t}^e$ for analytical tractability. This proxy is used in Phase 1 for layer-wise dimension allocation, rather than as the classification objective. After the dimensions and projectors are determined, the final detector is trained in Phase 2 by optimizing the BCE loss in Eq. (19), which defines the final probe’s classification objective. In practice, we use the empirical AUROC of a lightweight probe as a plug-in surrogate and perform leave-one-domain-out (LODO) over the M training domains: each fold holds out an entire training domain e , trains on $\{1, \dots, M\} \setminus \{e\}$,

and evaluates AUROC on e to obtain $\widehat{\text{AUROC}}_{d_t}^e$. We select d_t^* by maximizing $\mathbb{E}_e[\widehat{\text{AUROC}}_{d_t}^e] - \lambda \text{Std}_e[\widehat{\text{AUROC}}_{d_t}^e]$, no test information is used for selecting d_t .

Effect of Dimension-Allocation Metric. While $R_{d_t}^e$ provides a tractable theoretical proxy for stability, we empirically adopt AUROC for layer-wise dimension allocation. Table 6 demonstrates that AUROC yields superior mean performance and strictly lower variance, justifying this practical design choice.

More results are available at <https://github.com/Nellie179/SALA>.

Table 6: Effect of the dimension-allocation metric on LLaMA-3.1-8B with TruthfulQA. We report the final detector AUROC over random seeds.

Selection Metric	Mean AUROC	Std.
AUROC	80.46	0.25
$R_{d_t}^e$	79.70	0.71

C Overall Framework

Algorithm 1 Stability-Aware Adaptive Layer Aggregation

- 1: **Parameters:** $\lambda, \lambda_s, \lambda_a, \mathcal{K}$, update-domain size m , domain-selection strategy
- 2: **Input:** Training datasets $\mathcal{D}_1^1, \dots, \mathcal{D}_\ell^M$.
- 3: Extract last-token embeddings $Z_\ell(\mathbf{Q}, \mathbf{A})$ from LLM $\mathbb{P}_\theta$ by Eq. (2) and Eq. (3).
- 4: **Phase I: Layer-wise dimension allocation**
- 5: **for** $\ell = 1$ **to** L **do**
- 6: **for** $d \in \mathcal{K}$ **do**
- 7: Generate choices of m training datasets by enumeration when feasible, or by sampling for efficiency.
- 8: **for** each generated choice **do**
- 9: Train the candidate projector $\mathbb{P}_d$ and lightweight scoring function $\mathbf{W}$ on the selected training datasets using Eq. (17) and Eq. (19).
- 10: Compute the held-out transfer score R_d introduced in Eq. (28) on the remaining training datasets.
- 11: **end for**
- 12: Aggregate the held-out scores for candidate dimension d .
- 13: **end for**
- 14: Select d_ℓ^* that maximizes the aggregated held-out score.
- 15: **end for**
- 16: Train the projectors $\{\mathbb{P}_{d_\ell^*}\}_{\ell=1}^L$ on all training datasets by minimizing Eq. (17).
- 17: Extract projected features $\{Z_\ell(\mathbf{Q}, \mathbf{A}; d_\ell^*)\}_{\ell=1}^L$.
- 18: **Phase II: Learning scoring function**
- 19: Obtain $\mathbf{Z}(\mathbf{Q}, \mathbf{A})$ based on Eq. (10).
- 20: Train $\mathbf{W}$ on all training datasets using Eq. (19).
- 21: Compute the hallucination score $\mathbf{W}^\top \mathbf{Z}(\mathbf{Q}, \mathbf{A})$ for unseen test-domain examples.

D Proof of Lemma 4.1

Proof.

Let $\bar{R}_{d_t} = \frac{1}{M} \sum_{e=1}^M R_{d_t}^e$ and $r_{d_t}^e = R_{d_t}^e - \bar{R}_{d_t}$. Then $\sum_{e=1}^M (r_{d_t}^e)^2 = M\beta_{d_t}^2$. For any two training domains e, e' , using $(a-b)^2 \leq 2(a^2+b^2)$, we have

$$\begin{aligned} |R_{d_t}^e - R_{d_t}^{e'}|^2 &= |r_{d_t}^e - r_{d_t}^{e'}|^2 \\ &\leq 2((r_{d_t}^e)^2 + (r_{d_t}^{e'})^2) \leq 2M\beta_{d_t}^2. \end{aligned}$$

Thus,

$$|R_{d_t}^e - R_{d_t}^{e'}| \leq \sqrt{2M} \beta_{d_t}.$$

For any unseen test domain t , training domain e , and any training domain e' , the triangle inequality gives

$$|R_{d_t}^t - R_{d_t}^e| \leq |R_{d_t}^t - R_{d_t}^{e'}| + |R_{d_t}^{e'} - R_{d_t}^e|.$$

Applying the training-domain bound above to the second term and minimizing the first term over e' yields

$$|R_{d_t}^t - R_{d_t}^e| \leq \min_{e' \in \{1, \dots, M\}} |R_{d_t}^t - R_{d_t}^{e'}| + \sqrt{2M} \beta_{d_t}.$$

This proves the claimed bound. A more detailed derivation is provided in the released GitHub repository.

E Lemmas

Lemma E.1. Under two assumptions described as follows,

(A1) (**Bounded score gap**). $|D_{d_t}^e| \leq B_D$ almost surely for some constant $B_D > 0$.

(A2) (**Non-degeneracy**). There exists $c > 0$ such that for all $e \in \{1, \dots, M\}$,

$$(M_{d_t}^e)^2 + V_{d_t}^e \geq c. \quad (29)$$

there exists a uniform constant C , such that for any domain indices two e, e' ,

$$|R_{d_t}^e - R_{d_t}^{e'}| \leq C d_{TV}(\mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e, \mathbb{P}_{Q,T}^{e'} \otimes \mathbb{P}_{Q,H}^{e'}),$$

where $d_{TV}(\cdot, \cdot)$ denotes total variation distance, and $\mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e$ and $\mathbb{P}_{Q,T}^{e'} \otimes \mathbb{P}_{Q,H}^{e'}$ denote product measures, indicating that truthful and hallucinated question-answer pairs are sampled independently from domains e and e' , respectively.

We first state the following standard lemma.

Lemma E.2. (TV controls expectation differences of bounded functions). Let f be measurable and bounded with $\|f\|_\infty \leq B$. Then for any probability measures P, Q ,

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq 2B d_{TV}(P, Q),$$

where $d_{TV}(P, Q) = \sup_A |P(A) - Q(A)|$.

Proof. Define $g = (f + B)/(2B)$. Then $0 \leq g \leq 1$ and

$$\mathbb{E}_P[f] - \mathbb{E}_Q[f] = 2B(\mathbb{E}_P[g] - \mathbb{E}_Q[g]).$$

For any $0 \leq g \leq 1$,

$$|\mathbb{E}_P[g] - \mathbb{E}_Q[g]| \leq \sup_{0 \leq h \leq 1} |\mathbb{E}_P[h] - \mathbb{E}_Q[h]| = d_{TV}(P, Q).$$

This proves the lemma.

Let

$\mathbb{P}^e = \mathbb{P}_{Q,T}^e \otimes \mathbb{P}_{Q,H}^e$, $\mathbb{P}^{e'} = \mathbb{P}_{Q,T}^{e'} \otimes \mathbb{P}_{Q,H}^{e'}$. Under Assumption 1, write $D_{d_t}^e$ as a bounded function $f(Q, T, Q', H)$ with $\|f\|_\infty \leq B_D$. Recall that

$$M_{d_t}^e = \mathbb{E}_{\mathbb{P}^e}[f], \quad V_{d_t}^e = \text{Var}_{\mathbb{P}^e}(f), \quad R_{d_t}^e = g(M_{d_t}^e, V_{d_t}^e),$$

where $g(M, V) = M^2/(M^2 + V)$.

We first control the sensitivity of g . Its partial derivatives are

$$\frac{\partial g}{\partial M} = \frac{2MV}{(M^2 + V)^2}, \quad \frac{\partial g}{\partial V} = -\frac{M^2}{(M^2 + V)^2}.$$

By Assumptions 1–2, $|M| \leq B_D$, $0 \leq V \leq B_D^2$, and $M^2 + V \geq c$. Hence

$$\left| \frac{\partial g}{\partial M} \right| \leq \frac{2B_D^3}{c^2}, \quad \left| \frac{\partial g}{\partial V} \right| \leq \frac{B_D^2}{c^2}.$$

The mean value theorem gives

$$|R_{d_t}^e - R_{d_t}^{e'}| \leq \frac{2B_D^3}{c^2} |M_{d_t}^e - M_{d_t}^{e'}| + \frac{B_D^2}{c^2} |V_{d_t}^e - V_{d_t}^{e'}|.$$

It remains to bound the mean and variance terms by total variation distance. By the lemma,

$$|M_{d_t}^e - M_{d_t}^{e'}| \leq 2B_D d_{TV}(\mathbb{P}^e, \mathbb{P}^{e'}).$$

For the variance term,

$$\begin{aligned} |V_{d_t}^e - V_{d_t}^{e'}| &\leq |\mathbb{E}_{\mathbb{P}^e}[f^2] - \mathbb{E}_{\mathbb{P}^{e'}}[f^2]| + |(M_{d_t}^e)^2 - (M_{d_t}^{e'})^2| \\ &\leq 2B_D^2 d_{TV}(\mathbb{P}^e, \mathbb{P}^{e'}) + 2B_D |M_{d_t}^e - M_{d_t}^{e'}| \\ &\leq 6B_D^2 d_{TV}(\mathbb{P}^e, \mathbb{P}^{e'}). \end{aligned}$$

Substituting the two bounds above, we obtain

$$\begin{aligned} |R_{d_t}^e - R_{d_t}^{e'}| &\leq \frac{2B_D^3}{c^2} \cdot 2B_D d_{TV}(\mathbb{P}^e, \mathbb{P}^{e'}) + \frac{B_D^2}{c^2} \cdot 6B_D^2 d_{TV}(\mathbb{P}^e, \mathbb{P}^{e'}) \\ &= \frac{10B_D^4}{c^2} d_{TV}(\mathbb{P}^e, \mathbb{P}^{e'}). \end{aligned}$$

This proves the desired bound. A more detailed derivation is provided in the released GitHub repository.

F Proof of Theorem 4.2

Proof. According to Lemma 4.1 and Lemma E.1, we can prove that

$$\begin{aligned} |R_{d_t}^t - R_{d_t}^e| &\leq \sqrt{2M} \beta_{d_t} \\ &+ C \min_{e' \in \{1, \dots, M\}} d_{TV}(\mathbb{P}_{Q,T}^t \otimes \mathbb{P}_{Q,H}^t, \mathbb{P}_{Q,T}^{e'} \otimes \mathbb{P}_{Q,H}^{e'}). \end{aligned} \quad (30)$$

It is easy to check that with the probability at least $1 - (1 - \mathcal{P}(\mathcal{N}_\epsilon^t))^M > 0$, there exists some $e^* \in \{1, \dots, M\}$ such that $\mathbb{P}_{Q,T}^{e^*} \otimes \mathbb{P}_{Q,H}^{e^*} \in \mathcal{N}_\epsilon^t$, which implies that

$$\begin{aligned} \min_{e' \in \{1, \dots, M\}} d_{TV}(\mathbb{P}_{Q,T}^t \otimes \mathbb{P}_{Q,H}^t, \mathbb{P}_{Q,T}^{e'} \otimes \mathbb{P}_{Q,H}^{e'}) \\ \leq d_{TV}(\mathbb{P}_{Q,T}^t \otimes \mathbb{P}_{Q,H}^t, \mathbb{P}_{Q,T}^{e^*} \otimes \mathbb{P}_{Q,H}^{e^*}) \leq \epsilon. \end{aligned}$$

We complete the proof.