LLM HALLUCINATION DETECTION VIA LOW-RANK MULTILINEAR DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) frequently generate fluent yet factually incorrect responses, raising concerns about their reliability. Recent hallucination detectors increasingly exploit internal representations, but extracting informative signals from high-dimensional hidden states remains challenging. Existing approaches often rely on selected activations or pooling-based compression, potentially overlooking discriminative information distributed across Transformer layers, feature modes, and tokens. We propose **LOTUS** (**L**ow-rank **O**rder-statistic **TU**cker-based **S**ubspace learning)), a multilinear low-rank decompsoition framework for hallucination detection that extracts compact representations from LLM hidden states while reducing dependence on labeled data. Our approach first summarizes variable-length token representations using complementary statistical descriptors that capture peak activations, activation distributions, and cross-layer changes. We then leverage low-rank multilinear decomposition to extract compact representations that capture the underlying structure of LLM hidden states, which are subsequently used to train a lightweight hallucination classifier. Experiments across multiple LLMs and question-answering benchmarks demonstrate improved hallucination detection performance over recent state-of-the-art baselines.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities in language understanding, generation, and reasoning (Huang & Chang, 2023), enabling widespread use in question answering, summarization, and dialogue systems (Brown et al., 2020; Min et al., 2025). They nonetheless remain susceptible to *hallucination*, generating fluent and plausible responses that are factually incorrect or unsupported by evidence (Huang et al., 2025a; Bang et al., 2025). Detection is difficult because hallucinated responses are often indistinguishable from correct ones at the surface level, remaining coherent and linguistically well formed even when their content is incorrect (Wang et al., 2024; Singh & Singh, 2025). Large reasoning models compound the problem by producing logically structured reasoning traces that appear sound yet reach an incorrect conclusion (Mei et al., 2026; Yao et al., 2025).

Existing hallucination detection methods are broadly either output-based (Farquhar et al., 2024; Manakul et al., 2023; Zhang et al., 2023; Shelmanov et al., 2025) or internal-representation-based (Duan et al., 2024; Hu et al., 2026; Bar-Shalom et al., 2026; Zeng et al., 2026). The former often exploit uncertainty or consistency across multiple generations, using signals such as lexical agreement (Lin et al., 2022b; Manakul et al., 2023) or semantic entropy (Farquhar et al., 2024), and therefore do not use the internal representations that give rise to those outputs. The latter treat hidden states as a richer source of hallucination-related information, e.g., Hu et al. (2026) projects hidden representations onto subspaces derived from the model's unembedding matrix, and Bar-Shalom et al. (2026) models representations across layers and tokens using a Vision Transformer (ViT)-inspired architecture. Both train supervised detectors on hallucination labels.

Despite their promise, hidden-state representations present several challenges for hallucination detection. First, they are extremely high-dimensional, spanning multiple Transformer layers, response tokens, and hidden features. Existing methods therefore reduce these representations before classification, e.g., by selecting particular hidden states (Hu et al., 2026; Zeng et al., 2026) or by applying fixed-resolution pooling (Bar-Shalom et al., 2026). Such reductions make detection tractable but

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.

may discard complementary information distributed across layers, tokens, and features. Directly learning from the full hidden-state tensor avoids these reductions but results in a substantially higher-dimensional learning problem, increasing computational cost and the risk of overfitting under limited supervision. Second, response lengths vary, so the number of hidden states along the token mode differs across generations, making a common fixed-size representation nontrivial. Third, hidden-state representations are model-specific. LLMs differ in both the number of layers and hidden dimensions, making detectors defined directly in the original hidden space architecture-dependent and difficult to transfer across LLMs.

Our contributions. We address these challenges with a framework based on a low-rank multilinear formulation of LLM hidden representations. Our key contributions are:

- ❏ *Structured low-rank hidden-state representation.* We introduce **LOTUS** (**L**ow-rank **O**rder-statistic **TU**cker-based **S**ubspace learning) for LLM hallucination detection, which converts variable-length token representations into complementary layer-by-feature summaries and applies an unsupervised Tucker-2 projection (Tucker, 1966; Kolda & Bader, 2009) over the layer and feature modes, preserving the multilinear structure of the hidden states.
- ❏ *Compact and efficient detection pipeline.* **LOTUS** learns its multilinear representation from unlabeled hidden states, with hallucination labels used only by a lightweight classifier. The low-rank decomposition reduces representation size and memory cost by mapping model-specific hidden states into fixed-dimensional cores, which also facilitates detector reuse across LLMs with different internal dimensions.
- ❏ *Strong empirical performance with comprehensive evaluation.* We evaluate **LOTUS** across two LLMs and four question-answering benchmarks, where it delivers strong and consistent detection performance, and report ablations on the statistical summaries, Tucker ranks, projection design, performance under limited labeled data, and cross-model transfer.

2 Problem Statement

Let $\mathcal{M}$ be a decoder-only language model with L transformer layers, each with hidden dimension D. Consider prompts $\{\boldsymbol{x}_i\}_{i=1}^N$, each associated with a reference answer $\boldsymbol{a}_i^*$, and for each prompt sample K responses $\mathcal{A}_i = \{\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,K}\}$ from $\mathcal{M}$. Let $s(\boldsymbol{a}_{i,k}, \boldsymbol{a}_i^*)$ denote the similarity between a generated answer and the reference answer. Given a similarity threshold λ, the hallucination label of the k-th response is

$$z_{i,k} = \mathbb{I}[s(\boldsymbol{a}_{i,k}, \boldsymbol{a}_i^*) < \lambda],$$

where $\mathbb{I}[\cdot]$ is the indicator function, so that $z_{i,k} = 1$ denotes a hallucinated response and $z_{i,k} = 0$ a correct one. The goal of hallucination detection is to learn a scoring function $c(\boldsymbol{a}_{i,k})$ that assigns larger values to hallucinated responses, from which a binary prediction follows as

$$g_\tau(\boldsymbol{a}_{i,k}) = \mathbb{I}[c(\boldsymbol{a}_{i,k}) \geq \tau],$$

where τ is a decision threshold, with $g_\tau(\boldsymbol{a}_{i,k}) = 1$ denoting a hallucinated response and $g_\tau(\boldsymbol{a}_{i,k}) = 0$ a non-hallucinated one.

Hidden-state tensor. Recent studies provide growing evidence that LLM internal representations encode cues relevant to hallucination detection across Transformer layers and token positions (Azaria & Mitchell, 2023; Chen et al., 2024; Orgad et al., 2025; Zhang et al., 2025a; Bar-Shalom et al., 2026). Since the location and strength of these cues vary across responses and datasets, detectors that rely on a few selected activations can miss useful information, which motivates modeling hidden representations jointly across layers, tokens, and feature dimensions.

Accordingly, we consider the internal representations generated by $\mathcal{M}$ for each response. For the k-th response to prompt $\boldsymbol{x}_i$, let $\boldsymbol{h}_{i,k,t}^{(\ell)} \in \mathbb{R}^D$ denote the hidden representation of its t-th response token after layer ℓ, with $\ell \in \{1, \ldots, L\}$ and $t \in \{1, \ldots, T_{i,k}\}$, where $T_{i,k}$ is the token length of that response. Since $\mathcal{M}$ is autoregressive, $\boldsymbol{h}_{i,k,t}^{(\ell)}$ is computed conditioned on the prompt $\boldsymbol{x}_i$ and the preceding generated tokens $\boldsymbol{a}_{i,k,<t}$. Collecting these representations across layers and answer tokens yields the *hidden-state tensor* given by

$$\underline{\boldsymbol{H}}_{i,k} \in \mathbb{R}^{L \times T_{i,k} \times D},$$

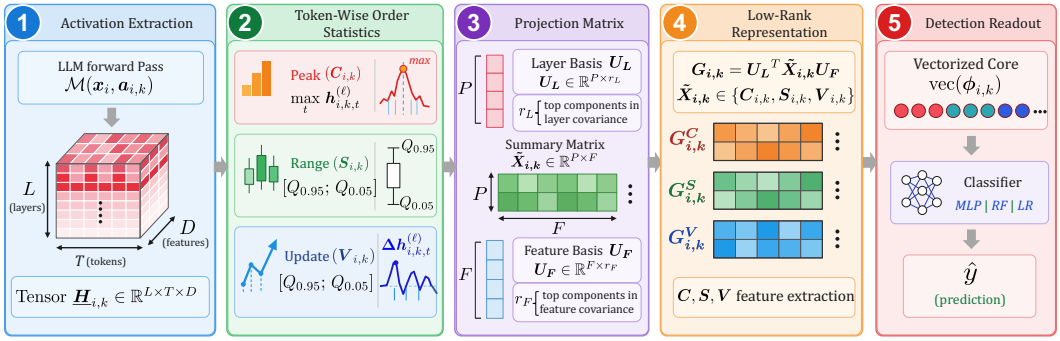


Figure 1: Overview of our proposed framework **LOTUS**. Hidden states are summarized across tokens using peak activations, activation quantiles, and cross-layer differences, projected with summary-specific Tucker-2 bases learned from unlabeled activations, and concatenated for lightweight hallucination classification.

with $\underline{\boldsymbol{H}}_{i,k}[\ell, t, :] = \boldsymbol{h}^{(\ell)}_{i,k,t}$, whose entries preserve three complementary modes of variation: Transformer layer, token position, and hidden feature dimension. Using these representations for hallucination detection poses three practical challenges: *(i)* the hidden-state tensor is extremely high-dimensional, so direct processing is memory- and computation-intensive and requires substantially more labeled data to avoid overfitting; *(ii)* the token dimension $T_{i,k}$ varies across responses, requiring a fixed-size representation without token alignment; and *(iii)* L and D vary across LLMs, so detectors defined directly in the original hidden space do not transfer readily across models. This raises a central research question:

> *How can we extract compact, informative hidden-state representations that enable computationally efficient and label-efficient hallucination detection while facilitating transfer across LLMs?*

3 Proposed Approach

Our approach is motivated by the hypothesis that hallucination-discriminative information can be captured by compact, low-rank multilinear representations of LLM hidden states. Rather than learning a detector directly from the high-dimensional hidden-state tensor, we propose a two-stage framework that separates representation learning from hallucination classification. The first stage constructs fixed-size statistical summaries of hidden states and learns label-free Tucker-2 projections over the layer and feature modes to obtain compact response representations. The second stage concatenates the resulting low-dimensional representations and trains a lightweight hallucination classifier. Fig. 1 gives the complete overview of **LOTUS**, and we describe each component below.

Order statistics over the token mode. The hidden-state tensor $\underline{\boldsymbol{H}}_{i,k} \in \mathbb{R}^{L\times T_{i,k}\times D}$ varies in size along the token mode, and hallucination-related activations may appear at different token positions. Hence, representations taken from predefined positions may overlook informative signals. We therefore summarize the token dimension using complementary statistical descriptors that produce fixed-size representations while retaining information about activation magnitude, distribution, and cross-layer changes. For each response $\boldsymbol{a}_{i,k}$ and each selected layer ℓ, we construct three descriptor matrices:

$$\boldsymbol{C}_{i,k}[\ell, :] = \left(\max_t \{\boldsymbol{h}^{(\ell)}_{i,k,t}\}\right)^\top, \quad \textbf{(Peak activation matrix)}$$

$$\boldsymbol{S}_{i,k}[\ell, :] = \left(\mathrm{Q}_{0.95}(\{\boldsymbol{h}^{(\ell)}_{i,k,t}\}_t)^\top,\ \mathrm{Q}_{0.05}(\{\boldsymbol{h}^{(\ell)}_{i,k,t}\}_t)^\top\right), \quad \textbf{(Activation quantile matrix)}$$

$$\boldsymbol{V}_{i,k}[\ell, :] = \left(\mathrm{Q}_{0.95}(\{\Delta\boldsymbol{h}^{(\ell)}_{i,k,t}\}_t)^\top,\ \mathrm{Q}_{0.05}(\{\Delta\boldsymbol{h}^{(\ell)}_{i,k,t}\}_t)^\top\right), \quad \textbf{(Cross-layer difference matrix)}$$

where $\max_t \{\boldsymbol{h}_t\} \in \mathbb{R}^D$ and $Q_\alpha(\{\boldsymbol{h}_t\}_t) \in \mathbb{R}^D$ denote the element-wise maximum and the element-wise empirical α-quantile of the hidden-state vectors over the token dimension, and the cross-layer difference $\Delta\boldsymbol{h}^{(\ell)}_{i,k,t} = \boldsymbol{h}^{(\ell+1)}_{i,k,t} - \boldsymbol{h}^{(\ell)}_{i,k,t}$ is the change in the hidden states between consecutive layers.

The three summaries capture complementary properties of hidden representations. The peak activation matrix captures localized activation peaks irrespective of their positions, since hallucination-related patterns may concentrate in a few tokens whose positions vary across responses (Orgad et al., 2025; Bar-Shalom et al., 2026). Because a maximum depends on a single token, the activation quantile matrix characterizes how feature activations are distributed across the response, distinguishing isolated peaks from sustained patterns. The cross-layer difference matrix summarizes how representations evolve as successive Transformer layers process them (Zhang et al., 2025b). Together, these position-independent summaries turn variable-length token representations into fixed-size matrices without token alignment.

Unsupervised Tucker-2 projection over layer and feature modes. The token summaries are of fixed-size but remain high-dimensional, particularly along the feature mode. To extract compact representations that capture dependencies across layers and hidden features, we employ a *Tucker-2 decomposition* (Tucker, 1966; Kolda & Bader, 2009), which learns a compact, separable projection comprising a layer basis and a feature basis, shared across all training responses and learned without any hallucination labels. Projecting each summary matrix onto these subspaces yields a low-dimensional core matrix that retains interactions between the layer and feature dimensions.

Let $\boldsymbol{X}_{i,k} \in \mathbb{R}^{P\times F}$ denote any one of the summary matrices $\boldsymbol{C}_{i,k}$, $\boldsymbol{S}_{i,k}$, or $\boldsymbol{V}_{i,k}$, where P is the layer-mode dimension and $F \in \{D, 2D\}$ the feature-mode dimension. The layer dimension P need not equal the total number of layers L—it can be set smaller to reduce computational cost, and the choice of which layers to retain can incorporate prior knowledge (see more details about the layer selection in §4.1). Before decomposition, we robustly scale each summary $\boldsymbol{X}_{i,k}$ with scaling parameters estimated from the training responses only, which prevents a small number of large-magnitude activations from dominating the projection, and denote the scaled summary by $\tilde{\boldsymbol{X}}_{i,k}$ (more details about this scaling is provided in Appendix D). From these scaled hidden representations $\tilde{\boldsymbol{X}}_{i,k}$, we seek *shared orthonormal bases* $\boldsymbol{U}_L \in \mathbb{R}^{P\times r_L}$ and $\boldsymbol{U}_F \in \mathbb{R}^{F\times r_F}$, with $r_L < P$ and $r_F < F$, that maximize the energy retained after projection onto the corresponding low-dimensional subspaces:

$$\max_{\boldsymbol{U}_L,\boldsymbol{U}_F} \quad \sum_{i=1}^{N}\sum_{k=1}^{K} \left\| \boldsymbol{U}_L^\top \tilde{\boldsymbol{X}}_{i,k} \boldsymbol{U}_F \right\|_F^2 \qquad \text{s.t.} \quad \boldsymbol{U}_L^\top \boldsymbol{U}_L = \boldsymbol{I}, \qquad \boldsymbol{U}_F^\top \boldsymbol{U}_F = \boldsymbol{I}. \tag{1}$$

where $\boldsymbol{I}$ is an identity matrix of appropriate size. This is the Tucker-2 decomposition of the third-order tensor formed by stacking the training summaries along a response mode (Tucker, 1966; Kolda & Bader, 2009). It compresses only the layer and feature modes, preserving a separate low-dimensional representation for each response.

We handle the optimization problem in equation 1 with truncated higher-order singular value decomposition (HOSVD) (De Lathauwer et al., 2000). Specifically, the layer basis $\boldsymbol{U}_L$ consists of the r_L leading eigenvectors of $\sum_{i=1}^{N}\sum_{k=1}^{K} \tilde{\boldsymbol{X}}_{i,k}\tilde{\boldsymbol{X}}_{i,k}^\top$, while the feature basis $\boldsymbol{U}_F$ is obtained from the r_F leading eigenvectors of $\sum_{i=1}^{N}\sum_{k=1}^{K} \tilde{\boldsymbol{X}}_{i,k}^\top\tilde{\boldsymbol{X}}_{i,k}$, We compute the feature eigenspace with randomized SVD on the vertically stacked rows of all training summaries (Halko et al., 2011) rather than forming the $F \times F$ Gram matrix, which reaches $F = 8192$ for LLaMA-3.1-8B. Once the bases are estimated, each response is mapped to a low-dimensional core matrix defined as

$$\boldsymbol{G}_{i,k} = \boldsymbol{U}_L^\top \tilde{\boldsymbol{X}}_{i,k} \boldsymbol{U}_F \in \mathbb{R}^{r_L\times r_F}, \tag{2}$$

where $\boldsymbol{G}_{i,k}$ contains $r_L r_F$ coefficients that represent interactions between the layer and feature directions. In this way, we learn a separate core representations for each summary type $\tilde{\boldsymbol{X}}_{i,k}$.

Remark 1 (Projection Parameter Efficiency) *For the same $r_L r_F$-dimensional output, Tucker-2 represents the projection using separate layer and feature bases, requiring only $Pr_L + Fr_F$ projection coefficients. Applying principal component analysis (PCA) to the vectorized summaries instead learns a single projection on the joint layer–feature space and requires $PFr_L r_F$ coefficients. Across the three summary types, this corresponds to approximately 1.15M versus 45.6M coefficients for Qwen2.5-7B-Instruct and 1.31M versus 52.2M for LLaMA-3.1-8B. The smaller parameterization also lowers the cost of storing the projection.*

Hallucination detector from low-dimensional representations. To learn the hallucination detector, we vectorize and concatenate the low-dimensional core representations $\boldsymbol{G}_{i,k}^C$, $\boldsymbol{G}_{i,k}^S$, and $\boldsymbol{G}_{i,k}^V$ of

$\boldsymbol{a}_{i,k}$ into the *response feature vector*:

$$\boldsymbol{\phi}_{i,k} = \left[\mathrm{vec}(\boldsymbol{G}^C_{i,k}), \mathrm{vec}(\boldsymbol{G}^S_{i,k}), \mathrm{vec}(\boldsymbol{G}^V_{i,k})\right]^\top \in \mathbb{R}^{d_\phi}, \tag{3}$$

where $d_\phi = r^C_L r^C_F + r^S_L r^S_F + r^V_L r^V_F$ is the dimension of the resulting feature vector. We then train a lightweight classifier $f_{\boldsymbol{\theta}}$, with parameters $\boldsymbol{\theta}$, on the feature vectors of the training responses and their hallucination labels $z_{i,k}$. At inference, the statistical summaries of an unseen response $\boldsymbol{a}_{\mathrm{test}}$ are computed from its hidden-state tensor, transformed with the learned projection bases using equation 2, and assembled as in equation 3 into $\boldsymbol{\phi}_{\mathrm{test}}$. The classifier returns the hallucination score $c(\boldsymbol{a}_{\mathrm{test}}) = f_{\boldsymbol{\theta}}(\boldsymbol{\phi}_{\mathrm{test}})$, which can be thresholded at τ to obtain the binary decision g_τ.

4 Experiments

4.1 Evaluation Setup

Models and datasets. We evaluate on two open-weight models of comparable scale: Qwen2.5-7B-Instruct, which is instruction-tuned, and the base LLaMA-3.1-8B. We use four question answering datasets: TruthfulQA (Lin et al., 2022a), TriviaQA (Joshi et al., 2017) without evidence documents, NQ-Open (Kwiatkowski et al., 2019), and the English questions of TyDiQA-GP (Clark et al., 2020), in which each question comes with a passage containing its answer. For each question we generate $K = 10$ responses by beam sampling with 10 beams, returning 10 sequences each capped at 64 new tokens. The datasets contain 817, 9,960, 3,610, and 440 questions, respectively. A question is *model-known* if at least one of its K responses is labeled correct, and *model-unknown* otherwise. Only model-known questions are eligible for detector training; their number varies substantially across models and datasets, and Qwen2.5-7B-Instruct has notably fewer on NQ-Open, which contributes to the larger variance in that setting. Appendix D provides the exact counts, prompts, generation parameters, and reproducibility details.

Hallucination labels. We label responses following the procedure in (Hu et al., 2026). A response $\boldsymbol{a}_{i,k}$ is compared with the reference answers of question i and labeled correct ($z_{i,k} = 0$) if its ROUGE-L score (Lin, 2004), averaged over the references, is at least 0.7, or if its BLEURT-20 score (Sellam et al., 2020) exceeds 0.5 for at least one reference; otherwise it is labeled hallucinated ($z_{i,k} = 1$). All methods are evaluated on the same generated responses and labels. Appendix P compares these labels with those of the judge released by Hu et al. (2026).

Train–test splitting protocol. Since each question is associated with multiple generated responses, the set of question-response pairs can be divided by question or by response. We use the *question-level* train-test split throughout, where we randomly assign 75% of the model-known questions to training and the remaining 25% to testing, keeping all responses to each question in the same split. The test set additionally includes all responses to model-unknown questions. Consequently, no question appears in both the training and test sets. The *answer-level* alternative pools responses instead, which puts responses to the same question on both train and test splits and evaluates a detector on questions it was trained on (e.g., HARP Hu et al. (2026) follows this protocol in their released implementation). Appendix E reports both splits and demonstrates how question overlap can inflate reported hallucination detection performance. The choice affects only detectors that score individual responses, since one producing a single score per question is unaffected.

Evaluation metrics. We use pooled AUROC as our primary metric, following standard practice in hallucination detection (Hu et al., 2026; Du et al., 2024; Bar-Shalom et al., 2026). Pooled AUROC compares responses across different questions, so it can reward a detector for identifying questions that are generally more prone to hallucination rather than for separating correct and hallucinated responses to the same question. We therefore also report *within-question AUROC*, which restricts the pairwise comparison to responses of the same question and so measures response-level discrimination while controlling for question-specific effects. A detector that assigns an identical score to every response of a question obtains 50%, however high its pooled AUROC. It is closely related to the within-question discrimination of Taubenfeld et al. (2025). Question- and dataset-level shortcuts in hallucination detection have likewise been reported (Dubanowska et al., 2025; Seo et al., 2025). Within-question AUROC is defined only for questions containing at least one correct and one hallucinated response—more details in Appendix N.

LOTUS implementation details. LOTUS uses a fixed window of nine consecutive Transformer blocks, 15–23, for both models and all datasets, motivated by prior evidence that hallucination- and

factuality-related information is prominent in intermediate Transformer representations (Azaria & Mitchell, 2023; Orgad et al., 2025; Zhang et al., 2025b). We use ranks $(r_L, r_F) = (5, 64)$ for $\boldsymbol{C}$ and $\boldsymbol{S}$, and $(4, 64)$ for $\boldsymbol{V}$, producing cores of dimensions 5×64, 5×64, and 4×64. After vectorization and concatenation, each response is represented by an 896-dimensional feature vector. We then train a random-forest classifier with 200 trees and balanced class weights on these features and use its predicted hallucination probability for AUROC evaluation. Layer window, Tucker ranks, and classifier settings are fixed across models and datasets rather than tuned per benchmark. Appendix C evaluates alternative windows of the same width and Appendix I the rank choices.

Baselines. We compare **LOTUS** against eight different hallucination detection baselines.

Label-free baselines. We consider six methods that require no hallucination-labeled training responses: answer-level *Perplexity*; LN-Entropy (Malinin & Gales, 2021), which aggregates predictive uncertainty across generated responses; Semantic Entropy (Kuhn et al., 2023), over semantically equivalent answer clusters; Lexical Similarity (Lin et al., 2022b), based on pairwise agreement among sampled responses; EigenScore (Chen et al., 2024), based on the eigenspectrum of hidden-state representations; and HaloScope (Du et al., 2024), which estimates a hallucination-related subspace from unlabeled generations.

Supervised baselines. We additionally compare with two recent hidden-representation-based detectors trained on hallucination labels. HARP (Hu et al., 2026) projects final-layer token representations onto a low-dimensional subspace derived from the model's unembedding matrix and applies a lightweight MLP to obtain token-level hallucination scores. ACT-ViT (Bar-Shalom et al., 2026) operates on layer–token activation tensors, reducing them to a fixed resolution and training a ViT-based classifier for response-level detection. Both are run with their released implementations and training settings; further implementation details are given in Appendix D.

4.2 Experiment Results

Comparison with baselines. Table 1 reports pooled AUROC under the question-level split across two LLMs and four QA datasets. LOTUS outperforms HARP in all eight model–dataset combinations, by 2.15 to 11.56 AUROC points and by 7.10 points on average, with the largest gains on NQ-Open and TruthfulQA. Compared with ACT-ViT, LOTUS achieves the higher AUROC in six of the eight settings: by 0.18, 2.43, and 5.04 points on NQ-Open, TruthfulQA, and TyDiQA-GP with Qwen2.5-7B-Instruct, and by 2.38 and 3.06 points on NQ-Open and TyDiQA-GP with LLaMA-3.1-8B, where the two are effectively tied on TruthfulQA (89.19 vs. 89.07). ACT-ViT is ahead on TriviaQA, by 1.08 and 3.11 points. TriviaQA is the largest of the four datasets by an order of magnitude, at $99{,}600$ responses per model, and ACT-ViT retains a layer–token activation tensor per response: on TriviaQA it stores 106–122 GB against 28.6–32.7 GB for LOTUS at the token resolution we report, and about nineteen times the LOTUS figure at its released resolution (Appendix K). The remaining margin on that dataset is therefore obtained at a considerably higher representation cost. LOTUS also outperforms every label-free baseline in all eight settings. Overall, the proposed low-rank multilinear representation gives strong and consistent detection performance across models and datasets at a small fraction of the representation size of the strongest supervised baseline.

Table 2 reports within-question AUROC under the question-level split. Label-free baselines that assign a single score per question attain 50% by construction and are omitted. LOTUS outperforms HARP in all four settings by 5.40–14.95 points and exceeds ACT-ViT in three of four, trailing by 0.77 points on TruthfulQA with LLaMA-3.1-8B. The larger gap between pooled and within-question AUROC for HARP on TruthfulQA indicates that more of its pooled performance is attributable to differences between questions, whereas LOTUS retains stronger response-level discrimination. Standard deviations are larger than for pooled AUROC because the metric uses only questions containing both correct and hallucinated responses; Appendix N reports the number of eligible pairs.

Different configurations of statistical summaries. Table 3 evaluates how each token-mode summary defined in § 3 contributes to LOTUS, which compresses and concatenates the peak activation ($\boldsymbol{C}$), activation quantile ($\boldsymbol{S}$), and cross-layer difference ($\boldsymbol{V}$) summaries independently (see equation 3). No single configuration is optimal in every model–dataset setting, but LOTUS is the most consistent. It is the best configuration in two of the eight settings and stays within 0.74 AUROC points of the best alternative in the others. Removing individual summaries generally reduces performance, so activation magnitude, distribution across tokens, and cross-layer variation carry com-

Table 1: Pooled AUROC (%) under the question-level split across two LLMs and four QA benchmarks. Results are averaged over five random trials; mean ± standard deviation is reported for supervised detectors. Best results are in bold.

Method	NQ-Open	TruthfulQA	TriviaQA	TyDiQA-GP
Qwen2.5-7B-Instruct				
Perplexity	57.58	65.36	75.42	55.72
LN-Entropy	63.24	59.61	73.96	68.60
Semantic Entropy	61.91	60.86	63.49	51.60
Lexical Similarity	65.95	58.44	71.09	72.86
EigenScore	62.14	56.24	70.84	63.48
HaloScope	60.49	63.62	74.19	66.75
HARP	67.73 ± 2.24	77.54 ± 1.51	88.75 ± 0.33	79.90 ± 2.87
ACT-ViT	79.11 ± 5.80	85.40 ± 1.01	$\mathbf{92.77 \pm 0.28}$	82.82 ± 1.46
LOTUS (Ours)	$\mathbf{79.29 \pm 5.24}$	$\mathbf{87.83 \pm 0.89}$	91.69 ± 0.10	$\mathbf{87.86 \pm 1.95}$
LLaMA-3.1-8B				
Perplexity	41.37	68.21	51.13	59.30
LN-Entropy	66.25	65.43	54.82	57.18
Semantic Entropy	54.21	57.09	63.43	62.81
Lexical Similarity	60.56	58.73	60.72	55.15
EigenScore	71.34	47.60	62.24	57.02
HaloScope	61.74	69.37	71.35	51.52
HARP	80.56 ± 1.06	79.58 ± 5.38	82.54 ± 0.49	74.99 ± 0.88
ACT-ViT	82.98 ± 1.76	89.07 ± 1.30	$\mathbf{87.80 \pm 0.33}$	79.43 ± 1.41
LOTUS (Ours)	$\mathbf{85.36 \pm 0.42}$	$\mathbf{89.19 \pm 1.41}$	84.69 ± 0.20	$\mathbf{82.49 \pm 1.23}$

Table 2: Within-question AUROC (%) under the question-level split. Results are reported as mean ± standard deviation over five random trials.

	Qwen2.5-7B-Instruct		LLaMA-3.1-8B	
Method	**TruthfulQA**	**TyDiQA-GP**	**TruthfulQA**	**TyDiQA-GP**
HARP	67.84 ± 2.07	82.69 ± 1.44	67.78 ± 8.05	75.87 ± 1.63
ACT-ViT	78.51 ± 4.51	81.93 ± 1.93	83.50 ± 3.12	80.50 ± 2.34
LOTUS (Ours)	$\mathbf{82.07 \pm 2.18}$	$\mathbf{88.09 \pm 3.77}$	$\mathbf{82.73 \pm 1.92}$	$\mathbf{81.34 \pm 3.01}$

plementary signals. Projecting $\boldsymbol{S}$ and $\boldsymbol{V}$ with one shared pair of bases underperforms the default in seven of eight settings, supporting separate bases per summary type.

Cross-model transfer. LOTUS produces representations of the same dimensionality across LLMs, but in model-specific coordinate systems, since the Tucker bases are learned independently from each model's activations. Therefore, a detector trained on one model cannot directly interpret another model's LOTUS features even when their dimensions match. To enable transfer, we learn a lightweight ridge mapping from the target model's representation to the source model's on paired, unlabeled responses, and apply the source detector to the aligned target representations without any target hallucination labels (see more details of the implementation in Appendix J). Table 4 shows that the aligned detector transfers in both directions. Across the four model–dataset settings, its pooled AUROC stays within 1.78–6.03 points of a detector trained directly on the target model, whereas removing the alignment causes substantial degradation, particularly on TyDiQA-GP.

Spectral evidence for low-rank hallucination structure. LOTUS relies on a low-dimensional feature basis, raising a natural question: *how much of the hidden-state variation is concentrated in a small number of directions?* We analyze the peak-activation summary $\boldsymbol{C}_{i,k}$ one layer at a time. At layer ℓ, each response is represented by $\boldsymbol{C}_{i,k}[\ell, :] \in \mathbb{R}^D$, and separate covariance matrices are formed for correct and hallucinated responses. To control for class imbalance, *size-matched* sampling uses an equal number of responses from each class; *question-paired* sampling additionally controls for question identity by selecting one correct and one hallucinated response from each question containing both outcomes. Each scheme is repeated over 20 draws. We characterize each

Table 3: The full LOTUS configuration independently projects and concatenates the peak activation ($\boldsymbol{C}$), activation quantile ($\boldsymbol{S}$), and cross-layer difference ($\boldsymbol{V}$) summaries. The LOTUS row reports absolute pooled AUROC (%), while all ablated variants report the change in AUROC relative to LOTUS (Δ). Positive values indicate improvement over LOTUS and negative values indicate degradation. Results use the question-level split and are averaged over five random draws.

		AUROC / Δ vs. LOTUS			
Configuration	**Summaries**	**NQ-Open**	**TruthfulQA**	**TriviaQA**	**TyDiQA-GP**
Qwen2.5-7B-Instruct					
LOTUS (full)	$\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{V}$	**79.29**	**87.83**	**91.69**	**87.86**
Ablated variants					
w/o $\boldsymbol{S}$	$\boldsymbol{C}, \boldsymbol{V}$	$+0.15$	-0.04	-0.23	-0.06
Peak only	$\boldsymbol{C}$	-0.48	-0.75	-0.41	$+0.40$
Quantile only	$\boldsymbol{S}$	-0.23	-0.75	-0.02	-0.38
Difference only	$\boldsymbol{V}$	-1.09	-0.11	-0.77	-0.85
Shared bases	$\boldsymbol{S}, \boldsymbol{V}$	-0.84	-0.64	$+0.10$	-0.91
LLaMA-3.1-8B					
LOTUS (full)	$\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{V}$	**85.36**	**89.19**	**84.69**	**82.49**
Ablated variants					
w/o $\boldsymbol{S}$	$\boldsymbol{C}, \boldsymbol{V}$	-0.18	$+0.13$	-0.63	-1.04
Peak only	$\boldsymbol{C}$	-0.30	-0.60	-0.49	-0.82
Quantile only	$\boldsymbol{S}$	-0.05	-0.46	$+0.10$	$+0.74$
Difference only	$\boldsymbol{V}$	-1.11	$+0.08$	-2.28	-2.49
Shared bases	$\boldsymbol{S}, \boldsymbol{V}$	-0.37	-0.30	-1.03	-0.80

Table 4: AUROC (%) on the target model under the question-level split. *Target-trained* uses target labels, while transferred detectors reuse the source classifier with or without unlabeled ridge alignment. Pooled AUROC is mean $\pm$ standard deviation over five draws; within-question AUROC is also reported.

	TruthfulQA		**TyDiQA-GP**	
Method	**Pooled**	**Within**	**Pooled**	**Within**
Qwen2.5-7B-Instruct $\rightarrow$ *LLaMA-3.1-8B*				
Target-trained LOTUS	89.22 ± 1.44	82.73	82.47 ± 1.24	81.34
Transfer + alignment	$\mathbf{87.44 \pm 1.18}$	**80.67**	$\mathbf{77.26 \pm 2.32}$	**78.44**
Transfer w/o alignment	63.73 ± 4.93	66.10	48.79 ± 2.31	50.60
LLaMA-3.1-8B $\rightarrow$ *Qwen2.5-7B-Instruct*				
Target-trained LOTUS	87.76 ± 0.87	82.07	87.82 ± 1.96	88.09
Transfer + alignment	$\mathbf{82.94 \pm 1.26}$	**78.18**	$\mathbf{81.79 \pm 2.63}$	**85.05**
Transfer w/o alignment	58.05 ± 5.30	59.86	44.36 ± 6.05	43.42

covariance spectrum by the variance captured by its leading 64 eigen-directions and by the effective rank $r_{\text{eff}} = \exp(-\sum_j p_j \log p_j)$, where $p_j = \lambda_j / \sum_m \lambda_m$.

Across the nine LOTUS layers, the leading 64 of $D = 3584$ directions explain 64.7–66.4% of the variance of correct responses and 52.3–55.5% for hallucinated responses, indicating substantial low-dimensional concentration. Hallucinated responses nevertheless exhibit broader spectra: their effective rank is 2.01–2.22$\times$ that of correct responses under size-matched sampling and 1.52–1.60$\times$ under question-paired sampling (Fig. 2a,b). This ordering persists across all layers, suggesting that hallucinated responses distribute their variance over a broader set of feature directions even after controlling for question identity. Spectral separation alone, however, does not determine the most discriminative layers. The embedding representation has the largest rank ratio but only 76.5% single-layer AUROC, whereas blocks 15–23 average 86.8% (Fig. 2c). We therefore use this analysis to motivate low-rank compression while selecting the layer window separately.

Performance with limited labeled data. We further examine whether the compact Tucker-2 representation helps when fewer labeled training questions are available. Table 11 in Appendix G reports the AUROC difference between Tucker-2 and an equal-dimensional flat-PCA representation as the training set is reduced to 5%–50%, with the test set held fixed. With the full training set, the two representations perform similarly, within 0.80 AUROC points in all settings. As the training data

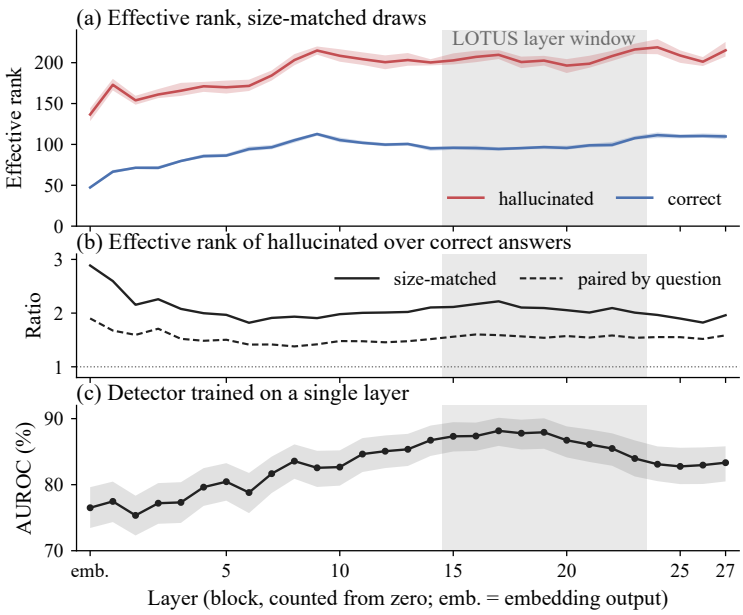


Figure 2: Covariance spectra of the peak activation matrix for Qwen2.5-7B-Instruct on TyDiQA-GP, at every layer. (a) effective rank of correct and hallucinated responses, (b) hallucinated-to-correct effective-rank ratio under size-matched and question-paired sampling, and (c) single-layer AUROC. Shaded bands indicate 95% intervals; the gray region marks the LOTUS layer window.

decrease, Tucker-2 gains a clear advantage in several settings: on LLaMA-3.1-8B with TruthfulQA it improves AUROC by 2.66–3.98 points in every reduced-data regime, and on Qwen2.5-7B-Instruct with TyDiQA-GP the gain reaches 5.06 points. The remaining settings are mixed, so the benefit is not universal, but the more compact, structured Tucker-2 parameterization is advantageous when labeled training data are scarce.

Tucker-3 token-mode extension. LOTUS aggregates the token mode before decomposition, trading explicit token-order information for a compact layer–feature representation. To examine this trade-off, we construct a Tucker-3 variant that partitions each response into B contiguous token bins and additionally decomposes the resulting token mode with rank r_T. Table 10 in Appendix F shows that retaining this coarse temporal structure provides only limited benefit. With the proposed order-statistic summaries, no Tucker-3 configuration improves pooled AUROC over the Tucker-2 baseline by more than 0.37 points across the four settings, while increasing the representation from 896 to as many as 7,168 features, and increasing the token rank gives no consistent gain. The same trend holds with average pooling, supporting the choice to aggregate the token mode before decomposition, i.e., the Tucker-2 representation is substantially more compact at comparable detection performance.

5 Conclusion

In this work, we introduced **LOTUS**, a low-rank multilinear framework for hallucination detection from LLM hidden representations. **LOTUS** converts variable-length token representations into complementary layer–feature summaries and learns compact Tucker-2 subspaces from unlabeled activations, requiring hallucination labels only for a lightweight downstream classifier. Across two LLMs and four question-answering benchmarks, our method achieves strong and consistent detection performance while substantially reducing representation size and supporting transfer across models with different internal dimensions. Extensive ablations further validate the contributions of the proposed summaries, low-rank projection, rank choices, and transfer design. Overall, our results show that structured low-dimensional representations provide an effective and efficient basis for detecting hallucinations from LLM internal states.

AI USE STATEMENT

We used a generative AI assistant in two ways. First, for code: most of the pipeline, job and analysis scripts in the released repository were written or revised with AI assistance under the authors' direction. Every number we report is produced by that code, the scripts carry self-tests that check them against independent implementations where one exists (the ACT-ViT preprocessing, for instance, is asserted bit-identical to the released implementation of its authors), and the authors reviewed both the code and its outputs. Second, for writing: the assistant was used to draft and revise text throughout the paper, and to locate candidate related work, which the authors then verified against the original papers. The assistant was not used to generate experimental results, data or labels: all generations, activations, metrics and tables come from running the released code. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artifacts.

ETHICS STATEMENT

This work uses four public question answering benchmarks and two openly available language models, each under its own license. It involves no human subjects, no new data collection, and no personally identifying information beyond what those public datasets already contain. Hallucination labels are assigned by automatic metrics rather than by human annotation (§4.1), so they inherit the biases of those metrics; we therefore report their agreement with an independently released judge in Appendix P rather than treating them as ground truth. A detector of this kind reduces, but does not remove, the risk of acting on an incorrect model output, and accuracy on these benchmarks does not imply reliability in deployment, in other languages, or on other domains; we make no such claim. We are not aware of dual-use concerns raised by this work beyond those already present in the publicly released models it analyses.

REPRODUCIBILITY STATEMENT

The detector is specified in §3: the token summaries, the entry-wise scaling, the two Tucker-2 bases and the classifier. The configuration behind every reported number is given in §4.1, which states the layer window, the ranks, the classifier, the split and the two metrics, and Appendix D adds the generation settings, the prompts, the per-question seeding and how the two trained baselines were run. A single harness owns the splits, the seeds and the metrics so that no method computes its own, and Appendix E defines both split procedures. Code for generation, extraction, evaluation, the ablations and the analysis is available at `https://anonymous.4open.science/r/LOTUS-E298/`, one script per stage, each with a self-test. The generated responses and activations are not released because of their size; each question is seeded from a global seed and its identifier, so they can be regenerated exactly from the released configuration.

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, 2023.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, 2025.

Guy Bar-Shalom, Fabrizio Frasca, Yaniv Galron, Yftah Ziser, and Haggai Maron. Beyond token probes: Hallucination detection via activation tensors with ACT-ViT. *Advances in Neural Information Processing Systems*, 38:85531–85561, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

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*, 2024.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, and Feng Zheng. Enhancing uncertainty modeling with semantic graph for hallucination detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23586–23594, 2025.

Pei Cheng, Xiayang Shi, and Yinlin Li. Enhancing translation ability of large language models by leveraging task-related layers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6110–6121, 2024.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972, 2024.

Hanyu Duan, Yi Yang, and Kar Yan Tam. Do LLMs know about hallucination? an empirical investigation of LLM's hidden states. *arXiv preprint arXiv:2402.09733*, 2024.

Zuzanna Dubanowska, Maciej Żelaszczyk, Michał Brzozowski, Paolo Mandica, and Michał Karpowicz. Representation-based broad hallucination detectors fail to generalize out of distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17563–17575, 2025.

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

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *International Conference on Learning Representations*, 2026.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

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.

Yiming Huang, Junyan Zhang, Zihao Wang, Biquan Bie, Yunzhong Qiu, Xuming Hu, Yi R. Fung, and Xinlei He. RePPL: Recalibrating perplexity by uncertainty in semantic propagation and language generation for explainable QA hallucination detection. *arXiv preprint arXiv:2505.15386*, 2025b.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

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.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8235–8246, 2024.

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32310–32322, 2025.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51 (3):455–500, 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

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

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022a.

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

Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, Yubing Jian, Yao Qin, and Roland Memisevic. Enhancing hallucination detection through noise injection. In *International Conference on Learning Representations*, pp. 128245–128262, 2026.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

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, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Sho, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know? In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3408–3458, 2026.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. *Advances in Neural Information Processing Systems*, 37:61372–61418, 2024.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. Towards multi-dimensional evaluation of LLM summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14417–14450, 2025.

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

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, 2020.

Yeongbin Seo, Dongha Lee, and Jinyoung Yeo. Quantifying genuine awareness in hallucination prediction beyond question-side shortcuts. *arXiv preprint arXiv:2509.15339*, 2025.

Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, et al. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in llm outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35700–35719, 2025.

Damanjot Singh and Amritpal Singh. Ubiquity of LLM hallucinations across critical domains: a survey. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 115–132. Springer, 2025.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, July 2025.

Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33377–33385, 2026.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19519–19529, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1088. URL https://aclanthology.org/2024.emnlp-main.1088/.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

Xinyue Zeng, Junhong Lin, Yujun Yan, Feng Guo, Liang Shi, Jun Wu, and Dawei Zhou. Halluguard: Demystifying data-driven and reasoning-driven hallucinations in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZURs3YZclt.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. Detecting hallucination in large language models through deep internal representation analysis. In *IJCAI*, pp. 8357–8365, 2025a.

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

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, 2025b.