# Bolster Hallucination Detection via Prompt-Guided Data Augmentation

**Wenyun Li**[1,2], **Zheng Zhang**[1,2 *], **Dongmei Jiang**[2], **Xiangyuan Lan**[2,3*]

[1]Harbin Institute of Technology, Shenzhen, China
[2]Pengcheng Laboratory , Shenzhen, China
[3]Pazhou Laboratory (Huangpu), Guangzhou, China
liwy@pcl.ac.cn, darrenzz219@gmail.com, jiangdm@pcl.ac.cn, lanxy@pcl.ac.cn

## Abstract

Large language models (LLMs) have garnered significant interest in AI community. Despite their impressive generation capabilities, they have been found to produce misleading or fabricated information, a phenomenon known as hallucinations. Consequently, hallucination detection has become critical to ensure the reliability of LLM-generated content. One primary challenge in hallucination detection is the scarcity of well-labeled datasets containing both truthful and hallucinated outputs. To address this issue, we introduce **P**rompt-guided data **A**ugmented ha**L**lucination d**E**tection (PALE), a novel framework that leverages prompt-guided responses from LLMs as data augmentation for hallucination detection. This strategy can generate both truthful and hallucinated data under prompt guidance at a relatively low cost. To more effectively evaluate the truthfulness of the sparse intermediate embeddings produced by LLMs, we introduce an estimation metric called the Contrastive Mahalanobis Score (CM Score). This score is based on modeling the distributions of truthful and hallucinated data in the activation space. CM Score employs a matrix decomposition approach to more accurately capture the underlying structure of these distributions. Importantly, our framework does not require additional human annotations, offering strong generalizability and practicality for real-world applications. Extensive experiments demonstrate that PALE achieves superior hallucination detection performance, outperforming the competitive baseline by a significant margin of 6.55%.

**Extended version** — https://arxiv.org/abs/2510.15977

## Introduction

Recently, large language models (LLMs) have emerged as one of the most significant breakthroughs in the field of artificial intelligence (Hurst et al. 2024). LLMs have found widespread application across a variety of tasks, including logical reasoning (Dong et al. 2024), visual question answering (Jian, Yu, and Zhang 2024), and speech-to-text transcription (Zhang et al. 2023), often surpassing human performance in many scenarios. Due to their exceptional reasoning and generative capabilities, LLMs have been integrated into numerous high-trust systems, such as those in the medical domain (Li and Pun 2023; Zhou et al. 2025; Li et al. 2025). However, despite these impressive capabilities, even state-of-the-art LLMs frequently generate factually incorrect or nonsensical content—a phenomenon known as hallucination (Huang et al. 2025). This inherent risk makes LLM outputs potentially unreliable in mission-critical applications (Wang et al. 2024b). Therefore, a reliable LLM should not only produce text that is coherent with the given prompt but also possess the ability to detect hallucinations in its output.

This concern underscores the importance of hallucination detection technology, which determines whether a generated output is truthful or not (Zhang et al. 2024b; Sriramanan et al. 2024; Su et al. 2024). Most approaches to hallucination detection rely on devising uncertainty scoring functions. For instance, logit-based methods (Ren et al. 2023; Malinin and Gales 2021) employ token-level probabilities as uncertainty scores, consistency-based methods (Lin, Trivedi, and Sun 2024a; Manakul, Liusie, and Gales 2023) quantify uncertainty by evaluating the consistency across multiple responses, and verbalized methods (Lin, Trivedi, and Sun 2024a; Kadavath et al. 2022) prompt LLMs to express their uncertainty in natural language. Recently, there has been increasing work (Wei and Zhang 2024; Du, Xiao, and Li 2024; Arteaga, Schön, and Pielawski 2024) aimed at leveraging the internal states of LLMs to assess the veracity of their outputs. For example, contrast-consistent search (CCS) (Burns et al. 2022) trains a binary truthfulness classifier to satisfy logical consistency properties, and HaloScope (Du, Xiao, and Li 2024) seeks to identify hallucination subspaces. However, all these methods face the challenge of a scarcity of well-labeled datasets containing both truthful and hallucinated outputs, which significantly limits their performance in detecting hallucinations in real-world scenarios.

We argue that the primary challenge in hallucination detection is the lack of labeled datasets containing both truthful and hallucinated generations. In practice, creating a high-quality ground truth dataset for hallucination detection requires substantial human labor to annotate a large number of generated samples. However, collecting such large-scale data and labeling it is extremely costly. In particular, the vast landscape of generative models and the diverse range of content they produce further complicate the process. Moreover, maintaining the quality and consistency of labeled data for hallucination detection requires continuous annotation ef-

*Corresponding authors.

forts and robust quality control measures. These formidable obstacles underscore the need to reduce reliance on human-labeled data for hallucination detection.

To address this issue, we introduce **P**rompt-guided data **A**ugmented ha**L**lucination d**E**tection (PALE), a novel framework that leverages state-of-the-art LLMs for data augmentation using non-parametric prompts. The augmented data generated through LLM prompting can be acquired at a relatively low cost. A state-of-the-art language model, such as GPT-4o, contains a vast amount of hidden knowledge, enabling us to harness LLMs to produce diverse truthful and hallucinated data via prompt engineering.

From this perspective, we introduce a novel metric, the **Contrastive Mahalanobis Score**, to assess the truthfulness of augmented data. Our key idea is inspired by prior work (Du, Xiao, and Li 2024; Burns et al. 2022), which leverages the latent state of a language model as a representation that encodes information related to truthfulness. Specifically, PALE employs a matrix decomposition approach to model the distributions of truthful and hallucinated data in the embedding activation space. It then computes the distance of a test sample to both distributions. Both truthful and hallucinated distributions are modeled as Gaussian, with their means and covariance matrices estimated from the embedding activations. Notably, the covariance matrix computation is facilitated by factorizing the LLM embeddings. The Mahalanobis distance between a test sample and each distribution is used to infer its truthfulness. The resulting score is straightforward to implement in practical applications.

The primary contributions of our research are summarized as follows:

- Our proposed framework, **P**rompt-guided data **A**ugmented ha**L**lucination d**E**tection (PALE), harnesses augmented data generated through prompt engineering for hallucination detection in LLMs. This approach not only yields significant performance improvements but also reduces reliance on costly human annotation.
- We present a scoring function called **Constrastive Mahalanobis Score**, which leverages the distances to both truthful and hallucinated data distributions in the activation space to effectively determine the truthfulness of test data.
- Extensive experiments demonstrate the superior effectiveness of our method compared to other state-of-the-art approaches. Moreover, comprehensive ablation studies assess the impact of various design choices in PALE and confirm its scalability across larger LLMs and diverse datasets. These findings provide a systematic and thorough understanding of leveraging LLM-augmented data for hallucination detection, paving the way for future research.

## Related Works

### Hallucination in LLMs

Large Language Models (LLMs) have made remarkable achievements across various AI domains, including logical reasoning (Dong et al. 2024; Wang et al. 2024a), visual question answering (Jian, Yu, and Zhang 2024), text generation (Min et al. 2023), and speech-to-text transcription (Zhang et al. 2023). Despite their impressive performance, LLMs still face multiple challenges (Lin et al. 2024). Notably, one of the most critical failures is the presence of factual errors in generated text, formally referred to as hallucinations (Zhao et al. 2025; Xu, Jain, and Kankanhalli 2025; Zhang et al. 2025). The existence of hallucinations poses a significant risk in security-critical scenarios such as medical computer-aided diagnosis (Wang et al. 2024b) and financial decision-making (Li et al. 2024). The issue of hallucinations in natural language generation was recognized by NLP researchers even before the widespread adoption of LLMs (Ji et al. 2023). First, the large-scale data corpora used to train LLMs inevitably contain erroneous information (Shumailov et al. 2024), which may contribute to hallucinated outputs. Second, the decoder component of LLMs is typically trained using maximum likelihood estimation, where each token is predicted based on the previously generated sequence, making hallucinations prone to compounding over time (Zhang et al. 2024a). Overall, hallucinations in LLMs remain an unresolved issue, calling for further in-depth research.

### Hallucination Detection

Hallucination detection (Arteaga, Schön, and Pielawski 2024; Liu et al. 2024) has recently gained significant attention as a means to address safety concerns in LLMs and ensure their reliability in real-world deployments. A wide range of studies approach hallucination detection by designing uncertainty scoring functions. For instance, logit-based methods (Ren et al. 2023; Malinin and Gales 2021) utilize token-level log probabilities as uncertainty scores, consistency-based methods (Lin, Trivedi, and Sun 2024a; Manakul, Liusie, and Gales 2023) assess uncertainty by comparing multiple generated responses, and verbalized methods (Lin, Trivedi, and Sun 2024a; Kadavath et al. 2022) prompt LLMs to express confidence in natural language. More recently, internal state-based methods (Du, Xiao, and Li 2024) have been proposed, which leverage hidden activations to detect hallucinations. Notable examples include contrast-consistent search (CCS) and HaloScope (Du, Xiao, and Li 2024). However, these methods still suffer from a scarcity of well-labeled datasets containing both truthful and hallucinated generations. For example, CCS relies on manually curated factual datasets, which require substantial human annotation. HaloScope (Du, Xiao, and Li 2024), on the other hand, employs unlabeled data from the same distribution, limiting its generalizability to more diverse and practical scenarios. In contrast, our method performs hallucination detection with minimal human supervision, making it more applicable to real-world settings. It is important to note that our research problem differs from prior work. For data augmentation, we utilize prompt engineering techniques to generate large-scale hallucinated data. The latent knowledge within LLMs can be harnessed to enrich the diversity of hallucination samples.

## Methodology

In this section, we first formally introduce the LLM generation process and define the problem of hallucination detec-

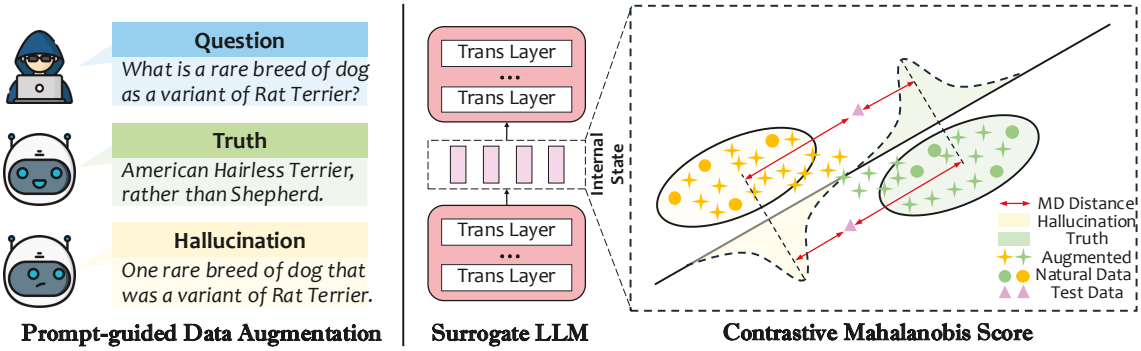



Figure 1: The pipeline of our proposed PALE. Our PALE comprises three steps: 1) Utilizing state-of-the-art LLMs to generate both truthful and hallucinated data; 2) Employing the Contrastive Mahalanobis Score to determine the truthiness of test samples.

tion. We then present the overall framework, followed by a detailed description of the hallucination detector used during inference.

## Problem Formulation

**Definition 1 (LLM generation)** *Given an $L$-layer causal LLM, which takes a sequence of tokens $\mathrm{x} = \{x_1, \ldots, x_n\}$ as input, an output $\mathrm{y} = \{y_{n+1}, \ldots, y_{n+m}\}$ with Markov property is generated in an autoregressive manner. Each output token $y_j$, $j \in [n+1, \ldots, n+m]$ is sampled from a distribution over the model vocabulary $\mathcal{V}$, conditioned on the prefix $\{x_1, \ldots, x_n\}$, which aims maximizing the conditional probability:*

$$y_j = \underset{y \in \mathcal{V}}{\operatorname{argmax}} P\left(y \middle| \{x_1, \ldots, x_{j-1}\}\right), \tag{1}$$

*and the probability $P$ is calculated as:*

$$P\left(y \middle| \{x_1, \ldots, x_{j-1}\}\right) = \operatorname{softmax}\left(\mathbf{w} \mathrm{f}_L\left(y\right) + \mathbf{b}\right), \tag{2}$$

*where $\mathrm{f}_L(y) \in \mathbb{R}^d$ denotes the representation at the $L$-th layer of LLM for token $y$, and $\mathbf{w}$ and $\mathbf{b}$ are the weight and bias parameters at the final output layer, respectively.*

**Definition 2 (Hallucination detection)** *Let $\mathbb{P}_{true}$ and $\mathbb{P}_{hal}$ denote the joint distributions over question and answer pairs $(x, y)$, respectively, referred to as the truthful distribution and the hallucinated distribution. Given any question and answer pairs $(x, y) \in \mathcal{X}$, the goal of hallucination detection is to learn a binary predictor $G$: $\mathcal{X} \to \{-1, 1\}$ such that:*

$$G(x, y) = \begin{cases} 1, & (x, y) \sim \mathbb{P}_{hal}, \\ -1, & (x, y) \sim \mathbb{P}_{true}. \end{cases} \tag{3}$$

## Proposed Framework

**Prompt-guided data augmentation** Hallucination in LLMs is a complex phenomenon in which models produce responses that are coherent yet factually inaccurate across multiple contexts, making hallucination detection particularly challenging. A primary obstacle is the lack of labeled datasets containing both truthful and hallucinated generations. We propose leveraging the LLM's own generation capabilities for data augmentation via prompt guidance. This approach substantially reduces the need for extensive human annotation to assess the authenticity of large numbers of generated samples. Formally, prompt-guided data augmentation can be characterized as follows:

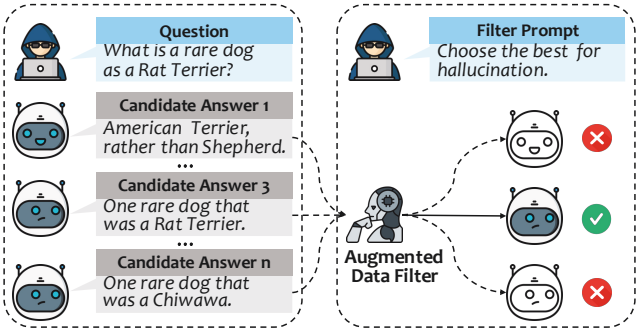



Figure 2: Prompt engineering technology illustration for data augmentation on state-of-the-art LLMs.

**Definition 3 (Prompt-guided data Generation)** *LLMs can perform a variety of tasks in a zero-shot manner under prompt guidance. Given a pretrained LLM $\mathcal{L}_\theta$ and an input question $x$, we define two prompts: the truth prompt $x_t$ and the hallucination prompt $x_h$. The corresponding generations for a truthful answer and a hallucinated answer are formulated as:*

$$y_{true} = \mathcal{L}_\theta\left([x_t, x]\right) \tag{4}$$

$$y_{hal} = \mathcal{L}_\theta\left([x_h, x]\right) \tag{5}$$

*where $y_{true}$ and $y_{hal}$ denote the generated truthful and hallucinated answers, respectively.*

For the design of the truth prompt $x_{true}$ and the hallucination prompt $x_{hal}$, we draw inspiration from (Li et al. 2023). Our framework employs a state-of-the-art LLM $\mathcal{L}_\theta$ to automate the generation of both truthful and hallucinated QA pairs without human annotation. Leveraging prompt engineering techniques, the model produces a series of QA pairs labeled as truthful or hallucinated. Additionally, we incorporate a verification mechanism to filter and retain only high-quality, truthful and hallucinated QA data, as illustrated in Figure. 2. Detailed descriptions of the prompt are provided in the Appendix.

**Definition 4 (Empirical dataset)** *Let* $\mathcal{M} = \{\mathcal{M}_{true}, \mathcal{M}_{hal}\}$ *denotes our empirical dataset, comprising a truthful subset*

$$\mathcal{M}_{true} = \left\{\left(x^1, y^1_{true}\right), \ldots, \left(x^N, y^N_{true}\right)\right\} \quad (6)$$

*and a hallucinated subset*

$$\mathcal{M}_{hal} = \left\{\left(x^1, y^1_{hal}\right), \ldots, \left(x^N, y^N_{hal}\right)\right\} \quad (7)$$

*where* $x^i$ *is the* $i$*-th input question and* $y^i_{true/hal}$ *denotes its corresponding truthful or hallucinated response. Here,* $N$ *denotes the number of samples in each subset.*

Despite the generation of both truthful and hallucinated data, evaluating the truthfulness of test samples remains a separate challenge. The internal embedding representations of the empirical dataset are typically sparse and relatively low-dimensional compared to those of natural data. Prior studies (Kapur, Marwah, and Alterovitz 2016; Gillis 2014) have indicated that MLP-based classifiers are prone to overfitting under such conditions. To address this, we propose a novel metric called the **C**ontrastive **M**ahalanobis (**CM**) score to determine the truthiness of prompt-guided augmented data from LLMs.

**Contrastive Mahalanobis Score** As indicated in prior work(Chen et al. 2024; Du, Xiao, and Li 2024; Azaria and Mitchell 2023), we employ hidden-state representations of LLMs for hallucination detection. For the $i$-th data sample $m^i = \left(x^i, y^i_{true/hal}\right)$ in dataset $\mathcal{M}$, let $h^l_i \in \mathbb{R}^d$ denote the hidden-state embedding at the $l$-th layer, where $d$ is the the dimension of the hidden embedding. Following (Azaria and Mitchell 2023; Ren et al. 2023), we compute a sentence embedding by averaging its token embeddings:

$$\mathbf{z} = \frac{1}{T}\sum_{i=1}^{T} h_i. \quad (8)$$

where $T$ is the number of tokens. For the $N$ samples in the truthful and hallucinated subsets, we collect their sentence embeddings into matrices $\mathbf{Z}_{true} \in \mathbb{R}^{N\times d}$ and $\mathbf{Z}_{hal} \in \mathbb{R}^{N\times d}$, respectively. We then perform singular value decomposition on these two embedding matrices:

$$z^i_{true} := z^i_{true} - \mu_{true} \quad \mathbf{Z}_{true} = \mathbf{U}_{true}\Sigma_{true}\mathbf{V}^T_{true} \quad (9)$$

$$z^i_{hal} := z^i_{hal} - \mu_{hal} \quad \mathbf{Z}_{hal} = \mathbf{U}_{hal}\Sigma_{hal}\mathbf{V}^T_{hal} \quad (10)$$

where $\mu_{true} \in \mathbb{R}^d$ and $\mu_{hal} \in \mathbb{R}^d$ denote the mean embeddings of the $N$ truthful and hallucinated samples, respectively, and are used to center each embedding matrix. The columns of $U_{true/hal}$ and $V_{true/hal}$ are the left and right singular vectors, forming orthonormal bases for the truthful and hallucinated subspaces. This factorization serves two purposes: 1) It identifies the principal spanning directions of the point sets in the hidden-state space; 2) It enables dimensionality reduction to reduce computational cost. Specifically, we truncate to the top $k$ components of singular vectors.

The covariance matrices for the truthful and hallucinated embedding matrices, $\mathbf{Z}_{true}$ and $\mathbf{Z}_{hal}$, are computed as

$$\mathbf{C}_{true} = \frac{1}{N}\mathbf{Z}^T_{true}\mathbf{Z}_{true} = \frac{1}{N}V_{true}\Sigma^T_{true}\Sigma_{true}V^T_{true} \quad (11)$$

$$\mathbf{C}_{hal} = \frac{1}{N}\mathbf{Z}^T_{hal}\mathbf{Z}_{hal} = \frac{1}{N}V_{hal}\Sigma^T_{hal}\Sigma_{hal}V^T_{hal} \quad (12)$$

Intuitively, if the embedding of an output instance lies close to the truthful embedding distribution, the instance is more likely to be reliable. Conversely, if it aligns more closely with the hallucinated distribution, it is more likely to be suspicious. Assuming both distributions are Gaussian in the embedding space, we denote the truthful and hallucinated Gaussians as $\mathcal{N}(\mu_{true}, \mathbf{C}_{true})$ and $\mathcal{N}(\mu_{hal}, \mathbf{C}_{hal})$, respectively. One way to quantify the distance of an embedding $z$ to a Gaussian distribution is via the Mahalanobis distance (MD).

When considering both the truthful and hallucinated distributions, the **C**ontrastive **M**ahalanobis (**CM**) Score is defined as

$$\delta = MD(z; \mu_{hal}, \mathbf{C}_{hal}) - MD(z; \mu_{true}, \mathbf{C}_{true}) \quad (13)$$

where $MD(z; \mu_{hal}, \mathbf{C}_{hal})$ and $MD(z; \mu_{true}, \mathbf{C}_{true})$ denote the MD of $z$ to the hallucinated and truthful Gaussian distributions, respectively. In this work, both Gaussians are estimated from hidden-state embeddings in the activation space.

The score defined in Eq. 13 serves as a truthfulness score, indicating how closely a sample $z$ aligns with the hallucinated domain relative to the truthful domain. A score $\delta \geq \tau$ suggests that $z$ is closer to the hallucinated distribution, while $\delta < \tau$ implies a greater proximity to the truthful distribution. In other words, higher values of $\delta$ indicate a greater likelihood of hallucination, whereas lower values correspond to higher confidence in truthfulness. $\tau$ represents the decision threshold.

# Experiments

## Datasets

We evaluate our method on four generative QA benchmarks: two open-domain datasets: TruthfulQA (Lin, Hilton, and Evans 2022b) and CoQA (Reddy, Chen, and Manning 2019), and two domain-specific datasets TriviaQA (Joshi et al. 2017) and TyDi QA-GP (Clark et al. 2020). Specifically, TruthfulQA and TyDi QA-GP contain 817 and 3,696 QA pairs, respectively. For CoQA, we follow (Lin, Trivedi, and Sun 2024b) and split the data into 7,983 QA pairs. For TriviaQA, we use the 9,960 QA pairs in the validation set's (*rc.nocontext subset*). Consistent with prior work (Du, Xiao, and Li 2024), we reserve 25% of QA pairs in each dataset for testing; the remaining pairs are used for LLM-based data augmentation.

## Models

For base models used to extract hidden-state representations, we employ three popular open source LLM families: LLaMA-3.1-chat (7B and 13B) (Touvron, Martin et al. 2023), OPT (6.7B and 13B) (Zhang, Roller et al. 2022), and Qwen-2.5 (7B and 14B) (Yang et al. 2024).

For data augmentation, we employ not only the open source LLaMA-3.1-chat-7B (Touvron, Martin et al. 2023) and Qwen-2.5-7B (Yang et al. 2024) models but also state-of-the-art commercial LLMs: GPT-4o (Brown et al. 2020) and Claude.

## Baselines

We compare our approach with 11 comprehensive baselines. These state-of-the-art baselines are categorized as follows: 1) uncertainty-based methods: Perplexity (Ren et al. 2023), Length-Normalized Entropy (LN-entropy) (Malinin and Gales 2021), and Semantic Entropy (Kuhn, Gal, and Farquhar 2023). 2) consistency-based methods: Lexical Similarity (Lin, Trivedi, and Sun 2024a), SelfCKGPT (Manakul, Liusie, and Gales 2023) and EigenScore (Chen et al. 2024). 3) verbalized methods: Verbalize (Lin, Hilton, and Evans 2022a) and Self-evaluation (Kadavath et al. 2022). 4) internal state-based methods: MIND (Su et al. 2024), Contrast-Consistent Search (CCS) (Burns et al. 2022) and HaloScope (Du, Xiao, and Li 2024). To ensure a fair comparison, all methods are evaluated on the same test datasets and the default experimental configurations are adopted as specified in their respective papers.

## Evaluation

Consistent with prior work (Kuhn, Gal, and Farquhar 2023; Du, Xiao, and Li 2024), we evaluate all methods using the area under the receiver operating characteristic curve (AUROC). A generated response is considered truthful only if its similarity to the reference answer exceeds 0.5. Following (Lin, Hilton, and Evans 2022b), we use BLUERT (Sellam, Das, and Parikh 2020) to compute this similarity. Additionally, we report semantic-similarity evaluations using GPT-4o.

## Implementation and Settings

Consistent with (Kuhn, Gal, and Farquhar 2023; Du, Xiao, and Li 2024), we generate the most likely answer using beam search with five beams for evaluation and employ multinomial sampling to produce ten samples per question. Following (Chen et al. 2024; Azaria and Mitchell 2023), we concatenate each question with its generated answer and use the last-token embedding as the representation for hallucination detection. The $k$ and $\tau$ are set to 5 and 0.15, respectively. More details and hyperparameter sensitivity study are provided in the Appendix.

## Main Results

As shown in Table 1, we compare PALE against competitive hallucination-detection methods from the literature. PALE achieves state-of-the-art performance across LLaMA-3.1-7B, OPT-6-7B, and Qwen-2.5-7B. Notably, PALE outperforms uncertainty-based and consistency-based methods by an average of 19.3% and 23.6%, respectively, over Semantic Entropy and EigenScore on TruthfulQA. Unlike these methods, which require $K$ repeated samples per question at test time, incurring $O\left(Km^2\right)$ computational overhead (where $m$ is the number of generated tokens). PALE requires no repeated sampling, yielding $O\left(m^2\right)$ complexity. PALE also surpasses verbalized methods, achieving a 22.5% improvement over prompt-based language model approaches, likely due to reduced overconfidence as discussed in prior work (Zhou, Jurafsky, and Hashimoto 2023; Wen et al. 2024). Finally, compared to internal-state methods MIND, CCS and HaloScope, PALE performs better than CCS, demonstrating that prompt-guided augmented data better captures the true versus hallucinated distribution than limited human-written examples and exceeds HaloScope by 6.5% on TruthfulQA with LLaMA-3.1. Whereas HaloScope uses only unlabeled in-domain data, PALE leverages non-parametric prompt engineering to augment both truthful and hallucinated examples, leading to significantly improved detection performance.

## Result on GPT-4o Evaluation

Besides the BLEURT score is used to determine whether a generation is considered truthful when its score exceeds a predefined threshold with answer, we also adopt GPT-4o for truthfulness evaluation, following the *LLM-as-a-judge* approach (Zheng et al. 2023). Specifically, we assess the semantic equivalence between LLM-generated responses and the corresponding reference answers. The results in Appendix show that our method consistently outperforms competitive baselines, demonstrating its robustness across different metrics for evaluating generation truthfulness. Further details are provided in the Appendix.

## Generalization to Practical Application

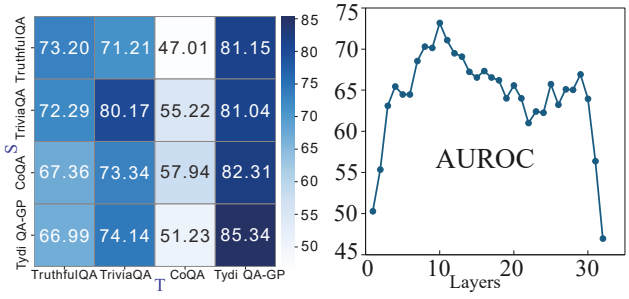
Our proposed PALE method demonstrates strong generalization capabilities in practical applications. In this section, we investigate the *transferability* of PALE across different data distributions, as well as its *scalability* when applied to larger LLMs.

**Can our PALE deal with different data distribution?** In real-world applications, the training data distribution often does not align well with the test data distribution. This discrepancy raises a critical question: **Does our proposed PALE method exhibit transferability across different data distributions?** To investigate this, we conduct experiments by training the hallucination detector on a source dataset and evaluating its performance on a target dataset with a different distribution. The results, presented in Figure 3a, demonstrate that PALE exhibits strong cross-dataset transferability. Notably, our method achieves impressive performance even under distribution shift, with an average accuracy of approximately 72%. This strong transferability highlights the robustness and practical utility of PALE in real-world scenarios.

**Can our PALE salable to larger LLMs?** With the continuous release of increasingly larger LLMs (Grattafiori et al. 2024; Hurst et al. 2024; Team et al. 2023), an important question arises: **Does PALE scale effectively to larger language models?** To evaluate the scalability of our approach, we conduct experiments on three larger LLMs: LLaMA-3.1-13B, OPT-6-13B, and Qwen-2.5-14B. As shown in Table 2, PALE maintains strong performance and even achieves improvements compared to results obtained with smaller models. These findings demonstrate the scalability and adaptability of PALE to more powerful LLM architectures.

| Model | Method | Single Sampling | TruthfulQA | TriviaQA | CoQA | TyDi QA-GP |
|---|---|---|---|---|---|---|
| LLaMA-3.1-7B | Perplexity $_{(ICLR'23)}$ | ✓ | 57.62 | 73.11 | 69.77 | 78.54 |
| | LN-entropy $_{(ICLR'21)}$ | ✗ | 60.15 | 71.31 | 73.02 | 76.15 |
| | Semantic Entropy $_{(ICLR'23)}$ | ✗ | 62.16 | 73.18 | 63.24 | 73.87 |
| | Lexical Similarity $_{(TMLR'24)}$ | ✗ | 55.64 | 75.69 | **74.36** | 44.46 |
| | SelfCKGPT $_{(EMNLP'23)}$ | ✗ | 52.90 | 73.97 | 71.76 | 46.38 |
| | EigenScore $_{(ICLR'24)}$ | ✗ | 51.92 | 73.99 | 71.75 | 46.38 |
| | Verbalize $_{(TMLR'22)}$ | ✓ | 53.07 | 52.54 | 48.46 | 48.12 |
| | Self-evaluation $_{(Arxiv'22)}$ | ✓ | 81.75 | 49.11 | 50.15 | 55.48 |
| | CCS (Burns et al. 2022) $_{(Arxiv'22)}$ | ✓ | 61.33 | 60.22 | 50.37 | 76.82 |
| | MIND $_{(ACL'24)}$ | ✓ | 67.53 | 74.66 | 53.96 | 75.64 |
| | HaloScope $_{(NeurIPS'24)}$ | ✓ | 70.16 | 76.13 | 55.47 | 78.38 |
| | **PALE (Our)** | ✓ | **73.20** | **80.17** | 57.94 | **85.34** |
| OPT-6-7B | Perplexity $_{(ICLR'23)}$ | ✓ | 59.16 | 69.69 | 70.36 | 63.94 |
| | LN-entropy $_{(ICLR'21)}$ | ✗ | 54.24 | 71.43 | 71.28 | 52.10 |
| | Semantic Entropy $_{(ICLR'23)}$ | ✗ | 52.06 | 71.45 | 71.24 | 52.07 |
| | Lexical Similarity $_{(TMLR'24)}$ | ✗ | 49.69 | 71.09 | 66.63 | 60.42 |
| | SelfCKGPT $_{(EMNLP'23)}$ | ✗ | 50.15 | 71.45 | 64.68 | 74.98 |
| | EigenScore $_{(ICLR'24)}$ | ✗ | 54.93 | 47.67 | 50.30 | 45.28 |
| | Verbalize $_{(TMLR'22)}$ | ✓ | 50.47 | 50.73 | 55.26 | 57.23 |
| | Self-evaluation $_{(Arxiv'22)}$ | ✓ | 51.07 | 53.91 | 47.28 | 52.08 |
| | CCS (Burns et al. 2022) $_{(Arxiv'22)}$ | ✓ | 60.17 | 51.33 | 53.19 | 65.37 |
| | MIND $_{(ACL'24)}$ | ✓ | 65.06 | 63.32 | 66.35 | 70.37 |
| | HaloScope $_{(NeurIPS'24)}$ | ✓ | 67.82 | 65.39 | 67.24 | 72.36 |
| | **PALE (Our)** | ✓ | **74.17** | **74.34** | **78.64** | **81.49** |
| Qwen-2.5-7B | Perplexity $_{(ICLR'23)}$ | ✓ | 65.17 | 50.23 | 53.47 | 54.33 |
| | LN-entropy $_{(ICLR'21)}$ | ✗ | 66.73 | 51.15 | 52.74 | 55.38 |
| | Semantic Entropy $_{(ICLR'23)}$ | ✗ | 58.76 | 48.58 | 63.71 | 65.72 |
| | Lexical Similarity $_{(TMLR'24)}$ | ✗ | 49.05 | 63.17 | 48.96 | 61.23 |
| | SelfCKGPT $_{(EMNLP'23)}$ | ✗ | 61.75 | 62.34 | 62.28 | 63.48 |
| | EigenScore $_{(ICLR'24)}$ | ✗ | 53.73 | 61.30 | 63.37 | 58.54 |
| | Verbalize $_{(TMLR'22)}$ | ✓ | 60.07 | 54.33 | 59.46 | 52.33 |
| | Self-evaluation $_{(Arxiv'22)}$ | ✓ | 73.71 | 50.12 | 53.85 | 52.87 |
| | CCS (Burns et al. 2022) $_{(Arxiv'22)}$ | ✓ | 67.96 | 53.08 | 51.94 | 51.77 |
| | MIND $_{(ACL'24)}$ | ✓ | 70.63 | 71.95 | 68.00 | 65.27 |
| | HaloScope $_{(NeurIPS'24)}$ | ✓ | 73.42 | 70.73 | 70.61 | 67.52 |
| | **PALE (Our)** | ✓ | **83.68** | **78.33** | **79.00** | **72.82** |

Table 1: Main results. Comparison with comprehensive baseline hallucination-detection methods across different datasets. "Single Sampling" indicates whether the approach requires multiple generation during inference. All values are percentages (AUROC), and the top-1 result is highlighted in bold.




(a) Generalization across four dataset distribution.

(b) Effect of the different layer to extract.

Figure 3: (a) Generalization across four datasets. Rows indicate the source dataset, while columns indicate the target dataset. (b) Impact of different layers. All values represent AUROC scores based on LLaMA-3.1-7B.

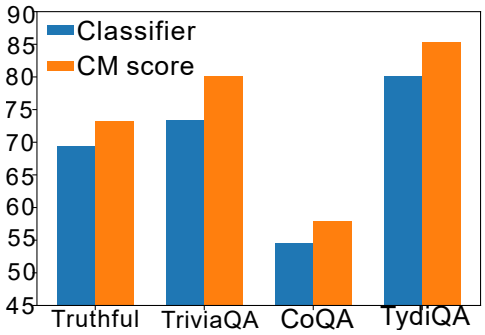



Figure 4: Comparison of direct classifier based detection versus CM Score based detection. All results are reported as AUROC using LLaMA-3.1-7B.

| Method | LlaMA-3.1-13B | | OPT-6-13B | | Qwen-2.5-14B | |
|---|---|---|---|---|---|---|
| | TruthfulQA | TyDi QA-GP | TruthfulQA | TyDi QA-GP | TruthfulQA | TyDi QA-GP |
| Perplexity | 52.32 | 77.33 | 58.34 | 64.12 | 64.33 | 55.12 |
| CCS | 63.11 | 77.11 | 60.00 | 66.73 | 68.21 | 53.66 |
| HaloScope | 71.42 | 79.33 | 68.31 | 73.24 | 74.11 | 68.44 |
| PALE | **75.51** | **86.72** | **75.33** | **82.77** | **84.37** | **73.23** |

Table 2: Hallucination detection results on larger LLMs.

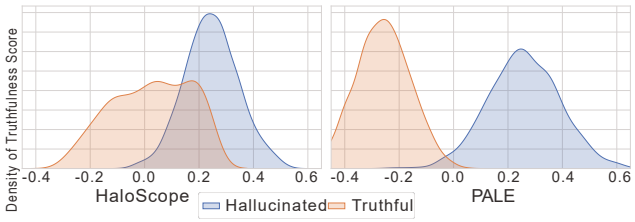



Figure 5: Score distribution visualization for HaloScope vs. our method.

## Analysis

**Effect of LLM-generated data augmentation** To evaluate the impact of LLM-generated data augmentation on hallucination detection, we conduct experiments on the TruthfulQA dataset. The experimental settings include training only on truthful and hallucinated data, using various LLMs for data augmentation. The results are presented in Appendix. Our key observations are as follows: 1) Compared to training solely on either truthful or hallucinated data, PALE demonstrates substantial improvements—achieving average AUROC gains of 11.7% and 39.0%, respectively. This highlights the significance of utilizing both truthful and hallucinated examples during training. 2) We experiment with several LLMs for data augmentation. While the Claude model yields the best performance, we find that the choice of LLM has a relatively minor influence on overall results. Consequently, we employ Claude throughout the paper for consistency. This demonstrates that our method, when paired with LLM-generated data augmentation, can significantly boost the performance of hallucination detection.

**Effect of prompt for data augmentation** We focus on investigating the effect of prompts for data augmentation. Specifically, we conducted experiments using PALE across all 10 prompt templates presented in the Appendix. The results, also shown in the Appendix, reveal a key insight: the quality of the prompt template does not significantly affect hallucination detection performance. PALE consistently demonstrates strong performance across all prompt templates, which not only validates the effectiveness of our data augmentation strategy but also highlights the robustness of the method to prompt variations.

**Effect of Contrastive Mahalanobis Score** Figure 4 presents a comparison between using a direct binary classifier and the CM score for hallucination detection. The CM score approach regresses the representation of a test sample to a truthfulness score, thereby bypassing the need to train a separate binary classifier. Across all four datasets, PALE consistently outperforms the direct classification method, demonstrating the superior generalizability of the matrix decomposition approach on sparse embedding data.

**Effect of different layers's representation** In Appendix, we analyze the impact of the layer selection for extracting internal representations in PALE, using LLaMA-3.1-7B as the backbone. All other experimental configurations remain consistent with our main setup. We observe a clear trend: hallucination detection performance improves from the lower to the middle layers (e.g., from the 1st to the 11th layer), followed by a performance decline in the higher layers. This observation suggests that early layers primarily perform information aggregation, while later layers may exhibit overconfidence due to the autoregressive nature of language modeling focused on vocabulary prediction. This finding aligns with prior studies (Jawahar, Sagot, and Seddah 2019; Hewitt and Manning 2019), which show that intermediate layer representations tend to be most effective for downstream tasks.

**Visualization study** Figure.5 visualizes the score distributions of HaloScope and our proposed PALE on the TruthfulQA dataset using the LLaMA-3.1-7B model. Compared to HaloScope, PALE exhibits a more distinct separation between the distributions of truthful and hallucinated data. This clearer differentiation highlights the superior discriminative capability of PALE, underscoring its effectiveness in hallucination detection.

## Conclusion

The paper introduces **P**rompt-guided data **A**ugmented ha**L**lucination d**E**tection (PALE), a novel framework designed to detect hallucinations in LLMs by leveraging prompt-guided data augmentation. PALE first generates both truthful and hallucinated data under the guidance of prompt engineering. It then introduces the Contrastive Mahalanobis Score to evaluate the truthfulness of augmented data based on their distances to the distributions of truthful and hallucinated samples. Empirical results demonstrate that PALE achieves superior performance across a comprehensive set of hallucination detection benchmarks. Furthermore, our in-depth ablation studies provide valuable insights into the practical effectiveness of PALE. We hope this work inspires future research on hallucination detection using augmented data, particularly in exploring multimodal hallucination detection in MLLMs to further enhance real-world applicability.

## Acknowledgements

This research was partially supported by the National Natural Science Foundation of China (Grant Nos. 62372132, 62402252, and 62536003), the Shenzhen Science and Technology Program (Grant No. RCYX20221008092852077) and Guangdong High-Level Talent Programme (Grant No. 2024TQ08X283). The authors would also like to thank Huawei Ascend Cloud Ecological Development Project for providing high-performance Ascend 910 processors.

## References

Arteaga, G. Y.; Schön, T. B.; and Pielawski, N. 2024. Hallucination Detection in LLMs: Fast and Memory-Efficient Fine-Tuned Models. arXiv:2409.02976.

Azaria, A.; and Mitchell, T. 2023. The Internal State of an LLM Knows When It's Lying. In *EMNLP*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; et al. 2020. Language models are few-shot learners. *NeurIPS*.

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Chen, C.; Liu, K.; Chen, Z.; Gu, Y.; Wu, Y.; Tao, M.; Fu, Z.; and Ye, J. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *ICLR*.

Clark, J. H.; Palomaki, J.; Nikolaev, V.; Choi, E.; Garrette, D.; Collins, M.; and Kwiatkowski, T. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *TACL*.

Dong, Y.; Liu, Z.; Sun, H.-L.; Yang, J.; Hu, W.; Rao, Y.; and Liu, Z. 2024. Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models. arXiv:2411.14432.

Du, X.; Xiao, C.; and Li, Y. 2024. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. In *NeurIPS*.

Gillis, N. 2014. The Why and How of Nonnegative Matrix Factorization. arXiv:1401.5226.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hewitt, J.; and Manning, C. D. 2019. A Structural Probe for Finding Syntax in Word Representations. In *NAACL*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM TOIS*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What Does BERT Learn about the Structure of Language? In *ACL*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*

Jian, P.; Yu, D.; and Zhang, J. 2024. Large Language Models Know What is Key Visual Entity: An LLM-assisted Multimodal Retrieval for VQA. In *EMNLP*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *ACL*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kapur, A.; Marwah, K.; and Alterovitz, G. 2016. Gene expression prediction using low-rank matrix completion. *BMC bioinformatics*, 17(1): 243.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *ICLR*.

Li, H.; Cao, Y.; Yu, Y.; Javaji, S. R.; Deng, Z.; He, Y.; et al. 2024. INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent. arXiv:2412.18174.

Li, J.; Cheng, X.; Zhao, X.; Nie, J.-Y.; and Wen, J.-R. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *EMNLP*.

Li, W.; and Pun, C.-M. 2023. ELF: An End-to-end Local and Global Multimodal Fusion Framework for Glaucoma Grading. In *IEEE BIBM*.

Li, W.; Yu, G.; Li, Q.; et al. 2025. Elevating Medical Image Security: A Cryptographic Framework Integrating Hyperchaotic Map and GRU. *arXiv*.

Lin, S.; Hilton, J.; and Evans, O. 2022a. Teaching Models to Express Their Uncertainty in Words. *TMLR*.

Lin, S.; Hilton, J.; and Evans, O. 2022b. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *ACL*.

Lin, Y.; He, P.; Xu, H.; Xing, Y.; Yamada, M.; Liu, H.; and Tang, J. 2024. Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis. In *EMNLP*.

Lin, Z.; Trivedi, S.; and Sun, J. 2024a. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *TMLR*.

Lin, Z.; Trivedi, S.; and Sun, J. 2024b. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *TMLR*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Malinin, A.; and Gales, M. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *ICLR*.

Manakul, P.; Liusie, A.; and Gales, M. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *EMNLP*.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A Conversational Question Answering Challenge. *TACL*.

Ren, J.; Luo, J.; Zhao, Y.; Krishna, K.; Saleh, M.; Lakshminarayanan, B.; and Liu, P. J. 2023. Out-of-Distribution Detection and Selective Generation for Conditional Language Models. In *ICLR*.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *ACL*.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*.

Sriramanan, G.; Bharti, S.; Sadasivan, V. S.; Saha, S.; Kattakinda, P.; and Feizi, S. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *NeurIPS, 2024*.

Su, W.; Wang, C.; Ai, Q.; Hu, Y.; Wu, Z.; Zhou, Y.; and Liu, Y. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of ACL*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Martin, L.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, L.; Hu, Y.; He, J.; Xu, X.; Liu, N.; Liu, H.; and Shen, H. T. 2024a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *AAAI*.

Wang, S.; Zhao, Z.; Ouyang, X.; Liu, T.; Wang, Q.; and Shen, D. 2024b. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1): 133.

Wei, J.; and Zhang, X. 2024. DOPRA: Decoding Over-accumulation Penalization and Re-allocation in Specific Weighting Layer. *ACM MM*.

Wen, B.; Xu, C.; Wolfe, R.; Wang, L. L.; Howe, B.; et al. 2024. Mitigating overconfidence in large language models: A behavioral lens on confidence estimation and calibration. In *NeurIPS Workshop*.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Findings of EMNLP*.

Zhang, M.; Press, O.; Merrill, W.; Liu, A.; and Smith, N. A. 2024a. How Language Model Hallucinations Can Snowball. In *ICML*.

Zhang, S.; Roller, S.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, X.; Quan, Y.; Gu, C.; et al. 2025. Shallow Focus, Deep Fixes: Enhancing Shallow Layers Vision Attention Sinks to Alleviate Hallucination in LVLMs.

Zhang, X.; Shen, C.; Yuan, X.; Yan, S.; Xie, L.; Wang, W.; Gu, C.; Tang, H.; and Ye, J. 2024b. From Redundancy to Relevance: Enhancing Explainability in Multimodal Large Language Models. *NAACL*.

Zhao, Q.; Zhang, X.; Li, y.; Xing, Y.; Xiaosong, Y.; Tang, F.; Fan, S.; Chen, X.; Zhang, X.; and Wang, D. 2025. MCA-LLaVA: Manhattan Causal Attention for Reducing Hallucination in Large Vision-Language Models. *ACM MM*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, J.; Li, Q.; Li, W.; et al. 2025. TDADL-IE: A Deep Learning-Driven Cryptographic Architecture for Medical Image Security. *arXiv*.

Zhou, K.; Jurafsky, D.; and Hashimoto, T. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. *arXiv preprint arXiv:2302.13439*.