IDENTIFYING TRUTHFUL INHERITANCE IN FAMILIY MODELS AND ENHANCING TRUTHFULNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

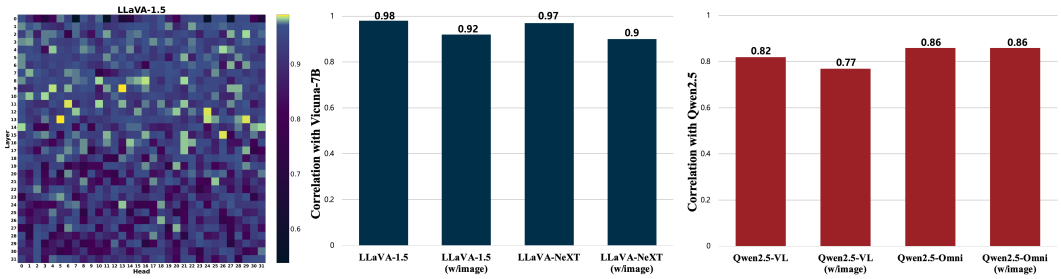
Recent advances in large language models (LLMs) have led to emergence of specialized multimodal LLMs (MLLMs), creating distinct model families that share a common foundation language models. This work investigates whether a core traits like **truthfulness** are inherited along this evolutionary trajectory. To quantify this trait, we employ linear probing on the models' internal representations. Our analysis of Vicuna and Qwen model families reveals a key finding: a strong correlation in truthfulness scores between LLMs and their finetuned MLLMs counterparts, even when they are finetuned or probed with different modalities and datasets. Building on this findings, we propose a soft gating method using the *Truthfulness* score to amplify the influence of these context-truthful heads to improve the context grounding ability while preserving the contributions of other heads. We validate our approach on base LLMs on HaluEval benchmark, demonstrating an improved ability for context truthful reasoning. We then show that the Truthfulness scores obtained from base LLMs can be effectively transferred and applied as a soft gate to its finetuned MLLMs, demonstrating its improved performance on POPE benchmark. The performance gain from this transfer is comparable to that obtained by probing the MLLMs directly, highlighting the potential for a unified approach to enhance truthfulness across an entire model family. Our work demonstrates a novel method for leveraging a model's inherent, inherited traits to systematically improve its truthfulness.

1 INTRODUCTION

Recent advancements in large language models (LLMs) has given rise to a wide range of specialized models, all of which are originated from a core foundational LLMs. This pattern reflects a broader trend: rather than building entirely new models from scratch, base LLMs are often refined through fine-tuning or multimodal extensions to serve domain-specific needs—ranging from mathematical reasoning to vision-language understanding, or even multi-sensory processing. Such evolutionary trajectories highlight that many advanced multimodal LLMs (MLLMs) share a clear lineage with their base LLMs.

Do these models inherit traits like truthfulness? If so, could we leverage this inherited trait to develop an unified method that enhances truthfulness not only in base LLMs but also their finetuned MLLMs?

Inspired by ITI (Li et al., 2023b), we hypothesized that if the model component, specifically the Attention Head, can classify whether the given context is truthful or not, that component can be regarded as truthful. To investigate this, we analyze how such characteristics are preserved and correlated across models sharing the same language model. Specifically, we study Vicuna-7B (Chiang et al., 2023) as a base LLM and its fine-tuned counterparts, LLaVA-1.5 (Liu et al., 2024a) and LLaVA-NeXT (Li et al., 2024) as well as Qwen2.5 family (Qwen et al., 2025), including Qwen2.5-VL-Instruct (Bai et al., 2025) and Qwen2.5-VL-Omni (Xu et al., 2025). Our analysis reveals the key property within model families: **Inheritance.** Truthfulness scores of MLLMs are highly correlated with those of their base LLMs, regardless of their specialization for different modalities, such as image and audio. Even when proved with different modality data for LLMs and MLLMs, the truthfulness correlation within a shared model family is much higher than that between models from unrelated families.



(a) **(b)** **(c)**

Figure 1: (a) Truthfulness scores of all attention heads across layers obtained via linear probing on LLaVA-1.5. (b) Cross-model similarity of probing results for MLLMs derived from the Vicuna-7B base language model. (c) Comparative similarity of probing results for MLLMs adapted from the Qwen2.5 base language model.

This finding suggests that MLLMs inherit truthful heads from the base LLMs. In addition, heads that accurately reference the provided context maintain consistent truthfulness across different datasets, indicating that this property is largely dataset-independent.

Building on these findings, we propose to amplify the influence of context-truthful heads so that the model's final outputs are more faithfully grounded in the given context, while still preserving the complementary roles of other heads. Importantly, we show that this gating strategy is not only effective within a single model, but also generalizes consistently across model families sharing the same backbone.

To begin, we validate the obtained truthfulness score from the base LLMs on HaluEval benchmark, showing their improved ability for truthful reasoning within a given context. Further, we explored whether the truthfulness scores identified in base LLMs can serve as a soft gate in their finetuned MLLMs, and observed performance gains on POPE benchmark due to the inherited traits. This performance improvement is comparable to that obtained by applying truthfulness scores derived from MLLMs itself.

Unlike prior approaches that targeted hallucination reduction through model-specific interventions, or head-level studies that remained descriptive without actionable refinement, our work identifies transferable components of truthfulness and operationalizes them for model improvement. By showing that truthfulness scores can be stably inherited and transferred within model families, we establish a principled foundation for refining both LLMs and their multimodal extensions toward greater truthfulness.

Our contributions are summarized as follows:

- We perform linear probing to systematically identify attention heads that ground responses truthfully in the given context.
- Through cross-model analysis, we demonstrate that context-truthful heads are largely preserved when LLMs are fine-tuned into MLLMs, revealing strong correlations between base models and their multimodal descendants.
- We introduce a refinement strategy, *TruthProbe*, which amplifies the influence of context-truthful heads. This approach yields more reliable and grounded outputs, and notably, achieves comparable performance gains in MLLMs even when the truthfulness scores are transferred directly from their base LLMs.

2 Identifying Components for Context-based Truthful Reasoning

Recent research has made significant strides in demystifying the internal mechanisms of Large Language Models (LLMs). A particularly compelling line of inquiry suggests that abstract concepts are encoded in interpretable directions within the model's activation space. For example, (Li et al.,

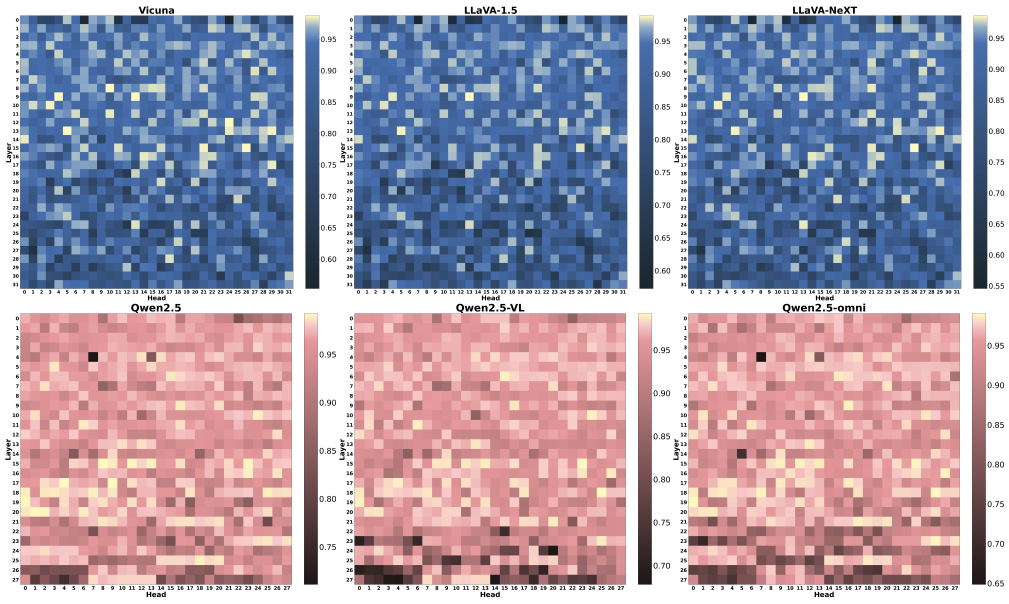


Figure 2: Heatmaps of head-level probing accuracy for two model families. (Top) Vicuna-based models, including LLaVA-1.5 and LLaVA-NeXT, fine-tuned from Vicuna-7B. (Bottom) Qwen2.5-based models, including Qwen2.5-VL and Qwen2.5-Omni, fine-tuned from Qwen2.5.

2023b) introduced Inference-Time Intervention (ITI), a technique that enhances model truthfulness by identifying and shifting activations in specific attention heads. Their findings indicate that models may possess latent "knowledge" of the truth, even when their generated outputs are false.

However, in many real-world applications, truthful reasoning requires more than accessing parametric knowledge—it also depends critically on how well the model leverages the given context. For instance, in Multimodal Large Language Models (MLLMs), tasks such as the widely studied "Where is Wally?" question require accurate grounding in the provided image, rather than relying solely on pre-trained internal knowledge. Motivated by this distinction, we move beyond ITI and focus on identifying attention heads that are not only truthful but also context-referential. Specifically, we aim to characterize and intervene on heads that reliably attend to context in a manner that supports truthful and grounded responses.

2.1 Preliminary

Formally, in a Transformer layer l, the Multi-Head Attention (MHA) mechanism is composed of H attention heads, each applying an independent linear projection to the residual representation. Given an input $x_l \in \mathbb{R}^d$, the h-th head projects it into query, key, and value subspaces via learned matrices Q_l^h, K_l^h, V_l^h. The head output is computed as:

$$Att_l^h(x_l) = \text{softmax}\left(\frac{Q_l^h x_l (K_l^h x_l)^\top}{\sqrt{d_k}}\right) V_l^h x_l, \tag{1}$$

where d_k denotes the key dimension. The outputs of all heads are then aggregated through an output projection W_l^o and added back to the residual stream:

$$o_l = W_l^o \cdot \text{Concat}_{h=1}^{H}(Att_l^h(x_l)) \tag{2}$$

$$x_{l+1} = x_l + o_l. \tag{3}$$

This formulation shows that each head contributes a distinct contextual transformation, which is subsequently integrated by the Multi-Layer Perceptron (MLP) through nonlinear operations.

2.2 FINDING CONTEXT TRUTHFUL HEAD

As introduced in Section 2, evaluating whether a Transformer layer truthfully leverages contextual information is most precise at the granularity of individual attention heads. Each head selectively references tokens from the context and adds its transformed representation into the residual stream. By analyzing heads individually, one can assess whether the contextual information is faithfully preserved or distorted.

We adopt the emerging view that neural networks encode interpretable directions in activation space and hypothesize that certain heads correspond to **truthfulness**. Specifically, we examine whether each head integrates context in a reliable manner or propagates misleading signals. To test this, we apply **linear probing** (Alain & Bengio, 2017) at the head level: a probe of each head is trained to discern whether the given sequence is truthful or not.

Our framework extends beyond Large Language Models (LLMs) to Multimodal Large Language Models (MLLMs), where contextual grounding is even more critical. For this setting, we structure the input as $x = \{x_{\text{knowledge}}, x_{\text{question}}, x_{\text{answer}}\}$, where the knowledge can be text of world knowledge or the real image, and analyze the activations at the position of the final answer token. The probe of each head is trained as a binary classifier, distinguishing heads that truthfully reference the input context.

Concretely, for each attention head h in layer l, we collect the attention head output vector x_l^h that contributes to the residual stream at the final answer position. The probe takes the form

$$p_\theta(x_l^h) = \sigma(\langle \theta, x_l^h \rangle), \tag{4}$$

where $\theta \in \mathbb{R}^D$ is the probe parameter and σ denotes the sigmoid function. We construct probing datasets $\mathcal{D} = (x_l^h, y_i)$, where $y_i = \mathbf{1}\{\text{answer is truthful}\}$ by labeling each activation with $y = 1$ when truthful answers are given and $y = 0$ for hallucinated ones. Each dataset is randomly split into training and validation sets with a 4:1 ratio. Probes are then trained on the training sets with a binary classification objective.

We evaluate this approach on the Vicuna-7B model (Chiang et al., 2023) using the HaluEval dataset (Li et al., 2023a) , which is specifically designed to measure sensitivity to contextual grounding. Probes are trained across all 32 transformer layers and their associated heads, enabling a fine-grained analysis of head-level contributions to truthful outputs.

As shown in 1, the truthfulness score can be observed for each layer and head according to the image. For instance, the sixth head of the first layer demonstrates a high capability for truthfully grounding the given context.

2.3 FINE-TUNED MLLMS INHERIT TRUTHFUL REASONING FROM FOUNDATIONAL LLMS.

To examine whether the phenomenon of truthfulness heads identified in Large Language Models (LLMs) persists when these models are adapted into Multimodal Large Language Models (MLLMs), we extended the analysis presented in Section 2.2. Specifically, we asked:

To what extent do truthfulness heads remain consistent when the same LLM is fine-tuned for multimodal objectives?

To address this, we evaluated representative MLLMs from two major model families: (i) LLaVA-1.5 and LLaVA-NeXT, both fine-tuned from Vicuna-7B (Chiang et al., 2023), and (ii) Qwen2.5-VL and Qwen2.5-Omni, both fine-tuned from Qwen2.5 (Qwen et al., 2025). This cross-family analysis allows us to test whether the inheritance of truthfulness heads generalizes beyond a single backbone architecture.

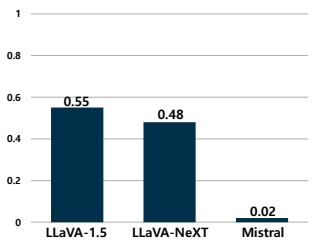


Figure 3: Cross-modal correlation with Vicuna-7B

We first investigated family-level correlations using the HaluEval dataset, where models must ground their predictions in the provided context. Since finetuned MLLMs are trained to process image tokens, we evaluated two conditions: (a) with identical textual inputs as their base LLMs, and (b) with an additional black image containing no informative content. As shown in Figure 1(b), LLaVA-1.5

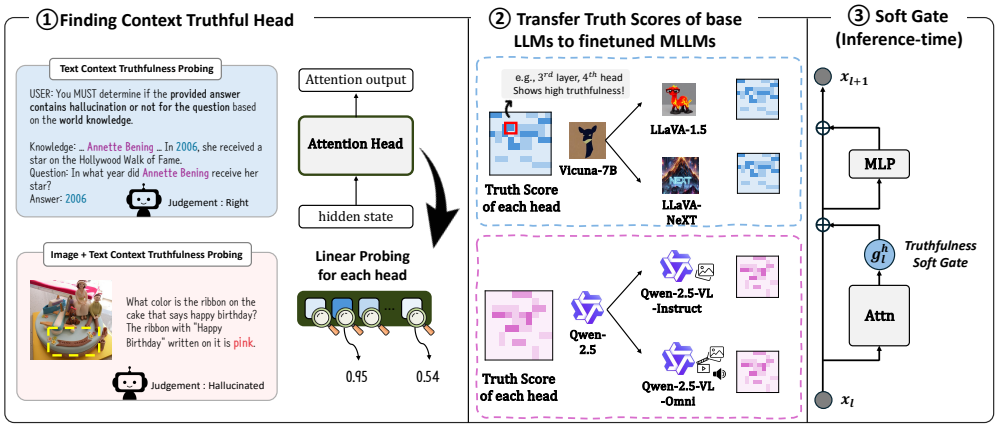


Figure 4: (1) Truthfulness scores of individual attention heads obtained via linear probing. (2) Cross-family similarity of head-level truthfulness scores when evaluated on models sharing the same baseline. (3) by applying the proposed soft gating mechanism, which adjusts head contributions based on their truthfulness scores.

and LLaVA-NeXT exhibited near-perfect correlation (≈ 1.0) with Vicuna-7B when given text-only inputs. Even when a non-informative image was introduced, the correlation remained consistently high (> 0.9), demonstrating robustness to added visual noise. For Qwen2.5-based models (Figure 1(c)), correlations were slightly lower than those of Vicuna but still substantial, exceeding 0.77 across conditions. Detailed head-level truthfulness scores for each model are shown in Figure 2.

Beyond single-dataset evaluations, we examined cross-modal correlations between the truthfulness scores of LLM heads (text-only) and their MLLM counterparts (with image context) as shown in Figure3. Using HaluEval (Li et al., 2023a) and PhD (text-context) (Liu et al., 2025) datasets, as well as RLHF-V (Yu et al., 2024) and PhD (image-context) (Liu et al., 2025) datasets, we found that within-family correlations remained consistently high (> 0.55), in sharp contrast to cross-family baselines such as Mistral, which showed negligible correlation (≈ 0.02). These results indicate that inheritance extends not only across datasets but also across modalities.

Taken together, our analysis shows that fine-tuned MLLMs preserve the structural role of truthfulness heads from their foundational LLMs. This inheritance holds even under multimodal adaptation and persists across both text- and image-grounded settings. These findings suggest that truthfulness heads represent a stable architectural property, providing a foundation for cross-family interventions aimed at improving contextual grounding and truthfulness.

3 Refining LVLMs towards Truthfulness

Building on the head-level analysis described in Section 2.2, we introduce a refinement strategy *truthprobe* that leverages the identified truthfulness scores of attention heads to guide model behavior. Specifically, our method amplifies the contribution of heads with consistently high truthfulness scores by increasing their additive influence in the residual connection, while attenuating or reweighting the contributions of less truthful heads. This selective adjustment ensures that the residual stream is more strongly shaped by context-faithful signals, thereby mitigating the effect of misleading activations. By systematically steering the residual pathway toward information from reliable heads, we aim to enhance the overall truthfulness of Multimodal Large Language Models (MLLMs).

3.1 Soft Head Gating for Truthfulness Amplification

To further refine the residual pathway with respect to context-faithful reasoning, we propose a soft gating mechanism that amplifies or attenuates the contribution of each attention head according to its estimated truthfulness score. Unlike hard masking, which discards information from untrusted

heads, our approach preserves the expressive capacity of multi-head attention (MHA) while softly steering the residual stream toward reliable signals.

Formally, in a Transformer layer l, the attention outputs of individual heads are aggregated as in equation 2. To apply the Truth Score as a soft gate, we take the projected attention before the residual connection, $o_l \in \mathbb{R}^d$, reshape it into head-wise components $\tilde{o}_l^h \in \mathbb{R}^{nh \times hd}$, and scale each by its corresponding gate value g_l^h. The gated representations are then concatenated back and added to the residual stream, thereby modulating each head's contribution according to its Truth Score:

$$x_{l+1} = x_l + \text{Concat}_{h=1}^{H}(g_l^h \cdot \tilde{o}_l^{(h)}), \tag{5}$$

$$g_l^h = 1 + \lambda \cdot \text{norm}(S), \tag{6}$$

Here, g_l^h denotes the soft gate for head h at layer l, parameterized by the normalized Truth Score S and scaled by a parameter λ. Specifically, when the norm-based score S is larger, the corresponding head output is amplified beyond the baseline level, whereas smaller values reduce its relative impact. This formulation enables the model to selectively strengthen more reliable heads while suppressing less informative ones. Importantly, the proposed soft gating mechanism ensures that all heads remain active; their influence on the residual connection is adaptively modulated in proportion to their truthfulness score, thereby preserving diversity while promoting context-faithful reasoning.

By embedding this gating mechanism into the residual update, the model effectively prioritizes trustworthy contextual cues without sacrificing the diversity of representations contributed by different heads. This design allows Multimodal Large Language Models (MLLMs) to more faithfully propagate context-grounded information and mitigates the propagation of misleading or hallucinated activations.

4 Experiments

4.1 Experimental Setting

Models. To investigate the transferability of truthfulness heads across model families, we focus on models that share a common backbone. Specifically, we use Vicuna-7B (Chiang et al., 2023) as the base LLM and evaluate its fine-tuned counterparts, LLaVA-1.5 (Liu et al., 2024a) and LLaVA-NeXT (Li et al., 2024). In parallel, we conduct experiments on the Qwen2.5 family, comparing the base Qwen2.5 (Qwen et al., 2025) model with its vision–language variants, Qwen2.5-VL-Instruct (Bai et al., 2025) and Qwen2.5-VL-Omni (Xu et al., 2025). This setup allows us to systematically analyze whether the identified truthful components remain consistent when adapted to multimodal tasks within the same architectural lineage.

Benchmarks. We evaluate our method on the Polling-based Object Probing Evaluation (POPE) benchmark (Li et al., 2023c). This benchmark is constructed from the MSCOCO (Lin et al., 2014), A-OKVQA (Marino et al., 2019), and GQA datasets (Hudson & Manning, 2019), and is specifically designed to assess whether Vision–Language Models (VLMs) correctly recognize objects in images. The evaluation is framed as a binary classification task, enabling the measurement of hallucination by testing whether the model provides truthful answers regarding object presence. We conduct experiments under three settings provided by POPE: *random*, *popular*, and *adversarial*.

Implementation Details. For the soft gating mechanism, we use scaling parameter λ and a normalization method to control the effect of the Truth Score. We maintain the consistent setting for each model across all benchmarks. For the MLLMs, we use min-max normalization and set $\lambda = 0.1$ for LLaVA-1.5, and $\lambda = 0.3$ for LLaVA-NeXT, Qwen2.5-VL-Instruct, Qwen2.5-VL-Omni. As for the LLMs in Tab. 1, we adopt a centered normalization, and $\lambda = 4.5$ for Vicuna-7B and $\lambda = 7.5$ for Qwen2.5.

Probing Dataset for Truthfulness-aware Head Gating. We employ the RLHF-V dataset to construct a probing dataset for the head gating experiment of MLLMs. The RLHF-V dataset provides sentence-level responses to image–question pairs, including model-generated incorrect responses and human-corrected responses. For the probing the MLLMs, we augmented the dataset by modifying the incorrect responses to match the sentence structure of the corresponding correct responses

	HaluEval			
Model	**Acc**	**F1**	**Prec**	**Rec**
Vicuna-7B	38.33	13.42	22.62	9.54
Vicuna-7B + Truth LLM	34.91	**44.65**	**38.89**	**52.4**
Qwen2.5	27.39	36.58	32.52	41.8
Qwen2.5 + Truth LLM	**36.5**	**41.07**	**38.38**	**44.17**

Table 1: **Validation of Truth Scores.** Comparison between vanilla LLM models and our truth-enhanced models (Ours) on the HALUEVAL benchmark, where Truth Score are obtained via Linear Probing.

using Qwen2.5-VL, a state-of-the-art vision-language model. This procedure yielded 292 incorrect-correct pairs that differ only in the final 1–2 words. We matched the number of MLLM probing data by also using 292 examples from the HaluEval dataset for our LLM probing experiments. For further details regarding the HaluEval dataset, please refer to the Appendix.

4.2 EVALUATION OF THE PROPOSED METHODS

Validation of Truth Scores. To validate the effectiveness of our proposed Truth Scores, we first analyze their impact on base LLMs. We obtain the Truth Scores for each model—Vicuna-7B and Qwen2.5—by performing linear probing on the HaluEval dataset as in Sec 2.2. These scores are then applied as a soft gate to the original base LLMs. We evaluate the models' truthfulness on the same HaluEval benchmark, which was used for probing. As demonstrated in Table 1, applying our method significantly enhances performance, with the models showing an improved ability to judge the truthfulness of given sequences. The results, especially for Qwen2.5, prove that our Truth Score effectively captures and enhances a model's truthful reasoning ability.

Transfer Truth Scores of base LLMs to finetuned MLLMs. Building upon our findings that the Truth scores of base LLMs and their finetuned MLLMs are highly correlated—even finetuned or proved with different modalities—we explored the direct transferability of these scores. We applied the Truth Scores obtained from the base LLMs (Vicuna-7B and Qwen2.5) as a soft gate to their corresponding finetuned MLLMs. Our experiments included LLaVA-1.5 and LLaVA-NeXT (finetuned from Vicuna-7B), as well as Qwen2.5-VL-Instruct and Qwen2.5-VL-Omni (finetuned from Qwen2.5). When evaluated on the POPE benchmark, which requires judging truthfulness from image and text inputs, we observed performance improvement over the vanilla models in most cases. This result suggests that Truth Scores of base LLMs can be effectively transferred to their finetuned MLLM counterparts. Furthermore, the performance gain from this transfer was found to be on par with the gain achieved by probing MLLMs on their own. This results highlights the potential for a unified approach: leveraging the Truth Scores from a single base LLM to enhance the truthfulness of multiple specialized MLLMs developed from the same foundation.

4.3 ABLATION OF ATTN HEAD GATING

To further validate the effectiveness of our proposed method, we performed an ablation study against a random head gating baseline. We used a baseline where the gating term $\lambda \cdot \text{norm}(S)$ in eq. 6 was replaced with a random value between -1 and 1. We assessed the performance of MLLMs—LLaVA-1.5 and LLaVA-NeXT with our method, which transfers its base LLM (Vicuna-7B)'s Truth Score as a soft gate, and the random head gate baseline using the POPE benchmark. As shown in Fig. 5, the random head gating method consistently leads to a notable decrease in performance than that of vanilla model. This degradation in performance indicates that randomly enhancing or suppressing head contributions disrupts the model's pretrained functions, particularly its ability of truthful reasoning for the given inputs. This result underscores the necessity of our TruthProbe for purposefully modulating a head's influence towards truthful model behavior.

Model	POPE(MSCOCO)			POPE(A-OKVQA)			POPE(GQA)		
	Acc	F1	Rec	Acc	F1	Rec	Acc	F1	Rec
LLaVA-1.5(Vanila)	**86.9**	**85.8**	79.1	**86.3**	**86.5**	87.8	**85.1**	85.3	86.1
LLaVA-1.5 + TruthProbe $_{\text{LLM}}$	86.8	**85.8**	**79.7**	86.0	86.4	88.8	85.0	85.3	**87.3**
LLaVA-1.5 + TruthProbe $_{\text{MLLM}}$	86.8	**85.8**	79.6	86.0	**86.5**	**89.1**	**85.1**	**85.4**	87.1
LLaVA-NeXT(Vanila)	87.7	86.5	78.8	87.4	87.4	86.8	86.6	86.4	84.9
LLaVA-NeXT + TruthProbe $_{\text{LLM}}$	**88.4**	**87.5**	**81.1**	87.7	87.9	**89.6**	86.6	86.7	87.5
LLaVA-NeXT + TruthProbe $_{\text{MLLM}}$	88.1	87.1	80.5	**87.8**	**88.0**	89.3	**86.8**	**86.9**	**87.6**
Qwen2.5-VL-Instruct(Vanila)	87.6	86.3	78.2	86.9	87.1	86.6	87.3	**87.1**	85.7
Qwen2.5-VL-Instruct + TruthProbe $_{\text{LLM}}$	**88.1**	**87.0**	79.6	**87.9**	**87.9**	**87.8**	**87.1**	**87.1**	**86.5**
Qwen2.5-VL-Instruct + TruthProbe $_{\text{MLLM}}$	**88.1**	**87.0**	**79.9**	**87.6**	87.6	87.2	86.8	86.8	86.1
Qwen2.5-VL-Omni(Vanila)	85.1	84.7	75.0	87.0	87.4	84.7	87.0	86.5	82.9
Qwen2.5-VL-Omni + TruthProbe $_{\text{LLM}}$	**87.5**	**86.2**	**78.1**	**87.7**	**87.6**	**87.3**	**87.5**	**87.4**	**86.7**
Qwen2.5-VL-Omni + TruthProbe $_{\text{MLLM}}$	87.3	85.9	77.3	**87.7**	87.5	86.3	87.3	87.1	85.6

Table 2: **Main Result.** TruthProbe $_{\text{LLM}}$ denotes the refinement process using head-level Truth Scores obtained from the base language models (Vicuna-7B and Qwen2.5). TruthProbe $_{\text{MLLM}}$ denotes the refinement process using Truth Scores derived directly from the corresponding MLLMs.

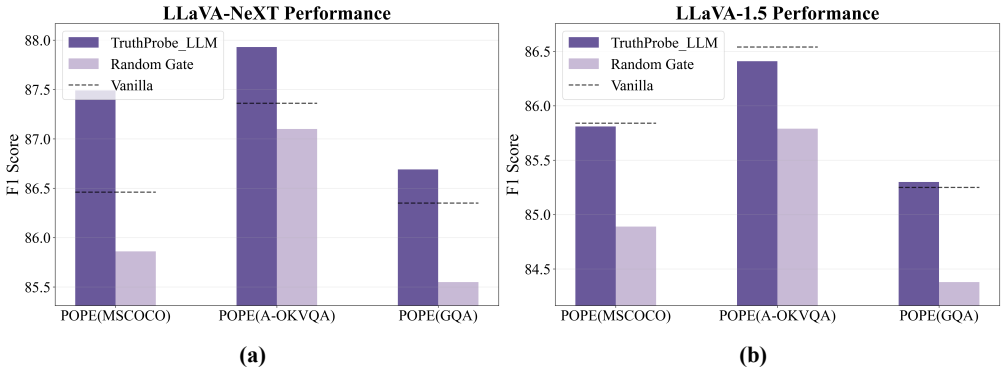


Figure 5: **Comparison of our method's effectiveness with Random Head Gating on POPE benchmarks.** Performance of the vanilla model is indicated by a dashed line. Compared to random head gating, our proposed TruthProbe gating consistently improves performance.

5 Discussions

5.1 Perspective of Model Families

Our findings demonstrate that the components responsible for truthfulness are not confined to a single model instance. Even after fine-tuning to adapt the backbone model to different modalities, the structural role of these components remains preserved. While our study primarily focused on identifying context-truthful heads, this invariance suggests that other well-studied head functions may exhibit similar stability across model families.

By establishing that truthfulness heads are both inherited and input-invariant, we provide a foundation for designing intervention strategies that generalize across related architectures. This opens the door for principled refinement approaches—such as soft gating—where interventions developed for one model can be seamlessly transferred to its variants. In real-world deployment, such cross-model stability not only reduces engineering overhead but also minimizes the risk of unintended behaviors, ultimately contributing to the development of safer and more interpretable LVLMs.

6 Related Works

6.1 Hallucination Mitigation in Large Vision-Language Models

Hallucination in LVLM refers to the generation of text that is inconsistent with the visual input, and numerous studies have analyzed its causes and proposed methods to address it. For example, LURE (Zhou et al., 2024) investigates several underlying factors of hallucination, including statistical bias introduced during pre-training—which can lead to the model's over-reliance on intrinsic knowledge or modality bias—uncertainty in token generation probability, and the positional bias of generated tokens in auto-regressive models. To mitigate these problems, some studies (Deng et al., 2024; An et al., 2025; Huo et al., 2025; Wang et al., 2025) employ contrastive decoding to improve the reliability of LVLMs; for instance, VCD (Leng et al., 2024) leverages the distributional differences between distorted and clean images to reduce distributional bias and suppress hallucination. On the other hand, training-based approaches (Yang et al., 2025; Sarkar et al., 2025) involve training dedicated modules to alleviate hallucination during inference. However, these approaches either overlook visual attention patterns in LVLMs or require substantial additional training data, which results in higher computational costs.

6.2 Attention-focused Methods

Owing to the transformer-based architecture of LVLMs, many recent studies have examined their attention mechanisms. Since LVLMs must effectively integrate visual information, several works have proposed modifying the model's attention distribution to mitigate hallucinations. Prior research suggests that an over-allocation of attention to textual input can induce hallucinations, motivating approaches that amplify visual attention (He et al., 2025; Zhou et al., 2025). For example, PAI (Liu et al., 2024b) demonstrates that increasing the allocation of attention to visual tokens can effectively reduce hallucinations. In addition, LVLMs often exhibit the attention sink phenomenon, where certain tokens receive disproportionately high attention regardless of their relevance, which is also linked to hallucinations. To address these issues, recent approaches (Kang et al., 2025) introduce adaptive mechanisms that reallocate attention more effectively toward visual tokens.

6.3 Attention Head

Since the transformer architecture consists of multiple attention heads and layers, several studies have demonstrated that each layer and head plays a distinct role in Large Language Models (LLMs) (Zheng et al., 2024). Some works identify the role of attention heads via linear probing (Li et al., 2023b), which trains linear classifiers to distinguish their functions. On the other hand, other studies (Wu et al., 2025; Yu et al., 2025) design custom scoring functions based on criteria such as attention weights or task-specific performance to characterize the roles of different heads. In the field of Vision-Language Models (VLMs), a growing body of research has also aimed at identifying attention heads that are particularly associated with visual information (Bi et al., 2025; Nam et al., 2025).

7 Conclusion

Our analysis reveals that truthfulness heads identified in Large Language Models (LLMs) are consistently inherited by their fine-tuned Multimodal Large Language Models (MLLMs), maintaining strong correlations across modalities and datasets. Leveraging this property, we introduced a soft head gating mechanism that amplifies context-faithful heads, improving grounding and reducing hallucination without losing complementary signals. Experiments on HaluEval and POPE benchmarks confirmed that truthfulness scores from base LLMs can be directly transferred to their multimodal descendants, achieving comparable gains to probing MLLMs themselves. These results establish truthfulness heads as a stable and transferable inductive bias, enabling unified interventions to enhance the reliability of both LLMs and MLLMs.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR*, 2017.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *CVPR*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. Technical report, Qwen Team, 2025.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *CVPR*, 2025.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *ICLR Workshop*, 2024.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. *ACL*, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *ICLR*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *ICLR*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv:2407.07895*, 2024.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP*, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Neurips*, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset. In *CVPR*, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*, 2024b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. *arXiv preprint arXiv:2505.13737*, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. Technical report, Qwen Team, 2025.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *ICLR*, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *ICLR*, 2025.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *ICLR*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *ICLR*, 2025.

Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seongho Joe, Taehee Lee, Youngjune L Gwon, and Sungroh Yoon. Correcting negative bias in large language models through negative attention score alignment. *NAACL*, 2025.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, 2024.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *ICLR*, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *ICLR*, 2024.

A Appendix

B Additional Experimental Details

B.1 Linear Probing Data for LLMs

We used two different datasets, HaluEval (Li et al., 2023a) and PhD (Liu et al., 2025), for LLM linear probing. HaluEval is a benchmark designed to evaluate hallucination in LLMs, comprising four components: knowledge, question, hallucinated answer, and right answer. Here, knowledge serves as a query for answering the given question. The LLM is evaluated by selecting the correct answer from the two provided options.

PhD is a VLM hallucination benchmark consisting of three tasks: visual ambiguity, incorrect context(icc), and counter common sense (ccs). The visual ambiguity task examines the capability of VLMs to leverage visual modality under ambiguous image inputs for vision question answering (vqa). Incorrect context task provides inconsistent textual and image modalities, requiring the model to rely on only one modality for answering. Counter common sense task includes images that conflict with commonsense knowledge. Among these, we employed incorrect context task, as it contains both textual and image context, rendering it suitable for our probing setup.

Both datasets share a (context text, question, answer) structure. For HaluEval, we constructed balanced (knowledge, question, right answer) and (knowledge, question, hallucinated answer) pairs, 10,000 samples in total. Similarly, for PhD, we created a balanced dataset by selecting 5,000 samples each for (context, question, right answer) and (context, question, hallucinated answer). Since PhD's answers are originally image-based, the yes/no labels were inverted when adapting the dataset for LLM probing.

B.2 Linear Probing Data for MLLMs

Since MLLM probing requires datasets consisting of multiple modalities, we selected two datasets: PhD (Liu et al., 2025) and RLHF-V (Yu et al., 2024). As described above, PhD provides three evaluation tasks, and we employed the incorrect context (icc) task for probing. This setting shares the same questions as LLM probing but introduces a different modality, making it suitable for multimodal evaluation.

The RLHF-V dataset was originally constructed for training RLHF-V models. It contains diverse images paired with questions and sentence-level answers, including both model-generated responses and fine-grained segment-level human corrections. Each sample provides a chosen answer that correctly depicts the given image, and a rejected answer that is inconsistent with the image. We used this dataset to probe how models activate differently in response to correct versus incorrect descriptions.

As both datasets share the (image context, question, answer) structure, we constructed MLLM probing datasets in a manner consistent with the LLM probing setup. For PhD, we created a balanced dataset of 10,000 samples, comprising (image, question, right answer) and (image, question, hallucinated answer) pairs. For RLHF-V, we similarly built balanced (image, question, right description) and (image, question, hallucinated description) pairs. To avoid confounding effects from overly long responses, we restricted RLHF-V to question-answering samples only, resulting in 2,726 instances.

C Linear Prober Training Details

We adopt the linear probing methodology from the ITI paper (Li et al., 2023b). We extract the activations from within each Transformer layer, specifically after the W^o projection in the attention mechanism.

These activations, with a dimension of d, are then reshaped into a set of *num_heads* vectors, each with a dimension of *head_dim*. A dedicated linear layer (probe) with dimensions of (*head_dim* $\times$ 1) is attached to each head. The reshaped, head-specific vectors are passed through their corresponding probe to produce features. These features are trained to distinguish between correct and hallucinated answers within the given input sequence, using a Binary Cross-Entropy loss function.

Right Answer

Question: What star of Now You See Me was born in Oman?

Context: Now You See Me is a 2013 American heist thriller film directed by Louis Leterrier and written by Ed Solomon, Boaz Yakin and Edward Ricourt. The film features an ensemble cast of Jesse Eisenberg, Mark Ruffalo, Woody Harrelson, Mélanie Laurent, Isla Fisher, Dave Franco, Michael Caine, and Morgan Freeman.Isla Lang Fisher (; born 3 February 1976) is an Australian actress. Born to Scottish parents in Oman, she moved to Australia at age 6.

Answer: Isla Fisher

Hallucinated Answer

Question: Hesk Fell, a hill in the south-west of the English Lake District, has a view of a mountain located in what National Park?

Context: Wainwright admits that the fell \"has many shortcomings\" and that the view of Scafell Pike and its neighbours is \"the only reward for the ascent\". It is located in the Lake District National Park, in Cumbria, and is part of the Southern Fells.

Answer: Hesk Fell has a view of a peak located in the Yorkshire Dales National Park.

Figure 6: Example from the HaluEval dataset. Top (blue) shows a correct answer, bottom (red) shows a hallucinated answer.

Right Answer

Question: Is there a tall tree in front of the train in the image?

Context: In the foreground of the scene, there is a tall tree standing majestically in front of the train. Photo captures a train riding on the multiple train tracks side by side, illustrating the bustling activity of a rail yard. Amidst this, a blue train can also be seen traveling past a set of traffic lights, highlighting the integration of rail and road transport.

Answer: yes

Hallucinated Answer

Question: Is there a tall tree in front of the train in the image?

Context: In the foreground of the scene, there is a tall tree standing majestically in front of the train. Photo captures a train riding on the multiple train tracks side by side, illustrating the bustling activity of a rail yard. Amidst this, a blue train can also be seen traveling past a set of traffic lights, highlighting the integration of rail and road transport.

Answer: yes

Figure 7: Example from the PhD dataset. Top (blue) shows a correct answer, bottom (red) shows a hallucinated answer.

We trained the probers for 200 epochs using the AdamW optimizer. On a single A6000 GPU, the training process for approximately 10,000 data points took about 10-20 minutes for LLMs and 30-40 minutes for MLLMs.

D Gating Details

For our soft gating mechanism, we apply normalization to the Truth Scores for the heads within each layer. As mentioned in the main paper, the LLMs in our study use a centered normalization approach. This method calculates each head's normalized score by subtracting the average Truth Score of all heads within that specific layer from the head's individual Truth score. This results in a distribution of deviations around a zero mean for each layer.

We selected the optimal λ value and normalization strategy for each model by performing a grid search on a held-out validation set, which comprised 20% of the full dataset. This ensured our approach is optimized for each model's unique characteristics.

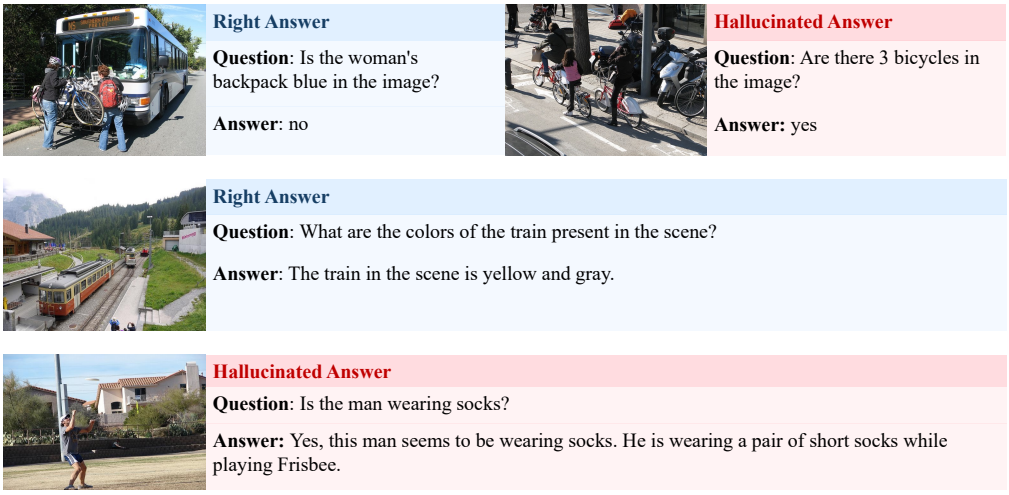


Figure 8: Examples from the MLLM probing datasets. Blue denotes a correct answer, while red denotes a hallucinated answer. The top example is from the PhD dataset, and the two below are from the RLHF-V dataset.

Model	POPE (MSCOCO)		
	Acc	F1	Rec
LLaVA-1.5	**86.9**	**85.8**	79.1
LLaVA-1.5 + TruthProbe $_{\text{LLM}}$	86.8	85.8	**79.7**
LLaVA-1.5 + Random Gate (3 Trials)	86.1 $\pm$ 0.18	84.9 $\pm$ 0.21	77.8 $\pm$ 0.28
LLaVA-NeXT(Vanila)	87.7	86.5	78.8
LLaVA-NeXT + TruthProbe $_{\text{LLM}}$	**88.4**	**87.5**	**81.1**
LLaVA-NeXT + Random Gate (3 Trials)	87.1 $\pm$ 0.08	85.8 $\pm$ 0.08	78.1 $\pm$ 0.1

Table 3: Performance comparison with TruthProbe vs. Random Head Gating on POPE (MSCOCO).

E Experimental Setup

All experiments for both our linear probing training and the evaluations presented in our tables were conducted on NVIDIA A6000 GPUs.

F Ablation Study: Full Results

Table 3 through 5 present the detailed ablation results for the LLaVA-1.5 and LLaVA-NeXT. We compare our TruthProbe method with a Random Head Gate baseline. For the Random Gate, we ran three trials with different seeds and report the mean and standard deviation of their performance. Across all evaluations, the Random Gate consistently underperforms the vanilla models, highlighting that arbitrarily modifying a head's contribution is detrimental to performance. This result confirms that our method's targeted approach is crucial for performance gains, as opposed to random manipulation.

Model	POPE (A-OKVQA)		
	Acc	F1	Rec
LLaVA-1.5	**86.3**	**86.5**	87.8
LLaVA-1.5 + TruthProbe $_{\text{LLM}}$	86.0	86.4	**88.8**
LLaVA-1.5 + Random Gate (3 Trials)	85.6 $\pm$ 0.12	85.7 $\pm$ 0.11	86.4 $\pm$ 0.07
LLaVA-NeXT(Vanila)	87.4	87.4	86.8
LLaVA-NeXT + TruthProbe $_{\text{LLM}}$	**87.7**	**87.9**	**89.6**
LLaVA-NeXT + Random Gate (3 Trials)	87.2 $\pm$ 0.07	87.1 $\pm$ 0.09	86.3 $\pm$ 0.22

Table 4: Performance comparison with TruthProbe vs. Random Head Gating on POPE (A-OKVQA).

Model	POPE (GQA)		
	Acc	F1	Rec
LLaVA-1.5	85.1	85.3	86.1
LLaVA-1.5 + TruthProbe $_{\text{LLM}}$	**85.0**	**85.3**	**87.3**
LLaVA-1.5 + Random Gate (3 Trials)	84.3 $\pm$ 0.28	84.3 $\pm$ 0.25	84.5 $\pm$ 0.10
LLaVA-NeXT(Vanila)	**86.6**	86.4	84.9
LLaVA-NeXT + TruthProbe $_{\text{LLM}}$	**86.6**	**86.7**	**87.5**
LLaVA-NeXT + Random Gate (3 Trials)	85.8 $\pm$ 0.04	85.5 $\pm$ 0.04	83.7 $\pm$ 0.16

Table 5: Performance comparison with TruthProbe vs. Random Head Gating on POPE (GQA).