The Truth Stays in the Family:
Enhancing Contextual Truthfulness via Inherited Heads in Model Lineages

Miso Choi[1], Seonga Choi[1], Mincheol Kwon[1], Woosung Joung[1], Jinkyu Kim[2], Jungbeom Lee[1]

[1] Korea University, Republic of Korea, [2] Kakao Mobility Corp., Republic of Korea

ICML International Conference On Machine Learning

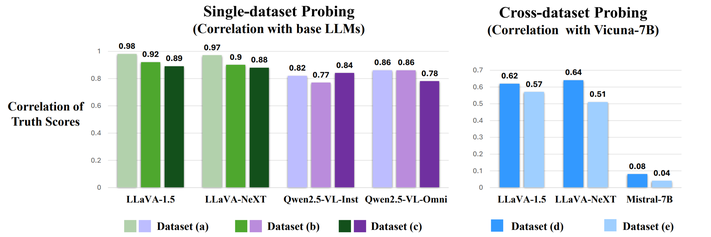
Motivation

- Existing hallucination mitigation methods often analyze or adapt each model separately.
- MLLMs (e.g., LLaVA-, Qwen2.5-VL) are derived from their base LLMs through multimodal finetuning.

Does a fundamental behavioral link exist between foundational LLMs and downstream variants?

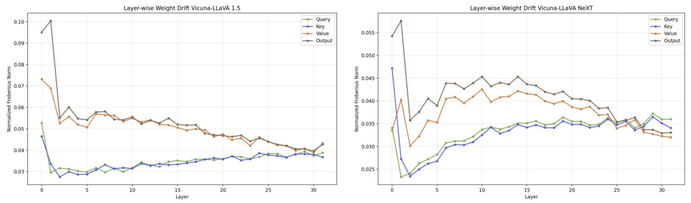
Our Findings: Truthful heads are inherited within families

- We define **Truth Score**, measured by applying linear probing at the head-level to quantify how much each attention head encodes context-truthful information.
- The Truth Score within the same family remain highly correlated, while unrelated model families have near-zero correlation.



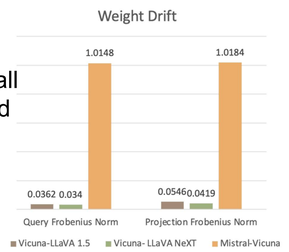
Mechanistic clue: Inheritance follows weight preservation.

1. Weight drift is concentrated in early layers.



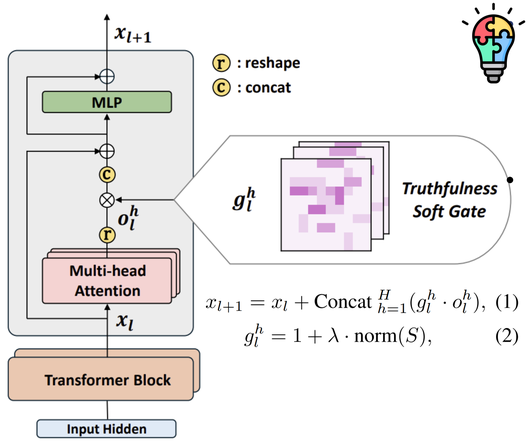
2. Top-20 truthful heads in LLaVA-1.5 appear in middle to deeper layers (10-31).

3. The weight difference is extremely small within the same family, whereas unrelated families diverge substantially.



TL;DR: Truthfulness-related heads are inherited within model families and can be leveraged to improve both LLMs and MLLMs.

TruthProbe: leveraging inherited Truth Scores as soft gates



$$x_{l+1} = x_l + \text{Concat}_{h=1}^{H}(g_l^h \cdot o_l^h), \quad (1)$$

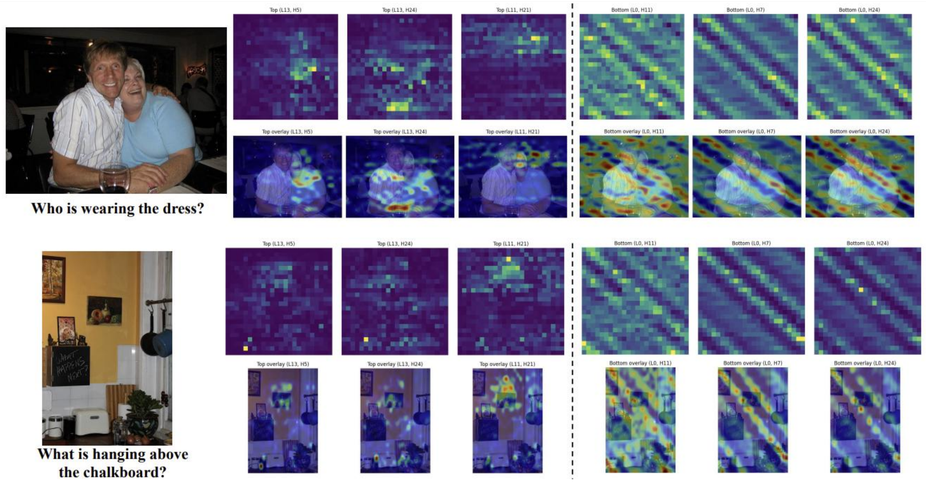
$$g_l^h = 1 + \lambda \cdot \text{norm}(S), \quad (2)$$

Probe once on the base LLM, then transfer obtained Truth Scores to derived LLMs / MLLMs

TruthProbe scales each attention-head output according to its Truth Score before the residual update.

TruthProbe as a lightweight plug-and-play gates that can be applied across multiple descendants in the same model family.

Analysis: Truthful heads attend to query-relevant evidence



Context-truthful heads attend to **query-relevant visual evidence**, whereas low-truthfulness heads tend to exhibit weakly semantic attention patterns

Arxiv | Github | Project Page

Experiments: improves contextual truthfulness on HaluEval

Model	HaluEval Acc	F1	Prec	Rec
Vicuna-7B (Chiang et al., 2023)				
Baseline	**38.89**	13.37	22.93	9.44
+ **TruthProbe**$_{\text{LLM}}$	38.53	**29.15**	**34.38**	**25.30**
Qwen2.5 (Qwen et al., 2025)				
Baseline	27.65	36.69	32.60	41.96
+ **TruthProbe**$_{\text{LLM}}$	**35.04**	**46.54**	**39.52**	**56.59**

Method	HaluEval Acc	F1	Prec	Rec
Qwen2.5-7B-Instruct (Bai et al., 2025)				
Baseline	34.90	16.29	22.79	12.68
+ **TruthProbe**$_{\text{Base LLM}}$	**37.35**	**17.24**	**25.36**	**13.05**
Vicuna-7B (Chiang et al., 2023)				
Baseline	38.89	13.37	22.93	9.44
+ **TruthProbe**$_{\text{Base LLM}}$	**48.47**	**57.17**	**48.90**	**68.82**

✓ On LLMs, TruthProbe improves contextual truthfulness on HaluEval.

✓ the Truth Score can be transferred from base LLM to its finetuned LLMs.

Experiments: reduces object hallucination on POPE/CHAIR

Method	POPE(COCO) Acc	F1	Rec	POPE(A-OKVQA) Acc	F1	Rec
LLaVA-1.5 (Liu et al., 2024a)						
Baseline	**86.9**	**85.8**	79.1	**86.3**	**86.5**	87.8
+ **TruthProbe**$_{\text{LLM}}$	86.7	**85.8**	**80.1**	85.7	86.3	**90.1**
+ **TruthProbe**$_{\text{MLLM}}$	86.8	**85.8**	79.6	86.1	**86.5**	89.0
LLaVA-NeXT (Li et al., 2024)						
Baseline	87.7	86.5	78.8	87.4	87.4	86.8
+ **TruthProbe**$_{\text{LLM}}$	**88.3**	**87.3**	**80.9**	**87.7**	**88.0**	**89.7**
+ **TruthProbe**$_{\text{MLLM}}$	88.2	87.2	80.1	**87.7**	87.9	89.5
Qwen2.5-VL-Instruct (Bai et al., 2025)						
Baseline	87.6	86.3	78.2	87.4	87.2	86.0
+ **TruthProbe**$_{\text{LLM}}$	**88.1**	**87.0**	79.9	**87.8**	**87.8**	**87.7**
+ **TruthProbe**$_{\text{MLLM}}$	**88.1**	**87.0**	**80.0**	87.7	87.7	87.4
Qwen2.5-VL-Omni (Xu et al., 2025)						
Baseline	85.1	84.7	75.0	87.0	87.4	84.7
+ **TruthProbe**$_{\text{LLM}}$	**87.3**	**86.0**	**77.7**	**87.8**	**87.8**	**87.1**
+ **TruthProbe**$_{\text{MLLM}}$	87.1	85.7	77.3	87.7	87.6	86.7

Method	CHAIR CHAIR_I (↓)	CHAIR_S (↓)
LLaVA-1.5 (Liu et al., 2024a)		
Baseline	6.99	23.00
+ **TruthProbe**$_{\text{LLM}}$	**5.36**	**17.40**
+ **TruthProbe**$_{\text{MLLM}}$	6.20	21.60
LLaVA-NeXT (Li et al., 2024)		
Baseline	6.91	13.40
+ **TruthProbe**$_{\text{LLM}}$	**4.94**	**11.20**
+ **TruthProbe**$_{\text{MLLM}}$	6.56	12.60
Qwen2.5-VL-Instruct (Bai et al., 2025)		
Baseline	6.14	13.20
+ **TruthProbe**$_{\text{LLM}}$	5.56	12.20
+ **TruthProbe**$_{\text{MLLM}}$	**5.26**	**7.80**
Qwen2.5-VL-Omni (Xu et al., 2025)		
Baseline	**5.26**	11.40
+ **TruthProbe**$_{\text{LLM}}$	5.94	**10.80**
+ **TruthProbe**$_{\text{MLLM}}$	5.54	11.00

✓ TruthProbe generally improves object-presence reasoning and reduces hallucination in generation setting.

✓ A key point is that base-LLM Truth Scores perform comparably to MLLM-probed scores, suggesting that probing a base LLM can provide useful gates for multiple downstream variants.