Lightweight Hallucination Detection via Semantic Token-Group Logit Trajectories

Seongmin Lee
The University of Tokyo
lee-s@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science, The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract

Although LLMs often generate coherent but incorrect responses, the internal mechanisms of such hallucinations remain unclear. While recent work probes LLMs' full hidden states for truthfulness, processing these thousand-dimensional representations is computationally expensive and lacks direct interpretability. We instead analyze semantic logit trajectories, the layer-wise evolution of logits for contrasting target token groups (*e.g.*, True/False). This approach unveils interpretable patterns with significantly reduced feature dimensionality. Our results demonstrate that a lightweight MLP trained on these trajectories effectively predicts hallucinations, offering a consistent and efficient alternative to raw hidden-state features.

1 Introduction

Large Language Models (LLMs) increasingly suffer from hallucinations, which are responses that appear coherent but are factually incorrect [1, 2, 3]. As these models are deployed in knowledge-intensive and high-stakes applications, ensuring the reliability of generated content is paramount. However, addressing this challenge requires more than just high detection accuracy. A practical detection mechanism must be computationally lightweight to enable real-time monitoring without introducing latency, and sufficiently interpretable to provide insights into the model's internal uncertainty, rather than merely treating the generation process as an opaque black box.

Recent research attempts to open this black box by probing internal hidden states. Studies have shown that intermediate representations encode distinct signals for truthfulness and uncertainty [4, 5]. For instance, supervised approaches like SAPLMA [6] train linear classifiers on hidden vectors to distinguish factual statements from hallucinations. Parallel efforts explore unsupervised methods to elicit latent knowledge [7] or detect deviations in internal states without labeled data [8, 9]. However, relying on these high-dimensional vector features presents significant limitations: (1) architecture-specificity hinders generalization; (2) training probes is computationally expensive; and (3) focusing on isolated layers misses the dynamic emergence of hallucinations.

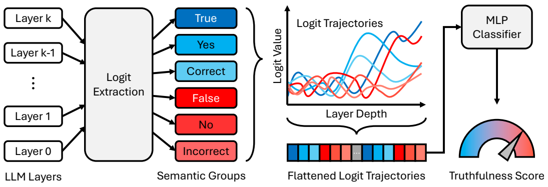

Figure 1 Overview of our proposed framework. We extract layer-wise logits for affirmative (blue) and negative (red) semantic groups to construct **Logit Trajectories**. These dynamic patterns are flattened into a feature vector and processed by a MLP classifier to estimate the **Truthfulness Score** of the statement.

In this study, we focus on binary verification tasks and propose a simpler, interpretable alternative: analyzing **semantic logit trajectories**. We track the layer-wise evolution of logits corresponding to target token groups (*e.g.*, True vs. False). This approach illuminates the internal flow of confidence, revealing how evidence for these targets accumulates or reverses, offering compressed yet informative features.

We validate our approach on the True-False dataset [6] across Llama3 (3.1-8B, 3.2-1B/3B) [10, 11] and Qwen (2.5-3B/7B, 3-4B) [12, 13]. Comparing against a strong baseline, SAPLMA [6], our lightweight trajectory MLP achieves competitive accuracy while drastically reducing the dimensionality of input features (*e.g.*, 4096 → 192 for Llama-3.1 8B). This confirms confidence evolution patterns as a robust foundation for efficient hallucination detection.

2 Method

Hallucinations in LLMs often stem from complex failures in knowledge recall or reasoning, which are not localized to a single layer but are distributed across various depths of the network [14]. Consequently, capturing these errors requires monitoring the model's internal dynamics across its entire depth, rather than relying on a static snapshot from a specific layer. However, directly aggregating high-dimensional hidden states from all layers introduces prohibitive computational costs, rendering the detector too heavy for real-time applications.

To resolve this dilemma, we propose a lightweight alternative: tracking **semantic logit trajectories**, defined as the layer-wise evolution of probability masses assigned to **target token groups**. By compressing the complex, multi-layer decision-making dynamics into low-dimensional logit patterns, our approach captures essential signals of hallucination, such as hesitation or conflict, while maintaining high computational efficiency.

2.1 Logit Trajectory Extraction

Following the Logit Lens approach [15], we project intermediate hidden states directly into the vocabulary space to inspect the model's internal confidence evolution. For a transformer with L layers, we extract the hidden state $h_\ell \in \mathbb{R}^d$ from each layer ℓ. In this work, we define h_ℓ as the output of the ℓ-th transformer block, taken after the final feed-forward network and residual connection. We project the normalized hidden states onto the vocabulary space using the unembedding matrix W_U to obtain the full vocabulary distribution:

$$z_\ell = \mathrm{RMSNorm}(h_\ell) W_U^\top \in \mathbb{R}^{|V|}. \tag{1}$$

Our method focuses on binary verification tasks where the model is strictly prompted to output a definitive binary response (specifically "True" or "False"). Despite this strict constraint, the model's internal confidence is often distributed across semantically related concepts. To capture this co-activation, we monitor six distinct semantic groups:

- Affirmative indicators: **True, Yes, Correct**
- Negative indicators: **False, No, Incorrect**

Beyond semantic synonyms, tokenization often fragments a single concept into multiple surface forms due to case sensitivity and leading whitespace. Therefore, for each target word w in our list, we aggregate its morphological variants (*e.g.*, case, spacing) into a unified group G_w. Crucially, we treat the six semantic concepts independently; we do *not* merge "True" and "Yes" into a super-group, but rather track their individual trajectories to preserve granular internal dynamics.

Table 1 Example of surface-form variants aggregated for the "True" group. We apply this same aggregation strategy to all six selected concepts (True, Yes, Correct, False, No, Incorrect).

Category	**Tokens in Group ($G_{\mathbf{True}}$)**
Title Case	`True`, `_True`
Upper Case	`TRUE`, `_TRUE`
Lower Case	`true`, `_true`

Table 1 illustrates the aggregation logic using the "True" group as an example. The same logic applies to the other five groups, collecting all tokenizer variants for each respective concept.

For each group G, we compute the representative logit at layer ℓ:

$$z_\ell(G) = \frac{1}{|G|} \sum_{t \in G} z_\ell(t). \tag{2}$$

Computing this for all six semantic groups yields a **6-dimensional representative logit vector** at each layer. This aggregation drastically compresses the feature space from the original high-dimensional hidden states (*e.g.*, $d_{model} = 4096 \rightarrow 6$), significantly reducing computational overhead while preserving essential semantic signals.

2.2 Detection via Logit Trajectories

To predict hallucination risks based on the extracted internal dynamics, we employ a multi-Layer Perceptron (MLP). The input to the classifier is the flattened trajectory feature vector $x \in \mathbb{R}^{6k}$, constructed by concatenating the representative logits of the six semantic groups from the first layer up to a target layer k.

The network processes these features through ReLU-activated hidden layers to output a scalar probability $\hat{y}$ via a final sigmoid activation, representing the estimated likelihood that the statement is True.

3 Experimental Setup

We evaluate whether our trajectory-based classifier can achieve performance comparable to an existing

Table 2 Classification accuracy for hallucination detection across Llama and Qwen model families. *SAPLMA* uses hidden-state probes at the best-performing layer, while *Ours* uses a lightweight MLP over semantic logit trajectories concatenated across layers. *Model Output* denotes the accuracy of the model's **actual** responses (baseline without classifier).

Model	Llama3.2-1B	Llama3.2-3B	Llama3.1-8B	Qwen2.5-3B	Qwen3-4B	Qwen2.5-7B
Model Output	0.5081	0.8192	0.8998	0.8289	0.8973	0.8910
SAPLMA (Hidden States)	0.7503 (Layer 11)	0.8673 (Layer 15)	0.9047 (Layer 16)	0.8847 (Layer 26)	0.9126 (Layer 24)	0.9076 (Layer 26)
Ours (Logit Trajectories)	0.7385 (Layer 0-15)	0.8642 (Layer 0-20)	0.9069 (Layer 0-20)	0.8647 (Layer 0-24)	0.9078 (Layer 0-28)	0.8985 (Layer 0-25)

method that directly leverages high-dimensional hidden-state probes [6]. To ensure robust generalization, we conduct experiments across multiple model families using a strict leave-one-domain-out protocol.

Dataset. We use the **True-False dataset** [6], comprising 6,084 statements across six domains (*e.g.*, Cities, Scientific Facts). False statements are generated via entity swapping or negation. Following the original protocol, we employ a **leave-one-domain-out** split for evaluation, holding out one domain for testing while training on the others to assess generalization capability.

Models. We evaluate our method across diverse open-source instruction-tuned models, including **Llama3.1-8B-Instruct** [10], **Llama3.2-1B/3B-Instruct** [11], **Qwen2.5-3B/7B-Instruct** [12], and **Qwen3-4B-Instruct** [13]. All experiments use greedy decoding to ensure deterministic outputs. To analyze internal logit formation consistent with our method (§ 2), we use a strict instruction template forcing a binary response:

> ***system:*** *You are a strict logic engine. You must answer with exactly one word: 'True' or 'False'. Do not provide any explanation.*
> ***user:*** *Statement: {proposition}*
> *Is this statement True or False? Answer:*

Baseline. We compare our method against **SAPLMA** [6], a standard supervised probe trained on raw hidden states. The probe employs a feedforward neural network with three hidden layers of decreasing sizes (256, 128, and 64) and ReLU activations. To ensure a fair comparison, we utilize this identical MLP architecture for our trajectory-based classifier as well. We strictly follow the original training setup (*e.g.*, 5 epochs, Adam optimizer) to serve as a representative high-dimensional baseline.

Metrics. Because the task is balanced binary classification, we report classification **accuracy** as the primary metric, following prior work. Additional analysis includes measuring layerwise logit gaps and visualizing semantic trajectories.

4 Results

We analyze semantic token-group logit trajectories to understand internal confidence formation and evaluate their utility for hallucination detection compared to hidden-state baselines.

4.1 Hallucination Prediction

Table 2 compared our trajectory-based classifier against the hidden-state-based SAPLMA baseline across both Llama and Qwen model families.

Results show that semantic logit trajectories offer robust hallucination detection across Llama and Qwen architectures. In the Llama series, our method rivals the baseline as capacity increases, even slightly outperforming SAPLMA on Llama3.1-8B (0.9069 vs. 0.9047). Similarly, in Qwen models, our approach consistently surpasses raw outputs and achieves comparable accuracy to the heavy probe (within 0.9-2.0%).

Crucially, by using an **identical MLP architecture** for both methods, we confirm that this competitive performance stems not from classifier capacity but from the **high informational density** of the trajectories. This proves that despite significant compression, trajectory features preserve the essential dynamics required for effective detection.

Beyond predictive performance, a key advantage of our approach is the drastic reduction in feature dimensionality. While SAPLMA requires processing raw hidden vectors (*e.g.*, $d_{model} = 4096$ for Llama-3.1 8B), our trajectory input consists only of $6 \times L$ dimensions (*e.g.*, $6 \times 32 = 192$).

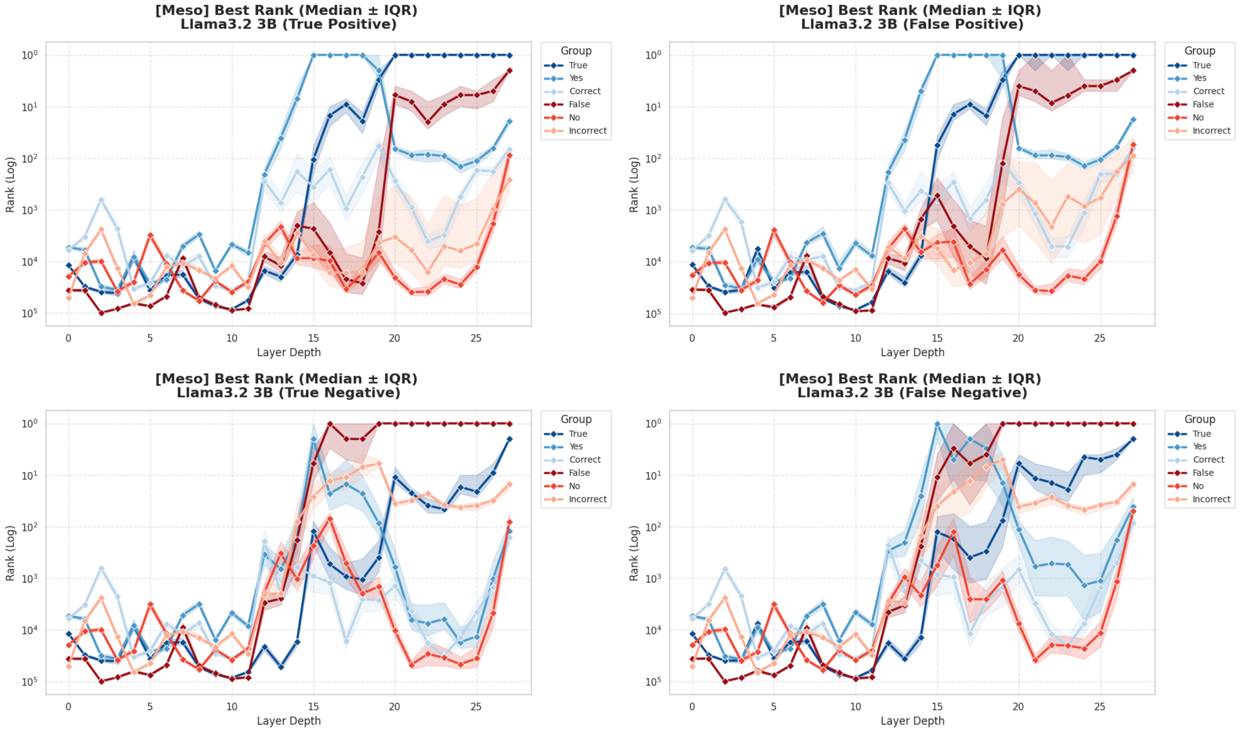

Figure 2 Layer-wise rank evolution of the highest-scoring token within each semantic group (Llama-3.2 3B). **Left (TP, TN):** Correct predictions show sharp divergence. **Right (FP, FN):** Hallucinations exhibit narrower gaps and higher volatility (wider IQR shading).

This represents a compression ratio of over 95%, significantly reducing the memory and computational overhead required to train and deploy the detector compared to full hidden-state probes.

4.2 Interpreting Logit Trajectories

Figure 2 visualizes the layer-wise rank evolution of semantic groups. Our analysis reveals consistent patterns distinguishing factual generations from hallucinations.

Intrinsic Affirmative Bias. We observe a consistent **intrinsic affirmative bias** across models. The affirmative group (*e.g.*, True) tends to rise earlier than negative concepts, regardless of the final truth value. This suggests the model defaults to affirmation and requires additional depth to suppress this signal when generating negative responses.

The Logit Gap and Signal Instability. Comparing factual versus hallucinatory trajectories reveals distinct dynamics in confidence formation:

Narrower Logit Gap: In correct predictions (Figure 2, left column), the target group rapidly ascends to Rank 1, creating a wide gap against the suppressed non-target group. In contrast, hallucinations (right column) exhibit a significantly **narrower gap**, where the opposing group lingers at relatively high ranks. This lack of clear separation reflects the model's internal hesitation.

High Variance in Non-Target Labels: Beyond gap magnitude, hallucinated instances show **significantly higher variance** (wider IQR shading) in the non-target group's trajectory. This indicates that the internal rejection mechanism remains volatile compared to the stable suppression observed in correct responses.

5 Conclusions

We proposed semantic logit trajectories as a lightweight, interpretable alternative to hidden-state probing. Our analysis reveals that hallucinations are characterized by narrower logit gaps and signal instability compared to factual generations. Experiments across Llama and Qwen architectures demonstrate that our method achieves detection accuracy comparable to high-dimensional baselines while significantly reducing computational overhead.

Future work will extend this approach from strict binary verification to open-ended generation to enable real-time hallucination monitoring in diverse scenarios.

References

[1] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2025.

[2] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, Vol. 43, No. 2, p. 1–55, January 2025.

[3] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 6449–6464, Singapore, December 2023. Association for Computational Linguistics.

[4] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

[5] Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen, editors, **Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**, pp. 88–104, Miami, Florida, US, November 2024. Association for Computational Linguistics.

[6] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics.

[7] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024.

[8] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms'internal states retain the power of hallucination detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, **International Conference on Representation Learning**, Vol. 2024, pp. 3056–3076, 2024.

[10] Meta AI. Introducing llama 3.1: Our most capable models to date, 2024.

[11] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024.

[12] 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, 2025.

[13] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[14] Xin Zhao, Zehui Jiang, and Naoki Yoshinaga. Neuron empirical gradient: Discovering and quantifying neurons' global linear controllability. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 21446–21477, Vienna, Austria, July 2025. Association for Computational Linguistics.

[15] Nostalgebraist. Interpreting gpt: The logit lens, 2020.