P^{2}LD: 隠れ状態差分による Few-shot ハルシネーション検知

井上 耕太朗
株式会社 SmartHR
kotaro.inoue@smarthr.co.jp

概要

大規模言語モデル（LLM）は質問応答や要約など多様なタスクで高性能を示す一方，流暢だが根拠のない出力（ハルシネーション）を生成しうる．出力ベースの検知に加え，内部表現を用いるプロービングは有効であるが，Mixture-of-Experts（MoE）モデルでは層ごとの表現が入力により変動し，単一層に依存する手法は層選択の不安定性を抱えやすい．またハルシネーションのラベル付きデータは収集コストが高く，Few-shot 条件では高次元特徴に起因する過学習が顕在化する．そこで本研究では，回答トークンの層間差分特徴に対し，各層で Robust scaling と PCA により冗長性を抑えた後，PLS 回帰により目的変数と共分散が大きい 1 次元潜在変数を抽出，全層を連結してロジスティック回帰で判定する PCA-PLS Layerwise Difference（P^{2}LD）を提案する．HaluEval の QA データセット上で 2 種の MoE モデルを評価した結果，全サンプル学習では従来手法（SAPLMA）と同等の性能を維持しつつ，Few-shot 条件で一貫して性能と学習安定性を改善した．

1 はじめに

LLM は質問応答や要約など多様なタスクで高性能を示す一方，流暢だが根拠のない出力 (ハルシネーション) を生成しうる．ハルシネーションは事実性を損なう重大なリスクであるため，その検知や抑制に関する研究が盛んに行われている [1, 2].

ハルシネーションには忠実性および事実性に関する大きく 2 つのカテゴリがあり [1]，特に事実性に関するものは利用者が気づきにくく，有害な場合が多い．事実性ハルシネーションの推論時の要因は主に 2 つある．1 つはモデルが知識を持たず捏造するケースで，RAG[3] や生成後の編集 [4] が有効な対策となる．2 つ目は，知識を保持していても適切に想起できなかったり，文脈や誘導により誤るケースである [5, 6]．モデルは自らの知識状態を完全には反映できず，確信のない内容を断定することもある [7]．こうしたケースでは，内部状態や生成過程の兆候を用いた検知が依然として重要である．

検知手法にはモデル出力のみを評価するブラックボックス型の手法に加え，内部表現で分類器を学習するホワイトボックス型の手法が提案されている [8, 9]．特に特定層の隠れ状態を用いる線形プローブは，計算効率が高く実装も容易である [10, 11]．しかし近年の LLM は MoE 化が進み [12, 13]，内部表現が入力により多様に変動することで，最適な層の選択が不安定になる懸念がある [14]．また，高次元の隠れ状態から学習するプロービングは少数のラベル付きデータでの過学習のリスクが高い．

そこで本研究は Few-shot 学習が可能なプロービング手法，P^{2}LD を提案し，MoE アーキテクチャの LLM を対象としてその有効性を検証する．提案手法は，層ごとの差分ベクトルを PCA により低次元に圧縮し [15]，さらに PLS により目的変数との共分散が大きい潜在変数を抽出する [16]．最後にこれらを層方向に連結し，ロジスティック回帰により適切なレイヤーへの重み付けを学習する．これらの工夫により提案手法は隠れ状態の全層情報を使うことで特定層の選択問題を回避しつつ，必要な学習パラメータ数を抑えることで，学習サンプル数が極めて限られる状況でも有効な手段となりうる．

提案手法は HaluEval[17] の QA データセットにおいて，従来手法である SAPLMA[9] と多角的な性能比較を行い，Few-shot 条件において大きな性能改善を示した．本研究の主な貢献は Few-shot 学習条件でありながらも，学習サンプルで過学習しづらい手法を明らかにした点にあり，これらの結果は信頼性のある LLM アプリケーションをより低コストかつ高速に提供する技術として寄与する．

2 Few-shot ハルシネーション検知

本論文ではあるクエリに対して得られた回答 a がハルシネーションを含むかを二値ラベル $y \in \{0, 1\}$

(1: hallucination) で判定する問題を扱う．提案手法および従来手法ともに生成済みの回答に対して内部表現を抽出し，軽量な分類器を学習後，未知のテストデータで y を推定する．

2.1 層別隠れ状態と層間差分の抽出

回答トークン列 $a = (t_1, \ldots, t_T)$ に対し，各層 $\ell \in \{1, \ldots, L\}$ の d 次元で構成される隠れ状態 $\mathbf{h}_\ell(t) \in \mathbb{R}^d$ を取得する．回答トークン位置における平均プーリングにより

$$\bar{\mathbf{h}}_\ell = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_\ell(t) \in \mathbb{R}^d \tag{1}$$

を得る．また近年は直接隠れ状態を用いず，層間のダイナミクスがより有効性の面で着目されており，本研究でも隣接層差分

$$\mathbf{\Delta}_\ell = \bar{\mathbf{h}}_{\ell+1} - \bar{\mathbf{h}}_\ell \in \mathbb{R}^d \qquad (\ell = 1, \ldots, L-1) \tag{2}$$

を導入する [18].

2.2 層ごとの潜在変数の抽出

各層の差分 $\mathbf{\Delta}_\ell$ に対し，外れ値の影響を抑制するため，特徴量ごとに中央値と四分位範囲を用いたスケーリングを適用する．スケーリング後の差分を $\tilde{\mathbf{\Delta}}_\ell$ とすると，次に PCA を用いて分散の 95%を保持する次元 d'_ℓ に射影する．

$$\mathbf{z}_\ell = \mathbf{P}_\ell^\top \tilde{\mathbf{\Delta}}_\ell \in \mathbb{R}^{d'_\ell} \tag{3}$$

ここで $\mathbf{P}_\ell$ は PCA 基底である．次に $\mathbf{z}_\ell$ を入力として PLS 回帰を学習し，目的変数 y と共分散が大きい 1 個の潜在変数 s_ℓ を抽出する．

2.3 多層統合と判定

全層の潜在変数を連結し，

$$\mathbf{s} = [s_1; \ldots; s_{L-1}] \in \mathbb{R}^{L-1} \tag{4}$$

連結した潜在ベクトル $\mathbf{s}$ を構成する．最終的に標準化を行った $\tilde{\mathbf{s}}$ に対し，L2 正則化 ($\lambda = 0.01$) を伴うロジスティック回帰

$$\hat{y} = \sigma(\mathbf{w}^\top \tilde{\mathbf{s}} + b) \tag{5}$$

を学習する．その際，ハルシネーション・ラベルの不均衡に対応するため，各クラスのサンプル数に反比例した重み付けを適用する．ロジスティック回帰は線形分離の枠内で解釈性が高く，正則化項を強めることで Few-shot 条件でも過学習を抑えやすい．

2.4 比較手法（SAPLMA）

SAPLMA は単一層の隠れ状態を入力とし，3 層の多層パーセプトロン（MLP）で真偽を判定するプローブである [9]．MLP は標準化した隠れ層の特徴量から 256, 128, 64 次元の ReLU 活性化の隠れ層を経てシグモイド出力を行う．最適化には Adam[19] を用い，5 エポックの短期間学習を行う．本論文では Few-shot 条件においてプロービングに適した層選択がサンプルにより変動しうる点に注目し，層選択の安定性も併せて分析する．

3 実験設定

生成モデルは MoE アーキテクチャの LLM を対象とするため Qwen3-30B-A3B-Instruct-2507(以降 qwen3) [20] および gpt-oss-20b(以降 gpt-oss) [21] を利用した．モデルごとの構成については表 1 に示す．

データセットには HaluEval の QA データセットを用い，各生成モデルで 5000 件のサンプルの回答を生成した．HaluEval にはハルシネーション例の回答と正解ラベルが用意されているが，提案手法が隠れ層出力の変化を特徴として利用する都合，生成モデルが本来生成し得ない文脈のトークンでの検出性能への影響を考慮し，本論文ではそれらを利用しない．そこで生成した回答に対する真偽ラベルは OpenAI の GPT-5.2 の LLM-as-a-judge によりデータセットの知識との整合性を判定したものを利用する．

評価指標は AUROC と AUPRC を中心とし，補助的に Accuracy, F1, マシューズ相関係数 (MCC) も計測した．全サンプル学習では 5000 サンプルのうち学習に 4000，テストに 1000 サンプル利用し，Few-shot 学習では学習のみ K=4/8/16/32 サンプルをハルシネーションの有無のラベルで 1:1 にバランスして学習し，テストは全サンプル学習時と同様である．

4 実験結果

4.1 全サンプル学習時の性能

表 2 に十分な学習データ数を用意できる状況における AUROC と AUPRC の比較を示す．両生成モ

表 1 生成モデルのアーキテクチャ．

Model	Total Params.	Active Params.	#Experts	#Layers
qwen3	30B	3B	128	48
gpt-oss	20B	3.6B	32	24

デルで AUROC は提案手法が僅かに上回った一方，AUPRC は僅差で下回る．実用上はほぼ同等性能と見られ，全サンプル学習時の性能の有意な優劣は見られなかった．

表 2　全サンプル学習時（$N = 4000$）の性能比較．

Model	Method	AUROC	AUPRC
qwen3	Proposed	**0.938**	0.776
	SAPLMA	0.927	**0.786**
gpt-oss	Proposed	**0.922**	0.703
	SAPLMA	0.903	**0.718**

4.2　Few-shot 学習時の性能

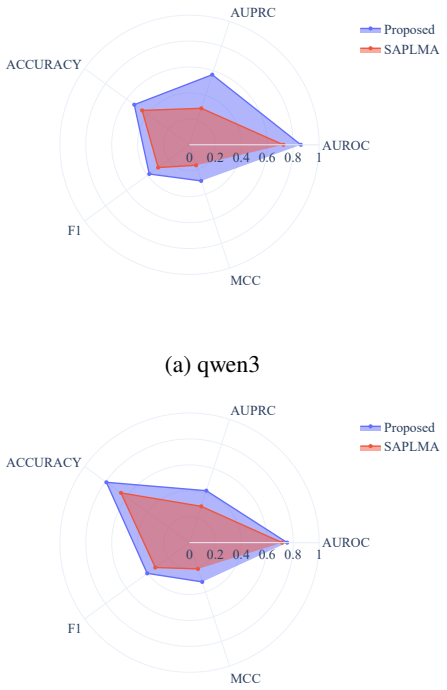
表 3 に Few-shot 学習時の性能比較を示す．Few-shot 学習では学習サンプルで性能のブレが大きいため，学習サンプルは再現性のあるシード値変化によって 5 回ランダムサンプリングし，合計 5 つのモデルで全サンプル学習と同様のテスト 1000 サンプルを評価した結果の平均を示している．提案手法は全学習サンプル数で SAPLMA を上回り，特に小サンプル数で大きく改善している．図 1 に学習 4 サンプルにおける AUROC, AUPRC, Accuracy, F1 値, MCC の平均をレーダーチャートによる可視化を示す．レーダーチャートより極端にサンプル数が限られた状況においても提案手法は多角的に従来手法を上回る性能であることがわかる．

表 3　Few-shot 学習における 5 回学習の平均性能（括弧内は標準偏差）．

		AUROC		AUPRC	
Model	K	Prop.	SAPLMA	Prop.	SAPLMA
qwen3	4	**0.86 (.03)**	0.72 (.05)	**0.57 (.09)**	0.30 (.07)
	8	**0.85 (.02)**	0.70 (.11)	**0.51 (.08)**	0.28 (.09)
	16	**0.87 (.01)**	0.79 (.05)	**0.54 (.07)**	0.42 (.13)
	32	**0.89 (.01)**	0.85 (.04)	**0.61 (.08)**	0.54 (.12)
gpt-oss	4	**0.78 (.09)**	0.71 (.06)	**0.44 (.09)**	0.29 (.06)
	8	**0.78 (.09)**	0.74 (.07)	**0.41 (.13)**	0.34 (.08)
	16	**0.85 (.02)**	0.82 (.04)	**0.51 (.03)**	0.47 (.05)
	32	**0.84 (.02)**	0.81 (.04)	**0.52 (.04)**	0.47 (.06)

5　考察

提案手法が Few-shot で強い理由として，全隠れ層を対象とした層選択問題の回避と，目的に即した低次元の潜在変数を PCA と PLS で層ごとに抽出することで，全層対象としながらも学習パラメータ数を圧倒的に削減している点が挙げられる．各層ごとの潜在変数は 1 次元にしており，ロジスティック回帰



(a) qwen3

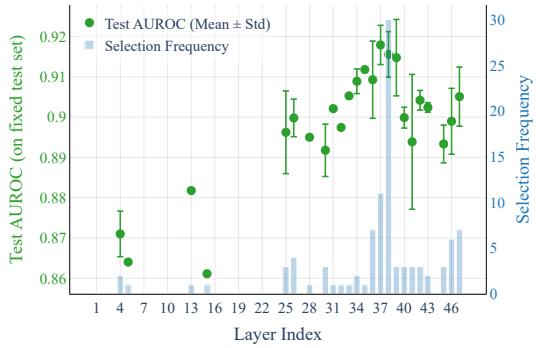
(b) gpt-oss

図 1　Few-shot（$K = 4$）でのレーダーチャート．

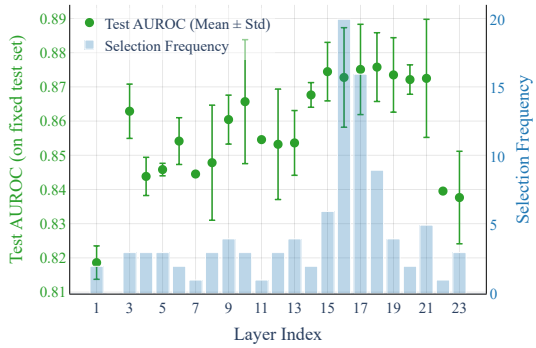
が学習するパラメータ数は隠れ層のレイヤー数より 1 少ない数で済む．特に K が小さいほど改善幅が大きいことは，この設計が Few-shot 耐性に寄与していることを示唆する．

5.1　プロービングにおける層選択問題

提案手法が全層を解析対象とする合理性は，層選択における学習安定性の観点から確認することができる．図 2 に，100 個のランダムサンプル（ハルシネーション比 1:1）を用いた SAPLMA の 100 回試行における層選択の分布と，テストデータでの AUROC の推移を示す．層の選択は各試行の 5 分割交差検証における最高 AUROC に基づいている．可視化結果より，100 サンプル程度の学習データであっても層選択の変動は大きく，選択頻度が最も高い層がテストデータで最良の汎化性能を達成するとは限らないことがわかる．極端な Few-shot 条件ではこの不安定性が一層増大し，単一層のみを利用するプローブの限界が露呈すると考えられ，全層の情報を統合して利用する提案手法の導入動機を強く支持する結果である．

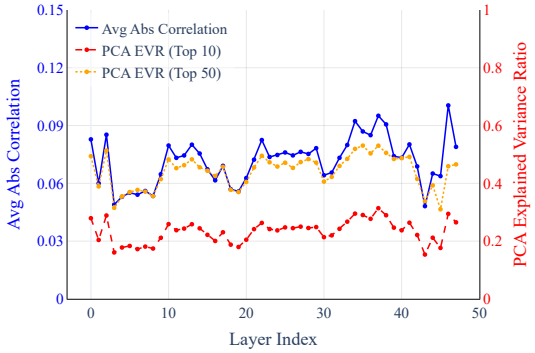


(a) qwen3

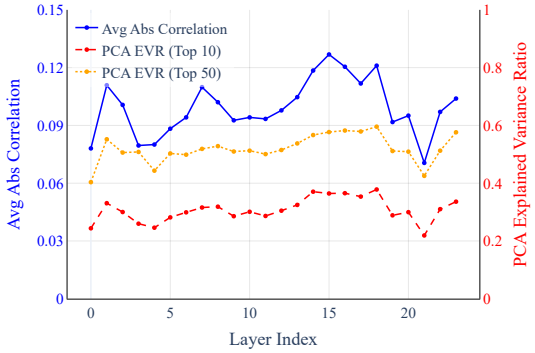


(b) gpt-oss

図 2　SAPLMA の学習によって選ばれる最良層の分布と AUROC の可視化.



(a) qwen3



(b) gpt-oss

図 3　層間差分の特徴独立性と PCA の説明分散比の層別可視化.

5.2　特徴分布と相関構造

提案手法の PLS は目的変数に寄与する潜在変数を抽出できるが，Few-shot 条件では情報量の少ない特徴に過適合する懸念がある．そこで，教師なし手法である PCA を前段に置き，目的変数と独立して情報量を最大化する空間へ射影することで過学習リスクを低減する．この設計の妥当性を特徴分布と相関構造の観点から検証する．

PCA による正規分布性の改善と変数間の独立性から確認する．入力が高次元かつ独立であれば，中心極限定理により PCA 出力は正規分布に近づき，PLS の重み付けが安定すると期待されるためである．各層および PCA 後の特徴量に対し Shapiro-Wilk 検定 [22] を行ったところ，正規分布とみなせる次元の割合は，gpt-oss で PCA 前後にて平均 12.6%から 12.1%と変化がなかったが，qwen3 では 6.2%から 9.1%へ改善した．この低い正規性は，隠れ層がガウス分布に従わず，裾が長い外れ値を含む可能性を示唆している．またこれは標準化の適用が不適切であることを意味し，提案手法で中央値と四分位範囲によるスケーリングを採用した根拠である．

次に層間差分特徴量の独立性をピアソン相関係数の非対角成分平均で評価した．図 3 の通り，平均絶対相関は qwen3 で 0.072（最大 0.100），gpt-oss で 0.100（最大 0.127）と極めて小さく，特徴間の相関は限定的である． また少数の主成分だけでは説明分散を回収しきれないことから，情報は広く分散していると言える．これらは PCA で冗長性を抑えつつ，PLS で有効な方向を探索する二段構成の妥当性を支持する結果である．

6　おわりに

本研究では隠れ状態差分で Few-shot 学習可能なハルシネーション検知手法を提案した．提案手法は極めて学習サンプル数が少ない状況においても優れたハルシネーション検知性能を示した．残課題としては QA 以外の別ドメインへの汎化性能の検証や，MoE アーキテクチャのルータなど他の内部状態との統合の検討余地がある．

参考文献

[1] Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 19519–19529, 2024.

[2] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM computing surveys**, Vol. 55, No. 12, pp. 1–38, 2023.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in neural information processing systems**, Vol. 33, pp. 9459–9474, 2020.

[4] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. RARR: Researching and revising what language models say, using language models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 16477–16508, 2023.

[5] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In **Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)**, pp. 2463–2473, 2019.

[6] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In **Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1)**, pp. 3214–3252, 2022.

[7] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. **CoRR**, 2022.

[8] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. **Nature**, Vol. 630, No. 8017, pp. 625–630, 2024.

[9] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it' s lying. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 967–976, 2023.

[10] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In **5th International Conference on Learning Representations (ICLR 2017), OpenReview submission**. OpenReview.net, 2017.

[11] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In **ACL 2018-56th Annual Meeting of the Association for Computational Linguistics**, Vol. 1, pp. 2126–2136. Association for Computational Linguistics, 2018.

[12] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In **5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings**. OpenReview.net, 2017.

[13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. **Journal of Machine Learning Research**, Vol. 23, No. 120, pp. 1–39, 2022.

[14] Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Probing semantic routing in large mixture-of-expert models. In **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 18263–18278, 2025.

[15] I. T. Jolliffe. **Principal Component Analysis**. Springer Series in Statistics. Springer, 2 edition, 2002.

[16] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. **Analytica chimica acta**, Vol. 185, pp. 1–17, 1986.

[17] 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 **The 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

[18] Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 17986–18002, 2025.

[19] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. 2015.

[20] Qwen Team. Qwen3 technical report, 2025.

[21] OpenAI. gpt-oss-120b&gpt-oss-20b model card, 2025.

[22] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). **Biometrika**, Vol. 52, No. 3-4, pp. 591–611, 1965.