視覚的文書理解モデルの内部機序に基づくチューニング

川﨑春佳　田中涼太　壹岐太一　西田京介
NTT 株式会社 人間情報研究所
{haruka.kawasaki, ryota.tanaka, taichi.iki, kyosuke.nishida}@ntt.com

概要

社会や産業におけるデータの多くは文書として存在し，視覚的に文書を理解する（視覚的文書理解）能力は大規模視覚言語モデル（LVLM）の研究・産業応用において不可欠である．LVLM の性能向上に向けた内部機序調査は進められているが，視覚的文書理解に関する調査は不十分である．また，内部機序調査の知見を活用したチューニング方法は提案されていない．本研究は，視覚的文書理解に関する LVLM の内部機序を調査し，それに基づくチューニング手法を初めて提案する．提案手法はデータやモデル構造を変更せずに適用可能である．複数の文書関連タスクでの実験の結果，訓練パラメータ数と訓練時間を削減しながら性能向上を確認した．

1　はじめに

社会や産業におけるデータの多くは文書として存在し，視覚的に文書を理解する（視覚的文書理解）能力は，大規模視覚言語モデル（LVLM）の研究・産業両面で重要となっている．強力な大規模言語モデル（LLM）や大規模データセットの登場により，LVLM 開発は急速に進展している [1, 2]．特に，データやモデル構造の工夫による性能向上が試みられている [3] が，モデルの本質的な能力や内部機序に関する理解が不十分なまま性能改善が進められている可能性がある．こうした工夫を施す前のモデルは，能力を最大限発揮できているのだろうか．この疑問を解明するため内部機序調査が進められている [4]．また，タスク性能向上目的でも調査されている [5]．しかし，これらは自然画像中心の調査であり，視覚的文書理解に関しては十分に調査されていない．

また，モデルが持つ能力を引き出す方法として，推論・学習時の工夫が考えられる．推論時の工夫 [6, 7] では性能低下の原因部分を直接改善できない．学習時の工夫は様々な方法が提案されている [8, 9] が，内部機序に基づく方法は提案されていない．

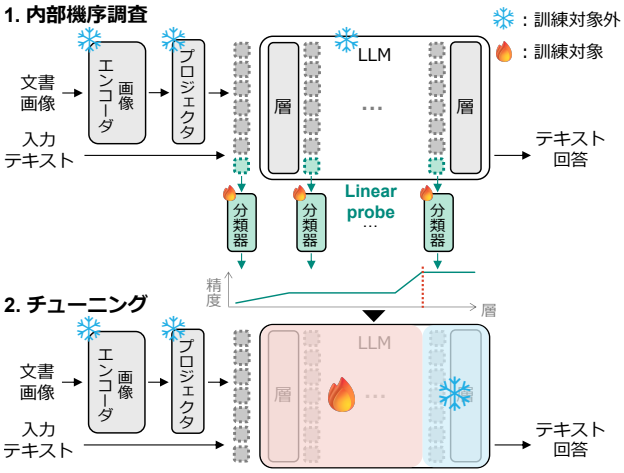


図 1　本研究の全体像．

本研究ではこれらの課題に着目し，**内部機序に基づくチューニング**を提案する．図 1 に示すように，視覚的文書理解に関する内部機序調査を行った後，それに基づくチューニングを行う．複数の文書関連タスクで有効性を検証し，性能向上を確認した．

2　関連研究

LVLM の内部機序　LVLM において，性能が低くなる原因が LLM 部分である可能性が示唆されている [4, 10, 11]．また，LLM 部分の層の段階ごとに役割があること [12, 13, 14, 15, 16, 17] や，LLM 部分で理解したことをテキストで出力できないこと [4] が確認されているが，LLM 全層の調査や文書画像での調査は行われていない．本研究では，初めて文書画像に関して LLM 全層を対象とした詳細な内部機序調査を行う．どの層で役割の変化があるか，性能向上を妨げる様子が見られるかを調査する．

チューニングの工夫　指標と閾値を用いたチューニング層の決定 [18, 19, 20, 21, 22] や，層位置に基づくチューニング方法 [9, 23] が提案されているが，内部機序に基づくチューニングは行われていない．本研究では，内部機序調査を踏まえ，モデル内部の役割に基づく層分割によるチューニングを提案する．

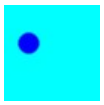
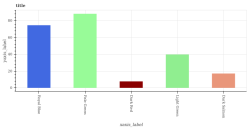
タスク	easy-VQA	Word recognition	Structuring	FigureQA
画像				
クエリ	Is the background black?	Is the text in the image "Spencerlan"?	Is the red boxed region a title?	Is Pale Green the minimum?
視覚処理の性質	局所的認識	局所的認識	空間・構造的認識	空間・構造的認識

図 2　Yes/No 分類タスクの例.

3　内部機序調査

3.1　Yes/No 分類タスク

Yes/No で分類できる問題を用意する．タスク間の難易度を合わせるため，Qwen2.5-VL 3B に解かせて誤った問題を使用する．図 2 に例を示す.

easy-VQA　easy-VQA データセット[1)]を参考に，画像・質問を作成した，色や形などの視覚的特徴を問うタスクである．例えば「*Is the background black?*」といったクエリがある．画像サイズ 64 × 64，訓練データ 10 万件，テストデータ 1 万件である．局所的な視覚認識が必要なタスクである.

Word recognition　MJSynth Text Recognition データセット [24, 25] の画像を使用し，クエリを付与した単語認識タスクである．例えば「*Is the text in the image "Spencerlan"?*」といったクエリがある．負例は，Qwen2.5-VL 3B に画像を読ませたときの誤答を用いる．画像サイズは 30 × 120 程度，訓練データ 10 万件，テストデータ 1 万件である．局所的な視覚認識が必要なタスクである.

Structuring　PubLayNet データセット [26] の画像とアノテーションを用いて赤枠をレンダリングし，赤枠内の構成要素を問うタスクである．構成要素は，タイトル，テキスト，リスト，図，表のいずれかである．クエリを付与し，例えば「*Is the red boxed region a title?*」がある．画像サイズは 792 × 612，訓練データ 10 万件，テストデータ 1 万件である．空間・構造的な視覚認識が必要なタスクである.

FigureQA　グラフ内の要素について比較を伴う推論タスクである FigureQA データセット [27] を使用する．例えば，「*Is Pale Green the minimum?*」といったクエリがある．画像サイズは 500 × 500 程度，訓練データ 74913 件，テストデータ 6111 件である．空間・構造的な視覚認識が必要なタスクである.

1）　https://github.com/vzhou842/easy-VQA

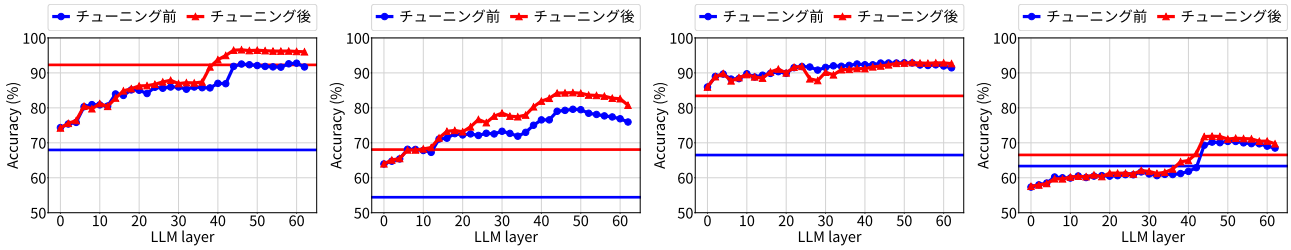
3.2　実験

実験設定　図 1 上部のように，LLM 各層最終トークンに対して Linear probe [28] を行う．Linear probe 精度の変化で，役割変化の層を確認する．また，Linear probe 精度とテキスト回答精度の比較で，LLM が性能向上を妨げる原因の一つか調査する．Linear probe の分類器はタスク・層ごとに作成し，それぞれ 1 エポック訓練する．また，調査の結果がチューニング不足に起因するのか，チューニング有無に依存しない現象なのかの区別のため，チューニングした LVLM でも同様の実験を行い，比較する．LVLM は，各タスクで 10 エポックチューニングする．画像エンコーダ - プロジェクタ - LLM の構造の LVLM である，Qwen2.5-VL 32B を対象に実験を行う.

層ごとに役割が変化するか？　図 3 のチューニング前の Linear probe 精度より，段階的に理解が進み，役割が変化したと考えられる．また，Gemma3 27B [29]，LLaVA-NeXT 13B [30] でも同様の傾向が見られた．各段階の具体的な役割は 4.1 節で述べる.

LLM 部分が性能向上を妨げているか？　図 3 のチューニング前の，テキスト回答精度は Linear probe 精度よりも低くなった．これは内部で理解しているがテキストに出力できない現象であり，LLM が更なる性能向上を妨げていると考えられる．Gemma3 27B，LLaVA-NeXT 13B でも同様の傾向が見られた.

チューニング不足が原因か？　図 3 の Linear probe 精度から，チューニング後も内部処理の段階が存在しており，段階が変化する層はチューニング前後で一致していることが分かる．チューニング後のテキスト回答精度はチューニング前の Linear probe 最高精度よりも低く，依然として内部理解をテキストで出力できていないことを確認した．以上より，処理の段階が存在すること，理解したことをテキストで完全には出力できず更なる性能向上を妨げていることは，チューニング不足が根本的原因ではなく，現在の LVLM の内部機序の影響だと考えられる.



easy-VQA　Word recognition　Structuring　FigureQA

図 3　各層最終トークンの Linear probe 精度（折れ線グラフ）とテキスト回答精度（水平線）．縦軸は精度，横軸は LLM の層を表し，2 層ごとにプロットしている．

4　チューニング

4.1　役割ごとの層分割

図 3 より，LLM 内部で理解が段階的に進んでいることが見て取れた．また，従来研究 [12, 13, 14, 15, 16, 17] から，LLM 低層で全体的な視覚情報を処理，中層で回答の根拠となる視覚情報を処理，高層で言語的な処理を行うと考えられる．これらを踏まえ，図 1 下部のように，内部処理の段階に合わせて層を分割し，分割単位でのチューニングで性能向上するか調査する．具体的には，Linear probe 精度の最初の上昇終了までの 0〜12 層を低層，次の上昇終了までの 13〜43 層を中層，上昇終了後の 44〜63 層を高層とし，それぞれの役割に対応するように分割する．

4.2　チューニングの設定

本実験では，一段階または二段階でチューニングを行う．一段階の場合は，「低」や「低〜中」のように表記する．「低〜中」は，低層と中層の，0〜43 層を一段階でチューニングすることを意味する．二段階の場合は，一段階目と二段階目でチューニングする層に被りがないようにし，一段階目と二段階目は同じエポック数ずつ訓練する．二段階の場合は，「低→中」のように表記する．これは，一段階目に低層をチューニングし，二段階目に中層をチューニングすることを意味する．すべてのモデルで合計訓練エポック数が同じになるようにする．また，チューニング前のモデル（base）と，全層のチューニングを行ったモデルとも比較する．

4.3　Yes/No 分類タスクでの実験

実験設定　フルファインチューニング（FT）と Low-Rank Adaptation（LoRA）チューニング [31] を行う．FT の学習率は 1e-6，LoRA の学習率は 1e-4 とする．3.1 節に記載のデータを用い，10 エポック訓練する．評価指標は，精度（Accuracy）である．

役割ごとの層分割チューニングに効果があるか？　表 1 より，役割ごとの層分割チューニングで全層よりも性能向上する傾向が見られ，効果があることが示された．これは，過剰適合や学習済み表現の破壊が抑制されたためだと考える．

どの役割の層が特に重要か？　表 1 で精度が高いモデルの多くは，低層と中層のみをチューニング範囲に含んでいた．特に一段階の場合では，中または低〜中が最も精度が高くなった．これは，Yes/No 分類タスクは言語的推論よりも視覚処理が重要となるためだと考えられ，視覚処理の役割の部分をチューニング範囲に含めることが重要であると分かる．

タスクとチューニングすべき層の関係は？　Structuring と FigureQA のみ，中〜高で全層よりも高くなるモデルが見られた．二段階の場合も，Structuring と FigureQA のみ高層を含むモデルが全層よりも高くなる場合があった．これは，タスクを解くのに局所的な認識と空間・構造的な認識のどちらが重要かで分かれているのではないかと考える．easy-VQA と Word recognition では，局所的な認識に精緻な視覚特徴が必要となるため，低〜中層の視覚処理段階が特に重要となり，Structuring と FigureQA では，空間・構造的な認識のための広域の構造や要素間の関係の理解が必要となるため，高層の重要度が増したのではないかと考える．これらは，図 3 の低〜中層部分の上昇幅の傾向とも一致した．

フルファインチューニングと LoRA の違いは？　違いは特に二段階のチューニングに現れた．二段階の FT では，チューニングするパラメータ数は一段階で同じ範囲をチューニングした場合に比べ少ないが，性能向上するタスクが多く見られた．一方で，二段階の LoRA では一段階の場合以上の性能向上が見られたのは Structuring のみだった．これは，

表 1　役割ごとの層分割をしたチューニングの比較．**太字**は各タスク・チューニング方法において最も精度が高いモデル，下線は二番目に精度が高いモデルである．また，青字は base よりも低いモデル，赤字は全層よりも高いモデルである．() 内の数値は，全層チューニングを 100 としたときのチューニングパラメータ数である．Accuracy（%）を示す．

Task	Tuning	base (0)	全層 (100)	低 (20)	中 (48)	高 (31)	低〜中 (69)	中〜高 (80)	低→中 (35)	中→低 (35)	中→高 (40)	高→中 (40)
easy-VQA	FT	67.96	92.39	65.76	93.25	68.48	90.53	91.48	76.50	**95.69**	89.10	88.54
	LoRA		92.40	69.75	92.21	69.75	**93.67**	87.07	86.64	88.25	88.88	88.80
Word recognition	FT	54.43	68.30	58.30	70.19	54.57	**74.59**	63.21	71.08	73.32	65.97	64.95
	LoRA		**76.71**	58.37	61.78	55.43	72.82	68.40	65.72	65.66	62.42	61.99
Structuring	FT	66.50	83.67	66.03	82.96	67.70	85.13	82.24	**88.25**	82.13	83.29	68.59
	LoRA		81.12	66.94	82.71	68.88	86.29	82.58	80.13	80.39	**87.96**	87.13
FigureQA	FT	63.34	64.38	65.52	66.45	63.43	66.83	66.19	68.01	**70.30**	66.14	68.37
	LoRA		66.11	65.28	65.73	63.64	**66.26**	65.42	65.62	65.72	65.47	65.13

表 2　実用的タスクにおける役割ごとの層分割をしたチューニングの比較．ANLS を示す．

Task	Tuning	base (0)	全層 (100)	低 (20)	中 (48)	高 (31)	低〜中 (69)	中〜高 (80)	低→中 (35)	中→低 (35)	中→高 (40)	高→中 (40)
DocVQA	FT	0.9380	0.9474	0.9379	0.9451	0.9356	0.9430	**0.9492**	0.9418	0.9386	0.9477	0.9410
	LoRA		**0.9423**	0.9401	0.9396	0.9405	0.9403	0.9408	0.9388	0.9396	0.9400	0.9401
InfographicVQA	FT	0.8352	0.8337	0.8320	0.8341	0.8299	0.8298	**0.8393**	0.8295	0.8306	0.8349	0.8306
	LoRA		0.8307	**0.8362**	0.8300	0.8306	0.8304	0.8260	0.8345	0.8309	0.8288	0.8335

LoRA ではチューニングパラメータ数が少なく，一段階目のチューニングモデルを二段階目でさらに矯正することが難しいためだと考える．また，LoRA では部分的なチューニングで全層での精度を上回ることが難しく，LoRA での工夫には課題が残る．LoRA の場合にはプロジェクタも併せたチューニング等の工夫が必要だと考える．

4.4　実用的タスクでの実験

実験設定　DocVQA[32] と InfographicVQA[33] を用いる．従来研究 [34] を参考にし，FT の学習率は 1e-5 とする．LoRA は 1e-4 である．合計 2 エポック訓練する．評価指標は Average Normalized Levenshtein Similarity（ANLS）である．

実用的タスクでも効果があるか？　表 2 より，LoRA を用いた DocVQA 以外では，役割ごとの層分割をした方が全層よりも性能向上し，実用的なタスクにおいても効果があると示された．

どの層をフルファインチューニングするべきか？　FT の場合，中層に加え高層を対象に含めることで性能向上した．これは，言語的推論が必要なタスクであるため，言語的推論を行っている高層を学習したことで向上したのではないかと考える．

どの層を LoRA チューニングするべきか？　LoRA の場合，DocVQA では高層を含めた場合に高くなる傾向が見られ，InfographicVQA では，低層を対象とすることで向上した．これは，LoRA では更新されるパラメータ数が少ないことで，もとの能力は保持したまま，それぞれ言語推論能力と画像処理能力がタスクに適応したためではないかと考える．LoRA では，文字が中心のタスクでは全層が最も性能が良く，もとのパラメータを変更せずに文字認識能力を向上させるのは難しい可能性が示唆された．

どのくらい効率的か？　DocVQA で最も向上した中〜高では，約 20%の訓練パラメータ数削減，約 35%の訓練時間削減しながら性能向上を達成した．

5　おわりに

LVLM の視覚的文書理解能力の向上を目指し，内部機序を調査し，役割に基づくチューニングを行った．複数の文書関連タスクで，データやモデル構造を変更せずに性能向上可能であることを確認した．

本研究の独自性　本研究は，LVLM において，初めて内部機序を踏まえた学習方法を提案した．また，視覚的文書理解に関する LVLM の LLM 全層の詳細な内部機序調査を初めて行った．

本研究の重要性　本研究は，データの増強やパラメータ数の増大による性能向上の潮流を脱却し，省リソースで視覚的文書理解能力を向上可能な方法の確立に向けた重要な知見を提供する．また，内部機序分析から解決策までつなげた研究が少ない状況に，一石を投じる研究である．

参考文献

[1] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. **arXiv preprint arXiv:2412.05271**, 2024.

[2] 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. **arXiv preprint arXiv:2502.13923**, 2025.

[3] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. **arXiv preprint arXiv:2501.14818**, 2025.

[4] Shivam Chandhok, Wan-Cyuan Fan, Vered Shwartz, Vineeth N. Balasubramanian, and Leonid Sigal. Response wide shut? surprising observations in basic vision language model capabilities. In **ACL**, pp. 25530–25545, 2025.

[5] Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. **arXiv preprint arXiv:2502.17516**, 2025.

[6] Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: Self logits evolution decoding for improving factuality in large language models. In **NeurIPS**, pp. 5188–5209, 2024.

[7] 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 **EMNLP**, pp. 19519–19529, 2024.

[8] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large models: A survey of methodologies. **arXiv preprint arXiv:2410.19878**, 2024.

[9] Taewook Hwang, Hyein Seo, Jeesu Jung, and Sangkeun Jung. Exploring selective layer freezing strategies in transformer fine-tuning: Nli classifiers with sub-3b parameter models. **Applied Sciences**, Vol. 15, No. 19, 2025.

[10] Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. In **Findings of ACL**, pp. 4830–4843, 2025.

[11] Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. On the perception bottleneck of VLMs for chart understanding. In **Findings of EMNLP**, pp. 10829–10841, 2025.

[12] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In **CVPR**, pp. 19781–19791, 2025.

[13] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In **CVPR**, pp. 25004–25014, 2025.

[14] Jitesh Jain, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. Elevating visual perception in multimodal llms with visual embedding distillation. In **NeurIPS**, 2025.

[15] 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**, pp. 4135–4144, 2025.

[16] Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In **EMNLP**, pp. 3512–3534, 2025.

[17] Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. **arXiv preprint arXiv:2511.05923**, 2025.

[18] Guangyuan Shi, Zexin Lu, Xiaoyu Dong, Wenlong Zhang, Xuanyu Zhang, Yujie Feng, and Xiao-Ming Wu. Understanding layer significance in llm alignment. In **COLM**, 2025.

[19] Kai Yao, Penglei Gao, Lichun Li, Yuan Zhao, Xiaofeng Wang, Wei Wang, and Jianke Zhu. Layer-wise importance matters: Less memory for better performance in parameter-efficient fine-tuning of large language models. In **Findings of EMNLP**, pp. 1977–1992, 2024.

[20] Chenxing Wei, Yao Shu, Ying Tiffany He, and Fei Yu. Flexora: Flexible low-rank adaptation for large language models. In **ACL**, pp. 14643–14682, 2025.

[21] Alessio Devoto, Federico Alvetreti, Jary Pomponi, Paolo Di Lorenzo, Pasquale Minervini, and Simone Scardapane. Adaptive layer selection for efficient vision transformer fine-tuning. **arXiv preprint arXiv:2408.08670**, 2024.

[22] Zeyu Liu, Souvik Kundu, Anni Li, Junrui Wan, Lianghao Jiang, and Peter Beerel. AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models. In **ACL**, pp. 161–167, 2024.

[23] Max Reinhardt, Gregor Geigle, Radu Timofte, and Goran Glavaš. Improving vision-language cross-lingual transfer with scheduled unfreezing. In **ALVR**, pp. 155–166, 2024.

[24] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In **Workshop on Deep Learning, NIPS**, 2014.

[25] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. **IJCV**, Vol. 116, No. 1, pp. 1–20, 2016.

[26] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In **ICDAR**, pp. 1015–1022, 2019.

[27] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. **arXiv preprint arXiv:1710.07300**, 2017.

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

[29] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. **arXiv preprint arXiv:2503.19786**, 2025.

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[31] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In **ICLR**, 2022.

[32] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In **WACV**, pp. 2200–2209, 2021.

[33] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In **WACV**, pp. 1697–1706, 2022.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **CVPR**, pp. 26296–26306, 2024.

A　内部機序調査

3 章の Gemma3 と LLaVA-NeXT の結果を図 4 と図 5 に示す．縦軸は精度，横軸は LLM の層を表し，2 層ごとにプロットしている．折れ線グラフは Linear probe 精度を，水平な線はテキスト回答精度を表す．タスクごとに色分けしている．Gemma3 はチューニング前のみ，LLaVA-NeXT はチューニング前後の結果を示す．

また，Linear probe を，最終トークン以外のトークンも用いて詳細に行った結果を図 6 から図 8 に示す．トークンは，テキストトークン，画像トークン，最終トークン，全トークンに分け，トークンが複数になる場合は Mean pooling する．縦軸は精度，横軸は LLM の層を表し，2 層ごとにプロットしている．折れ線グラフは Linear probe 精度を，水平な破線はテキスト回答精度を表す．トークンの種類ごとに色分けしている．

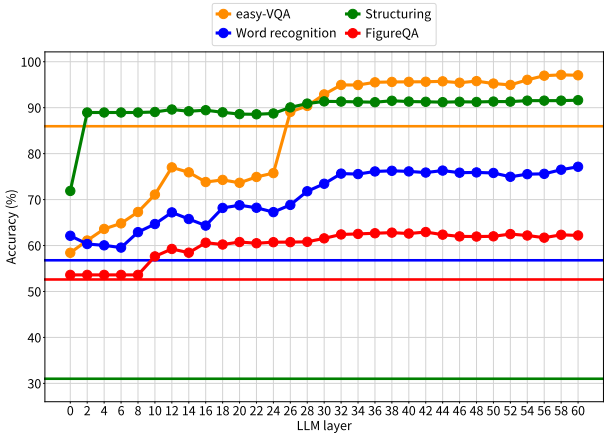


図 4　Gemma3 27B における各層最終トークンの Linear probe 精度とテキスト回答精度．

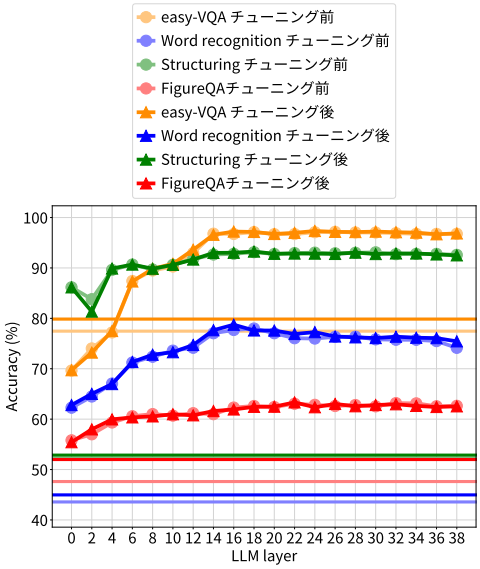


図 5　LLaVA-NeXT 13B における各層最終トークンの Linear probe 精度とテキスト回答精度．

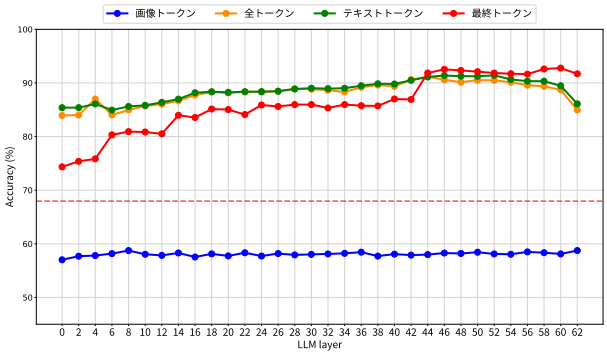


図 6　easy-VQA 各層・各トークンからの Linear probe 結果．

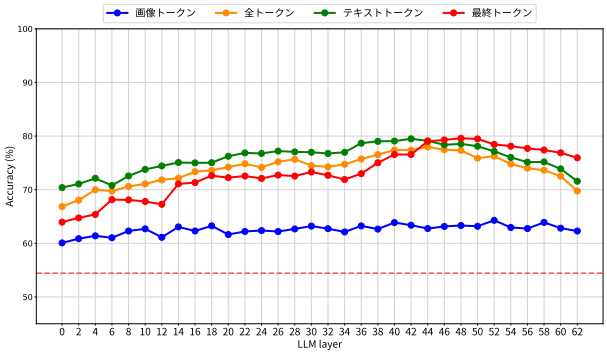


図 7　Word recognition 各層・各トークンからの Linear probe 結果．

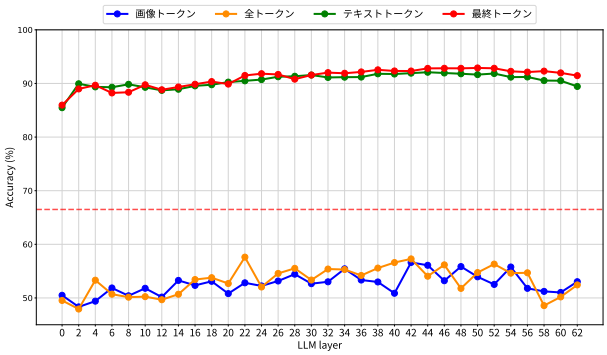


図 8　Structuring 各層・各トークンからの Linear probe 結果．

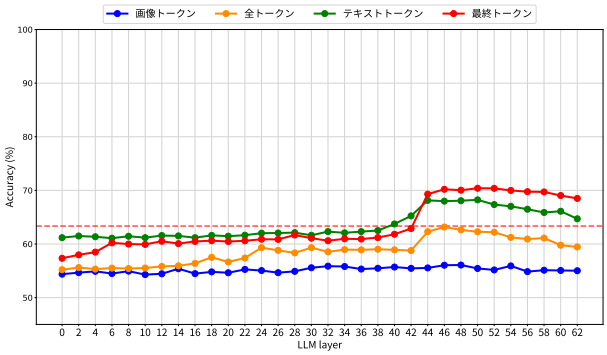


図 9　FigureQA 各層・各トークンからの Linear probe 結果．