文脈内学習におけるタスク指向情報除去のメカニズム

趙羽風[1]　楊昊霖[2]　峰岸剛基[3]　井之上直也[1,4]
[1]北陸先端科学技術大学院大学　[2]シカゴ大学　[3]東京大学　[4]理化学研究所
yfzhao@jaist.ac.jp　完全版 arXiv:2509.21012　Verb_subspace

概要

本研究では，情報除去の観点から文脈内学習（**I**n-**c**ontext **L**earning, ICL）のメカニズムを分析する．まず，言語モデルは，0-shot クエリを複数タスクの情報が混在した非選択的な表現として符号化するため，意図した出力を得難いことを示す．次に，低ランクフィルタに基づく選択的な情報除去により，0-shot のモデル出力を意図したタスクへと誘導できること，及び few-shot ICL はこの情報除去操作を自然に実現していることを示す．最後に，情報除去操作を担う重要なアテンションヘッドを同定し，これらが ICL の性能に大きく寄与することを確認する．

1　はじめに

ICL は，入力-ラベル対からなるデモとクエリを言語モデル（LM）に与え，クエリのラベルを予測させる枠組みである．既存研究では，ICL のメカニズムを特定の入力要素 [1, 2, 3, 4]，事前学習分布 [5, 6, 7, 8]，または単純なアルゴリズムとの類似 [9, 10, 11] として説明してきた．さらに，Mechanistic Interpretability における研究 [12, 13, 14, 15, 16, 17, 18] では，ICL をアテンションヘッド等の構成要素へ還元し，誘導ヘッドによるデモからのラベルコピー機構をその中核的メカニズムと位置づけている．しかし，クエリに対する正解ラベルがデモに含まれない状況（**Unseen Label シナリオ**）においても [17, 15]，ICL は 0-shot 設定を上回る性能を示すことが報告されている．この性能差は，デモからのラベルコピーだけでは説明できず，誘導ヘッド以外にも ICL を支える追加的なメカニズムが存在することを示唆する．

以上を踏まえ，本研究では 図 1(B) に示す新たな視点を提案する．すなわち，ICL を**「新しい情報を出力へコピーする [15]」「新たなタスクを学習する [19]」過程として捉えるのではなく，クエリ中のタスク非関連情報を除去し，デモで指定されたタスクを強調する過程**として捉える．具体的には，既存研究 [20, 21] より，クエリの隠れ状態には様々な情報が異なる部分空間に混在しており（図 1(A)），ICL は，デモの情報を利用して隠れ状態からタスク非関連情報を除去し，タスク関連情報が集中的に存在する部分空間（**タスク言語化部分空間**; **T**ask-**V**erbalization **S**ubspace; TVS）を強調しているというのが我々の仮説である．デモのない 0-shot 入力で尤もらしいがタスク指向的でない非選択的な出力が生じてしまう（図 1(D)）のは，隠れ状態がノイズまみれだからというわけである．この仮説を検証するため，図 1(C) に示すように，0-shot 入力の最終トークンの残差流に低ランクフィルタを人工的に注入し，望ましい出力から TVS を特定した．その結果，低ランクな TVS でも精度は大幅に向上し，TVS に直交する情報の除去が非選択的出力を抑制し，タスク指向の出力を強制できることが示された．

これらの知見に基づき，本研究では，低ランクフィルタを注入した人工的な 0-shot 設定におけるタスク指向情報除去を，自然な（注入を行わない）few-shot シナリオへと一般化する．図 1(B) に示すように，few-shot のデモに導かれることで，LM は，先行する低ランクフィルタ注入実験で独立に算出された TVS に向かって隠れ状態を暗黙的に駆動し，タスク特化の出力を生成すること，さらに，Unseen Label シナリオにおける同様の挙動も確認した．

加えて，誘導ヘッドから独立した，情報除去を実行するアテンションヘッドの一群を特定した．アブレーション実験により，精度が大幅に低下すること，特に Unseen Label シナリオでは精度がゼロ近傍まで落ちることが確認された．この結果は，タスク指向情報除去が ICL の，特に誘導ヘッドのみでは説明できない Unseen Label シナリオにおける有効なメカニズムであることを示している．

2　手法：タスク指向情報除去の観測

本節では，タスク指向情報除去を観測するためのフレームワークを提案する．なお，本論文では ICL

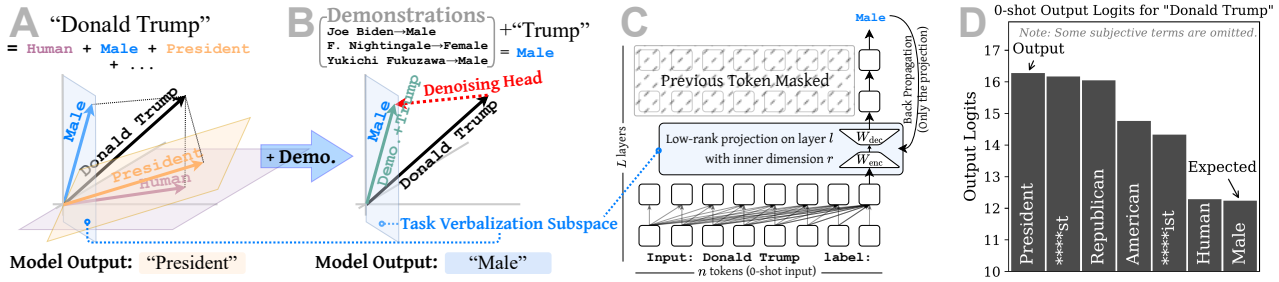

図 1: **(A)** 0-shot クエリは，多様な部分空間にまたがるすべてのタスク情報を含む非選択的な意味表現として符号化され，出力はこれらの情報間で非選択的となる．**(B)** デモは，クエリ情報をフィルタリングし，タスクに関連する情報のみを特定の部分空間（TVS）に保持させることで特化的な出力を導く．**(C)** 0-shot 入力の残差流に低ランクフィルタを注入し，正解出力の上にそのフィルタのみを学習することで，TVS を明示的に特定する．**(D)** 0-shot 入力「Donald Trump, label:」に対する出力ロジット分布（一部のトークンのみ表示）．

のメカニズムを検証対象とするが，本フレームワークはより広範な設定へも拡張可能であると考える．

前提: タスク言語化部分空間 (TVS)． 図 1(A) に示すように，LM は 0-shot クエリ x_q を，l 層の Transformer ブロック通過後に，非選択的な意味表現 $h_q^l \in \mathbb{R}^d$ として符号化する．この h_q^l から残りの層によって導かれる出力は，図 1(D) に示すように，x_q に対して本来タスクが期待する出力とはなりにくい．この状況において，我々は，タスク関連情報を符号化する低ランク部分空間が表現空間内に存在すると仮定する．この部分空間は，タスク非関連情報を除去するための射影行列 $W \in \mathbb{R}^{d\times d}$ を備え，h_q^l を $h_q^l W$ へ写像することで，最終出力をタスク特化の出力へと誘導する．次節で述べるように，W はタスクそのものに加え，「positive」と「+」のような出力の表現形式も規定するため，W によって定義される部分空間をタスク言語化部分空間（TVS）と呼ぶ．

Step 1：0-shot 隠れ状態における明示的 TVS の特定． 本研究では，TVS が低ランクフィルタ $W_{\text{enc}}W_{\text{dec}} \in \mathbb{R}^{d\times r}\cdot\mathbb{R}^{r\times d}$ によってパラメタ化されて存在すると仮定し，この仮定をフィルタ注入によって検証する．まず，図 1(C) に示すように，$W_{\text{enc}}W_{\text{dec}}$ を特定の層 l における入力列の最終トークン（出力が生成される位置）の残差流へ注入し，さらに，後続層における最終トークンから先行トークンへのアクセスを遮断することで，フィルタ注入後に入力文脈から得られる追加情報を排除し，因果関係を明確化する．この上で，0-shot 入力とその正解応答を与え，LM のパラメタを固定して $W_{\text{enc}}W_{\text{dec}}$ のみを学習し，出力の精度を検証する（詳細は付録 A を参照）．

Step 2：TVS への暗黙的なタスク指向情報除去のトレーシング． 次に，本研究では，LM がデモの助けを借りて，**フィルタ注入を行わない**自然な推論過程においても，隠れ状態を TVS へと駆動していると仮定する．この仮定を検証するため，Step 1 で学習した $W_{\text{enc}}W_{\text{dec}}$ を用い，few-shot 入力の隠れ状態に対する二つの幾何学的指標を慎重に設計する．具体的には，l 層の Transformer ブロック通過後に得られる k-shot ICL 入力の最終トークンから，N 個の隠れ状態 $H^{l,k} = \{h_i^{l,k}\}_{i=1}^N$ が与えられるとする：

- **離心率** (Eccentricity)：情報除去の**マグニチュード**を測定するため，$H^{l,k}$ の第一主成分方向における共分散負荷率として離心率を算出する．これは，単一の線形表現上での情報の集中度，すなわち異方性を表す [22]．離心率が高いほど，より純度の高い表現であることを示す．
- **TVS 上の共分散流束** (Covariance Flux)：情報除去の**正確性**を測定するため，タスク関連情報の割合を算出する．具体的には，Step 1 において層 l で学習された次元数 $r = 8$ の $W_{\text{enc}}W_{\text{dec}}$ を用い，$H^{l,k}$ 中のすべての $h_i^{l,k}$ をこの $W_{\text{enc}}W_{\text{dec}}$ へ投影し，$H^{l,k}W_{\text{enc}}W_{\text{dec}}$ と元の $H^{l,k}$ との間の共分散比を計算する（詳細は 付録 A を参照）．共分散流束が高いほど，$H^{l,k}$ に含まれる情報のうち，より大きな割合がタスク関連情報であることを示す．

Step 3：情報除去と ICL 性能の因果関係の確立． 最後に，Step 2 で観察された情報除去の過程が ICL のメカニズムの中核的要素であることを確認する．本研究では，アテンションヘッドが情報除去の役割を担っていると仮定し，まず各ヘッドをアブレーションし，隠れ状態上で前述の指標を再計算することでタスク指向情報除去への寄与を定量化し，寄与の大きいヘッド（**D**enoising **H**eads; DH）を特定する．さらに，全 DH をアブレーションして ICL 性能を評価し，因果関係を検証する．詳細な操作手順について

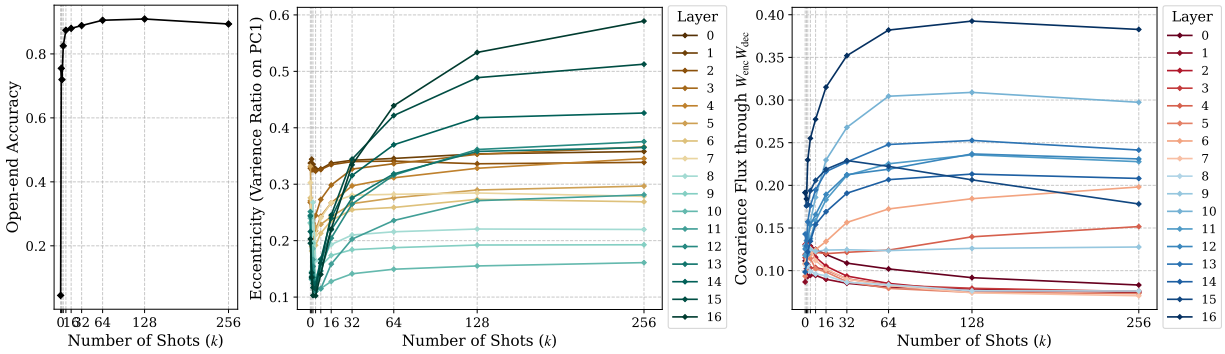

図 2: **（左）** デモ数 k に対する精度，**（中）** 各層・k における隠れ状態点群の離心率，**（右）** 各層・k において取得された隠れ状態点群に対する共分散流束．各層において学習された $W_{\text{enc}}W_{\text{dec}}$ を用いて計算した．

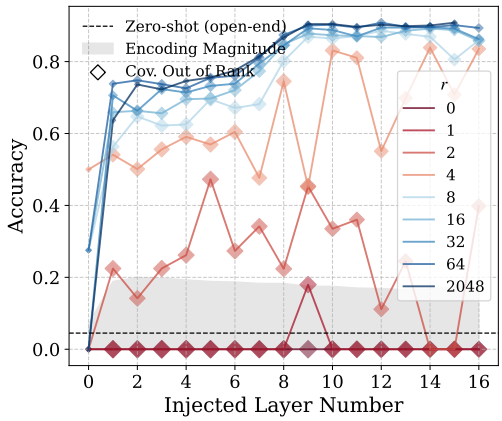

図 3: 各層およびランクにおけるフィルタ注入評価．**マーカーサイズ**：h_q^l の点群に対する上位 r 主成分に由来する残余共分散．**Encoding Magnitude**：現在の層における隠れ状態のクエリの文埋め込みとしての品質 [8]．正解率の向上と同時に上昇が始まっている．

は，§3.3 および付録 A を参照されたい．

3　実験：ICL における情報除去

本稿では，紙幅の制約により，主として SST-2 [23] における Llama 3.2-1B [24] の実験結果のみを報告する．6 種類のモデルおよび 9 種類のデータセットの結果については，完全版論文を参照されたい．

3.1　Step 1: TVS の特定

隠れ状態における明示的 TVS の存在． §2 の Step 1 で述べた注入・学習実験を実施し，さまざまな l および r に対する精度を図 3 に示す．全体として，ゼロショット精度と比較すると，最大ランク[1] 2 の TVS の存在を明示的に仮定して隠れ状態をフィルタリングすることで，出力精度が大幅に向上することが確認された．これらの結果は，0-shot 入力の隠れ状態から情報を選択的に除去することで出力がタスクへと誘導されるという仮説を支持するものであり，ICL がこの処理を暗黙的に適用している可能性を示唆している．この点については上記の Step 2 に従って §3.2 にて検証する．

$W_{\text{enc}}W_{\text{dec}}$ は情報除去である． 一つの懸念として，h_q^l の主成分方向が $W_{\text{enc}}W_{\text{dec}}$ の固有ベクトルと強く平行している場合，$W_{\text{enc}}W_{\text{dec}}$ が十分な情報除去を行っていない可能性が考えられる．この可能性を排除するため，h_q^l の上位 r 主成分に対する残余共分散負荷率を計算し，最大ランク r における情報除去量の下限として評価した（図 3 のダイヤサイズ）．結果より，$W_{\text{enc}}W_{\text{dec}}$ は 0-shot 精度を向上させつつ，有効に情報を除去していることが確認できた．

さらに付録 B において，TVS の名前の通り，W_{enc} が「タスク」の関連情報を抽出し，W_{dec} が出力の「言語化」パターンを制御していることを確認した．

3.2　Step 2: 暗黙的情報除去のトレース

LM は隠れ状態を自然に TVS へ圧縮する． §3.1 で学習した TVS を用い，§2 の Step 2 に従って，さまざまな k と l において両指標を評価した．結果および出力精度を図 2 に示す．全体として，デモ数 k が増加するにつれ，タスク指向情報除去は主に中〜後段の層で生じ，離心率と共分散流束が増大する．これは，デモにより強い異方的，タスク指向に圧縮が生じることを示す．特に SST-2 では，k に対して離心率が一度減少してから増加する非単調な挙動が観測される．これは，このデータセット上では情報除去が 0-shot 隠れ状態の第一主成分から始まることを示唆する（詳細は完全版を参照）．

指示文の影響． 実際には，タスク記述の指示文（instruction, 例：「この文の感情を予測せよ」）は

1)　r が十分に大きい場合でも，$W_{\text{enc}}W_{\text{dec}}$ の実効ランクは小さくなり得るため，r は最大可能ランクを表す（元の埋め込み次元は 2048）．

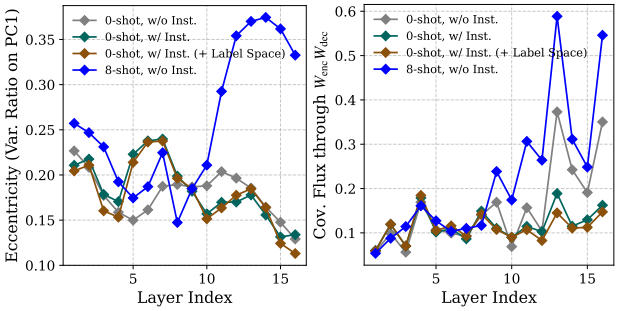

図 4: MR [25] における，指示文あり・デモあり・通常の 0-shot プロンプトに対する離心率および共分散流束.

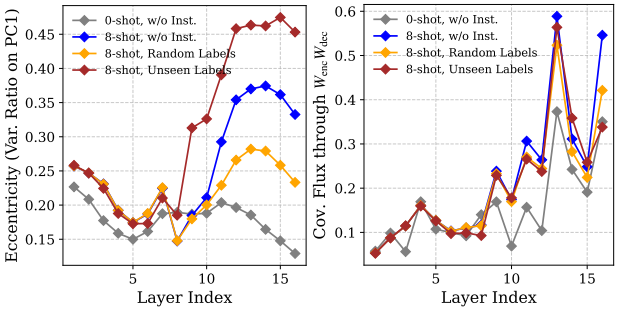

図 5: MR における，ランダムラベル・Unseen ラベル・通常設定に対する離心率および共分散流束.

表 1: DH のアブレーション後の推論精度（%）（6 データセットで平均）. **Layer**：DH を探索した層数/総層数. **w/o DH**：DH をアブレーションした精度. **w/o RH**：各層で DH と同数のランダムに抽出した注意ヘッドをアブレーションした精度.

Model / Layer	Demonstration Configuration	8-shot	8-shot w/o DH	8-shot w/o RH
Llama 3.2-1B 16 / 16	Random Sample	71.02	55.04	$67.29_{3.84}$
	Seen Label	73.40	56.52	$69.02_{3.72}$
	Unseen Label	14.54	1.06	$10.65_{4.65}$
Llama 3-8B[2] 7 / 32	Random Sample	77.63	72.05	$76.38_{0.96}$
	Seen Label	80.47	75.11	$79.70_{1.30}$
	Unseen Label	21.66	7.52	$18.94_{3.98}$
Qwen 2.5-3B[26][2] 18 / 36	Random Sample	73.97	68.41	$74.31_{1.29}$
	Seen Label	76.41	70.31	$76.36_{1.57}$
	Unseen Label	23.49	11.59	$26.82_{3.82}$
Qwen2.5 3B-Ins[2] 18 / 36	Random Sample	77.24	75.57	$77.28_{1.90}$
	Seen Label	78.86	77.52	$78.05_{0.81}$
	Unseen Label	48.70	37.51	$46.32_{4.52}$

few-shot デモの代替として用いられることが多い．そこで，指示文付きの入力に対して両指標を測定し，few-shot デモと同様のタスク指向情報除去が生じるかを検証した．図 4 に示すように，8-shot デモは 0-shot より強い情報除去と形態差を示す一方，指示文ありではラベル空間を明示しても，精度向上にも関わらず挙動は 0-shot 推論とほぼ同一である．以上より，情報除去は few-shot デモによってのみ誘発され，指示文とは異なることが示唆される．

ランダムラベルおよび Unseen ラベルの影響. デモ中のラベルの正確性は推論ダイナミクスや精度に与える影響が小さいことが報告されている [1, 2, 3, 15]. 本節では，デモのラベルをランダム化した場合 (Random Labels) と，Unseen Label シナリオで両指標を測定した（図 5）．離心率に着目すると，Random Labels では，通常の 8-shot 推論より情報除去が弱く，誤ったラベルがデモの機能を弱めるという報告 [15] と整合する．一方，Unseen Labels ではより強い情報除去が観測された．これは，ラベル空間の縮小によりタスクが狭く指定され，冗長情報が増えるためと考えられる．この傾向は multi-way より 2-way タスクで顕著であり，ラベル除去がタスク範囲に大きく影響することを示す（詳細は完全版参照）．総じて，タスク指向情報除去は全ての設定で観測され，違いは主にその強度に現れることが確認された．

3.3 Step 3: ICL 性能への因果関係

DH の役割. §2 の Step 3 に基づいて DH を同定し（詳細は付録 A 参照），一般 (Random Sample)・Unseen・Seen（少なくとも一度は正しいラベルが出現する）ラベル設定において，全 DH を同時にアブレーションした際の精度を評価した（表 1）．その結果，一般または Seen ラベル設定では，DH 除去による精度低下は限定的だが，Unseen ラベル設定では精度がほぼ消失した．これは，Unseen Label シナリオにおいて DH が主要な精度源として誘導ヘッドを補完することを示す [15]．さらに付録 B において，DH が誘導ヘッドと高い独立性を持つことを示す．

4 結論

本論文は，ICL 推論をクエリの隠れ状態からのタスク指向情報除去として解釈する新たな視点を提案した．具体的には，まず 0-shot 入力の隠れ状態に低ランクの明示的フィルタを注入して冗長情報を削減することで，精度が向上することを示した．次に，few-shot ICL がこの過程を自発的に実行していることを示し，そのハンドルである DH を同定するとともに，アブレーション実験によって ICL のメカニズムにおける情報除去の重要性を確認した．

2) 計算資源の制約により，これらのモデルの一部の層においてのみ DH を同定しているため，ここで示すアブレーション結果は全ての DH を網羅しているわけではない．

謝辞

本研究は，JST 創発的研究支援事業 JPMJFR232K，および中島記念国際交流財団の助成を受けたものです．

参考文献

[1] Sewon Min, et al. Rethinking the role of demonstrations: What makes in-context learning work? In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pages 11048–11064, 2022.

[2] Kang Min Yoo, et al. Ground-truth labels matter: A deeper look into input-label demonstrations. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pages 2422–2437, 2022.

[3] Jane Pan. What in-context learning "learns" in-context: Disentangling task recognition and task learning. Master's thesis, Princeton University, 2023.

[4] Jannik Kossen, et al. In-context learning learns label relationships but is not conventional learning. In **The Twelfth International Conference on Learning Representations**, 2024.

[5] Xiaonan Li and Xipeng Qiu. Finding support examples for in-context learning. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pages 6219–6235, 2023.

[6] Yuxian Gu, et al. Pre-training to learn in context. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 4849–4870, 2023.

[7] Xiaochuang Han, et al. Understanding in-context learning via supportive pretraining data. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 12660–12673, 2023.

[8] Hakaze Cho, et al. Mechanistic fine-tuning for in-context learning. **arXiv preprint arXiv:2505.14233**, 2025.

[9] Ruiqi Zhang, et al. Trained transformers learn linear models in-context. **arXiv preprint arXiv:2306.09927**, 2023.

[10] Damai Dai, et al. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In **ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models**, 2023.

[11] Sang Michael Xie, et al. An explanation of in-context learning as implicit bayesian inference. **arXiv preprint arXiv:2111.02080**, 2021.

[12] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In **The Twelfth International Conference on Learning Representations**, 2024.

[13] Lean Wang, et al. Label words are anchors: An information flow perspective for understanding in-context learning. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pages 9840–9855, 2023.

[14] Aaditya K Singh, et al. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In **Forty-first International Conference on Machine Learning**, 2024.

[15] Hakaze Cho, et al. Revisiting in-context learning inference circuit in large language models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[16] Haolin Yang, et al. Unifying attention heads and task vectors via hidden state geometry in in-context learning. **arXiv preprint arXiv:2505.18752**, 2025.

[17] Gouki Minegishi, et al. Beyond induction heads: In-context meta learning induces multi-phase circuit emergence. In **Forty-second International Conference on Machine Learning**, 2025.

[18] Haolin Yang, et al. Localizing task recognition and task learning in in-context learning via attention head analysis. **arXiv preprint arXiv:2509.24164**, 2025.

[19] Jiaoda Li, et al. What do language models learn in context? the structured task hypothesis. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 12365–12379, 2024.

[20] Baturay Saglam, et al. Large language models encode semantics in low-dimensional linear subspaces. **arXiv preprint arXiv:2507.09709**, 2025.

[21] Haiyan Zhao, et al. Beyond single concept vector: Modeling concept subspace in LLMs with gaussian distribution. In **The Thirteenth International Conference on Learning Representations**, 2025.

[22] Joshua Engels, et al. Not all language model features are one-dimensionally linear. In **The Thirteenth International Conference on Learning Representations**, 2025.

[23] Richard Socher, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[24] Aaron Grattafiori, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[25] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)**, pages 115–124, 2005.

[26] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[27] Diederik P Kingma. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

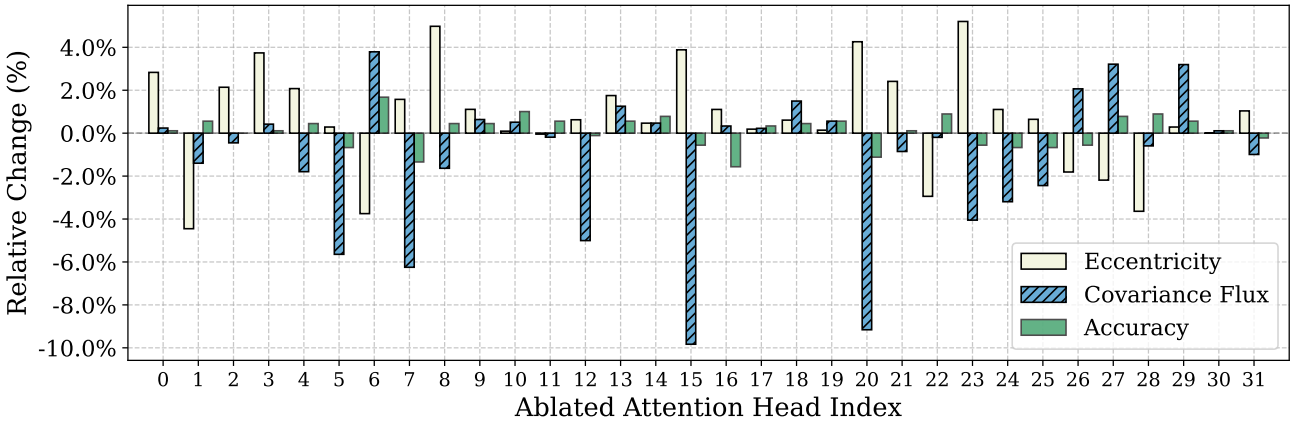

図 6: Llama3.2-1B の Layer 9 における全ヘッドについての，8-shot SST-2 上での二つの指標および出力精度のアブレーション．値が小さいほどアブレーション時の低下が大きい，すなわち該当ヘッドが該当指標へ大きく寄与することを意味する．

A　実験詳細

Step 1: 訓練設定. 図 1（C）に示すように，§3.1 の実験では，残差流の上に連続する 2 層の線形層を学習し，第一層（W_{enc}）のみにバイアス項を付与する．学習ではこの 2 層のみを更新対象とし，ゼロショット例 2048 件を用いて，Adam [27]（学習率 10^{-4}，$\beta_1 = 0.9$，$\beta_2 = 0.999$）で訓練する．32 件ごとに勾配計算を行い，パラメタを 1 回更新する．エポック数は 4．学習後，語彙全体で厳密なトークン一致により，512 件のホールドアウトテストで評価する．

Step 2: 共分散流束の計算. 与えられた最終トークン隠れ状態集合 $H^{l,k} = \{h_i^{l,k}\}_{i=1}^N$ に対し，§3.1 で学習した低ランクフィルタ $W_{\text{enc}}, W_{\text{dec}}$ を用い，写像後の集合 $H_1^{l,k} = \{h_i^{l,k} W_{\text{enc}} W_{\text{dec}}\}_{i=1}^N$ を得る．このとき，バイアスは共分散に影響しないため無視する．元の集合の共分散 $D^{l,k}$ と写像後の共分散 $D_1^{l,k}$ を計算し，以下で共分散流束を定義する：

$$\text{Covariance Flux} = \frac{\|D_1^{l,k}\|_*}{\|D^{l,k}\|_*}, \tag{1}$$

ここで $\|\cdot\|_*$ は特異値の総和（核ノルム）である．

Step 3: Denoising Head の同定. §2 のアブレーション再測定法によって計算した，Layer 9 の代表的結果を図 6 に示す．離心率が層によって相転移的に変化するため[3)]，本研究では，出力ゼロ化アブレーションを用い，共分散流束を指標に，閾値を –5% に設定し，強く寄与する注意ヘッド（DH）を特定する．いくつかのヘッド（例：#5, 7, 12, 15, 20, …）は両指標および精度に明確な寄与を示す．

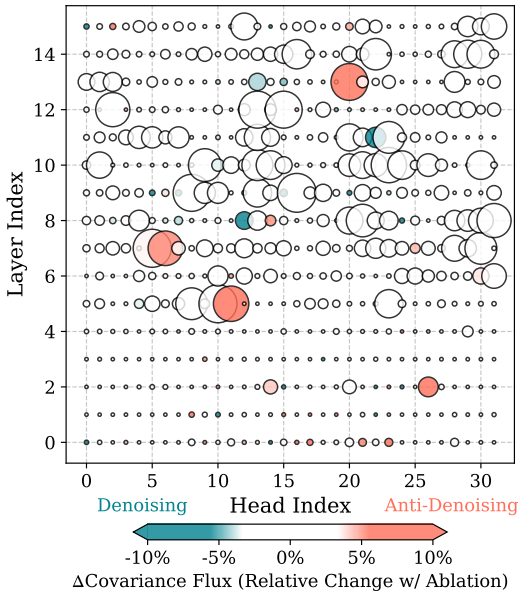

図 7: ヘッド毎（**散布点の大きさ**）の誘導ヘッドらしさと（**色**）アブレーション後の共分散流束の相対変化．

B　補足実験

W_{dec} はタスクの出力パターンを制御する. すなわち，$W_{\text{enc}} W_{\text{dec}}$ の学習ではタスク情報だけでなく，その情報をどの語彙（例：“positive/negative” や “A/B”）で出力するかというパターンも決定されるため，これを「タスク-言語化（TVP）」と呼ぶ．加えて，出力パターンは主に W_{dec} が担うことが分かった．具体的には，SST-2 で通常の語彙（“positive/negative”）により学習した後，W_{enc} または W_{dec} のどちらかを凍結し，残りのみを “A/B” への出力にファインチューニングしたところ（表 2），W_{dec} のみをファインチューニングした場合に限り言語化の転換が成功した．これは，W_{enc} がタスク情報のみを抽出し，語彙依存の情報を保持しない一方，W_{dec} が unembeddings と整合させて言語化を実現していることを示唆する．

DH は誘導ヘッドと独立である. 情報除去挙動および DH が誘導ヘッドから独立であることを示すため，各ヘッドに対し，ラベルトークンから最終トークンへの注意スコア総和を誘導マグニチュードとして算出し，共分散流束アブレーションの結果と併せて可視化した（図 7）．その結果，誘導ヘッドと DH が同じ層付近で現れるにもかかわらず，両者はほとんど重ならないことが分かった．これは，DH およびその情報除去機構が，ICL により新たに誘発される独立した操作であることを明確に示している．

表 2: フィルタの一部を凍結したままでのファインチューニング結果．

Trained Part	Both	W_{enc}	W_{dec}
Accuracy	0.88	0.00	0.84

3)　§3.2 参照．冗長性が十分除去され，タスク情報が第 1 主成分へ align すると，寄与的除去が離心率を増加させる．