大規模言語モデルにおける In-context Learning の推論回路

趙羽風 [1,☆]　加藤万理子 [1]　坂井吉弘 [1]　井之上直也 [1,2]
[1] 北陸先端科学技術大学院大学　[2] 理化学研究所　☆ 主な貢献者, yfzhao@jaist.ac.jp

概要

In-context Learning (ICL) は，言語モデルにおける新たな少数ショット学習パラダイムとして注目されているが，その内在的メカニズムは十分に解明されていない．本研究では，ICL の推論ダイナミクスを 3 つの基本操作に分解し，それらを基盤として推論回路を構築した上で精密な測定を行い，従来の研究で観察されてきた現象を統一的に説明することを試みた．さらに，提案した回路を無効化するアブレーション分析の結果，ICL の性能が顕著に低下することが確認され，提案した推論回路が ICL の主要なメカニズムであることが示唆された[1)]．

1　はじめに

In-context Learning (ICL) [1, 2] は，言語モデル (LM) における少数ショット学習パラダイムである．ICL では，通常，図 1 のような入力列 $[x_1][s_1][y_1][x_2][s_2][y_2]\ldots[x_q][s_q]$ を構築する．ここで，x_i は入力テキスト ($x_{1\sim k}$ はデモ，x_q はクエリと呼ばれる)，y_i はラベルトークン，s_i は，その直後にラベルトークンが来ることを意味する記号 (以後「予兆トークン」; 例えば “Label: ” 内のコロン) である．この入力列を LM に入力することで，次のトークンの予測確率を計算し，その予測確率を x_q のラベル予測値として利用する．

ICL は広い関心を集めており，先行研究では，ICL のメカニズムの理論的または実証的な説明が試みられてきた [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]．しかしながら，大規模言語モデル (LLM) の ICL における推論ダイナミクスや興味深い推論現象を十分に理解できる包括的な説明は未だに存在しない．

そこで本論文では，ICL の推論ダイナミクスとして，3 つの基本操作からなる推論回路を提案する．具体的には，ICL の推論プロセスを図 1 のように分割する．**Step 1: 入力テキストエンコード**．各入力テキスト x_i を，対応する予兆トークン s_i の隠れ状態に線形表現としてエンコードする．**Step 2: セマンティクス統合**．各デモについて，Step 1 で作られた s_i の内部表現を，対応するラベルトークン y_i の隠れ状態と統合する．**Step 3: 特徴検索とコピー**．クエリ s_q の内部表現に類似する，Step 2 で作られたラベルトークン $y_{1:k}$ の内部表現を，タスク関連部分空間内で検索し，それをクエリ s_q の内部表現と統合する．Step 2-3 は ICL の主要メカニズムである Induction Circuit に相当するが，これらの先行研究は人工的な実験に留まっており，より大規模で複雑な LLM の推論ダイナミクスを説明できるかは自明でない．

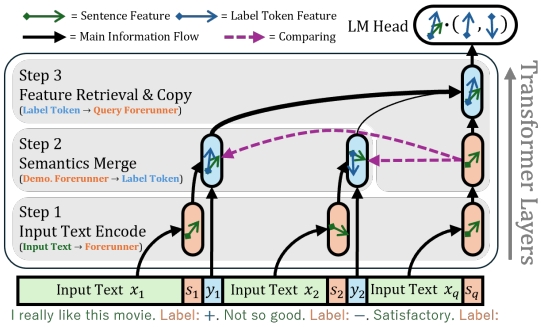

図 1　提案する ICL の 3 ステップ推論回路．**Step 1**: LM は各入力テキストをエンコードする．**Step 2**: デモの内部表現を，対応するラベルセマンティクスと統合する．**Step 3**: クエリの内部表現に類似する統合ラベルの内部表現を検索し，それをクエリの内部表現にコピーする．

本研究の貢献は，(1) ICL の推論ダイナミクスを説明する 3 ステップの推論回路を提案し，LLM を用いて各ステップの存在を実証的に確認したこと (§2, 3, 4)，(2) アブレーション分析により，提案した回路が主要な役割を果たしていることを確認したこと (§5)，(3) ICL の推論プロセスの既存研究で観察された現象を解明するためにより詳細な測定を実施したこと (§2, 3, 4, 付録 B)，である．

1.1　実験設定

最新の LLM である Llama 3 (8B, 70B) [14]，Falcon (7B, 40B) [15] を用いた．特に断りのない限り，本論

1)　本論文の完全版は ICLR 2025 に投稿されており，https://arxiv.org/abs/2410.04468 からアクセスできる．

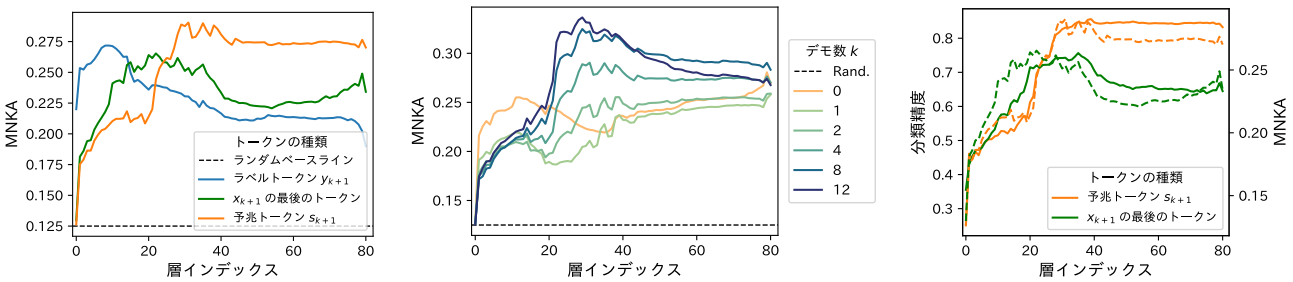

図 2　各層における隠れ状態の x_{k+1} のエンコード強度 (MNKA で測定). 左: x_{k+1} の最後のトークン，s_{k+1}，y_{k+1} の隠れ状態におけるエンコード強度. 中: s_{k+1} における異なる k のエンコード強度. 右: 隠れ状態を用いて訓練された重心分類器のテスト結果. **実線**: セントロイド分類の精度，**点線**: MNKA.

文の結果は Llama 3 70B に基づくものである．これは，その深く狭い構造が階層的な推論ダイナミクスを示しやすいためである．また，文分類データセット SST-2 [16], MR [17], Financial Phrasebank [18], SST-5 [16], TREC [19, 20], AGNews [21] を用いて ICL 形式のテスト入力を構築した．特に断りのない限り，ICL への入力には $k = 4$ 個のデモを使用し，これらの結果の平均値を報告する．その他のモデルの結果，テストデータのサンプリング方法やプロンプトテンプレート等の詳細は，完全版論文[1] を参照されたい．

2　Step 1: 入力テキストエンコード

本節では，s_i の隠れ状態内で，タスクに関して線形分離可能な x_i の内部表現が生成されていることを確認する．この線形表現は，Induction Circuit に基づいた ICL を説明するための重要な基盤であり，簡略化されたモデルに基づく既存研究 [8, 22, 4] でも，この線形表現の存在が仮定されている．

入力テキスト x_i は，予兆トークン s_i の内部表現にエンコードされる． 左記を確かめるため，k 個のデモとテキスト x_{k+1}，予兆トークン s_{k+1}，ラベル y_{k+1} からなる入力事例を作成し，x_{k+1} の情報が s_{k+1} の隠れ状態に存在するかを調べる．エンコード強度の指標として，最新のエンコーダ LM の一つである BGE M3 [23] を用いて x_{k+1} の埋め込み表現を生成し，これと s_{k+1} の隠れ状態の間の **M**utual **N**earest-neighbor **K**ernel **A**lignment (MNKA)[2] を計算する．さらに対照群として，s_{k+1} だけでなく，y_{k+1}，および x_{k+1} の最後のトークンを用いる．実験結果を図 2 (左) に示す．対照群と比較して，s_{k+1} が最も高い MNKA を示していることから，LM が s_{k+1} の隠れ状態において入力テキスト x_{k+1} をエンコードしていることが示唆される．これは，y_{k+1} が x_{k+1} の情報を収集すると示唆した先行研究 [25] を補完する重要な知見である．

エンコーディングはデモにより強化される． x_{k+1} のエンコードに対する先行文脈の影響を調査するため，異なる k で s_{k+1} のエンコード強度を測定した (図 2 (中))．直感的には，先行文脈が長くなるほどノイズ，すなわち x_{k+1} 以外の情報が加わるため，s_{k+1} の内部表現の品質が劣化すると予想されるが，これに反して，k の増加がエンコードの強化に繋がることがわかった．この発見は，LM が (1) 文脈情報を活用して x_{k+1} のエンコードを強化し，(2) 異なるデモを正確に分割していることを示している．

予兆トークンの内部表現は線形分離可能でタスク関連性がある． 未知の 256 個の入力事例を用い，s_{k+1} (対照群として x_{k+1} の最後のトークンも用いる) の各層の隠れ状態を特徴量とした重心分類器を訓練し，ラベル y_{k+1} を予測させる [26]. 結果を図 2 (右, 実線) に示す．s_{k+1} の非常に高い分類精度より，重心分類器が線形であることから，タスクのセマンティクスに関する部分空間における，隠れ状態の高い線形分離可能性が示唆される．また，分類精度と MNKA は似た傾向を示しており，先の MNKA の実験の信頼性を裏付けている．

その他，(1) LM が簡単な入力テキストに対してより良い内部表現を生成すること，(2) ICL 入力における x_{k+1} の位置がエンコーディングに影響を与えることを確認している (詳細は付録 A を参照).

3　Step 2: セマンティクス統合

本節では，s_{k+1} から y_{k+1} への情報のコピーのプロセスを調査・測定する．また，s_{k+1} と y_{k+1} の相互作用を調べ，ICL が，デモ内の誤ったラベルに対してなぜ頑健であるのかの説明を試みる．

s_{k+1} の内部表現が y_{k+1} にコピーされる． コピープロセスの存在を確認するために，層 l における s_{k+1} の隠れ状態 (コピー元) と層 $(l+1)$ におけ

2)　直感的には，異なる空間における特徴量の間の類似度として解釈できる．詳細は [24] を参照されたい．

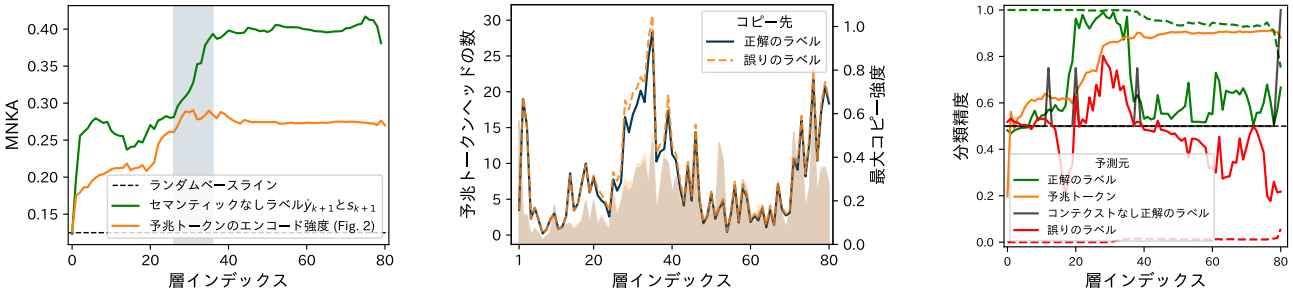

図 3 s_{k+1}からy_{k+1}への情報のコピー強度. **左**: s_{k+1}(コピー元) と次の層の$\hat{y}_{k+1}$(コピー先) との間の MNKA. **中**: **折れ線:** マークされた予兆トークンヘッドの数; **色付き領域:** s_{k+1}からy_{k+1}への最大 Attention Score. **右**: SST-2 および MR における正解のy_{k+1}と誤りのy_{k+1}の隠れ状態に基づいて予測された重心分類器の精度. (**実線**：s_{k+1}の隠れ状態で訓練された分類器 $\mathcal{C}_f$ による予測. **点線**：y_{k+1}の隠れ状態で訓練された分類器 $\mathcal{C}_l$ による予測.)

る$\hat{y}_{k+1}$の隠れ状態 (コピー先) との MNKA を計算する. ここで，$\hat{y}_{k+1}$は無意味な記号であり，これにより元のラベルのセマンティクスに影響されることなくs_{k+1}の情報のコピーの有無を純粋に評価する. 図 3 (左) に示すように，MNKA は徐々に増加し，s_{k+1}へのエンコーディングが終了した後に急上昇する様子が観察された. これは，s_{k+1}からy_{k+1}へのコピープロセスの存在を示唆すると同時に，入力テキストx_iは Step 1 でs_{k+1}にエンコードされ，さらにこの情報がy_{k+1}に統合されることを示している.

s_{k+1}の内部表現は選択性なくコピーされる. y_{k+1}からs_{k+1}への Attention Score $\alpha_{s_{k+1}\to y_{k+1}}$ を抽出し，$\alpha_{s_{k+1}\to y_{k+1}} \geqslant 5/n_{k+1}$ (n_{k+1}: y_{k+1}以前のトークンの長さ) であるヘッドを「予兆トークンヘッド」として，各層でそれらを数える. 図 3 (中，正解のラベル) に示すように，図 3 (左) のコピー期間と一致するピークが確認された. さらに，コピー先のラベルトークンの影響を調査するために，y_{k+1}を誤ったラベルトークンに置換した. 図 3 (中，誤りのラベル) に示すように，その結果は正しいラベルの場合とほぼ同一であった. これは，s_{k+1}とコピー先のラベルの意味的一貫性が無視され，単にs_{k+1}の内部表現がy_{k+1}に統合されていることを示唆する.

y_{k+1}の隠れ状態はテキスト表現とラベルセマンティクスの統合表現である. 上記の結果を踏まえ，コピーされるテキストの内部表現と元のラベルのセマンティクスがどのような相互作用を経てy_{k+1}の隠れ状態に合成されるのかを解析する. 第一に，s_{k+1}の情報がy_{k+1}の隠れ状態にどのような形で含まれるかを確認するため，ラベルy_{k+1}を予測する重心分類器 $\mathcal{C}_f$ を構築し，s_{k+1}の隠れ状態を用いて訓練する. 推論時にはy_{k+1}の隠れ状態を用いることで，テキストの情報の有無を確認する. 図 3 (右，実線，正解のラベル) に示すように，コピー処理中に高い分類精度が達成されることから，s_{k+1}のテキストの特徴が部分的かつ線形に，ラベルトークンから検出可能であることを示唆している. 誤ったラベルトークンをy_{k+1}として用いた場合 (誤りのラベル) も同様の傾向であり，前述の知見と一貫している. 第二に，y_{k+1}の隠れ状態にラベルの情報が含まれているかを確認するために，先の重心分類器をy_{k+1}の隠れ状態を用いて訓練し，$\mathcal{C}_l$ を得る. 図 3 (右，点線) に示すように，こちらはラベルの正誤に応じて精度が極端に変化しており，ラベルの情報がラベルトークン内に残存していることを示唆する. 以上より，ラベルトークンの隠れ状態は，テキスト表現とラベルのセマンティクスの結合表現であるといえる.

ラベルノイズの除去は，ラベルのセマンティクスとテキスト表現の重なり部分で行われる. 図 3 (右，実線) より，y_{k+1}の隠れ状態は，確かにs_{k+1}の情報を含むことが分かったが，コピー先のラベルの正誤により，ラベルの予測精度が改善もしくは悪化することも観察されている. これは，テキストの内部表現においてラベルのセマンティクスと一致する情報が，コピー先のラベルトークンによって選択的に強化されやすいことを示唆しており，逆もまた然りである. この特徴選択性は，ラベルのセマンティクスとテキスト特徴に関する情報が個別かつ線形に抽出可能であるという前述の実験結果より，これら二種類の情報が隠れ状態の異なる部分空間に位置し，予兆トークンヘッドのアテンション操作によって線形に統合されることによって起こると考えられる. さらに，予兆トークンヘッドの非選択性を考慮すると，この特徴選択性は，ラベルのセマンティクスとテキスト特徴の部分空間の重なりにおける特徴ベクトルの算術的相互作用に由来することが直感的

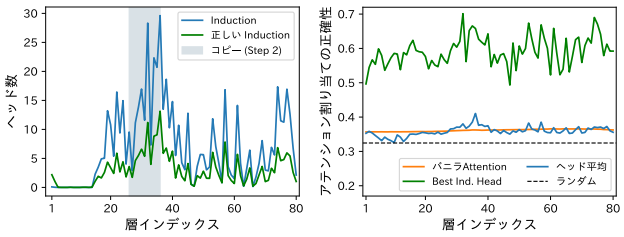

図 4　Induction Head の測定. 左：層ごとの Induction Head と正しい Induction Head の数. 右：アテンション割り当ての正確性 (正しいラベルトークンへの割り当ての合計を，すべてのラベルトークンへの割り当てで正規化した値). **バニラ Attention** は完全な次元で直接計算されたアテンションスコア，**ヘッド平均**はすべてのヘッド間で平均化されたアテンションスコア，**Best Ind. Head** は最も正確性の高いアテンションヘッドのスコアを示す.

に理解できる. この相互作用により，ICL がラベルノイズに対して安定するという現象 [9] が説明できる. さらに，[27] によると，大規模なモデルほどラベルノイズに対する安定性が低いことが示されている. これは，より大きな隠れ次元を持つ大規模モデルでは，ラベルのセマンティクスとテキスト表現の部分空間の重なりが減少し，相互作用が低下することが原因と推測される.

4　Step 3: 特徴検索とコピー

この節では，s_q の特徴と類似した $y_{1:k}$ の特徴を取得し，取得した特徴を s_q に書き込む Induction Head の存在を検討する. また，このプロセスにおける Multi-head Attention の必要性についても主張する.

ごく少数のヘッドの部分空間でのみ正しい Induction が起こる.　図 3 (中) と同様に，(1) すべてのラベルトークン $y_1, \ldots, y_k$ から s_q への Attention Score の合計が $5k/n_t$ を超えるアテンションヘッドを Induction Head とし，(2) x_q の正解ラベルと一致するすべてのラベルトークンからの Attention Score の合計が $5k/|\mathbb{Y}|n_t$ を超えるアテンションヘッドを「正しい Induction Head」とし，各層でこれらを数える ($\mathbb{Y}$ はラベル空間). 図 4 (左) に示すように，Step 2 のコピー処理より後に単峰パターンが観察された. さらに，Induction Head の半数以上は「正しい」ものではなく，タスク固有の特徴類似性は，正しい Induction Head の部分空間 (低ランク遷移行列 $W_Q^{h\top} W_K^h$ によって定義される) でのみ捉えられ，s_q の内部表現に統合されることを示唆している. この現象をさらに分析するために，正解のラベルトークンに対するアテンションスコアが全体に占める割合を計算し，図 4 (右) に示す. 最良の正しい Induction Head は，高い正確性を示している一方で，対照群となる，変換やヘッド分割を伴わない通常のアテンションと，ヘッド間で平均化されたアテンションスコアについてはその値は低い. 平均値を考慮すると，大多数のアテンションヘッドはほぼランダムにラベルトークン情報を s_q にコピーし，予測がプロンプト内のラベル頻度に偏る原因となっている [28]. これは，隠れ状態の情報は ICL に対して十分であるが (図 2 (右) 参照)，最小限ではなく，冗長な情報がアテンションの類似性計算を妨げていると推測される.

表 1　各推論ステップに対する Attention 接続を，下位層から上位層へ一定の割合で除去した場合の Llama 3 8B における精度変動（%）. 小さい数値は，ランダムに同等量の Attention 接続を除去した場合の結果 (mean ± std) を示す.

#	切断された Attention Key → Query	切断されたレイヤーの比率 25%	50%	75%	100%
1	**無し** (4-shot ベースライン)	±0 (Acc. 68.55)			
	– Step1: 入力テキストエンコード –				
2	**Demo.** x_i → **Forerunner** s_i	−4.98 ±0.89 ± 0.00	−15.82 ±1.19 ± 0.02	−23.43 ±3.29 ± 1.87	−30.60 ±1.61 ± 0.01
3	**Query** x_q → **Forerunner** s_q	−13.87 ±0.16 ± 0.00	−21.10 ±0.08 ± 0.00	−24.74 ±0.47 ± 0.04	−28.38 ±0.55 ± 0.00
	– Step2: セマンティクス統合 –				
4	**Demo.** s_i → **Demo.** y_i	−2.24 ±0.00 ± 0.00	−3.45 ±0.18 ± 0.00	−3.39 ±0.10 ± 0.04	−3.42 ±0.18 ± 0.01
	– Step3: 特徴検索とコピー –				
5	**Demo.** y_i → **Query** s_q	−5.14 +0.03 ± 0.00	−10.03 ±0.08 ± 0.00	−11.36 +0.00 ± 0.00	−10.22 −0.06 ± 0.00
	参考値				
6	**0-shot**	−17.90 (Acc. 50.65)			
7	**ランダム予測**	−36.05 (Acc. 32.50)			

さらに，Induction Head の部分空間がタスク特有であり，デモがこの空間上で飽和していることを確認した (詳細は付録 A を参照).

5　まとめ

アブレーション分析. 提案された回路が ICL において支配的であることを示すために，提案された回路の各ステップに関連するアテンション接続を切断した場合の精度を調べ，表 1 に示す. その結果，回路に関係のない接続をランダムに除去した場合の対照実験結果と比較して，提案された回路によって指定された重要な接続を除去した場合，ICL の精度が大幅に低下し，回路の存在を支持する結果が得られた.

結論. 本研究は，3 つの基本操作からなる ICL の推論回路を提案し，その存在とともに，提案した推論回路が ICL の主要なメカニズムであることを示した. また，精密な測定を通じて既存の多様な現象の説明に成功した. 本研究が ICL の実践に新たな洞察をもたらすことを期待している. なお，本文に記載しきれなかった既存研究の現象の説明については付録 B を，本研究の Limitation は付録 C を参照のこと.

謝辞

本研究は，JST 創発的研究支援事業 JPMJFR232K，および JSPS 科研費 19K20332 の助成を受けたものです．

参考文献

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[2] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. **arXiv preprint arXiv:2301.00234**, 2022.

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

[4] Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. **arXiv preprint arXiv:2404.07129**, 2024.

[5] Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[6] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Pre-training to learn in context. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4849–4870, 2023.

[7] Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 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)**, pp. 12660–12673, 2023.

[8] Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 18878–18891, 2022.

[9] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 11048–11064, 2022.

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

[11] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. **arXiv preprint arXiv:2306.09927**, 2023.

[12] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 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.

[13] Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Explaining emergent in-context learning as kernel regression. **arXiv preprint arXiv:2305.12766**, 2023.

[14] AI@Meta. Llama 3 model card. 2024.

[15] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. **arXiv preprint arXiv:2311.16867**, 2023.

[16] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

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

[18] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. **Journal of the Association for Information Science and Technology**, Vol. 65, , 2014.

[19] Xin Li and Dan Roth. Learning question classifiers. In **COLING 2002: The 19th International Conference on Computational Linguistics**, 2002.

[20] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. In **Proceedings of the First International Conference on Human Language Technology Research**, 2001.

[21] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. **Advances in neural information processing systems**, Vol. 28, , 2015.

[22] 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.

[23] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. **arXiv preprint arXiv:2402.03216**, 2024.

[24] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. **arXiv preprint arXiv:2405.07987**, 2024.

[25] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 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**, pp. 9840–9855, 2023.

[26] Hakaze Cho, Yoshihiro Sakai, Mariko Kato, Kenshiro Tanaka, Akira Ishii, and Naoya Inoue. Token-based decision criteria are suboptimal in in-context learning. **arXiv preprint arXiv:2406.16535**, 2024.

[27] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. **arXiv preprint arXiv:2303.03846**, 2023.

[28] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In **International conference on machine learning**, pp. 12697–12706. PMLR, 2021.

[29] Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 10136–10148, 2023.

[30] Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. **arXiv preprint arXiv:2402.10738**, 2024.

[31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. **Communications of the ACM**, Vol. 64, No. 3, pp. 107–115, 2021.

本論文で省略または簡略化された情報については，完全版 (https://arxiv.org/abs/2410.04468) を参照．

A　補足実験

テキストの Perplexity が高いほどエンコードが劣化する. 異なる k における x_{k+1} の Perplexity と s_{k+1} のカーネルアライメントの相関を調査する．図 5 (上) は，$k = 0$ の場合に負の相関を示しており，これは，デモが与えられない場合，LM がより複雑な入力テキストに対して劣ったエンコードを生成することを示している．一方で，文脈内にデモが与えられた場合 (図 5，下)，負の相関は消失し，LM が文脈内のデモを利用してより複雑なサンプルを効果的にエンコードしていることを示唆している．

入力テキストのエンコーディングは位置に偏っている. 理想的には，類似した x の内部表現は，入力内での位置に関係なく非常に類似しているべきであり，分類のための操作をサポートする必要がある．これを検証するために，各エンコーディング対象について，異なる数のデモを持つ予兆トークンの隠れ状態を抽出する．その後，同じ対象または異なる対象に対する隠れ状態のすべての可能なペア間のコサイン類似度を計算する．図 6 に示されるように，同じターゲットにおける全体的な類似度は異なるターゲットにおける類似度よりも高いが，両者とも位置が近い場合に特に高くなる．すなわち，クエリに近いデモが ICL に対してより強い影響を与える．

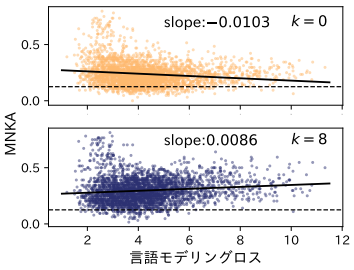

図 5　Llama 3 70B の第 24 層におけるエンコーディング強さを，x_{k+1} の言語モデル損失に対して，(上) $k = 0$ および (下) $k = 8$ の場合で比較する．

Induction Head の部分空間におけるデモの飽和. 図 7 では，遷移行列 $W_Q^{h\top} W_K^h$ と主成分分析を用いて，Induction Head の部分空間にマッピングされたデモのラベルトークンの内部表現を可視化している．これにより以下が示される：(1) 正しい Induction Head と比較して，誤った Induction Head ではラベルの内部表現が線形分離不可能な形でマッピングされやすい．(2) デモの初期段階 (k = 1 → 2) では，新しいデモが与えられると，クエリに対するアテンション割り当ての形態が大きく変化する．一方，後半 (k = 15 → 16) では，アテンション割り当ての形態は安定している．

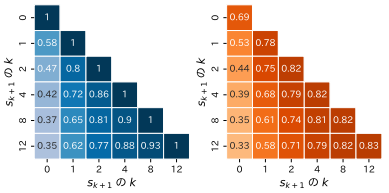

図 6　層 24 における異なる位置の ICL 隠れ状態の類似性．左: 同一のクエリ，右: 異なる 2 つのクエリ (SST-2 上)

B　先行研究で観察された推論現象の解釈

難易度に基づくデモ選択. 付録 A で，ゼロショットシナリオでは Perplexity の高いテキストはエンコードが難しく，低い Perplexity のデモを選ぶことで ICL のパフォーマンスが向上することが示されている．また，デモが増えることで，LM は複雑な入力を処理できるようになり，後に難易度の高いデモを入力することが有益である可能性が説明される．これは PPL-ICL [29] と ICCL [30] の観察結果を説明している．

予測バイアス. (1) 位置バイアス：付録 A に示すように，入力テキストが近いほどエンコードが類似し，クエリの近くにあるラベルトークンはクエリにより類似した情報を持つため，Induction Head においてより多くの注意が割り当てられ，予測に対する影響が大きくなる．**(2) 頻度バイアス**：図 4 (右) に示すように，いくつかのヘッドはラベルに対して選択性がなく，ラベルトークンからクエリへの平均的なコピー処理を引き起こし，たとえその貢献が小さくても，デモのラベル頻度に対する予測バイアスを引き起こす．これらの 3 つのバイアスは [28] によって観察され，キャリブレーション法によって除去可能である．

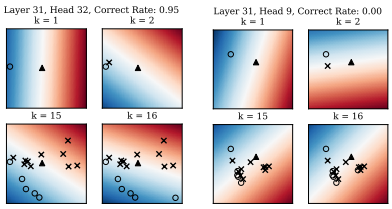

図 7　k 個のデモのラベルの内部表現を可視化したもの．左 4：正しい Induction Head，右 4：誤った Induction Head，SST-2 のサンプル 1 つに基づく．○:「ポジティブ」ラベル，×:「ネガティブ」ラベル，▲: ゼロベクトル．色：クエリに割り当てられた Score．

デモの役割と飽和性.　デモが ICL の性能を向上させることは広く知られている．この性能向上を 2 つの要素に分解する：**(1)** デモが初期層のエンコードを改善する (図 2)，**(2)** より多くのデモがより大きなラベルトークンの閉包を提供し，より正確なアテンション割り当てを可能にする (付録 A)．ただし，この閉包のボリュームはデモに対して劣モジュラ性を持つため，デモに対する ICL 性能の飽和を引き起こす．

誤ったラベルの影響.　ラベルノイズが勾配ベースの学習 [31] よりも ICL [9] において害が少ないことはよく知られている．§3 で説明したように，ICL はラベルノイズに対して ICL を安定化させるためにラベルをデノイズすることを意味するが，その効果は高い次元数によって弱められる．

C　Limitation

(1) デモにおいてクエリの真のラベルが提供されない場合には，Step 3 が発動せず，モデル内部のパラメタを利用した推論が行われる可能性があり，本研究の結論が適用できない場合がある．これらは通常，ICL ではなく IWL (**In-weight Learning**) と呼ばれ，標準的な ICL との間に顕著な違いや対立があることが以前の研究 [8, 22] で指摘されている．したがって，これは合理的かつ必然的に別の議論を必要とする．本論文は ICL 条件下でのモデルの推論挙動を説明しており，IWL の推論挙動については今後の研究に委ねる．(2) 本研究は分類タスクに焦点を当てているが，我々はその発見が非分類タスクにも適用できると考えており，そのギャップを埋めるためにはさらに努力が必要である．