Attention Head 間の共有構造を考慮したグループ単位の介入による LLM の迎合現象抑制

原 大晟 [1]　吉丸 直希 [2]　波多野 賢治 [1]
[1] 同志社大学 文化情報学部　[2] 同志社大学大学院 文化情報学研究科
{cgjh0027@mail4, yoshimaru@mil, khatano@mail}.doshisha.ac.jp

概要

大規模言語モデル（LLM）は高い性能を示す一方，事後学習の副作用として，ユーザの誤った主張や質問に同調してしまう迎合現象が生じる．この現象の抑制手法として，モデル内部の Attention Head の一部に対して追加学習を行う手法が提案されているが，この手法は各 Head が独立した機能単位として振る舞うという前提に基づいている．しかし，近年の LLM では内部構造が複雑化し，複数の Attention Head や構成要素が情報を共有して処理を行うため，迎合に関わる機能も複数の Head に分散して現れる．本研究では，これらの Head を一つの機能グループとして扱い，グループ単位で迎合現象に介入する手法を提案する．評価実験の結果，提案手法は既存手法と比較して，より高い迎合抑制効果を示した．

1　はじめに

近年，大規模言語モデル（LLM）は高度な能力を有している．これらのモデルは大規模な事前学習によって一般的な言語知識を獲得した後，ユーザとの円滑な対話を行うために Instruction Tuning [1] や RLHF [2] といった事後学習を行う．これらの学習過程により，LLM はユーザとの円滑な対話が可能になり，多様な応用場面での利用が進んでいる [3, 4, 5]．一方で，事後学習によって獲得される性質の中には，信頼性や安全性に悪影響を及ぼす挙動が含まれることが指摘されている [6, 7]．

その例の一つが Sycophancy（迎合現象）である．迎合現象とは，モデルが正しい知識を保持しているにもかかわらず，ユーザの誤った主張や質問に対して不適切に同意し，正しい回答から逸脱した誤った応答を生成する現象である [7, 8, 9]．この現象はモデルの知識不足ではなく，事後学習においてユーザに寄り添う応答や対話が過度に最適化される結果として生じる [7]．そのため，ユーザの発言が事実と異なっていても，それを否定せずに肯定する傾向が意図せず強化される [10]．このような現象は，専門的な知識が求められる医療相談や公平性が重視されるファクトチェックのような場面では，モデルがユーザに誤った情報を提供することで利用者の意思決定に悪影響を及ぼす恐れがある [11]．したがって，LLM の一般的な能力を維持しつつ，迎合現象を抑制することが必要である．

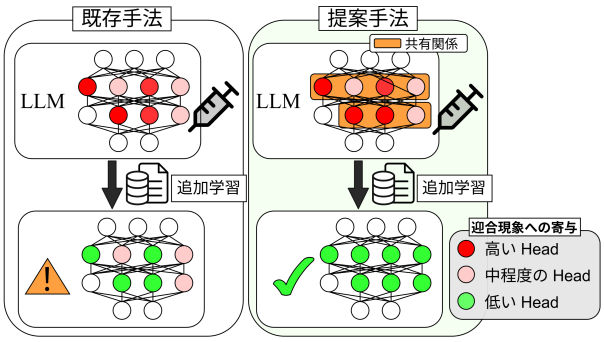


図 1: 既存手法（左）と提案手法（右）の比較

迎合現象を抑制する既存手法として，モデル内部の Attention Head 単位で迎合現象への寄与を分析し，寄与の高い部分のみに追加学習を行う手法がある [12]．この手法では，各 Head が独立した機能を行うという仮定に基づいている．しかし，近年の LLM では Grouped Query Attention（GQA）構造などが導入され，複数の Head がパラメータを共有して処理を行う [13]．このような共有構造では，迎合現象に関わる影響が特定の Head に集中せず，共有パラメータを通じて複数の Head に分散する傾向がある．実際に，3.1 節の基礎分析では，迎合への寄与は Head 単位では分散しており，同一 Layer 内で共有する Head のグループ単位で影響が集中することが確認された．この結果は，Head 単位で介入を行う既存手法では，迎合に関わる機能を十分に修正できず，抑制効果が不安定になる．

本研究では，この課題に対し，Attention Head 間の共有関係に着目した新たな迎合抑制手法を提案する．具体的には図 1 より，パラメータを共有する複数の Head を一つの機能的役割を担うグループとして扱い，グループ単位で迎合現象に介入する．これにより，モデル構造が複雑化した LLM に対しても，安定した迎合抑制を実現する．

2 関連研究

迎合現象の抑制手法として，モデル内部の構造を調整する Supervised Pinpoint Tuning（SPT）がある [12]．SPT は，Path Patching [14] という手法を用いて，各 Attention Head が迎合的な応答生成に与える寄与度を算出する．具体的には，各 Head の出力を非迎合時の値に差し替え，最終出力の変化を算出することで迎合現象に寄与する部分を特定する．寄与の高い Head のみを追加学習の対象とすることで，モデルの一般性能を維持しつつその抑制を実現している．

しかし，SPT は各 Head が機能的に独立しているという前提の下で設計されている．近年の LLM では GQA 構造が導入され，複数の Head がパラメータを共有しながら処理を行う [13]．このような共有構造では，Head 間の挙動は独立ではなく密接に関係しており，迎合現象への影響が複数の Head に分散して現れる．その結果，Head 単位で介入を行う SPT では迎合に関わる機能を十分に修正できず，単一の Head に対する局所的な介入が，他の Head を介した情報伝播によって相殺される可能性があるため，抑制効果が不安定になるという課題が生じる．

本研究では，この課題に対し，SPT の前提が成立しないモデルの内部構造に着目し，密接に関係する複数の Head を一つの機能的役割を担うグループとして扱う手法を提案する．これにより，構造が複雑化したモデルに対しても，安定した迎合現象の抑制を実現することを目指す．

3 提案手法

本研究では，GQA 構造を持つモデルにおいて，迎合現象の寄与が Attention Head 間に分散する点に着目し，個別の Head ではなく，パラメータを共有する複数の Head を介入単位とした迎合抑制手法を提案する．

表 1: 各モデルにおける Attention Head 寄与の尖度

対象モデル	尖度 (Kurtosis)
Llama-2-7B-Chat [15]	670.05
Mistral-7B-Instruct-v0.3 [16]	84.59
Llama-3-8B-Instruct [17]	72.15
Qwen-2.5-7B-Instruct [18]	26.91

3.1 基礎分析

本節では，SPT が複雑化したモデルにおいて迎合現象を安定して抑制できない要因を明らかにするため，モデル内部における迎合寄与の分布について分析を行う．Path Patching による各 Attention Head の迎合寄与度を算出し，分布の偏りを示す尖度を計算した結果を表 1 に示す．尖度が高いほど寄与が一部の Head に集中しており，低いほど寄与が複数の Head に分散していることを意味する．分析の結果，Llama-2-7B-Chat [15] では尖度が非常に高く，迎合への寄与が少数の Head に集中していることが確認された．これに対し，GQA 構造を持つモデルの Mistral-7B-Instruct-v0.3 [16]，Llama-3-8B-Instruct [17]，Qwen-2.5-7B-Instruct [18] では比較的低く，迎合寄与が一部の Head に集中せず，複数の Head に分散していることが明らかになった．可視化結果については付録 A に掲載するが，迎合に関わる機能が一部の Head に集中せず，同じ Layer 内の広範囲に分散していることが視覚的にも確認できる．この傾向は，GQA 構造において同じ Layer 内の複数 Head が同じ情報を参照していることが影響している可能性が高い．つまり，情報を共有する Head 間で迎合的な挙動をする機能的な役割を分担している．したがって，個別の Head を独立に扱う SPT ではこのような共有関係を十分に捉えきれない可能性があり，モデルの構造的特徴に基づいた手法が必要である．

3.2 共有構造に基づく迎合抑制

3.1 節の分析を踏まえ，提案手法では GQA 構造に対応可能な抑制手法を提案する．分析の結果，迎合現象は一部の Head に集中せず，複数の Head に分散して現れることが確認された．GQA 構造では，複数の Head が同一の Key および Value を共有しながら処理を行う [13]．この構造は，各 Head が参照する文脈情報や注意を向ける対象が共通することを意味しており，迎合現象に関わる表現も共有関係にあ

表 2: 各モデルの迎合抑制効果と一般性能の比較

	Sycophancy		General Ability					
	Truthfulness		StrategyQA		GSM8K		HumanEval	
Models (Ratio: 4%)	**Acc.**	Δ	**Acc.**	Δ	**Acc.**	Δ	**Acc.**	Δ
Llama-3-8B [17]	49.76	-	56.51	-	**77.33**	-	**57.71**	-
+ SPT	71.51	+21.75	**60.92**	+4.41	54.28	-23.05	54.48	-3.23
+ **提案手法**	**76.84**	+27.08	59.30	+2.79	48.22	-29.11	52.68	-5.03
Qwen-2.5-7B[18]	**86.70**	-	**73.41**	-	87.49	-	70.73	-
+ SPT	81.03	-5.67	70.70	-2.71	87.26	-0.23	74.39	+3.66
+ **提案手法**	82.99	-3.71	72.45	-0.96	**87.72**	+0.23	**74.44**	+3.71
Mistral-v0.3-7B[16]	61.29	-	**66.03**	-	**49.20**	-	**39.02**	-
+ SPT	77.23	+15.94	63.49	-2.54	24.18	-25.02	32.93	-6.09
+ **提案手法**	**82.09**	+20.80	65.37	-0.66	28.13	-21.07	32.94	-6.08

る複数の Head に分散して現れると考えられる．3.1 節の基礎分析においても，迎合現象の寄与は一部の Layer 内の広範な Head に分布しており，迎合現象に関わる機能がグループ単位の役割として保持されている可能性が示唆された．

そこで本研究では，共有関係にある複数の Head を，一つの機能的役割を担うグループとして定義し，グループ単位で迎合現象に介入する．この方法により，迎合寄与が複数の Head に分散する場合であっても，安定した抑制が可能となる．

モデル全体の Attention Head 数を N_{a}，Key/Value Head 数を N_{kv} とすると，一つのグループに含まれる Head 数 H_{g} は次式で与えられる．

$$H_{\mathrm{g}} = \frac{N_{\mathrm{a}}}{N_{\mathrm{kv}}} \tag{1}$$

同一の Key/Value を共有する H_{g} 個の Head の集合をグループとして扱い，迎合現象抑制の対象とする．

次に，定義した Head グループが迎合現象に与える影響を定量化するため，Path Patching をグループ単位に拡張して適用する．迎合現象が発生する入力 x に対し，グループ g の出力を，迎合現象が発生しない入力における理想的な出力 $\bar{g}$ に差し替え，最終出力の変化を算出する．グループ g の迎合現象への寄与度 $I(g)$ は次式で定義する．

$$I(g) = \frac{S(y \mid x, g \leftarrow \bar{g}) - S(y \mid x)}{S(y \mid x)} \tag{2}$$

上記の式は，ある Head グループを迎合しない状態に変更した時に，迎合スコアがどれだけ減少するかの変化率を表す．ここで $S(y \mid x)$ は元の入力に対する迎合スコア，$S(y \mid x, g \leftarrow \bar{g})$ はグループ g の出力を差し替えた際の迎合スコアである．$I(g)$ の負の値が大きいほど，そのグループが迎合的な出力に強く寄与していることを示す．本手法では，全グループの $I(g)$ を算出し，値が小さい上位 K 個のグループを追加学習の対象として選択する．

最後に，Path Patching により選択した Head グループに追加学習を行う．学習データには，モデルの正答に対してユーザが誤った反論を行う対話例を用いて，モデルが迎合せずに元の回答を維持しつつ正確な応答ができるように教師あり学習を行う．

4 評価実験

提案手法の有効性を確認するために，既存手法と提案手法のそれぞれを用いて迎合抑制を適用した後の言語モデルの挙動を比較する．各手法における迎合抑制と一般性能への影響を考慮した上で比較することで，実用に耐えうる手法であるかを検証する．

4.1 実験設定

評価ベンチマークとして，迎合現象の抑制を評価するために SycophancyEval [7] を使用する．これは，ユーザが事実に反する主張や前提を含めて質問した際，モデルがその誤りに同調せず，正しい回答を維持できるかを評価する．本研究ではその正答率を Truthfulness として用いる．モデルの一般性能を確認するために，複数の推論ステップを必要として論理的思考力を問う StrategyQA [19]，数学能力を測る GSM8K [20]，および Python コード生成能力を評価する HumanEval [21] を用いる．

評価対象の言語モデルとして，3.1 節の基礎分析において迎合寄与の分散が確認された Llama-3 および Qwen-2.5，Mistral-v0.3 を用いる．追加学習の対象数は，先行研究において迎合抑制と一般性能の維持が両立することが示された知見に基づき，モデル全

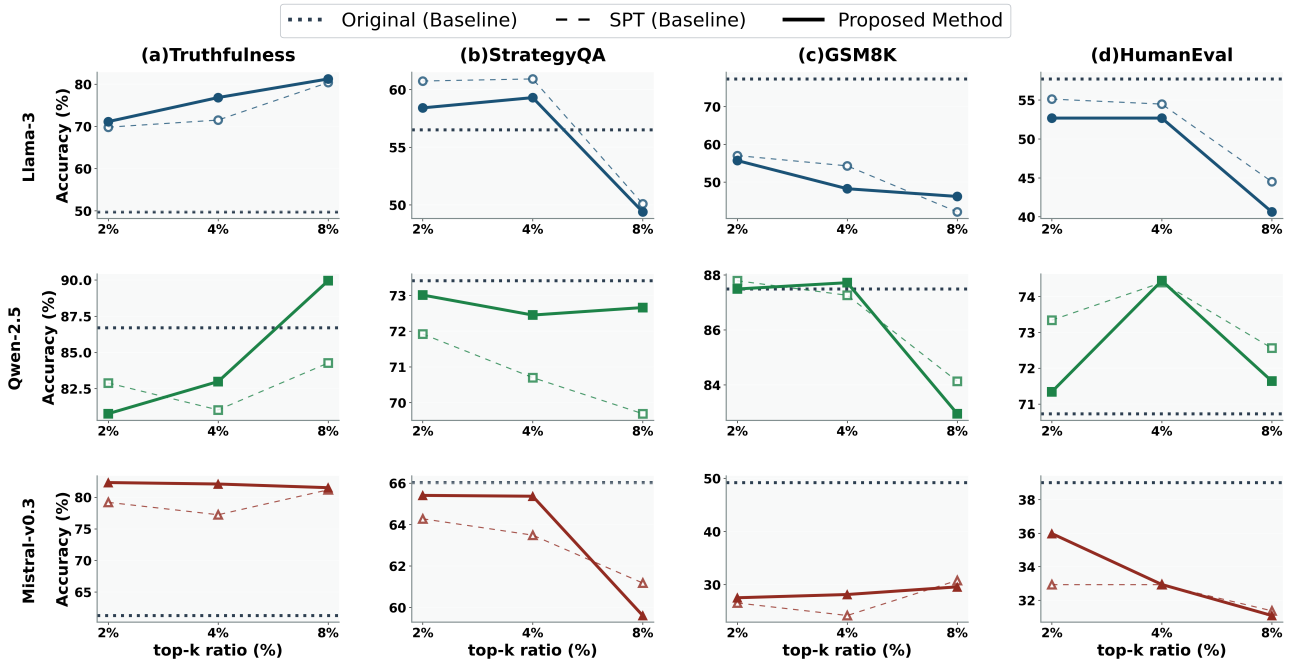


図 2: 各モデルにおける追加学習時の top-k の変化に伴う各指標の精度比較

体の Head の約 4%相当とする [12]．提案手法および比較手法はいずれも同一の学習設定で追加学習を行い，ハイパーパラメータは SPT [12] に従った．

4.2 結果・考察

表 2 の結果から，提案手法は SPT よりも高い抑制効果を示す傾向が確認できる．この結果は，グループ単位で介入を行う提案手法が特定のモデル構造に依存せず，迎合の寄与が分散するモデルに対しても一貫して有効であることを示唆している．一方で，一般性能の維持という観点では課題が残る．GSM8K や HumanEval において追加学習後の性能は元のモデルに比べて一定の低下が確認された．ただし，提案手法はその低下を SPT と同程度に抑えている．以上の結果から，迎合現象の抑制と一般性能の維持の両立は難しく，提案手法の課題である．

次に，一般性能と迎合抑制の関係を詳細に理解するため，追加学習の対象割合を 2%，4%，8% に変化させた際の影響を分析する．図 2 は，追加学習する Attention Head の割合を変化させた場合の推移を示しており，迎合抑制と一般性能の間に生じるトレードオフの傾向が確認できる．図 2(a) より，割合を増加させるにつれて Truthfulness が向上する傾向がある．しかし，一般性能では追加学習の対象が増えるほど性能が低下する傾向が確認された．この低下の程度はモデルごとに異なり，特に Qwen-2.5 においてはその傾向が顕著である．8%の更新で StrategyQA と HumanEval などのスコアが大幅に低下しており，追加学習の割合を増加させた際に，影響を強く受けやすいモデルであることが示唆される．

以上の結果から，提案手法は迎合現象の抑制において，SPT よりもモデルの構造に依存せずに安定した汎用性の高い手法である．しかし一方で，一般性能の維持はモデルの種類などによって変化するため，安定して維持することが難しい．このことから，迎合現象の抑制と一般性能の完全な両立は依然として難しく，本手法においても今後の課題として残る．

5 おわりに

本研究では，迎合現象への寄与が分散しやすい内部構造を持つ LLM において，Attention Head 間の共有構造に着目し，複数の Head をグループとして扱う迎合抑制手法を提案した．評価実験の結果，提案手法は先行研究と比較して高い迎合抑制効果を示し，複雑化したモデル構造に対しても有効であることを確認した．

今後の課題として，迎合抑制と一般性能のさらなる両立が挙げられる．具体的には，グループ内に含まれる各 Attention Head の役割をより詳細に分析し，一般性能への影響を最小化する介入方法へと拡張する必要がある．

謝辞

本研究の一部は，JSPS 科研費 JP25H01167 によるものである．

参考文献

[1] Jason Wei, Maarten Bosma, Vincent Y. Zhao, et al. Finetuned language models are zero-shot learners. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2022.

[2] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In **Advances in Neural Information Processing Systems (NeurIPS)**, Vol. 35, pp. 27730–27744, 2022.

[3] Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. Large language models are built-in autoregressive search engines. In **Findings of the Association for Computational Linguistics (ACL)**, pp. 2666–2678, 2023.

[4] Swaroop Mishra and Elnaz Nouri. HELP ME THINK: A simple prompting strategy for non-experts to create customized content with models. In **Findings of the Association for Computational Linguistics (ACL)**, pp. 11834–11890, 2023.

[5] Jinyue Feng, Chantal Shaib, and Frank Rudzicz. Explainable clinical decision support from text. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1478–1489, 2020.

[6] Stephen Casper, Xander Davies, Claudia Shi, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. **arXiv preprint arXiv:2307.15217**, 2023.

[7] Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards understanding sycophancy in language models. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2024.

[8] Avneet Kaur. Echoes of agreement: Argument driven sycophancy in large language models. In **Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2025**, pp. 22803–22812, 2025.

[9] Chien Hung Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Self-augmented preference alignment for sycophancy reduction in LLMs. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 12379–12391, 2025.

[10] Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In **Findings of the Association for Computational Linguistics (ACL) 2024**, pp. 12717–12733, 2024.

[11] Jillian Fisher, Shangbin Feng, Robert Aron, et al. Biased LLMs can influence political decision-making. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 6559–6607, 2025.

[12] Wei Chen, Zhen Huang, Liang Xie, et al. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. In **Proceedings of the 41st International Conference on Machine Learning (ICML)**, pp. 4028–4045, 2024.

[13] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4895–4901, 2023.

[14] Kevin Wang, Arthur Variengien, Adam Conmy, et al. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. **arXiv preprint arXiv:2211.00593**, 2022.

[15] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[16] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. **arXiv preprint arXiv:2310.06825**, 2023.

[17] Llama Team, Meta. The Llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[18] An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. **arXiv preprint arXiv:2412.15115**, 2024.

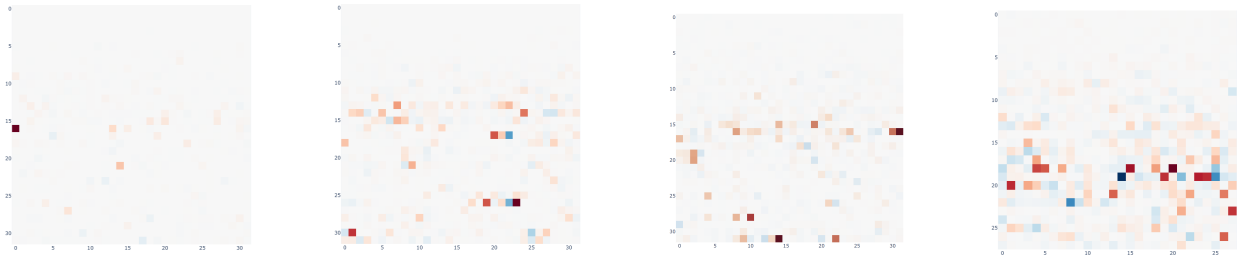
[19] Mor Geva, Daniel Khashabi, Elad Segal, et al. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. **Transactions of the Association for Computational Linguistics (TACL)**, Vol. 9, pp. 346–361, 2021.

[20] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[21] Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. **arXiv preprint arXiv:2107.03374**, 2021.

A　迎合寄与の可視化

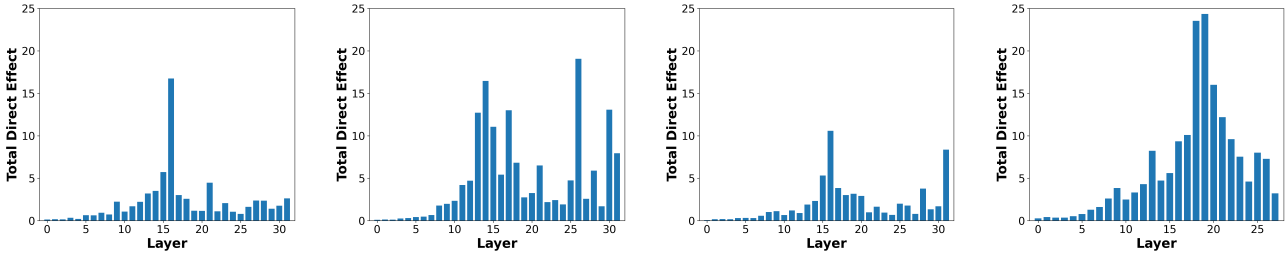
3.1 節で述べた各モデルにおける Attention Head と Layer 単位での迎合の寄与の可視化を示す．



(a) Llama-2　(b) Llama-3　(c) Mistral-v0.3　(d) Qwen-2.5

図 3: 各モデルにおける Attention Head 単位での迎合寄与度（縦軸：Layer，横軸：Head）

Llama-2 では少数 Head に寄与が集中しているのに対し，GQA を採用した Llama-3，Mistral，Qwen-2.5 では，寄与が広範囲に分散していることが確認された．



(a) Llama-2　(b) Llama-3　(c) Mistral-v0.3　(d) Qwen-2.5

図 4: 各モデルにおける Layer 単位での迎合寄与値の総和（縦軸：寄与値の総和，横軸：Layer）

いずれのモデルも，中間層から深い層にかけて寄与が大きい傾向が確認された．特に GQA モデルでは，特定の層においてグループ単位での強い寄与が見られ，提案手法のグループ単位の介入の妥当性を裏付ける．

B　パラメータサイズの違いによる結果

表 3: Llama-3.2-3B における迎合抑制効果と一般性能の比較

	Sycophancy		General Ability					
	Truthfulness		StrategyQA		GSM8K		HumanEval	
Models (Ratio: 4%)	**Acc.**	Δ	**Acc.**	Δ	**Acc.**	Δ	**Acc.**	Δ
Llama-3.2-3B	45.69	-	59.04	-	**68.92**	-	**54.27**	-
+ SPT	68.61	+22.92	**63.20**	+4.16	61.49	-7.43	51.22	-3.05
+ **提案手法**	**69.11**	+23.42	52.31	-6.73	56.02	-12.90	47.90	-6.37

3B の小規模モデルにおいても，提案手法は SPT を上回る抑制効果が得られた．一方，小規模モデルは一般性能の低下が比較的大きく，これは Head の数が少なく，迎合抑制機能と一般性能が同一 Head を共有している可能性がある．