Attention Sink および Massive Activation の発生機序の分解

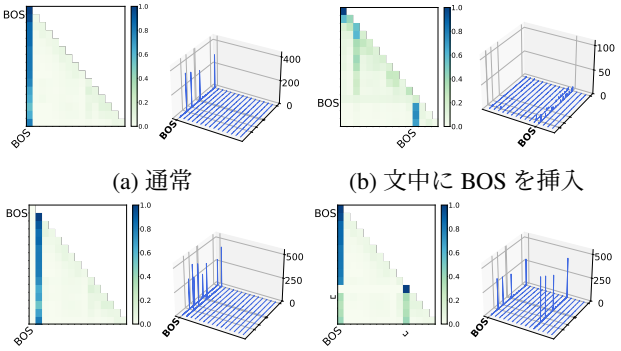
木谷頼斗 [1]　大橋諭貴 [1]　佐藤宏亮 [1]　鴨田豪 [2,3]　高橋良允 [1,4]
山本悠士 [2,3]　塩野大輝 [1]　坂口慶祐 [1,4]　小林悟郎 [1]
[1] 東北大学　[2] 総合研究大学院大学　[3] 国立国語研究所　[4] 理化学研究所
raito.kiya@dc.tohoku.ac.jp

概要

大規模言語モデル（LLM）では，特定トークンに注意が集中する Attention Sink や，それに伴う Massive Activation が知られている．これらはモデルの安定動作に寄与する一方，量子化精度の低下を招く重要な現象である．本研究では，(1) BOS トークン固有の埋め込み表現，(2) 位置情報の影響，(3) 注意を自身のみに集中する挙動，の三要素に着目し，現象の発生要因に迫る．これは LLM の解釈性向上やより堅牢な設計への指針となる．

1　はじめに

大規模言語モデル（LLM）の主なアーキテクチャには，トークン間で注意を向け合うことで文脈情報への参照を可能にする注意機構を核とした Transformer [1] が採用されている [2–4]．近年では LLM 内部の注意機構の多くで入力列の先頭（BOS）やピリオドなどの特定のトークンに対し，意味的な関連性の有無によらず注意が過度に集中する「Attention Sink」という現象が知られている [5–7]．Attention Sink は直感と反する特徴的な内部挙動として興味深いだけでなく，LLM の動作安定性および軽量化の観点でも注目すべき現象である．先行研究では，この現象が LLM の安定動作に重要な役割を果たすことや [5]，Attention Sink が発生するトークンで巨大な活性化値[1)]（Massive Activation）が生じることが報告されている [8]．このような巨大な値は LLM を量子化する際に精度低下の原因となる [9, 10]．最近ではこれらの抑制を図って Qwen チームによる Gated Attention [4] や，OpenAI による学習可能なバイアス付き Softmax [11] など，Attention Sink および Massive Activation を考慮または対策したアーキテクチャの工夫が導入され始めているが，完全な抑制には至っていない．これらの現象の観察や分析も行われつつあるが，その発生機序については未だ解明されていない点が多い．



(a) 通常　(b) 文中に BOS を挿入

(c) 先頭の注意をマスク　(d) 自身に注意を向ける介入

図 1: 本研究では，Attention Sink や Massive Activation (a) に対して 3 種類の介入 (b, c, d) を行い，これらの移動をもって現象の発生要因を明らかにする．[2)]

本研究では，入力や注意機構の計算過程に介入して挙動や出力を観察する因果的介入手法を用いて，Attention Sink および Massive Activation の発生条件を体系的に整理し，その発生原理の解明に向けた基盤的知見を提供する（図 1）．具体的には，(1) BOS トークン固有の埋め込み表現，(2) 位置情報の影響，(3) 注意を自身のみに集中する挙動，の三要素に切り分けて分析を行った．結果として，このうち (1) と (3) の二つが主な発生要因であることを特定し，その原理についても一部解明した．さらに，初期層の注意機構への介入によって，後続層における Attention Sink および Massive Activation の発生位置を制御可能であることを発見した．これらの知見は，LLM の解釈性向上や，量子化への適応性向上といった発展に寄与するものである．

1)　本論文では，「活性化値」は隠れ状態におけるものを指す．

2)　この図では Llama-3.2-3B モデルを使用した．また，特段断りがない限り，本稿で掲載する注意マップは対象とする層内の全ヘッド平均である．

2 背景

2.1 注意機構

注意機構の役割は，文脈情報を現在のトークンの表現に動的に混ぜ込むことである．入力列 $\{\boldsymbol{x}_i\}_{i=1}^{T}$, $\boldsymbol{x}_i \in \mathbb{R}^d$ に対して，位置 i の出力 $\boldsymbol{y}_i$ は以下のように計算される．ただし，T は入力列のトークン数，d はモデルの次元，d' は注意機構（ヘッド）の次元，$\boldsymbol{W}_{\{Q,K,V,O\}} \in \mathbb{R}^{d\times d'}, \boldsymbol{\gamma} \in \mathbb{R}^d, \varepsilon \in \mathbb{R}$ はモデルのパラメータ，$\boldsymbol{R}_i \in \mathbb{R}^{d'\times d'}$ は RoPE の回転行列であり，$\odot$ は要素積を表す[3]．

$$\boldsymbol{y}_i = \left(\Sigma_{j=1}^{i}\alpha_{i,j}\boldsymbol{v}_j\right)\boldsymbol{W}_O^\top \quad \in \mathbb{R}^d \tag{1}$$

$$\alpha_{i,j} = \mathrm{Softmax}_j\left(\boldsymbol{q}_i^\top \boldsymbol{R}_i \boldsymbol{R}_j^\top \boldsymbol{k}_j/\sqrt{d'}\right) \quad \in \mathbb{R} \tag{2}$$

$$\{\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i\} = \tilde{\boldsymbol{x}}_i\{\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V\} \quad \in \mathbb{R}^{d'} \tag{3}$$

$$\tilde{\boldsymbol{x}}_i = \mathrm{RMSNorm}(\boldsymbol{x}_i) = \frac{\boldsymbol{x}_i \odot \boldsymbol{\gamma}}{\sqrt{\|\boldsymbol{x}_i\|_2^2/d + \varepsilon}} \quad \in \mathbb{R}^d \tag{4}$$

2.2 関連研究

Attention Sink および Massive Activation は工学的にも重要な現象である．特定のトークンが注意の集約点として機能することを前提にすることで，KV キャッシュの再利用や圧縮が可能となり，メモリ使用量と推論安定性を同時に改善できる [5, 12, 13]．また，過剰に大きな活性化値は低精度量子化における外れ値問題 [8, 9] の主因であり，これの理解および制御はモデルの挙動を損なうことなく安全に軽量化するための理論的基盤となる．

Attention Sink および Massive Activation について分析する既存研究では，初期層での形成プロセスと中盤層での圧縮プロセスが注目されている．これらの現象は初期層において発生し始め，中盤層にわたって続き，最終層付近では弱まることが知られている [6, 14]．Queipo-de-Llano ら [15] は「Mix-Compress-Refine」という理論を提唱し，中盤層の Massive Activation が文脈情報の過剰混合（Over-mixing）を防ぎ，文脈を圧縮保持していると主張している．また，Gu ら [6] は BOS が余分な注意を引き受けるためのバイアスとして機能することを経験的に示し，Barbero ら [16] はこれが情報の過剰混合を防ぐ理論的メカニズムであると述べている．

しかし，これら既存研究の多くは現象の観察や機能的意義の考察に留まっており，発生要因の特定には未だ至っていない．また，Attention Sink と Massive Activation の共起関係は報告されているものの，両者の因果関係については，実証的な検証が乏しい．本研究では，入力や注意機構への介入実験を通じてこれらの要因を分離し，現象の発生原理および両者の因果的関係の解明に迫る．

3) Multi-head 注意機構の場合も同様である

3 発生要因の分離と特定

Attention Sink および Massive Activation の発生要因として考えられる条件を整理し，それぞれが実際に発生のトリガーとなっているのかを検証することで発生機序の解明を目指す．これらの現象は入力の先頭に挿入される BOS トークンで発生することが知られている．これを手がかりに，本研究では BOS トークンの特徴を次の 3 つに切り分け，各要因候補について調査する：1) BOS トークンに対応する埋め込みを使用する（§ 3.1），2) RoPE に渡される絶対位置が常に 0 である（§ 3.2），3) 常に自分自身に 100% の注意を向ける（§ 3.3）．実験は Llama-3.2-3B [3][4] を対象に行い，入力には WikiText [17][5] を用いた．また，Llama-2-7B [18][6] についても実験を行った結果，Attention Sink および Massive Activation の発生という点では共通するが，発生する層については差異が見られた．詳細は付録 A に掲載する．

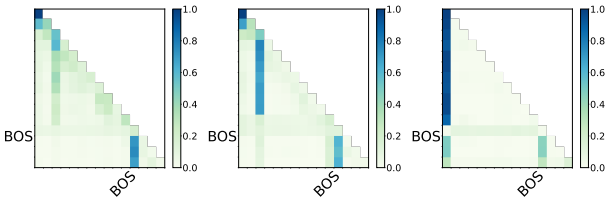
3.1 BOS トークン埋め込みによる影響

Attention Sink が BOS トークン固有の性質に起因するのか，あるいは入力列における位置に起因するのかを調査するため，入力トークン列に対して BOS トークンと文中のトークンを入れ替える介入を行う．

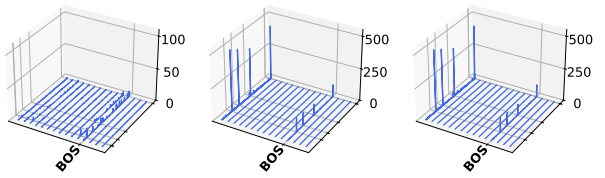
注意マップの観察　入れ替え前の入力では，図 1a のように BOS トークンで Attention Sink および Massive Activation が観察できる．

図 2 に介入後の結果を示す．図 2a に示すように，Llama-3.2-3B では第 1 層で先頭でなくとも BOS トークンに対して Attention Sink および Massive Activation が発生している．一方，BOS ではない差し替えられた先頭トークンに対しては Attention Sink，Massive Activation いずれも発生していない．これは，第 1 層において Attention Sink および Massive Activation は位置情報よりも BOS トークン固有の性質（埋め込

4) https://huggingface.co/meta-llama/Llama-3.2-3B
5) https://huggingface.co/datasets/Salesforce/wikitext の wikitext-2-raw-v1 サブセットを用いた．
6) https://huggingface.co/meta-llama/Llama-2-7b-hf



(a) 第 1 層　(b) 第 2 層　(c) 第 3 層



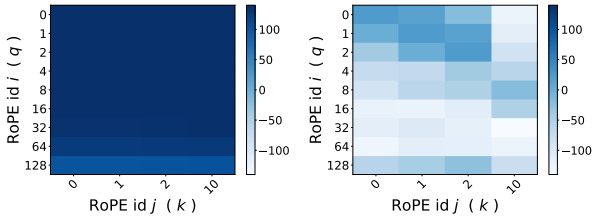
(d) 第 1 層　(e) 第 2 層　(f) 第 3 層

図 2: 入力段階で BOS トークンと任意のトークンを入れ替えた場合の注意スコア（上段）と活性化値（下段）．

み表現など）に強く依存することを示唆する．一方，第 3 層では，差し替えられた先頭トークンにも Attention Sink や Massive Activation が発生している（図 2c, 2f）．以上より，本現象の発生要因は複合的であり，BOS トークンに起因するものは初期層から，先頭位置に起因するものは層を重ねることで顕在化することが示唆された．なお，Llama-2 系列では第 1 層での Attention Sink は見られないが，数層後には同様の現象が発生する（付録 A 参照）．これは，埋め込み由来の要因と層を重ねることによる要因が異なるタイミングで顕在化することを示唆しており，本研究の「要因の分解」というアプローチの妥当性を裏付けている．しかし，モデル非依存的な分析は，今後の課題である．

次に，Attention Sink と Massive Activation の関係性について考察する．図 2e, 図 2c の関係に着目すると，先頭トークンで Massive Activation が発生した直後の層において，Attention Sink が形成されていることが確認できる．この順序関係は，Massive Activation が Attention Sink を誘発している可能性を示唆するものである．

埋め込みレベルの分析　第 1 層から Attention Sink が発生する原理を解明するため，我々は BOS トークンの埋め込みと他のトークンの埋め込みの違いを分析した．BOS トークンの埋め込みと，その他の埋め込みベクトルの平均をそれぞれ $\boldsymbol{e}_{\text{bos}}, \boldsymbol{e}_{\text{avg}} \in \mathbb{R}^d$ とする．ここでは，第 1 層におけるそれぞれの Key ベクトル $\boldsymbol{k}_{\{\text{bos,avg}\}} := \tilde{\boldsymbol{e}}_{\{\text{bos,avg}\}} \boldsymbol{W}_K \in \mathbb{R}^d$ に着目



(a) 内積: $\boldsymbol{q}_{\text{avg}}^\top \boldsymbol{R}_i \boldsymbol{R}_j^\top \boldsymbol{k}_{\text{bos}}$　(b) 内積: $\boldsymbol{q}_{\text{avg}}^\top \boldsymbol{R}_i \boldsymbol{R}_j^\top \boldsymbol{k}_{\text{avg}}$

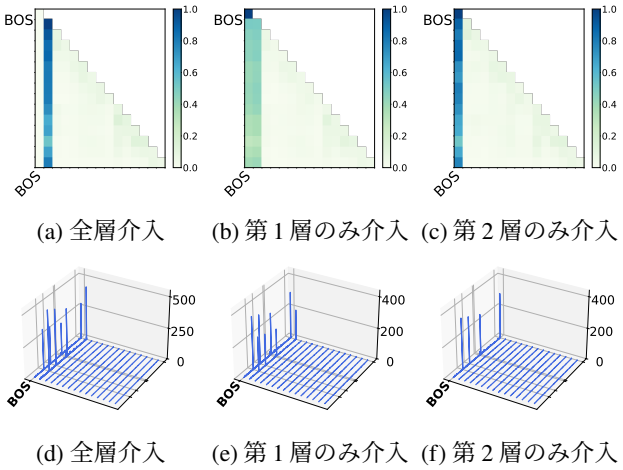
図 3: RoPE による位置埋め込みを含めた第 1 層での $\boldsymbol{q}$–$\boldsymbol{k}$ 関係．

する[7]．まずノルムを比較したところ，$\|\boldsymbol{k}_{\text{bos}}\| = 26.155, \|\boldsymbol{k}_{\text{avg}}\| = 100.611$ であり，$\boldsymbol{k}_{\text{bos}}$ は他のトークンに比べノルムが小さかった．続いて，第 1 層における平均クエリベクトル $\boldsymbol{q}_{\text{avg}} = \tilde{\boldsymbol{e}}_{\{\text{bos,avg}\}} \boldsymbol{W}_Q \in \mathbb{R}^d$ と，$\boldsymbol{k}_{\{\text{bos,avg}\}}$ の内積およびコサイン類似度を比較したところ，$\cos(\boldsymbol{q}_{\text{avg}}, \boldsymbol{k}_{\text{bos}}) = 0.521, \cos(\boldsymbol{q}_{\text{avg}}, \boldsymbol{k}_{\text{avg}}) = 0.010$ であり，$\boldsymbol{k}_{\text{bos}}$ は $\boldsymbol{q}_{\text{avg}}$ と極めて高いコサイン類似度を持つことが確認できた．つまり BOS トークンの埋め込みは，Key ベクトルに変換された後において大きさ（内積への寄与）ではなく平均的な Query ベクトルとの向きの一致によって注意を集めていることが示唆される．Gu ら [6] は，入力列の先頭位置にあるトークンの Key ベクトルについて同様の性質を報告している．これに対し本研究は，この特性が文脈や位置情報が付与される以前の，埋め込み表現から射影された段階で既に獲得されていることを明らかにした．これは，BOS トークン固有の静的な性質こそが，第 1 層目から Attention Sink を引き起こす主要因であることを裏付けている．

3.2　RoPE の位置インデックスの影響

RoPE の位置インデックスが Attention Sink の発生要因として機能しているのかを検証する．前節と同様に埋め込みベクトルを用い，Query ベクトルと Key ベクトルそれぞれの RoPE 位置インデックス i, j を変化させた際の内積 $\boldsymbol{q}_{\text{avg}}^\top \boldsymbol{R}_i \boldsymbol{R}_j^\top \boldsymbol{k}_{\{\text{bos,avg}\}}$ を図 3 に示す．図 3a より，BOS に対する注意計算においては位置インデックス i と j がいかなる値であっても一貫して高い内積を維持していた．対照的に，図 3b の平均的なトークンに対する注意計算においては両方の位置インデックス i, j の変化に伴い内積の変動が見られた．ただし，一貫して強い内積をもたらすような Key 位置インデックス j は観測されなかった．これらの結果から特に初期層においては，BOS

7)　$\tilde{e}$ は RMSNorm を適応したベクトルである（式 4）．



(a) 全層介入　(b) 第 1 層のみ介入　(c) 第 2 層のみ介入

(d) 全層介入　(e) 第 1 層のみ介入　(f) 第 2 層のみ介入

図 4: Llama-3.2-3B において BOS トークンの注意ロジットを $-\infty$ でマスクした際の第 14 層における注意スコア（上段）と活性化値（下段）．

トークンに注意が集まる要因として RoPE の位置インデックスはあまり重要でなく，先頭の位置インデックス（$j = 0$）が適用されることは Attention Sink の発生要因でない可能性が高い．

3.3　自身への注意集中による影響

Attention Sink および Massive Activation が BOS 以外のトークンであっても先頭位置に発生した（図 2c, 2f）ことから，ここまでで検証した「BOS 固有の埋め込み」と「RoPE 位置インデックス」以外で発生要因が存在するはずである．本節では先頭位置トークンが持つ最後の特徴である「常に自分自身に 100% の注意を向ける」という性質がもう 1 つの発生要因であることを示す．

まず，先頭トークンへの注意をマスク（Softmax 適用前の注意スコアを $-\infty$ で置換）する介入を行うことで，非 BOS かつ 2 番目の位置にあるトークンに「自身に全ての注意を集中させる」性質を付与して Attention Sink と Massive Activation の挙動を観察した[8]．図 4 に示すように，全層あるいは第 1 層にこの介入をすることで，Attention Sink が 2 番目のトークンに移動し，Massive Activation が先頭と 2 番目のトークンの両方で発生することがわかる．一方で，第 2 層のみへの介入ではこれらの移動は起きなかった（図 4c）．ここで Attention Sink が第 3 層以降で発生する Llama-2-7B では，第 2 層のみへの介入で

8）本研究では Softmax 適用後の注意スコアへの介入実験も行ったが，モデル出力の崩壊が確認された．そのため，分量の都合上，本稿では注意ロジットへの介入に関する実験結果のみを掲載する．

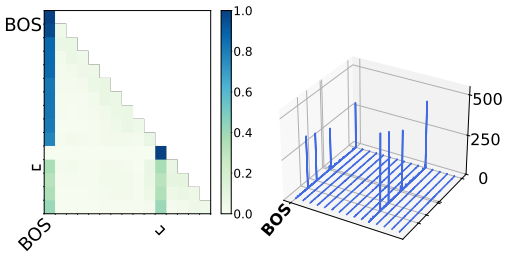


図 5: 任意位置のクエリに対して注意スコアを自身のキーに集中させた時の第 14 層における注意スコア（左）と活性化値（右）．

も同様の移動が確認できた（付録 A）．これらの結果から，特に Attention Sink が初めて形成されるまでの初期層において，自身に注意を集中させるという性質が Attention Sink および Massive Activation の発生に寄与していることが示唆される．

続いて設定をさらに一般化し，入力列中の任意位置のトークンに対して，注意を全て自身に向けさせる介入（Softmax 適用前の自身への注意スコアを ∞ で置換）を行った．図 5 に示すように，この介入によって **Attention Sink および Massive Activations の発生を任意位置に移動させることができた**．この結果は，自身に注意を集中させるという特徴が Attention Sink および Massive Activation の発生要因の 1 つであることをさらに強く支持する．また，このような発生位置の制御可能性に関する知見は，今後のモデルの解釈性や制御において重要な意味を持つ可能性がある．本稿では要因の特定までに留まったが，「自身への注意集中」がどのようにしてそれら現象の発生に結びついているかの原理解明に取り組むことは今後の展望の一つである．

4　おわりに

本研究では因果的介入に基づく分析により，LLM における Attention Sink および Massive Activation の発生要因を分析し，BOS トークン固有の埋め込み表現が要因の一つであることを実証した．さらに，クエリが自身へ全注意を払うことで Attention Sink や Massive Activation が発生することを確認し，先頭位置で現象が起こる要因の解明に貢献した．今後は，本稿で特定した要因が Attention Sink や Massive Activation を発生させる具体的な内部機序の解明に取り組む．また，モデル系列間で発生層が異なる要因の同定を通じて，よりモデル非依存な理解を目指す．

謝辞

本研究は JSPS 科研費 JP25K03175，JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援） JPMJBS2412，JPMJBS2421，JST 次世代研究者挑戦的研究プログラム JPMJSP2104 の助成を受けたものです.

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. **Advances in Neural Information Processing Systems**, Vol. 30, , 2017.

[2] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and others. Gemma 3 Technical Report. **arXiv [cs.CL]**, 25 March 2025.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and others. The Llama 3 herd of models. **arXiv [cs.AI]**, 31 July 2024.

[4] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, and others. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems**, 29 October 2025.

[5] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In **Proceedings of the Twelfth International Conference on Learning Representations (ICLR)**, 2024.

[6] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When Attention Sink Emerges in Language Models: An Empirical View. In **The Thirteenth International Conference on Learning Representations**, 4 October 2024.

[7] Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention Sinks and Outlier Features: A 'Catch, Tag, and Release' Mechanism for Embeddings. **arXiv preprint**, Vol. arXiv:2502.00919, , 2025.

[8] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In **First Conference on Language Modeling**, 2024.

[9] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.Int8(): 8-Bit Matrix Multiplication for Transformers at Scale. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2022.

[10] Ruikang Liu, Haoli Bai, Haokun Lin, Yuening Li, Han Gao, Zhengzhuo Xu, Lu Hou, Jun Yao, and Chun Yuan. IntactKV: Improving Large Language Model Quantization by Keeping Pivot Tokens Intact. In **Findings of the Association for Computational Linguistics: ACL 2024 (Findings of ACL)**, pp. 7716–7741, 2024.

[11] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, and others. gpt-oss-120b & gpt-oss-20b model card. **arXiv preprint**, Vol. arXiv:2508.10925, , 2025.

[12] Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. In **Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)**, 2025.

[13] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, and others. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2023.

[14] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive Activations in Large Language Models. 27 February 2024.

[15] Enrique Queipo-de-Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention Sinks and Compression Valleys in LLMs Are Two Sides of the Same Coin. **arXiv preprint**, Vol. arXiv:2510.06477, , 2025.

[16] Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why Do LLMs Attend to the First Token? 2025.

[17] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In **International Conference on Learning Representations (ICLR)**, 2017.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and others. Llama 2: Open foundation and fine-tuned chat models, 2023.

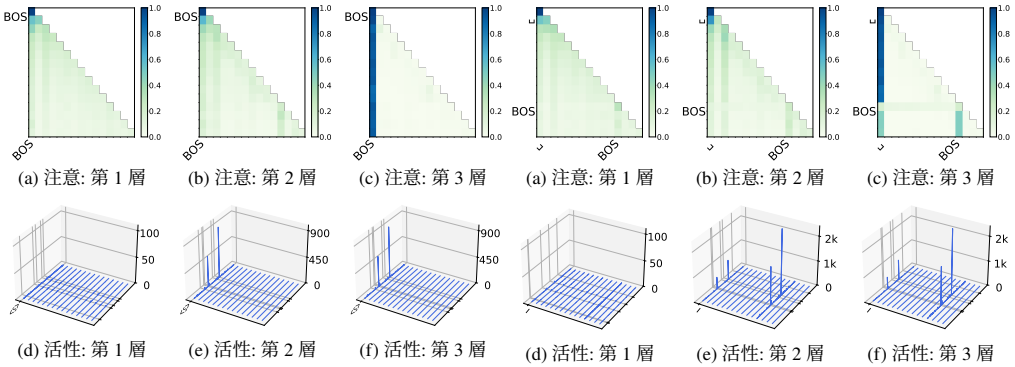
A　Llama-2-7B-hf における詳細分析

本文 § 3 では主に Llama-3.2-3B を用いた分析結果を報告したが，Attention Sink および Massive Activation の発生機序の普遍性を検証するため，Llama-2-7B-hf においても同様の調査を行った．

A.1　通常時の挙動と BOS トークン入れ替えの影響

図 6 に通常時の注意スコアおよび活性化値を示す．Llama-3.2-3B とは異なり，Llama-2-7B-hf の第 1 層では BOS トークンにおいて Attention Sink と Massive Activation は発生しない (図 6a)．しかし，層を重ねるにつれ，BOS トークンへの Attention Sink や Massive Activation が徐々に顕著となる．これは，モデルによって両現象が発生する層の深さに差異があることを意味する (図 6b, 6c, 6e and 6f)．

BOS トークンを文中に移動させた結果を図 7 に示す．Llama-3.2-3B と異なり，第 1 層では Attention Sink および Massive Activation は発生しない．しかし，層を重ねると通常のトークンに入れ替えたにも関わらず先頭トークンにおいて Attention Sink と Massive Activation がいずれも発生する．この結果は，Llama-2-7B-hf における Attention Sink および Massive Activation の発生要因が，静的な埋め込み表現ではなく，層を重ねる計算過程にあることを示唆している．



(a) 注意: 第 1 層　(b) 注意: 第 2 層　(c) 注意: 第 3 層
(d) 活性: 第 1 層　(e) 活性: 第 2 層　(f) 活性: 第 3 層

図 6: 通常時の挙動

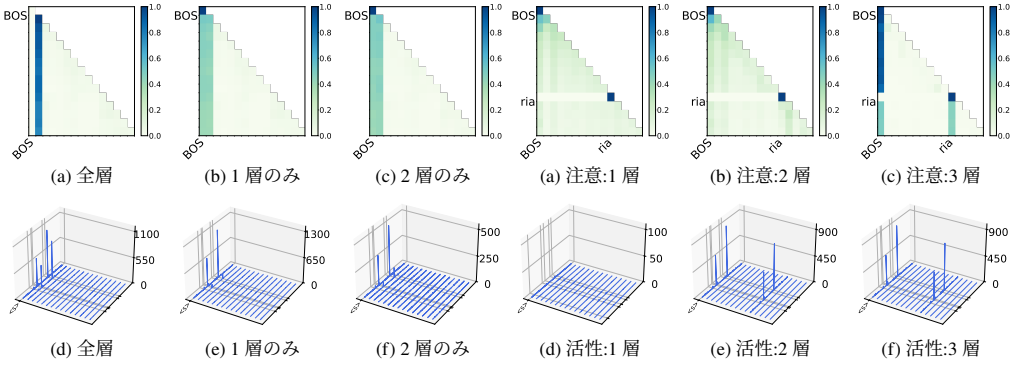
(a) 注意: 第 1 層　(b) 注意: 第 2 層　(c) 注意: 第 3 層
(d) 活性: 第 1 層　(e) 活性: 第 2 層　(f) 活性: 第 3 層

図 7: BOS 入れ替え介入

A.2　介入による Attention Sink の位置移動と発生要因の検証

Attention Sink の位置移動と発生要因の検証を行った結果を示す．図 8 は注意ロジットへのマスク介入の結果である．通常時の Massive Activation 形成が顕著になる第 2 層以降では介入が有効に機能し，第 2 層 (図 8c) での介入により Attention Sink が 2 番目のトークンへ移動していることが確認できる．

また，図 9 は自身へ全注意を払う介入を行ったものである．第 2 層や第 3 層において，介入を行ったトークンで Attention Sink および Massive Activation が誘発されている．この結果は，Attention Sink および Massive Activation の発生において，BOS トークンや先頭位置という属性は必須ではなく，自身への注意集中という構造的メカニズムこそが発生要因である可能性を示唆している．



(a) 全層　(b) 1 層のみ　(c) 2 層のみ
(d) 全層　(e) 1 層のみ　(f) 2 層のみ

図 8: 注意マスク介入（第 14 層の出力）

(a) 注意:1 層　(b) 注意:2 層　(c) 注意:3 層
(d) 活性:1 層　(e) 活性:2 層　(f) 活性:3 層

図 9: 注意スコアを自身のキーに集中させた時