注意機構における Attention Sink のバイアス項的解釈

大橋諭貴 [1]　木谷頼斗 [1]　佐藤宏亮 [1]　高橋良允 [1,2]　鴨田豪 [3,4]
山本悠士 [3,4]　塩野大輝 [1]　坂口慶祐 [1,2]　小林悟郎 [1]
[1] 東北大学　[2] 理化学研究所　[3] 総合研究大学院大学　[4] 国立国語研究所
ohashi.satoki.t8@dc.tohoku.ac.jp

概要

大規模言語モデルでは，文頭トークンに注意が集中する Attention Sink 現象が知られている．この現象の機能的役割や，モデルアーキテクチャへの依存性については未だ十分に解明されていない．本研究では，注意機構において文頭トークンから各トークン表現へ加算される更新ベクトルの幾何学的な特性を分析し，それが文脈に依存しない定数ベクトルのようであることから，Attention Sink 現象がバイアスのように機能している可能性を示す．さらに，推論時に該当の更新ベクトルを事前計算した定数ベクトル（固定のバイアス項）で置換する介入実験を行い，モデルの出力品質が維持されることを確かめる．

1　はじめに

Transformer [1] ベースの大規模言語モデル（LLM）において，内部の注意機構が文頭トークンなど特定のトークンに対して意味的重要度とは無関係に注意が偏る Attention Sink 現象が報告されている [2]．この現象を考慮した工夫を導入することで性能劣化を抑えつつ注意機構の計算量を削減できるなど工学的に注目が集まっている [2–4]．その一方，この現象の発生メカニズムや機能については未だ完全には解明されていない．

いくつかの既存研究は，この現象を Softmax 関数の正規化制約やモデルの構造的な制約に起因する学習の副産物として解釈している．例えば，Attention Sink を不要な注意重みを退避させる「ゴミ捨て場」[2] や実質的な情報の更新を行わない「操作無し（no-op）」[5,6] として扱うものから，Vision Transformer [7] において情報を一時的に保存するための「レジスタ」[8] として扱うものまで多岐にわたる．また，Attention Sink と密接に関連するものとして，中間層において注意が偏る特定のトークンで一部のニューロンが極端に大きな活性値を示す

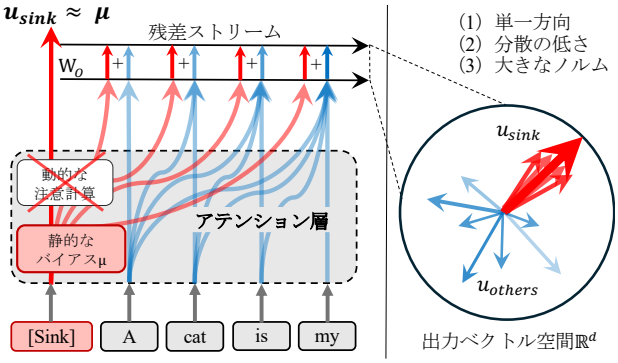

図 1: Sink トークン由来の更新ベクトル分析概要．(§ 3) 文脈に依存せず常に一定の方向と巨大なノルムを持ち，実質的なバイアス項として機能しているならば，（§ 4）Sink トークンによる動的な注意機構の出力を，静的な平均ベクトル $\boldsymbol{\mu}$ で置換することが可能である

Massive Activations [9] が報告されている．Sun ら [9] は，これらの異常活性値がモデル内部でバイアス項として機能する可能性を示唆している．この知見を踏まえると，Attention Sink についても同様にバイアス項としての機能を持つ可能性が考えられる．しかし，この現象が、バイアス項を持たない近年のモデルアーキテクチャに特有の機能補完なのか，それともアーキテクチャに依存せず発生し，既存のバイアス項と協調する普遍的な機能なのかについては，統一的な見解が得られていない．

本研究では，Attention Sink がモデル構造の欠落を補完する暗黙的なバイアス項のように機能しているという仮説を検証する．具体的には，注意機構において Sink トークンに由来する更新ベクトルの影響力，情報の単純性，および時間的な安定性を分析することで，文脈に依存しない定数ベクトルとしての性質を幾何学的に検証する．さらに，その動的計算を定数で置換する介入手法 “Static Sink

Patching (SSP)” を通して実験的にも検証する．実験の結果，明示的なバイアス項を持たないモデルにおいて，Sink トークンからの表現更新を定数に置換しても性能が維持されることを明らかにした．これは Attention Sink が機能的役割のない単なる副産物ではなく，実質的なバイアス項として機能していることを示唆しており，LLM の解釈性向上および将来的な推論効率化に寄与する知見である．

2 準備

2.1 注意機構と Attention Sink

ある入力トークン列 s を言語モデルに入力したときの，第 ℓ 層，位置 i のトークンでの注意機構の出力 $\boldsymbol{y}_i^{(\ell)} \in \mathbb{R}^{d_{\text{model}}}$ は，以下のように「先頭 (BOS) トークンからの更新ベクトル $\boldsymbol{u}_{\text{sink},i}^{(\ell)}(s)$」と「それ以外のトークンからの更新ベクトル $\boldsymbol{u}_{\text{others},i}^{(\ell)}(s)$」と「バイアス項 $\boldsymbol{b}^{(\ell)}$」に分解できる．

$$\boldsymbol{y}_i^{(\ell)}(s) = \underbrace{\boldsymbol{u}_{0,i}^{(\ell)}(s)}_{\boldsymbol{u}_{\text{sink},i}^{(\ell)}(s)} + \underbrace{\sum_{j=1}^{i} \boldsymbol{u}_{j,i}^{(\ell)}(s)}_{\boldsymbol{u}_{\text{others},i}^{(\ell)}(s)} + \boldsymbol{b}^{(\ell)} \tag{1}$$

ここで，d_{model} はモデルの次元数，$\boldsymbol{u}_{j,i}^{(\ell)} \in \mathbb{R}^{d_{\text{model}}}$ は，注意機構で位置 j から i に加算されるベクトルであり，$\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{d_{\text{model}}}$ は注意機構に存在するバイアス項である．本稿では Attention Sink は BOS トークンで発生するものに限定して扱い，Sink トークンと呼称する．また以降では各項を青字で示した表記で参照する． また，N 件の系列からなるデータセット $\mathcal{D} = \{s_k\}_{k=1}^{N}$ 上で計算された $\boldsymbol{u}_{\text{sink},i}^{(\ell)}(s)$ の平均 $\boldsymbol{\mu}_{\text{sink}}^{(\ell)} \in \mathbb{R}^{d_{\text{model}}}$ を以下で定義する．

$$\boldsymbol{\mu}_{\text{sink}}^{(\ell)} = \frac{1}{\sum_{k=1}^{N} |s_k|} \sum_{k=1}^{N} \sum_{i=0}^{|s_k|-1} \boldsymbol{u}_{\text{sink},i}^{(\ell)}(s_k) \tag{2}$$

2.2 モデルとデータセット

本研究ではパラメータ数が 7B〜8B 程度の同規模のモデルから，注意機構内にバイアス項を持つモデルとして Qwen2.5-7B [10] と Pythia-6.9b [11]，バイアス項を持たない ($\boldsymbol{b}^{(\ell)} = \boldsymbol{0}$) モデルとして Llama-2-7b-hf [12]，Mistral-7B-v0.1 [13]，Qwen3-8B-Base [14] を調査対象とする．データセットには OpenWebText [15] から抽出した 100 件を使用し，入力長は 512 トークンに統一した．

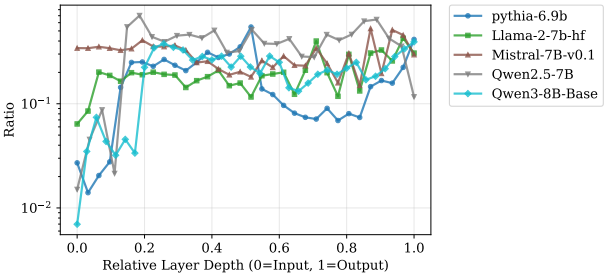

図 2: 各層における Sink トークンの影響比率 $\text{Ratio}_{(\ell)}$．Ratio が 1 に近づくほど Sink トークン単体の寄与が他のトークンと同程度となり，小さくなるほど相対的な影響は弱まる．

3 注意機構の更新ベクトルの分析

Sink トークンが持つ更新ベクトル $\boldsymbol{u}_{\text{sink},i}^{(\ell)}$ の幾何学的特性を分析し，それが「動的な文脈処理」ではなく「静的なバイアス付与」としての性質を強く持っていることを示す．

3.1 Sink トークンの影響力

Sink トークンが注意機構の出力に与える影響の大きさを，他のトークンと比較して以下のように定量化し，その結果を図 2 に示す．

$$\text{Ratio}^{(\ell)} = \frac{\mathbb{E}_i[\|\boldsymbol{u}_{\text{sink},i}^{(\ell)}\|_2]}{\mathbb{E}_i[\|\boldsymbol{u}_{\text{others},i}^{(\ell)}\|_2]} \tag{3}$$

いずれのモデルにおいても，ほとんどの層で Ratio = 0.2 ~ 0.5 であり，512 トークンの入力に対して Sink トークン単体から加算されるベクトルがかなり強く寄与していた．この結果はモデルが明示的なバイアス項を持つかどうかに関わらず，Attention Sink 現象を介して擬似的なバイアス成分を各トークン表現に注入していることを示唆している．

3.2 Rank-1 Structure: 情報の単純性

続いて，Sink トークン由来の更新ベクトル $\boldsymbol{u}_{\text{sink},i}^{(\ell)}$ が，トークン位置 i に依らずほぼ一定の方向を向いていることを定量的に示す．まず入力 s の各トークン位置において，第 ℓ 層で Sink トークンに由来する更新ベクトルを結合した行列 $\boldsymbol{C}_{\text{sink}}^{(\ell)} \in \mathbb{R}^{|s| \times d_{\text{model}}}$ を以下で定義する．

$$\boldsymbol{C}_{\text{sink}}^{(\ell)}(s) = \left[\boldsymbol{u}_{\text{sink},1}^{(\ell)}(s) \quad \cdots \quad \boldsymbol{u}_{\text{sink},|s|-1}^{(\ell)}(s)\right]^{\top} \tag{4}$$

次に，データセット内の 100 件の入力に対して計算されたこの行列を重ねて特異値分解（SVD）を適用

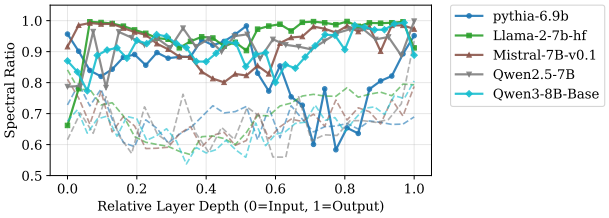

図 3: Sink トークン由来の更新ベクトル $\boldsymbol{u}^{(\ell)}_{\text{sink},i}(s)$（実線）と Sink 以外の全ての文脈トークン由来の更新ベクトルの平均 $\boldsymbol{u}^{(\ell)}_{\text{ctx},i}(s)$（点線）における第一特異成分の支配率．値が 1 に近いほど，更新ベクトルが静的なバイアスとして機能していることを示唆する．

し，最大の特異値 $\sigma_1^{(\ell)}$ を持つ方向が全体に占める寄与率（SpectralRatio）を次のように計算する．

$$\text{SpectralRatio}^{(\ell)} = \left(\sigma_1^{(\ell)}\right)^2 \Big/ \sum_k \left(\sigma_k^{(\ell)}\right)^2 \tag{5}$$

これを，Sink トークン以外の全てのトークン由来の更新ベクトル (以降 $\boldsymbol{u}^{(\ell)}_{\text{ctx},i}(s)$ と呼ぶ) に対しても同様に計測し，その結果を図 3 に示す．いずれのモデルにおいても Sink トークン由来の更新ベクトル（実線）では多くの層で SpectralRatio が 80%以上であり，文脈トークン由来の更新ベクトル（点線）に比べて明確に高かった．これは Sink トークン由来の更新ベクトルが，入力文脈の内容に関わらず，高次元空間内でほとんど単一の方向（1 次元部分空間）に縮退していることを示している．したがって，Sink トークンは複雑な言語的特徴量ではなく，まるでバイアス項のように，単純なスカラー倍で表現可能な文脈非依存の定数ベクトルを各トークンの表現へ足し込んでいると解釈できる．

3.3 Normalized Variance: 時間的静止性

Sink トークンが真にバイアス項として機能しているのであれば，その出力は方向だけでなく，大きさも含めて時間的にほぼ一定のはずである．この仮説を検証するため，Sink トークン由来の更新ベクトルのうち，時間的に変動する成分（分散）が占める割合 $R^{(\ell)}$ を以下のように定量化し，結果を図 4 に示す．

$$R^{(\ell)} = \frac{\mathbb{E}_t\left[\|\boldsymbol{u}^{(\ell)}_{\text{sink},t} - \boldsymbol{\mu}^{(\ell)}_{\text{sink}}\|_2^2\right]}{\mathbb{E}_t\left[\|\boldsymbol{u}^{(\ell)}_{\text{sink},t}\|_2^2\right]} \tag{6}$$

通常の文脈トークン由来の更新ベクトル（点線）は文脈に応じた変動により $R^{(\ell)} \approx 1.0$ 付近を推移するのに対し，Sink トークン由来の更新ベクトル（実線）は $R^{(\ell)}$ が $10^{-2} - 10^{-1}$ という小さい値を示した．

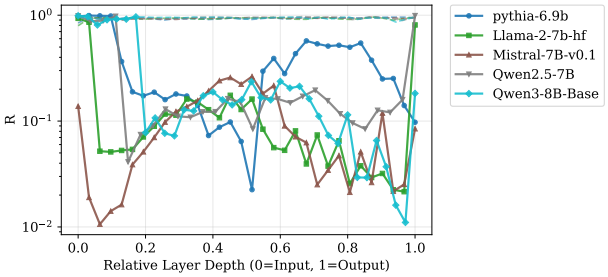

図 4: Sink トークン由来の更新ベクトル（実線）と文脈トークン由来の更新ベクトル（点線）における変動成分の支配率．値が小さいほど，位置によらず一定のベクトルを足し込むことを示す．

これは，Sink トークンの出力が方向だけでなく，その大きさも含めて時間的にほぼ不変（静的）であることを決定づける証拠である．特に Mistral-7B-v0.1 などの明示的なバイアス項を持たないモデルにおいてこの傾向は顕著であり，Sink トークン由来の更新ベクトルの加算が純粋な定数オフセットとして機能していることを裏付けている．

4 因果的介入

本章では，注意機構における Sink トークンからの動的なはずの更新を，事前計算された静的な平均ベクトルで強制的に置き換える介入手法 “Static Sink Patching (SSP)” を提案する．

4.1 Static Sink Patching

注意機構の計算において，各位置 i の Sink トークン由来の動的な項 $\boldsymbol{u}^{(\ell)}_{\text{sink},i}$ を $\boldsymbol{\mu}^{(\ell)}_{\text{sink}}$ で置き換える．

$$\underset{\text{new}}{\boldsymbol{y}_i^{(\ell)}} := \underset{\text{original}}{\boldsymbol{y}_i^{(\ell)}} - \boldsymbol{u}^{(\ell)}_{\text{sink},i} + \boldsymbol{\mu}^{(\ell)}_{\text{sink}} \tag{7}$$

ここで $\boldsymbol{\mu}^{(\ell)}_{\text{sink}}$ は i に依存しない定数である．本介入後もモデルの出力が維持されるならば，Sink トークン由来の成分は実質的に位置に依存しない定数項として作用していると解釈でき，Sink トークンが注意機構を介して文脈非依存なバイアス項を実装していることを示唆する．

SSP の効果を検証するため，対照として Sink トークン由来の寄与を単純に削除するアブレーション実験（式 7 で $\boldsymbol{\mu}^{(\ell)}_{\text{sink}}$ を加算しない設定）も行う．この対照実験により，SSP による性能維持が Sink トークン自体が不要であるためなのか，あるいは静的ベクトルによる代替が機能しているためなのかを識別することが可能となる．評価には § 2.2 と同様の手法

表 1: **モデル仕様と SSP 介入およびアブレーション適用時の PPL 変化**．左側は各モデルの構成要素（バイアス項の有無）を示す．右側は，介入手法を適用した際のベースラインからの PPL の変化量を示す．層範囲は，序盤層（30%），中盤層（40%），終盤層（30%）として定義した．

モデル	b（バイアス）	ベース	序盤層		中盤層		終盤層		全層	
			SSP	Abl.	SSP	Abl.	SSP	Abl.	SSP	Abl.
Mistral-7B-v0.1 [13]	False	6.94	+0.06	+1.19	±**0.00**	+0.56	±**0.00**	+0.22	+0.16	+2.38
Llama-2-7b-hf [12]	False	7.00	+0.28	+0.75	±**0.00**	+0.22	+0.06	+0.06	+0.28	+1.75
Qwen3-8B-Base [14]	False	10.44	+6.19	+4.75	±**0.00**	+0.63	±**0.00**	+2.13	+7.31	+4.94
Qwen2.5-7B [10]	True	11.25	+1.69	+1.13	+0.19	+0.56	+1.31	+1.31	+2.31	+1.31
Pythia-6.9b [11]	True	11.81	+10.56	+12.44	+1.38	+93.19	+0.19	+0.38	+6.81	+58.19

で構築したデータセットを用い，各モデルにおける PPL の変化を計測した（表 1 ）．このとき，$\boldsymbol{\mu}_{\text{sink}}^{(\ell)}$ の計算に使用したものとは異なる OpenWebText [15] から取得した長さ 1500 – 2000 文字のテキストを含む 100 個のプロンプトを使用した．

4.2 実験結果

表 1 に各モデルにおける介入結果を示す．まず，明示的なバイアス項 $\boldsymbol{b}$ を持たない Mistral-7B-v0.1 および Llama-2-7b-hf に注目すると，SSP の適用による PPL の悪化は見られずベースラインとほぼ同等の性能が維持された．一方で，単純な削除を行った場合，PPL は Mistral で +2.38，Llama-2 で +1.75 だけ悪化した．この結果は，Sink トークンが推論に不可欠な存在である一方で，その寄与が文脈に依存した動的計算を必要としないことを示している．この示唆は § 3 で示した幾何学的分析とも一貫し，Sink トークン由来の更新ベクトルは文脈非依存の静的なバイアス項として機能していると考えられる．対照的に，バイアス項が存在する Pythia-6.9b では SSP による全層介入で +6.81 と，性能低下が確認された．バイアス項がないモデルにおける $\boldsymbol{u}_{\text{sink},i}^{(\ell)}$ 静的な要素を，バイアス項があるモデルではバイアス項に任せ，$\boldsymbol{u}_{\text{sink},i}^{(\ell)}$ は動的な成分を主として扱っている可能性を示唆している．また，多くのモデルに共通して，序盤層における介入の影響が最も大きく，中盤層・終盤層では比較的影響が小さいという傾向が確認された．この層依存性は，Attention Sink が主として序盤層（特に第二層以降）で顕在化するという既存研究の観察 [2] と整合的であり，Sink トークンが初期段階で残差ストリームの基準点を形成している可能性を示唆しているなお，Qwen2.5-7B および Qwen3-8B-Base では，全層介入において SSP 適用時の PPL がアブレーション時の悪化を上回るという挙動が観測されており，注意機構のバイアス項の有無以外のアーキテクチャの観点を含めたさらなる分析が必要である．

5 考察と結論

本研究では，Attention Sink 現象に対し，幾何学的分析（§ 3）と介入実験（§ 4.1）を組み合わせることで，その機能的役割の解明を試みた．その結果は，Attention Sink を単なる学習の副産物ではなく，モデル内部で創発したバイアス項として捉える新たな解釈を支持するものである．まず幾何学的分析では，Sink トークンが生成する更新ベクトルは– (1) 層全体に影響を与える巨大なノルム，(2) 単一方向への縮退，(3) 時間的な不変性 –という三つの特性を示した．SSP による介入実験では，Sink トークン由来の動的な計算を，層ごとに事前計算した定数ベクトルで置換しても，明示的なバイアス項を持たない Mistral-7B-v0.1 や Llama-2-7b-hf において PPL がほぼ維持されることが示された．この結果は，Sink トークンの寄与が，位置や文脈に依存しない定数項として振る舞っていることを因果的に裏付けるものであり，Sink トークンがバイアス項の役割を担っている事を示唆する．本研究の知見は Attention Sink を Softmax の正規化制約や確率質量の退避先として捉える従来の解釈を補完し，LLM が内部的にどのような定数的バイアス成分を実装しているかを示す一つの具体例を与えるものである．今後は，本研究で示した創発的バイアスとういう視点を，他の特殊トークンや中間層の異常活性（Massive Activation）[9] へと拡張することで，LLM の内部表現の設計原理や推論効率化に関する理解がさらに深まると期待される．今後の課題として，Bias 項が存在するモデルの Bias 項の特性調査や層正規化層 (RMSNorm, LayerNorm) の特性等を考慮して，Attention Sink のバイアス的な解釈を深めていきたい．

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. In **Advances in Neural Information Processing Systems (NeurIPS)**, pp. 5998–6008, 2017.

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

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

[4] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)**, pp. 3991–4008, 2024.

[5] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable Transformers: Removing Outliers by Helping Attention Heads Do Nothing. In **Advances in Neural Information Processing Systems 36 (NeurIPS 2023)**, 2023.

[6] Evan Miller. Attention Is Off By One, 2023. Blog post.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and others. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In **International Conference on Learning Representations (ICLR)**, 2021.

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In **The Twelfth International Conference on Learning Representations**, 2024.

[9] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive Activations in Large Language Models. In **First Conference on Language Modeling**, 2024.

[10] Qwen, An Yang, Baosong Yang, and others. Qwen2.5 Technical Report. **arXiv:2412.15115**, 2025.

[11] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, and others. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. **arXiv:2304.01373**, 2023.

[12] Hugo Touvron, Louis Martin, Kevin Stone, and others. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv:2307.09288**, 2023.

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and others. Mistral 7B. **arXiv:2310.06825**, 2023.

[14] An Yang, Anfeng Li, Baosong Yang, and others. Qwen3 Technical Report. **arXiv:2505.09388**, 2025.

[15] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. OpenWebText Corpus, 2019.

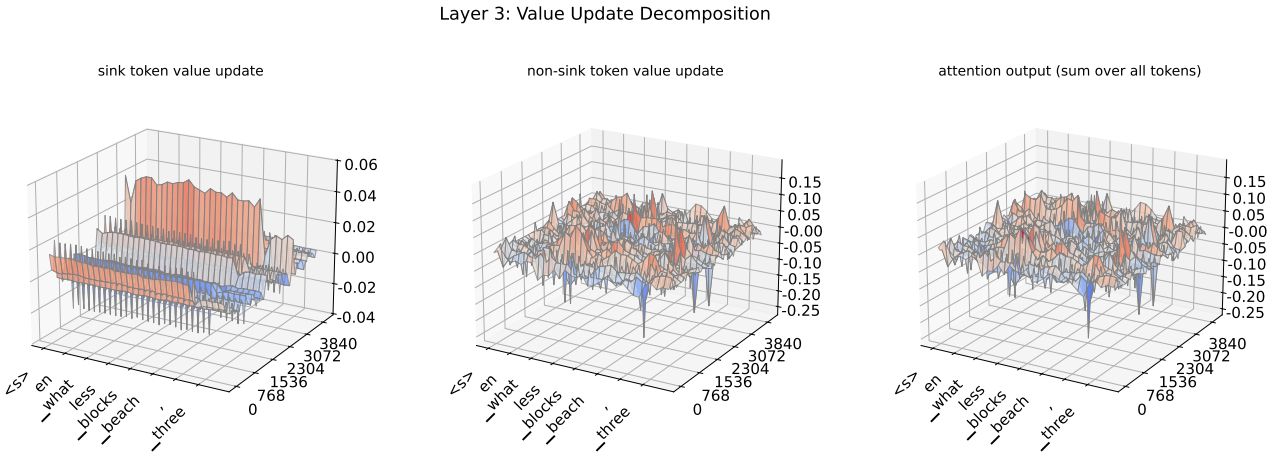

図 5: Llama-2-7b-hf における Layer 3 の Value Update 分解可視化. （左）Sink トークン由来の更新, （中）非 Sink トークン由来の更新, （右）全トークンからの更新の総和（Attention 出力）.

A Additional Visualization of Value Updates

本節では，本文で述べた Sink トークン由来の更新ベクトルの性質を直感的に補足するため, Value Update の分解可視化を示す. 具体的には，Llama-2-7b-hf を用い, ある代表的な層（Layer 3）において, Attention 出力を以下の 3 成分に分解して可視化した.

- Sink トークン由来の更新
- Sink 以外のトークン由来の更新
- 全トークンからの更新の総和（Attention 出力）

設定 入力系列長は 256 トークンとし, 各位置 i における Attention 出力 $\boldsymbol{y}_i^{(\ell)} = \sum_j \boldsymbol{u}_{j,i}^{(\ell)}$ を対象とした. ここで $\boldsymbol{u}_{j,i}^{(\ell)}$ は，第 ℓ 層においてトークン j から位置 i へ寄与する Value Update を表す.

可視化結果 図 5 に, Layer 3 における Value Update の分解結果を示す. 左図は Sink トークン由来の更新, 中央は非 Sink トークン由来の更新, 右図はそれらの和である Attention 出力を表す.

Sink トークン由来の更新は，トークン位置に依らず非常に類似した形状を示しており，高次元空間においてほぼ同一方向に 大きな成分を持つことが確認できる. 一方で，非 Sink トークン由来の更新は，位置や次元に応じて符号・大きさが大きく変動しており，文脈依存な情報を担っている様子が観察される.

これらの可視化結果は, 本文で示した Sink トークンの幾何学的特徴, すなわち **(i) Rank-1 構造, (ii) 時間的安定性, (iii) 大きなノルム** が，実際の Attention 出力の分解においても 直感的に確認できることを示している.