

B6-15

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

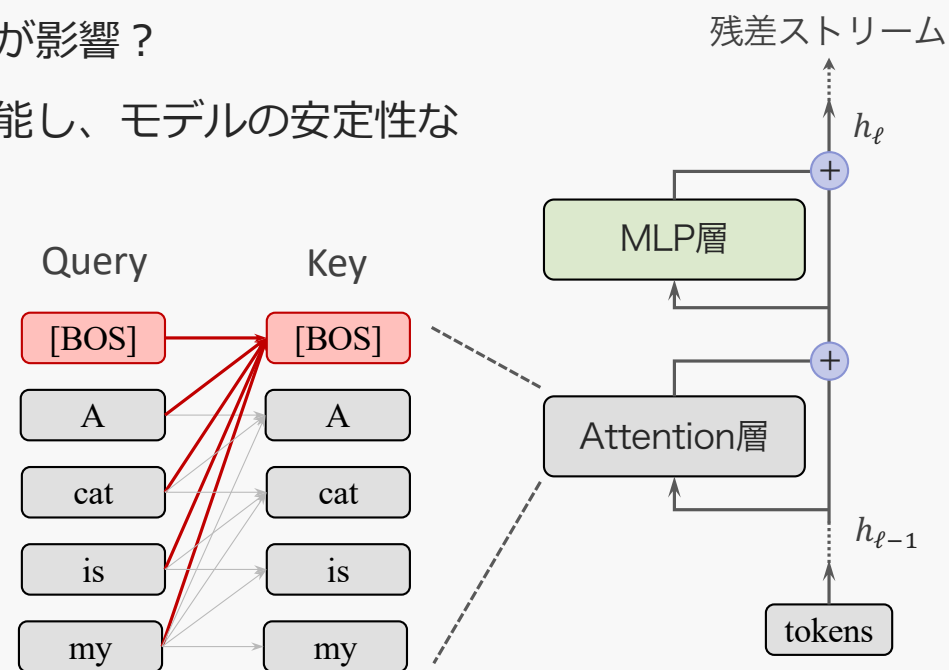
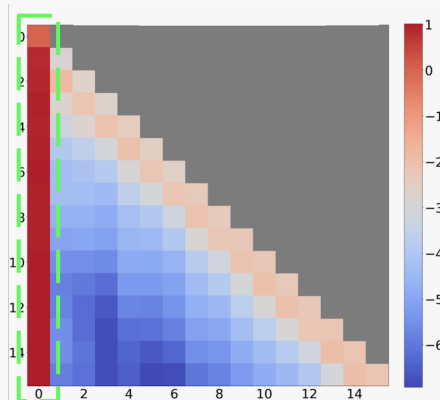
言語処理学会第32回大会@宇都宮, 2026年03月11日



注意の過剰な集中 - Attention Sink -

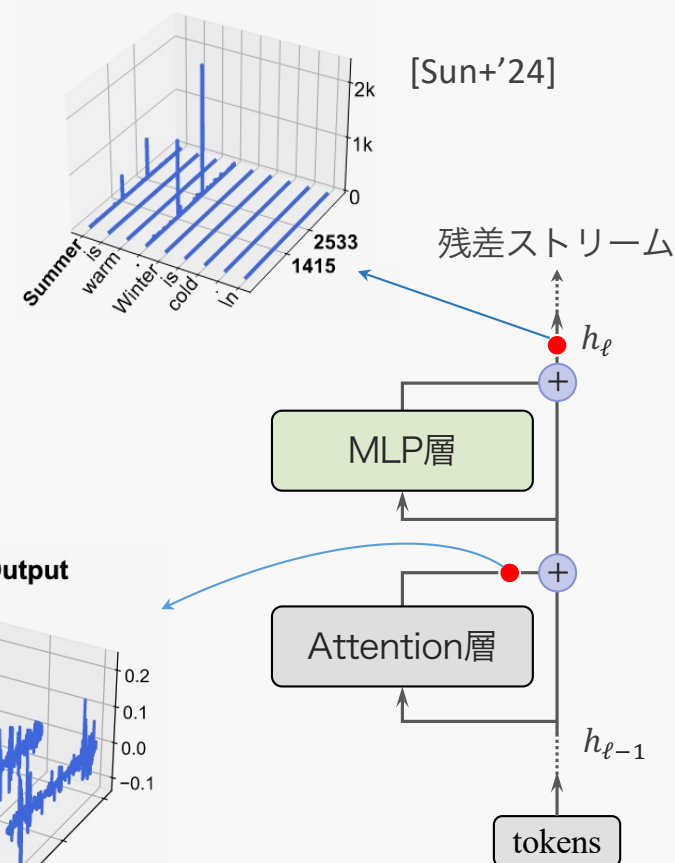
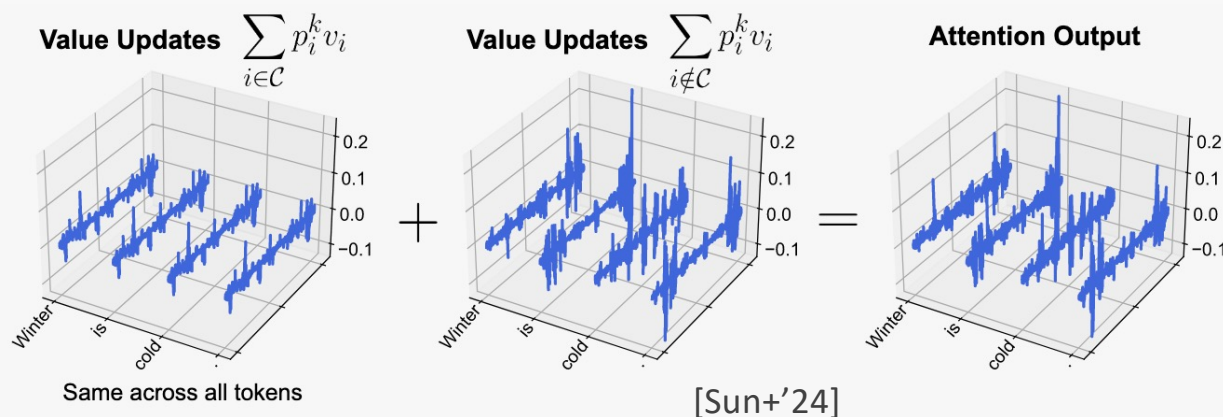
- 先頭トークン(等)に注意重みが集中する現象（区切り文字などでも起きる）
- Softmax による総和を1にするための正規化制約が影響？
- いらぬ重みを捨てるための「ゴミ箱」として機能し、モデルの安定性などに寄与

[Xiao+'24]



Attention Sink と Massive Activation

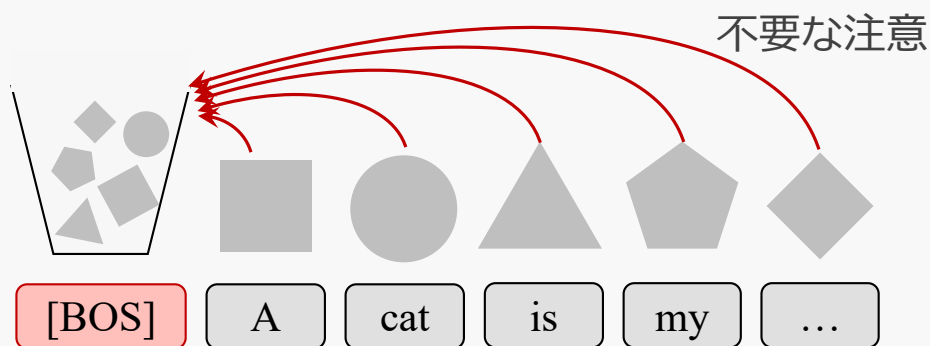
- Massive Activation
 - MLP 後の隠れ状態 (h_ℓ) での特定トークンへの異常活性化現象
 - 異常値が起きたトークンでは、更新ベクトルが一定
- Attention Sink と同様のトークンで発生



Attention Sink は単なる「ゴミ箱」じゃない

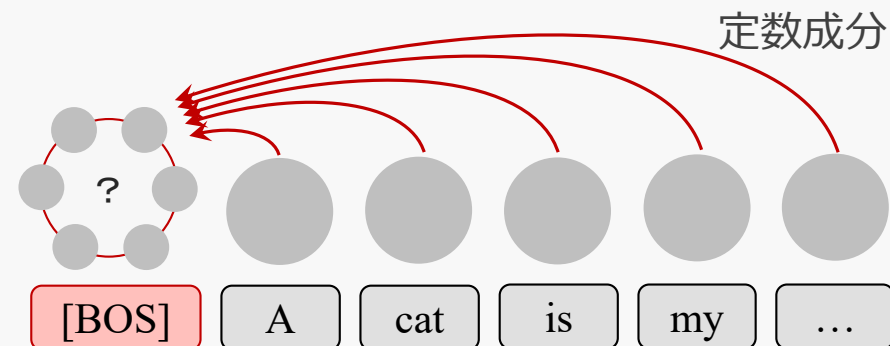
[BOS] A cat is my favorite animal.

従来の解釈（ゴミ箱説）



- 不要な注意重みを捨てるための場所

近年の知見と本研究の着眼点

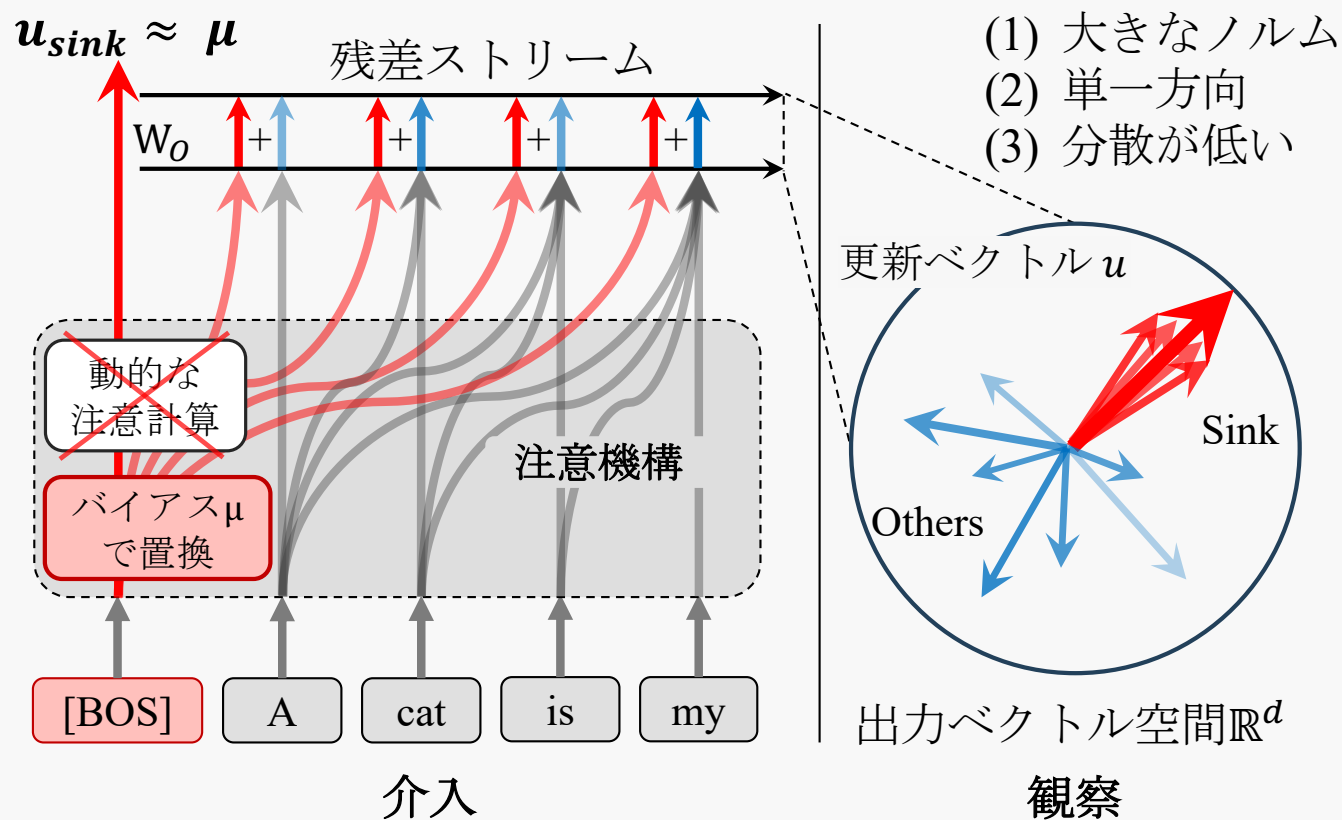


- 実質的にバイアスとしての機能
- 高次元空間でのベクトルの実態は？
- モデルで汎化する知見なのか？

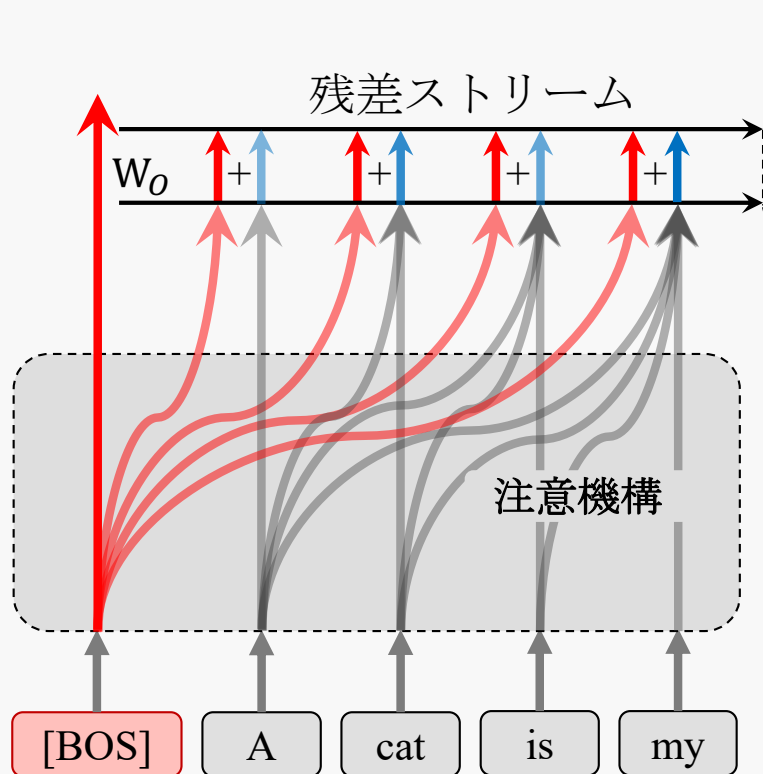
研究概要

- Sink トークン由来の更新ベクトルを全層を通じて観察
- 先頭トークン（Sink）由来の更新ベクトルは、**大きなノルム・単一方向性・時間的静止性**を持つ
- 定数置換介入（SSP）により、本現象が**バイアス項を持たないモデル特有の補完メカニズム**である可能性を示唆

研究概要



幾何学的観察と定数置換 (SSP) による検証



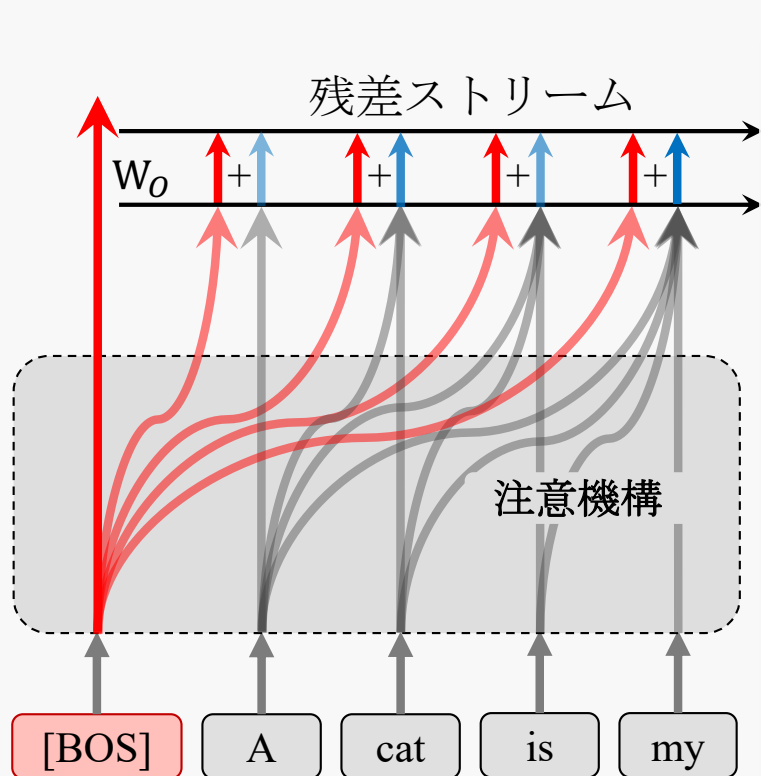
$$y_i^{(\ell)} = \overbrace{u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)}}^{\text{更新ベクトル}} + b^{(\ell)}$$

注意機構からの出力を2つに分解

- Sink トークン由来の更新ベクトル
- その他トークン由来の更新ベクトル

出力ベクトル空間 $\mathbb{R}^d$
観察

幾何学的観察と定数置換 (SSP) による検証



更新ベクトル

$$y_i^{(\ell)} = u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)} + b^{(\ell)}$$

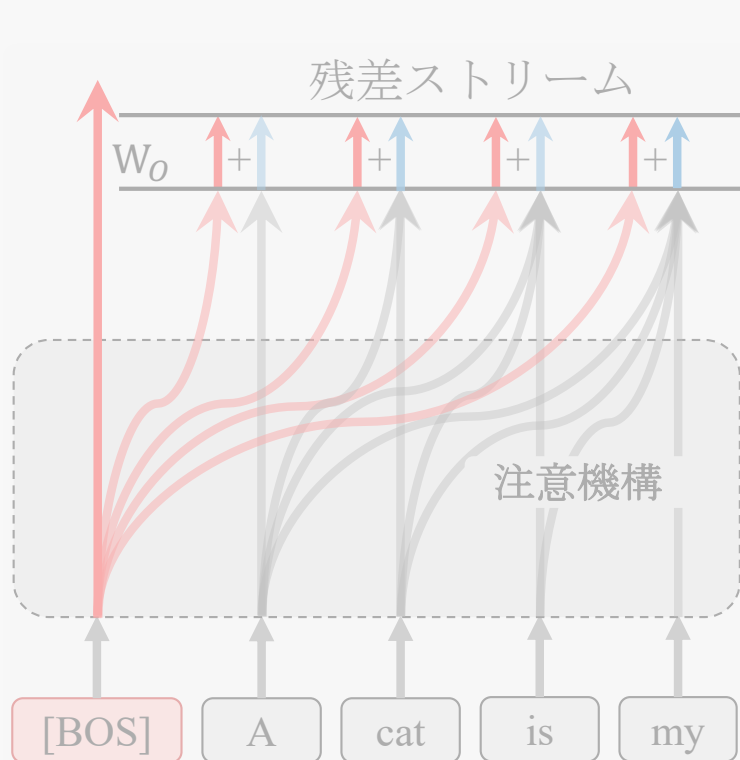
更新ベクトル u

バイアス項がないモデルでは、Sinkがこれを代替する

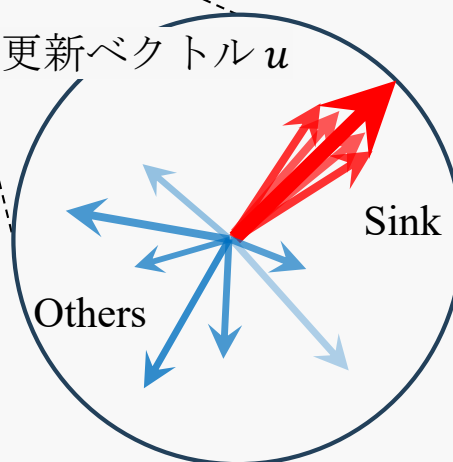
仮説

Attention Sink は、モデルがアーキテクチャ上の**バイアス項の欠落を補完**するための機構である

幾何学的観察と定数置換 (SSP) による検証



更新ベクトル u



出力ベクトル空間 $\mathbb{R}^d$

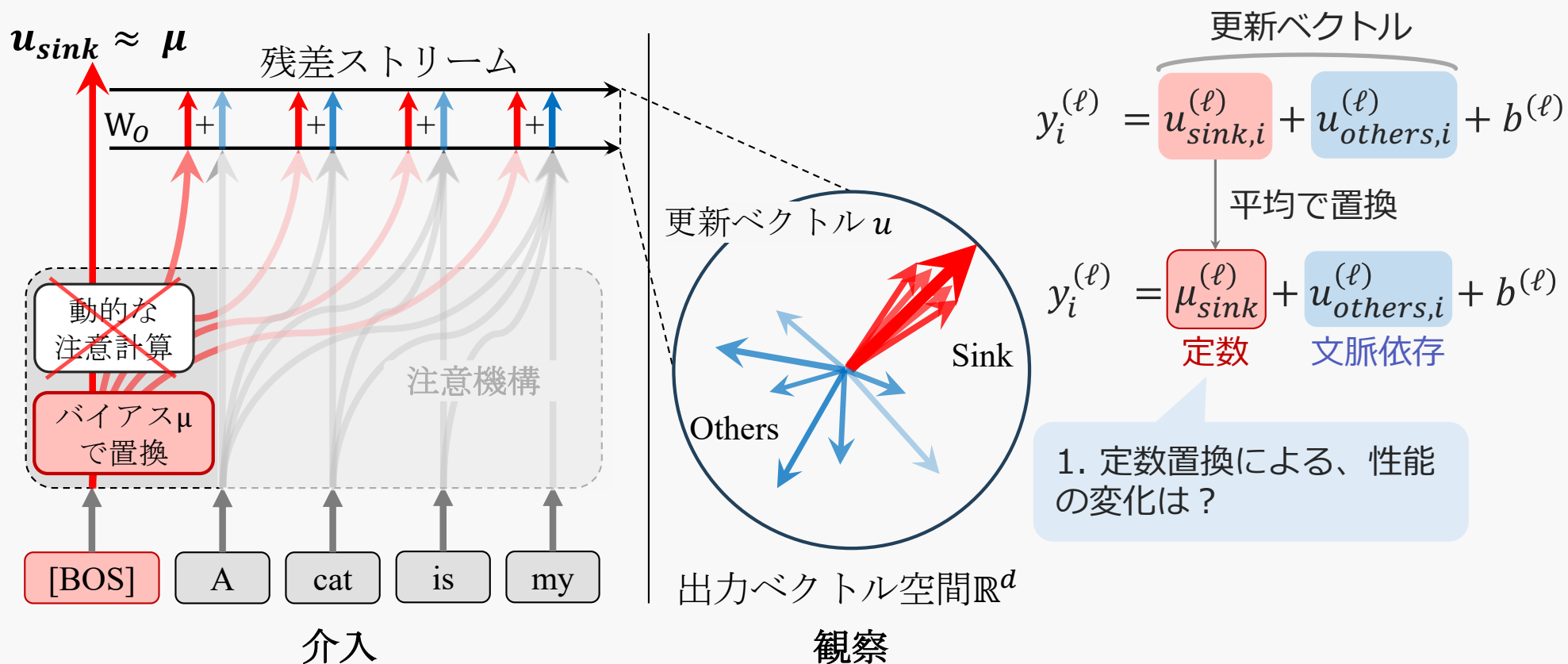
観察

$$y_i^{(\ell)} = \overbrace{u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)}}^{\text{更新ベクトル}} + b^{(\ell)}$$

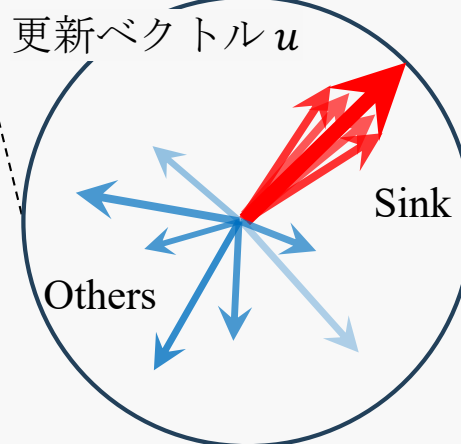
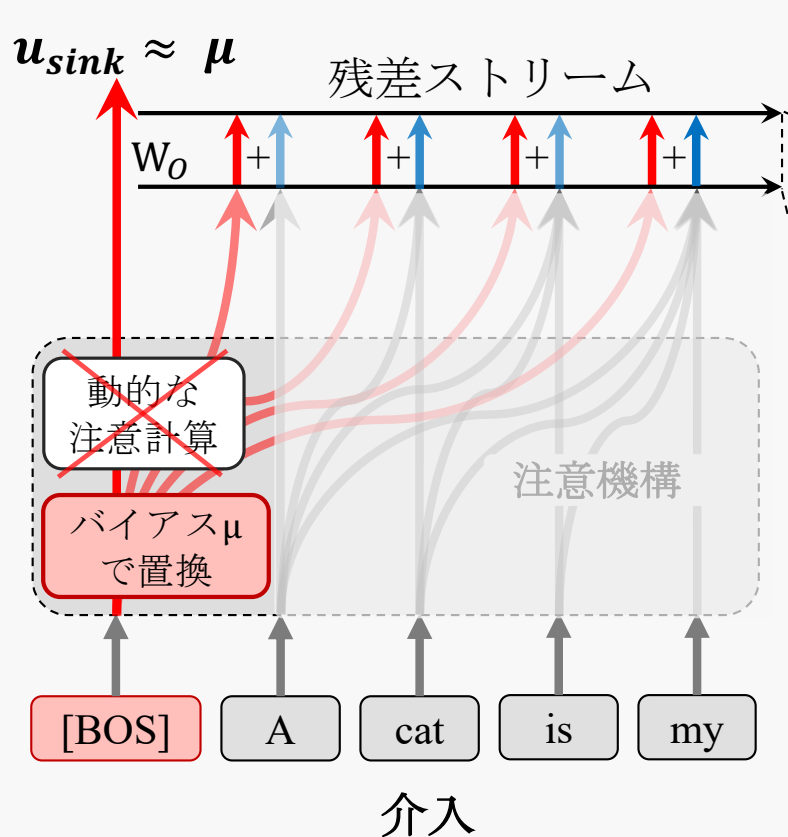
先頭更新ベクトルの性質

- (1) 大きなノルム
- (2) 単一方向性
- (3) 時間的静止性

幾何学的観察と定数置換 (SSP) による検証



幾何学的観察と定数置換 (SSP) による検証



出力ベクトル空間 $\mathbb{R}^d$

観察

更新ベクトル

$$y_i^{(\ell)} = u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)} + b^{(\ell)}$$

平均で置換

$$y_i^{(\ell)} = \mu_{sink}^{(\ell)} + u_{others,i}^{(\ell)} + b^{(\ell)}$$

定数 文脈依存

1. 定数置換による、性能の変化は？

2. 明示的なバイアス項との関係は？

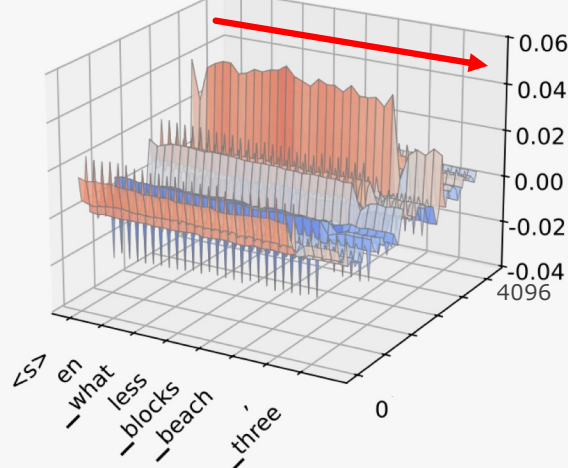
実験設定

- モデル
 - バイアス無し : Llama2-7b-hf, Mistral-7B, Qwen3-8B-Base
 - バイアス有り : Qwen2.5-7B, Pythia-6.9b
- データセット
 - OpenWebText から抽出した100 サンプルのテキスト
 - 512トークンで長さを統一
- 評価指標
 - Perplexity (PPL) でモデル性能を検証

Sink トークン由来の更新ベクトルは「定数」として振る舞う

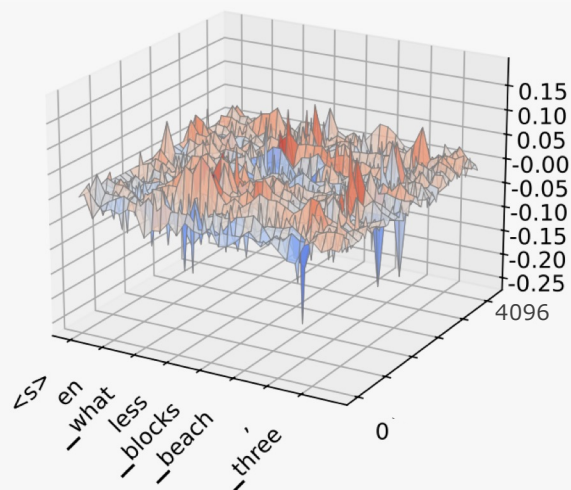
分散の低さ

$$\|u_{sink} - \mu_{sink}\| \approx 0$$



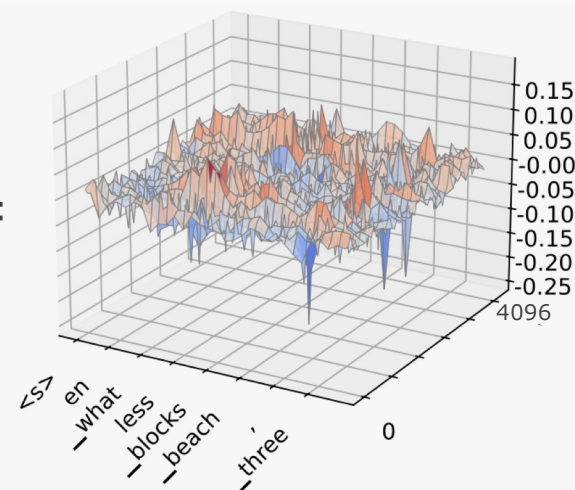
Sink 由来の更新ベクトル

+



Non-Sink 由来の更新ベクトル

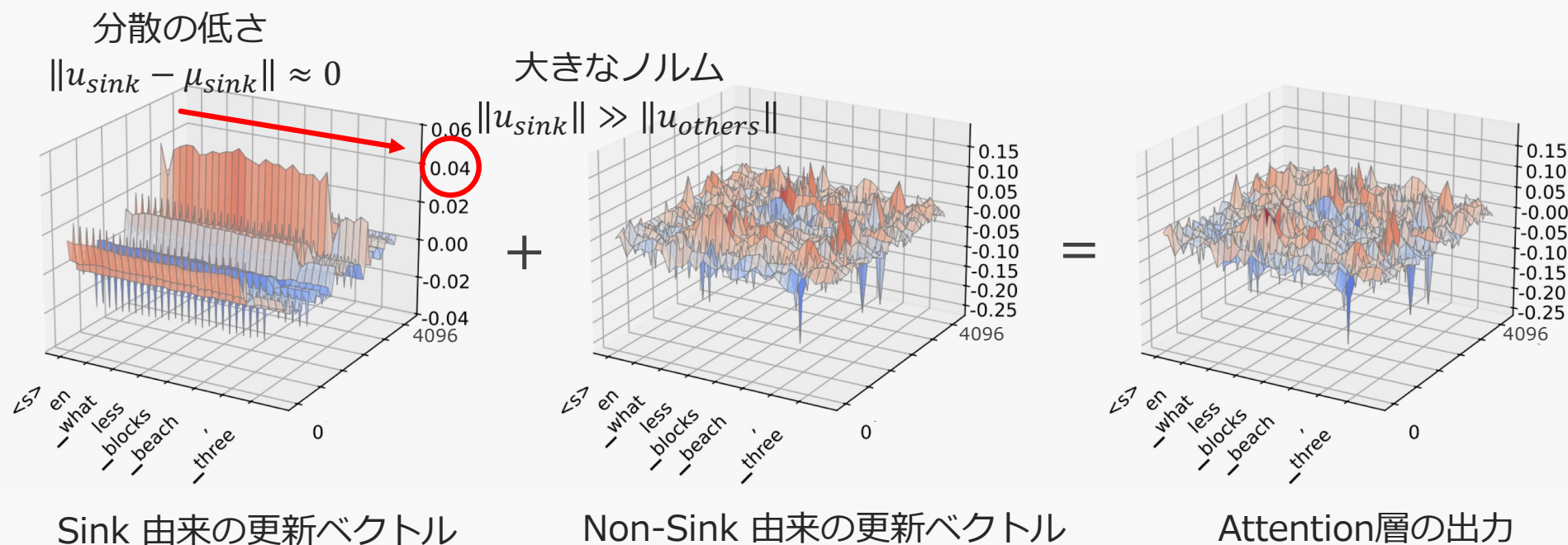
=



Attention層の出力

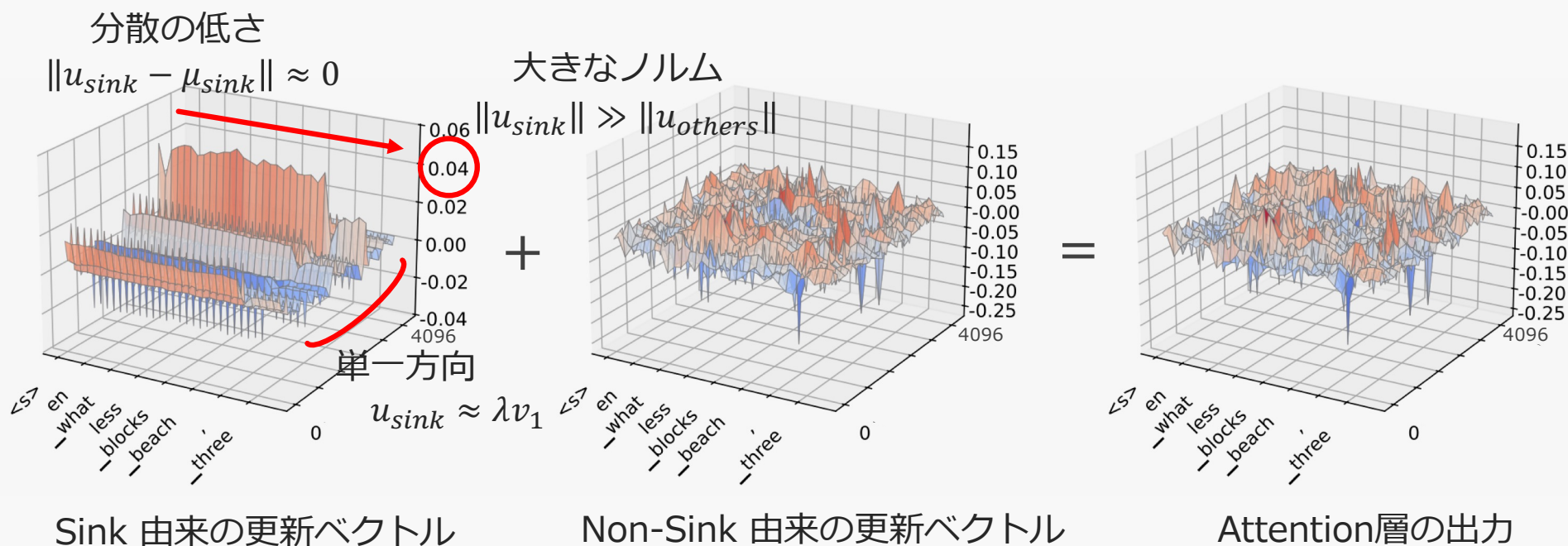
$$y_i^{(\ell)} = u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)} + b^{(\ell)}$$

Sink トークン由来の更新ベクトルは「定数」として振る舞う



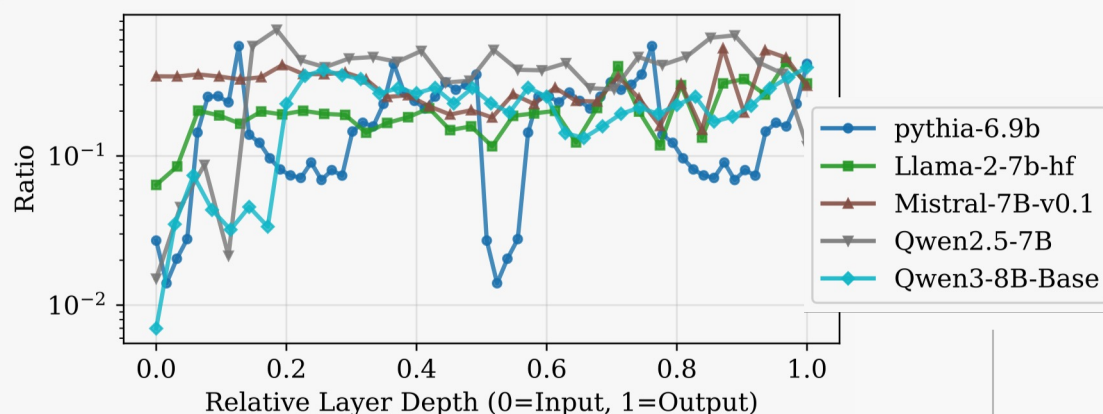
$$y_i^{(\ell)} = u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)} + b^{(\ell)}$$

Sink トークン由来の更新ベクトルは「定数」として振る舞う



$$y_i^{(\ell)} = u_{sink,i}^{(\ell)} + u_{others,i}^{(\ell)} + b^{(\ell)}$$

Sink トークン由来の更新ベクトルは「定数」として振る舞う



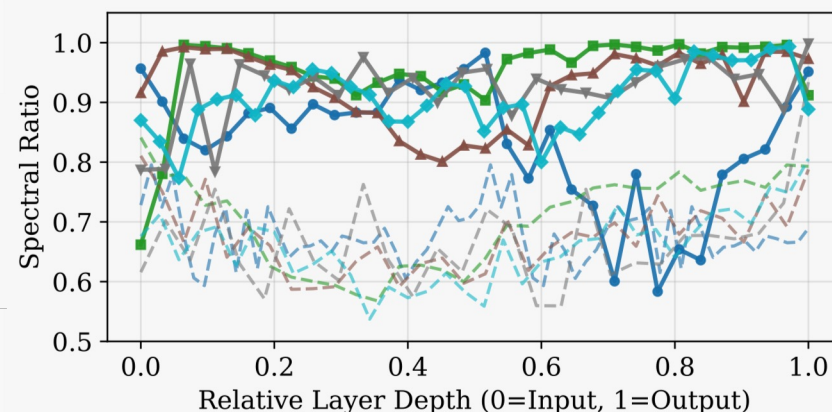
$$\frac{\mathbb{E} \left[\left\| u_{sink}^{(\ell)} \right\| \right]}{\mathbb{E} \left[\left\| u_{others}^{(\ell)} \right\| \right]}$$

ノルムの比較

- 全体に対してSinkが占める割合

- Sink 単体で文脈全体の**20~30%**の大きさ

➡ モデル出力に対して影響力を持つ



$$\frac{\sigma_1^2}{\sum_k \sigma_k^2}$$

方向の偏り（特異値分解）

- 第1 特異成分が占める割合

- Sink の更新ベクトル第1 特異成分だけで**80~100%**を説明できる

➡ Rank-1 に縮退する

定数ベクトルへの置換でも性能維持、実質的なバイアス項として機能

SPP (置換) < Abl. (削除) ➡ 定数成分が含まれることを示す

モデル	バイアス項の有無	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	False	6.94	+0.06	+1.19	±0.00	+0.56	±0.00	+0.22	+0.16	+2.38
Llama-2-7b-hf	False	7.00	+0.28	+0.75	±0.00	+0.22	+0.06	+0.06	+0.28	+1.75
Qwen3-8B-Base	False	10.44	+6.19	+4.75	±0.00	+0.63	±0.00	+2.13	+7.31	+4.94
Qwen2.5-7B	True	11.25	+1.69	+1.13	+0.19	+0.56	+1.31	+1.31	+2.31	+1.31
Pythia-6.9b	True	11.81	+10.56	+12.44	+1.38	+93.19	+0.19	+0.38	+6.81	+58.19

- Mistral, Llama2 等では、置換を行っても**性能が維持**される
- Pythia では、性能は低下するが削除にくらべて大幅な性能改善
➡ バイアス無しモデル、Sink がバイアス成分を**強く補完**

$$y_i^{(\ell)} = \underbrace{\mu_{sink}^{(\ell)}}_{\text{定数}} + \underbrace{u_{others,i}^{(\ell)}}_{\text{文脈依存}} + b^{(\ell)}$$

定数ベクトルへの置換でも性能維持、実質的なバイアス項として機能

SPP (置換) < Abl. (削除) ➡ 定数成分が含まれることを示す

モデル	バイアス項の有無	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	False	6.94	+0.06	+1.19	±0.00	+0.56	±0.00	+0.22	+0.16	+2.38
Llama-2-7b-hf	False	7.00	+0.28	+0.75	±0.00	+0.22	+0.06	+0.06	+0.28	+1.75
Qwen3-8B-Base	False	10.44	+6.19	+4.75	±0.00	+0.63	±0.00	+2.13	+7.31	+4.94
Qwen2.5-7B	True	11.25	+1.69	+1.13	+0.19	+0.56	+1.31	+1.31	+2.31	+1.31
Pythia-6.9b	True	11.81	+10.56	+12.44	+1.38	+93.19	+0.19	+0.38	+6.81	+58.19

- Qwenシリーズにおいて、仮説通りの結果は得られず
- 原因は、Qwenシリーズでは<BOS>トークンを使用せず、学習を行なっているから？

$$y_i^{(\ell)} = \underbrace{\mu_{sink}^{(\ell)}}_{\text{定数}} + \underbrace{u_{others,i}^{(\ell)}}_{\text{文脈依存}} + b^{(\ell)}$$

おわりに

まとめ

- Transformer モデルにおける Attention Sink の機能的役割を調査
- 先頭トークン（Sink）由来の更新ベクトルは、**大きな影響力・単一方向性・時間的静止性**を持ち静的である
- 定数置換介入（SSP）により、本現象が**バイアス項を持たないモデル特有の補完メカニズム**である可能性を示唆

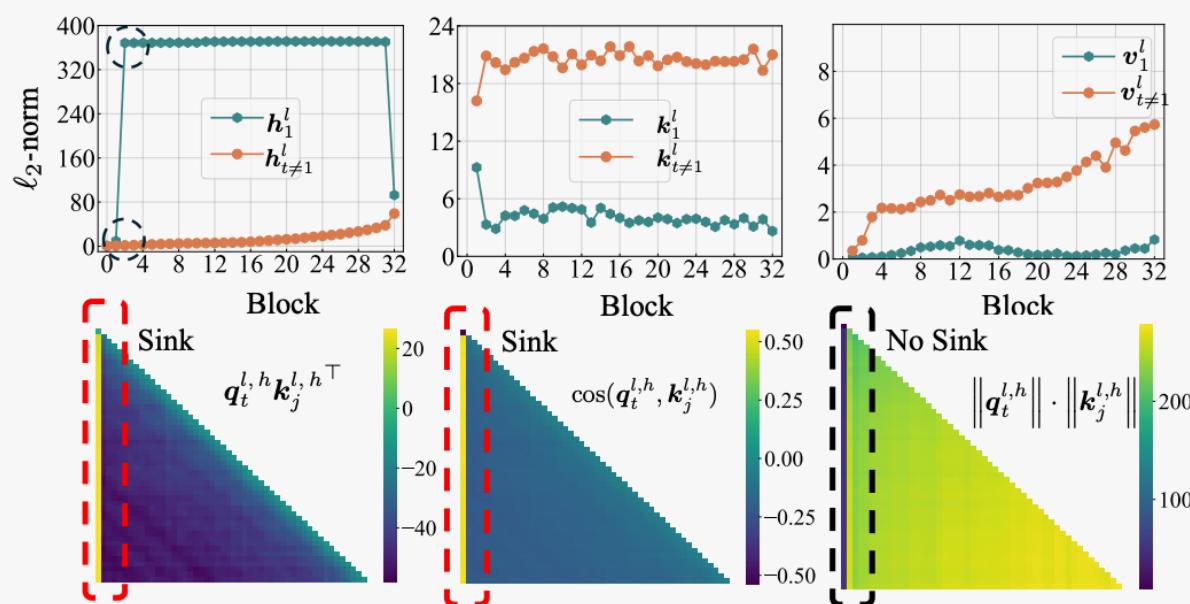
展望

- 学習条件を統一して比較検証
- 他のアーキテクチャ要素（MLP層のバイアス、層正規化）との相互作用

| 付録 / Appendix

Attention Sink は「いつ、どのように、なぜ」発生する？

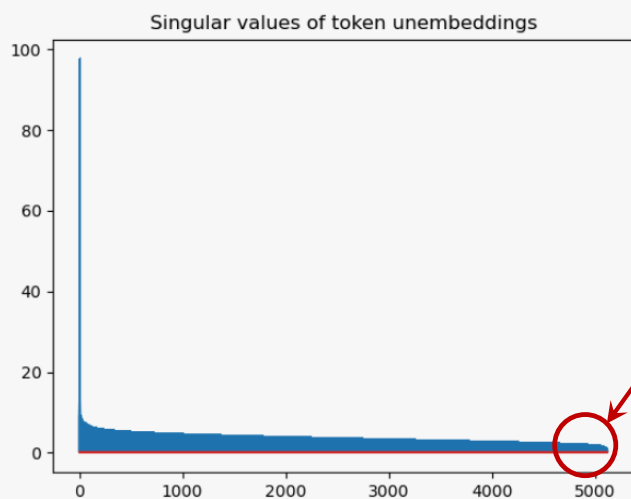
- 学習がある程度進んだ後の（序盤層）で発生する
- **Key**ベクトルが他のトークンのQueryベクトルと**近い方向**を向くことで発生
- Softmax関数の正規化制約により発生する



[Xiangming+'25]

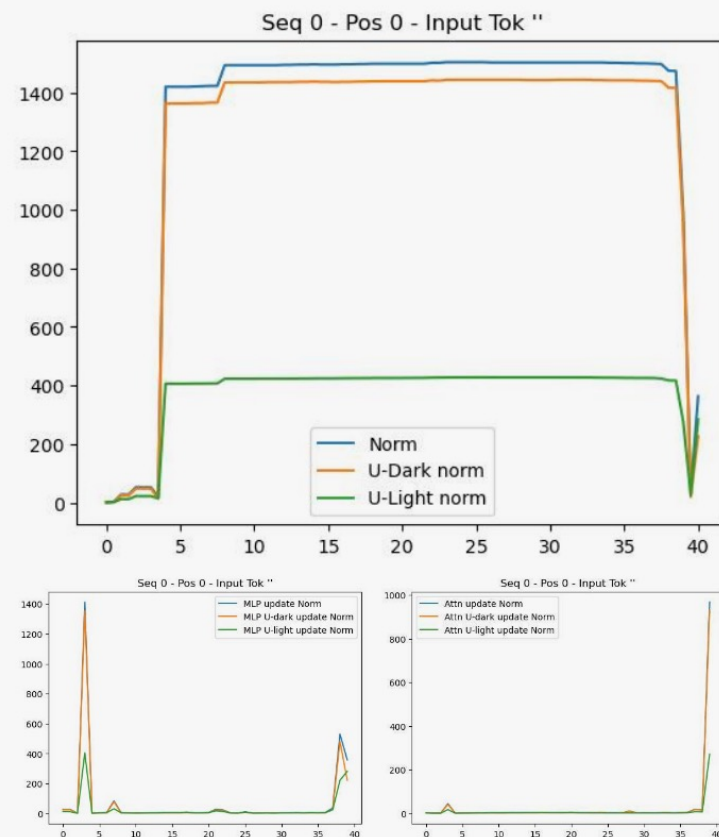
Attention Sink はどこで発生する？

- 中間表現を語彙空間に射影する際の Unembedding 行列で使われない「**暗黒空間**」を利用して実装される
- 主に**序盤のMLP層**にて発生し、**終盤のAttention層**によって打ち消される



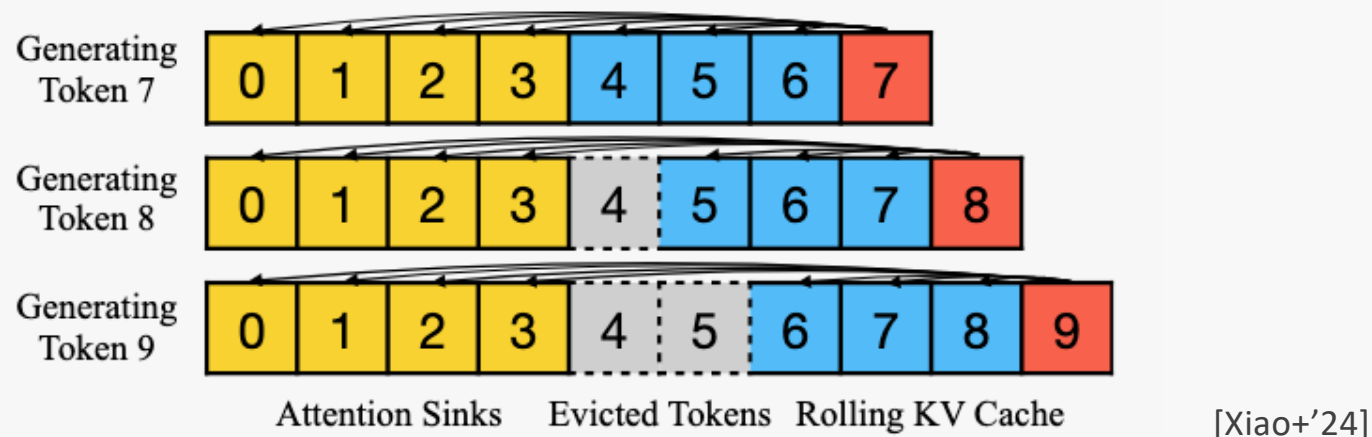
暗黒空間

[Nicola+'24]



Attention Sink の利用

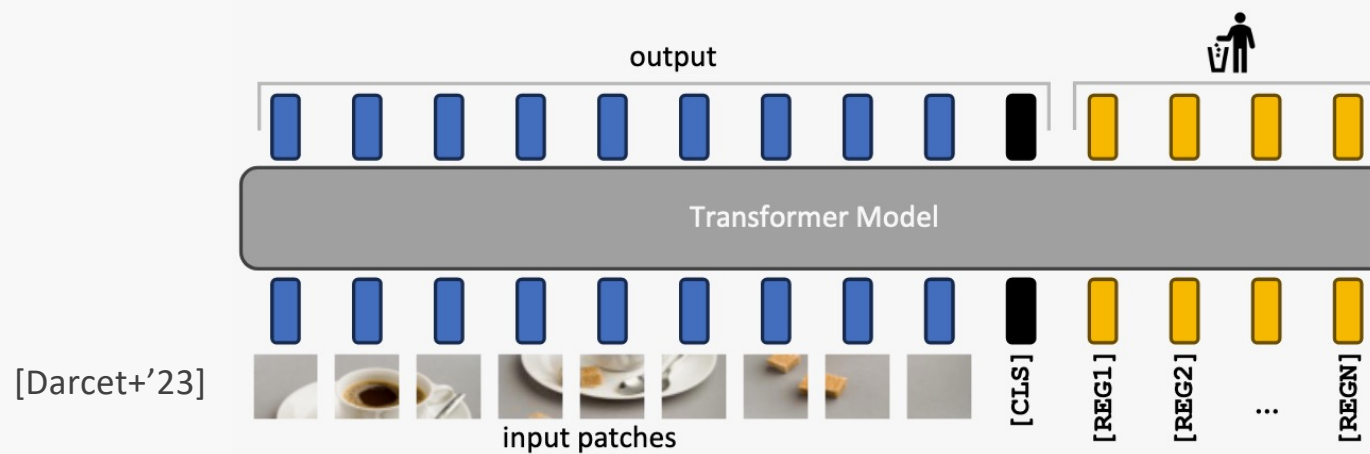
- StreamingLLM
 - 先頭数トークンをPrefixトークンとして維持することで無限長の入力を可能にする



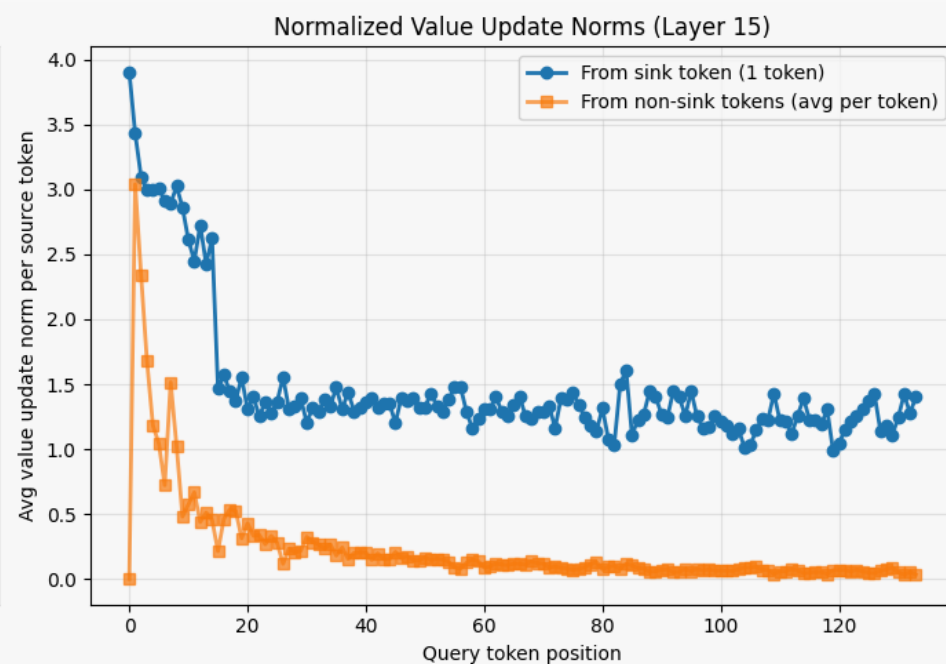
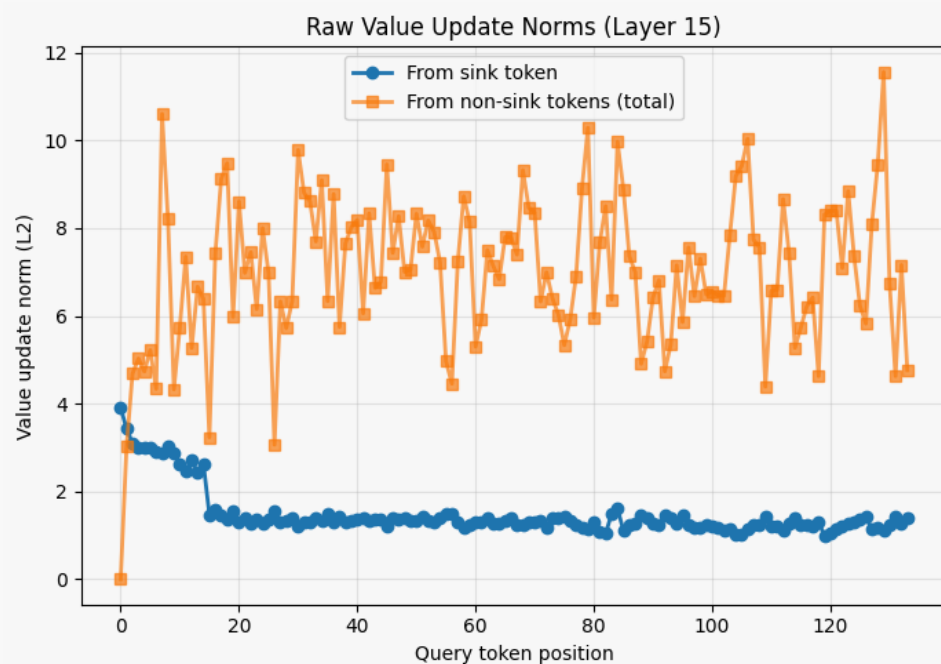
Vision Transformer と Attention Sink

- Vision Transformer

- 注意の偏りが起きるトークンは、グローバルな情報を保持してる可能性がある
- 最後にレジスタトークンを用意することで、注意の偏りを解決する



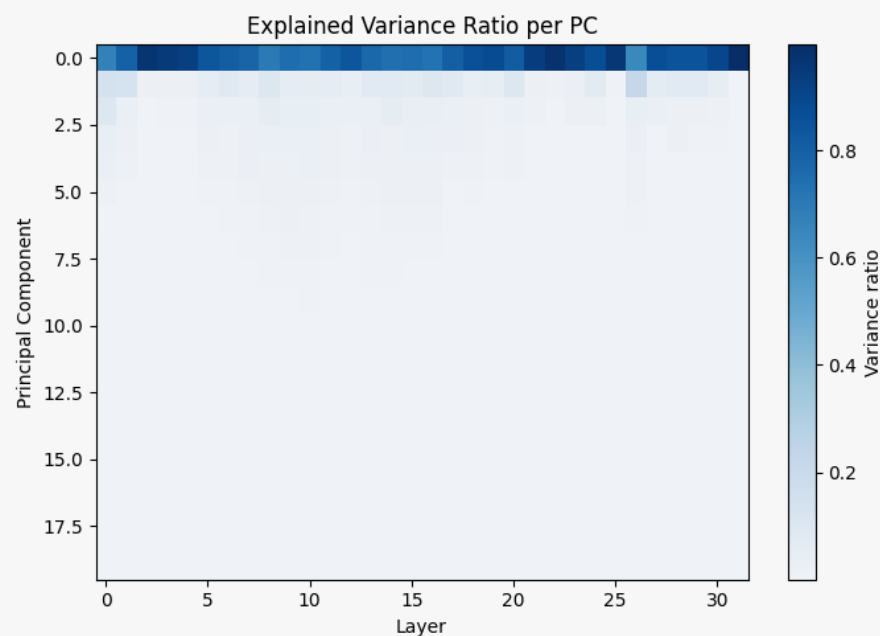
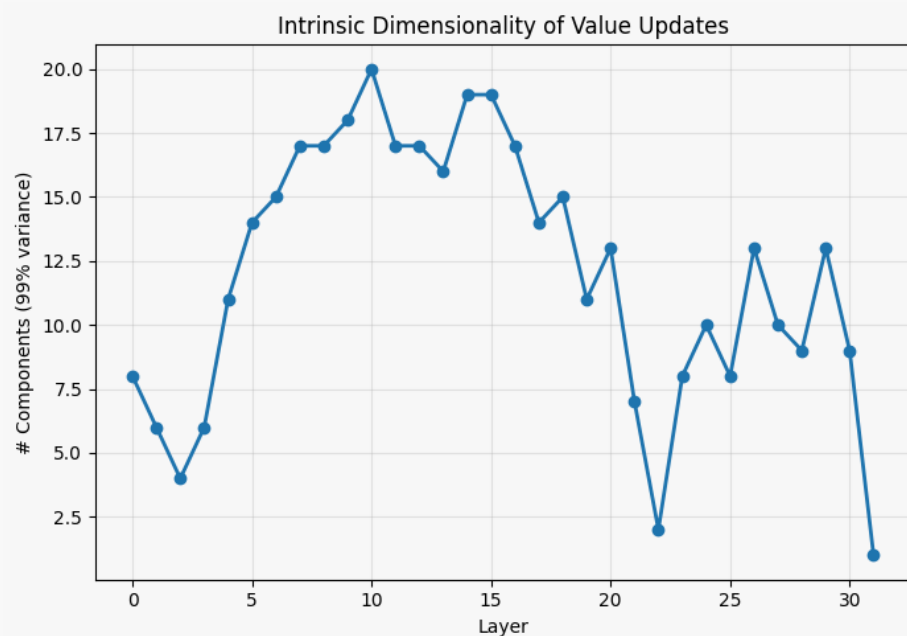
Sink 由来の更新ベクトルのノルムについて（補足）



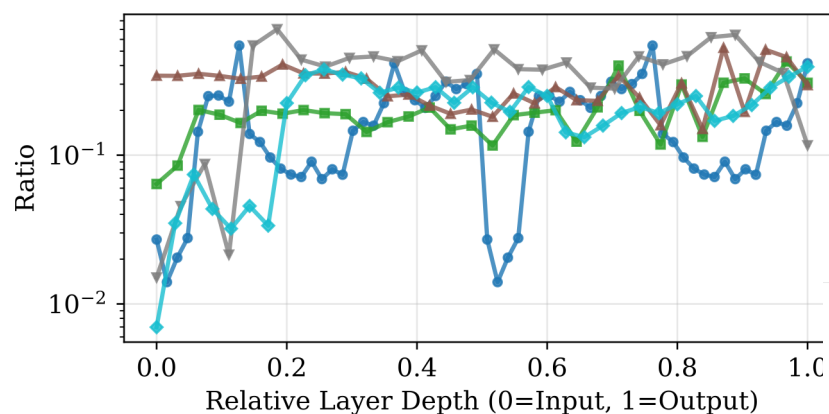
Llama2-7b-hf

Sink 由来の更新ベクトルの単一方向性について（補足）

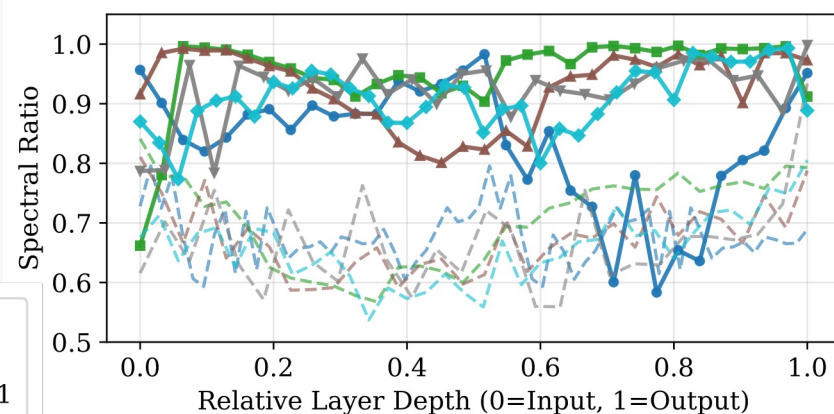
- 4096次元のうち実際は20次元程度しか利用していない
- 多くは1次元の成分が占める



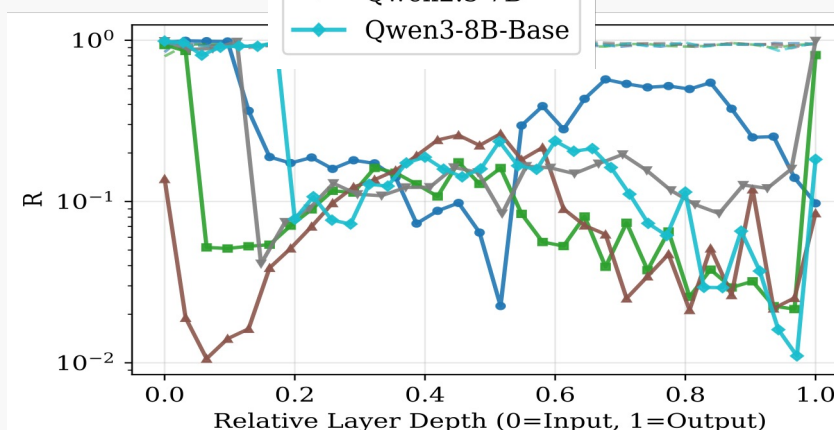
Sink トークン由来の更新ベクトルは「定数」として振る舞う



$$\frac{\sigma_1^2}{\sum_k \sigma_k^2}$$



$$\frac{\mathbb{E} \left[\left\| u_{sink}^{(\ell)} \right\| \right]}{\mathbb{E} \left[\left\| u_{others}^{(\ell)} \right\| \right]}$$



$$\frac{\mathbb{E} \left[\left\| u_{sink}^{(\ell)} - \mu_{sink}^{(\ell)} \right\|^2 \right]}{\mathbb{E} \left[\left\| u_{others}^{(\ell)} \right\|^2 \right]}$$

注意機構における Attention Sink のバイアス項的解釈 | NLP2026