

概要

- 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

背景・動機

Attention Sink 現象：

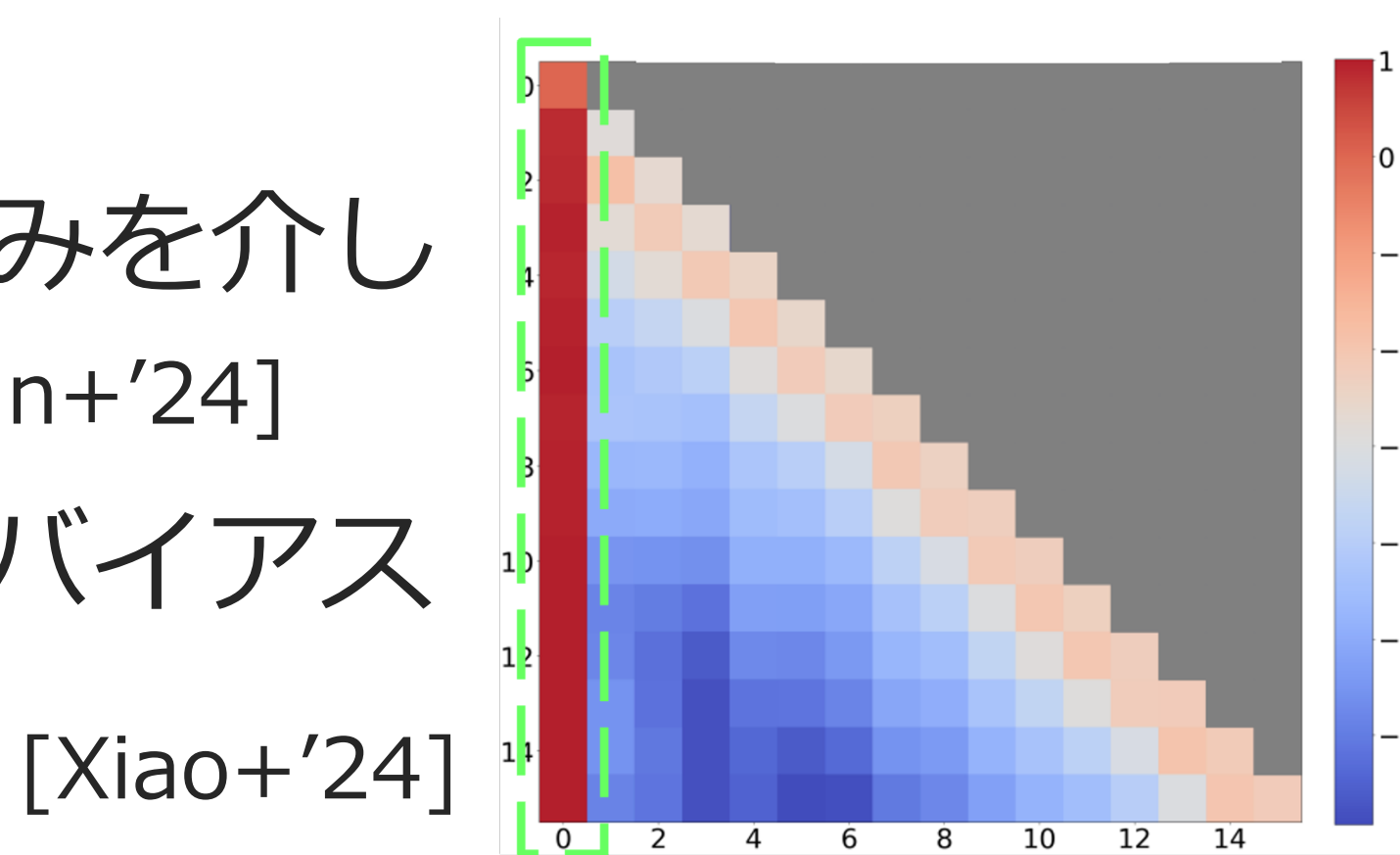
- 先頭トークン(等)に注意が集中し、Softmax の「ゴミ箱」として機能 (モデルの安定性に寄与)

関連研究：

- Massive Activation は重みを介してバイアスとして機能 [Sun+'24]
- Attention Sink も同様にバイアスとして機能??

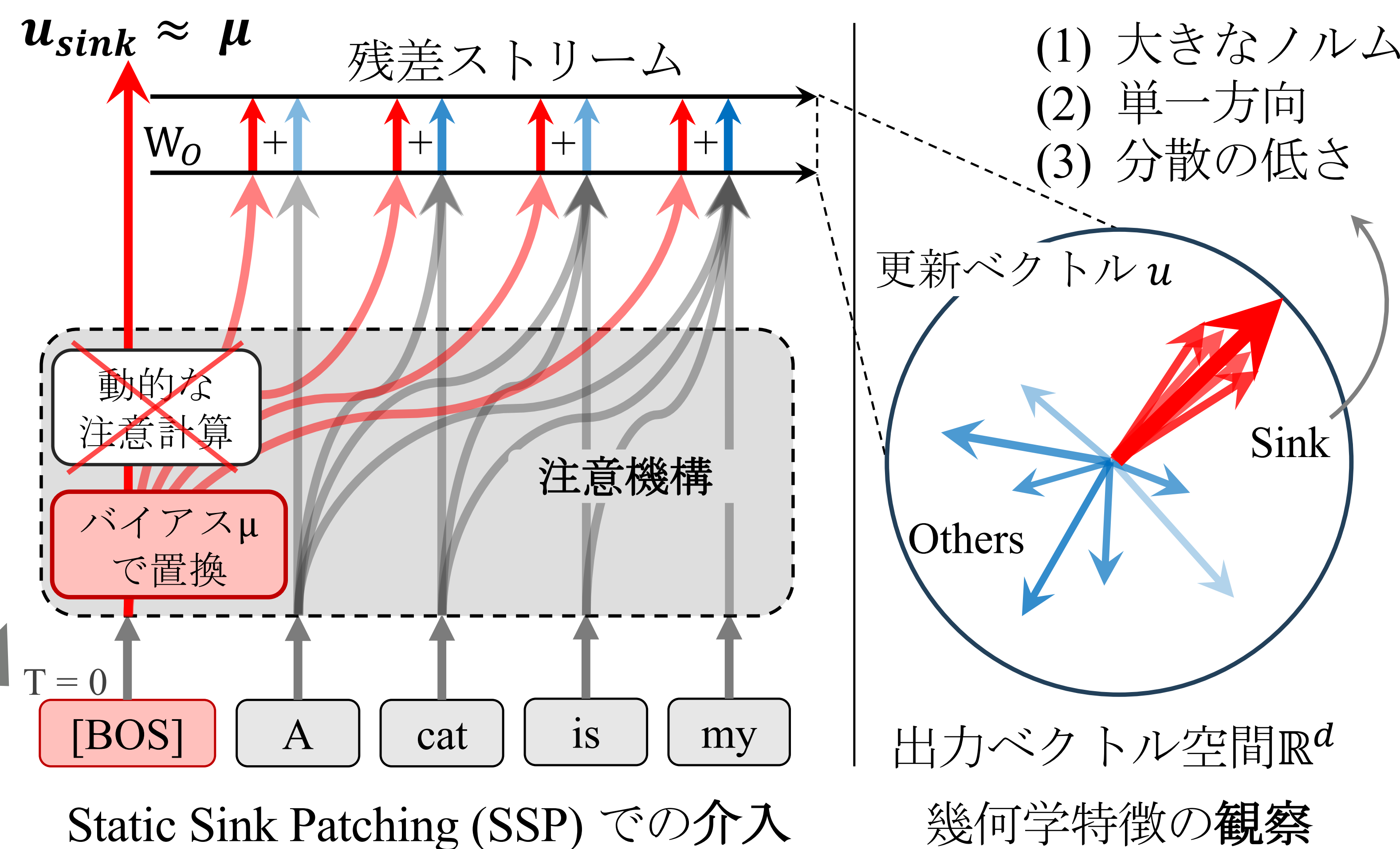
目的：

- 幾何学的特性 / モデルアーキテクチャとの関連性？



[Xiao+'24]

実験概要・設定



T=0

Static Sink Patching (SSP) での介入

データセット：

- OpenWebText から抽出 (100サンプル)
- 入力長：512トークン

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

定式化と評価

観察 注意機構の出力を分解

更新ベクトル

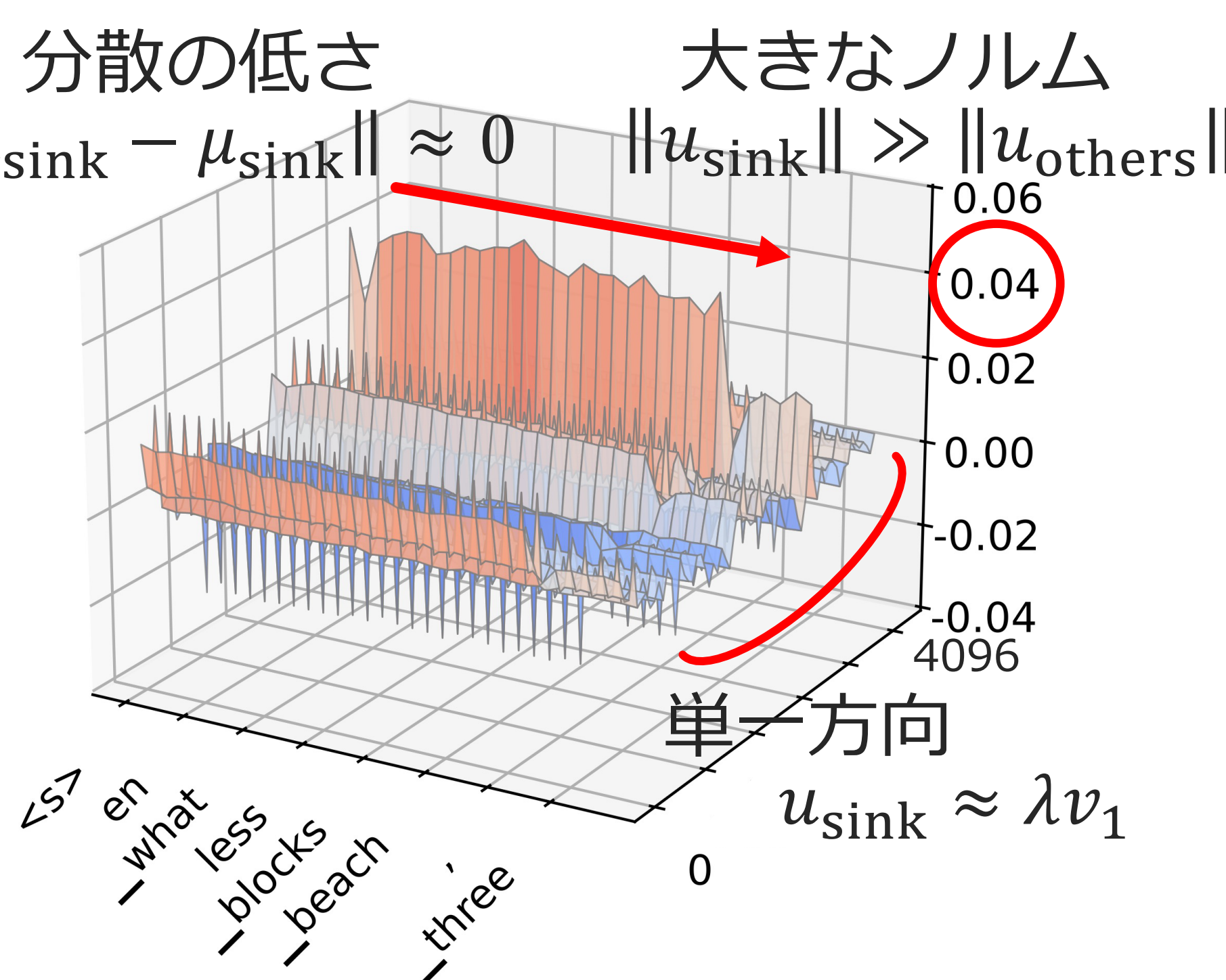
$$y_i^{(\ell)} = u_{\text{sink},i}^{(\ell)} + u_{\text{others},i}^{(\ell)} + b^{(\ell)}$$

介入

平均で置換

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

定数 文脈依存



先頭トークンが持つ更新ベクトルの性質

- 大きな影響力、Rank-1縮退、時間的静止性
- LLama2、Mistralでは、SSPで性能維持
- ✓ 欠落したバイアス項を補完している

課題と例外

- Qwen では、バイアス有無と結果が不整合
- ✗ バイアス項以外の構造的要因も影響

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

今後の展望

- 学習条件を統一した比較検証
- 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

結果の見方：

ベースラインと同等の性能 ($|\Delta\text{PPL}| < 0.5$)大幅な性能劣化 ($|\Delta\text{PPL}| > 2.0$)

SSP < Abl. → 定数的成分が含まれることを示唆