

B6-13

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

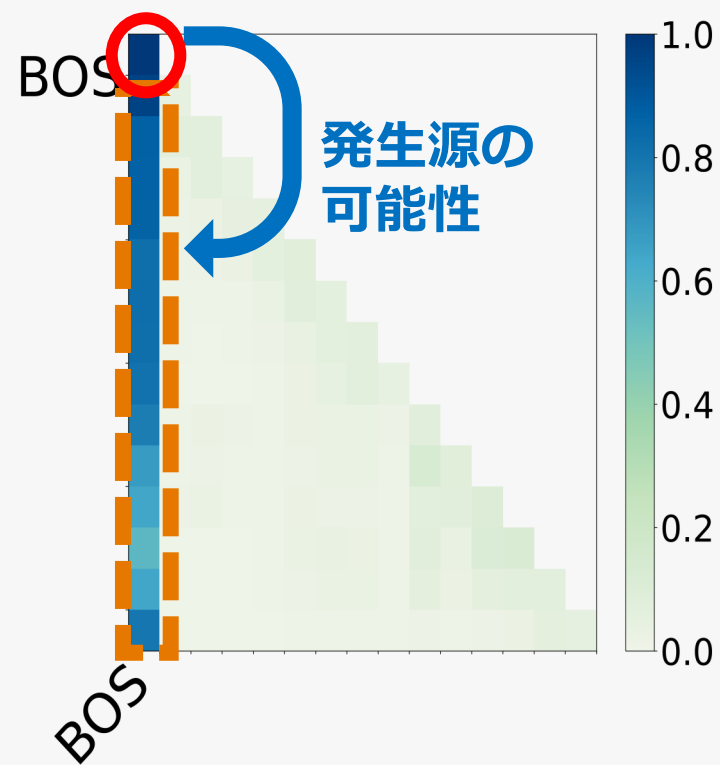
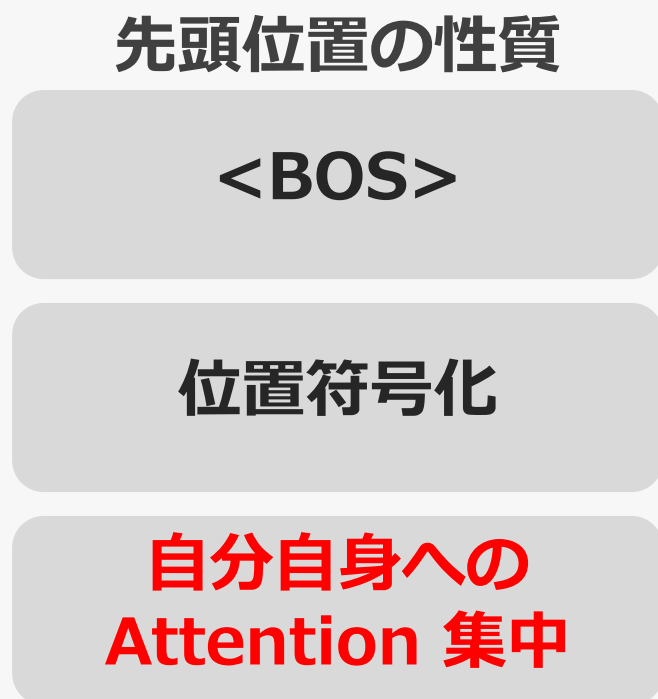
(¹ 東北大学 ² 総合研究大学院大学 ³ 国立国語研究所 ⁴ 理化学研究所)

✉ raito.kiya@dc.tohoku.ac.jp

言語処理学会第32回年次大会2026年3月12日

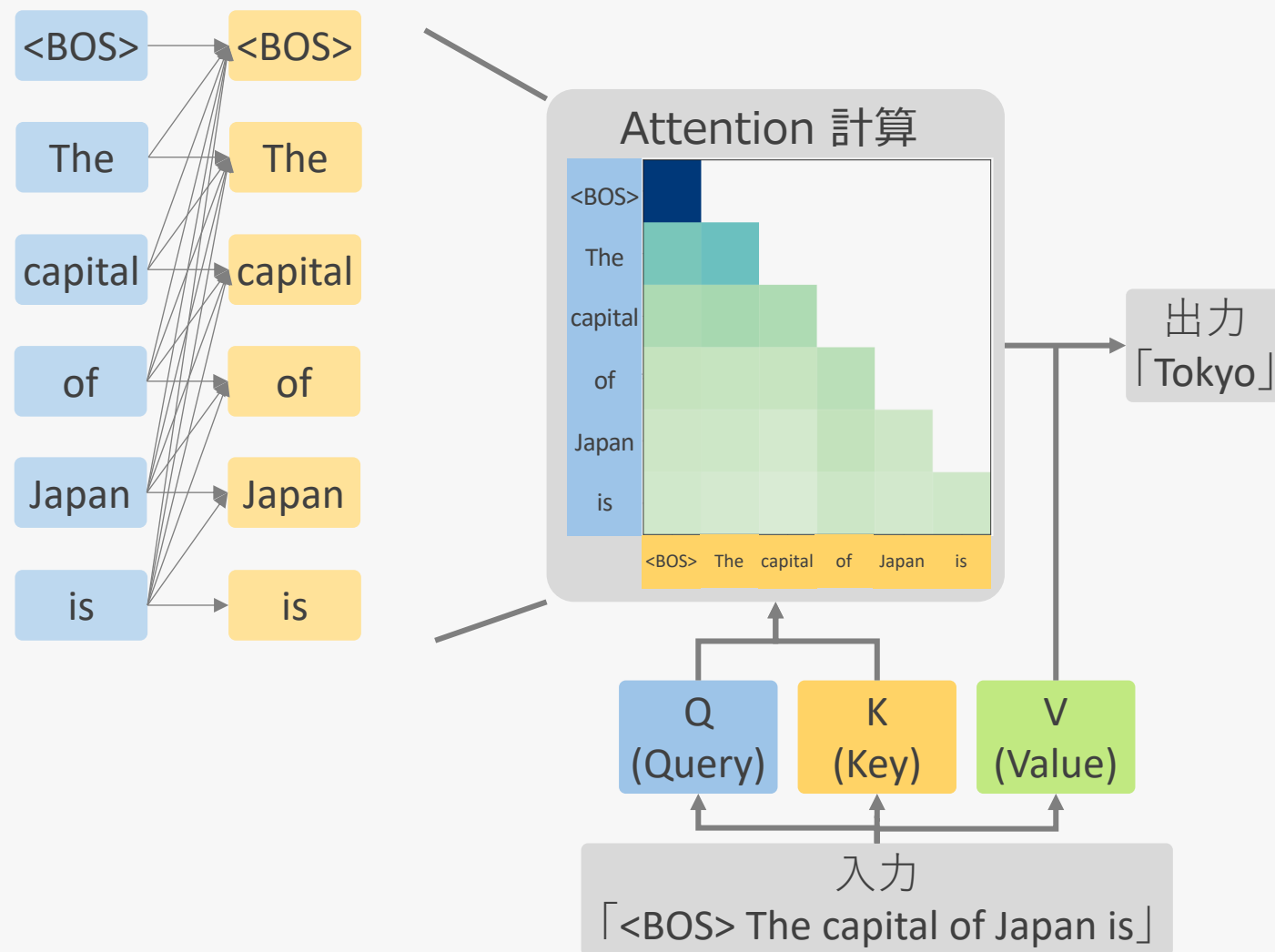
LLM の先頭注目現象：発生源の分解と特定

- Attention Sink（先頭_等 への過度な注目）の発生源を分解
- 自身への Attention 集中が発生源である可能性を示唆



Attention 機構とは

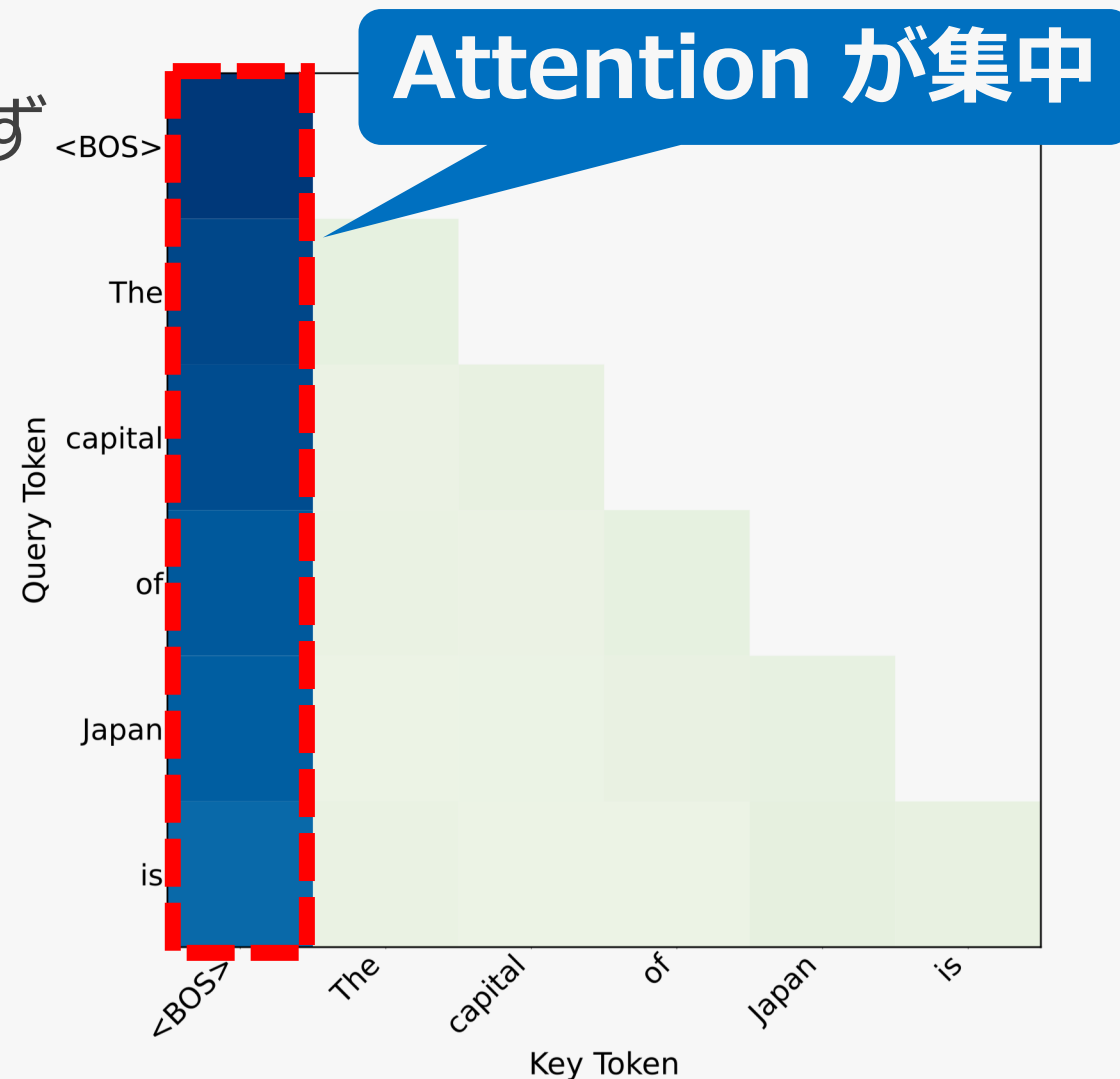
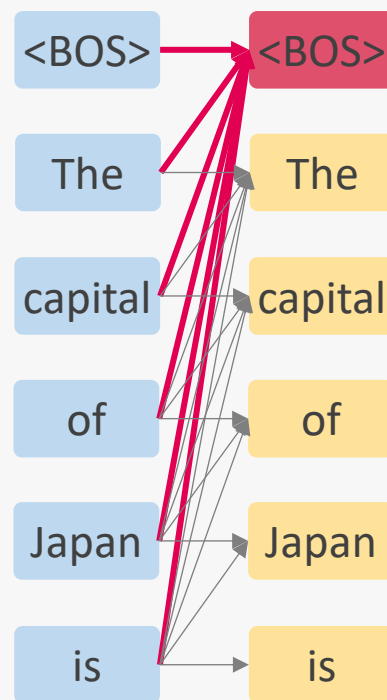
- 文脈情報を現在のトークンの表現に動的に混ぜ込む役割



先頭への過度な注目 –Attention Sink–

- LLM では、意味的関連性に関わらず先頭(等)に Attention スコアが集中

➡ **Attention Sink** 現象 [Xiao+'24]
(以降 Sink と呼ぶ)

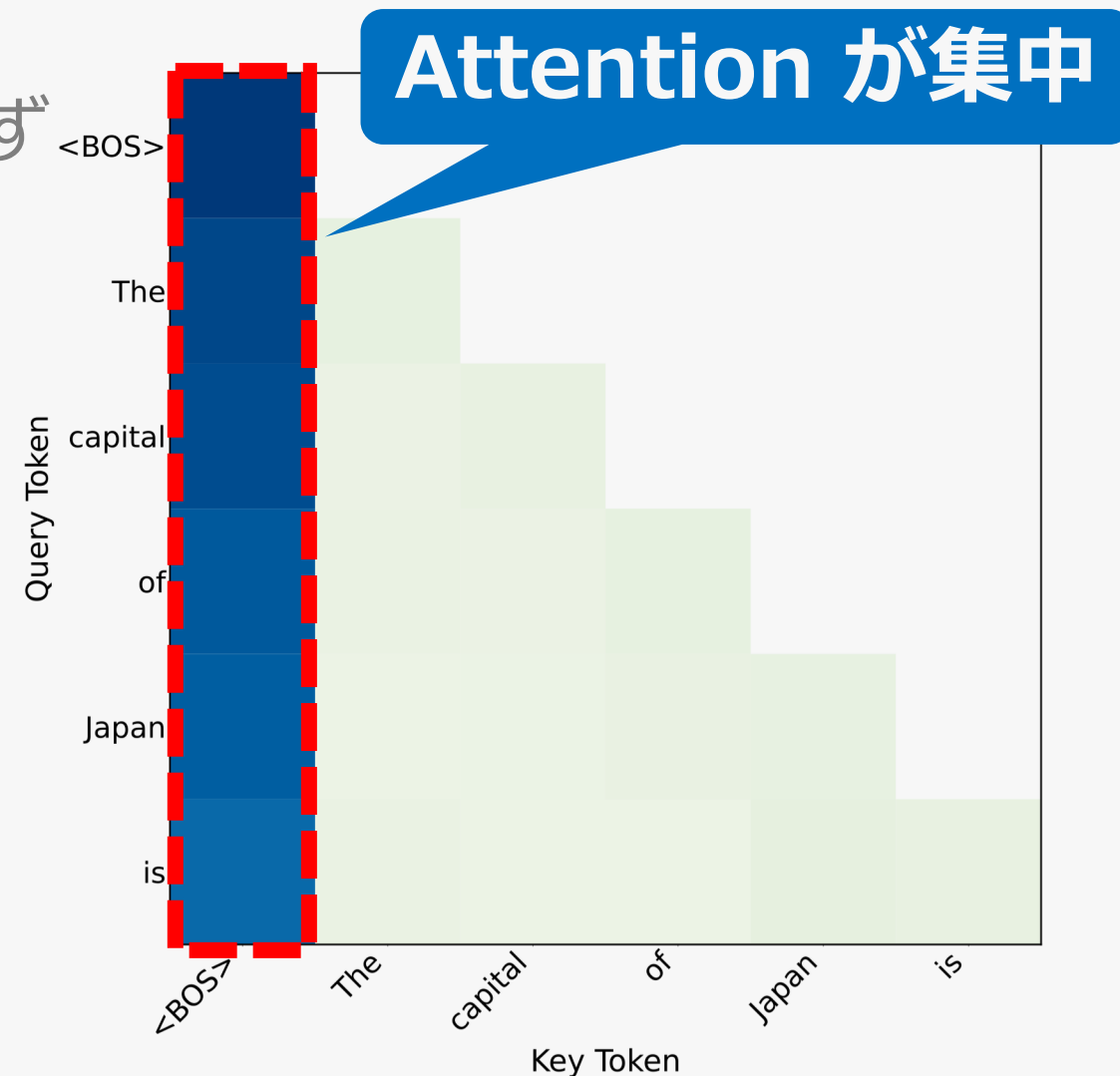


先頭への過度な注目 —Attention Sink—

- LLM では、意味的関連性に関わらず先頭(等)に Attention スコアが集中
➡ **Attention Sink** 現象 [Xiao+'24]

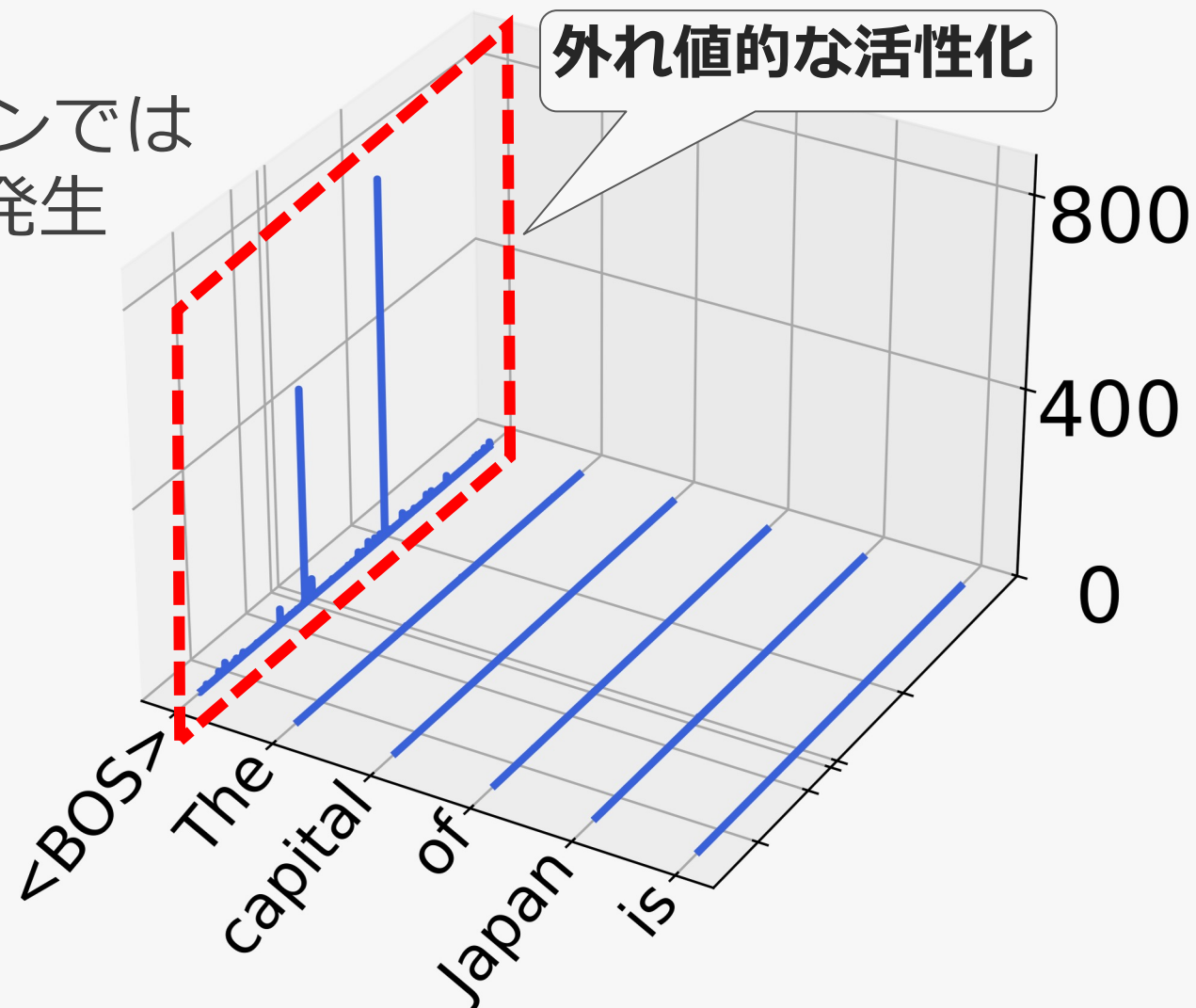
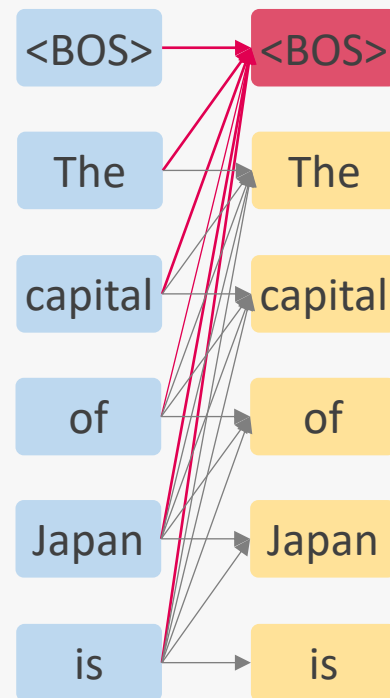
文脈情報を混ぜ
込めていない? !

意味に関わらず集中...
解釈性に欠ける? !



巨大な活性化値 —Massive Activation—

- Attention Sink が起こるトークンでは **Massive Activation** [Sun+'24] も発生

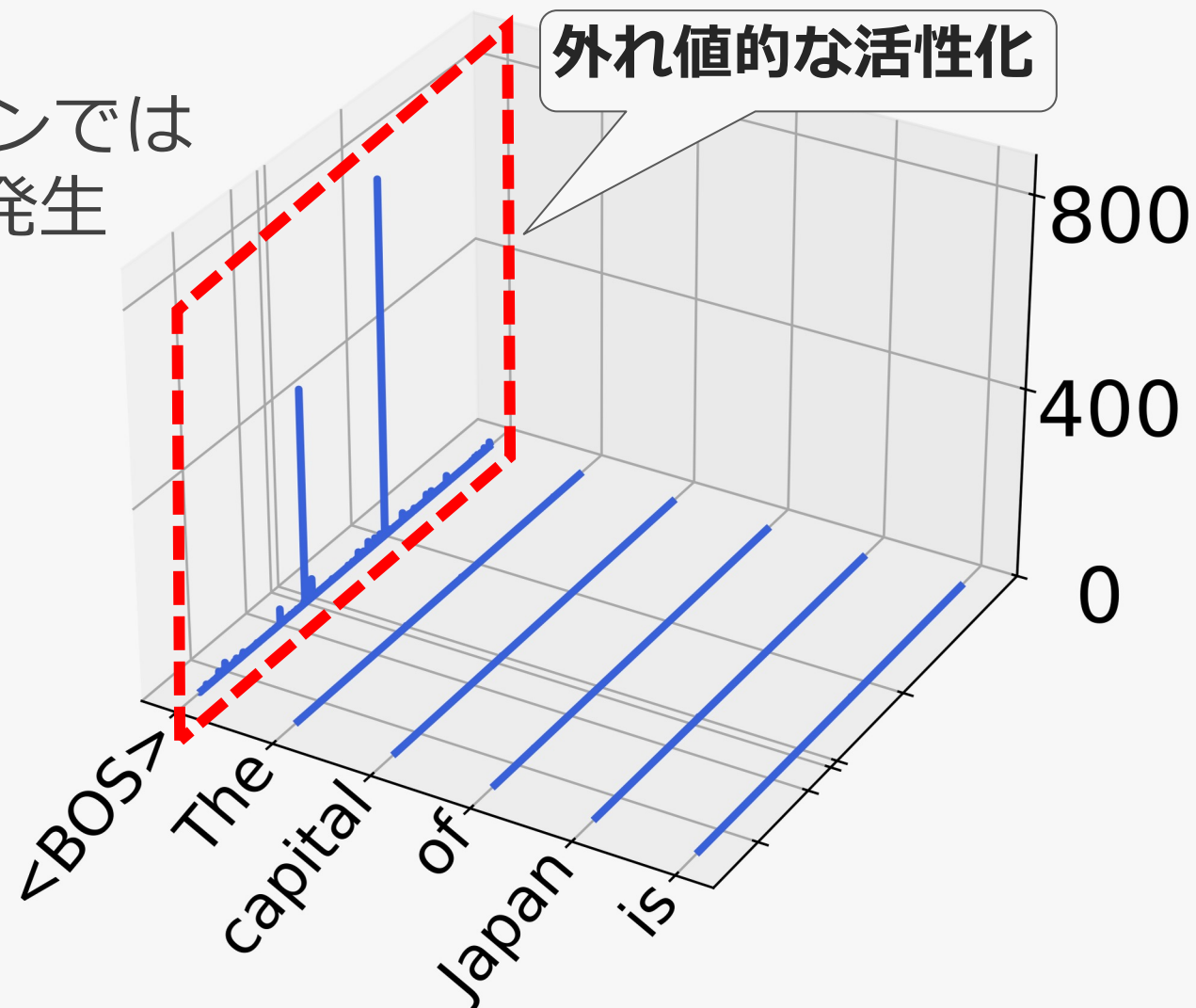


巨大な活性化値 —Massive Activation—

- Attention Sink が起こるトークンでは **Massive Activation** [Sun+'24] も発生

➡ 量子化の妨げ

工学的にも
嬉しくない



Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

位置符号化

自分自身への
Attention 集中

Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

位置符号化

自分自身への
Attention 集中

発生源は 先頭位置 or <BOS> ?

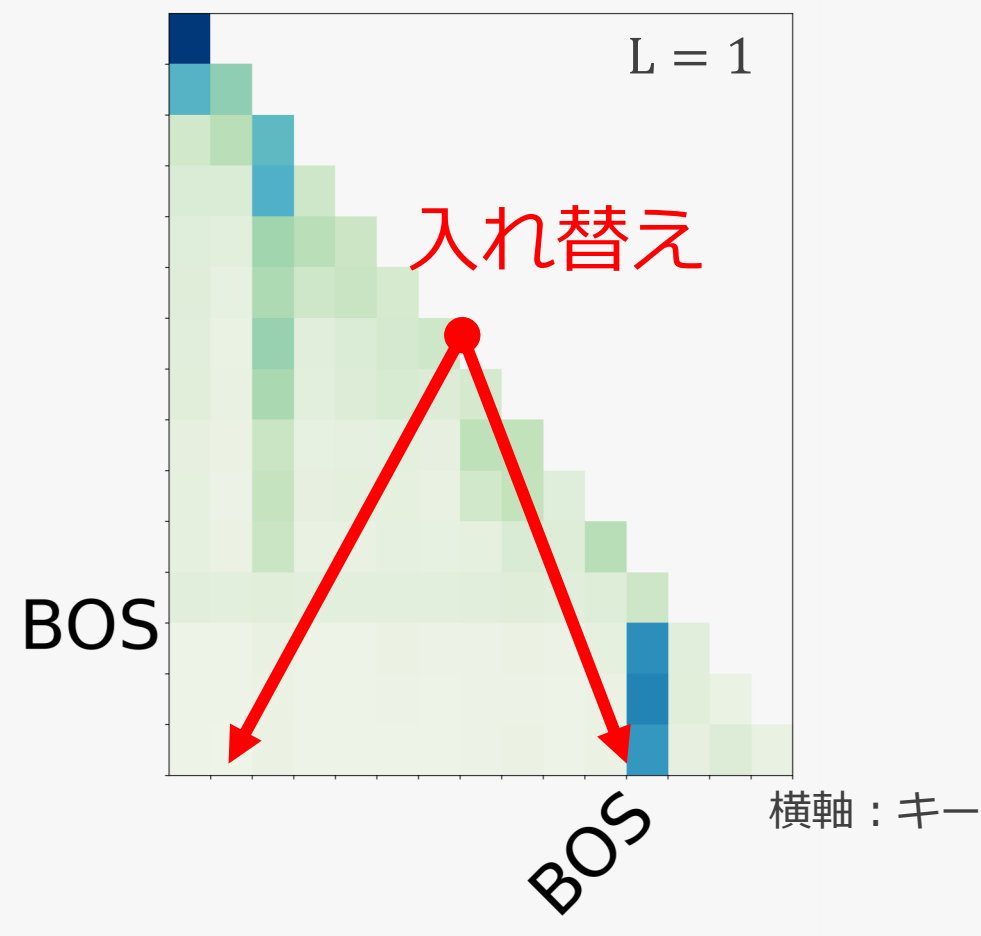
- 入力トークン列の <BOS> と文中のトークンを入れ替えて実験

例：

「<BOS> The capital of Japan is」

「of The capital <BOS> Japan is」

縦軸：クエリ

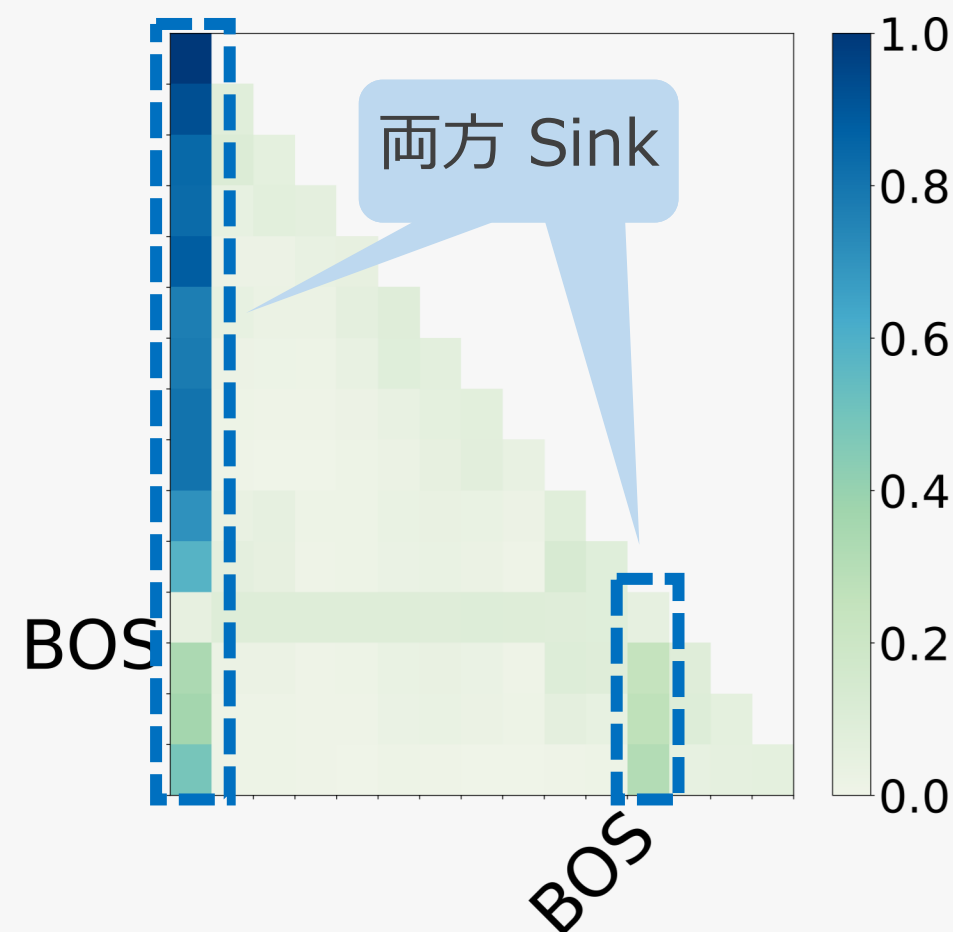


※Llama-3.2-3B を使用.

先頭位置・<BOS> どちらも発生源

- 先頭位置
➡ <BOS> でなくとも Sink が発生
- <BOS>
➡ 先頭位置でなくとも Sink が発生

それぞれが**単独で発生源となり得る**



※ 以降、特に断りがない限り図は第14層（中間層）について掲載。

Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

位置符号化

自分自身への
Attention 集中

- ✓ 先頭位置 + <BOS>
両方発生源

Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

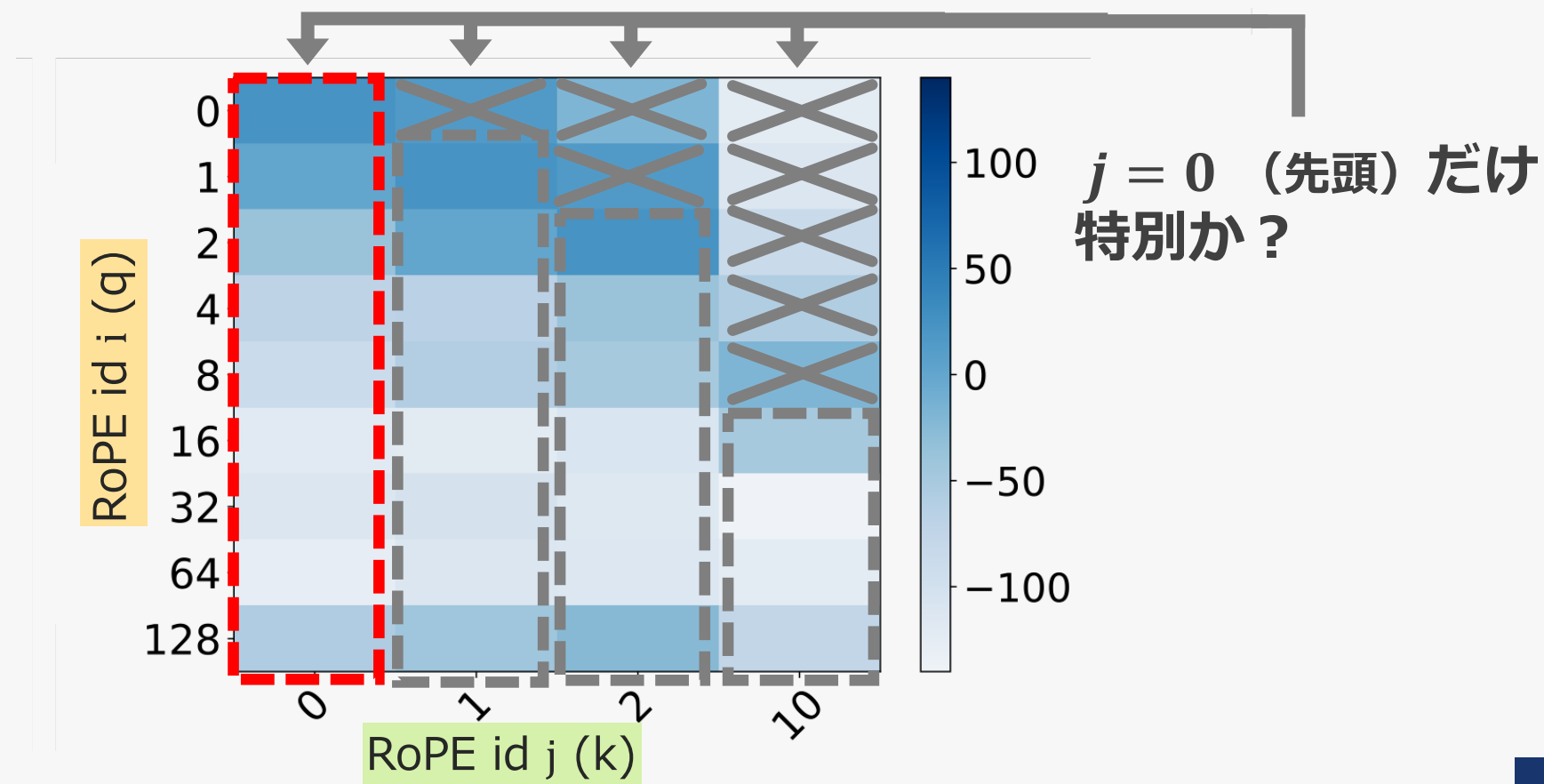
位置符号化

自分自身への
Attention 集中

- ✓ 先頭位置 + <BOS>
両方発生源

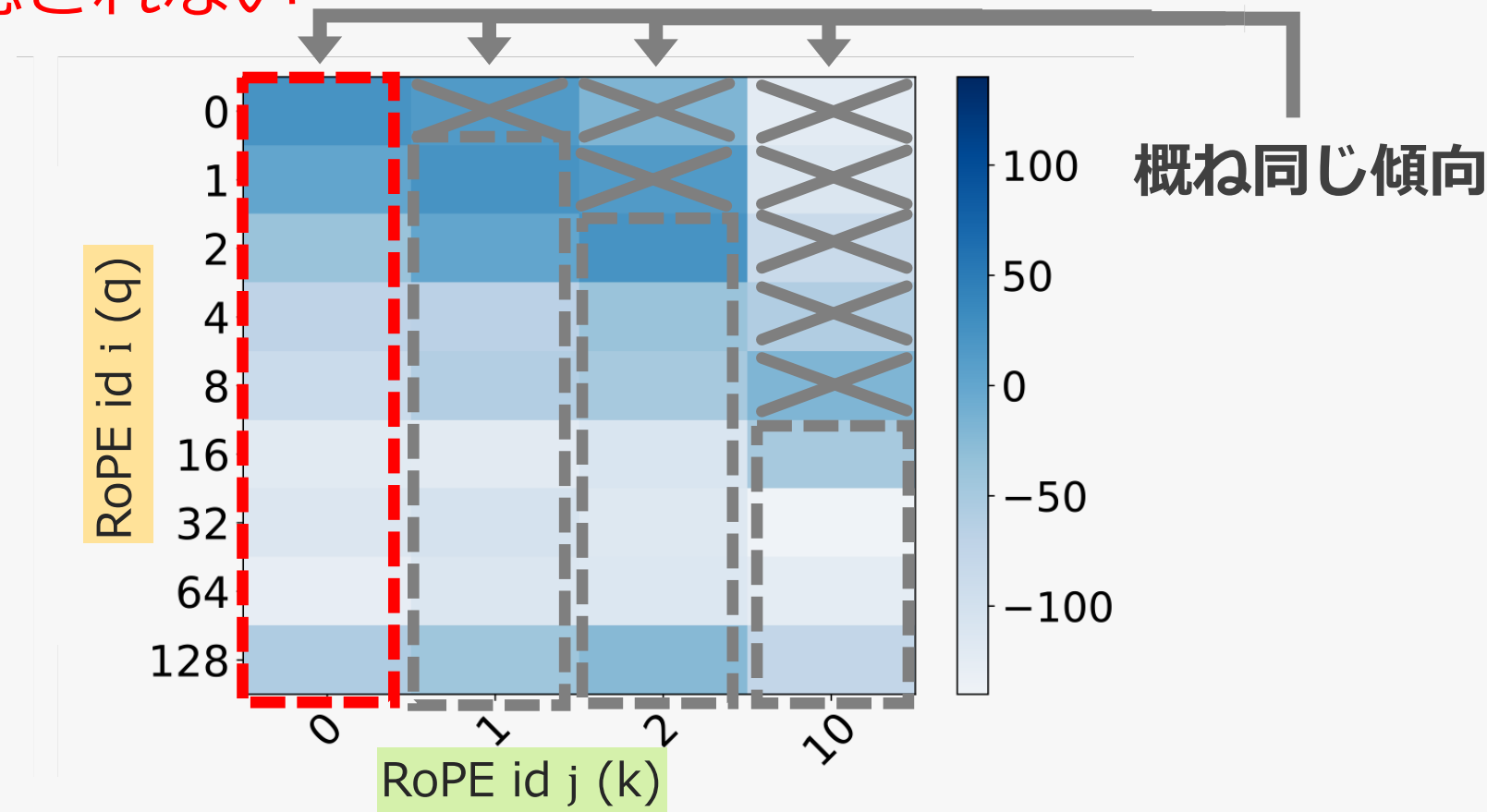
位置符号化は影響するか？

- 位置インデックス (i, j) を変化させて平均的なクエリ・キーの内積を分析



位置符号化による Sink への寄与は見られない

- $j = 0$ (キーの先頭) と他のインデックスの特徴に
有意な差は確認されない



Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

✓ 先頭位置 + <BOS>
両方発生源

位置符号化

✗ 位置符号化 (RoPE) の
影響は確認されず

自分自身への
Attention 集中

Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

✓ 先頭位置 + <BOS>
両方発生源

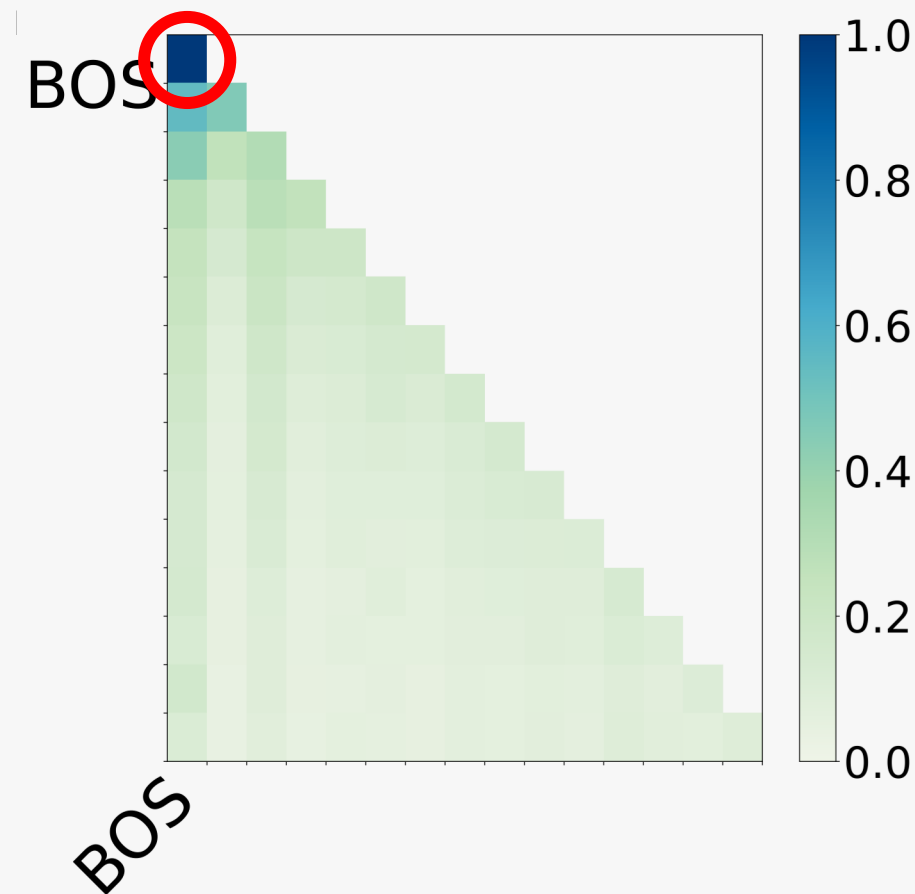
位置符号化

✗ 位置符号化 (RoPE) の
影響は確認されず

**自分自身への
Attention 集中**

先頭特有の性質「自分自身への Attention 集中」

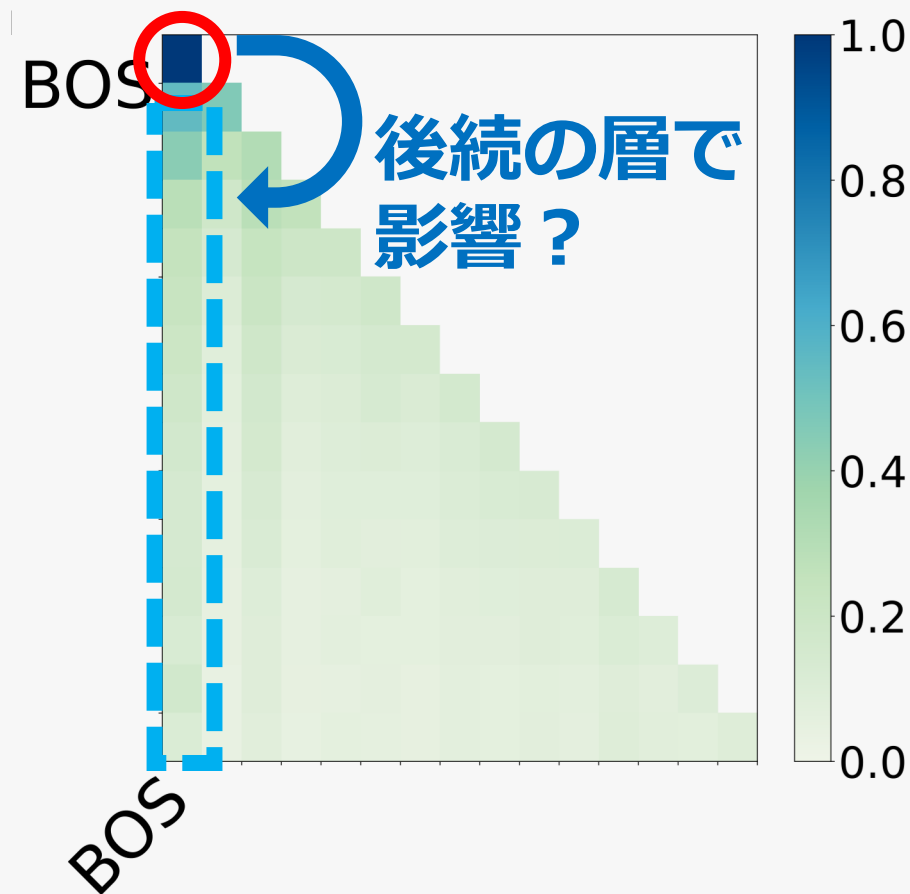
- 先頭トークンのクエリは自身のキーに対して **Attention スコア = 1**（全集中）を割り当てる



※上図：Llama-2-7B-hf 第1層（例示用） / 実験：Llama-3.2-3B

先頭特有の性質「自分自身への Attention 集中」

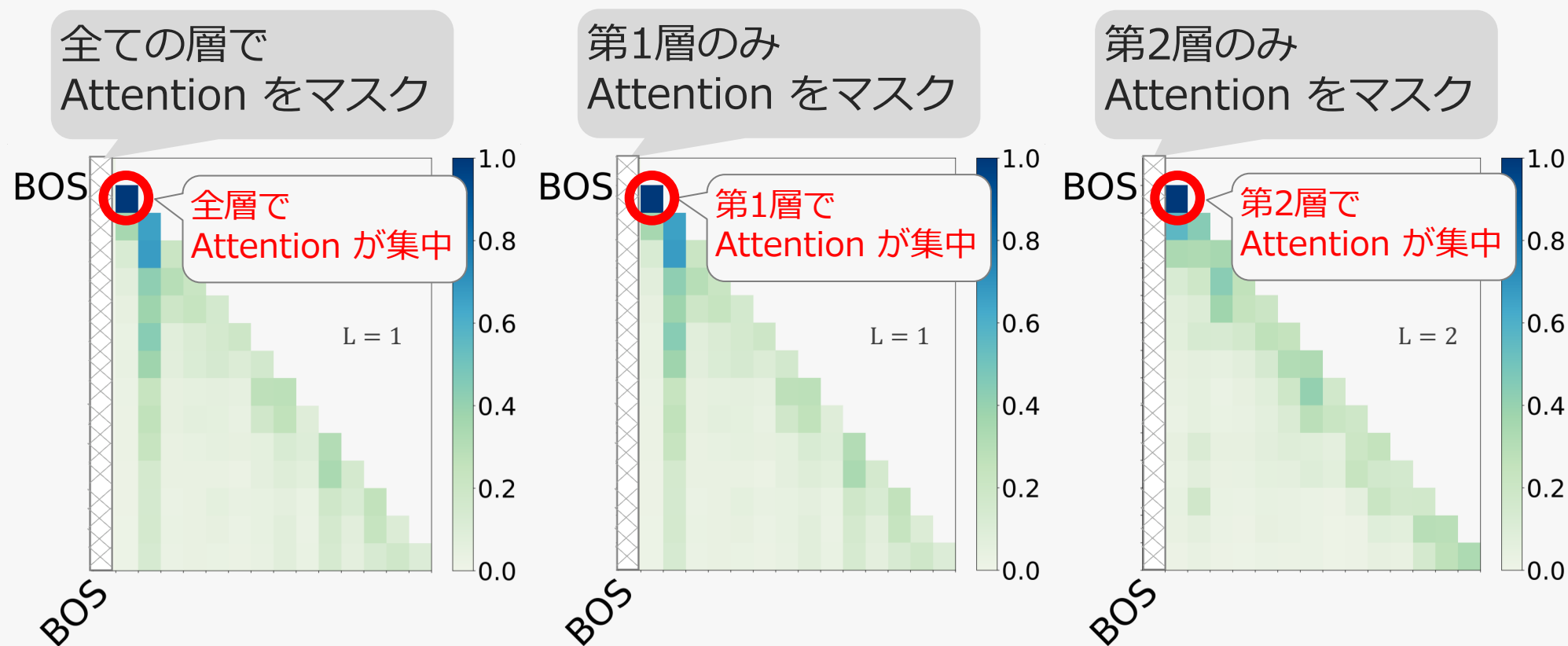
- 仮説：
先頭クエリの
「自分自身への Attention 集中」が
他のクエリの Attention スコアにも
影響を与えている
→ **Attention Sink の発生源と**
なっている



※上図：Llama-2-7B-hf 第1層（例示用） / 実験：Llama-3.2-3B

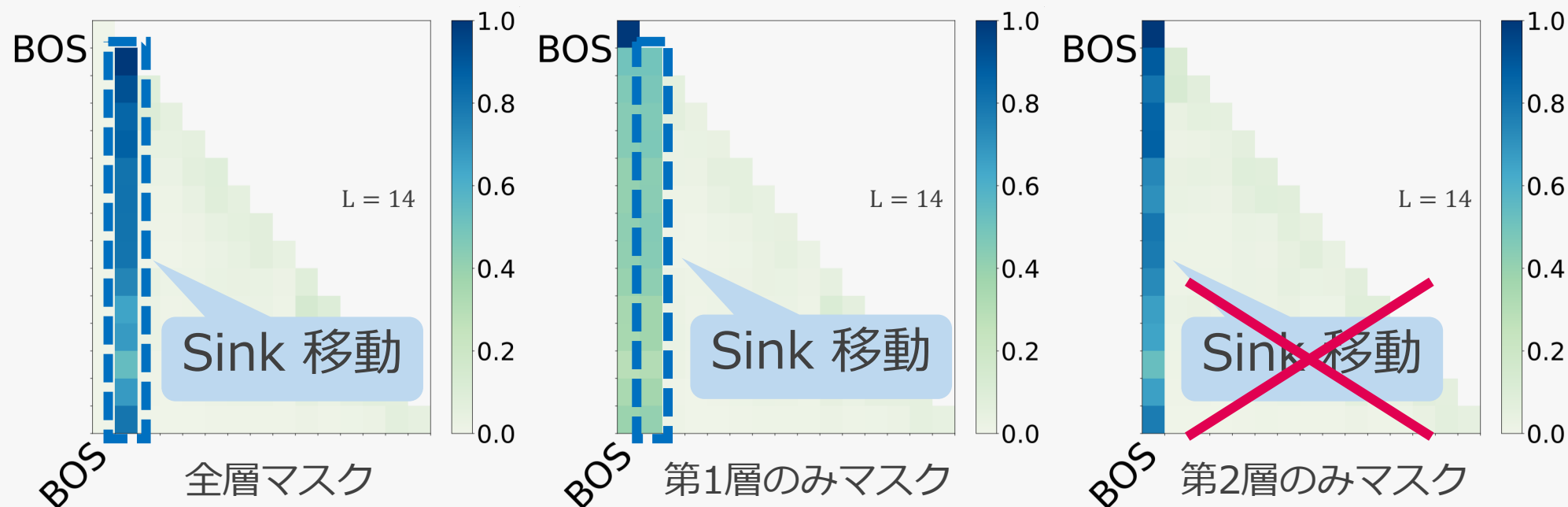
先頭の Attention をマスク

- 先頭の Attention をマスクすることで
2番目のトークンで「自分自身への Attention 集中」が起こる



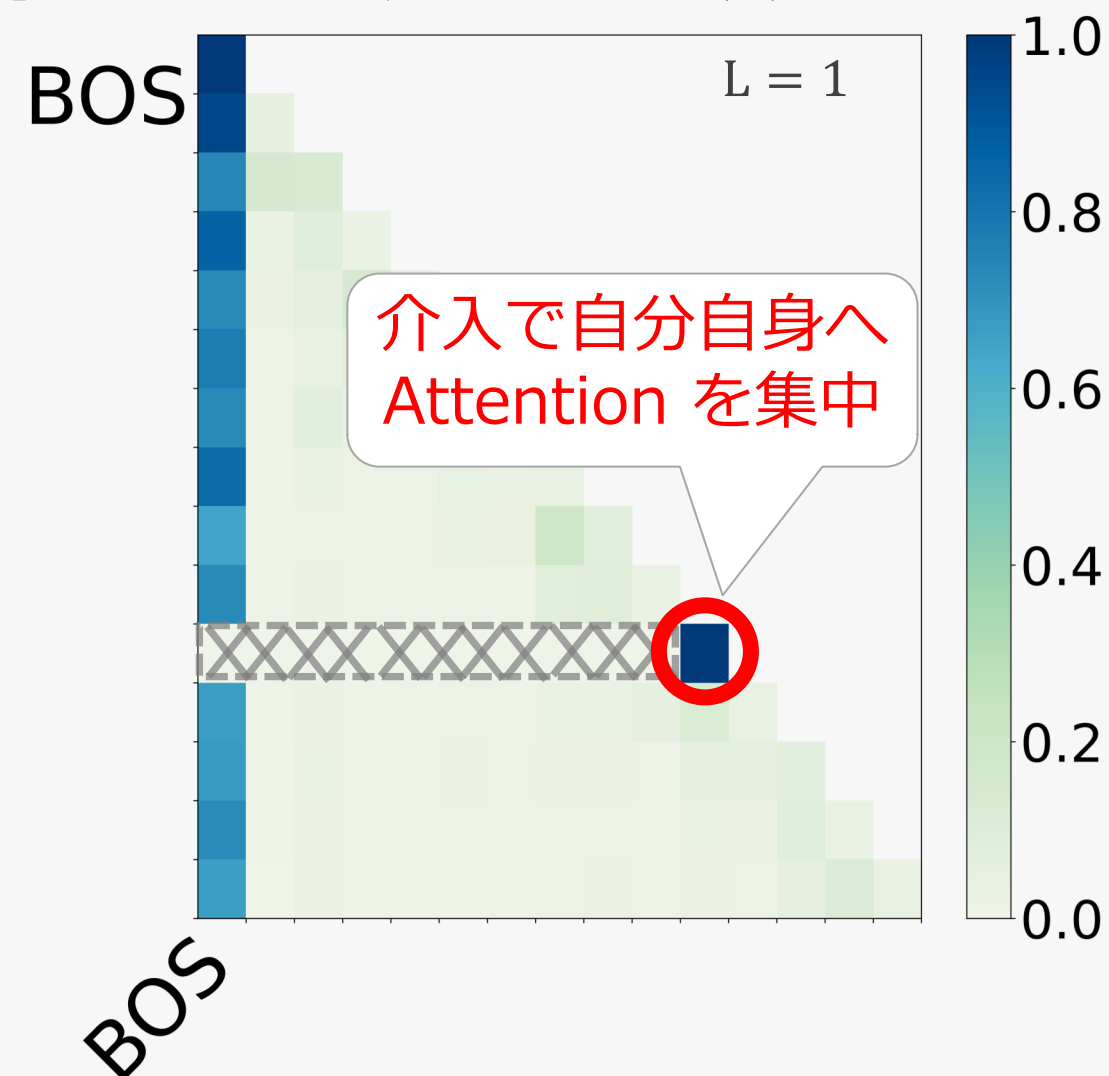
介入により Attention Sink の位置が移動

- 全層 / 第1層のマスキで Sink が移動
→ 第1層で「自分自身へ Attention を集中」していることが Attention Sink の発生源である可能性を示唆



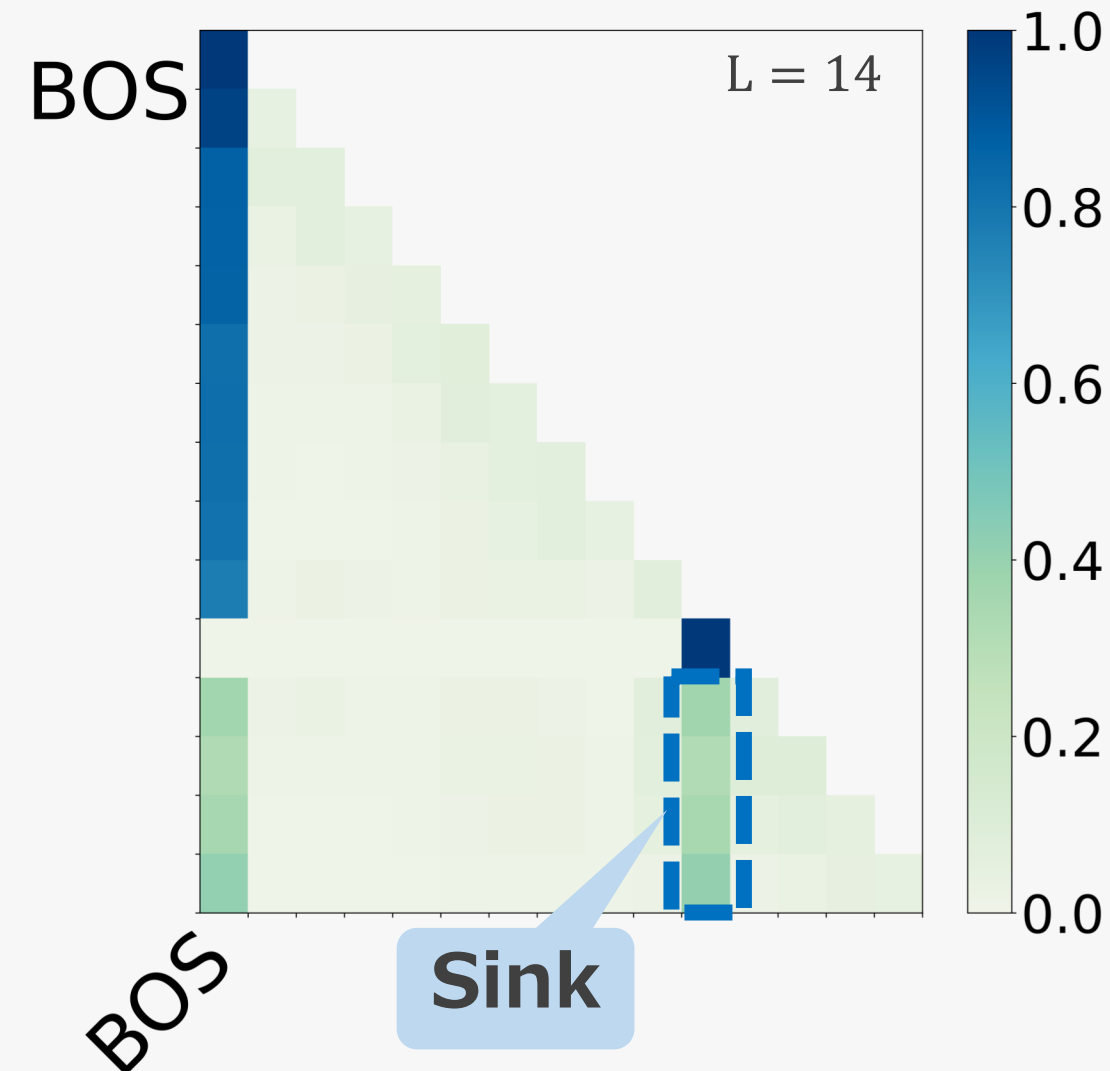
自分自身に Attention を集中させる介入実験

- 文中クエリの Attention スコアに介入.
自分自身のキーへの「Attention スコア = 1」を強制.



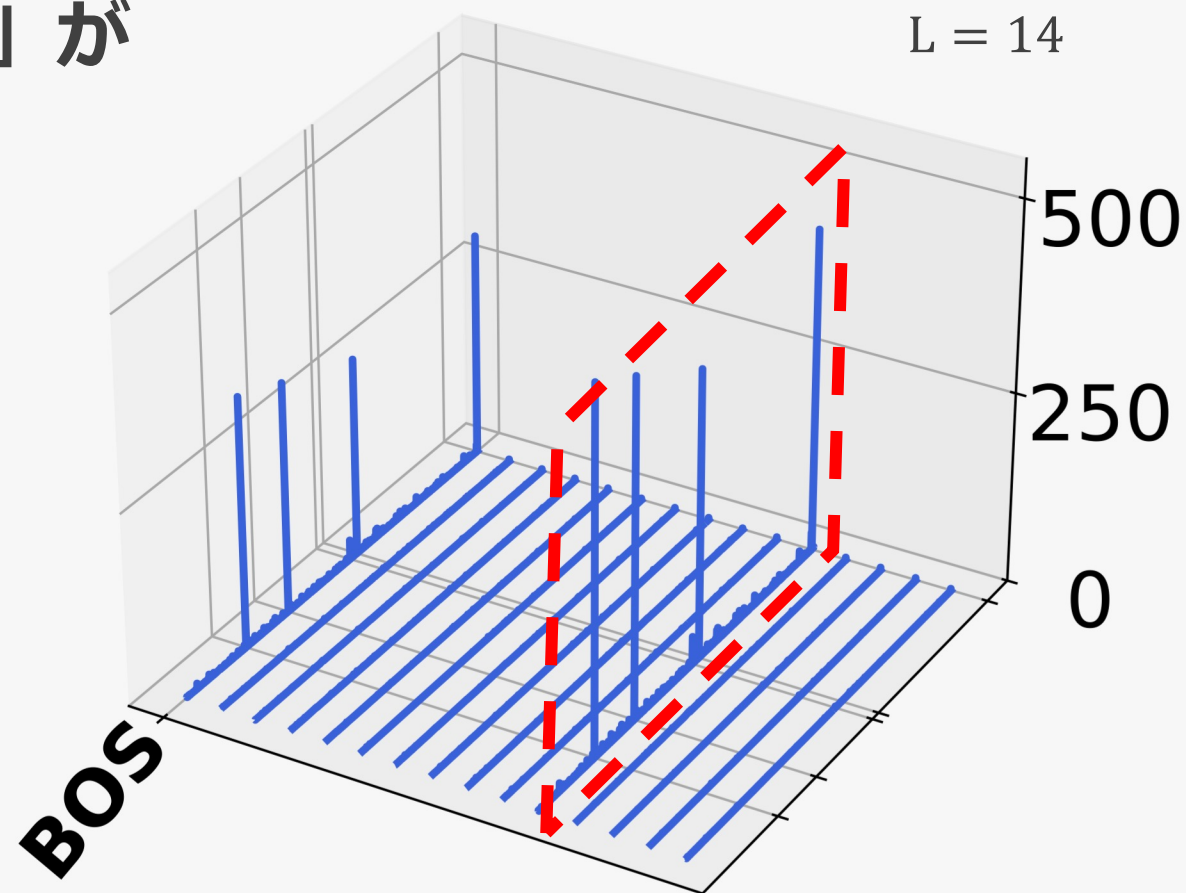
介入トークンにも Attention Sink が発生

- 「自分自身への Attention 集中」が Attention Sink の発生源である可能性を強く示唆



Massive Activation も介入トークンで発生

- 「自分自身への Attention 集中」が Attention Sink だけでなく, **Massive Activation の発生源** でもある可能性を強く示唆



Attention Sink の発生源は？

- 本研究では、**先頭位置**で Attention Sink が起こる発生源を分解

先頭位置に関する性質

<BOS>

✓ 先頭位置 + <BOS>
両方発生源

位置符号化

✗ 位置符号化 (RoPE) の
影響は確認されず

自分自身への
Attention 集中

✓ Attention Sink および
Massive Activation の
発生源である可能性を
示唆

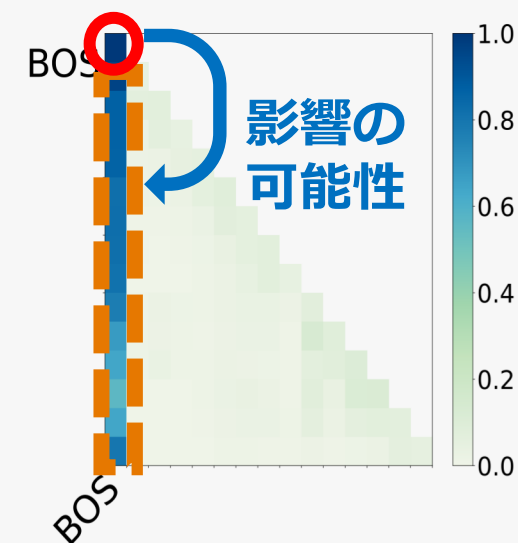
おわりに

まとめ

- Attention Sink, Massive Activation の発生源を分解
- 「自分自身への Attention 集中」が発生源である可能性を示唆

展望

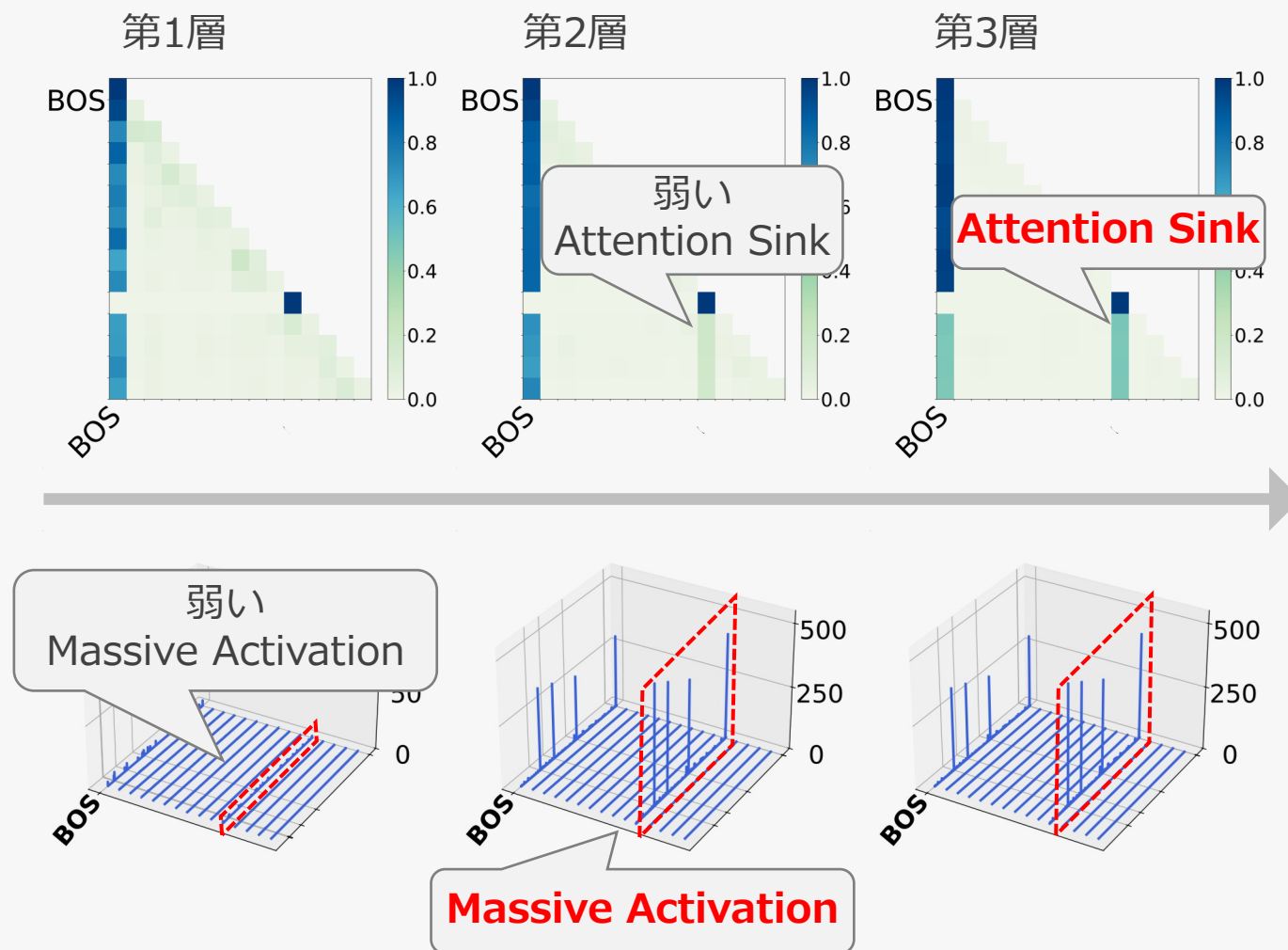
- 「自分自身への Attention 集中」で Sink が発生する具体的な内部機序を分析
- Attention Sink を考慮したモデル (e.g. gpt-oss, Qwen3-Next) でも検証



| 付録 / Appendix

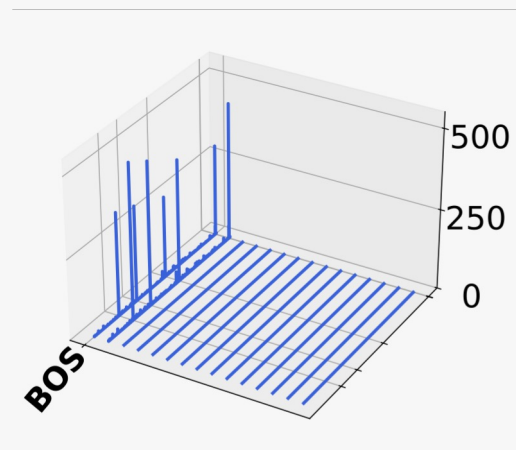
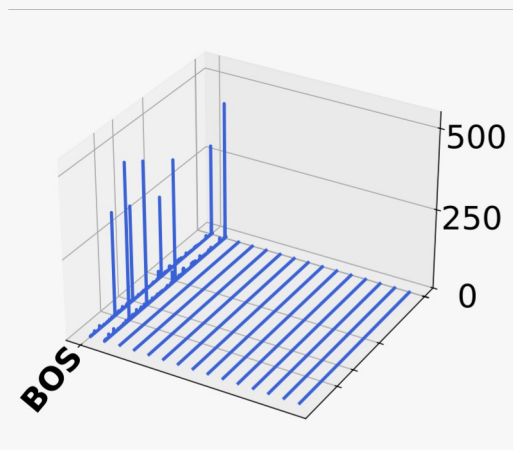
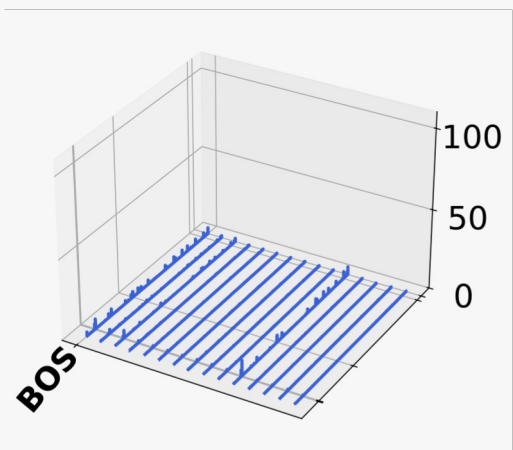
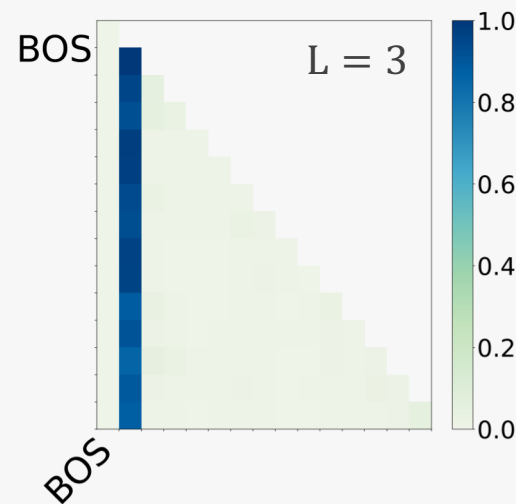
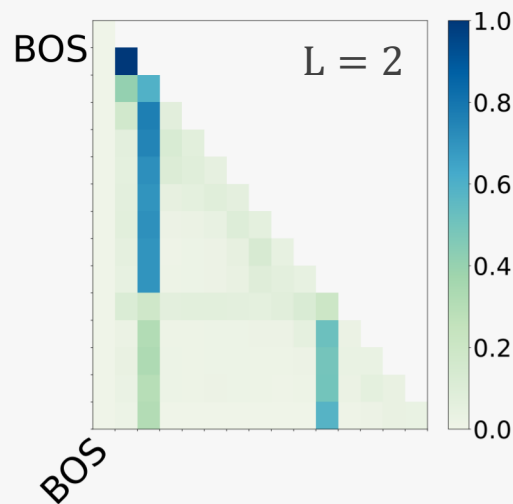
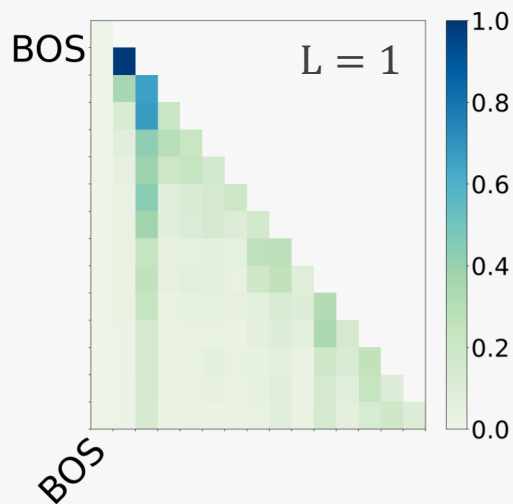
Attention Sink と Massive Activation の因果関係

- 発生順序
Massive Activation 発生
↓
Attention Sink 発生
- Massive Activation が
Attention Sink を
発生させている



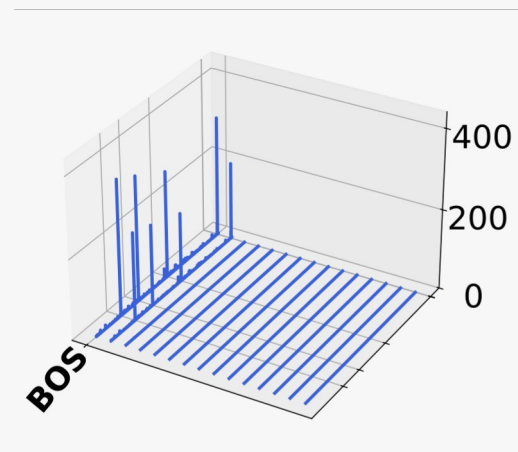
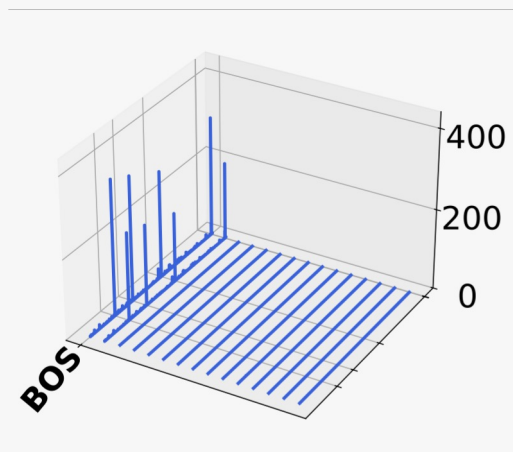
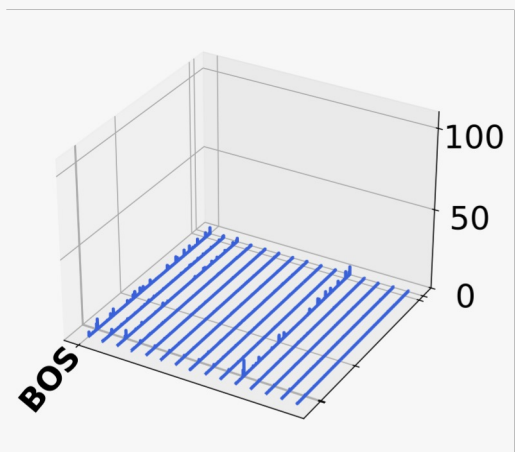
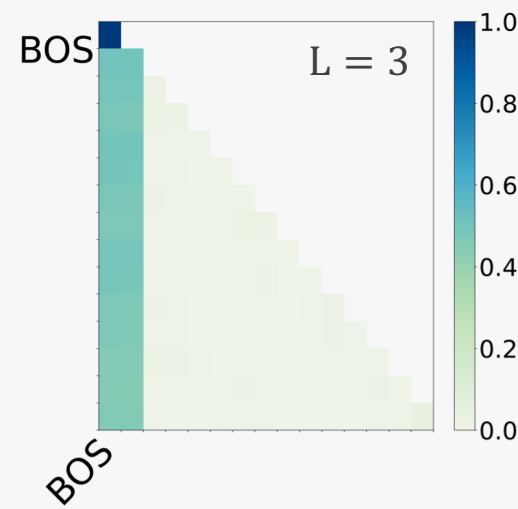
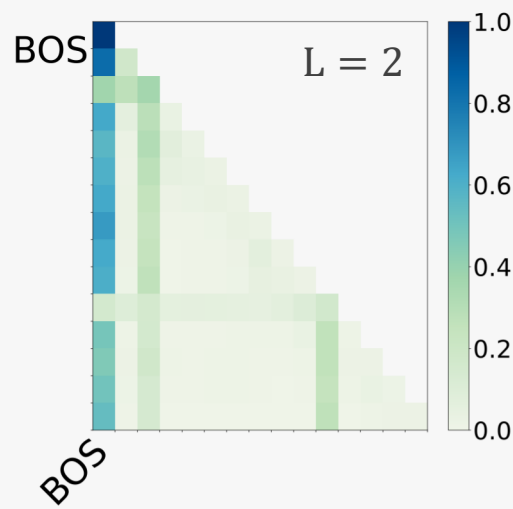
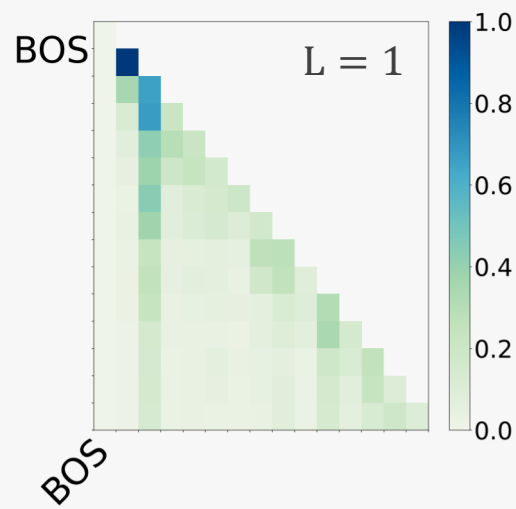
先頭 Attention マスクにおける Sink 形成過程

全ての層でマスクした時



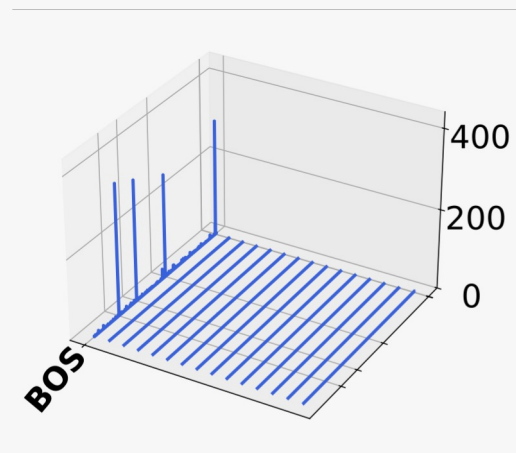
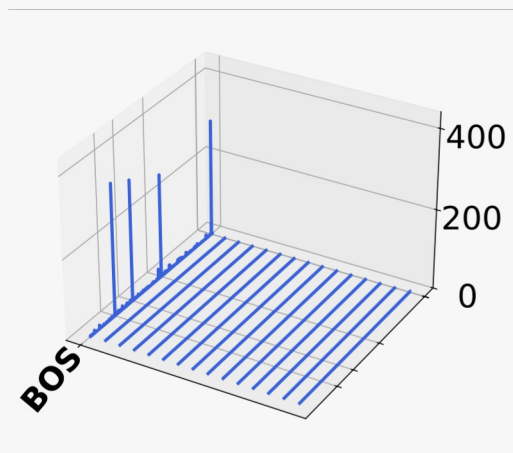
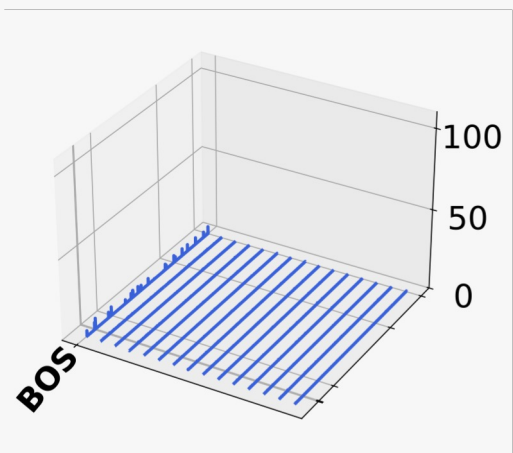
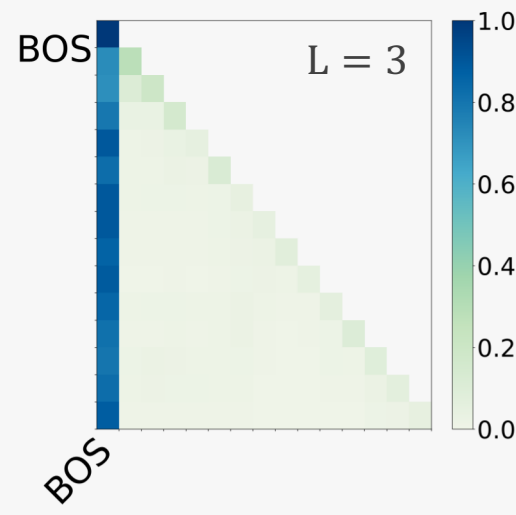
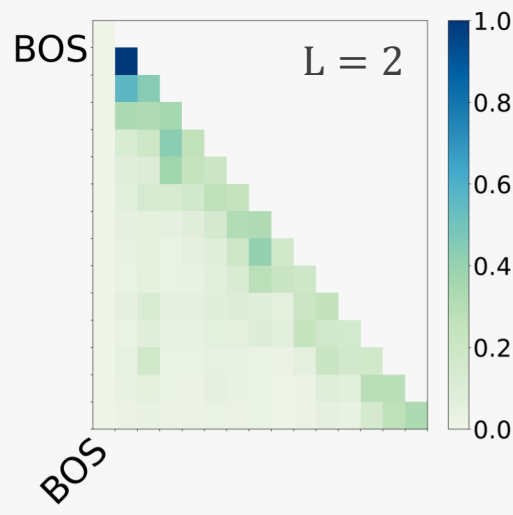
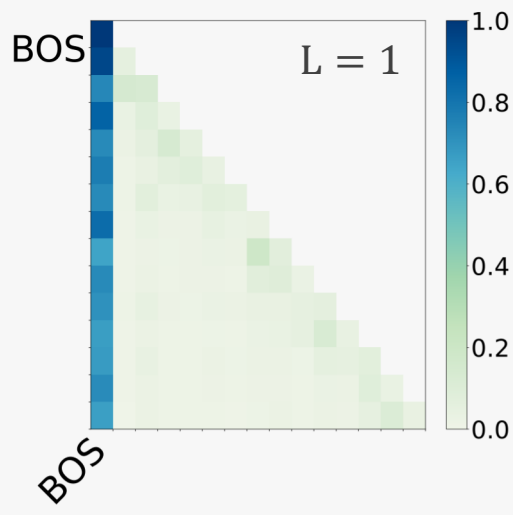
先頭 Attention マスクにおける Sink 形成過程

第1層でのみマスクした時



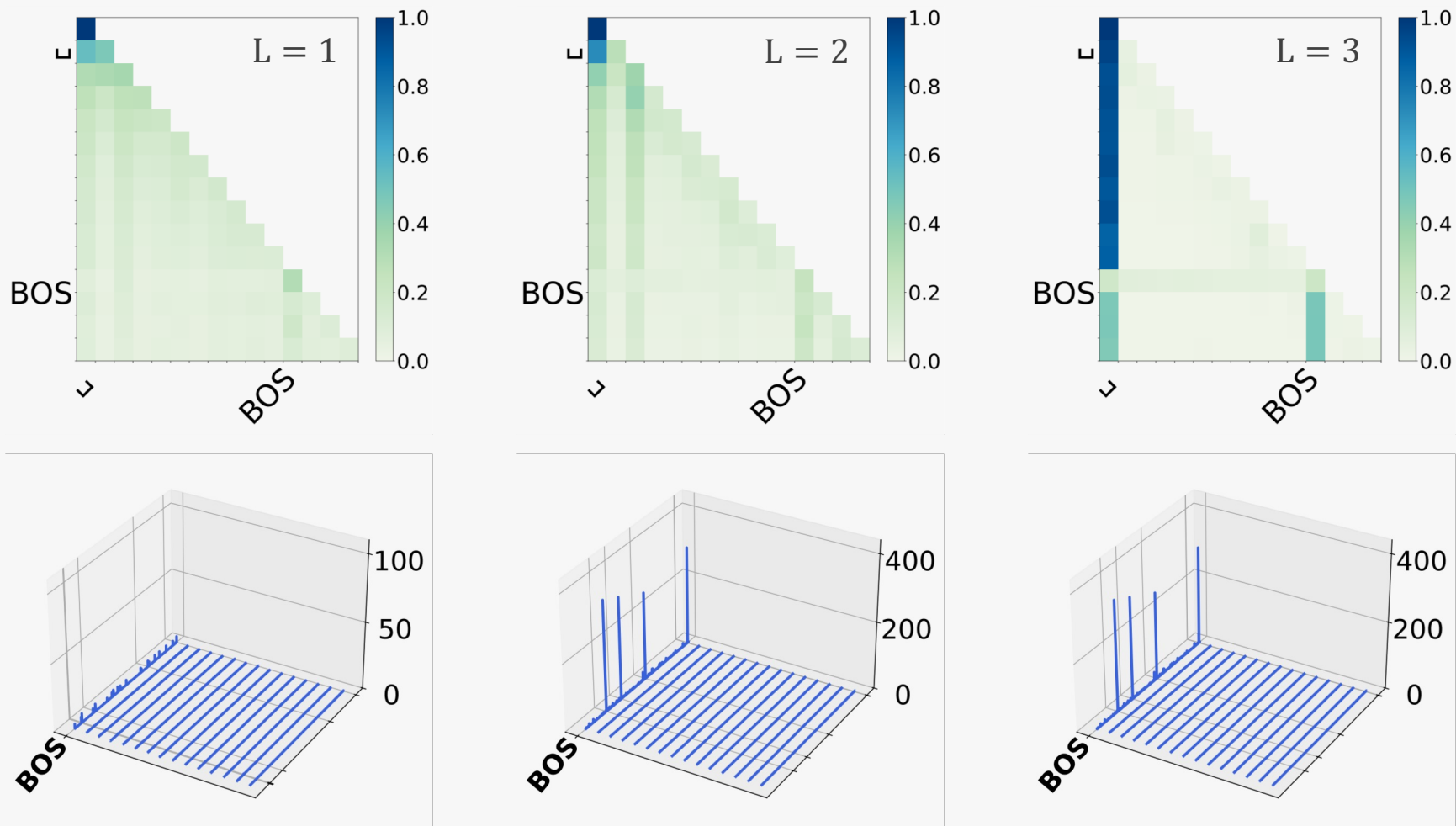
先頭 Attention マスクにおける Sink 形成過程

第2層でのみマスクした時



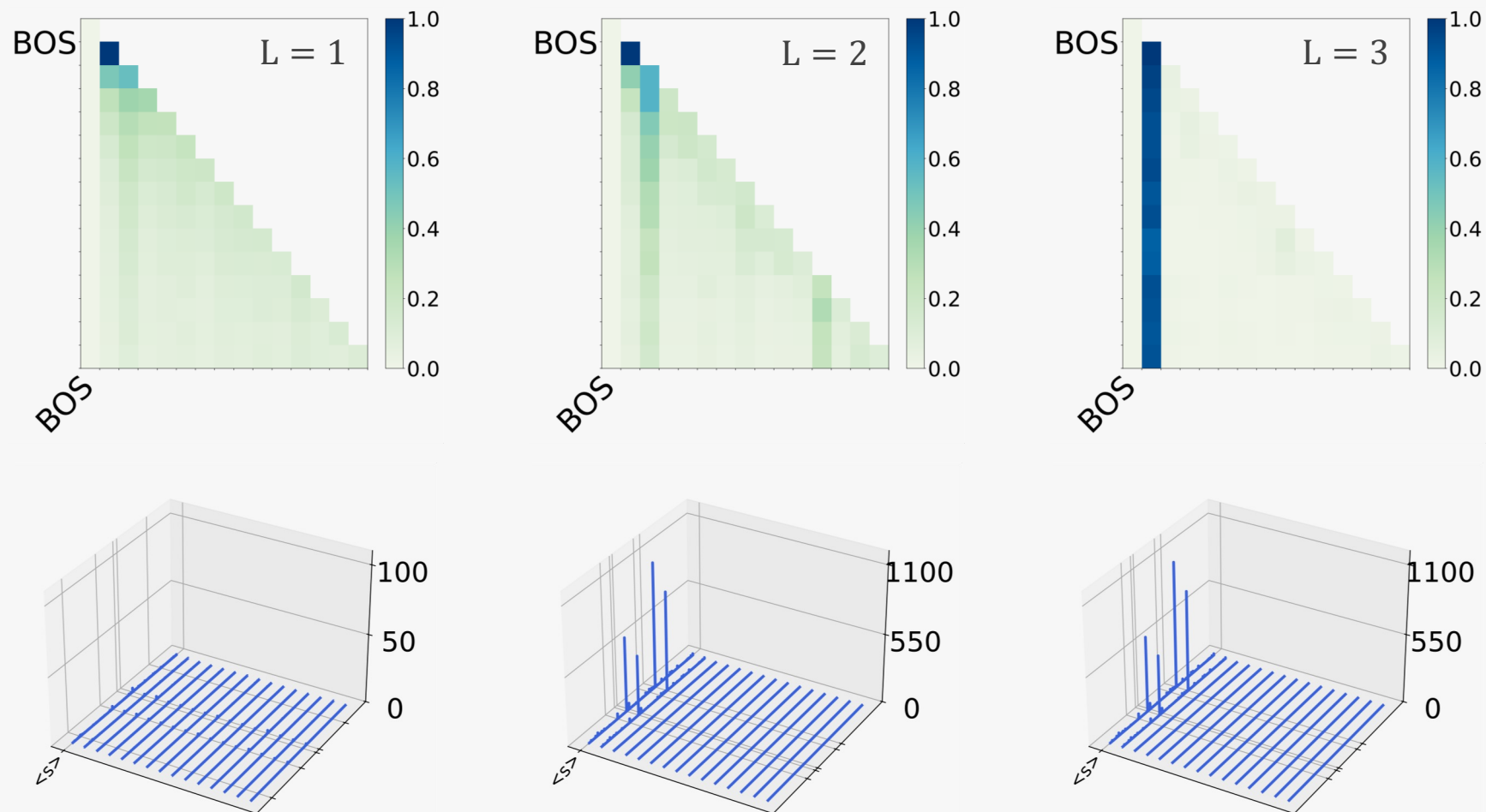
Llama-2-7B-hf における分析

<BOS> と文中のトークンを入れ替えた時



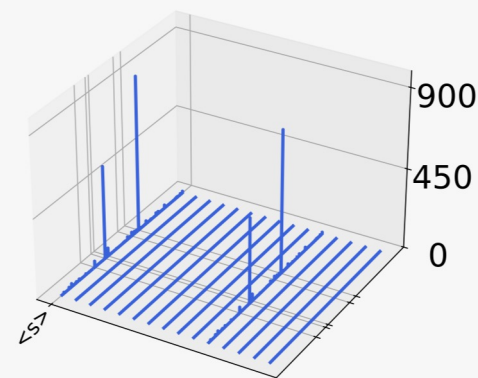
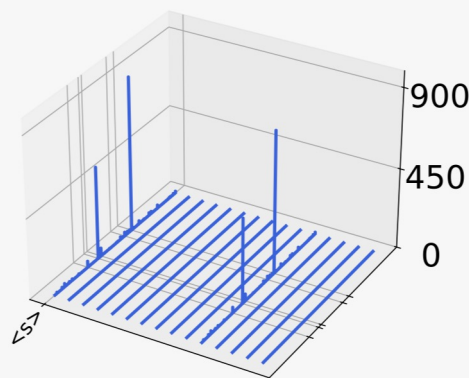
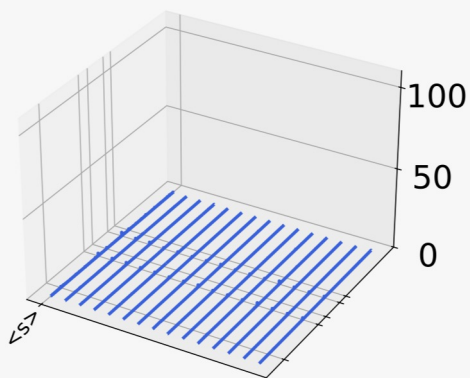
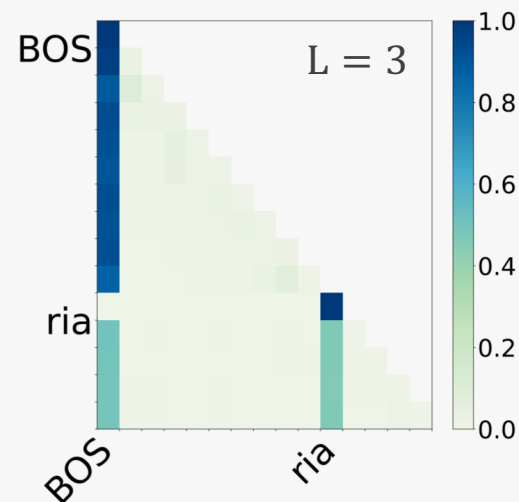
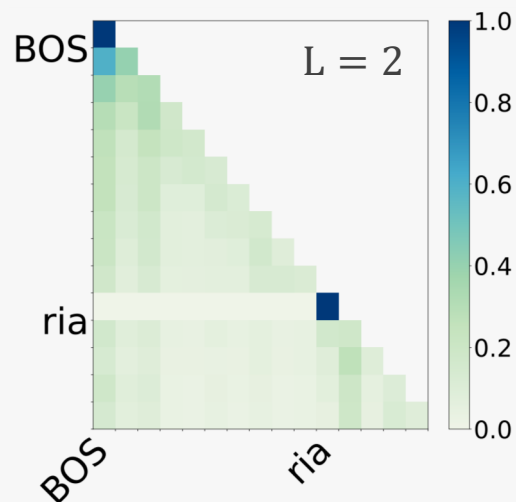
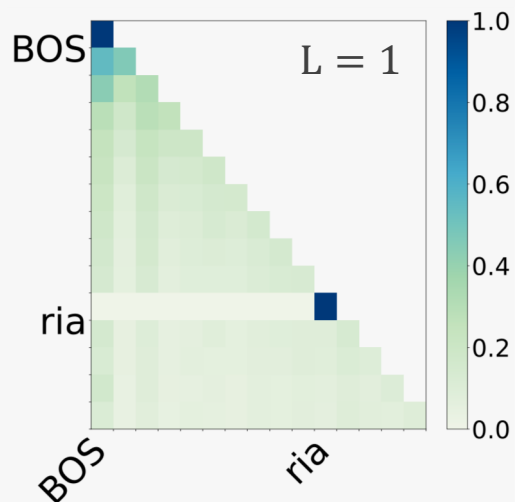
Llama-2-7B-hf における分析

先頭の Attention をマスクした時



Llama-2-7B-hf における分析

文中のトークンの Attention を自身に集中させた時



参考文献 / References

- [Xiao+'24] 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.
- [Sun+'24] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In First Conference on Language Modeling (COLM), 2024.