

概要

- Attention Sink の発生源は **先頭位置** か **<BOS>** か？
✓ **先頭位置 + <BOS>**
- 位置符号化 (RoPE) が原因か？
✗ **RoPE の影響は確認されない**
- なぜ先頭位置で Sink するか？
✓ **「自身への注意集中」が発生源**である可能性を示唆

先頭の性質

- <BOS>
- 位置インデックス $j = 0$
- 「自身へ注意集中する瞬間」がある

背景・目的

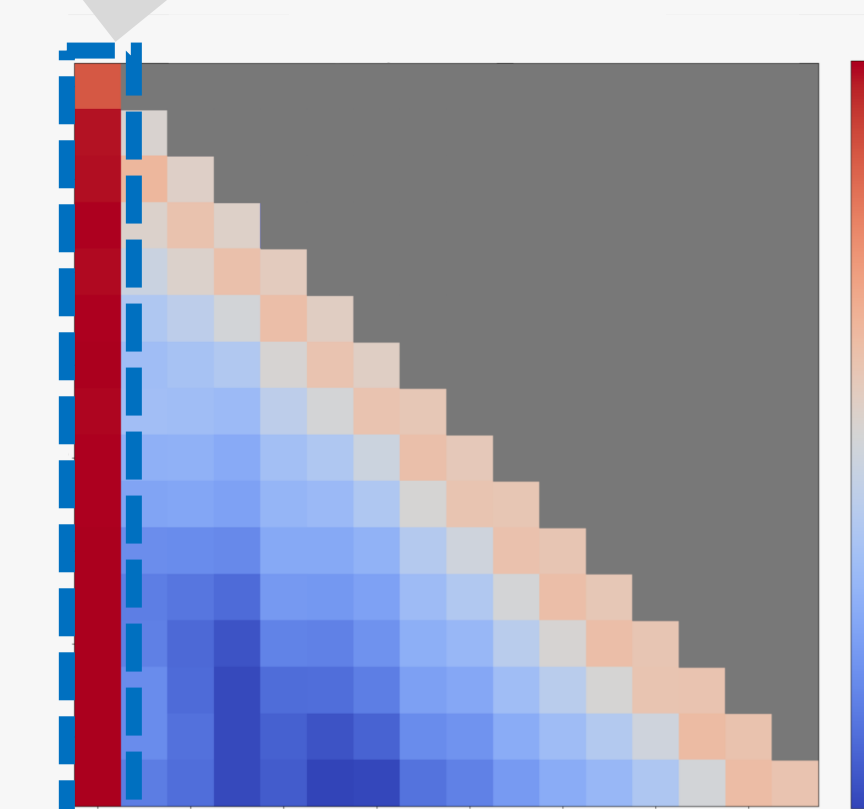
LLM は先頭 (等) へ
不自然に注意を集中

→ **Attention Sink** (Sink)

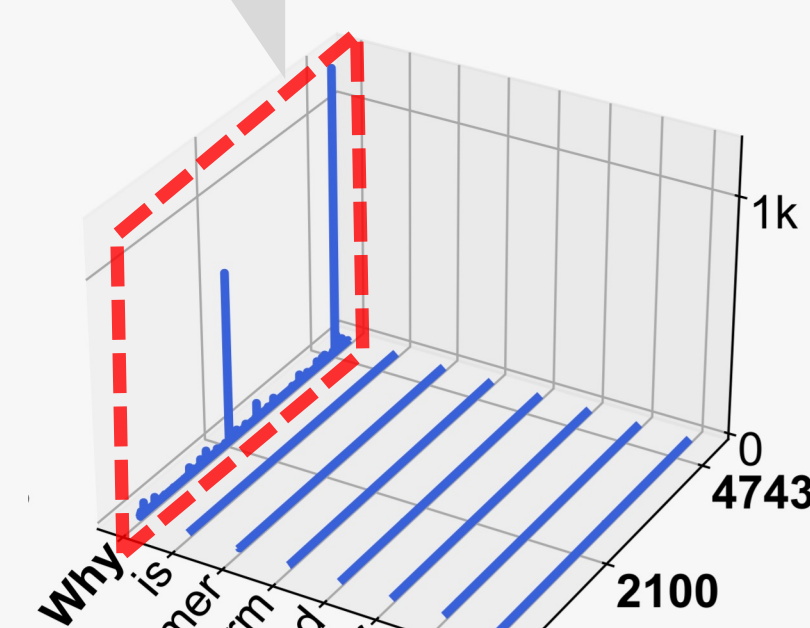
Sink 発生トークンに
Massive Activation

→ 量子化の妨げ

- 発生源の特定
- 将来的な
 - 解釈性向上
 - 量子化適応性への貢献



[Xiao+'24]

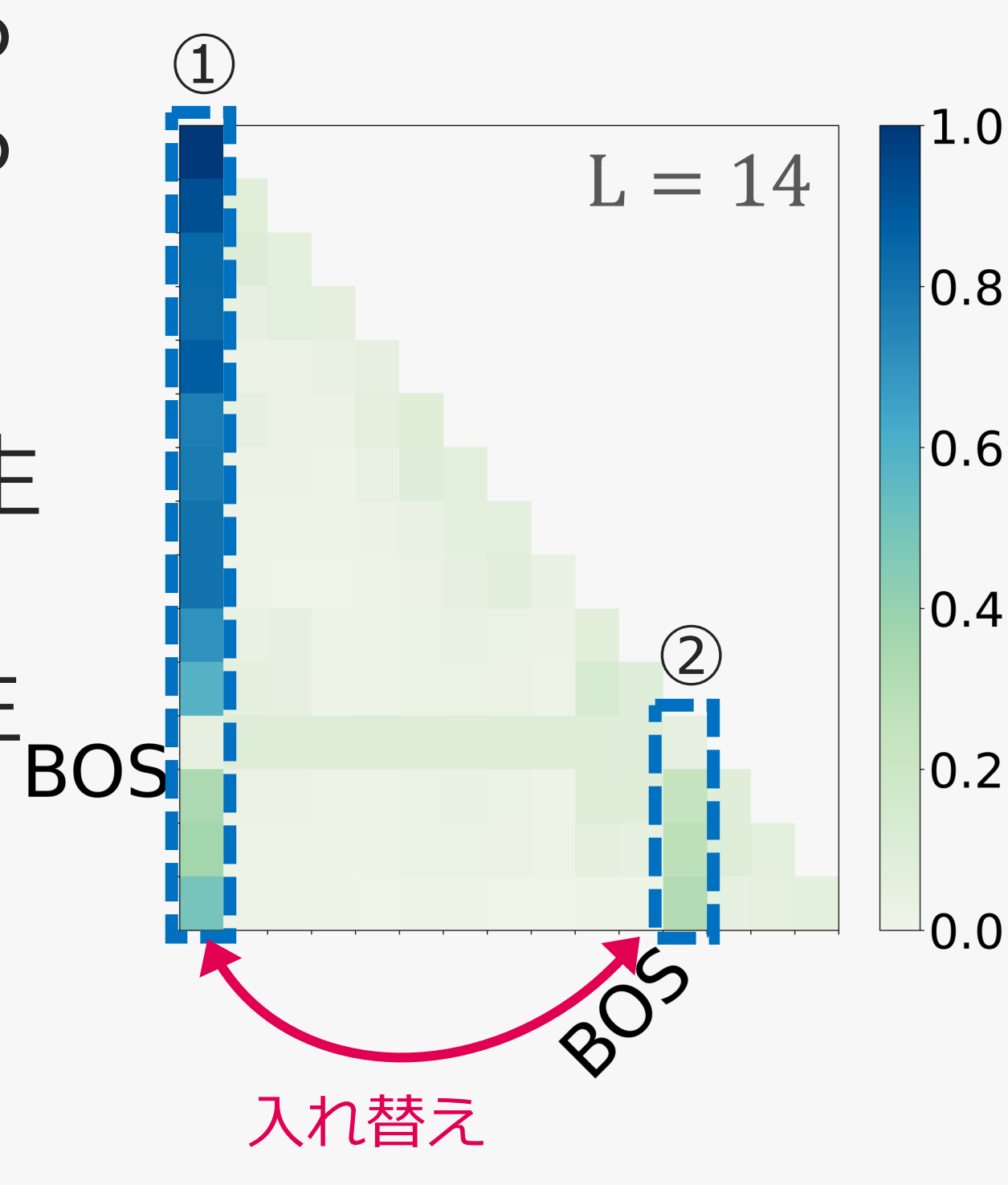


[Sun+'24]

1. 発生源は 先頭位置 or <BOS> ?

実験：入力トークン列の <BOS> と
文中のトークンを入れ替える
結果：それぞれ**単独で発生源**となる

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

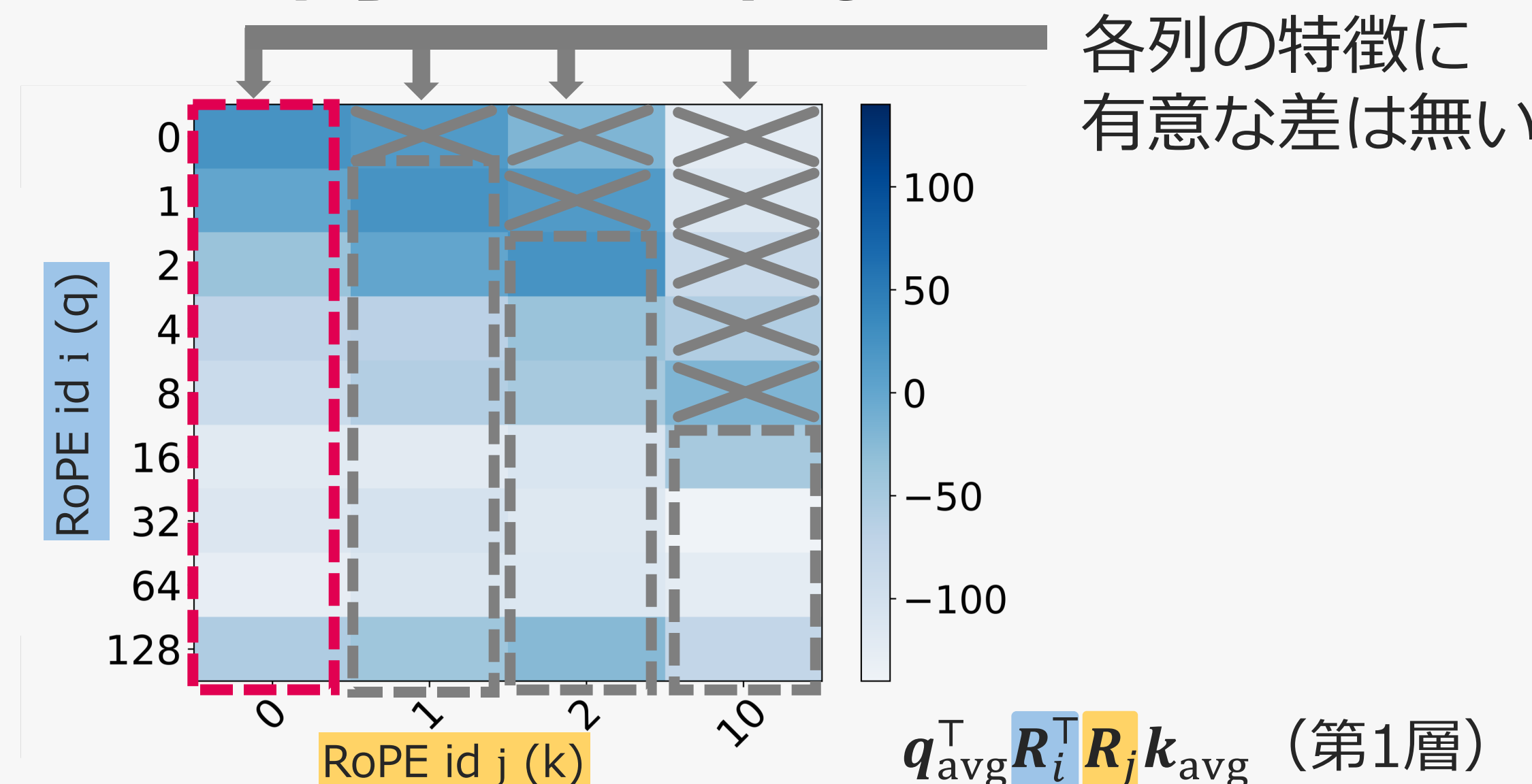


※Llama-3.2-3B (全28層) を使用.
注意ヒートマップはいずれも第14層 (中間層)

2. 位置符号化は発生に寄与するか？

RoPE の位置インデックスを変化させながら
平均的なクエリ・キーの内積を観察

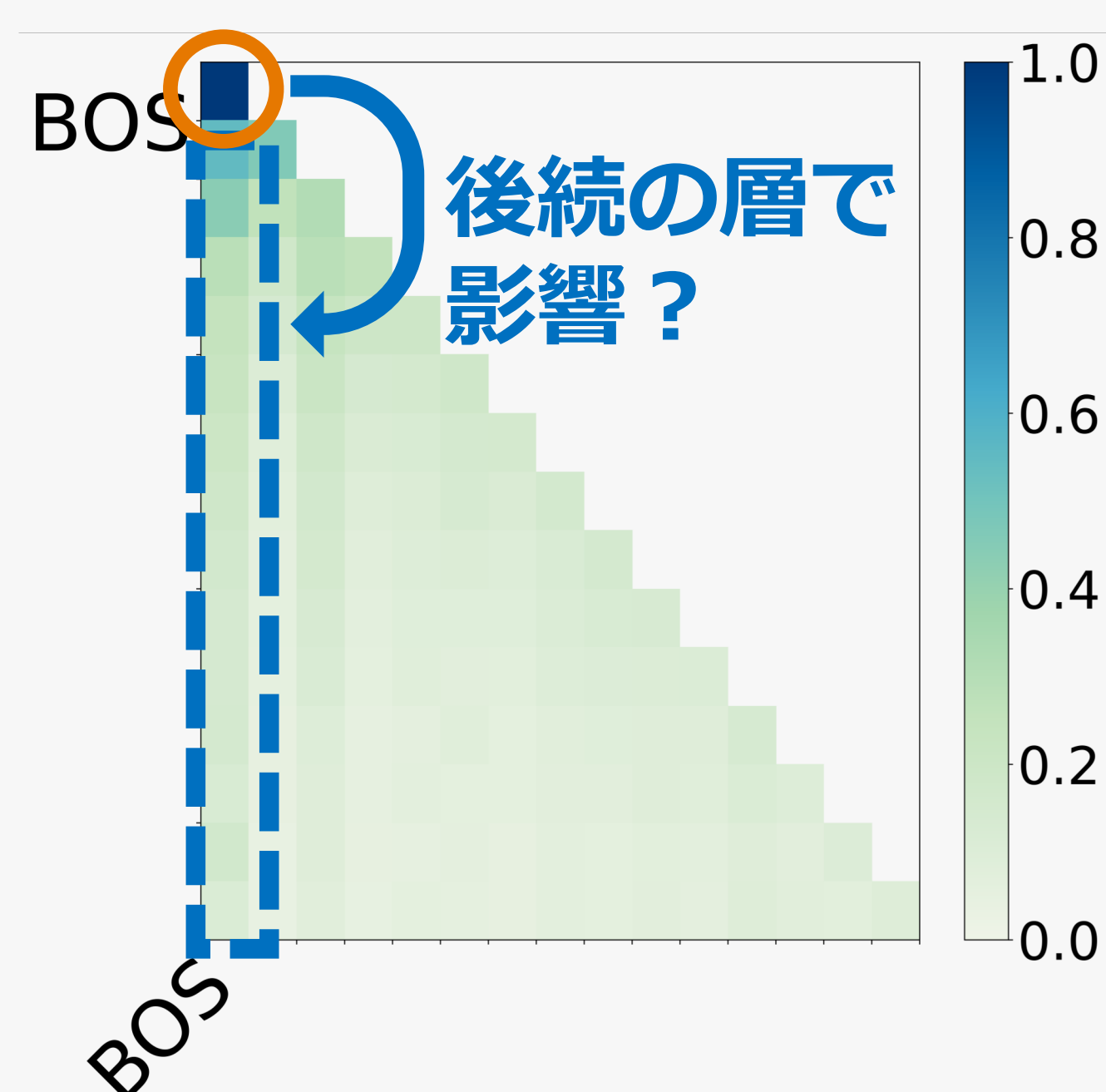
→ ✗ $j = 0$ (キーの先頭) という位置インデックス自体に
Sink 発生への寄与は確認されない



3. (RoPE の影響はないのに) なぜ先頭は <BOS> 以外も Sink するか？

仮説

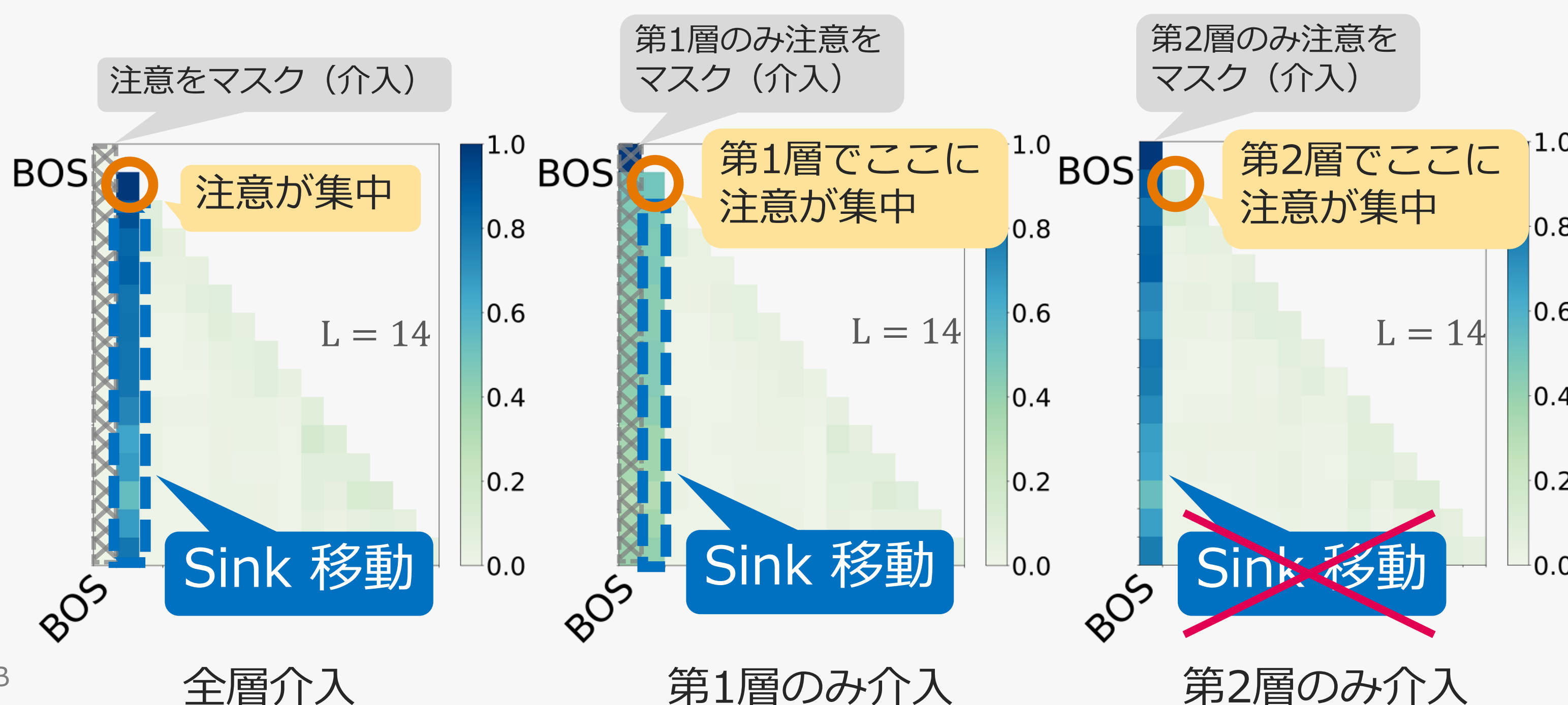
「自身へ注意を
集中する瞬間」があるから



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

先頭への注意介入実験

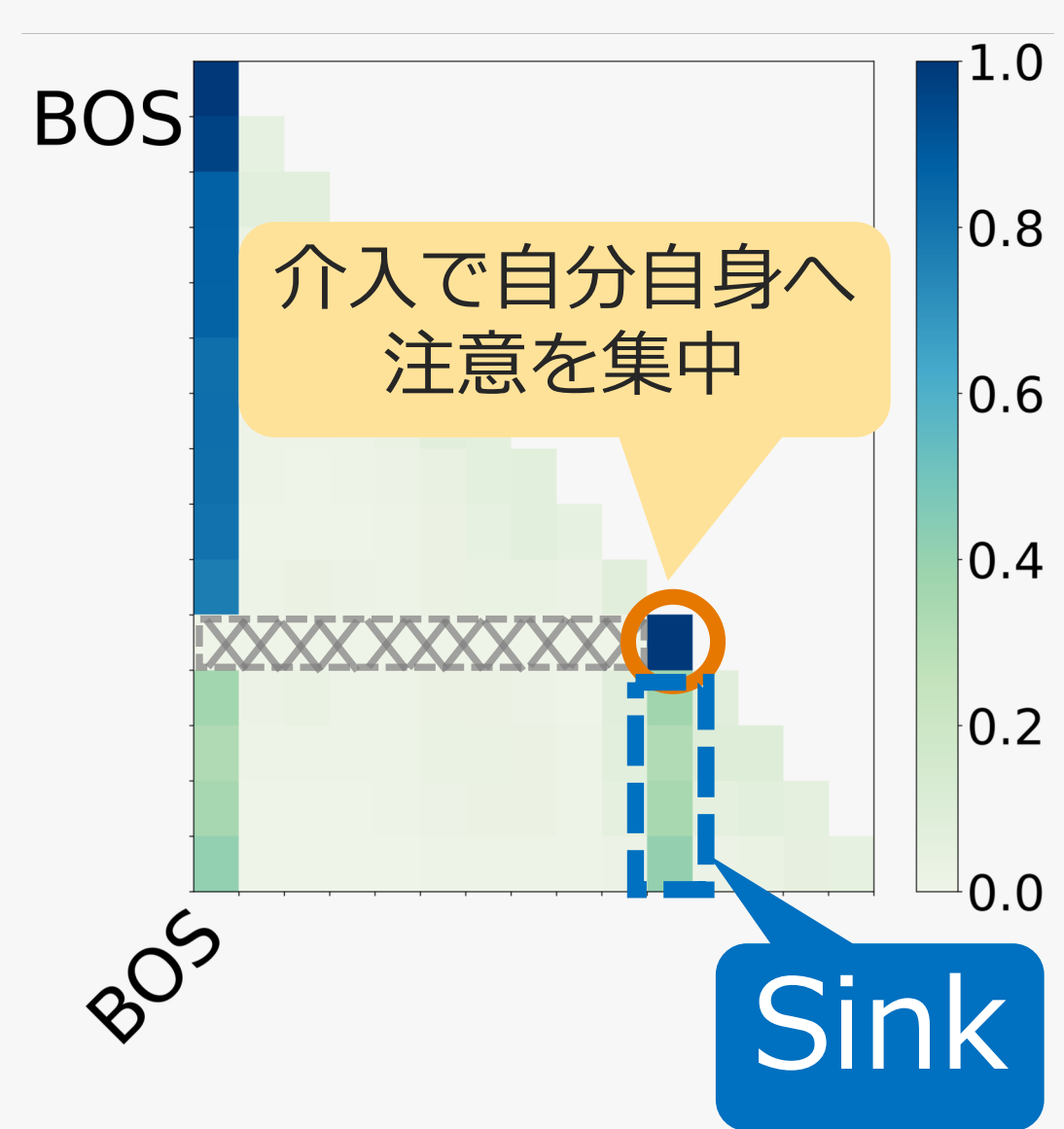
全層 / 第1層介入で Sink の位置が移動
→ **「初期層における自身への注意集中」**が
発生源である可能性を示唆



自身に注意を 集中する介入実験

介入によって Sink が
発生

→ **自身への注意集中**が
Sink 発生に強く寄与



今後の展望

- 具体的な発生**機序**の分析
 - 「自身への注意集中」で Sink が発生する内部機序を解明
- モデル非依存な要因の解明
- Attention Sink や Massive Activation を考慮した最近のモデルを分析 (e.g. Qwen3-Next, gpt-oss)