

一様拡散言語モデルの能力獲得過程

佐々木 睦史^{1*}, 葉 夢宇¹, 趙 羽風², 工藤 慧音^{1,2}, 坂口 慶祐^{1,2}, 鈴木 潤^{1,2,3}(¹ 東北大学 ² 理化学研究所 ³ NII LLMC)

*mutsumi.sasaki@dc.tohoku.ac.jp

概要

- 一様拡散言語モデル Sumi を用いて能力獲得順序・内部表現収束過程を調査
- 自己回帰言語モデル (AR) と能力獲得順序、表現の収束の仕方が異なる可能性が示唆

背景

- 一様拡散言語モデルは一様なノイズから系列全体を反復更新させる
- 能力獲得順序 (Liu+26) や内部表現収束過程 (Martinez+24) は AR で分析されてきた

実験1. 能力獲得順序

ベンチマーク間の能力獲得順序

arc (easy, challenge), race, piqa, winogrande, hellaswag

- 正解率に対し正規化スコア $\tilde{s}_t$ を計算
 - $\tilde{s}_t = (s_t - s_{\text{start}}) / (s_{\text{final}} - s_{\text{start}})$
- 正規化スコアが閾値を初めて上回る点での順位を計算
 - 複数閾値での平均順位に対しモデル間の順位相関を計算

AR同士の能力獲得順序は似る (Liu+26) が

AR と Sumi は類似性が比較的低い

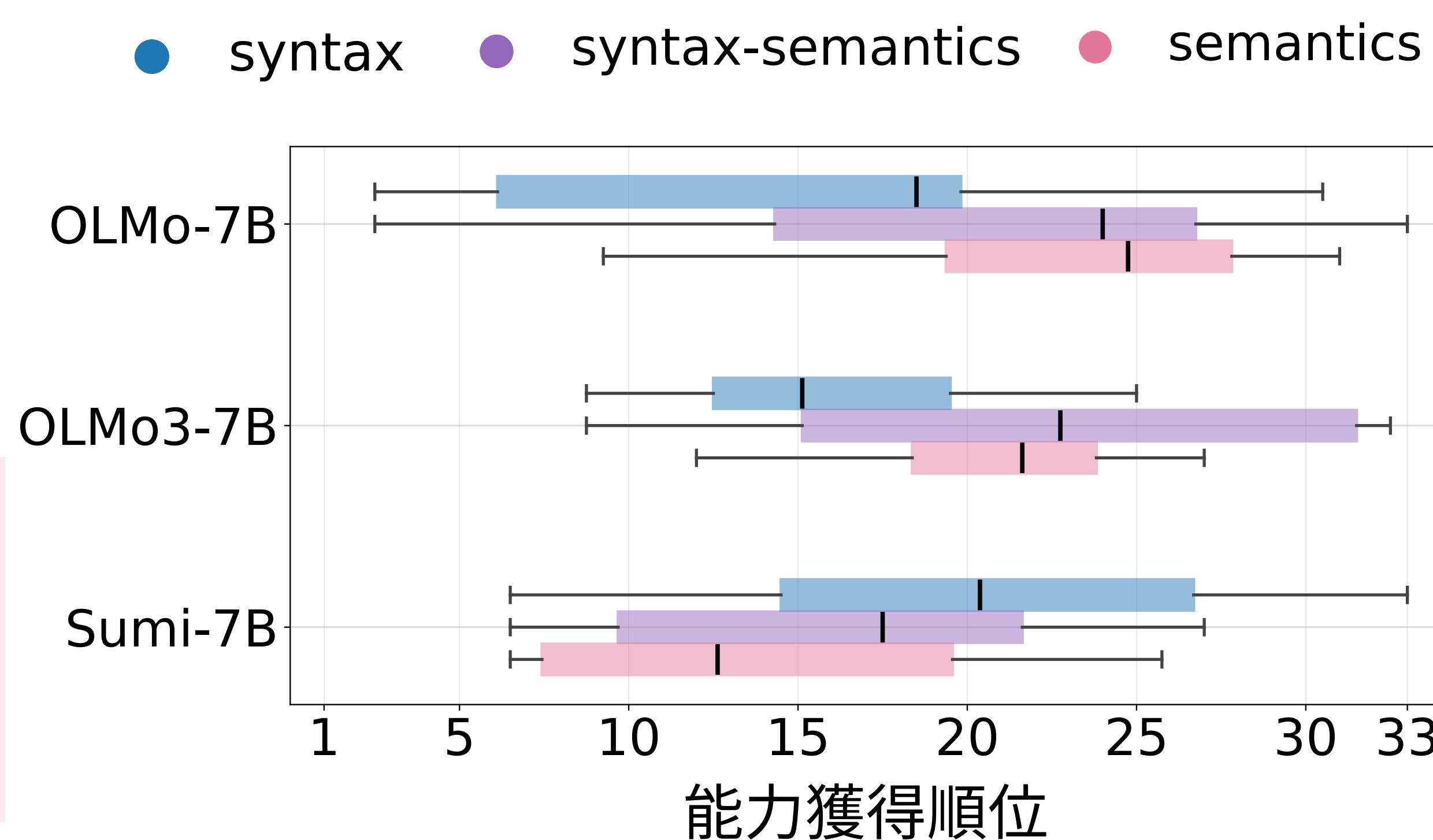
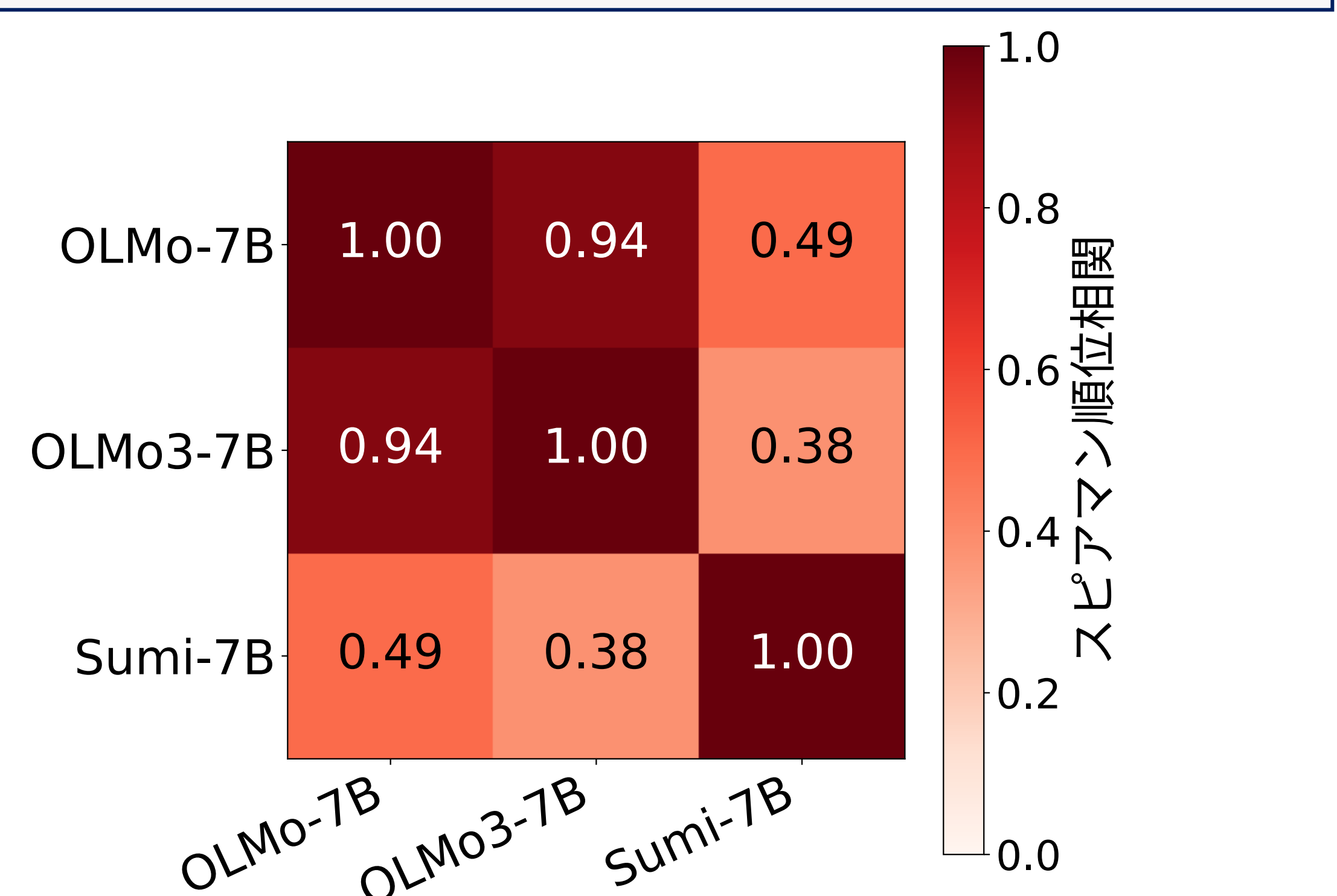
文法能力の獲得順序

- BLiMP の正解率 が 0.6 以上の 33 サブセット を使用
- 正規化スコアが閾値を初めて上回る点での順位を計算
 - 複数閾値で平均順位を報告
 - BLiMP 付属の 3 分野各々で順位の統計をとる

Sumi は構文・語法が遅れて学習される傾向

各ノイズ除去ステップで任意トークンを更新できるため

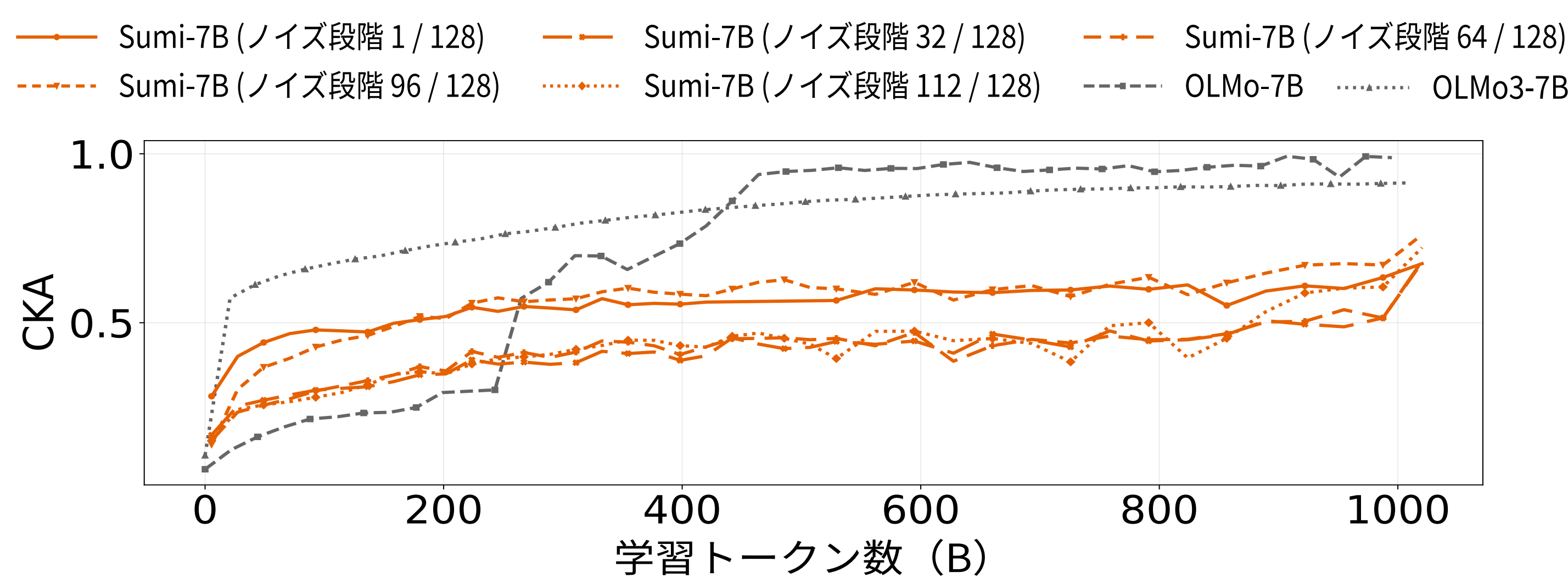
ARより語順の信号が弱い可能性



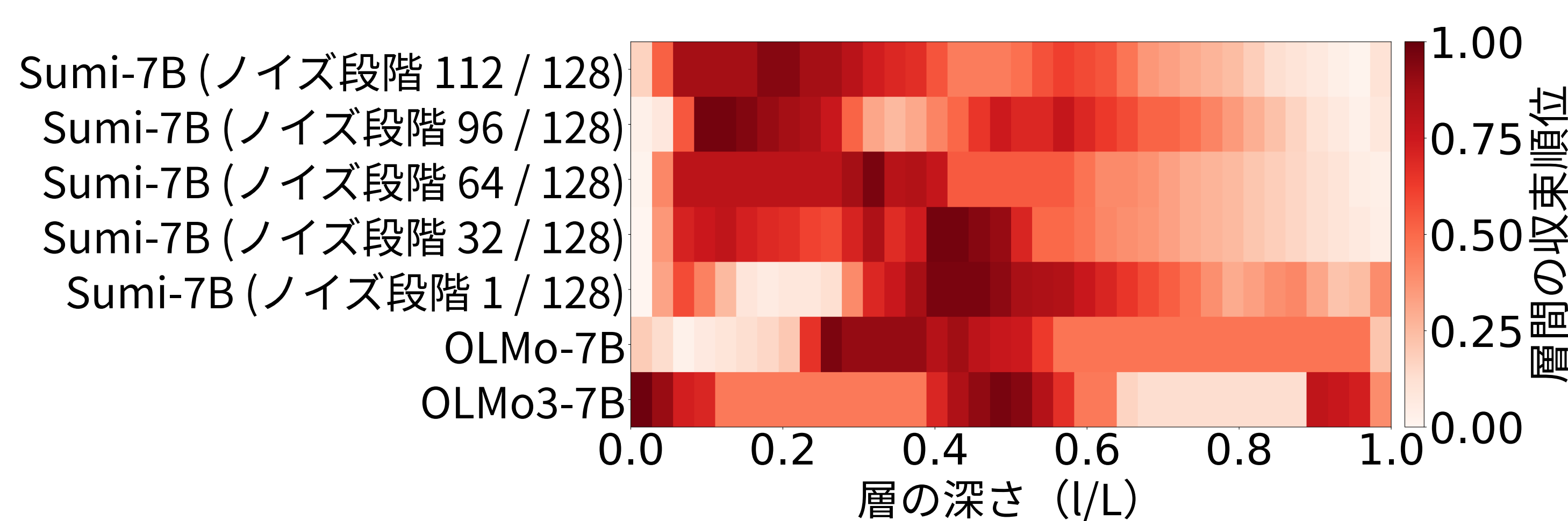
実験2. 内部表現収束過程

- race データセットを入力した全トークンの隠れ状態行列 **H** を取得
- 隠れ状態間の表現類似度を測る指標として CKA (Kornblith+19) を層ごとに計算

層間の平均 CKA の推移



層方向の収束順序の傾向

**Sumi は隠れ状態の収束が遅く、ノイズ段階が大きい隠れ状態ほど AR と異なる収束傾向**

任意トークンを更新できる機構やノイズの少ない状態を学習する訓練方法が影響する可能性

訓練過程の更なる理解にはノイズ状態間の内部表現の写像を追跡する必要がある可能性