

2025.11.26 (Wed)

10:00-12:00

https://llmc.nii.ac.jp/llmc_sympo2025/



Japanese Symposium on Open Large Language Models
LLM-jp成果発表

大規模言語モデルの事前学習と中間学習

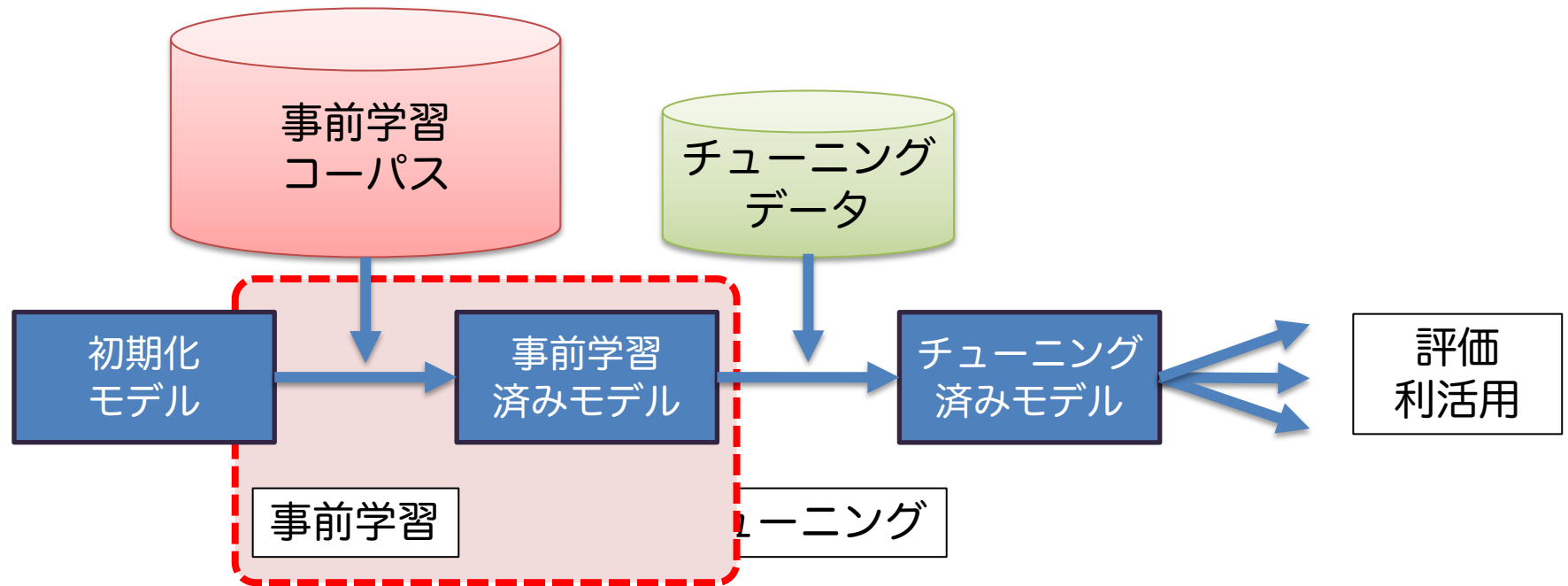
鈴木潤

国立情報学研究所 客員教授

東北大学 言語AI研究センター センター長・教授



大規模言語モデル(LLM)構築の流れ





事前学習の前提

- **事前学習**：LLM構築において計算機利用の観点で最も時間もコストもかかる工程
- 多数の**試行**は困難 + 失敗した時の**損失** (時間/費用) が膨大
=> 定型の「**失敗しない設定**」が通常用いられる

LLM-jpにおいても世の中の多くの知見に従って
事前学習を設計・実施



LLM-jpでの最近の取り組み

- (データの質を上げる取り組み)
 - 一つ前の河原先生からの発表で報告済み
- 1. 大規模データでの学習実験
- 2. 学習率スケジューラーの見直し
- 3. 中間学習 (Mid-training) の導入
- 4. その他：MoEモデル

- 1. 大規模データでの学習実験



実験設定

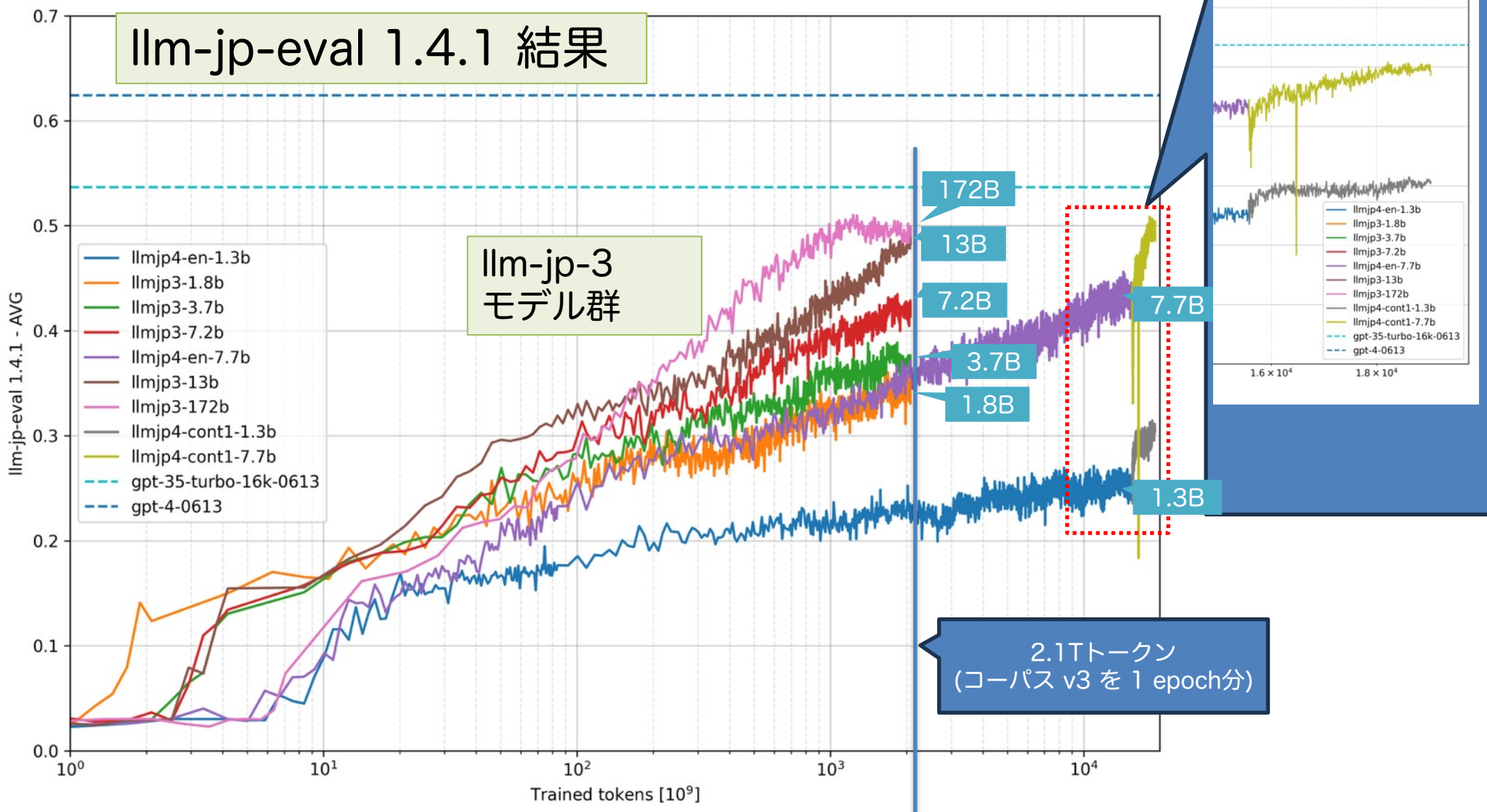
- 事前学習データ [参考] llm-jp-3 (昨年度) → 2.1T トークン
 - ①英語データ： 15.6T トークン (合計 19.1T トークン)
 - ②日本語を含むデータ：3.5T トークン
- 学習方式
 - 第1段階：①英語データ で事前学習
 - 第2段階：②日本語を含むデータ で継続事前学習

● モデル設定

- 二種類のモデルサイズ
 - 1.3B パラメータ
 - 7.7B パラメータ

	LLM-jp-4 1.3B (w/ tokenizer v3)	LLM-jp-4 8B (w/ tokenizer v3) (LLaMA 3 8B compat.)
Vocabulary size	99487	99487
--hidden-size	2048	4096
--ffn-hidden-size	7168	14336
--num-layers	16	32
--num-attention-heads	16	32
--num-query-groups	8	8
--tensor-model-parallel-size	1	1
--make-vocab-size-divisible-by	128	128
--untie-embeddings-and-output-weights	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

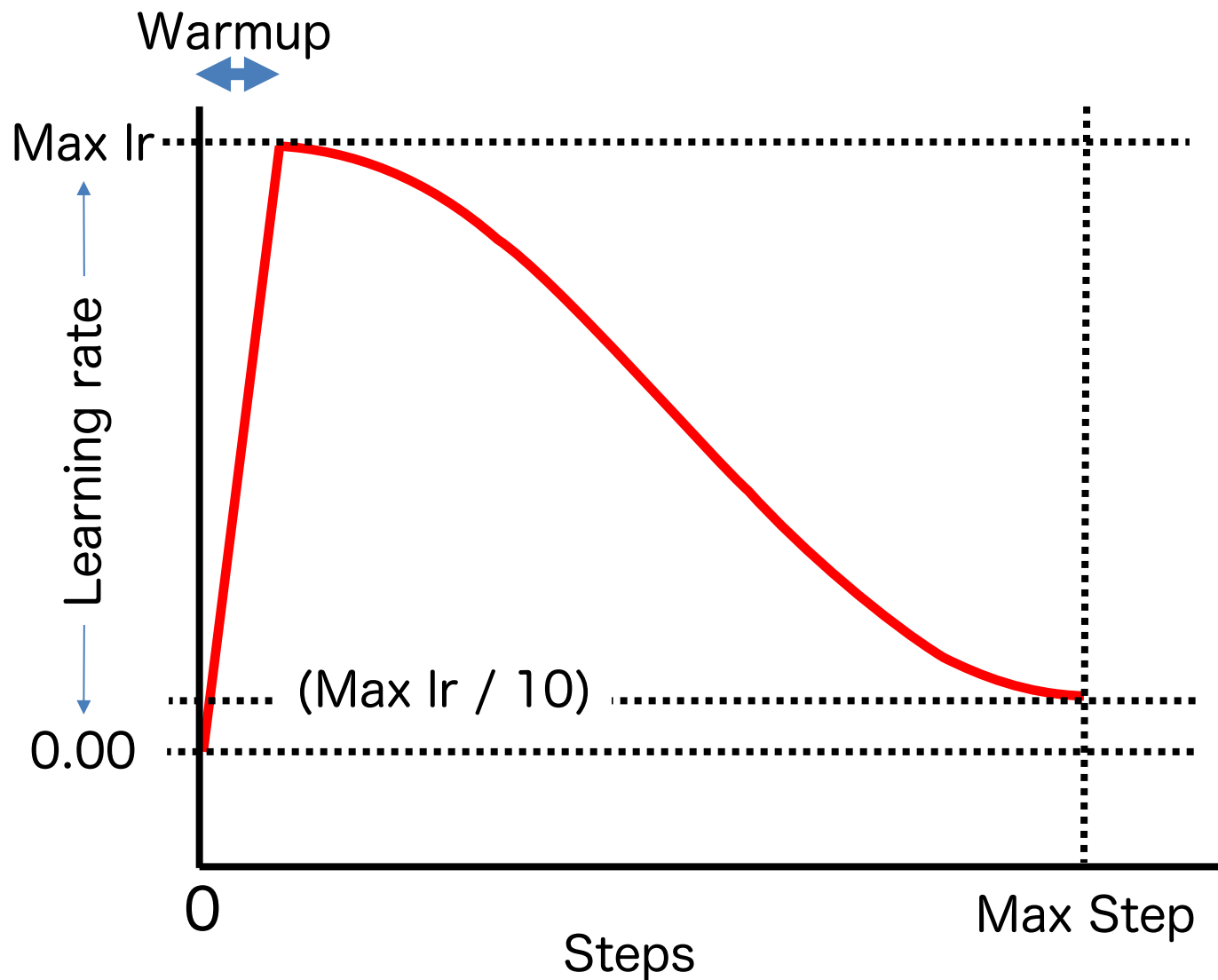
実験結果



- 2. 学習率スケジューラーの見直し

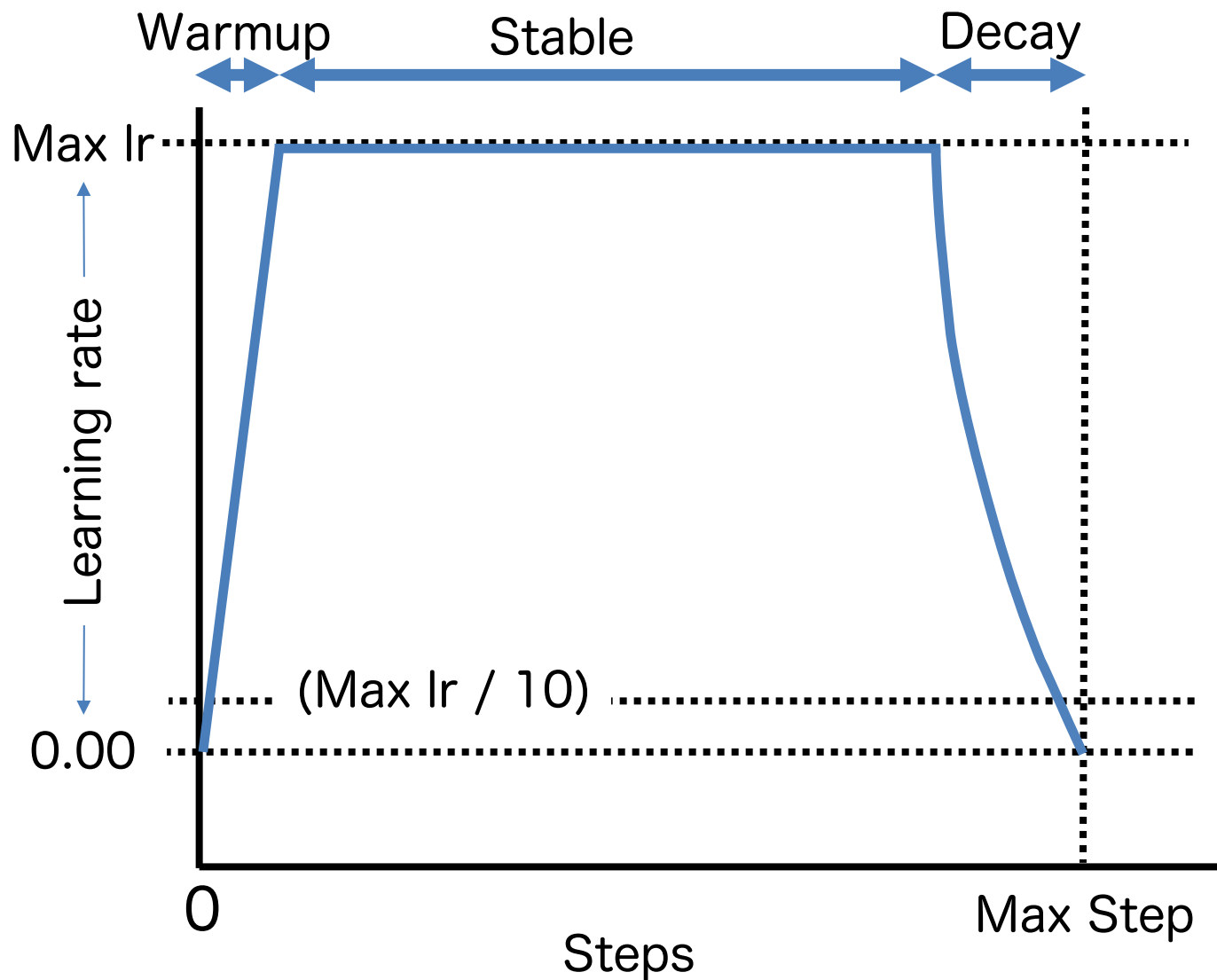


Cosine Scheduler (従来よく使われていたスケジューラー)



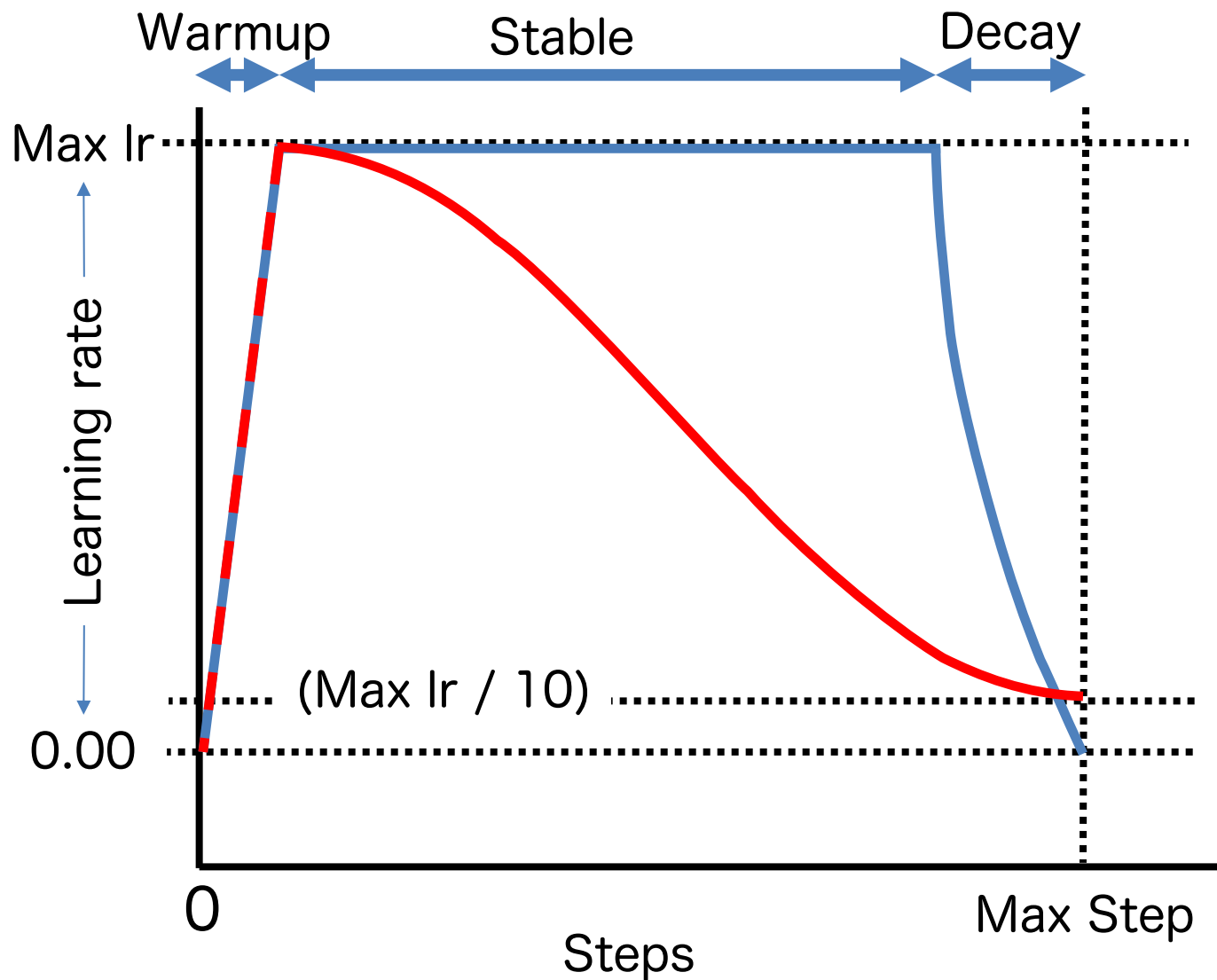


WSD Scheduler





WSD Scheduler





WSD Scheduler

● 目的

- 性能向上ではない
- 事前学習を取り扱いやすくする

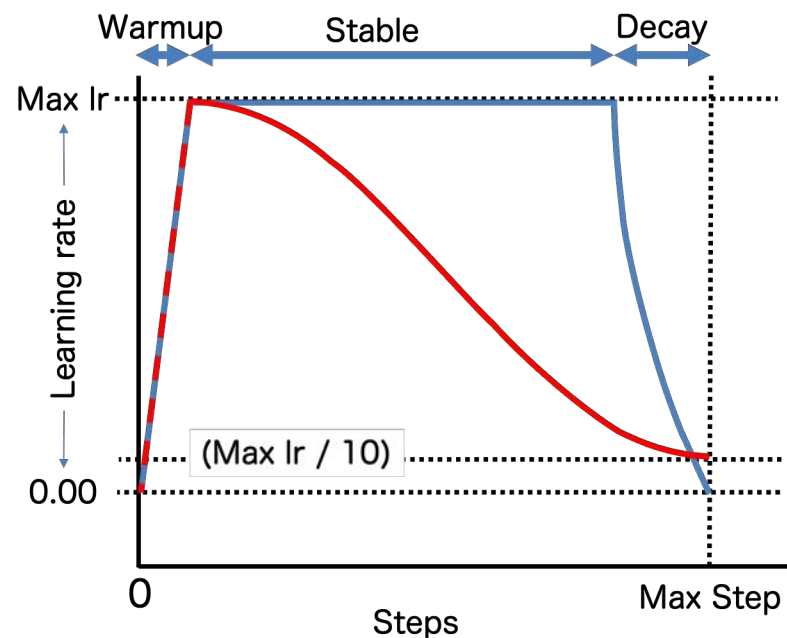
● メリット

(従来よく使われていた
Cosine Scheduler に対して)

- 事前に**最大Step数**を決めなくとも学習開始可能
 - 学習途中で学習データ量を変更可能
 - 学習過程の**再利用**が容易

● デメリット

- 学習が**不安定**になる可能性
- モデルの **性能/出来栄** を予測しにくい



(学習率が高い状態が続くので
Loss-spikeなどが発生する可能性)

(Stable期間はlossの値が高止まり
するため Decay する必要あり)



スケジューラーに対する実験

● WSD Scheduler

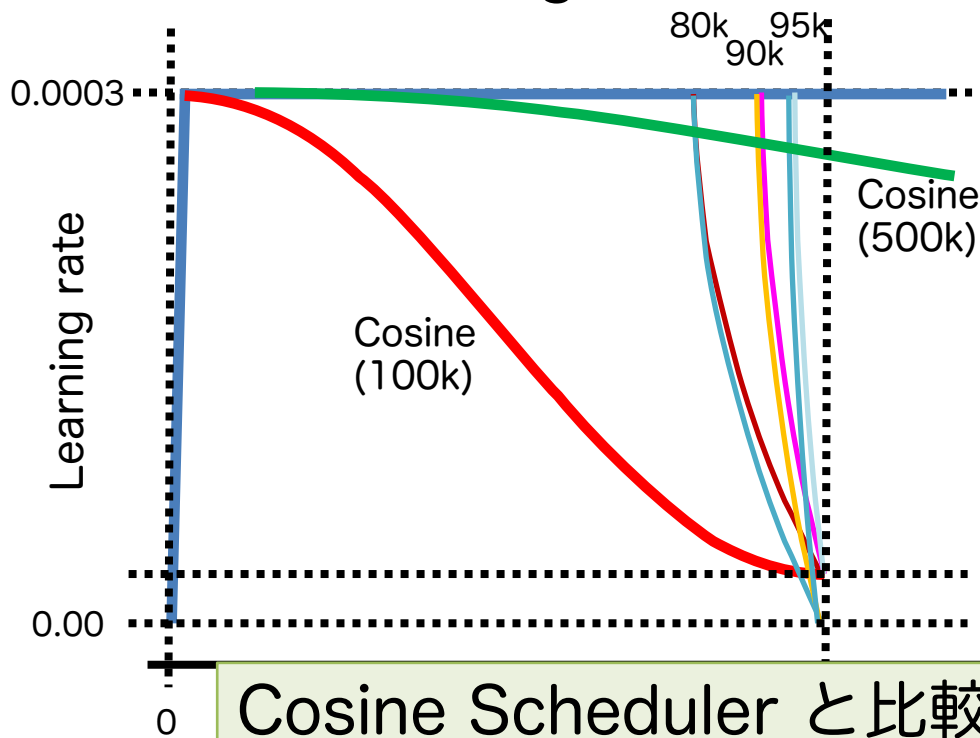
- Warmup 2000 steps
- Decay to 0 or to max lr /10

● Cosine Scheduler

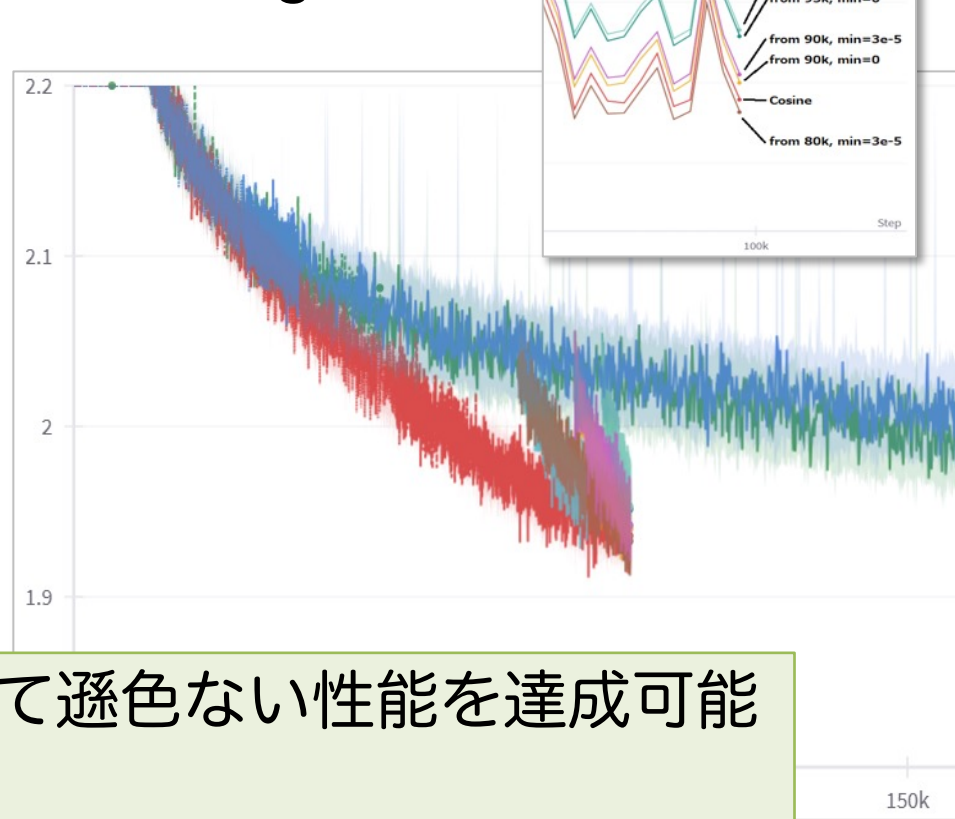
- Warmup 2000 steps
- Decay to max lr /10

100k Steps 付近の Training Loss の挙動

Learning rate



Training loss



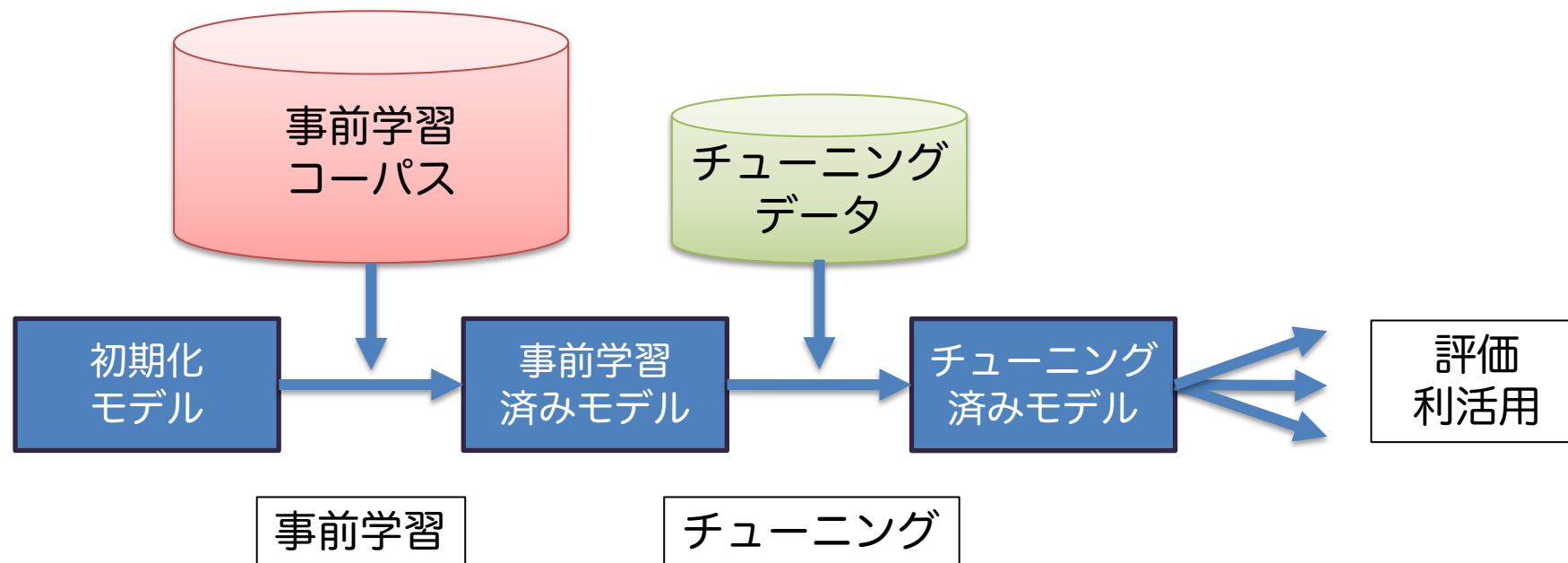
Cosine Scheduler と比較して遜色ない性能を達成可能
=> WSD Scheduler を採用

- 3. 中間学習 (Mid-training) の導入



中間学習：簡単な説明

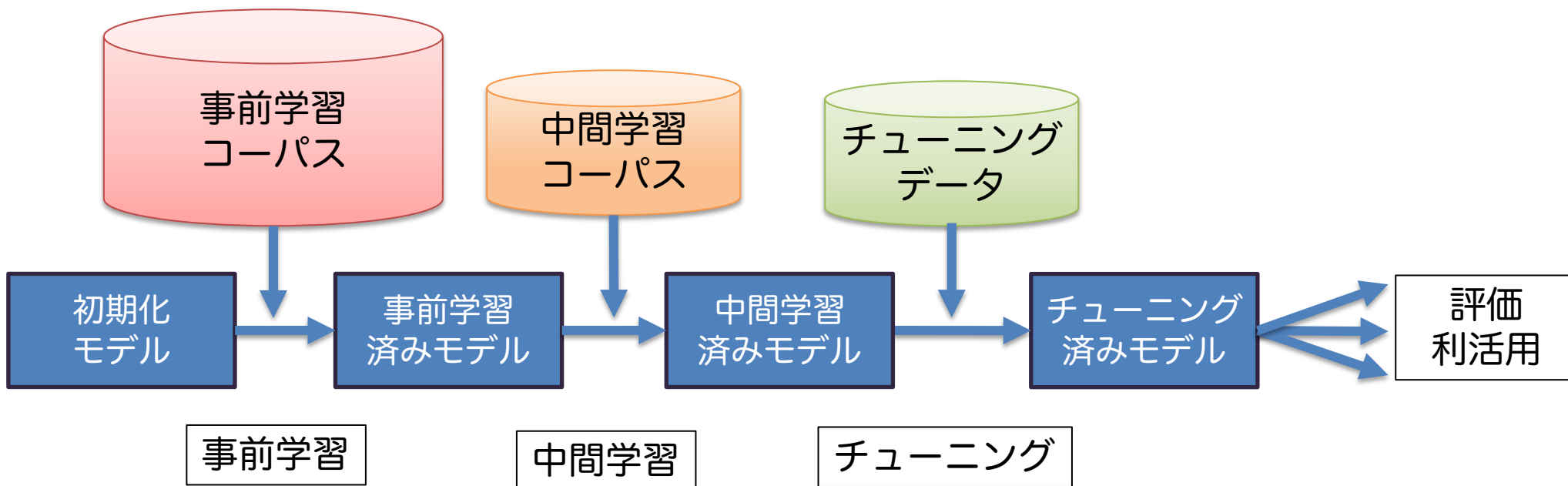
- 事前学習(Pre-training) 後に 継続学習のような形式で実施





中間学習：簡単な説明

- 事前学習の(最)終盤の試行錯誤から派生





中間学習：簡単な説明

- 事前学習の(最)終盤の試行錯誤から派生
 - 事前学習後の **継続事前学習** と同等
 - 学習データ / 学習率スケジューラー / 系列長などを事前学習時の設定から変更
 - 特にデータは性能を向上させたい**特定のタスク/ドメインのデータ**や**良質なデータ**を利用するが多い
- (OLMo2の論文が公開されて以降 中間学習 (Mid-training) という用語をよく聞くようになった?)

<https://arxiv.org/pdf/2501.00656?>

- **Mid-training Recipe.** OLMo-0424 (Ai2, 2024), DBRX (Databricks, 2024), and Llama 3 (Grattafiori et al., 2024) demonstrated the usefulness of data curricula for pretraining, as discussed by Blakeney et al. (2024). We discuss the advantages of splitting pretraining into two stages, with the latter *mid-training* stage being used to infuse new knowledge and patch deficiencies in capabilities. Further, we show how data sources for mid-training can be independently assessed to reduce experimentation cost through a technique we call *micro-annealing* (Section §4).



Blog公開

<https://llm-jp.nii.ac.jp/ja/blog/blog-1039/>

LLM-jp

ホーム ニュース リリース 資料 ブログ メンバー 参加申請 謝辞 English

Blog
ブログ

NII-LLMC

2025年08月27日

LLM-jpモデルに対するOLMo2ベースの中間学習の検討

Date: 2025.8.27
Author: Koshi EGUCHI, Sosuke Hosokawa, Kouta NAKAYAMA

0. はじめに

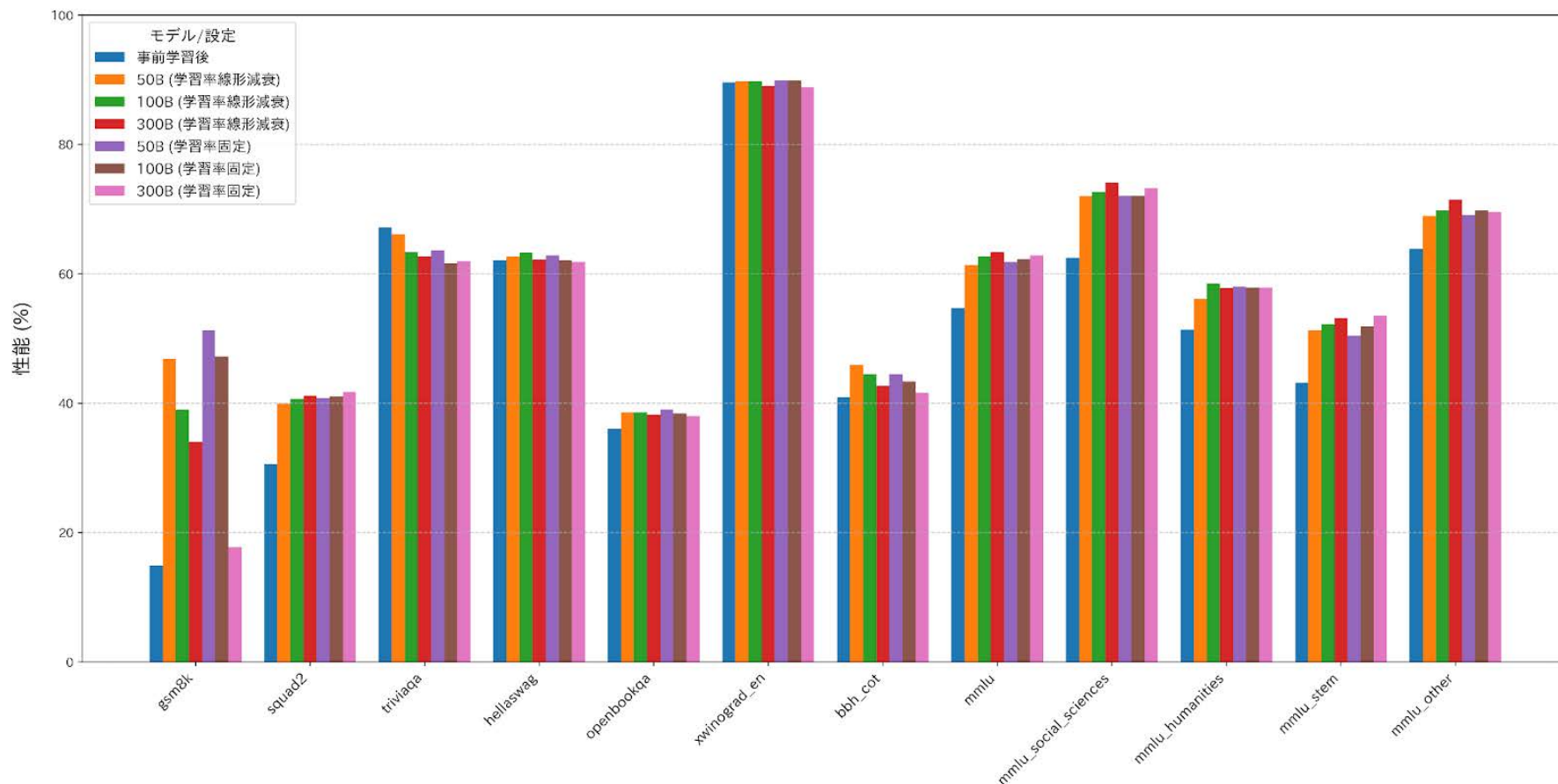
LLM-jpではオープンかつ日本語に強い大規模言語モデルの開発を進めています。LLM-jp-3シリーズの公開に続き、現在新しいモデルシリーズの公開に向けて活動しています。



7.7Bモデルの結果 (Blogから抜粋)

- OLMo2 の報告を概ね再現
=> 中間学習を採用

7.7Bモデルのベンチマーク別性能比較



- 4. その他：Mixture-of-Experts (MoE) モデル



MoEモデルに関するLLM-jpの取り組み

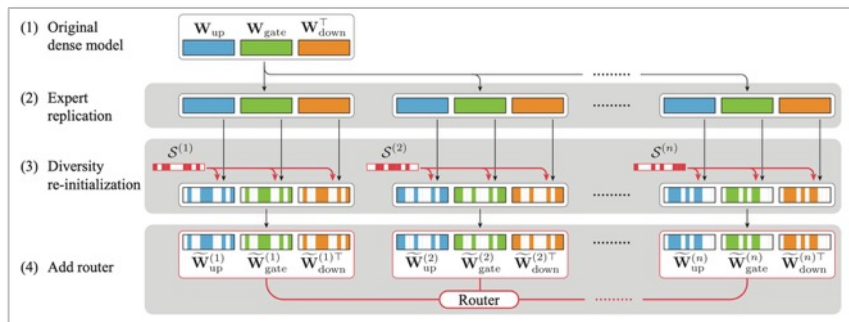
● ICLR-2025 採録

DROP-UPCYCLING: TRAINING SPARSE MIXTURE OF EXPERTS WITH PARTIAL RE-INITIALIZATION

Taishi Nakamura^{1,2,3}, Takuya Akiba², Kazuki Fujii¹, Yusuke Oda³,
Rio Yokota^{1,3}, Jun Suzuki^{4,5,3}

¹Institute of Science Tokyo, ²Sakana AI, ³NII LLMC, ⁴Tohoku University, ⁵RIKEN

https://proceedings.iclr.cc/paper_files/paper/2025/hash/d24b7366d714b09a977946ef0d9bf3ad-Abstract-Conference.html



● LLM-jp-3 MoE シリーズの公開 2025年03月26日

<https://llm-jp.nii.ac.jp/ja/blog/blog-603/>



モデル, (学習データ), コード, 学習ログ 全て公開済み

- Weights** huggingface.co/collections/llm-jp/drop-upcycling-674dc5be7bbb45e12a476b80
- Data** gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3
- Code** github.com/Taishi-N324/Drop-Upcycling
- Logs** wandb.ai/taishi-nakamura/Drop-Upcycling

さらにMoEに関する研究を進め
新たなMoEモデルを学習中

完成後公開予定

- まとめ



まとめ：LLMの事前学習と中間学習

- 大規模コーパスによる事前学習
 - (学習コーパスを洗練させることで性能向上を目指す)
 - 今年度新たなコーパスで学習されたモデルを公開予定
- 学習率スケジューラー：WSD Schedulerの採用
 - 事前学習全体を実用的かつ効率的に実行するため
- 中間学習の導入
 - LLMの性能向上に寄与することを確認
- MoEモデル
 - LLM-jp-3 MoE シリーズの公開 [2025年03月26日]
 - 新たなMoEモデルを公開予定