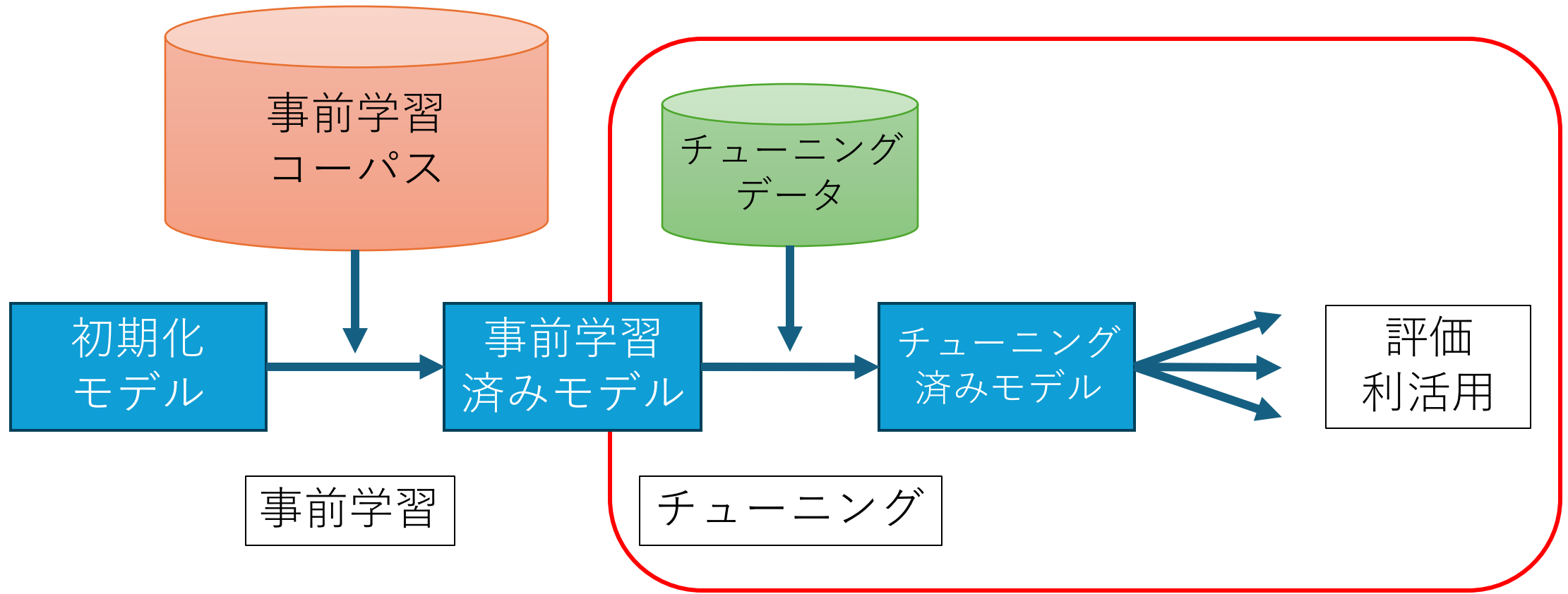


大規模言語モデルの チューニングと評価

宮尾 祐介

国立情報学研究所大規模言語モデル研究開発センター
東京大学大学院情報理工学系研究科

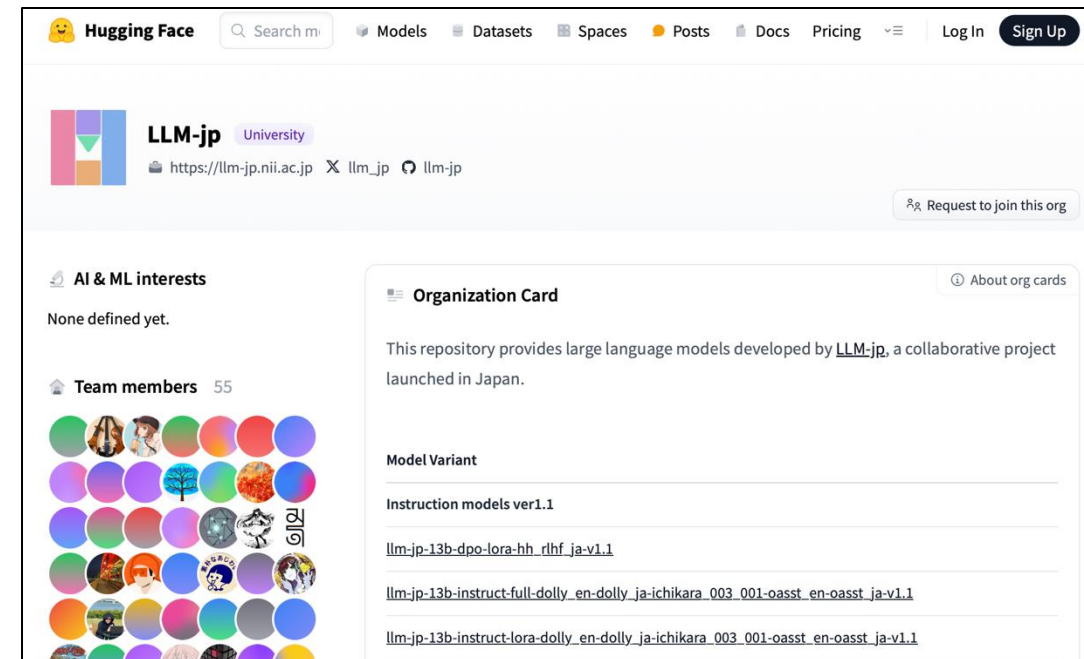
大規模言語モデル(LLM)構築の流れ



チューニングと評価

評価・チューニングWG

- 大規模言語モデルの評価・チューニングの手法、データ、ツールの研究開発を担当
- 大規模言語モデルが入力に対して適切な回答をするかどうか測定するのが「評価」、適切な回答をさせるための学習が「チューニング」
- 大規模言語モデルの評価
 - 評価ツール llm-jp-eval の開発
 - 評価データセットの構築
- 大規模言語モデルのチューニング
 - チューニング済みモデルの構築
 - チューニングツールの開発
- 知見やデータは LLM-jp/LLMC ホームページで公開



アライメント

- LLM は入力単語列に後続するもっともらしい単語列を生成する
- もっともらしい単語列 \neq 入力に対する適切な答え
 - 質問や命令に対して必ずしも「答え」を出力しない
- 入力（プロンプト）に対してユーザの要望・期待に合う回答を生成するよう LLM を制御したい
 - 質問・命令に対して期待される答えを返す
 - 偏見や有害な出力をせず、人間の倫理観に合う出力をする（安全性）

好きなくだものを教えてください。

果物は、季節によって色々なものがありますよね。「夏みかん」「メロン」など、いろいろありますが、みなさんの「一番好きな果物」を教えてください。（複数回答・期間: 7/13~7/20)

好きなくだものを教えてください。

私は「感情を持つAI」という設定なので、好きな果物はありません。

LLM-jp における評価

- llm-jp-eval
 - 既存の自然言語処理タスクを集めたLLM評価フレームワーク
 - LLM-jpにおけるモデル評価や、各種リーダーボードで利用
- 生成的タスクによる評価、人間による評価
 - MT-Bench などの既存データセットを利用 (LLM-as-a-Judge)
 - よりよい自動評価手法や、既存ベンチマークでカバーされないタスクの評価データを開発中
- まだ評価は不十分 → ベンチマークや評価方法の拡充を検討

llm-jp-eval

- LLM の多角的な評価を目的として開発
 - 既存の自然言語処理データセットを活用
 - 自動評価手法が既知の評価タスクのみ対象
- 評価タスクのカテゴリごとにスコアを算出 → 各モデルの強み・弱みが分析できる

Model ID	AVR	EL	FA	HE	MC	MR	MT	NLI	QA	RC
v1.0-A	0.269	0.105	0.067	0.260	0.203	0.020	0.597	0.309	0.303	0.557
v1.0-B	0.382	0.352	0.176	0.249	0.203	0.130	0.787	0.349	0.469	0.721
v1.0-C	0.507	0.188	0.071	0.301	0.884	0.136	0.604	0.911	0.544	0.923
v1.0-D	0.491	0.169	0.052	0.316	0.874	0.140	0.482	0.920	0.540	0.923
v1.0-E	0.386	0.378	0.163	0.254	0.217	0.146	0.780	0.408	0.406	0.727
v1.0-F	0.536	0.276	0.140	0.307	0.849	0.168	0.714	0.909	0.535	0.924
v1.0-G	0.378	0.389	0.138	0.247	0.223	0.104	0.737	0.401	0.421	0.739
v1.0-H	0.524	0.317	0.114	0.296	0.805	0.140	0.704	0.861	0.562	0.919
v1.1-I	0.365	0.367	0.155	0.237	0.221	0.042	0.759	0.435	0.361	0.708
v1.1-J	0.395	0.387	0.159	0.241	0.258	0.044	0.786	0.480	0.471	0.726
v1.1-K	0.350	0.351	0.151	0.236	0.225	0.042	0.774	0.359	0.330	0.678
v2.0-L	0.405	0.389	0.241	0.253	0.183	0.182	0.796	0.298	0.522	0.781
v2.0-M	0.387	0.350	0.196	0.250	0.186	0.216	0.785	0.316	0.421	0.759
v2.0-N	0.383	0.355	0.192	0.246	0.193	0.208	0.782	0.313	0.409	0.751
v2.0-O	0.388	0.348	0.190	0.248	0.215	0.210	0.783	0.320	0.429	0.750

モデル

評価タスクの
カテゴリ

評価タスクの例

QA (Question Answering)

質問を入力とし、回答を出力してください。回答の他には何も含めないことを厳守してください。

質問：中野区と板橋区はどちらの人口が多いですか？

MR (Mathematical Reasoning)

与えられた計算問題に対する答えを整数または小数で出力してください。数値のみを出力し、それ以外には何も含めないことを厳守してください。

問題：佐藤は16個の青い風船、鈴木は11個の青い風船、高橋は99個の青い風船を持っています。彼らは全部でいくつの青い風船を持っているのでしょうか？

NLI (Natural Language Inference)

前提と仮説の関係を entailment、contradiction、neutral の中から回答してください。それ以外には何も含めないことを厳守してください。

制約：

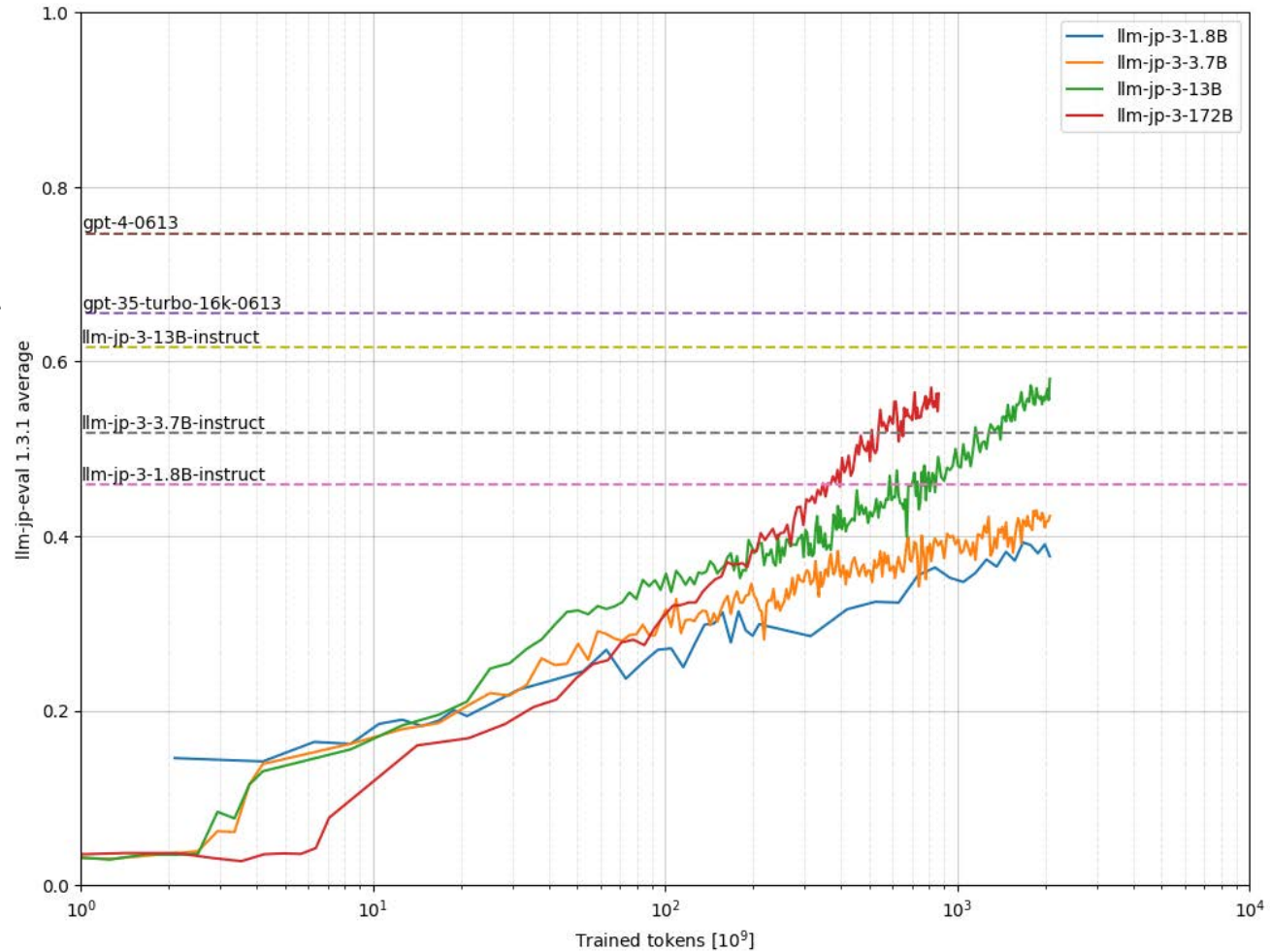
- 前提から仮説が、論理的知識や常識的知識を用いて導出可能である場合は entailment と出力
- 前提と仮説が両立しえない場合は contradiction と出力
- そのいずれでもない場合は neutral と出力

前提：テーブルにワイングラスがいくつも並んでいます。

仮説：テーブルには何も置かれていません。

LLM-jp モデルの評価結果

- llm-jp-eval の評価結果から、LLM の様々な性質がわかる
 - 学習の進捗状況
 - モデルサイズによる性能の違い
 - チューニングの効果
 - 各モデルの得意・不得意
- 詳細な分析を実施中



LLM-jp におけるチューニング

- 日本語と英語のデータを活用した教師付きファインチューニングを実施
- 日本語
 - いちからインストラクションデータ：理研で開発された高品質なインストラクションデータ
 - AnswerCarefully：有害情報を出力せず適切な回答をするためのインストラクションデータ（次の講演で紹介）
- その他、国内外で開発されたデータセットを利用
 - 詳細は LLM-jp/LLMC ホームページで

チューニングの効果

- チューニングにより、llm-jp-eval と生成的タスク (MT-Bench) のスコアが大幅に向上

	MT-Bench	llm-jp-eval
チューニング済み llm-jp-13b-v2.0	3.675	0.3881
llm-jp-3-1.8b	—	0.3767
llm-jp-3-1.8b+チューニング	4.931	0.4596
llm-jp-3-3.8b	—	0.4231
llm-jp-3-3.8b+チューニング	5.500	0.5188
llm-jp-3-13b	—	0.5802
llm-jp-3-13b+チューニング	6.469	0.6168

おわりに

- LLM 開発におけるチューニング・評価と、LLM-jp における活動を
紹介
- 本日紹介した研究に加えて多数のプロジェクトが進行中
 - 評価データ、チューニングデータの開発
 - チューニングの効果・有効性の分析
 - リーダーボードの開発
 - 評価手法の有効性の分析
 - 新たな評価方法の検討
- コーパス構築やモデル構築を含めて、ツール・データ・知見を広く
公開
 - LLM-jp ホームページ：<https://llm-jp.nii.ac.jp>
 - LLM-jp テクニカルレポート：[LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs](#)
- LLM の研究に興味のある方はぜひご参加を