

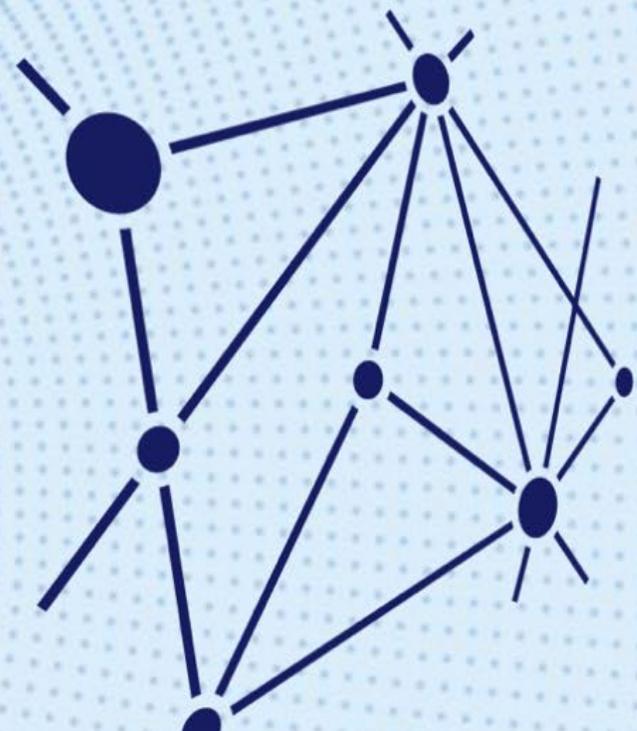
LLMC

国立情報学研究所

大規模言語モデル研究開発センター

(LLM研究開発センター)

Research and Development Center for Large Language Models



黒橋 禎夫

国立情報学研究所 所長／大規模言語モデル研究開発センター センター長

言語モデル

大規模コーパス

私はりんごを

…お店で私はりんごを食べた…

…彼と私はりんごを食べた…

…昨日私はりんごをかじった…

…私はりんごを食べた後…

…私はりんごを殴った夢を…

$$P(\text{食べた} | \text{私はりんごを}) = \frac{\text{コーパス中の頻度(私はりんごを食べた)}}{\text{コーパス中の頻度(私はりんごを)}} = 0.6$$

$$P(\text{殴った} | \text{私はりんごを}) = \frac{\text{コーパス中の頻度(私はりんごを殴った)}}{\text{コーパス中の頻度(私はりんごを)}} = 0.2$$

言語の理解

基礎解析

- 形態素解析、固有表現解析
- 構文解析、述語項構造解析
- 語義曖昧性解消
- 照応・省略解析、談話構造解析

応用システム

- 機械翻訳
- 情報検索
- 質問応答
- 対話システム

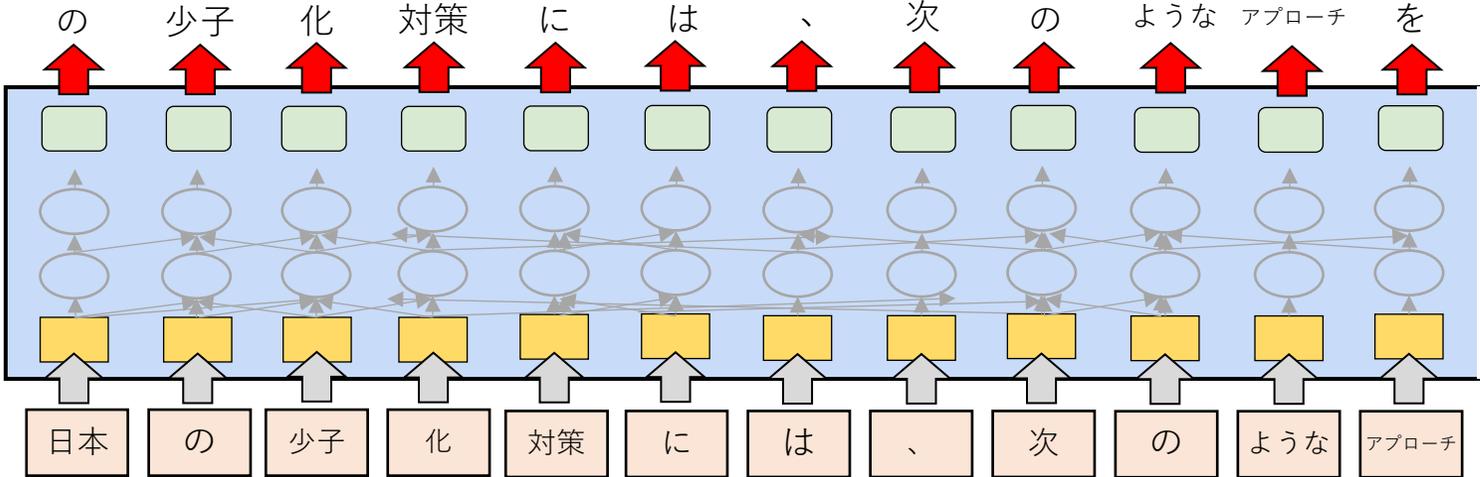
コーパス、アノテーション、機械学習、ニューラルネットワーク

2022年11月公開：大規模言語モデル(175B)に基づくチャットボット

ChatGPT

学習時

学習データ：日本の少子化対策には、次のようなアプローチを...

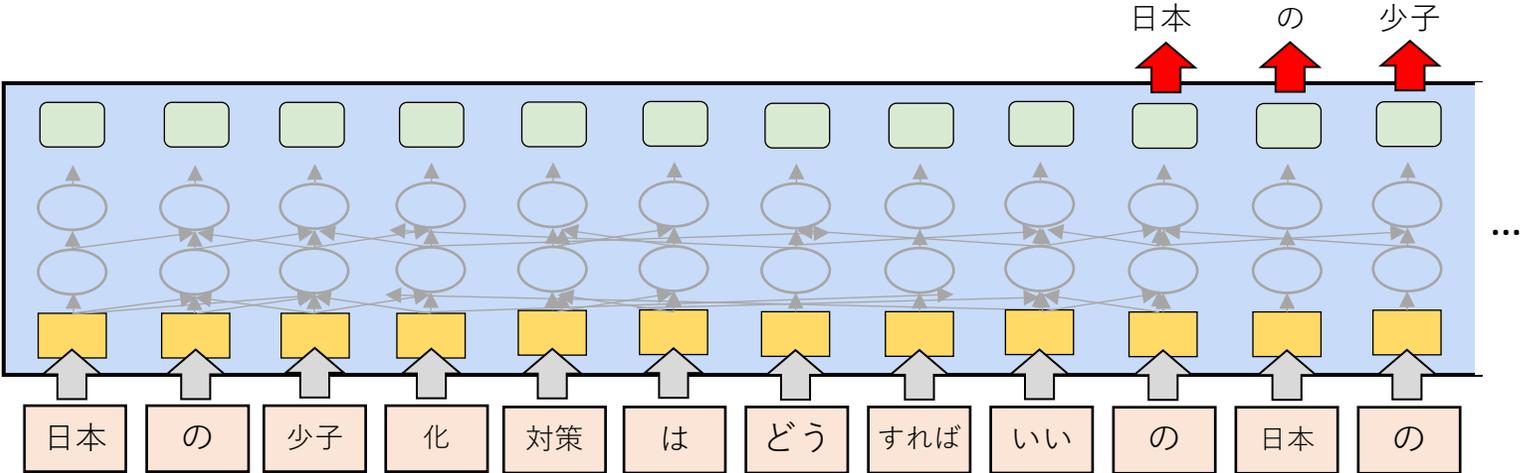


2022年11月公開：大規模言語モデル(175B)に基づくチャットボット

ChatGPT

推論時

プロンプト：日本の少子化対策はどうすればいいの



2022年11月公開：大規模言語モデル(175B)に基づくチャットボット

ChatGPT

基礎解析

- 形態素解析、固有表現解析
- 構文解析、述語項構造解析
- 語義曖昧性解消
- 照応・省略解析、談話構造解析

応用システム

- 機械翻訳
- 情報検索
- 質問応答
- 対話システム

すべて解決! ?

翻訳：うなぎ文

 DeepL

日本語 (自動検出) ▾ ⇄ 英語 (US) ▾ 用語集

部長はうなぎです × The chief is an eel.

その他の候補：
The director is an eel.

翻訳：うなぎ文

ChatGPT PLUS

GPT-4



「部長はうなぎです」を英語に訳してください



「部長はうなぎです」というフレーズは、直訳すると"The manager is an eel" となります。   

ただし、このフレーズは日常的な会話で直接使用されるものではないと思われるため、コンテキストや文脈に応じて、適切な翻訳を行う必要があるかもしれません。特定の文化的・慣用的な意味が含まれている場合、そのニュアンスを取り入れた訳が必要となることも考えられます。

翻訳：うなぎ文

ChatGPT PLUS



お店で注文している場合だとどうなりますか



お店で「部長はうなぎです」と注文する場面を想像すると、これはおそらく部長がうなぎを注文したいという意味になります。この場合、英語での翻訳は "The manager would like eel" もしくは "The boss orders eel" となるでしょう。



参考：『ChatGPT翻訳術 新AI時代の超英訳スキルブック』山田 優 著, 株式会社アルク 2023.

2022年11月公開：大規模言語モデル(175B)に基づくチャットボット

ChatGPT

基礎解析

- 形態素解析、固有表現解析
- 構文解析、述語項構造解析
- 語義曖昧性解消
- 照応・省略解析、談話構造解析

応用システム

- 機械翻訳
- 情報検索
- 質問応答
- 対話システム

すべて解決! ?

言語処理学会理事会主催 緊急パネル

緊急パネル:ChatGPTで自然言語処理は終わるのか？

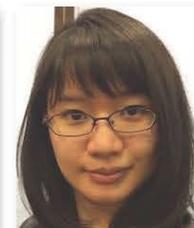
- ファシリテーター

- 乾 健太郎 氏 (東北大)



- パネリスト

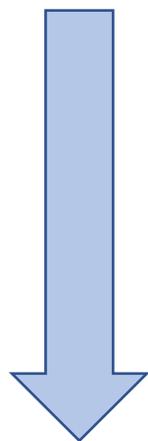
- 黒橋 禎夫 氏 (京大)
- 相良 美織 氏 (バオバブ)
- 佐藤 敏紀 氏 (LINE)
- 鈴木 潤 氏 (東北大)
- 谷中 瞳 氏 (東大)



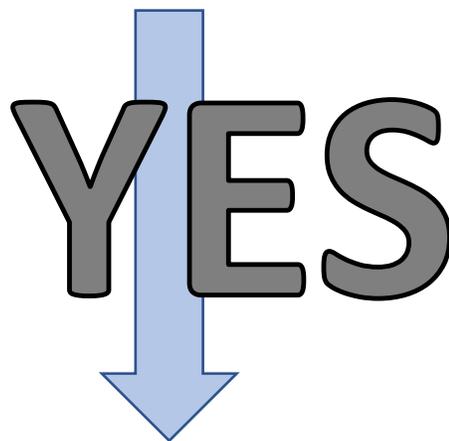
- 3月14(火) 13:10-13:50, H会場 (劇場ホール)

- Slack: [#0314-1310-緊急パネル-h会場](#) (質問はこちらへ)
- Twitter: [#ChatGPTで自然言語処理は終わるのか](#)

ChatGPTで自然言語処理は終わるのか？

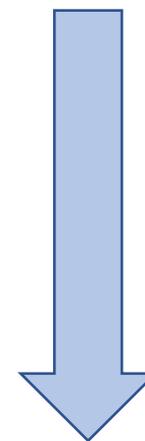


ChatGPT+



旧来のNLP

形態素解析、構文解析、省略解析 ...
関係抽出、要約、MT、対話システム ...



数年後には、
1年後にも

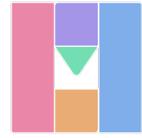
新しい自然言語処理

もっと人間や社会へ向かうもの

- 脳神経科学との関係
- 人の感情
- 人の理解、納得、説得
- **AI Native**世代の教育、批判的思考の涵養
- 情報の信頼性、関係付け（参照）

そのためには

- LLMが**完全にオープン**であること
- また、ChatGPTにおける日本語コーパスは圧倒的に少ない
- 日本文化や経済安全保障の視点
- 良質で十分な日本語テキスト（+英語テキスト）で学習したオープンな大規模言語モデルを日本が持つ必要性



LLM-jp (LLM勉強会)

- オープンかつ日本語に強い大規模モデルを構築し、LLMの原理解明に取り組む
- モデル・データ・ツール・技術資料等を議論の過程・失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可

2023.5
自然言語処理の研究者
30名程度による勉強会を開催

2023.10
mdxを用いて
13Bパラメータモデル
LLM-jp-13Bを公開

2023.11
ABCI第2回LLM構築支援
プログラムに参加
175Bパラメータモデルの
学習トライアル

2024.4
GENIAC環境で
172Bパラメータモデルの
学習開始

1700名超

国立情報学研究所に
LLM研究開発センター設置

- **mdx**: データ活用社会創成プラットフォーム. 9大学2研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境
- **ABCI**: AI橋渡しクラウド. 産業技術総合研究所 (AIST) が提供するAI向け計算用で現状国内最大の計算資源
- **GENIAC**: Generative AI Accelerator Challenge. 日本国内の基盤モデル開発力の底上げのために計算資源の提供等を行う経産省のプログラム

NII 大規模言語モデル研究開発センターの体制と研究課題

外部連携機関

東北大学
東京大学
東京工業大学
早稲田大学
京都大学

黒橋 禎夫 センター長 (NII所長)

相澤 彰子 副センター長 (NII副所長)

武田 浩一 副センター長 (NII特任教授)

科学 主幹

関根聡 特任教授、岡崎直観 客員教授、河原大輔 客員教授、空閑洋平 客員准教授、鈴木潤 客員教授、
田浦健次郎 客員教授、宮尾祐介 客員教授、横田理央 客員教授、尾形哲也 客員教授(10/1～)

事務局

課題1: 研究開発用LLM構築

- ① 新規コーパス開拓・整備、フィルタリングツール開発
- ② GPU並列計算環境整備
- ③ モデル構築

課題2: 透明性・信頼性に関する研究開発

【透明性】

- ④ コーパス検索基盤及びLLM入出力観察・分析基盤の構築
- ⑤ チューニング・評価、FT・LHFの効果分析

【信頼性】

- ⑥ 広島プロセス国際指針等を踏まえた安全性対策
- ⑦ データ改変、データバイアス等の影響抑制等
- ⑧ 外部知識利用、ハルシネーション防止技術

【社会受容性】

- ⑨ LLMモニタリング基盤の構築・運用
- ⑩ 法制度や倫理基準を踏まえたLLM評価手法の開発

課題3: 高度化に関する研究開発

- ⑪ ドメイン適応、モデル軽量化
- ⑫ Transformerアーキテクチャーの発展

※センター全体で研究者等30名規模

※データ作成等を担うアノテータを20名程度、大学院学生RA (リサーチ・アシスタント) を30名程度を雇用

※次世代AI人材育成プログラム BOOSTに採択された研究者がクロスアポイントメントで本センターに参画することも想定

2024年度の計算資源

mdx

4月～2025年3月、昨年度同様の資源を通年利用（専有利用 A100 16ノード（128枚）及びスポット利用）

経産省GENIACプログラムの支援によるGCPの計算資源

4月15日～8月15日、172Bパラメータモデルを学習

東工大TSUBAME 4.0

5月～9月の5ヶ月間、H100 50ノード（200枚）

さくらインターネット

7月9日～7月末 H100 33ノード（264枚） 8月～2025年3月 H100 100ノード（800枚）

リソースの開発

トークナイザー

- 10万語彙（日本語5万語、英語3万語、中国語1万語、韓国語1万語）

事前学習コーパス

- 日本語：約5,920億トークン（WebアーカイブCommon Crawlから日本語テキストを全量抽出・フィルタリング、国立国会図書館WARP事業で収集されたURLを基にクロールしたデータ、日本語Wikipedia、KAKEN (科学研究費助成事業データベース) の研究課題概要)
- 英語：約9,500億トークン（Dolma等）
- 中国語・韓国語：約10億トークン
- プログラムコード：約1,140億トークン

安全性関連のデータ

- AnswerCarefully：安全性・適切性に特化した指示学習データ
- JSocialFact：ソーシャルメディアからの偽誤情報データセット
- LLM-jp Toxicity Dataset：有害ウェブページデータセット

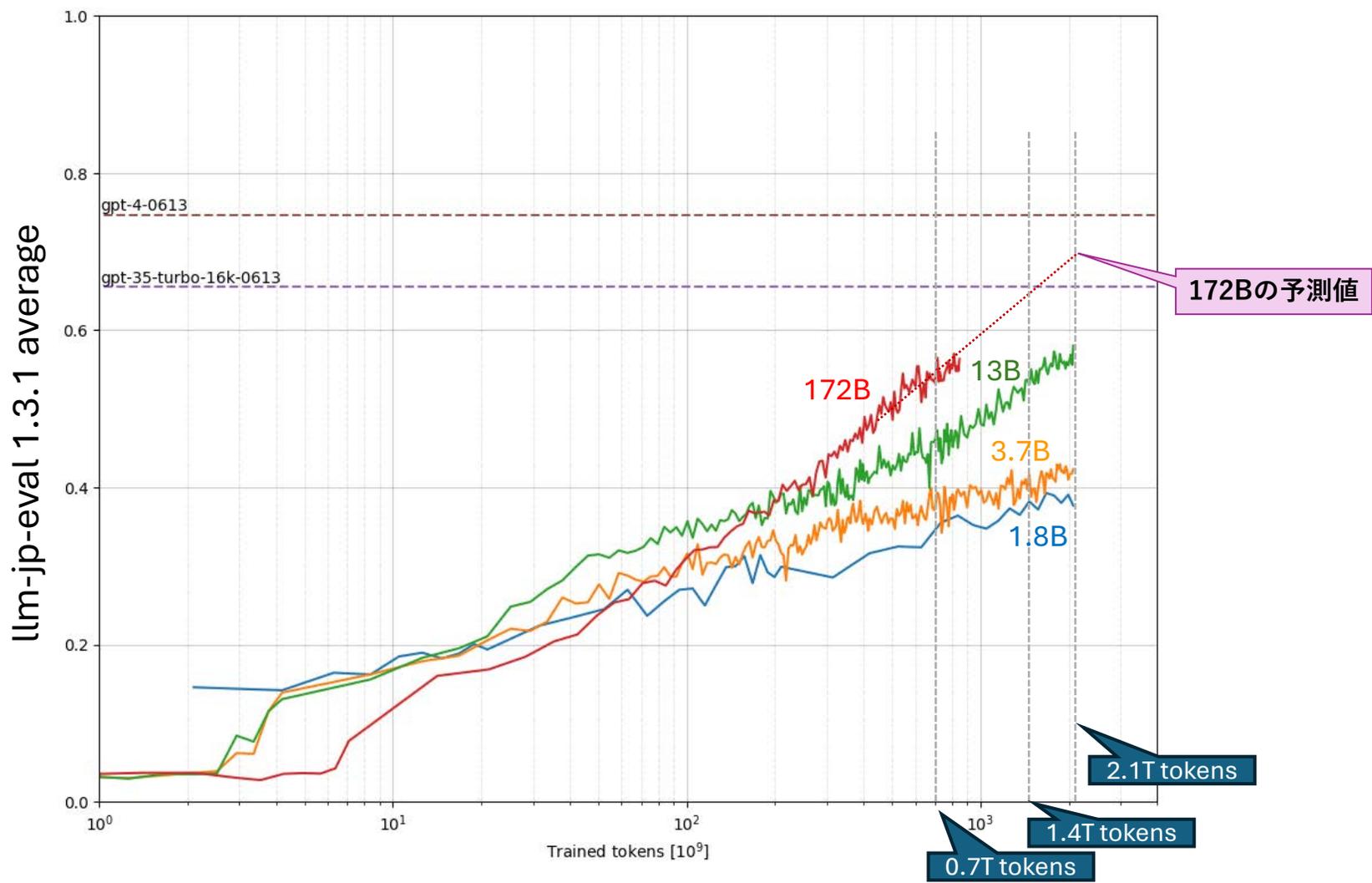
指示学習データ

- オープンな既存チューニング用データに加え、ichikara-instruction、AnswerCarefullyなど日本語オリジナルの高品質チューニングデータを人手で新たに作成、8データセット、計35万インスタンス

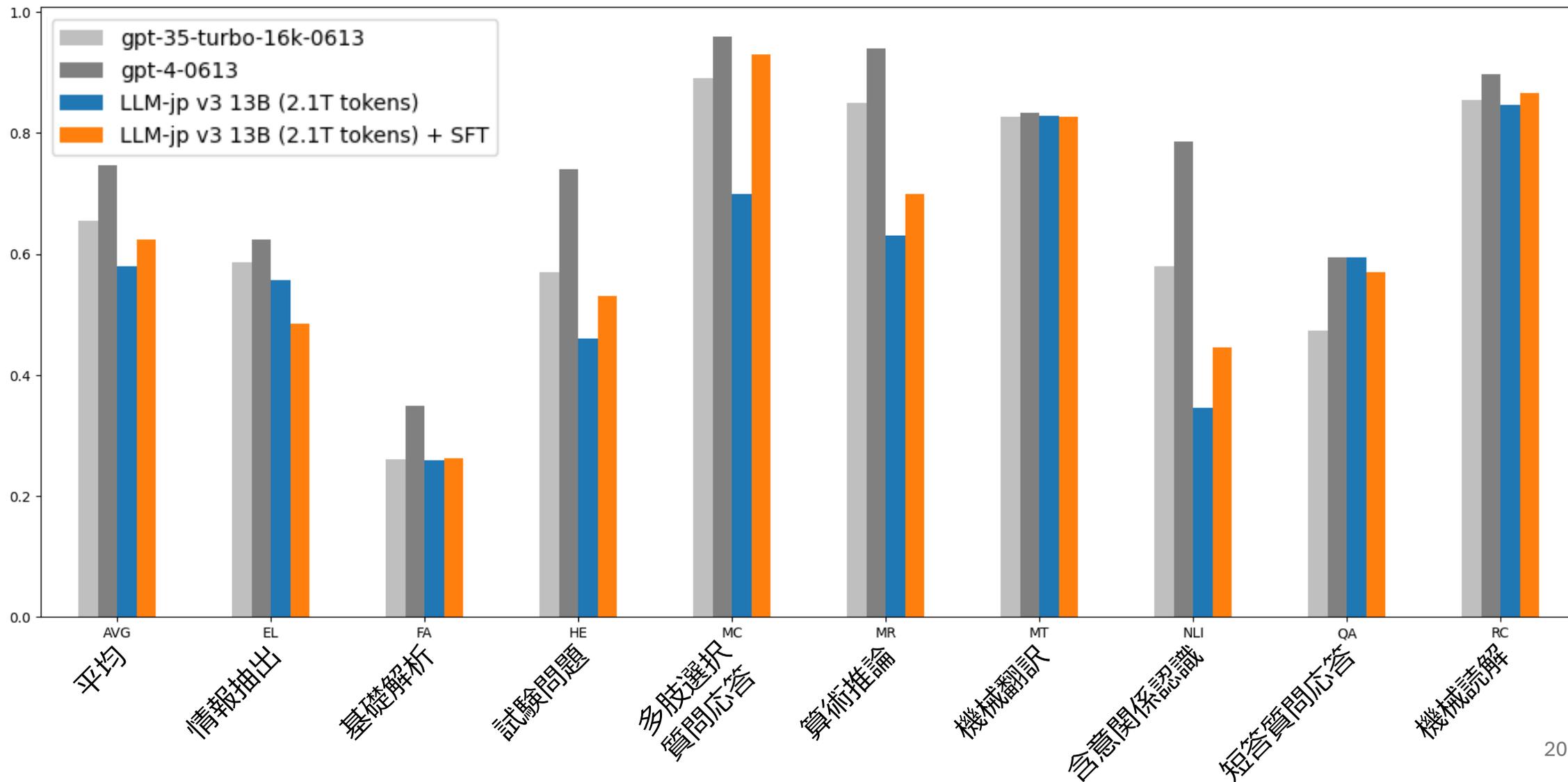
評価ベンチマーク (llm-jp-eval)

- 質問応答、算術推論、翻訳、要約など11カテゴリ、計6.4万インスタンス

モデル学習



- **llm-jp-3 13B** と **llm-jp-3 13B instruct** の性能 (llm-jp-eval 1.3)
- 平均値でGPT3.5に近い、いくつかのタスクではGPT4.0相当



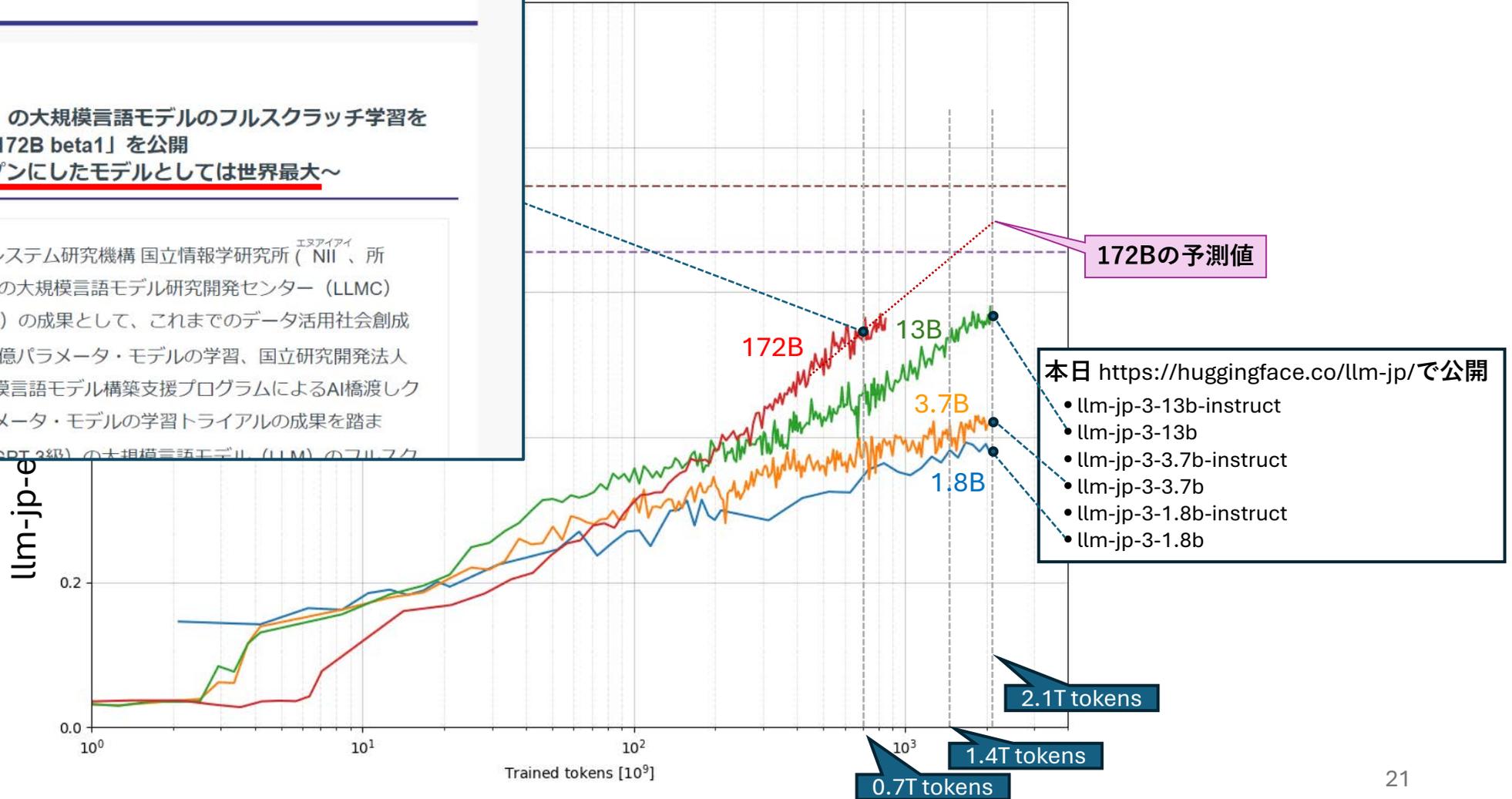
モデル学習

ニュースリリース

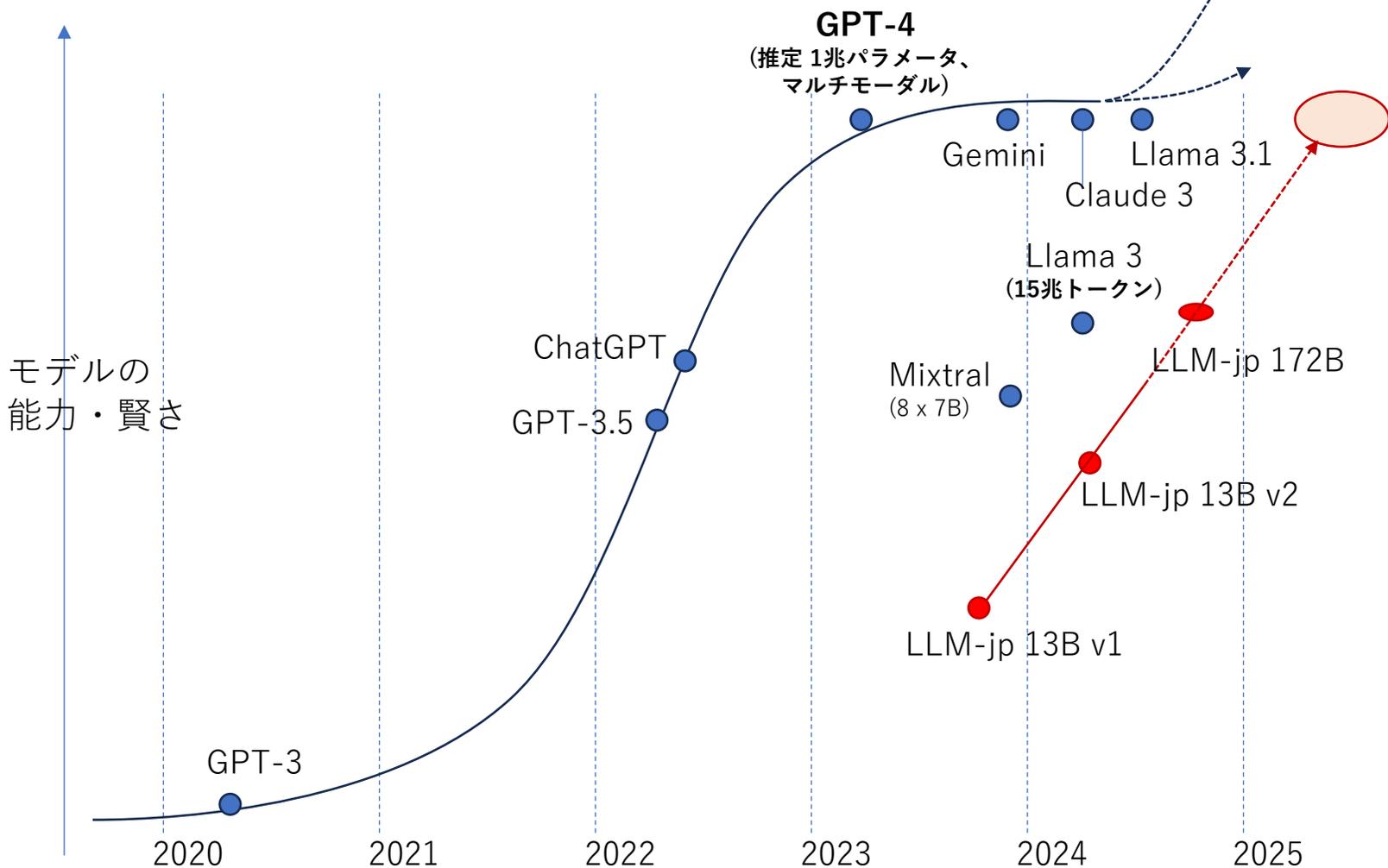
2024/09/17

約1720億パラメータ（GPT-3級）の大規模言語モデルのフルスクラッチ学習を行い、プレビュー版「LLM-jp-3 172B beta1」を公開
～学習データを含めすべてオープンにしたモデルとしては世界最大～

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 (NII、所長: 黒橋 禎夫、東京都千代田区) の大規模言語モデル研究開発センター (LLMC) は、主宰するLLM勉強会 (LLM-jp) の成果として、これまでのデータ活用社会創成プラットフォームmdx^{(*)1}での130億パラメータ・モデルの学習、国立研究開発法人産業技術総合研究所の第2回大規模言語モデル構築支援プログラムによるAI橋渡しクラウド (ABCI) での1750億パラメータ・モデルの学習トライアルの成果を踏まえ、パラメータ数^{(*)2}約1720億 (GPT-3級) の大規模言語モデル (LLM) のフルスク



今後のビジョン



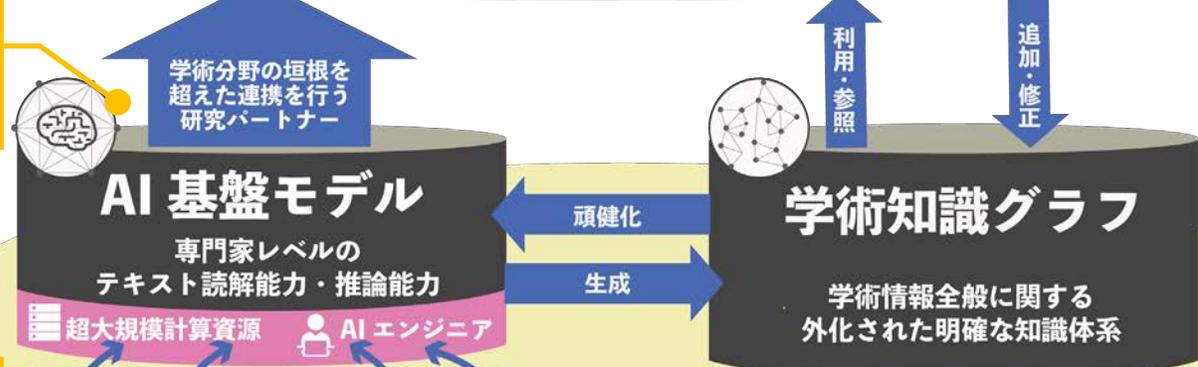
- 透明性・信頼性・安全性の向上
- ハルシネーション対策
- LLMの原理解明・高度化
- MoE(Mixture of Experts)など新たなアーキテクチャの検証
- マルチモーダルモデルの学習
- アジア圏の諸言語に拡張
- Github、HuggingFace、NII GitLab上でソースコード、モデル、データセットなどを公開
- AISI (AI Safety Institute)との協業
- LLM-jp等を通じた知見の共有

データ基盤から知識基盤へ

日本学術会議「未来の学術振興構想」の策定に向けた「学術の中長期研究戦略」に提案（2022年12月16日）



AI基盤モデルが出力する情報の信頼性を担保する知識トレーサビリティ



AI基盤モデル構築に必要なデータの信頼性・信憑性を確保する技術



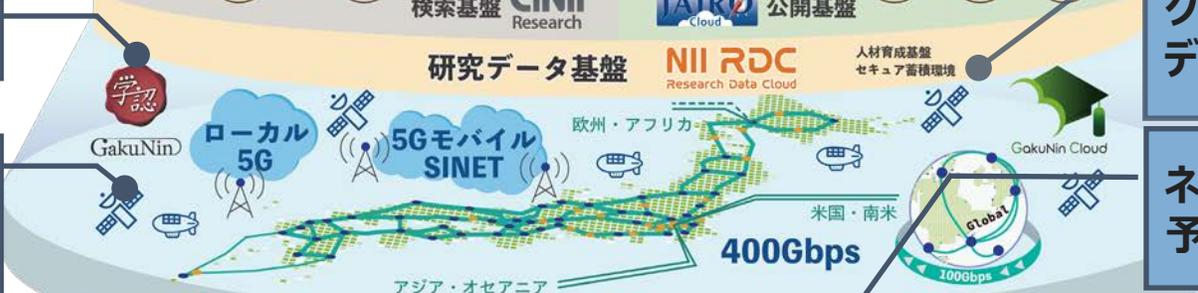
研究データ基盤の機能充実

異分野間の高度認証連携



クラウド・エッジサーバ・デバイス間のデータ収集及び資源最適化

国際協調に基づく非地上形ネットワークの構築による学術分野の発展・開拓



ネットワーク状態の高度診断・障害予兆検知

If you want to go **fast**, go **alone**

If you want to go **far**, go **together**



LLMC 研究員、RA 募集継続中

LLM-jpへの参加 いつでもwelcome