日本語に強い大規模言語モデルの開発のためのコーパス構築

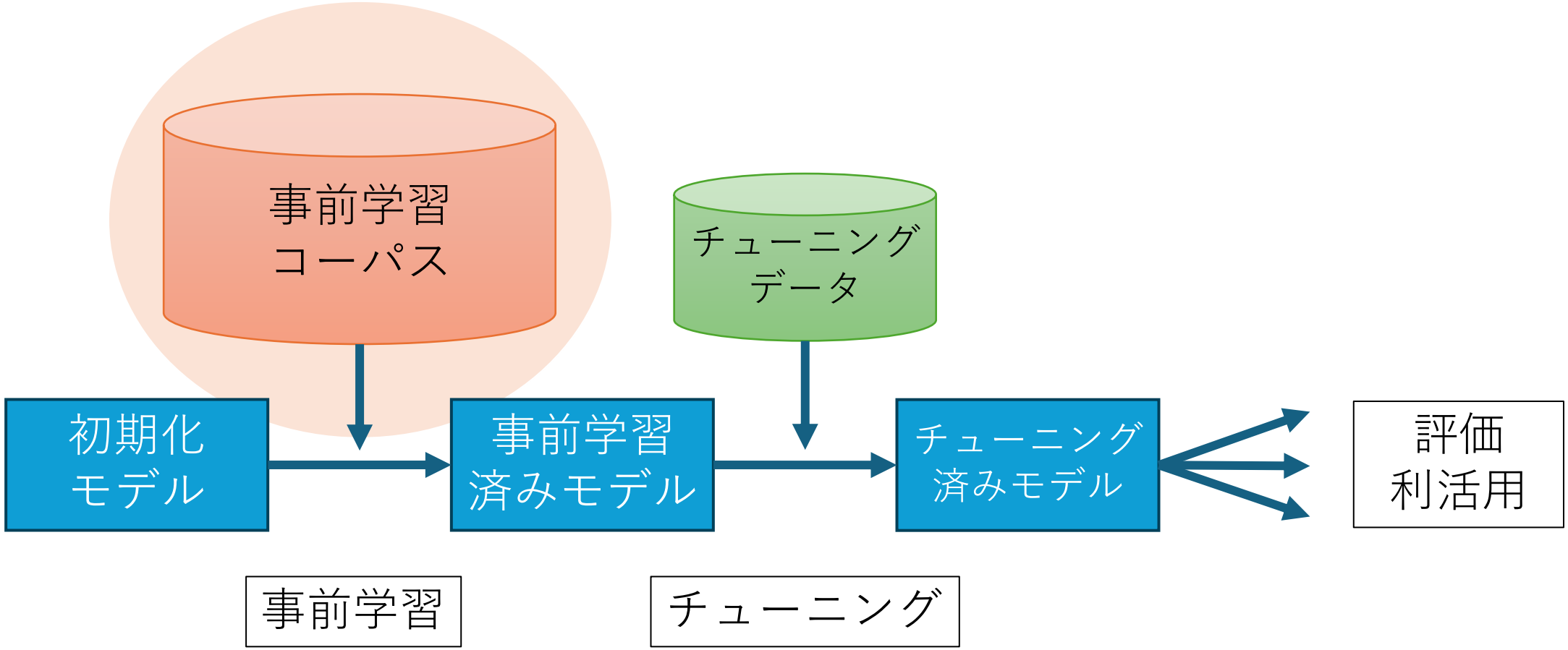
河原 大輔

早稲田大学，NII LLMセンター

LLM-jp コーパス構築WG

LLMシンポジウム2024「大規模言語モデルのあるべき姿とは？」(2024/9/25)

大規模言語モデル(LLM)構築の流れ



事前学習コーパス構築の課題

- 強力なLLM構築のためには良質かつ大量のテキストが必要
  - Meta Llama: 15兆トークン
  - Microsoft Phi: 教科書レベルの質のテキスト

  } 質と量のトレードオフ
- 日本語のコーパス候補
  - 日本語Wikipedia
  - WebアーカイブCommon Crawl (CC)に含まれる日本語ページ
  - 書籍、論文、特許文書など
- 課題
  - 大規模な日本語コーパスをどのように入手し整備していくか?
  - どの程度の質を求め、どのようなフィルタリングをすべきか?
  - 英語や他の言語の最適な混合比は?
  - 日本語を含む多言語に最適なトークナイザとは?
  - LLMの生成が著作物に酷似している可能性は?

LLM-jpコーパス構築の軌跡

LLM-jpコーパスv1 (2023/8)

- 多言語Webコーパス mC4の日本語部分をフィルタリング
  - ドメインフィルタ(レトリバ社)
  - 有害ワードフィルタ(LINE社)

言語	サブコーパス	トークン数
日本語	Wikipedia	1B
	mC4	136B
英語	Wikipedia	5B
	Pile	176B
コード	Stack	148B

https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v1

LLM-jpコーパスv2 (2023/12)

- Webアーカイブ CC全量から抽出した日本語テキスト
  - 言語モデルフィルタリング
  - 重複除去
    (ワークスアプリケーションズ社Uzushio)

言語	サブコーパス	トークン数
日本語	Wikipedia	1B
	CC	285B
英語	Wikipedia	5B
	Pile	176B
コード	Stack	148B

https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v2

LLM-jpコーパスの構築: 最近～今後の取組

LLM-jpコーパスv3 (2024/7)

- 国立国会図書館 (NDL) WARPから提供されたURLを基にクロール
- KAKEN研究課題の概要テキスト

言語	サブコーパス	トークン数
日本語	Wikipedia	1B
	CC	380B
	NDL PDF/HTML	207B
	KAKEN	1B
英語	Wikipedia	5B
	Dolma	945B
他言語	中韓Wikipedia	1B
コード	Stack	114B

https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

現在・今後の取組

- コーパス開拓 (日本語3兆トークン目標)
  - 論文などの科学技術テキスト
  - 書籍、雑誌などの出版物
  - 画像、映像を含むマルチモーダルデータ
- LLMの生成テキストから事前学習コーパスを検索、分析
  - LLMの暗記と忘却の分析
  - 生成テキストの自動ファクトチェック
- フィルタリングの改良
  - 質と量のトレードオフに関する分析
  - 有害文書フィルタリングの適用強度の検討 [安全性WGと連携]
- 英語や他の言語の最適な混合比の検討

事前学習コーパスの分類

- 事前学習に使用
  - L1:学習+検索+配布　コーパス検索時に原文表示可、配布可
  - L2:学習+検索　コーパス検索時に原文表示可、配布不可
  - L3:学習　原文表示不可、メタ情報のみ表示、配布不可
- その他
  - LX:非学習　事前学習に使用しない
  - LZ:不使用　LLM-jpで使用しない

コーパス名	利用レベル	言語（LLM-jp内）	ドメイン
Common Crawl	L1: 学習+検索+配布	日本語	Web
Dolma	L1: 学習+検索+配布	英語,コード	Web,書籍,GitHub
Stack	L1: 学習+検索+配布	コード	GitHub
Wikipedia	L1: 学習+検索+配布	英語,日本語,中国語,韓国語	辞書
NDL WARP	L1: 学習+検索+配布	日本語	Web
Kaken	L1: 学習+検索+配布	日本語	学術
日本語ウェブコーパス2010	L1: 学習+検索+配布	日本語	Web
Ceek News	L3: 学習	日本語	ニュース
Twitter/X	L3: 学習	英語,日本語	SNS
国語研日本語ウェブコーパス	L2: 学習+検索	日本語	Web
NDL DC インターネット公開資料	L1: 学習+検索+配布	日本語	図書など
J-STAGE	L3: 学習	日本語	論文

コーパス開拓

- クロールデータの拡充
  - 国語研日本語ウェブコーパス: 約6億ページ
  - NDL WARPから提供されたURLを基にクロール (残り): 約10億ページ
- 科学技術テキスト
  - J-STAGE論文: 520万PDF
- 書籍、雑誌などの出版物
  - NDL デジタルコレクション インターネット公開資料: 35万件
    - 著作権保護期間満了となった図書
    - 古典籍
- マルチモーダルデータ [マルチモーダルWGと連携]
  - 画像とテキストのインターリーブデータ
  - 映像データ

言語モデルの暗記の分析 (1/2)

言語モデルが訓練データをどのくらい暗記しているか検証

- 暗記の判定：訓練データの一部からその続きを抽出できるか (Ippolito et al., 2023)

  例：第九条 日本国民は、正義と秩序を基調とする国際平和を誠実に希求し、国権の発動たる...

  この部分を言語モデルに入力したとき　　この部分が抽出されるか（BLEU >= 0.75）調べる

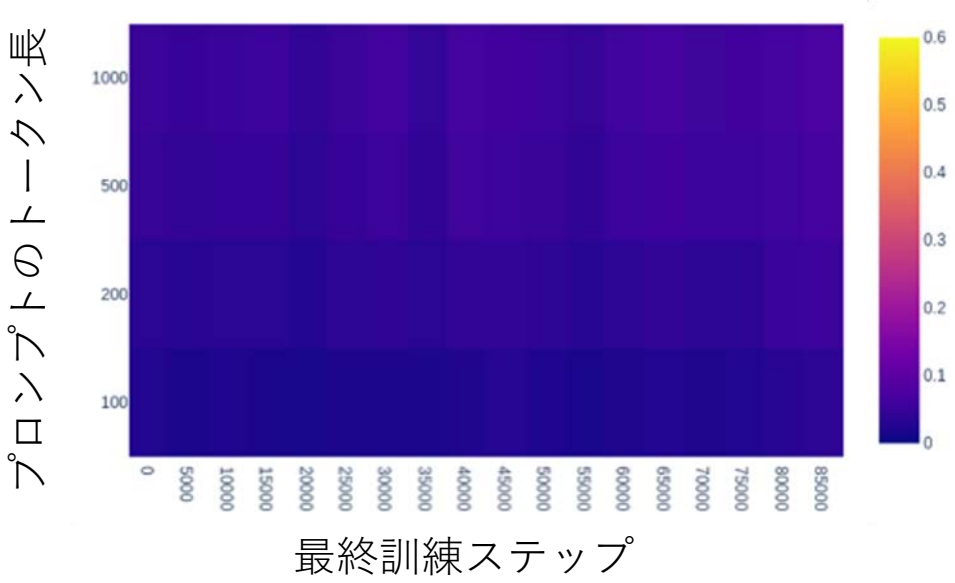
LLM-jp 13B v1.0 を対象に以下の観点から暗記を調査：

- 頻度 (Carlini et al., 2023)：何度も出てくるデータほど続きを抽出しやすい？
- プロンプト長 (Carlini et al., 2023)：プロンプトが長いほど続きを抽出しやすい？
- 最終訓練ステップ：終盤に学習したデータほど続きを抽出しやすい？

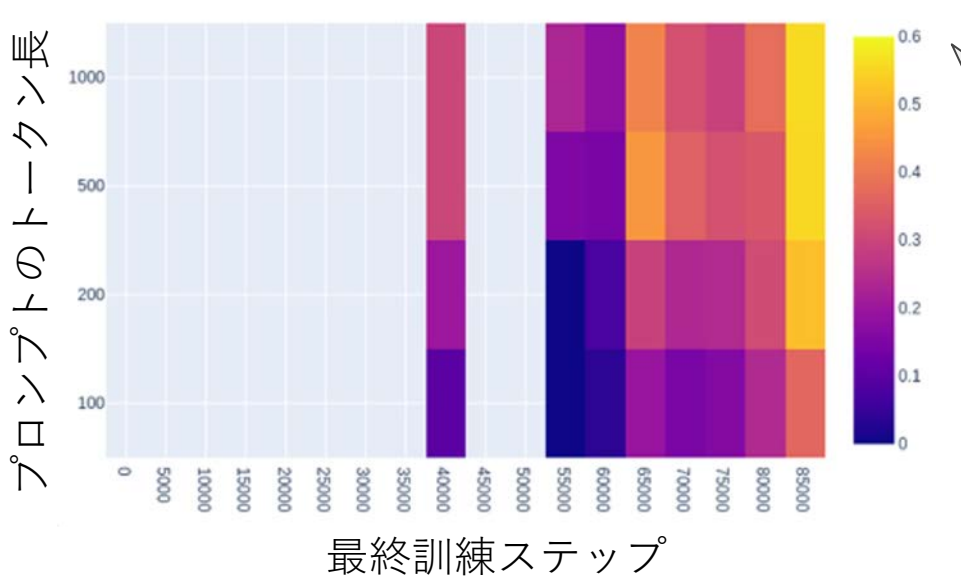
発表論文: Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi.
A Comprehensive Analysis of Memorization in Large Language Models.
In Proc of the 17th International Natural Language Generation Conference (INLG 2024), 2024.9. [PDF]

言語モデルの暗記の分析 (2/2)

低頻度（頻度: 1〜10）の訓練データの暗記割合



高頻度（頻度: 10〜100）の訓練データの暗記割合



一番明るいグリッドでは60%の訓練データが抽出される

- 頻度：高頻度の訓練データほど抽出しやすい（低頻度データの暗記はまず発生しない）
- プロンプト長：プロンプトが長いほど訓練データを抽出しやすい
- 最終訓練ステップ：終盤の訓練データほど抽出しやすい（高頻度データもしばらく経つと忘れる）

事前学習コーパスの検索

- 事前学習コーパスの検索システムを開発中；これにより：
  - 生成テキストの根拠と思われる文書を提示し、ユーザが生成テキストの信頼性を評価するのを補助
  - 生成テキストに事前学習コーパス中の著作物と強く関連のある文章が含まれる場合に適切なクレジットを表示
  - 事前学習コーパスの側面からLLMの最重要課題の一つであるHallucinationを分析

- これが出来るのは事前学習コーパスまで含めてオープンだから
  - GPT-4 のようなクローズドなモデルはもちろん、Llama 3のようにモデルパラメータはオープンだが事前学習コーパスはクローズドなモデルでは出来ない

事前学習コーパスの検索：実際の例

@llmjp-13b-finetuned 地震の原因はナマズが動くことですか？

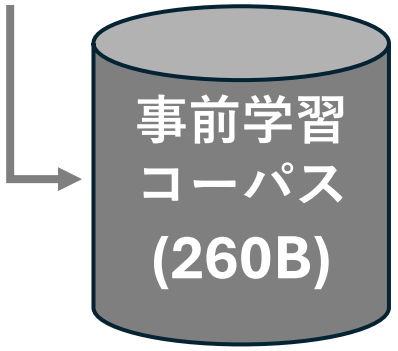
1 reply

llmjp-13b-finetuned APP 3 hours ago

いいえ、地震はナマズのせいではありません。地震は、地殻に蓄積されたエネルギーが突然解放されることで起こる自然現象です。地下の岩盤が、圧力に耐えきれずにずれ始めると、断層に力がかかり、地震が発生す

地震は通常、海底の隆起や火山の

根拠と思しき文書がどのデータセットに含まれていて、いつ学習されたかが分かる



[モデル構築WGと連携]

Dataset: Wikipedia #1
URL: https://ja.wikipedia.org/wiki/地震
Training step: 28,955 (/61,000)
地震の原因論とメカニズム論の展開 神話など日本（大和民族）では古来より「地中深くに大ナマズが存在し、その大ナマズが暴れることにより大地震が起きる」という俗説が信じられていた。現代においてもよく知られた俗説だが、ナマズが地震を予知できる根拠は見つかっていない[...]

Dataset: Common Crawl #2
URL: https://gbank.gsj.jp/wellweb/GSJ/kaisetsu/science/science.html
Training step: 168 (/61,000)
日本は世界有数の地震国です。そこで誰でも考えるのは、「地震発生が事前にわかればあれほどの被害は出ないのに」ということ。つまり地震予知です。この地震予知に関しては昔からいろいろな事が言われてきました。例えば「地震の前にはナマズが暴れる」、「地震の前には[...]

...

まとめ

- LLM学習の基盤となる事前学習コーパスを構築、公開
  - LLM-jpコーパスv3: https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3
    (日本語590Bトークンを含む合計1.7Tトークン)
- 日本語コーパスをさらに開拓中
  - 論文などの科学技術テキスト
  - 書籍、雑誌などの出版物
  - 画像、映像を含むマルチモーダルデータ
- LLMの透明性・信頼性の向上に向けた取組
  - LLMの生成テキストから事前学習コーパスを検索、分析