

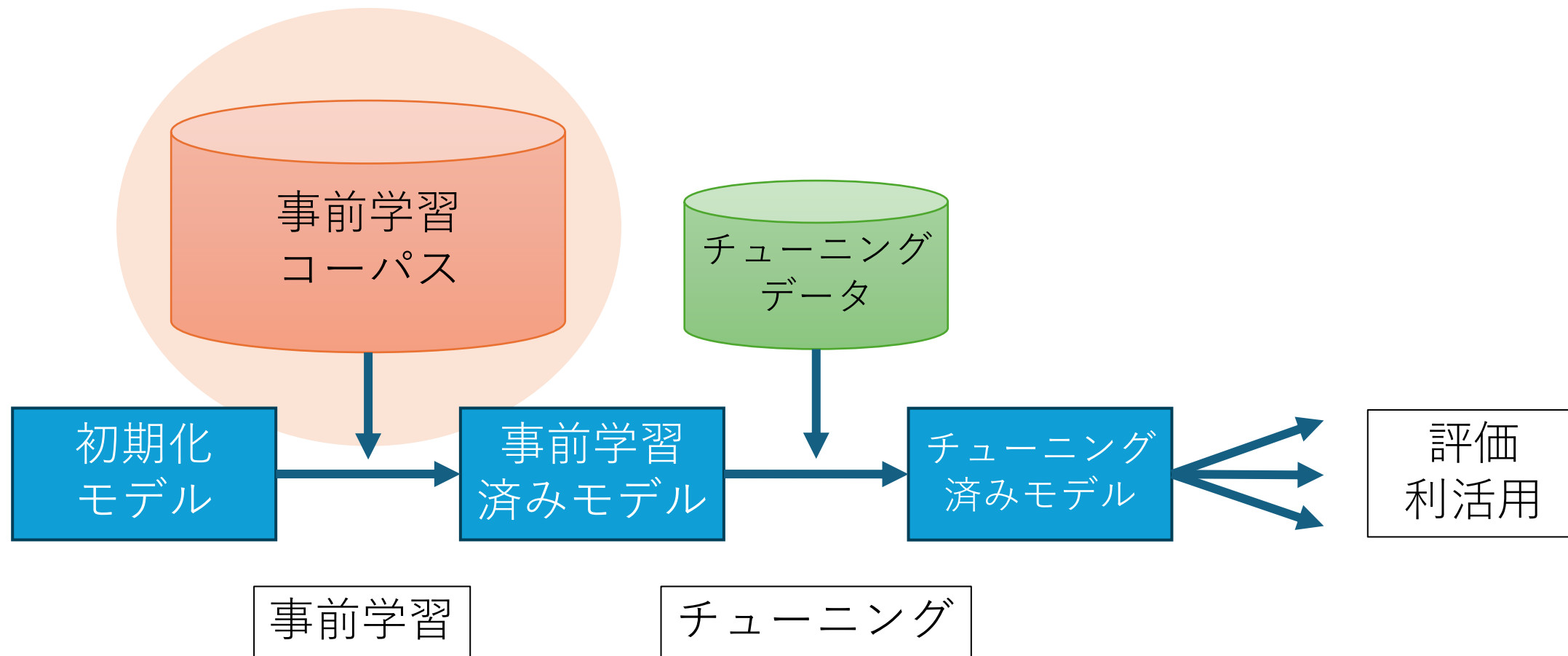
日本語に強い大規模言語モデルの 開発のためのコーパス構築

河原 大輔

早稲田大学, NII LLMセンター

LLM-jp コーパス構築WG

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の構築のためには**良質かつ大量のテキストが必要**
 - Meta Llama, Alibaba Qwen: 30兆トークン以上
 - 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
 - 図書・雑誌、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の生成テキストが著作物に酷似している可能性は?

コーパス構築WGの取り組み

- コーパス開拓（日本語3兆トークン目標）
 - **クロールデータ**
 - **図書・雑誌などの出版物、古典籍資料、論文**
 - 合成データ
- 事前学習コーパスの最適化
 - コーパスフィルタリングの改良
 - トークナイザの改良
 - 有害文書フィルタ、要配慮個人情報フィルタの開発
 - **サブコーパス比率の最適化**
 - **段階的な事前学習の導入**
- 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
 - **生成テキストの根拠の分析**

LLM-jpコーパスv4の構築 (2025/6公開)

- 日本語コーパスの拡充
 - Webクロールデータ
 - [日本語ウェブコーパス 2010](#), [国語研日本語ウェブコーパス](#), [Ceek.jp News](#), [FineWeb 2](#), [SIP3日本語汎用ウェブコーパス](#)
 - 特許、法律、国会議事録など
- 日本語コーパスの質的改善
 - サブコーパスごと、および、コーパス全体での類似重複除去
 - PDF系コーパスのさらなるクリーニング
- 英語・コードコーパスの拡充
 - [FineWeb \(Edu\)](#)
 - [OLMo 2 1124 MIX](#)
 - 論文(arXiv)、数学(OpenWebMath, Algebraic Stack)、コード(StarCoder)など
- 他言語コーパスの拡充
 - [FineWeb 2](#) 中国語・韓国語

言語	サブコーパス	トークン数
日本語	Wikipedia, CC, NDL PDF/HTML, KAKEN, NWC2010, NWJC, Ceek.jp, FineWeb 2, SIP Web, 特許, 法律, 国会議事録	0.7T (日本語全体で重複除去)
英語	Wikipedia, FineWeb, 論文, 数学	17.7T
他言語	中韓Wikipedia, FineWeb 2	0.85T
コード	Stack, StarCoder	0.2T

合計: 19.5T

<https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v4>

さらなるコーパスの開拓

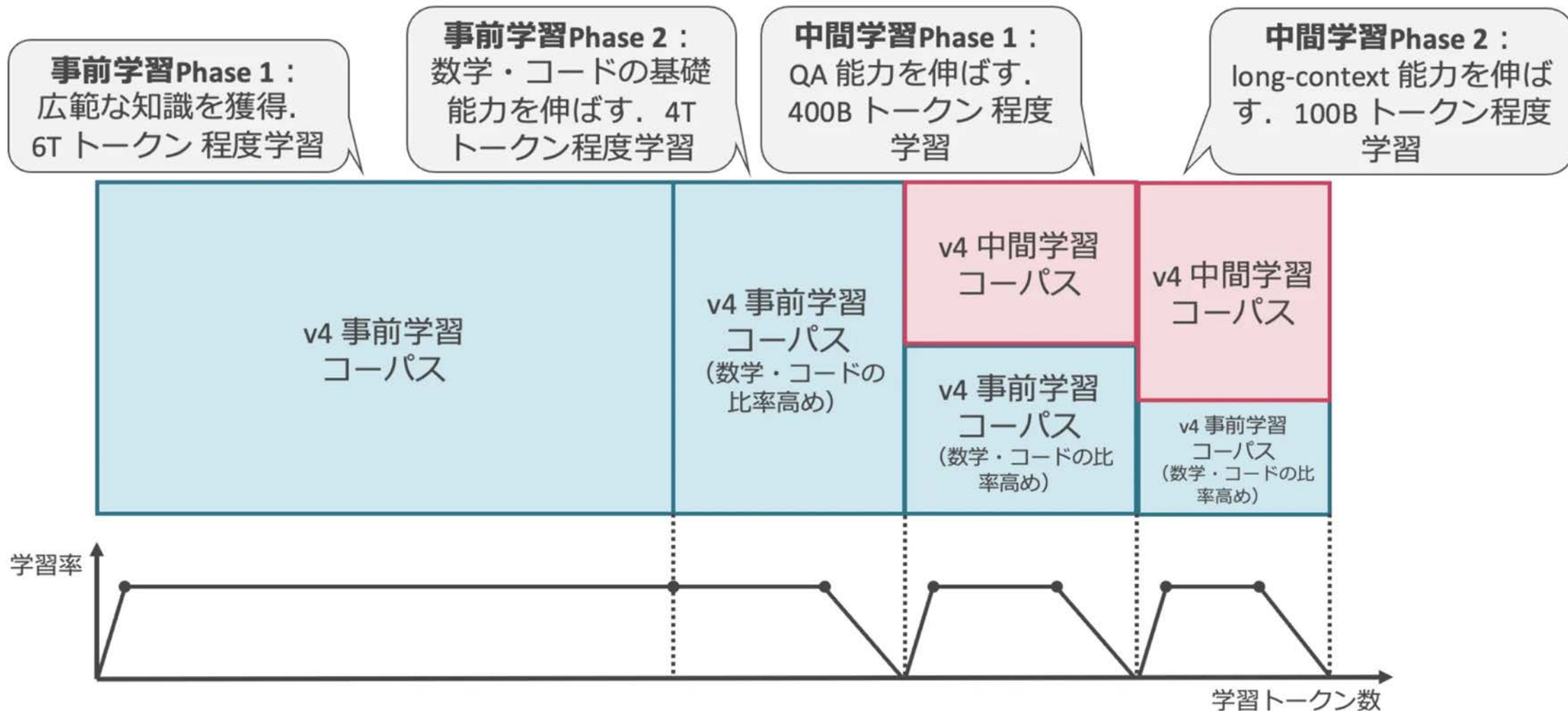
- 図書・雑誌などの出版物、古典籍資料
 - 国立国会図書館 官庁出版物テキスト
 - 主に1995年までに刊行された図書のほか、雑誌、官報を含め、合計約30万点のOCRテキスト
 - L3: 学習
 - 国文学研究資料館 古典籍データ
 - 画像・OCRテキストペア300万件
 - L2: 学習+検索
- 科学技術テキスト [学術ドメインWGと連携]
 - J-STAGE論文 (約500万PDF)
- マルチモーダルデータ [マルチモーダルWGと連携]
 - 画像とテキストのペア、インターリーブデータ
 - 映像データ



『たけとり物語』(国文学研究資料館所蔵)
出典: 国書データベース, <https://doi.org/10.20730/200017275>

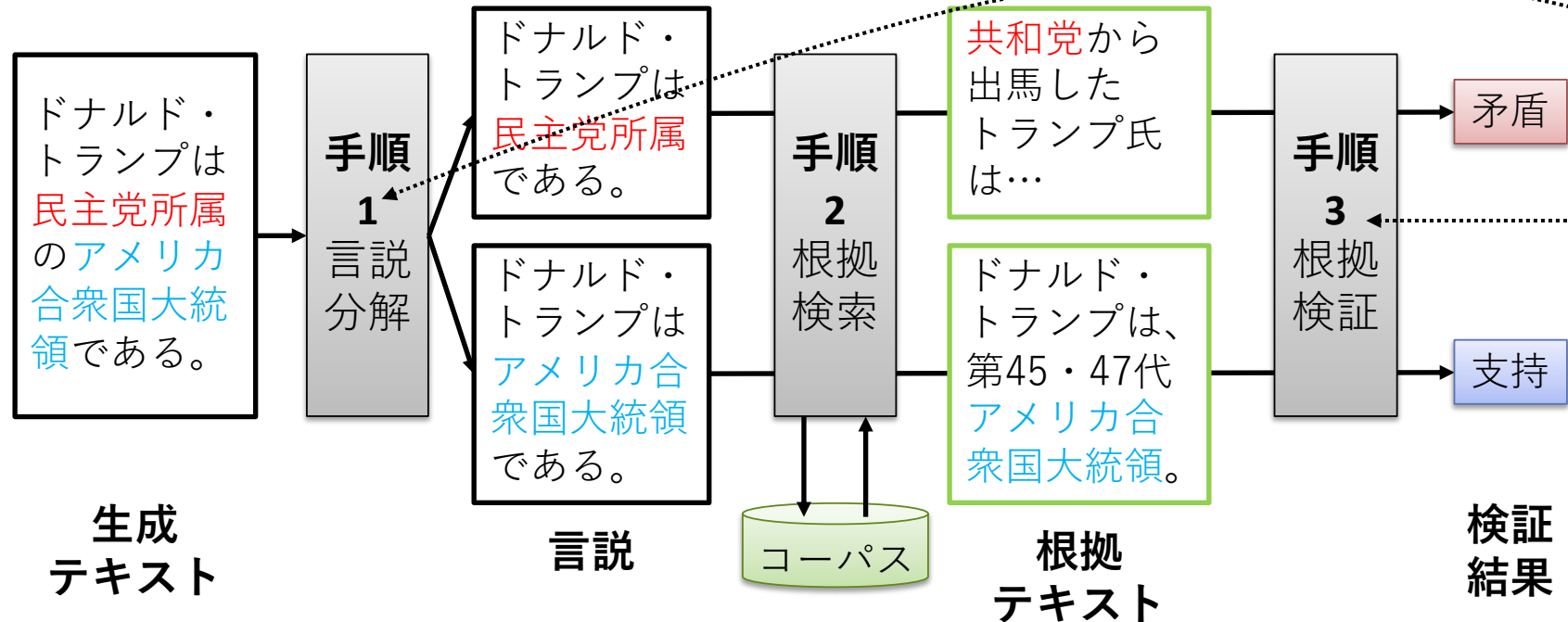
LLM-jp-4モデル事前学習のためのコーパス設計

SmolLM3 [[Bakouch+ 2025](#)]などを参考に、事前学習を段階的に実施



生成テキストの根拠の分析

- LLMの生成テキストにはハルシネーションが含まれる可能性
- 生成テキストの根拠を事前学習コーパスから検索し、ユーザが**生成テキストの信頼性を評価**するための補助システムが必要
- LLM-jp-3モデルの生成テキストを言説(一つの物事に関する性質や関係を表す独立した最小粒度の情報単位)に切り分け、**事前学習コーパスから根拠を検索、検証するシステムを構築**



1. 言説分解 と
3. 根拠検証 の
データセットを
構築中

言説分解

- LLMの生成テキストを言説に分解するタスク

格闘家ボブ・サップの出身国はアメリカ合衆国です。

彼は1973年9月22日にアメリカ合衆国のネブラスカ州オマハで
生まれました。

- 言説 1 : ボブ・サップは格闘家です。
 - 言説 2 : ボブ・サップの出身国はアメリカ合衆国です。
 - 言説 3 : ボブ・サップは1973年9月22日に生まれました。
 - 言説 4 : ボブ・サップはアメリカ合衆国のネブラスカ州オマハで生まれました。
- AI王データセットv2 (QA)の約400質問と LLM-jp Chatbot Arena (雑談)の約100質問に対するLLM-jp-3 13Bの応答にアノテーション

言説の根拠検証

- 言説が根拠テキストに支持されるかどうかを判定するタスク
 - **言説**：「天平文化」は聖武天皇の時代に栄えました。
 - **根拠テキスト**：ました。この文化を、聖武天皇のときの元号天平（てんぴょう）を採って「天平文化」と呼んでいます。文化史上では、この時代は「天平時代」と呼ばれています。奈良時代（天平時代）の前の日本史上の時代は、飛鳥時代。すでに飛鳥時代のページで書いた通り、飛鳥時代は文化史の面からみると、「飛鳥時代」と「白鳳時代」に分けられます。http://matiere.at.webry.info/200802/article_4.html 奈良時代といえば、「飛鳥時代」「白鳳時代」も含まれると誤解を招きそうなので、奈良時代と言うより、「天平時代」の方が妥当な呼称といえるかもしれません。**天平文化は、聖武天皇のころの文化です。**聖武天皇は、仏教によって国の平安を守ろうとし、平城京には多くの寺院が移建されたり、創建されました。飛鳥京や藤原京から移建された興福寺・大安寺・薬師寺・元興寺の四大寺と、奈良に創建された東大寺・西大寺の二大寺の六寺を総
 - **検証結果**：支持
- 言説分解のアノテーションと同じく、AI王データセットv2とLLM-jp Chatbot Arenaの質問に対するLLM-jp-3 13Bの応答(約5,000言説)にアノテーション中
- 年内に完成予定

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v4: <https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v4>
(日本語0.7Tトークンを含む合計19.5Tトークン)
- 日本語コーパスをさらに開拓中
 - 図書・雑誌などの出版物、古典籍資料
 - 論文などの科学技術テキスト
 - 画像、映像を含むマルチモーダルデータ
- 事前学習コーパスの最適化
- LLMの透明性・信頼性の向上に向けた取り組み
 - LLMの生成テキストから事前学習コーパスを検索し、根拠を分析