

# 知識基盤の構築

国立情報学研究所 所長  
黒橋 禎夫

2026年4月24日

# AIで新発見が続出

Rebalancing the Gut: How AI Solved a 25-Year Crohn's Disease Mystery

AIが「スーパー耐性菌」に有効な抗生物質を設計 淋菌とブドウ球菌で

医学・生命科学での新発見

長年の数学未解決問題の解決

フィールズ賞のテレンス・タオ氏、「GPT-5.2 Proが数学の未解決問題をほぼ自律的に解き切った」と評価——エルデシュ問題#728で示されたAIの新たな到達点



## AI Program Plays the Long Game to Solve Decades-Old Math Problems

February 13, 2025

February 13, 2026 Research Publication  
GPT-5.2 derives a new result in theoretical physics

In a new preprint, GPT-5.2 proposed a formula for a gluon amplitude later proved by an internal OpenAI model and verified by the authors.

Unlocking the sun's magnetic secrets: AI-powered mapping reveals intricate 3D details

宇宙の謎の解明

理論物理の問題の新しい定式化

40 Years

GPT-5.2

0

≠ 0

Key Insight: OpenAIのGPT-5.2が、40年間「ゼロ（存在しない）」とされていた物理現象（グルーオンの散乱振幅）が実は存在することを発見しました。

Impact: 人間の直感を越えた数式をAIが導出・証明した、科学史に残る転換点です。

2026年2月13日 公開

GPT-5.2が物理学の定説を覆す：40年の「ゼロ」を12時間で証明  
40年越しの物理学的ブレイクスルー



40年間の定説「振幅ゼロ」を否定  
長年不可能とされた特定のグルーオン相互作用が、実際には起こり得ることを発見。

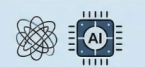
12時間

12時間で新たな公式を証明  
人間では困難な複雑なパターンの抽出と形式的な証明を、AIが短時間で完結。

人間とAIによる「共創」のプロセス



人間：基礎データの構築  
研究者が少数の粒子 (n=6) を用いた手計算を行い、AIに解析の素材を提供。



AI：汎用公式の推測と証明  
GPT-5.2が複雑な式を単純化し、あらゆる数に対応する公式を導出・証明。



「道具」から「研究」の同  
重力子の研究など、物理学最  
「万物の理論」への応用が期待

提供：情報システム研究機構

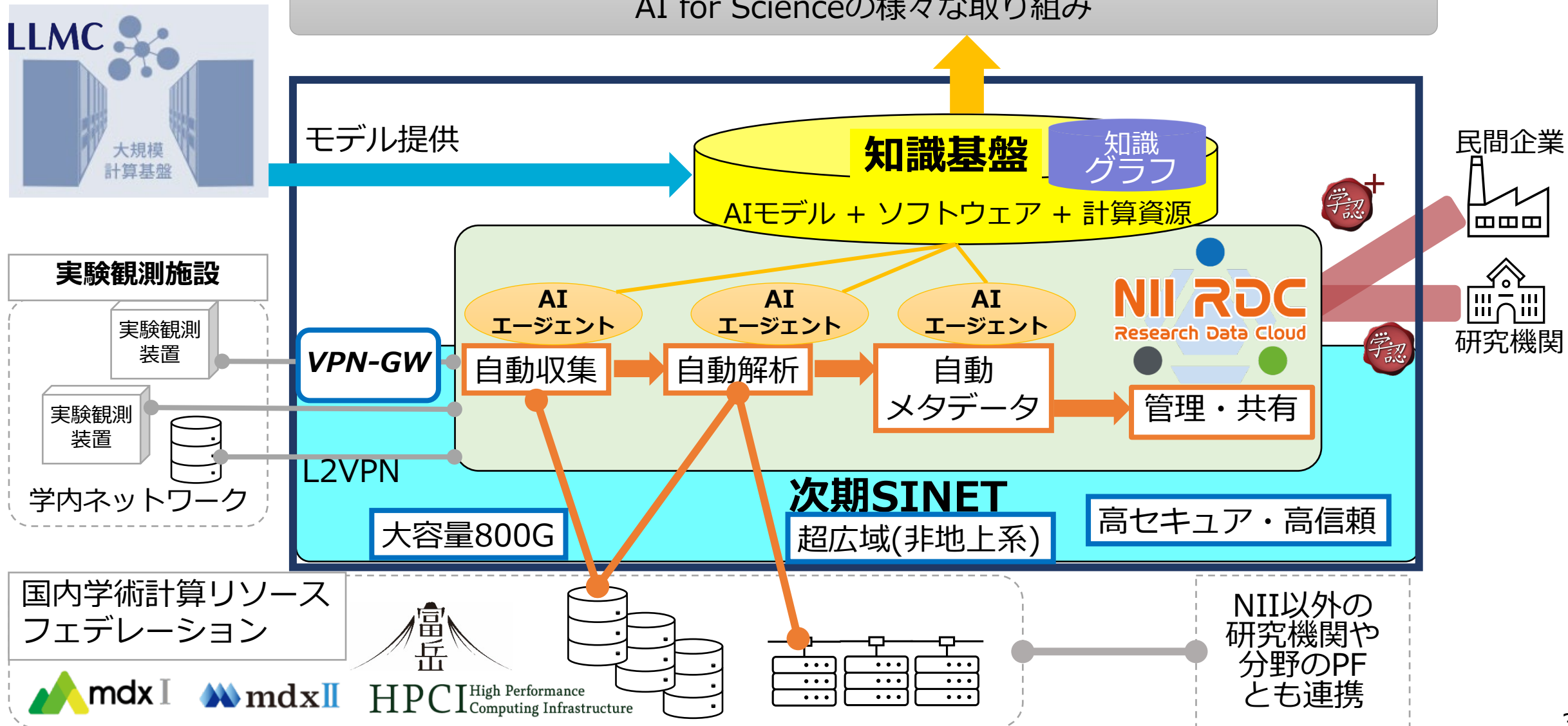
Rules of mysterious ancient Roman board game decoded by AI

考古学の新発見

# 学術研究プラットフォームの将来像

AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ(第1回) <令和7年12月24日> 資料より

AI for Scienceの様々な取り組み



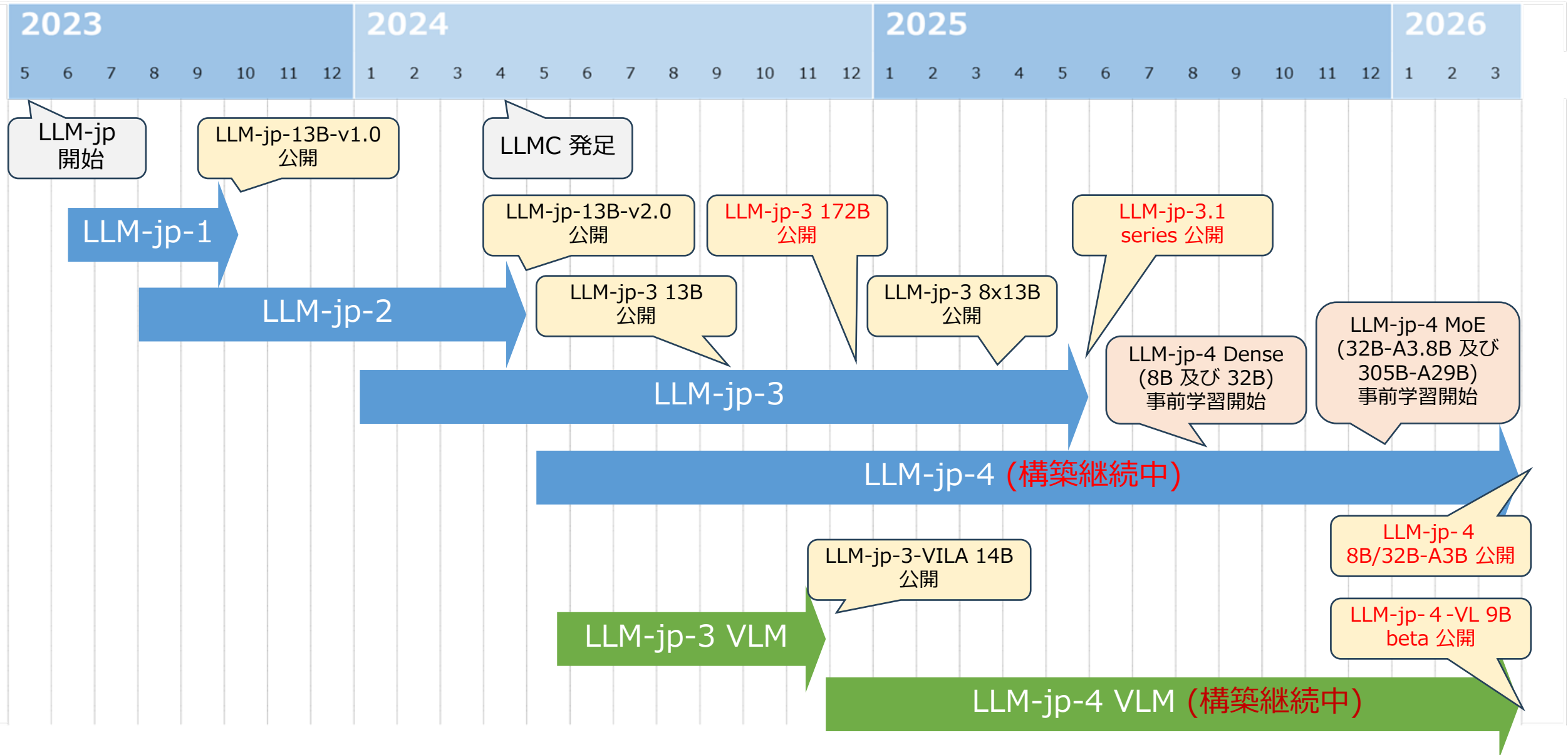
# LLM-jp/LLMC

—Team Scienceとしての生成AI研究開発—

# LLM・AIエージェントの主な発展（2025年以降）

- 2025.1 DeepSeek社が**低開発コスト**で世界トップレベルの推論能力をもつDeepSeek-R1を公開
- 2025.2 OpenAI社がGPT-4.5を発表、**Deep Research**モードを搭載
- 2025.4 Anthropic社がClaude Opus 4およびSonnet 4を公開。**Claude Code**が一般公開され、その後急速に普及
- 2025.5 OpenAI社がコーディングエージェントCodexを発表
- 2025.10 論文の執筆・査読にAIが参加する実験的国際会議 **Open Conference of AI Agents for Science**開催
- 2025.11 自律型AIエージェントOpenClaw (Clawdbot)がオープンソースとして公開
- 2026.1-2 Anthropic社がClaude Cowork、Claude Opus 4.6を公開、ソフトウェア業界に衝撃
- 2026.4 Google社がGemma 4をApache 2.0ライセンスで公開
- 2026.4 Anthropic社がClaude Mythos Previewを発表。**一般公開は見送りに**

# モデルの構築、公開



# LLM-jp-4 モデルの公開 (2026年4月3日)

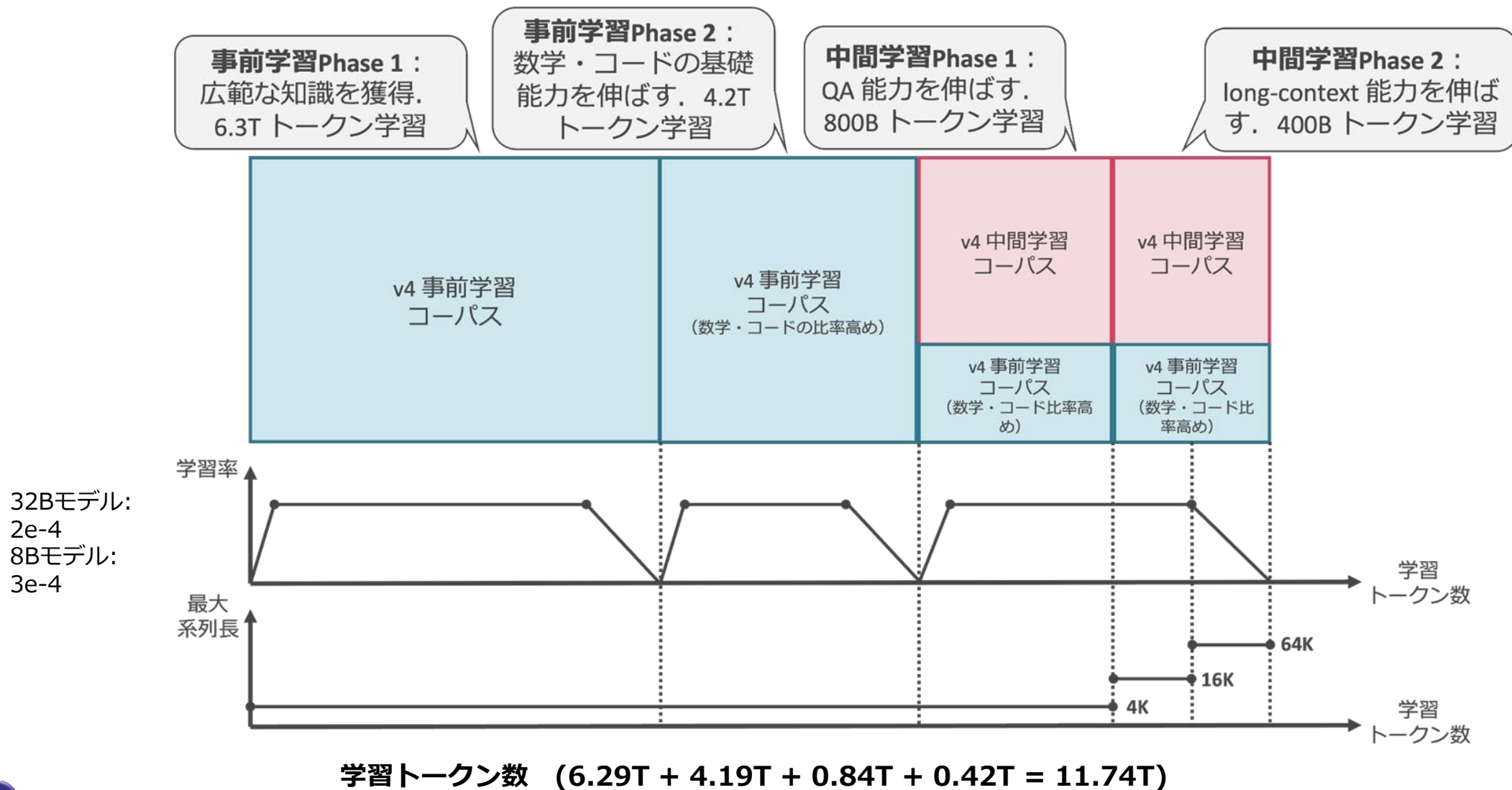
<https://www.nii.ac.jp/news/release/2026/0403.html>

The screenshot shows the NII website's news page. The header includes the NII logo (National Institute of Informatics) and navigation links for 'About NII', 'Research', 'Service', and 'Graduate Program'. The main content area is titled 'ニュース / News' and features a sidebar with a year selection menu (2018 to 2026) and a main article titled 'ニュースリリース' (News Release) dated 2026/04/03. The article headline reads: '約12兆トークンの良質なコーパスで学習した新たな国産LLM「LLM-jp-4 8Bモデル」「LLM-jp-4 32B-A3Bモデル」をオープンソースライセンスで公開～一部ベンチマークでGPT-4oやQwen3-8Bを上回る性能を達成～'. The body text, partially visible, describes the model's training on a 12 trillion token corpus and its performance relative to GPT-4o and Qwen3-8B.

- 8B denseおよび32B-A3B MoEモデルの base/thinkingの計4モデルを公開
- 日本語 MT-Bench、MT-Benchにおいて、GPT-4oやQwen3-8Bを上回る性能を達成
- 商用利用も可能なオープンソースライセンス (Apache 2.0)で提供
- 「LLM-jp-3.1」シリーズと比較して約6倍の規模となる学習コーパス(約19.5兆トークン)を構築し、うち合計約10.5兆トークンを事前学習に使用してフルスクラッチで学習
- 中間学習を実施し、事前学習コーパスに指示事前学習データ (Instruction Pre-training データ) を含む LLM による合成データを加えた、合計1.3兆トークンの学習コーパスを使用。文脈長は約6万5千トークン

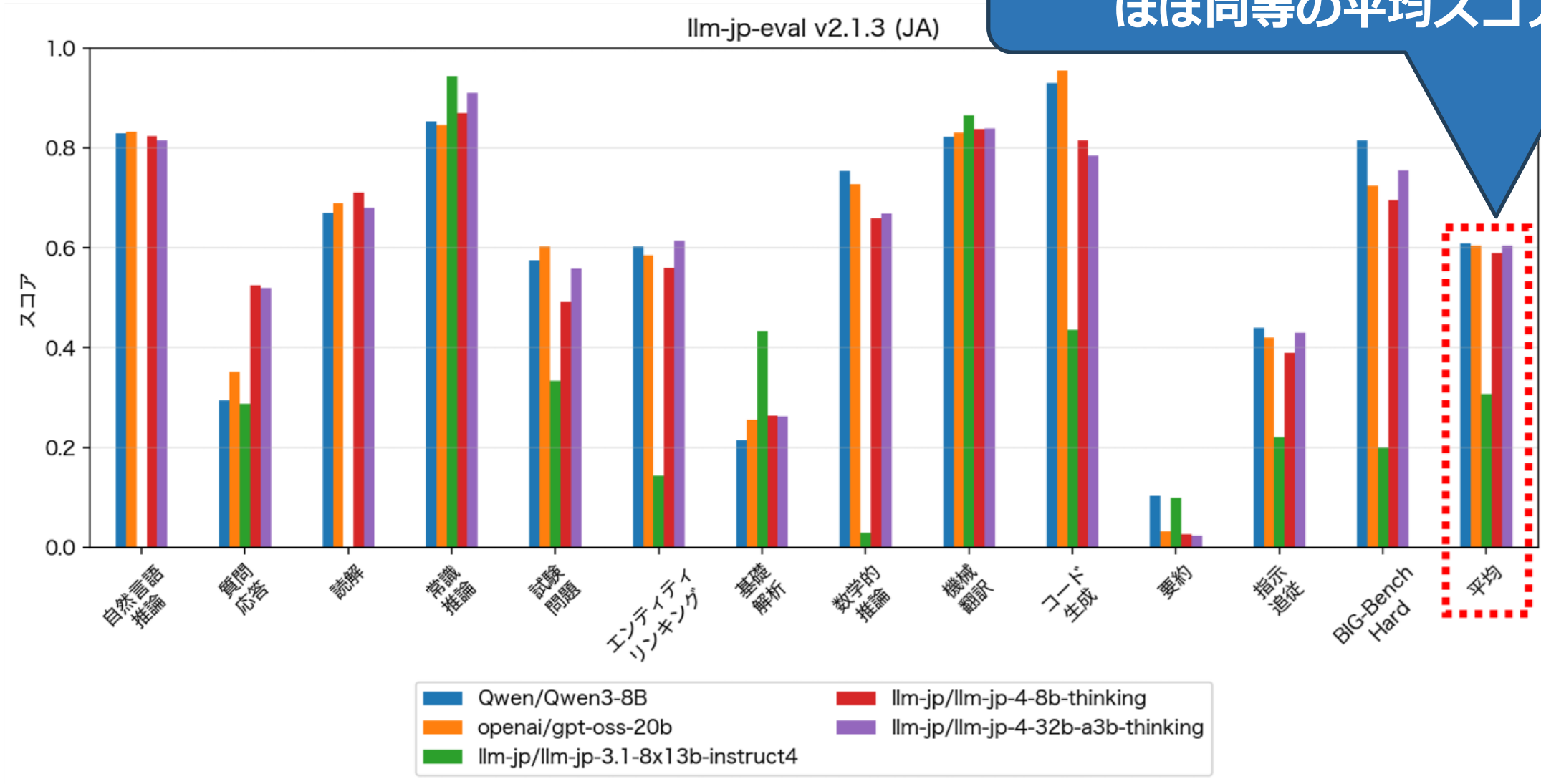
# LLM-jp-4モデルの事前学習レシピ

cf. SmoLLM3 mid-training



# ベンチマーク評価: llm-jp-eval v2.1.3

Qwen3-8B や gpt-oss-20b と  
ほぼ同等の平均スコア



# LLM-jp-4 32B-A3B Thinkingモデルの実行例(OpenWebUI)

**ローカルLLMの可能性**  
- オープンなモデルが拓くAI活用の展望 -

2025年8月4日  
株式会社日本総合研究所

<本資料に関するお問い合わせ>  
執筆者：先端技術ラボ 伊藤誠、近藤浩史  
本レポートに関するお問い合わせにつきましては、当社ホームページの [お問い合わせフォーム](#) よりご連絡ください。

<https://www.jri.co.jp/MediaLibrary/file/advanced/advanced-technology/pdf/16012.pdf>



添付資料

プロンプト

オープンなLLMはフロンティアLLMに比べてどんな長所がありますか？

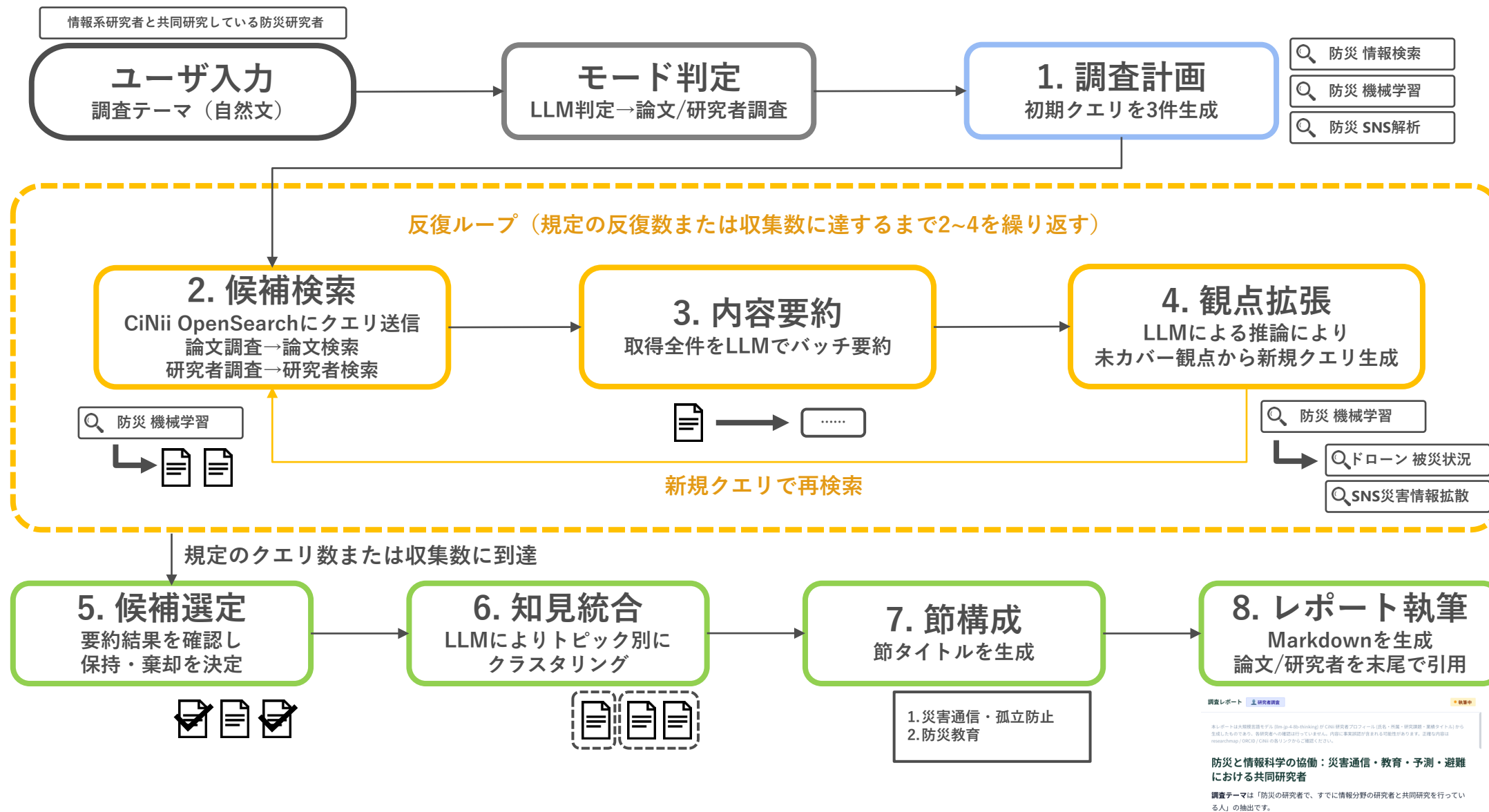
回答

オープンな大規模言語モデル（LLM）は、先進的な「フロンティア」LLM（例：商用の大規模事前学習モデル）に対して以下のような長所があります。

Web検索(3ページ取得)

長所	内容	根拠
無料または低コストで利用可能	オープンソースなので、研究・商用ともにライセンス料が不要です。	w3resource.com
自由にカスタマイズできる	モデルを自由にファインチューニングしたり、特定タスク向けに微調整できます。	w3resource.com
コミュニティ主導の改善	世界中の開発者が多数参加し、バグ修正や新機能追加が迅速に行われます。	w3resource.com
データプライバシーの制御	データは自社サーバーやローカル環境で保持でき、機密情報漏洩リスクを低減できます。	ローカルLLMの可能性.pdf（ローカルLLM利用の利点）
省スペース・省エネな小型モデル	大規模フロンティアモデルに比べて軽量版（TinyLlama 等）の開発・展開が容易です。	w3resource.com（Smaller, efficient modelsの傾向）
多様な用途へ柔軟に適応	特定産業（医療、金融、教育等）向けに最適化されたオープンLLMが多数公開されています。	w3resource.com（Industries: Healthcare, Finance, Education）

ユーザ入力 → モード判定 → 計画 → 反復収集 → 選定 → 統合 → 節構成 → 執筆



防災の研究者で、すでに情報分野の研究者と共同研究を行っている人を教えてください

検索・調査

調査例: Transformer × 日本語 NLP 2020年以降の情報検索技術 源氏物語の受容と注釈 情報検索を研究している研究者 日本文学の専門家

進行状況

経過: 5分41秒

53

収集済み

反復 2 回目

→ 10 件を選定

- ✓ 調査計画
- ✓ 候補検索
- ✓ 内容要約
- ✓ 観点拡張
- ✓ 候補選定
- ✓ 知見統合
- ✓ レポート執筆
- 完了  
完了

検索クエリ (6 件)

- 01 防災 情報検索
- 02 防災 機械学習 ビッグデータ
- 03 防災 ソーシャルメディア 解析
- 04 ソーシャルメディア 災害情報 拡散
- 05 ドローン 被災状況把握
- 06 ブロックチェーン 復旧資金管理

モデルの思考過程

reasoning\_content stream

Intro: "この節は、防災に関する最新の機械学習を用いた情報分野の手法で進展した防災研究を取り上げます。" Card: R9.

## 避難・リスクコミュニケーション・都市交通

Intro: "この節は、リスクコミュニケーションや都市交通における情報技術活用で情報分野の研究者と共同している防災研究者を取り上げます。" Card: R7.

Provide profile cards with details.

At end, ## 研究者一覧 with list from given.

Need to reference using [R1] etc when mentioning.

Let's write concise but thorough. Ensure no extra preamble. Use markdown headings.

Proceed.

検索ログ

10 件選定 53 件 / 6 クエリ

- [Q1] 防災 情報検索 10 件
- [Q2] 防災 機械学習 ビッグデータ 9 件
- [Q3] 防災 ソーシャルメディア 解析 10 件
- [Q4] ソーシャルメディア 災害情報 拡散 10 件
- [Q5] ドローン 被災状況把握 4 件
- [Q6] ブロックチェーン 復旧資金管理 10 件

調査レポート

研究者調査

.MD

.BIB

● 完成

本レポートは大規模言語モデル (llm-jp-4-8b-thinking) が CiNii 研究者プロフィール (氏名・所属・研究課題・業績タイトル) から生成したものであり、各研究者への確認は行っていません。内容に事実誤認が含まれる可能性があります。正確な内容は researchmap / ORCID / CiNii の各リンクからご確認ください。

## 防災と情報科学の協働：災害通信・教育・予測・避難における共同研究者

調査テーマは「防災の研究者で、すでに情報分野の研究者と共同研究を行っている人」の抽出です。

本レポートでは、情報技術（通信・ICT・データサイエンス）と防災課題の交差点に着目し、各節でその側面に直接貢献している代表的な防災研究者を [R1]-[R10] の番号で紹介します。

### 災害通信・情報孤立防止

この節は、災害時に情報インフラが遮断される状況で、情報分野の研究者と共同で通信・情報共有技術を開発している防災研究者を取り上げます。

#### R1. 高畑 一夫 (明海大学, 歯学部, 非常勤講師)

- 専門/主要テーマ：次世代ネバー・ダイ・ネットワークおよび認知無線通信方式の開発。災害現場でのリアルタイム情報共有を目的に、情報工学者と共同実証実験を実施。
- 代表業績：分散環境を考慮した大規模災害情報共有システムの構築と評価；岩手山防災訓練における安否情報検索システムの性能評価。
- リンク：<https://cir.nii.ac.jp/crid/1420564276175888384>

# LLM-jp-v4シリーズ（8B／32B-A3B）産業界の受け止め

2026年4月3日 NIIが公開した国産オープンソースLLMへのネット上の反応

## 総論：明確にポジティブ

学習コーパス・トークナイザーまで公開、高い透明性

## 産業界が特に評価する5点

- 日本語性能：MT-Bench 7.54～7.82（GPT-4o 7.29超）
- Apache 2.0：派生モデルの商用展開に高い自由度
- オンプレ・データ主権（金融／医療／官公庁）
- 推論コスト効率（8B単体GPU、32B-A3Bは活性3B）
- OSAID準拠の透明性

## 実装・検証を始めた企業・事例

### さくらインターネット

8BをZenn記事で自社GPU基盤（VRT/H100）上にホスティング

### GMOコネク

32B-A3Bで業務文書JSON抽出93%精度を達成・知見共有

### 株式会社一創 他コンサル系

段階導入戦略（8B→32B→332B）を発信

### 個人エンジニア／インフルエンサー層

note・Qiita・Xで検証記事・歓迎コメント多数

## 検討中のユースケース

- 中小企業：社内ナレッジRAG、議事録・契約書要約
- スタートアップ：自社サービス組み込み（追加ライセンス費不要）
- 金融・医療・官公庁：データ主権重視のオンプレ運用
- ハイブリッド構成：機密日本語処理はLLM-jp-4でローカル、複雑推論はGPT/Claude APIで併用

## 国産LLM「8Bでも世界と戦える」時代が来たかもしれない

「正直ナメてた」国産LLM—実際にHuggingFace Spaceで動かしたら認識がひっくり返った。8Bモデルが日本語MT-Benchで7.54、GPT-4oの7.29を上回る。

MoE版32B-A3Bは総320億パラメータだが推論時は38億だけ稼働し、7.82を達成。「安い計算コストで高いスコア」を実現する設計。

Apache 2.0で商用利用・改変自由。2,600名超の産学コミュニティで開発。今後32B-A31Bモデルも予定、「Claude最上位に挑む」ロードマップが本気。

**「LLM-jp-4は『国産LLMだから応援する』というレベルを超えて、実力でGPT-4oやQwen3-8Bと競合できるモデルになっています」**

— もるふお @morphox\_ai (AimanaVo / 2026.4.4)

[aimanavo.com/c/morphox\\_ai/a/yNsRmed4VTG-hg](https://aimanavo.com/c/morphox_ai/a/yNsRmed4VTG-hg)

## LLM-jp-4を「今更」と思ったのに、調べたら普通にすごかった

「国産LLMがGPT-4oを超えた」と聞いて「あ、4oレベルか」とスルーしかけた。でも調べたら認識が変わった—感覚が麻痺していたのは自分だった。

8.6BパラメータでGPT-4o（推定1.8兆パラメータ）を超えた事実。数百分の一のサイズで上回ったのは驚き。しかもApache 2.0で誰でも無料で使える。

**「日本語に強いオープンソースモデルが充実してくると、企業が自前のAI基盤を持ちやすくなります。海外サービスへのデータ送信リスクを避けながら、国内で完結した使い方ができます。これは地味ですが、じわじわ大事な変化です」**

— ハヤシタカサン (note / 2026.4.9)

[note.com/foat/n/nb220cd116b7d](https://note.com/foat/n/nb220cd116b7d)

「高性能な純国産モデルが出るのは個人的に本当に大事なこと。ベンチマーク結果はかなり良いみたいですね」

— 水谷@AIDB (X)

[x.com/mizutaniken\\_jp/status/2040115799852306591](https://x.com/mizutaniken_jp/status/2040115799852306591)

「海外LLM一強」から「国産×オープン×高性能」の新時代が、いよいよ始まりました

— 池田朋弘 (X)

[x.com/pop\\_ikeda/status/2044249355331207621](https://x.com/pop_ikeda/status/2044249355331207621)

GGUF形式のllm-jp-4-32b-a3b-thinkingを魔改造で作成。なんかそれっぽく動いているので良しとします

— あるふ@alfredplpl (X)

[x.com/alfredplpl/status/2040207663762698343](https://x.com/alfredplpl/status/2040207663762698343)

baseとthinkingモデルを統合、エキスパート数2倍・アクティブ約6BのMoEモデルを自作しました

— AI X Satoshi (X)

[x.com/AiXsatoshi/status/2040583476605522182](https://x.com/AiXsatoshi/status/2040583476605522182)

Ollamaで異常出力、vLLMでもパーサー設定見落としで出力が壊れる...最終的にgpt-oss相当の精度を安定して実現できた

— GMOコネク ト永田氏 (Qiita)

[qiita.com/ntaka329/items/7cb7e4565060b9f8691b](https://qiita.com/ntaka329/items/7cb7e4565060b9f8691b)

国産LLMがベンチマークを超える時代が来た。あとは「普通のGPUで普通に動く」までのラストワンマイル

— kenimo49 (Qiita)

[qiita.com/kenimo49/items/f7de2c47177986273494](https://qiita.com/kenimo49/items/f7de2c47177986273494)

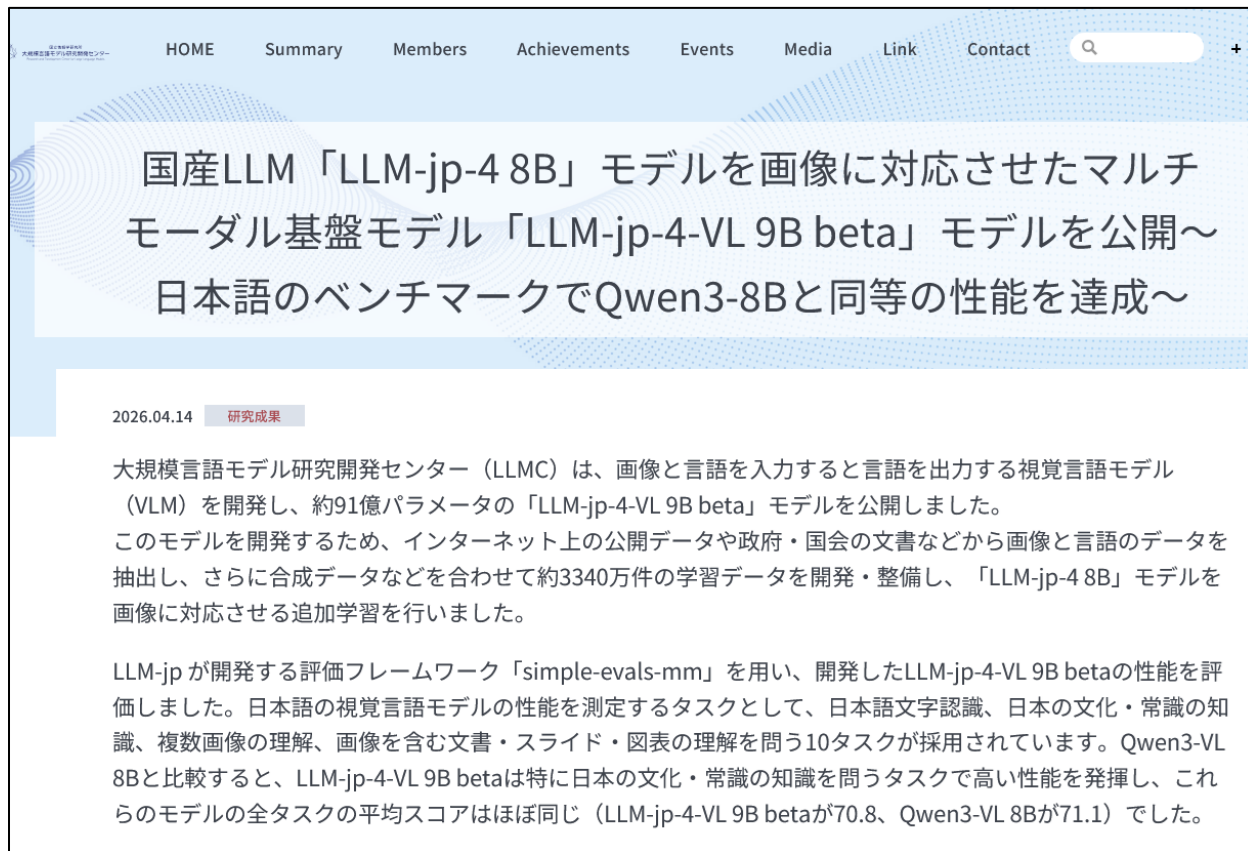
「国産だから」だけでは企業は使わない。現行オープンモデルの上をいく秀でた性能と実用性を出せるかがキモ

— Hodachi (X)

[x.com/hokazuya/status/1959389570652623041](https://x.com/hokazuya/status/1959389570652623041)

# 新たな視覚言語モデルLLM-jp-4-VL 9B beta の公開(4月14日)

<https://llmc.nii.ac.jp/topics/post-2700/>



HOME Summary Members Achievements Events Media Link Contact

## 国産LLM「LLM-jp-4 8B」モデルを画像に対応させたマルチモーダル基盤モデル「LLM-jp-4-VL 9B beta」モデルを公開～日本語のベンチマークでQwen3-8Bと同等の性能を達成～

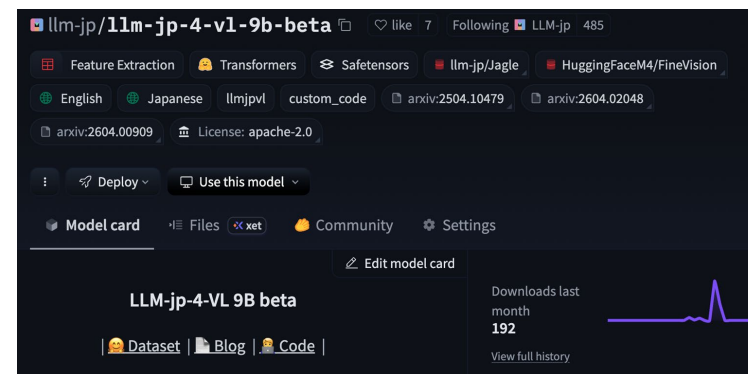
2026.04.14 研究成果

大規模言語モデル研究開発センター（LLMC）は、画像と言語を入力すると言語を出力する視覚言語モデル（VLM）を開発し、約91億パラメータの「LLM-jp-4-VL 9B beta」モデルを公開しました。このモデルを開発するため、インターネット上の公開データや政府・国会の文書などから画像と言語のデータを抽出し、さらに合成データなどを合わせて約3340万件の学習データを開発・整備し、「LLM-jp-4 8B」モデルを画像に対応させる追加学習を行いました。

LLM-jpが開発する評価フレームワーク「simple-evals-mm」を用い、開発したLLM-jp-4-VL 9B betaの性能を評価しました。日本語の視覚言語モデルの性能を測定するタスクとして、日本語文字認識、日本の文化・常識の知識、複数画像の理解、画像を含む文書・スライド・図表の理解を問う10タスクが採用されています。Qwen3-VL 8Bと比較すると、LLM-jp-4-VL 9B betaは特に日本の文化・常識の知識を問うタスクで高い性能を発揮し、これらのモデルの全タスクの平均スコアはほぼ同じ（LLM-jp-4-VL 9B betaが70.8、Qwen3-VL 8Bが71.1）でした。

- インターネット上の公開データや政府・国会の文書などから画像と言語のデータを抽出し、さらに合成データなどを合わせて約3340万件の学習データを開発・整備
- LLM-jp-4 8B instruct モデルを画像に対応させる追加学習を実施
- LLM-jpが開発する評価フレームワーク「simple-evals-mm」を用い性能評価。評価タスクは、日本語文字認識、日本の文化・常識の知識、複数画像の理解、画像を含む文書・スライド・図表の理解を問うタスクなど合計10タスク
- Qwen3-VL 8Bと比較すると、LLM-jp-4-VL 9B betaは特に日本の文化・常識の知識を問うタスクで高い性能を発揮。全タスクの平均スコアはほぼ同じ

HuggingFace上でモデルを公開  
<https://huggingface.co/llm-jp/llm-jp-4-vl-9b-beta>



llm-jp/llm-jp-4-vl-9b-beta

Feature Extraction Transformers Safetensors llm-jp/JagLe HuggingFaceM4/FineVision

English Japanese llmjpvl custom\_code arxiv:2504.10479 arxiv:2604.02048

arxiv:2604.00909 License: apache-2.0

Deploy Use this model

Model card Files xet Community Settings

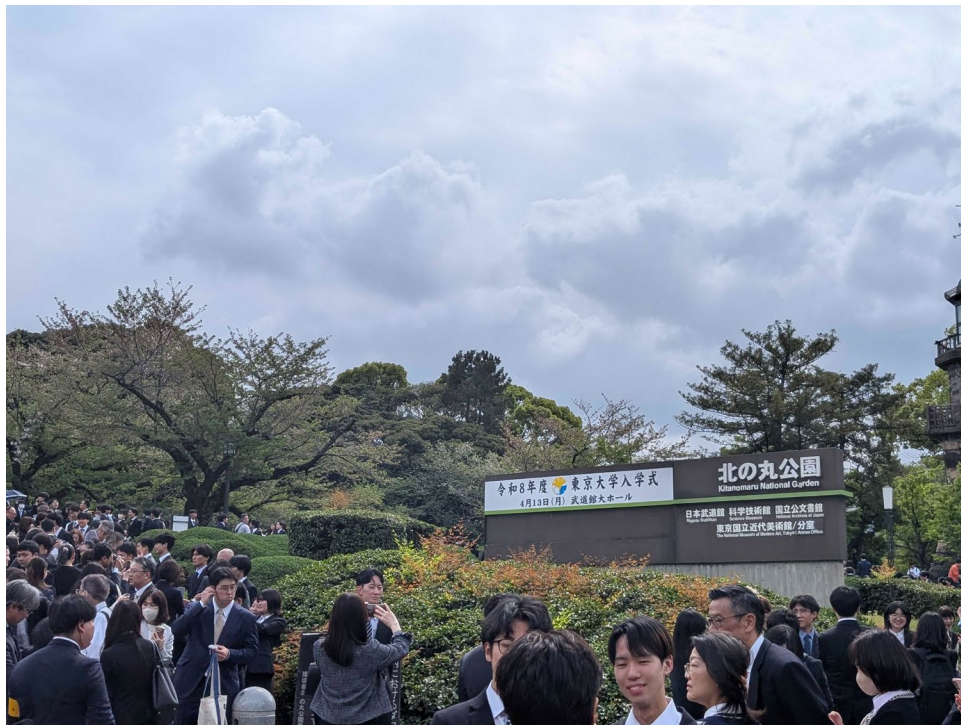
Edit model card

LLM-jp-4-VL 9B beta

Downloads last month 192

View full history

# LLM-jp-4-VL 9B betaの動作例



これは何ですか？

llm-jp-4-vl-9b-beta: 東京大学入学式



ここで釣りをするつもりです。

llm-jp-4-vl-9b-beta :残念ですが、ここは釣り禁止です。

# 新たな視覚言語モデルLLM-jp-4-VL 9B beta 詳細情報

## LLM-jp-4-VL 9B beta

- ベースLLM : **llm-jp/llm-jp-4-8b-instruct** (同時公開)
- 画像エンコーダ : google/siglip2-so400m-patch16-512
- プロジェクト: 2層のフィードフォワードネットワーク
- パラメータ数 : 約91億個 (9B)

「InternVL 3.0」モデルを参考にVLMアーキテクチャを設計。LLM対話履歴を与えるフォーマットとしてOpenAI Harmonyを拡張し、複数画像の入力に対応。Apache 2.0ライセンスで公開。

## 学習データ

以下の日英合計約3,340万事例のデータを使用。

英語: [HuggingFaceM4/FineVision](#) (約2420万事例)

日本語: [llm-jp/Jagle](#) (約920万事例)

ウェブページ、Wikipedia、日本の風景などを収めた画像コレクション、PDFデータ(政府公開データを含む)、図表などのオープンデータを収集し、画像と言語が対応付けられたデータを抽出

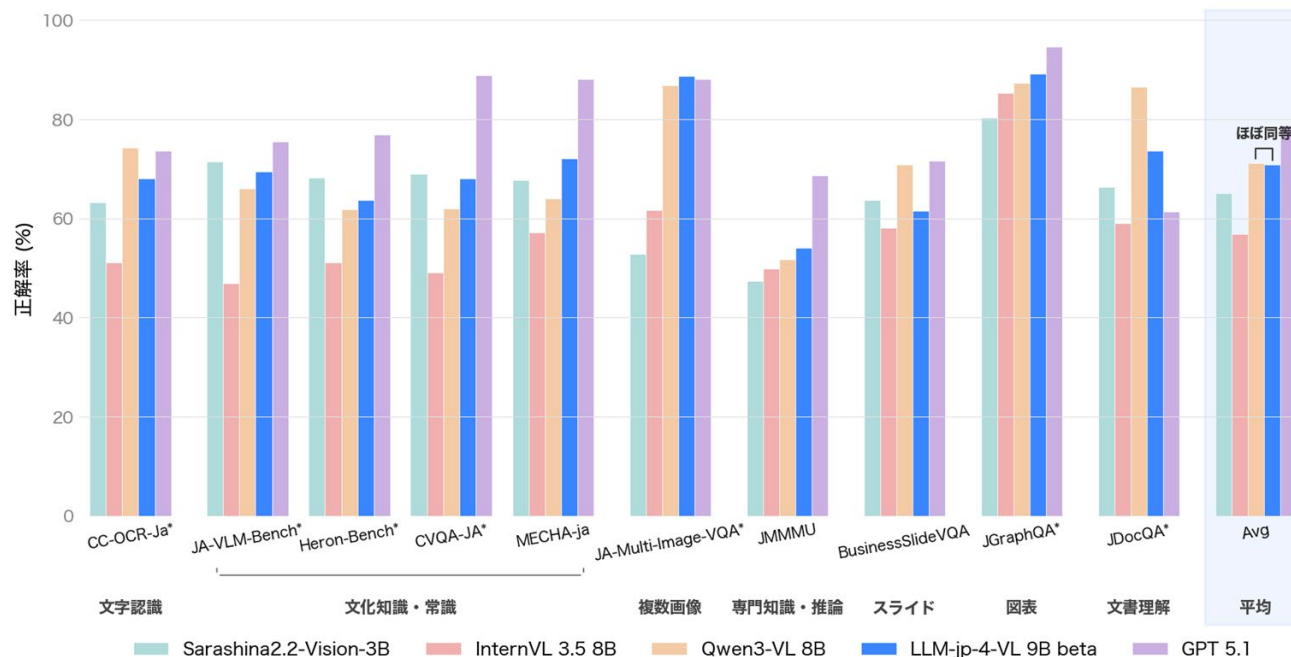
## リソース公開

- モデル: <https://huggingface.co/llm-jp/llm-jp-4-vl-9b-beta>
- 学習データセット: <https://huggingface.co/datasets/llm-jp/Jagle>
- 評価データセット: <https://huggingface.co/datasets/llm-jp/JAMMEval>
- 評価フレームワーク: <https://github.com/llm-jp/simple-evals-mm>

Jagle: 約920万事例からなる日本語マルチモーダル学習データセット

JAMMEval: 既存の日本語ベンチマーク7つを精緻化した評価コレクション

## 主な視覚言語モデルの日本語タスクでの評価結果



日英合計約180Bトークン(約3,340万事例)で学習。マルチモーダル事後学習における学習トークン数はQwen3-VLの約1/10。日本語タスク平均でQwen3-VL-8Bと同等の性能を達成

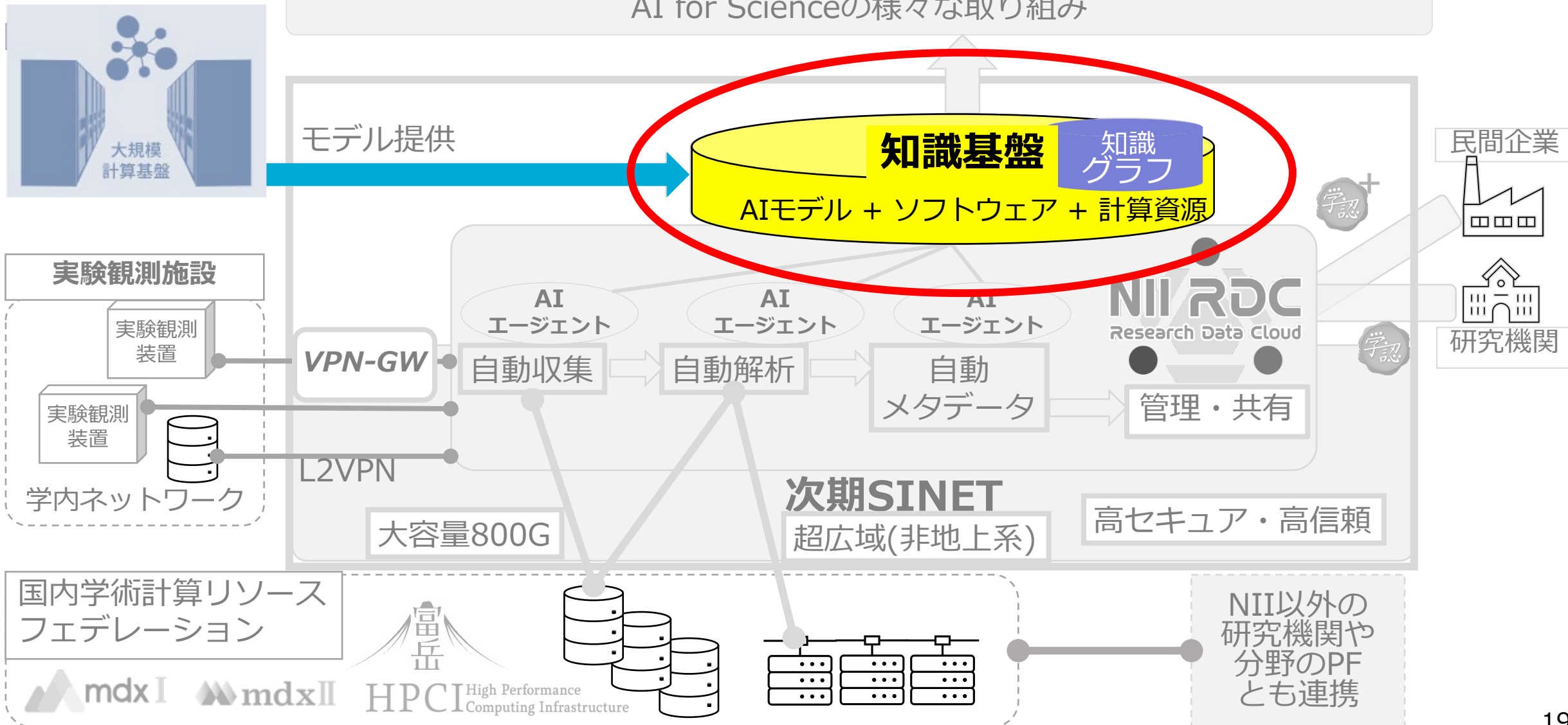
# まとめ

- NII-LLMC/LLM-jpでは、広く活用されているオープンウェイトモデルである Qwen3と同等レベルの性能を持つソブリンAIモデル（LLM、VLM）を開発。これらを基盤として社会実装に向けた取り組みも進行中
- Llamaの勢いの変化や、2026年3月のQwenチーム主要研究者の退職など、**企業主導によるオープンモデル開発の今後の見通しは不透明**
- クローズドな**フロンティアモデル**は著しい発展を遂げているが、**収益上重要なドメインや言語に集中**。その結果、AIと人類の共存のあり方に関わる人文・社会科学や文化的側面は、十分に扱われているとは言い難い
- クローズドモデルはもちろん、オープンウェイトモデルであっても、**学習データや開発レシピは公開されておらず**、Science for AIの観点からはほとんど知見はえられない
- 完全オープンなモデル開発としてAllen Institute for AIによるOLMoの取り組みがあるが、対象は英語限定。NII-LLMC/LLM-jpでは、韓国、インド、タイ、フランス、ブラジルなどとの連携を進めている。**ソブリンAIの開発は一国で完結するものではなく、国際的な協力のもとで推進していくことが不可欠**

# 学術研究プラットフォームの将来像


AI for Scienceを支える研究データの管理・利活用と流通の在り方ワーキンググループ(第1回) <令和7年12月24日> 資料より

AI for Scienceの様々な取り組み



# AI駆動型研究（AI for Science）における AI利用シナリオと必要となる計算環境の整理

**知識基盤が機能提供**



# 背景 — 学術研究のあらゆる側面がAI化していく

学術研究のあらゆる側面がAI化していくことは間違いない

研究者はどのように  
AIを利用するか

必要となるGPU計算環境は  
どの程度の規模か

NIIが提供すべき  
AI利用環境とは

## 本稿の構成

1. 現状の整理（クローズドモデル／オープンモデル）
2. AI駆動型研究におけるAI利用シナリオ
3. LLM-jp-4-8B（試算に用いる基盤モデル）
4. NIIにおいてAI利用環境を提供する場合に必要なGPU資源

LLM-jp-4-8Bを用いてバッチ型／対話的サーバ型のAI利用に必要なGPU資源量を試算  
人文社会科学を含む学術界全体に対するオープンモデルAI利用環境の提供を検討

# 1. 現状の整理① クローズドモデル利用の実態と課題

## 一般的な利用形態

- ChatGPT／Gemini／Claude等のクローズドモデル（無料版または有料版（個人・法人契約）を使用）
- RAGはNotebookLM等の範囲で限定的に利用
- LLM追加学習はあまり一般的でない
- 秘匿性の高いデータの取り扱い  
「入力データを学習に用いない」契約を信じて入力  
／ worst caseを想定してAI処理しない

## クローズドモデル依存の課題

- 無料版は機能制限（質的制限を含む）
- 有料版は組織単位で非常に大きなコスト
- 秘匿性の高いデータは実質AI処理できない
- RAG／追加学習の自由度が限定的
- 実験の再現性確保が困難

クローズドモデルに完全依存することは  
学術界として健全とはいえない

# 1. 現状の整理② オープンモデル活用とNIIの方向性

## オープンパラメータモデルのローカル活用 — 研究者による自発的な動き

- RAG／追加学習の自由度、秘匿データ処理、実験の再現性確保の観点から有効
- GPU環境とモデル運用ノウハウが必要 → 実践できる研究者は少数に限られる
- 選択モデルはQwen、gpt-oss等／強いオープンモデルが継続利用できるかは不透明  
(例：Qwen (Alibaba) がオープンモデル開発を継続するかは定かでない)

## NIIの方向性：学術界全体へオープンモデルによるAI利用環境を提供

- 人文社会科学分野を含む日本の学術界全体を対象とする共通環境の検討
- AI駆動型研究におけるAI利用シナリオの整理
- 複数の最先端オープンモデルを動かす前提（試算はLLM-jp-4-8Bを使用）
- 必要となるGPU資源量を定量的に見積もり、国としての整備の議論に資する

## 2. AI利用シナリオ（1）バッチ型処理

研究者・研究グループのAI利用は「バッチ型」と「対話的サーバ型」に大別 / 本スライドはバッチ型

### ① 論文・実験ノート等のRAG化

OCR → 埋め込みベクトル生成 → indexing  
研究者個人の知識ベース構築の基盤

### ② AI-readyデータの整備

実験ノート・実験データを対象に、  
クレンジング、構造化、メタデータ付与、  
チャンク化、統合等を実施

### ③ LLM追加学習による特化モデル構築

オープン学術データと自らのデータを用いたファインチューニング  
分野特化モデルで精度向上

### ④ 一定規模の実験データ等の解析・推論

一般モデルまたは特化モデルを用いた  
バッチ推論・解析処理  
大規模データ処理を一括実行

※ 特徴：GPUを一定時間専有して実行。処理時間・データ量が事前に見積もれる

## 2. AI利用シナリオ（2）対話的サーバ型

- 研究者・研究グループのRAG利用、学術論文DBに対するDeep Research機能等を前提
- ウェブインタフェース経由／AIエージェントによるAPI経由の両方を想定

### ① アイデア出し・研究計画・共同研究者検討

社会課題解決に資する分野横断的な研究テーマの探索を含む

### ② 日常の研究活動支援

実験ログの取りまとめ、実験ノートの構造化など、日常的な業務支援

### ③ 報告書・論文の執筆支援

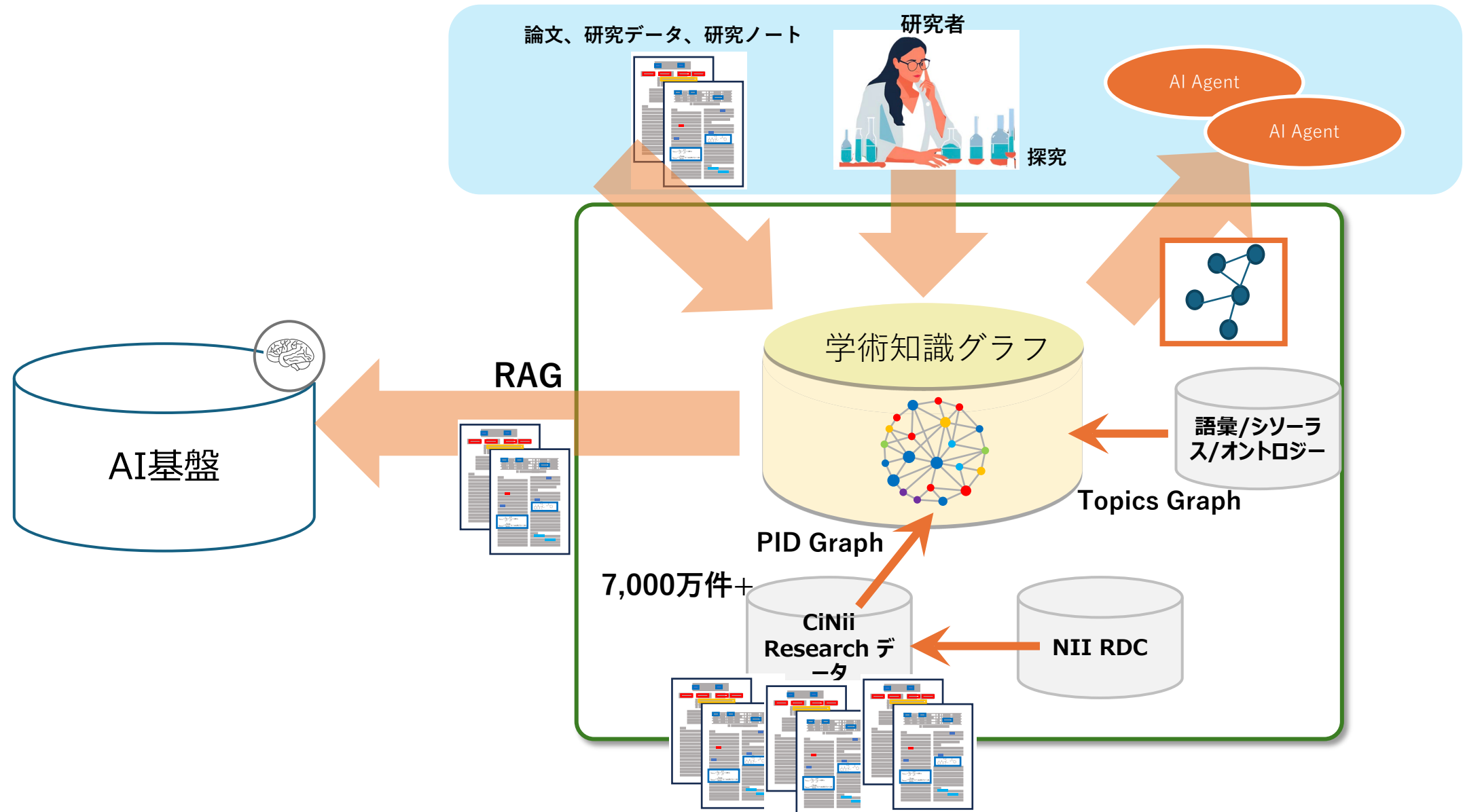
下書き作成、構成案、言い換え、翻訳、校正等の文章作成支援

### ④ 論文・研究提案書などの査読支援

査読観点の抽出、論点整理、コメント作成支援等

※ 特徴：多数のユーザからのリクエストに同時対応。レスポンスタイム確保が重要。

## 2. AI利用シナリオ 学術知識グラフ – 学術情報のRAG活用



### 3. LLM-jp-4-8B (試算に用いる基盤モデル)

#### 2026年4月にNII-LLMCから公開されたフルオープンモデル

学習データの出所・構成、学習レシピ等をすべて明示

一部ベンチマークで GPT-4o、Qwen3-8B を上回る性能

#### モデル仕様

- パラメータ数：8B (正確には 86億)
- 精度形式：bfloat16 (2bytes/parameter)
- 物理サイズ：約 17 GB (= 8.6B × 2bytes)
- **数十GBで扱える軽量クラスに属する**

#### 追加学習・推論スループット

- 追加学習：約 4B tokens/ノード/日(H200)  
大規模な追加学習でも数日程度で実行可能
- 推論:A6000などの1GPUで動作/起動は数十秒
- H200では 1GPUに複数モデル搭載も可能
- vLLM等で非同期・マルチセッション型サービス運用可

NII-LLMCでは 30Bモデル、305B (アクティブ29B) MoE モデルも構築中/2026年度に随時公開予定

## 4. GPU資源試算① バッチ型：RAG化／AI-readyデータ整備

※ 大型研究施設・大型プロジェクトのヘビーユーザは自前環境を使う想定。それ以外（大半）の研究者のためのAI利用環境として試算。

### (1) 論文・実験ノート等のRAG化

- 処理：OCR → 埋め込み生成 → indexing
- 想定：研究者等 約 80万人  
× 2,000 文書／年（各 10MB PDF）  
⇒ **総処理データ量 16PB**
- LLM-OCR：H200 で 4,500 ノード日
- 埋め込み生成（100M BERT級）：1ノード日  
（誤差）
- indexing：CPU処理（誤差）

**必要GPU資源 約 12.3 ノード／年**

### (2) AI-readyデータ整備

- 処理：クレンジング、構造化、  
メタデータ付与、チャンク化、統合
- 想定：5,000 研究グループ（約5万人）  
× 年 100 回
- 1回あたり：H200 × 1枚（= 1/8 ノード）× 30分

**計算式：5,000 × 100 × 1/8 × (0.5/24) ÷ 365**

**必要GPU資源 約 3.6 ノード／年**

## 4. GPU資源試算② バッチ型：追加学習／解析・推論

### (3) 追加学習（特化モデル構築）

- 想定：1,000 研究グループ × 年 12 回
- 1回あたり：4B tokens の追加学習
- スループット：約 4B tokens／ノード／日 (H200)

#### 計算式

$$1,000 \times 12 \times 1 \text{ ノード日} \div 365$$

- 研究プロジェクト等のオンプレGPU環境でも現実的

必要GPU資源 約 32.9 ノード／年

### (4) 実験データ等の解析・推論

- 想定：5,000 研究グループ × 年 20 回
- 1回あたり：H200 × 1 ノード × 2 時間
- 対象：一般モデルまたは特化モデル

#### 計算式

$$5,000 \times 20 \times 2/24 \text{ ノード日} \div 365$$

- 大規模データに対する一括解析に適用

必要GPU資源 約 22.8 ノード／年

## 4. GPU資源試算③ 対話的サーバ型：利用想定と推論性能

### 対象ユーザと利用頻度の想定

#### 研究者等（URA・研究支援者含む）

50 万人

平均 1 日 30 回（うち 6 回は PDF 添付）

#### 大学院生

28 万人

平均 1 日 20 回（うち 2 回は PDF 添付）

#### 大学の学部 4 年生

65 万人

平均 1 日 10 回（うち 0.5 回は PDF 添付）

**合計 約 143 万人**

### 推論性能の実測と試算の前提

- ABCI 1ノード（H200×8）上の vLLM実測  
gpt-oss-120b（Active 5.1B）で  
**約 30,000 tok／ノード秒**  
※ 電力キャップにより本来の2/3性能
- 試算では **42,000 tok／ノード秒** を使用
- LLM-jp-4-8B の推論速度は同等と仮定
- デコード（出力）は逐次処理でプロンプト処理より遅い  
GPT-5.4では6倍遅いとされ、この値で試算

**対話的サーバ型 約 127.8 ノード**

## 4. GPU資源試算④ 合計とサービス運用に必要な規模

区分	利用シナリオ	必要ノード数/年
バッチ型	(1) 論文・実験ノート等のRAG化	12.3
バッチ型	(2) AI-readyデータ整備	3.6
バッチ型	(3) 追加学習（特化モデル構築）	32.9
バッチ型	(4) 実験データ等の解析・推論	22.8
サーバ型	(5) 対話的なサーバ型の利用	127.8
合計（試算値）		199.4 ノード

100万人規模のユーザに妥当なレスポンスタイムで提供 → 少なくとも倍、約 400 ノード必要

### 民間クラウドで確保した場合

年間 約 140 億円

(NII-LLMCでの民間クラウド調達の実験で試算)

### オンプレで整備した場合

年間 約 105 億円

(mdx試算：システム構築300億円、保守・運用費年間45億円、5年間利用時の年平均)

経済的観点、および、大学での計算基盤運用力の維持・発展、人材育成の観点から、オンプレ整備には大きな意味があり、利用者負担の考え方を含め、mdxとの前向きな議論を開始している。

# まとめ — NIIが担うべき役割と今後の検討課題

## NIIが担う役割（AI活用を日本の学术界全体に広げるために）

- オープンモデルによる共通AI利用環境の提供（バッチ型＋対話的サーバ型）
- 追加学習による特化モデル構築の支援
- AI利活用に関するコンサルテーション、情報共有の場の提供

### 定量的な試算結果

試算合計

約 199 ノード

サービス運用に必要な

約 400 ノード

オンプレ整備

年約 105 億円

※ mdxと協力予定