

GakuNin RDM-mdx連携による AI活用プラットフォームとユースケース創出課題への展開 ～AI 推論基盤 mdx-MaaSの提案・構築

鈴木 豊太郎

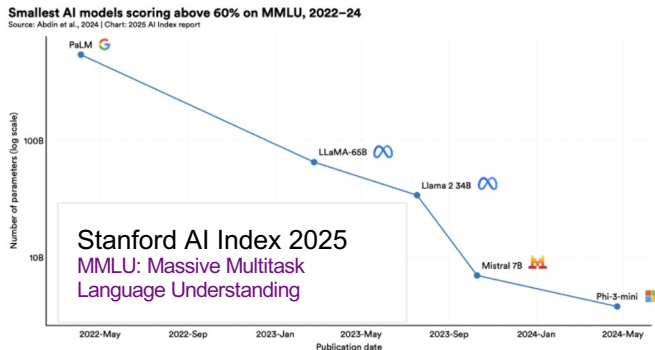
東京大学大学院情報理工学系研究科/情報基盤センター 教授

プロジェクトメンバー：福田敦史, 神沢雄大, 金刺宏樹

研究データエコシステム構築事業シンポジウム 2025
2025年10月9日

背景： AIの発展、学習から推論へ

- 学術研究にはAIが必須の時代へ (AI for Science)
- 学習から推論へ
 - SLM (Small Language Model) でも十分な精度が高くなり、**小規模化**と**高性能化**が進む
 - 推論時計算 (Inference-time compute)によって、より高度な Reasoningが可能に
- AI for Science時代では、学習済みモデルを安全・効率的に運用できる推論環境が、**研究再現性**と**知の信頼性**を左右する



inference

1 of 39

TEXT ONLY

AI: It's All About Inference Now

MICHAEL GSCHWIND

MODEL
INFERENCE HAS
BECOME THE
CRITICAL DRIVER
FOR MODEL
PERFORMANCE

Public discourse about artificial intelligence and neural networks until recently invariably centered on training. First and foremost, training is a prerequisite to anything that follows. Until very recently, a defining question for AI has been whether neural networks can deliver quality results that make them relevant, and how much computational capacity is necessary to train them.

As far back as the 1990s and 2010s, the works of LeCun et al. [1998],³³ Krizhevsky et al. [2012],³⁴ and Sutskever et al. [2014]³⁵ resoundingly answered this question: Neural networks can deliver meaningful capabilities over traditional systems. While the question of how far we can push training continues to be of great importance and the focus for many researchers, the question of whether AI model outputs can be produced affordably—a step known as *inference*—has become topical, now that it has been established that models have reached a quality that makes it worthwhile to deploy them.

acmqueue | march-april 2025 40

1

ACM Queue (<https://queue.acm.org/>), May 24, 2025 Volume 23, issue 2, AI: It's All About Inference Now Model inference has become the critical driver for model performance. Michael Gschwind (NVIDIA)

推論における課題

- **商用AI（推論）**

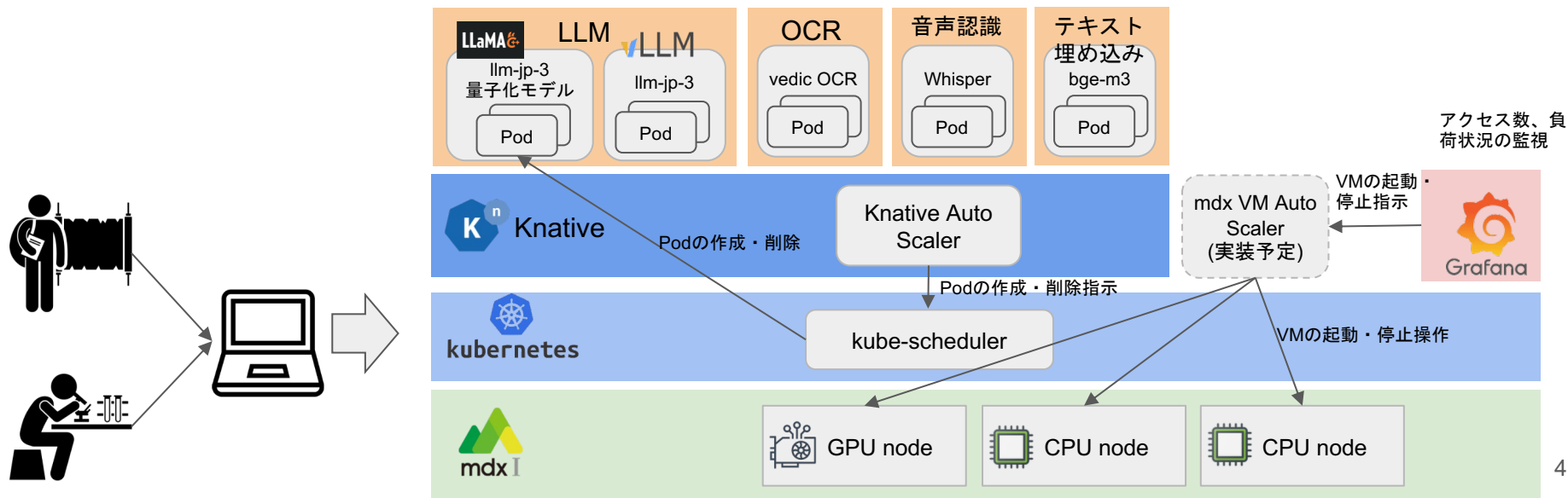
- 外部送信リスク（機密情報、著作権上の問題等）
- コスト増
- 仕様変更・停止リスク

- **個別ローカル運用**

- オンプレミスでのGPU/マルチノード確保に伴う様々なコスト（調達、環境構築）
- 性能・スケーラビリティ
 - ・ LLM等の量子化技術、キャッシュ技術（Key-Valueキャッシュ等）

mdx MaaS : 学術クラウド基盤mdx におけるスケーラブルなAI 推論基盤

- 安全・効率（環境構築/性能）を同時に満たす、mdx上のAI推論基盤
 - mdx内部で動作するため、商用AI等の外部サービスへの情報漏洩のリスクがない
 - 様々なAIモデルが利用可能
 - 負荷に応じた資源最適化、高性能・高いスケーラビリティ



アーキテクチャ概要

- ソフトウェアスタック

- VM層 (mdx VM)
- k8s コア層 (k0s + Knative + Grafana)
- 推論ランタイム層 (vLLM / llama.cpp + Webサーバ)

- AIモデル

- 言語、音声、画像認識, Embedding等の最新のAIモデルをサポート

- 性能・スケーリング

- 1) **Pod スケール** (秒～短時間の変動対応)
 - Knative Autoscaler : 負荷に応じ Pod 自動起動/停止
- 2) **VM スケール** (資源逼迫時に増減・)
 - VM Autoscaler (**実装中**) : Grafana指標で高負荷時にVM 自動起動し、低負荷時に自動停止し、余剰分を削減



デプロイ済みモデル（抜粋）

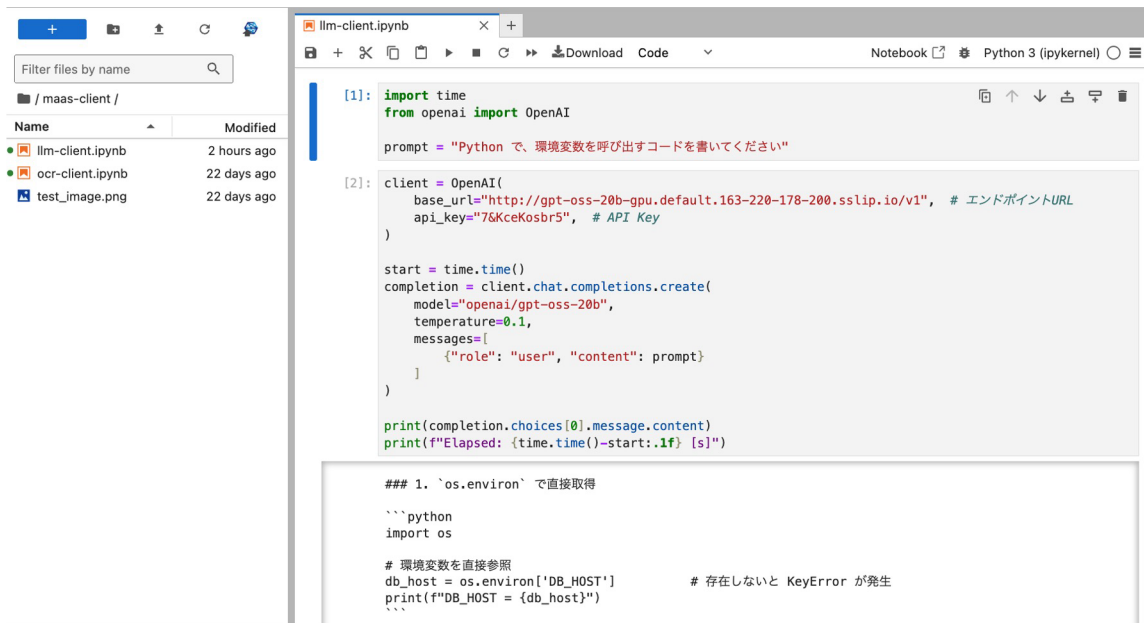
カテゴリ	モデル名	特徴
LLM	llm-jp-3.1-13b-instruct4 (llm-jp)	日本語の対話生成に最適化
LLM	gpt-oss-20b (gpt-oss)	OpenAI API の o3-mini に相当 [5]
LLM	DeepSeek-R1-14b-Japanese (DeepSeek)	長文生成と推論能力を強化
LLM	Qwen2.5-3b-instruct (Qwen2.5)	多言語対応、最大 128k トークンのコンテキスト長
OCR	trocr-large-printed-vedic (vedic)	汎用 OCR モデルをデーヴァナーガリー文字用に微調整
音声認識	faster-whisper-large-v3 (faster-whisper)	OpenAI の Whisper-Large-v3 モデルを軽量化
音声認識	whisper-large-v3-ja (whisper-ja)	軽量化の上、さらに日本語データで微調整
テキスト埋込	bge-m3	多機能性、多言語性、多粒度性を保持

随時最新モデルを追加し、ニーズに応じて追加する予定

利用イメージ：OpenAI互換API 経由

- エンドポイント + 発行されたAPIキー、組織内データ留置で安全
- エンドポイント例: <http://gpt-oss-20b-gpu.default.163-220-178-200.sslip.io/v1>

API呼び出し例



```
[1]: import time
    from openai import OpenAI

    prompt = "Python で、環境変数を呼び出すコードを書いてください"
```

```
[2]: client = OpenAI(
      base_url="http://gpt-oss-20b-gpu.default.163-220-178-200.sslip.io/v1", # エンドポイントURL
      api_key="76KceKosbr5", # API Key
    )

    start = time.time()
    completion = client.chat.completions.create(
        model="openai/gpt-oss-20b",
        temperature=0.1,
        messages=[
            {"role": "user", "content": prompt}
        ]
    )

    print(completion.choices[0].message.content)
    print(f"Elapsed: {time.time()-start:.1f} [s]")
```

```
### 1. `os.environ` で直接取得
```python
import os

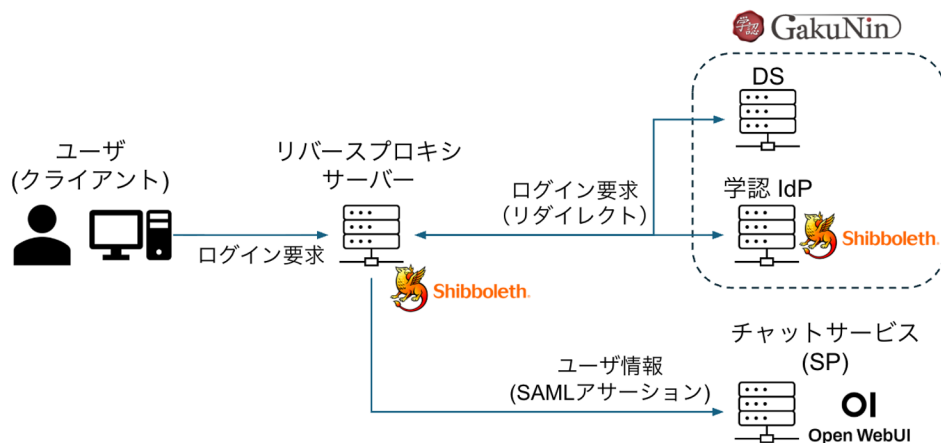
環境変数を直接参照
db_host = os.environ['DB_HOST'] # 存在しないと KeyError が発生
print(f"DB_HOST = {db_host}")
```

## 今後の実装タスク

- \* [ssliip.io](https://ssliip.io)は仮のDNSサービスなので、mdx1.jpに置き換え予定
- SSL証明書を発行し、https対応にする
- mdx のVMインスタンスの ACL設定で、現在は東大からのみアクセス可能
- APIキーは現在、1つしか発行できないが、今後ユーザー毎にAPIキーを発行する仕組みを実装予定

# 利用イメージ：チャットサービス & 学認認証

- **チャットUI (Open WebUI\*)** : ChatGPTライクなUIを提供
- **学認認証** : SAML : DS→IdP→SP、SSO 対応
- **リバースプロキシ** :
  - SAMLアサーションの安全性を確保し、不正ユーザー情報を遮断
  - SPはリバースプロキシ経由でのみ通信 (グローバルIP不要・LB機能付)





# 性能検証：スループットと同時アクセス数

複数クライアントによる同時アクセス（**並列数**と呼ぶ）を想定し、その際のスループット(トークン数/秒)とレイテンシ（秒）を、GPU及びCPU、そして2つのモデル (gpt-oss-20b, Qwen2.5-3b)で測定

## (i) 論文要約バッチ (gpt-oss-20b, 6,136 tokens)

### 最大スループットは？

- GPU 64同時処理がスループットが最大化し、11,821 tokens/sまで向上→1本につき10k tokenだとすると、**平均して論文1本1秒で処理可能**
- CPU(16コア) の場合は、8並列 438 tokens/sが最大（\*GPUとの理論性能差 ≈27x と整合）

表 2 論文要約タスクのスループット

環境	並列数	スループット
GPU	1	1116
	4	2869
	16	5609
	64	11821
CPU (16 コア)	8	438

## (ii) 単問応答 (≈1,000 tokens)

**(低レイテンシのまま) 何人の同時アクセスが可能？**

- gpt-oss-20 (GPU) の場合は**4から16**
- Qwen-3b (64GPU) の場合は**64**
- Qwen-3b (CPU)の場合は2

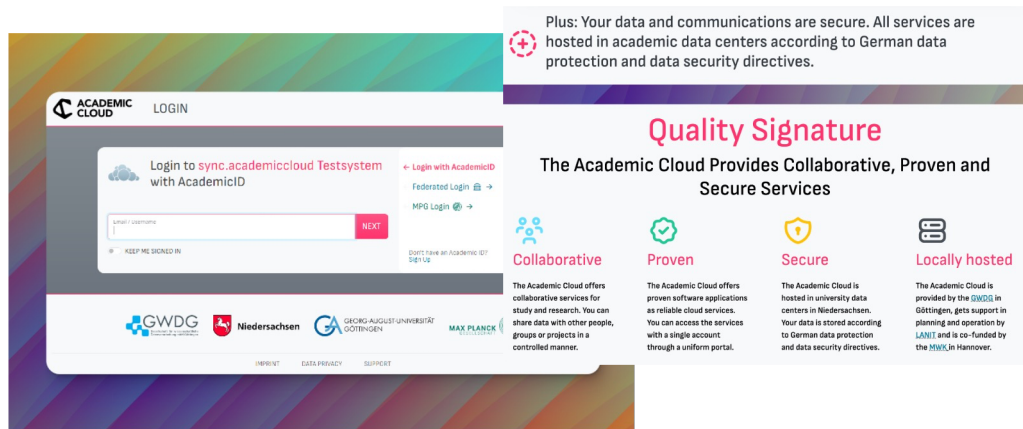
表 3 単問応答タスクの応答時間 (秒) とスループット (TP)

環境	並列数	gpt-oss-20b		Qwen2.5-3b	
		時間	TP	時間	TP
GPU	1	6.6	151	2.3	433
	4	8.2	494	3.2	1238
	16	12.4	1292	5.9	2735
	64	18.6	3450	9.7	6618
CPU (16 コア)	1	42.3	24	8.5	118
	2	41.9	48	12.4	161
CPU (32 コア)	1	33.9	30	7.5	134
	2	36.0	56	11.9	168
CPU (64 コア)	1	32.0	31	6.8	146
	2	45.2	44	9.3	214
OpenAI API	1	9.6	104		

# 関連プロジェクト

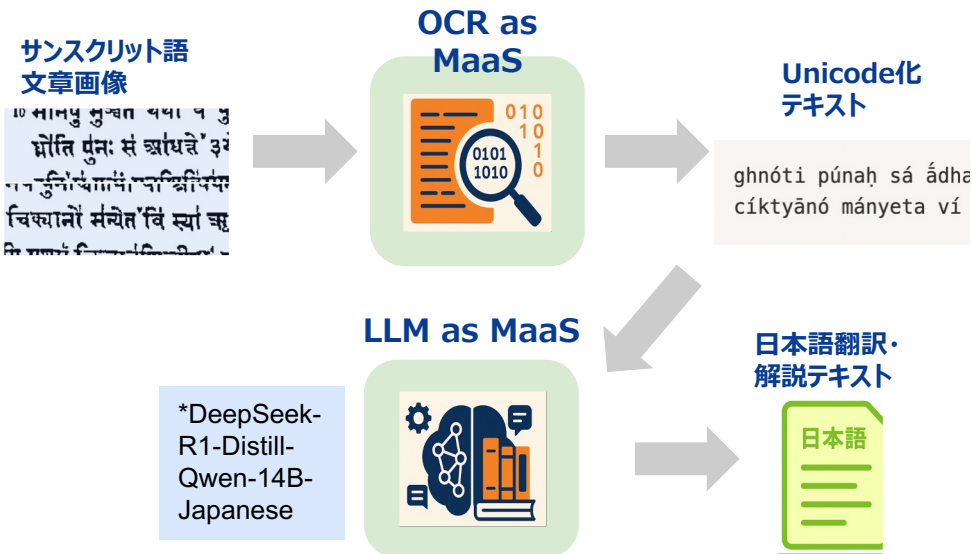
- **Jetstream2 (米国 Indiana大)** : SSO + Open WebUI + OpenAI互換API
- **Academic Cloud (独ゲッティンゲン学術データ処理センター)**
  - チャットUI + OpenAI互換API
  - 独ゲッティンゲン大学、**マックス・プランク研究所**等の研究者に計算機環境を提供

→ **mdx MaaS は同潮流に位置し、日本の学術要件へ最適化**



# Gakunin RDM連携 ～サンスクリット語の文字画像OCR認識

- 古文書は著作権上、商用AIに投げる事が出来ない
- 人文学領域の東京大学・塚越先生/大向先生と協業し、サンスクリット語（インドの古典語）の起源である**ヴェーダ語**（紀元前1200年～紀元前500年）の文章画像から文字認識ができるように、mdx MaaSに OCRモデルを整備
- LLM (DeepSeek\*) を使用し、OCRで得られたヴェーダ語を、日本語で解説する仕組みを構築

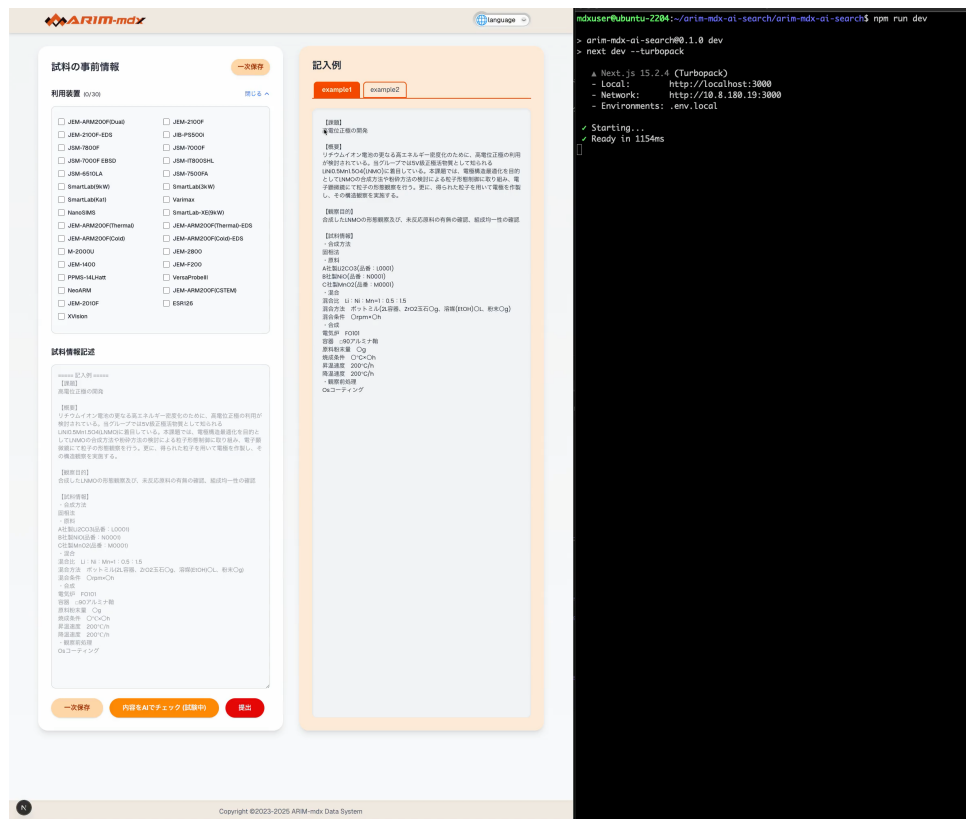


\*DeepSeek-R1-Distill-Qwen-14B-Japanese

The screenshot shows the GakuNin RDM web interface. The top navigation bar includes 'GakuNin RDM', 'toyolabo', 'ファイル', 'Wiki', '解析' (selected), 'メンバー', 'アドオン', and '設定'. The main content area is titled '私の解析環境' (My analysis environment) and lists a single environment: 'rugfly-osfstorage-1k1xh8x' with a status of '2025-10-04 02:12 PM' and a green '起動' (start) button. Below this is a section for '新しい解析環境' (New analysis environment) with a '基本イメージ' (Basic image) dropdown set to 'Python 3.12 + R 4.4 (JupyterLab 4.x)' and a list of installed packages including 'apt-get', 'conda', 'pip', 'R (CRAN)', and 'MATLAB (mpm)'. A green button at the bottom right says '新しい解析環境を作成' (Create new analysis environment).

# mdx MaaSと材料科学

- 材料試料データの収集システム
- ユーザーから提出された試料の情報(テキスト)を**LLMを用いてクオリティをチェック**
  - “十分な試料情報です”
  - “もう少しXXXやYYYの情報を加えるといいでしょう”
- OpenAI を mdx MaaSに置き換えテスト完了→本格運用はこれから
  - **機密情報への対処**、低コスト化
  - Fine-tuningなどによるモデルの最適化



## • テックブログ→ [URL](#)





### クラウド基盤mdxでのKubernetes構築に関する記事一覧

2025/03/09に公開 2025/09/15

Kubernetes

MDX

mdxjp

Tech

第1回 クラウド基盤mdxの使い方の基本

Atsushi Fukudaさんによる記事

zenn.dev

第2回 クラウド基盤mdxでKubernetes環境を構築する

Atsushi Fukudaさんによる記事

zenn.dev

第3回 クラウド基盤mdxで構築したKubernetesクラスタを管理するツールLensの紹介

Atsushi Fukudaさんによる記事

zenn.dev

第4回 クラウド基盤mdxで構築したKubernetesで、サーバレスWebアプリケーションを構築する

Atsushi Fukudaさんによる記事

zenn.dev

第5回 クラウド基盤mdx+ Kubernetesの上でLLM Webサービスを立ち上げる

Atsushi Fukudaさんによる記事

zenn.dev

第6回 クラウド基盤mdx+ Kubernetesでホスト側のディレクトリをマウントする方法

Atsushi Fukudaさんによる記事

zenn.dev

クラウド基盤mdxに構築したKubernetesクラスタ上で簡単に分岐学習を行う

Atsushi Fukudaさんによる記事

zenn.dev



Atsushi Fukuda

フォロワー

Atsushi Fukudaさん

バッジをもらう

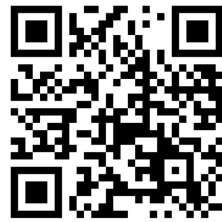
バッジを贈るとは →

Zennからのお知らせ

 ZennFes2025 - Auth0賞が追加！秋の祭典ZennFes開催中！

 Zenncafe出張イベントも開催！10/18 DevelopersIO 2025 TOKYO

## AXIES'25で発表予定→ [URL](#)



mdx MaaS: 学術クラウド基盤 mdx におけるスケーラブルな AI 推論基盤

鈴村 豊太郎<sup>1),2)</sup>, 福田 敦史<sup>3)</sup>, 神沢 雄大<sup>3)</sup>, 金刺 宏樹<sup>2)</sup>

1) 東京大学 大学院 情報理工学系研究科

2) 東京大学 情報基盤センター

3) Independent Researcher

suzumura@ds.itc.u-tokyo.ac.jp

### mdx MaaS: A Scalable AI Inference Platform on the Academic Cloud System mdx

Toyotaro Suzumura<sup>1),2)</sup>, Atsushi Fukuda<sup>3)</sup>, Yuta Kanzawa<sup>3)</sup>, Hiroki Kanezashi<sup>2)</sup>

1) Graduate School of Information Science and Technology, The University of Tokyo

2) Information Technology Center, The University of Tokyo

3) Independent Researcher

#### 概要

AI 技術の普及に伴い、研究・教育への活用需要が高まっている。一方で、商用 AI サービスの利用に伴うセキュリティリスクや、オープンソースモデルを個人で運用するための計算資源確保など、実運用上のハードルは高い。本稿では、学術機関で AI モデル推論を安全かつ効率的に活用するための推論基盤 mdx MaaS を提案する。データ活用社会創成プラットフォーム mdx I に推論ランタイムと OpenAI 互換 API エンドポイント、チャット UI、動的スケーリングを統合し、安全性と運用効率を両立する推論基盤を提供する。mdx MaaS を構成するソフトウェア、提供するサービスと AI モデル、および実運用に向けた予備検証実験結果について報告する。

#### 1 はじめに

AI 技術の急速な発展により、学術研究においてもその活用が不可欠となっている。特に、OpenAI や Gemini など企業が提供する商用モデルと、オープンソースモデル間の性能差が縮小し、多様な AI モデルを柔軟に利用できる環境が整いつつある。近年ではモデル学習よりも学習済みモデルの推論利用が増加しており、高性能なモデル推論環境をいかに安定的に利用できるかが重要となっている。また、大学・研究機関においても、分野を問わず AI モデルの活用が進んでいる。学術利用における文章作成・プログラム実装から、大規模なテキスト・画像等の研究データの変換・解析まで、その応用範囲は広がつつある。

しかし、従来の方法で AI モデルを研究利用するには限界がある。商用のパブリック AI サービスは手軽に利用できる一方で、機密性の高い研究データを外部に送信するセキュリティリスクやコスト増大等の問

コストに加え、構築・運用ノウハウの属人化が問題となる。

本稿では、この問題に対処し、学術機関で AI モデルを安全かつ効率的に活用するための推論基盤として、「mdx MaaS (Model as a Service)」を提案する。データ活用社会創成プラットフォーム mdx I [1] 上に、AI モデル推論、需要に応じた計算資源の動的スケーリングを含めた性能最適化機構、API・チャット UI を統合した基盤を提供することで、セキュリティ・性能最適化・効率化の両立を目指す。以下、mdx MaaS のアーキテクチャ設計と実装、さらに各種 AI モデルを実際に配備して行った予備的な性能検証の結果について報告する。

#### 2 現状の AI モデル活用の課題

AI モデルを利用する主な手段として、民間企業が運営する商用 AI サービスを利用する方法と、各自でローカル LLM を入手してデプロイする方法が考えら

# まとめ & 今後

## まとめ

- 安全・効率（環境構築/性能）を同時に満たす、mdx上の共通のAI推論基盤 mdx MaaSを提案し、そのプロトタイプシステムを構築・予備評価を行った

## 利用者募集

- ベータバージョンのテストユーザー・プロジェクトを募集していきたいので、準備が出来次第、データエコのトラックでアナウンスします

## ロードマップ

- サービスインに向けた機能拡充・テスト
  - 複数ユーザーへの API キー発行
  - 100人同時ユーザーを想定した耐久テスト（品質、スケーラビリティ、耐故障性等）
- サービスモデル（提供形態、料金等）
- モデルの拡充（マルチモーダル等）
- RAG連携、エージェント連携
- mdx II やスパコン、その他のシステムへの展開

研究データエコシステム構築事業  
融合・活用開拓チーム

# ユースケース・創出課題

- ・取り組み概要
- ・本シンポジウム初日企画

本事業 運営委員（東京大学）  
田浦・鈴木・古宇田

2025年10月9日



GakuNin RDM



mdx

# ユースケース創出課題 取り組み概要

## ◆事業概要

### <目的>

「全国的な研究データ基盤（NII-Research Data Cloud）を活用し、異なる分野間でのデータ連携を前提とした AI・データ駆動型研究のシーズ・ユースケースの創出」

### <公募事業>

- ・2か月に1回程度で公募実施。2025年7月までに多様な分野から40課題を採択
- ・2年間迄の研究資金助成を受けて研究データのエコシステム構築に向けた活動を実施

### <コミュニティ形成>

- ・採択課題は研究資金助成期間後も本シンポジウムにおいて情報交換
- ・本事業のslackを設けて情報交換の場を形成

## ◆今後の課題

- ・各ユースケース創出課題に関わる研究データ利活用コミュニティの拡大
- ・ユースケース創出課題間での分野融合の機会創出と支援強化
- ・AI活用プラットフォームを活かした分野融合・データ活用開拓を促進



# 本シンポジウム初日企画 ～新たな協創を導く～（１）

- ◆各カテゴリよりAIを駆使した課題に口頭発表 ⇒ データセットやAIツールの活用展開
- ◆人文課題は分野融合の可能性をパネルで議論 ⇒ 興味を持った他分野との接合を誘引
- ◆昨年度シンポ以降採択課題にフラッシュトーク ⇒ 新規課題の周知とポスターでの議論促進
- ◆全課題ポスター発表 ⇒ 研究コミュニティの拡大・新たな協創

No	カテゴリ	課題名	代表機関	提案代表者	発表形態	発表者
1	人文	人文学DXを指向する情報基盤の構築	京都大	原正一郎		馬場 弘樹(中央大)
2	人文	形態素解析・係り受け解析AIにおけるデータ管理とデモ環境の統合	京都大	安岡孝一		同左
3	人文	人文学研究における「読み」を共有するためのデジタルアーカイブ構築・AI活用ワークフローの確立	東京大	大向一輝	パネル	同左
4	人文	専門家・市民・AIの協働によるデジタルアーカイブ構築に向けたデータプロビナンス基盤のデザイン	東京大	大向一輝	パネル	同左
5	人文	異分野共創による史料学DXの確立	東京大	尾上陽介		渋谷 綾子(東京大)
6	人文	古典籍テキストデータを活用したデータ駆動型人文学のための研究資源構築プロジェクト	国文学研	菊池信彦	パネル	同左
7	人文	Archivematicaを用いた人文学データのキュレーションプロセスの自動化に関する試み	東京大	中村寛	パネル	同左
8	人文	オープン・エスノグラフィ: GakuNin RDMと連携したデータ管理ソフトウェアによる質的研究のコラボレーションとオープンデータ化の研究	大阪大	森田敦郎		同左
9	人文	楔形文字文献の統一デジタルアーカイブ化に向けたパイプライン構築	東京大	塚越柚季	フラッシュ	同左
10	人文	大規模言語モデルを使用した文学研究のための研究資源およびオープンプラットフォームの構築	中央大	橋本 健広	口頭	同左
11	人文	共創型情報システムによる紙文化財の多角的解析と活用	東京大	黒嶋敏	フラッシュ	同左
12	人文	「行為」の構造化を軸とするデジタル・ヒストリー研究データと研究プロセスの接続に関する研究	東京大	小川潤	フラッシュ	同左
13	人文	LJ aR+: 日琉語族を中心とする言語類型データベースのAI駆動型構築	筑波大	宮川創	フラッシュ	同左
14	実験	ESR装置群を基軸とする研究データ流通・利活用エコシステムの構築	大阪公立大	佐藤和信		同左
15	実験	大規模実験データの計測・解析・共有・公開を通じた知の創出のためのエコシステム構築	大阪大	小野寛太		鈴木 謙介
16	実験	核融合研究データのオープンな利活用基盤「プラズマ・核融合クラウド」の構築と整備	核融合研	中西秀哉		同左
17	実験	実験データのコンテキストを記述するメタデータ・文書等の共通スキーマ整備と相互参照用PIDの附与	核融合研	中西秀哉	フラッシュ	同左
18	実験	コアファシリティにおける研究データ管理アーキテクチャの構築	金沢大	松平拓也		高田 良宏(金沢大)
19	実験	全国規模の材料データプラットフォームにおけるAIベース検索システムの構築	東京大	華井雅俊	口頭	同左

# 本シンポジウム初日企画 ～新たな協創を導く～（２）

No	カテゴリ	課題名	代表機関	提案代表者	発表形態	発表者
20	バイオ・医療	乳児の股関節脱臼の見落としゼロを目指す 異常判別AI とコミュニティスクリーニングシステムの開発	東京大	吉岡京子		同左
21	バイオ・医療	国際的なRVD研究を実現する医療情報分析基盤の検討	愛媛大	木村映善		同左
22	バイオ・医療	健康医療データの質を保証する安全安心な医療AI サービスプラットフォームの構築	大阪大	林美加子	口頭	同左
23	バイオ・医療	マルチスピーシーズ霊長類脳画像の知識型データ共有エコシステムの構築: データ駆動型 脳科学を推進する共創型国際研究基盤	京都大/ 滋賀医科大	岡田知久/ 守村直子		同左
24	バイオ・医療	細根フェノロジーの解明に向けた森林土壌の多地点自動解析システムの構築	兵庫県立大	大橋瑞江		池野 英利( 福公大)
25	システム	高専における分析データ集約・配信モデルシステム構築	奈良工専	山田裕久		同左
26	システム	研究データの可視化・検索性の向上を目指したメタデータ変換システムの開発と実装	名古屋市立大	能勢正仁		同左
27	システム	インタラクションエンコーダによる研究データ基盤横断探索システムの研究開発	NII	高須淳宏		同左
28	システム	AI Qps によるネットワーク運用効率化に向けたデータ駆動型共有ナレッジベースの構築	広島大	近堂徹	口頭	同左
29	システム	分野横断型データベースAM DERの活用による次世代型データ利活用スキームの構築	ROIS	小財正義		同左
30	地域	地域文化資源データの共創のための汎用プラットフォームの開発	北海道大	山下俊介		橋本 雄太( 暦博)
31	地域	地域デザインのためのインクルーシブ・データプラットフォームの構築	京都大	徳地直子		同左
32	地域	地域資料データの継承とオープン化を目指した地域横断型データ共有基盤の構築	山形大	佐藤琴		同左
33	地域	地域コミュニティを基盤とした国際コミュニティ・データベースの構築	東北大	伊藤文人		同左
34	人流	都心部における人流変容メカニズムの解明と予測モデルの構築	東京大	蕭耕偉郎		同左
35	人流	擬似人流データを用いた身体活動量の推定と地理的・社会経済的環境から見た地理的地域特性の解明: 地域住民の健康増進に向けた活動量シュミレーションシステムの開発	琉球大	喜屋武享		同左
36	観測	テラヘルツ電波多波長観測データと学習・シミュレーションの融合による, 月・小惑星表層構 造リモートセンシング解析情報基盤の確立	東京大	飯野孝浩		同左
37	観測	衛星コンステレーション観測データ共有基盤の開発	群馬大	小川康一	フラッシュ	同左
38	環境	地球人間圏科学における 3D データ活用基盤の構築	北海道大	早川裕式	口頭	同左
39	環境	森林長期データを活用した分野横断的研究プラットフォームの設計	東京大	徳永友花	フラッシュ	藤原 章雄( 東京大)
40	教育	学習者の批判的思考を促進するオンライン教育アクセス基盤の構築: 質問応答支援方式の開発と大規模教育コンテンツ研究データセットの利活用推進	関西学院大	角谷和俊		同左