学術クラウド基盤基盤mdxの現在地

大学・研究機関で共創する
産学官連携のための
データプラットフォーム

東京大学 情報基盤センター データ科学研究部門
鈴村豊太郎

mdx の背景と目的

- 背景～科学研究の進め方の変化
  - データ駆動型研究
  - オープンサイエンス（透明性・再現性）
  - AI for Science
  - → 現在、 GPU整備は日本の成長戦略における最重要投資対象の１つ

- mdxとは～高性能計算機と大容量ストレージからなるクラウド型IaaS
  - データレポジトリとしての活用：
    SINETを活かしたリアルタイムデータ収集・集積とセキュア解析
  - 大規模データ処理・高性能計算機としての利用：
    計算科学×データ科学で高精度予測

さまざまな分野のデータ保持者・解析者・利用者コミュニティを育成し、
新たな価値形成へ

mdx 構築の経緯

- **これまでの経緯**

  第5期科学技術基本計画において日本の未来社会のコンセプトとして提唱されたSociety5.0の実現を支える基盤として、

  - 11機関（9大学2研究所）の共同プロジェクトとして2021.9にmdxI運用開始
  - 耐故障性や耐災害性の観点から東西２拠点体制とし2024.11にmdxII運用開始

- **運営組織～データ活用社会創成プラットフォーム協働事業体**

北海道大学 情報基盤センター	東京科学大学 情報基盤センター（旧称：東京工業大学 学術国際情報センター）
東北大学 サイバーサイエンスセンター	名古屋大学 情報基盤センター
筑波大学 人工知能科学センター	京都大学 学術情報メディアセンター
東京大学 情報基盤センター	大阪大学 D3センター（旧称：大阪大学 サイバーメディアセンター）
国立情報学研究所	九州大学 情報基盤研究開発センター
産業技術総合研究所 情報・人間工学領域	

mdx のシステム構成

システムの特徴

- 仮想化技術を用いた研究プロジェクト毎の環境
- データ収集公開や推論等のインタラクティブ環境
- スポット／起動保証VMによる資源の有効利用
- mdxIからmdxIIへの相互運用ノードの設置
- 「学認」による利用申請、GakuNin RDM連携

mdxII 実際の写真

主要スペック

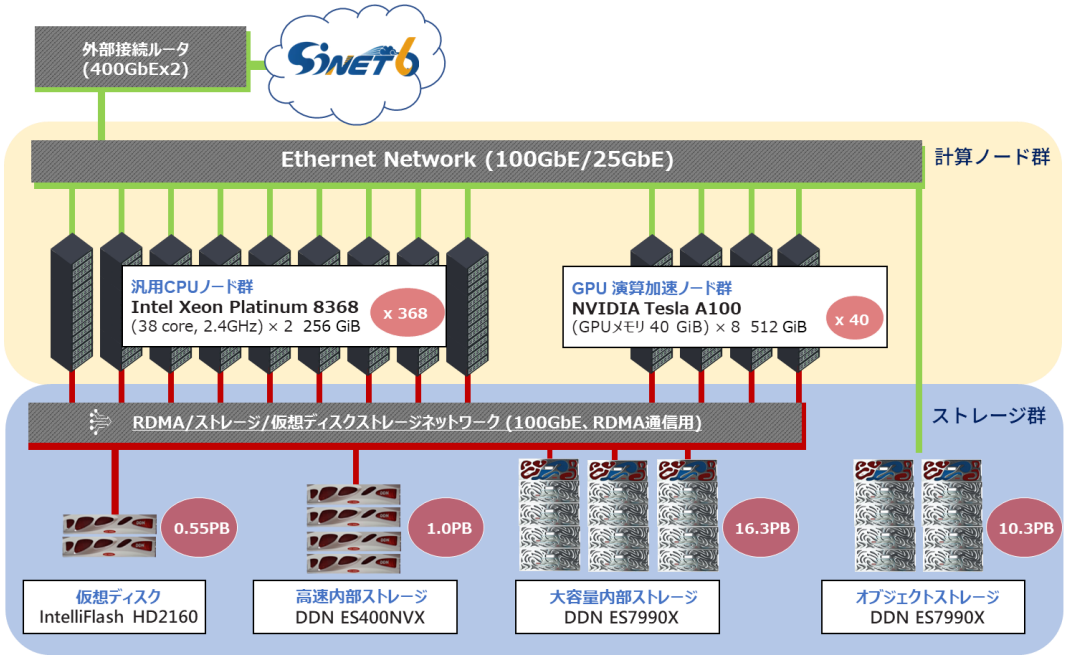
性能指標	mdxI	mdxII
CPU	4.2PFLOPS(FP32)	0.9PFLOPS(FP32)
GPU	6.7PFLOPS(FP32) A100x8基x40ノード	1.9PFLOPS(FP32) H200x4基x7ノード （2026年度より15ノード、4.1PFLOPS構成へ）
ストレージ （大容量ストレージ＋オブジェクトストレージ）	26.6 PB	1.5 PB
仮想環境	VMware	OpenStack, VMware(mdxIとの相互運用用)
ネットワーク	SINET6(400Gbps)、モバイルSINET (セキュアIoT広域データ収集基盤)	
運用会社	富士通	日本電気

昨今、旺盛なAI関連需要を反映し、mdxI,IIともGPU使用率は90％超の状況

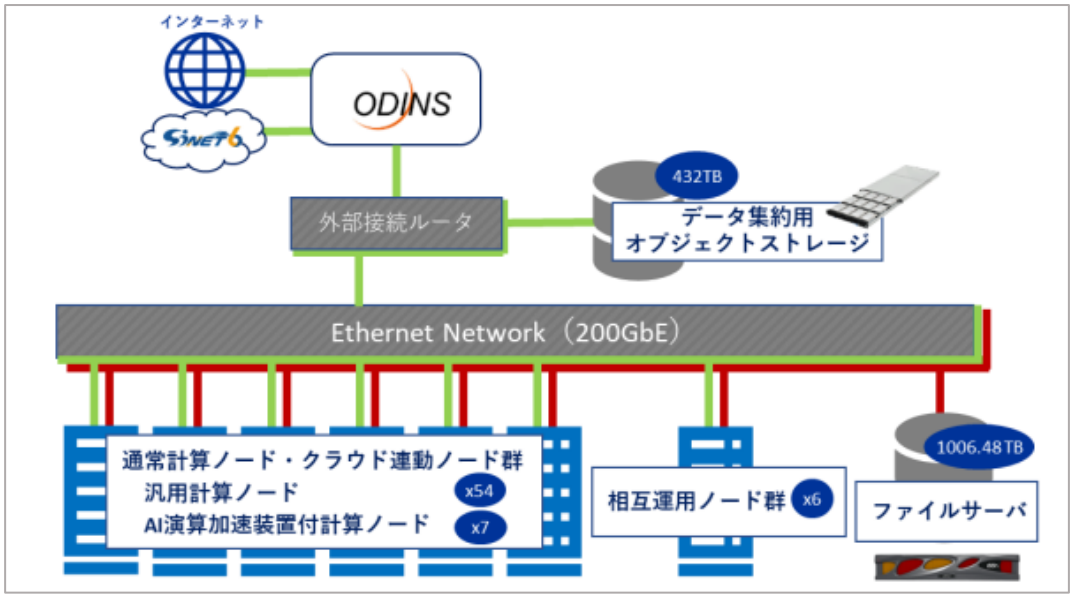
【ご参考】mdx システム構成図

mdx

mdx I



mdx II



mdx の活用事例

- 現在のプロジェクト数は約 175

セキュリティ要件の厳しい分野においては、パブリッククラウド等の利用が困難な場合が少なくないため、学術研究用クラウドの活用が盛ん

1) 大規模言語モデル / マルチモーダルAI / NLP
   LLM-jp / マルチモーダルLLMの構築と検証 / japanese-med-llm
2) 医療・生命科学データ基盤／医用画像／ゲノム
   医療画像解析基盤 / ハプロタイプカタログ/ 胆管の発生メカニズム解析
3) マテリアルズ / 計算化学・ポリマー
   ポリマーインフォマティクスのデータ基盤構築 / マテリアルズインフォマティクス
4) 流体力学・乱流・気候・地球科学
   Turbulence Database / 長周期地震動予測 / 世界の気候モデルデータアーカイブ
5) 天文データ / 観測データ基盤
   Tomo-e Gozen（全天サーベイカメラ）データプラットフォーム / 交通量推計モデルの開発
6) デジタル人文学 / OCR / アーカイブ
   古典写本（くずし字）OCRプロジェクト / デジタルアーカイブ公開プロジェクト
7) ロボティクス・センサ・IoT
   自動運転に関するAI応用 / ロボットセンサデータ収集解析基盤の構築 / Soft Robotic Simulator
8) 社会・経済・金融
   manga_ai / 財務ビッグデータの可視化と統計モデリング / 価値交換工学
9) 教育・実習・演習

具体的な事例をWebサイトにてご紹介中 https://mdx.jp/mdx1/p/doc/cases

さらなる活用促進に向けた取り組み

- **公募型研究課題**に採択されることで利用料金の補助、または、mdxを活用する取り組みに対して研究費を助成
  - 学際大規模情報基盤共同利用・共同研究拠点(JHPCN)（～2025 mdx I , 2026～mdx I mdx II）
  - AI等の活用を推進する研究データエコシステム構築事業 mdx I mdx II
  - 東京大学情報基盤センター 若手・女性利用 mdx I
- **価格設定の工夫**
  - 2025年度よりCPU利用料を**半額！** mdx I
  - 大容量ストレージ・オブジェクトストレージは利用料を**無償化！** mdx I
- **AI推論基盤 mdx-MaaS の提供** mdx I
- さまざまなVMテンプレートを提供しユーザビリティを向上 mdx I mdx II
  - Windowsインスタンスや、Gaussian内包インスタンスの提供 mdx II
- マルチノード環境構築のためのクラスタパックの提供 mdx I mdx II
- **研究用スーパーコンピューターとの連携**
  - Miyabi（国内研究用２位、80.1 PFLOPS）とのデータ連携（予定） mdx I
  - OCTOPUSのクラウドバースティング先としての活用（予定） mdx II

mdx の利用申込方法

mdx公式サイト https://mdx.jp

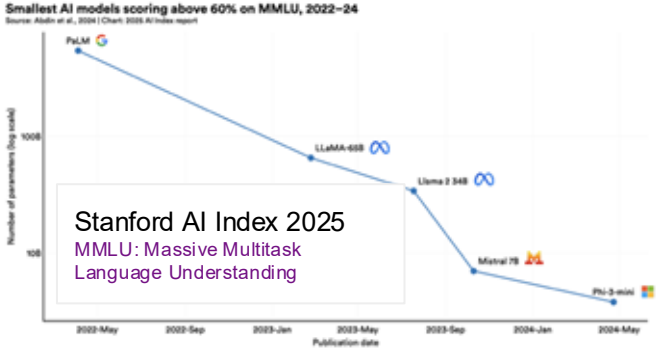
mdx MaaS: 学術クラウド基盤mdxにおけるスケーラブルなAI推論基盤

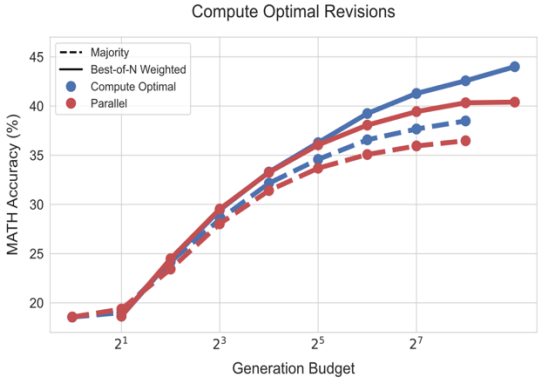
鈴村 豊太郎

東京大学　大学院情報理工学研究科電子情報学専攻
東京大学情報基盤センター
教授

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**
  - オープンウェイトモデルや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)を増やすことによって*、より高度なReasoningが可能になり、自律進化型AI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重要→研究再現性と知の信頼性を左右する**





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)

*Charlie, et.al, “Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning” , ICLR24

推論のオプションと課題

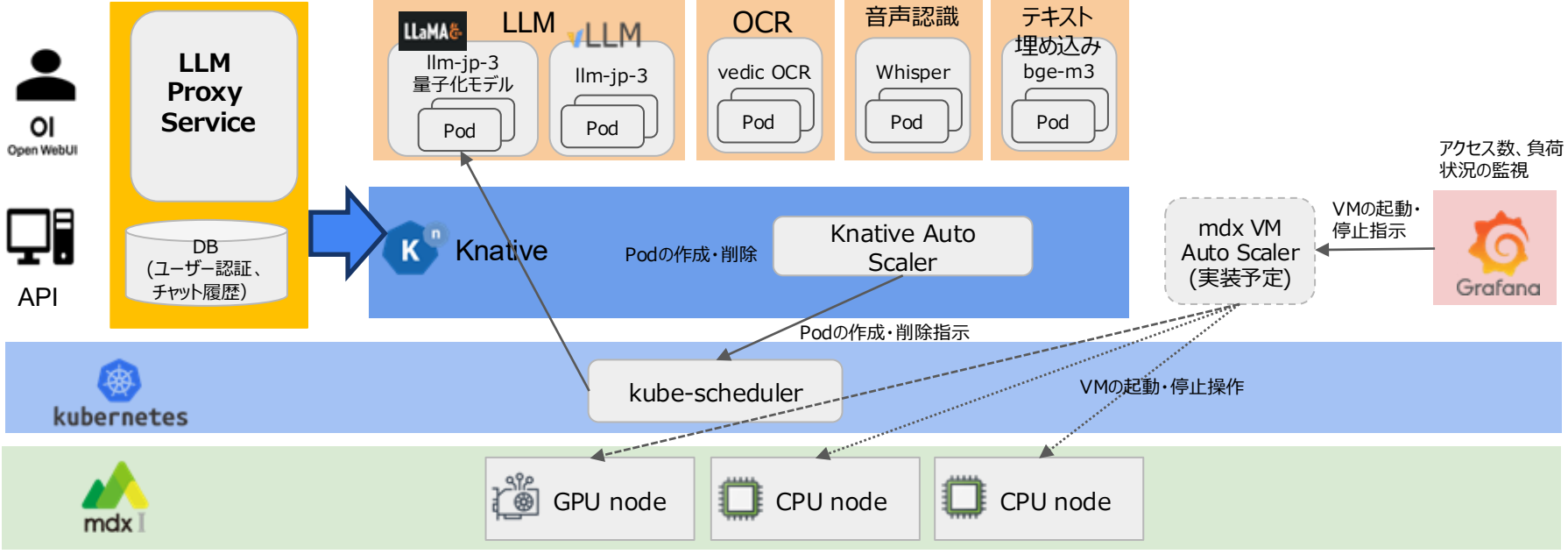
AI for Science時代では、学習済みモデルを安全・効率的（性能・コスト）に運用できる推論環境が必要

- **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
  - デプロイされたモデルは基本永続的に利用できるため、実験における再現性が担保される
  - 負荷に応じた資源最適化、高性能・高いスケーラビリティ

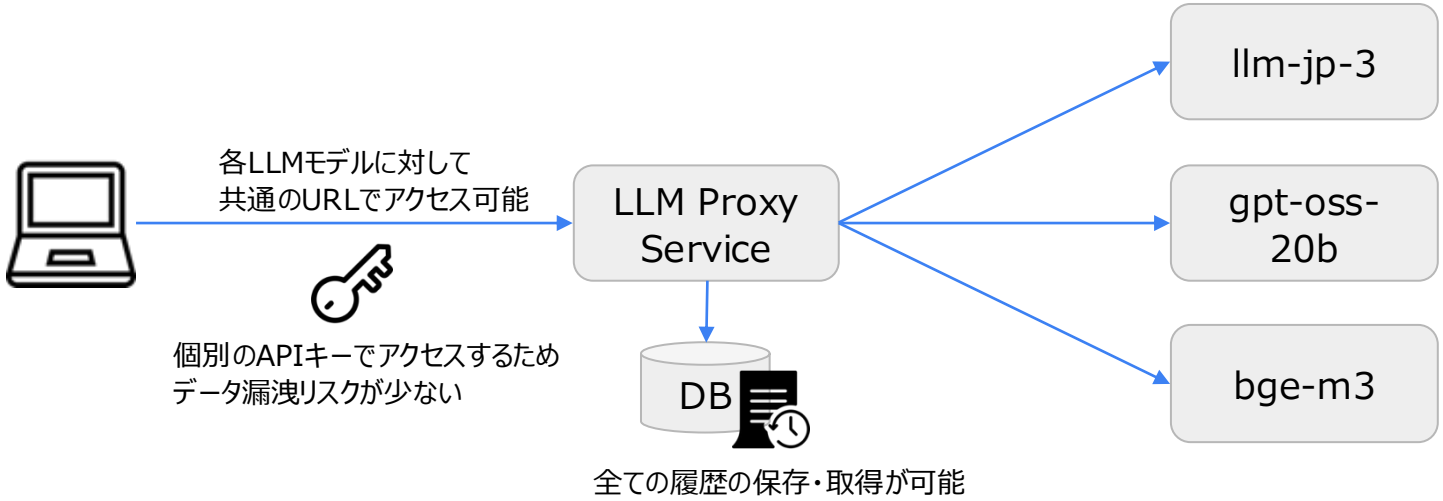


アーキテクチャ概要

- **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

LLM Proxy Service

- LLM Proxy Service の提供する機能
  - 各デプロイ済みモデルに対して共通のURLでアクセスが可能なゲートウェイサービス
  - ユーザーごとに個別のAPIキーを発行できるため、**パスワードの漏洩リスクが少なく(失効や再発行が容易)**、履歴データを保存できるため、実験データの保存が可能
  - 今後、過去のLLMとのやり取りを取得できるAPIを提供予定。これにより、**過去の全ての実験データをいつでも取得可能**



公開予定モデル

β版リリース時に公開予定のモデル

カテゴリ	モデル名	特徴
LLM	llm-jp-3.1-13b-instruct4	日本語の対話生成に最適化
LLM	gpt-oss-20b	OpenAI API の o3-mini に相当

随時公開予定のモデル

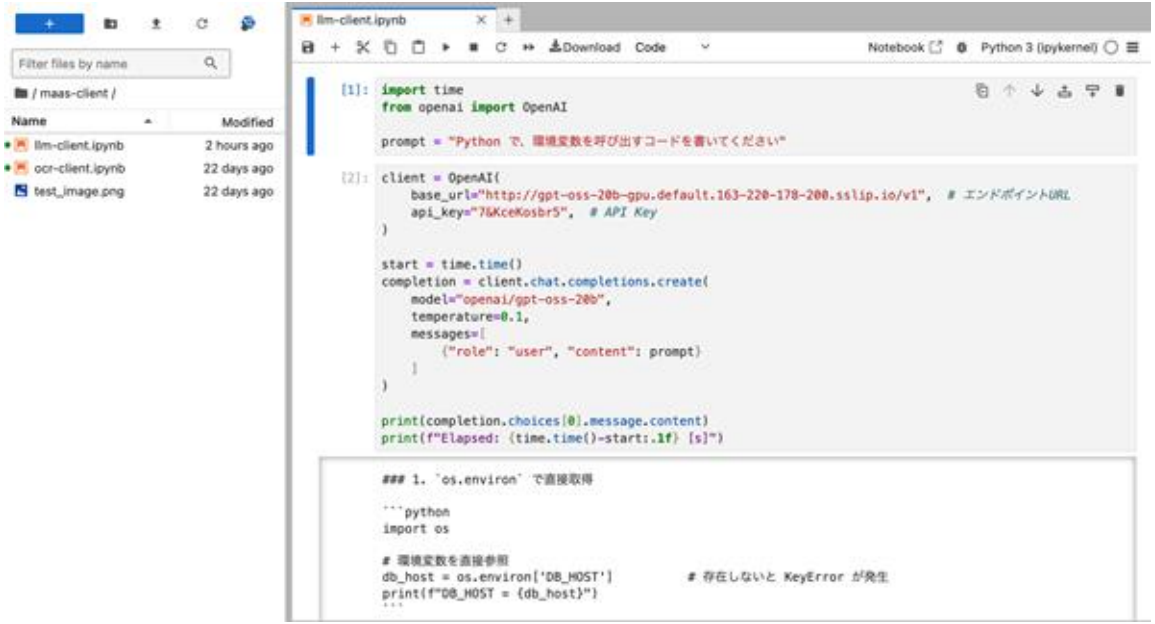
カテゴリ	モデル名	特徴
テキスト埋込	bge-m3	多機能性、多言語性、多粒度性を保持
VLM	Qwen3-VL-30B-A3B-Instruct	テキストと画像を扱えるモデル
音声認識	whisper-large-v3-ja	OpenAI の Whisper-Large-v3 モデルを軽量化し、日本語に最適化

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

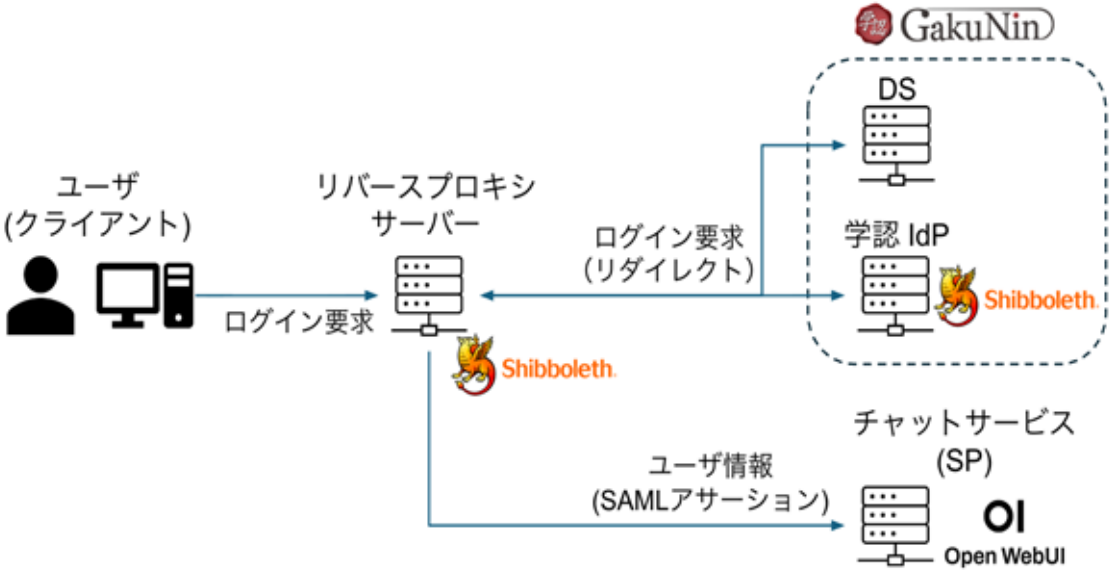
- OpenAI APIと互換性のあるインターフェースであるため、PythonのOpenAIライブラリを使って利用可能
- HTTPSで通信が暗号化されており、ユーザーごとにAPIキーの設定ができるため安全に利用可能
- 組織毎に過去のデータを蓄積できるため、実験データの管理にも有効

API呼び出し例



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
- **リバースプロキシによる学認認証機能の提供**
  - SAMLアサーションの安全性を確保し、学認登録機関以外の不正なユーザーからのアクセスを完全に遮断
  - チャットサービスのサーバはリバースプロキシ経由でのみアクセス可能なため、不正な攻撃からの防御能力が高い



*Open WebUI https://github.com/open-webui/open-webui

チャットサービスのデモ

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

・GPU64同時処理において、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、**理論上平均して論文1本1秒で処理可能**

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(<10s)を保ったまま、同時アクセスが可能な最大数

- Qwen-3b（2GPU）→ **100**
- gpt-oss-20b（2GPU）→ **8**

表 3　単問応答タスクの応答時間 (秒) とスループット (TP)

		gpt-oss-20b		Qwen2.5-3b	
環境	並列数	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	1	42.3	24	8.5	118
(16 コア)	2	41.9	48	12.4	161
CPU	1	33.9	30	7.5	134
(32 コア)	2	36.0	56	11.9	168
CPU	1	32.0	31	6.8	146
(64 コア)	2	45.2	44	9.3	214
OpenAI API	1	9.6	104		

今後のロードマップ

- 過去の履歴を参照、取得できるAPIの実装
- 有償化に向けたサービスモデル
  - Batch API の実装
  - VMオートスケーラーのスケーリング戦略の具体化と実装
- 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

まとめ

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

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
- ベータ版を 2026年1月19日に開始