

データ基盤から知識基盤へ

黒橋 禎夫

国立情報学研究所 所長／京都大学 特定教授

問題意識

- 環境問題、格差問題、地域紛争など、人類は矛盾に満ちた世界に苦しんでいる
- 学術が複合的な社会課題を解決し、真に人々の心の安寧をもたらすためには、人文・社会科学を含む様々な学術研究の協働が必要

⇒ オープンサイエンス → NII学術情報基盤

- この30年間、情報学が社会および学術界に広範なインパクトを与えてきた (WWW、深層学習、AlphaFold、DeepMind、ChatGPT)

⇒ 生成系AIによる社会変革 → LLM-jp

知識基盤



2023 G7科学技術大臣会合におけるOSの推進



G7科学技術大臣コミュニケ 2023年5月12日-14日（仙台）



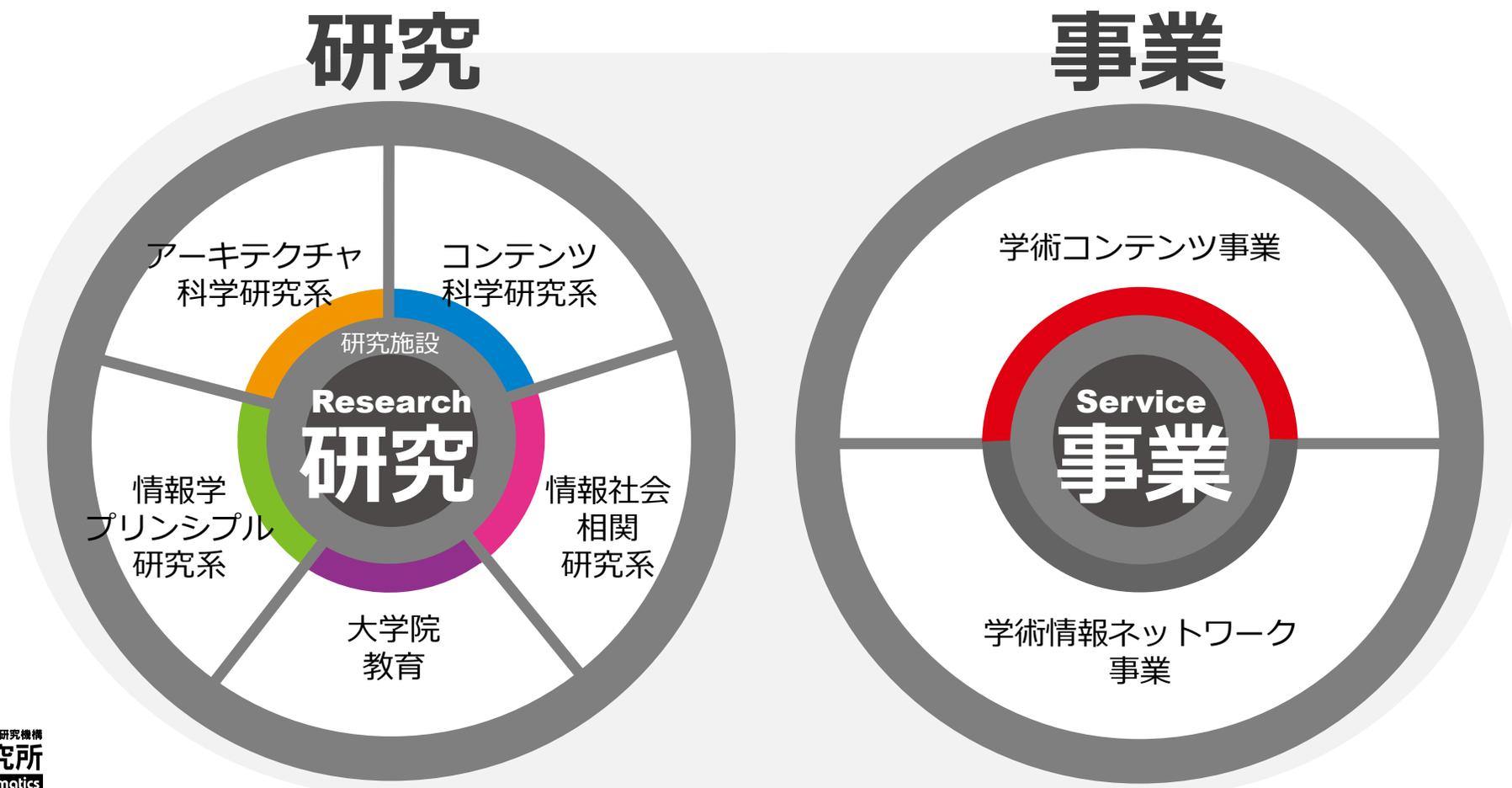
科技大臣会合でのNII RDCデモ

1. 科学研究における自由と包摂性の尊重およびオープン・サイエンスの推進

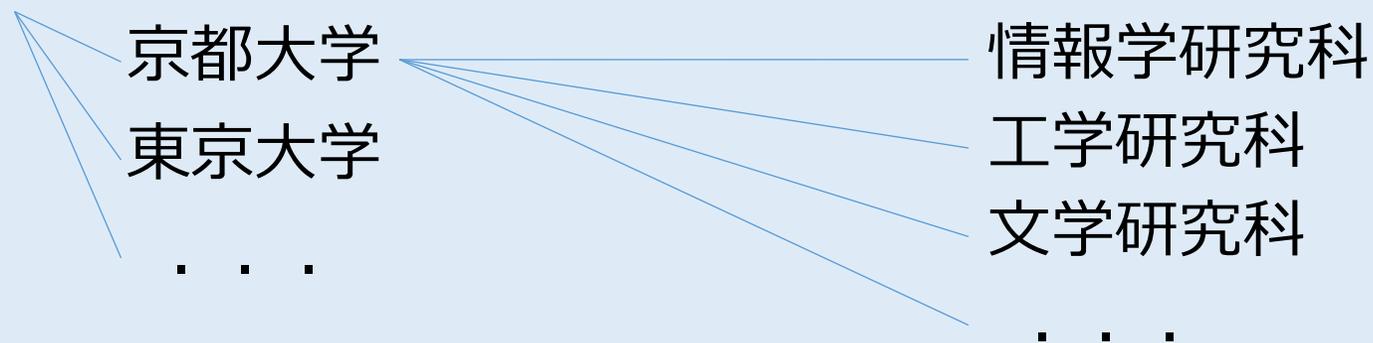
G7は、FAIR原則（Findable = 見つけられる、Accessible = アクセスできる、Interoperable = 相互運用できる、Reusable = 再利用できる）に沿って、科学的知識並びに研究データ及び学術出版物を含む公的資金による研究成果の公平な普及による、オープン・サイエンスの拡大のために協力する。これは、世界中の研究者や人々がその恩恵を受けるとともに、新しい知識の創造、イノベーションの促進、社会による知識へのアクセスの民主化及び地球規模の課題に対する解決策の開発に貢献するためである。これは、より再現性があり、信頼できる研究成果を構築することにも役立つ。

国立情報学研究所(NII) のミッション

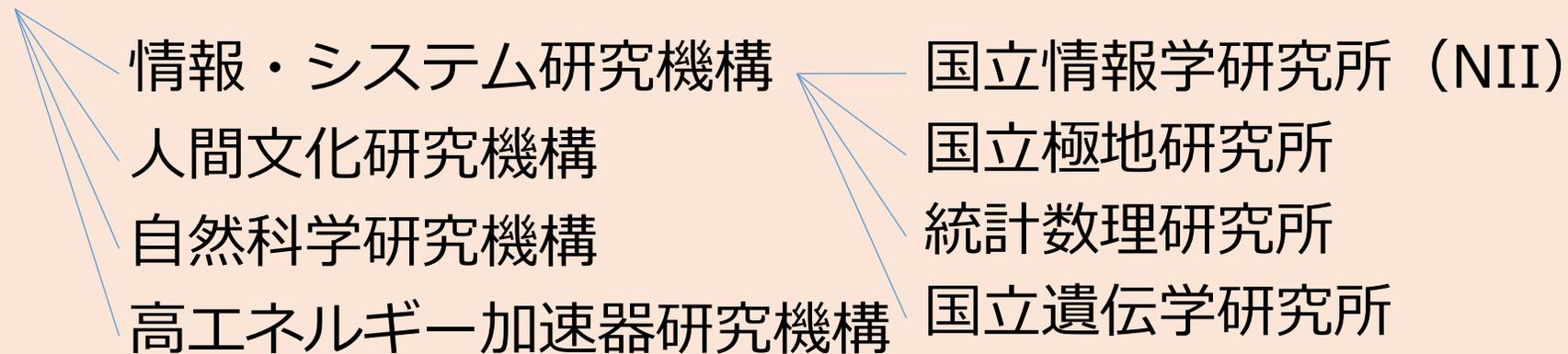
研究と事業を両輪として情報学による未来価値を創成



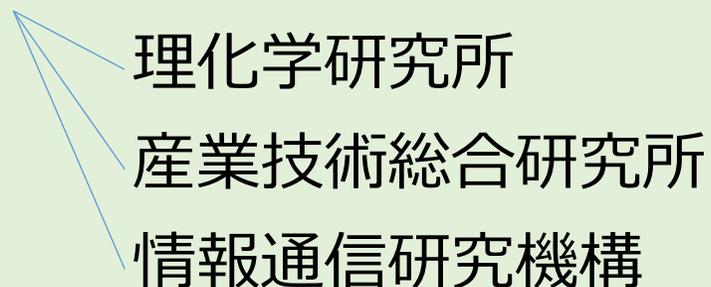
国立大学法人



大学共同利用機関法人



国立研究開発法人（独立行政法人）



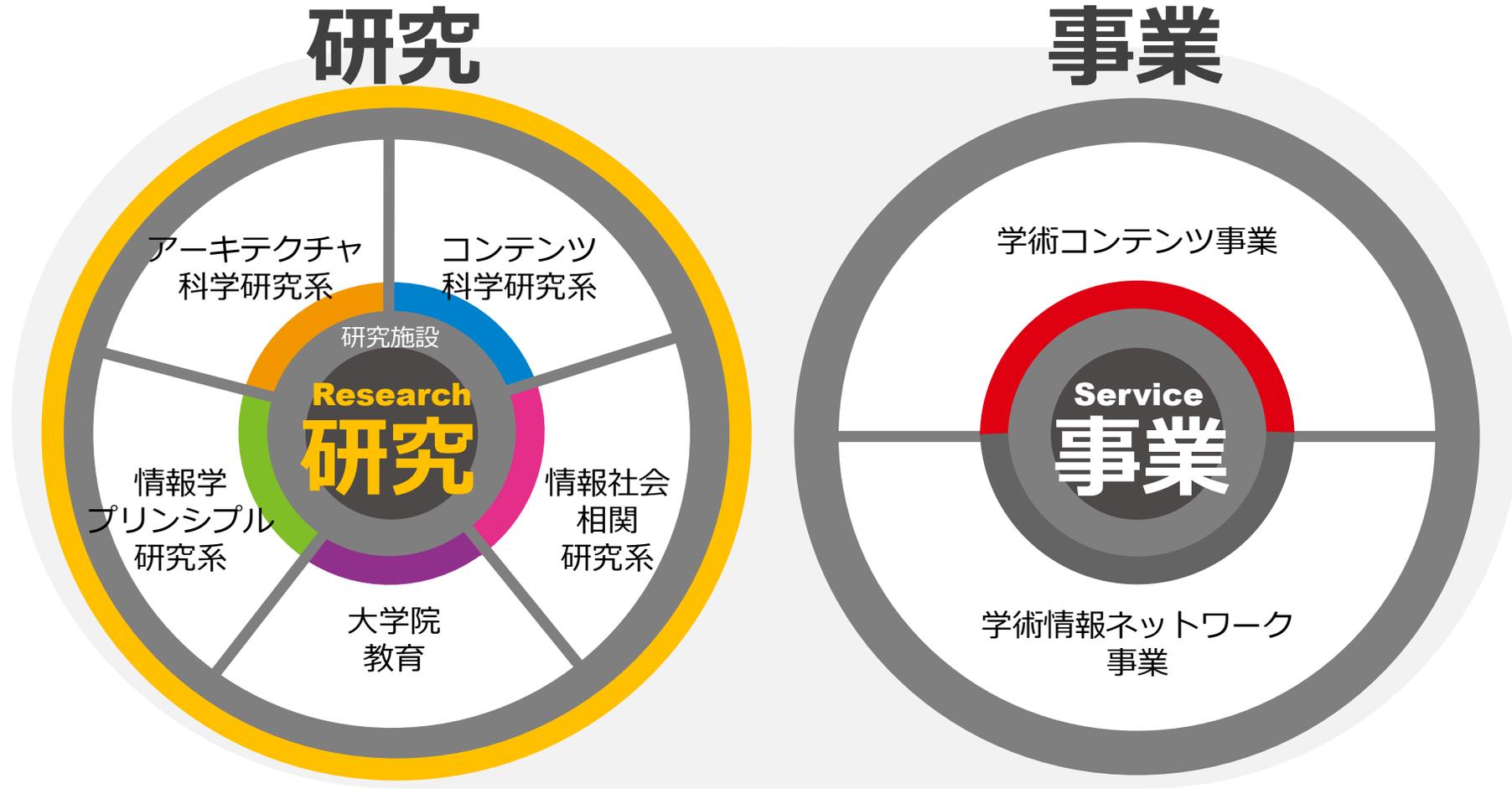
所員・予算

所員		(2023年4月現在)
研究教育職員		71
事務職員等		69
特任教授等 (プロジェクト型)		36
特定有期雇用職員等		301
客員教授等 <small>(2023.3時点)</small>		204
大学院生 (インターンシップ生を除く)		138

予算 (2023年度)		(単位：千円)
運営費交付金	12,445,065	ミッション実現戦略加速化経費等9,513,554 SINET：8,516,943 セキリティ：915,388 ストラテジックサイバーレジリエンス研究開発センター：69,223
受託事業等	1,437,492	(見込)
雑収入	138,073	(見込)



研究と事業



4つの研究系

情報学プリンシプル研究系



数理情報, 形式検証, 量子コンピューティング, アルゴリズムデザイン, データマイニング, 人工知能・機械学習, オントロジー

科研費
KAKENHI

(基盤研究S)

- グラフアルゴリズム基盤と最適化：理論研究とアルゴリズム開発 (河原林健一教授)

CREST

- 記号推論に接続する機械学習 (杉山磨人准教授)

アーキテクチャ科学研究系



ネットワーク・計算機アーキテクチャ, セキュリティ, ミドルウェア, ソフトウェア工学, プログラム言語

ERATO

- 運尾メタ数理システムデザインプロジェクト (運尾一郎准教授)

CREST

- 形式検証とシステムソフトウェアの協働によるゼロトラストIoT (竹房あつ子教授)

コンテンツ科学研究系



テキスト・自然言語処理, マシンビジョン, コンピュータグラフィックス, バーチャルリアリティ・パターン認識

CREST

- VoicePersonae : 声のアイデンティティクローニングと保護(山岸順一教授)
- 納得感のある人間-AI協調意思決定を目指す信頼インタラクションデザインの基盤構築(山田誠二教授)

情報社会相関研究系

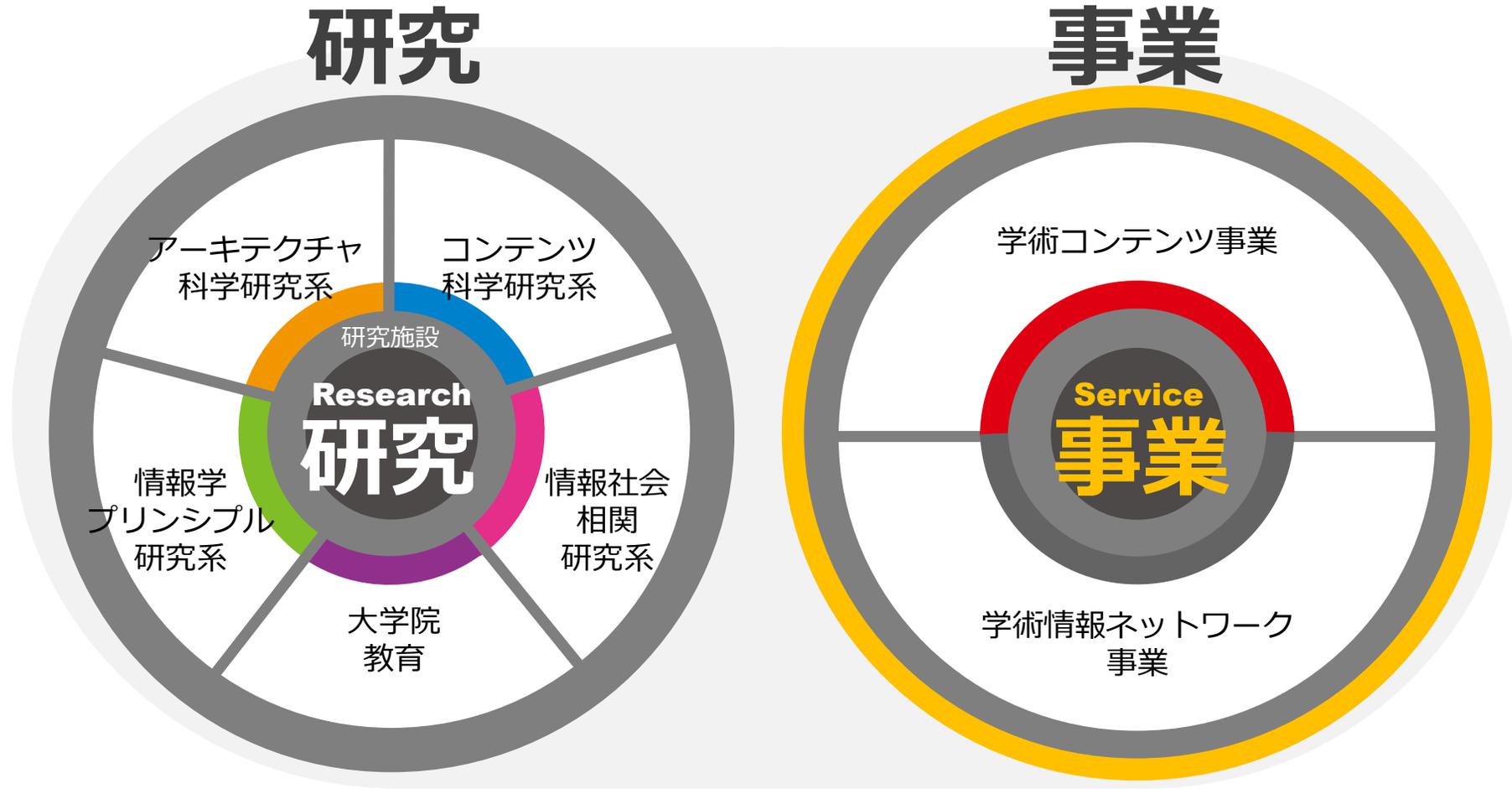


学術情報, 図書館情報学, 情報検索, 情報技術と社会, 情報制度

CREST

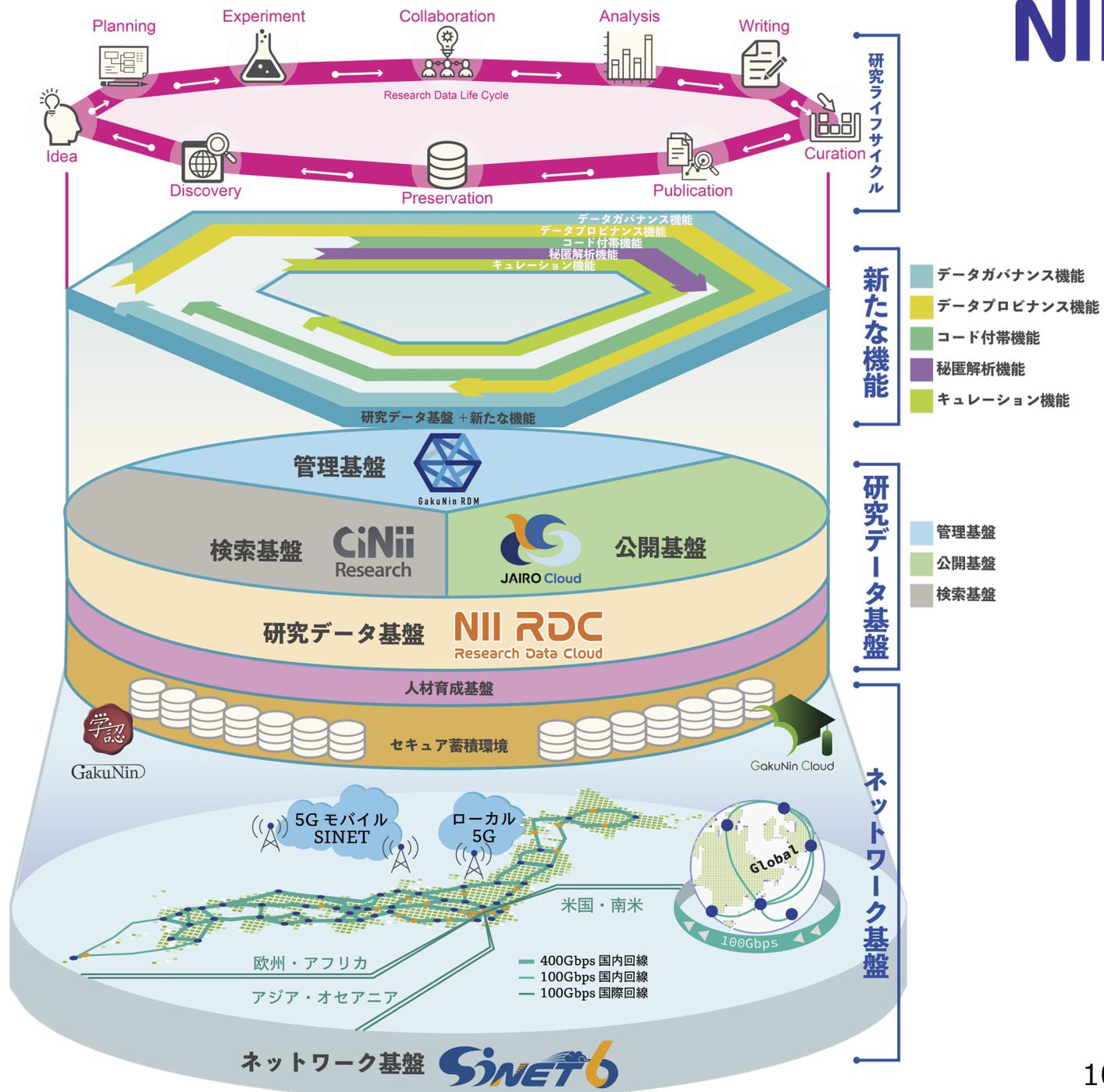
- インフォデミックを克服するソーシャル情報基盤技術 (越前功教授)

研究と事業



NII事業の全体像

- 超高速ネットワーク、研究データ基盤、サイバーセキュリティ、クラウド、認証に関するサービスを提供し、全国の大学や研究機関等の研究教育環境を高度化

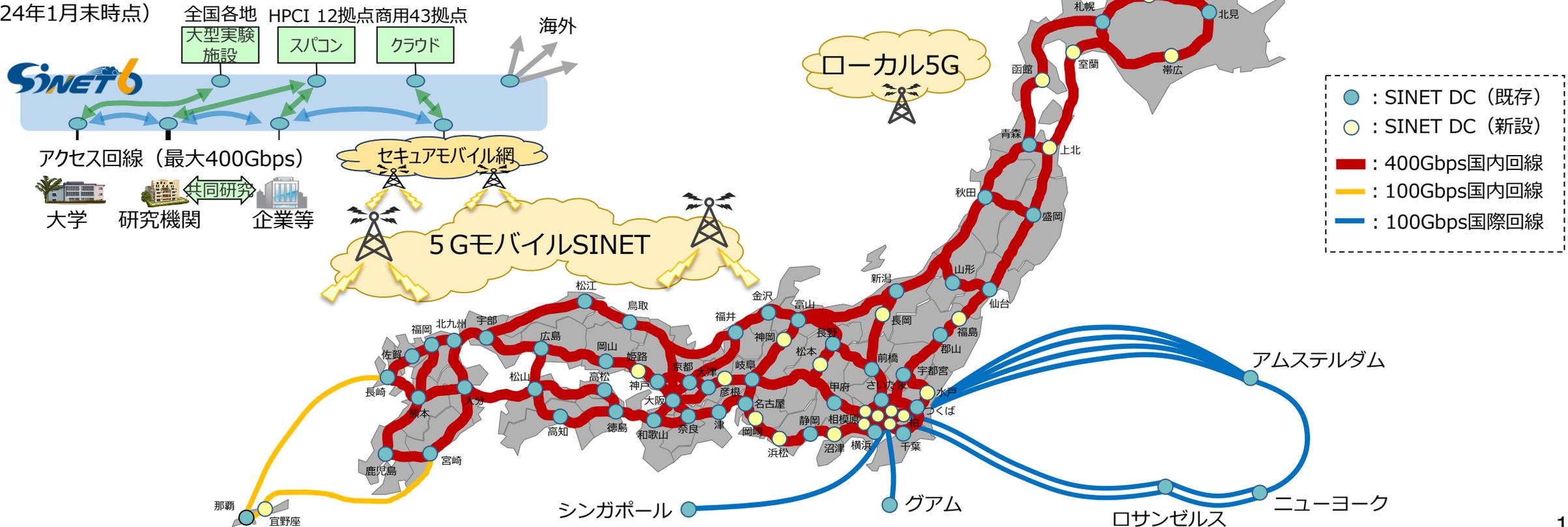


ネットワーク基盤 - SINET6

- 2022年4月より、新しいネットワーク基盤 SINET6 の運用を開始
- SINET6では、①全国400Gbps化と接続点の拡大、②5Gと400Gbpsの融合、③エッジ機能配備とサービス拡大、④国際回線の増強等を実現

	国立大学	公立大学	私立大学	短期大学	高等専門学校	大学共同利用機関	研究を目的とする組織	合計
加入機関数 (加入率)	86 (100%)	96 (94%)	441 (71%)	90 (30%)	56 (97%)	16 (100%)	231	1,016

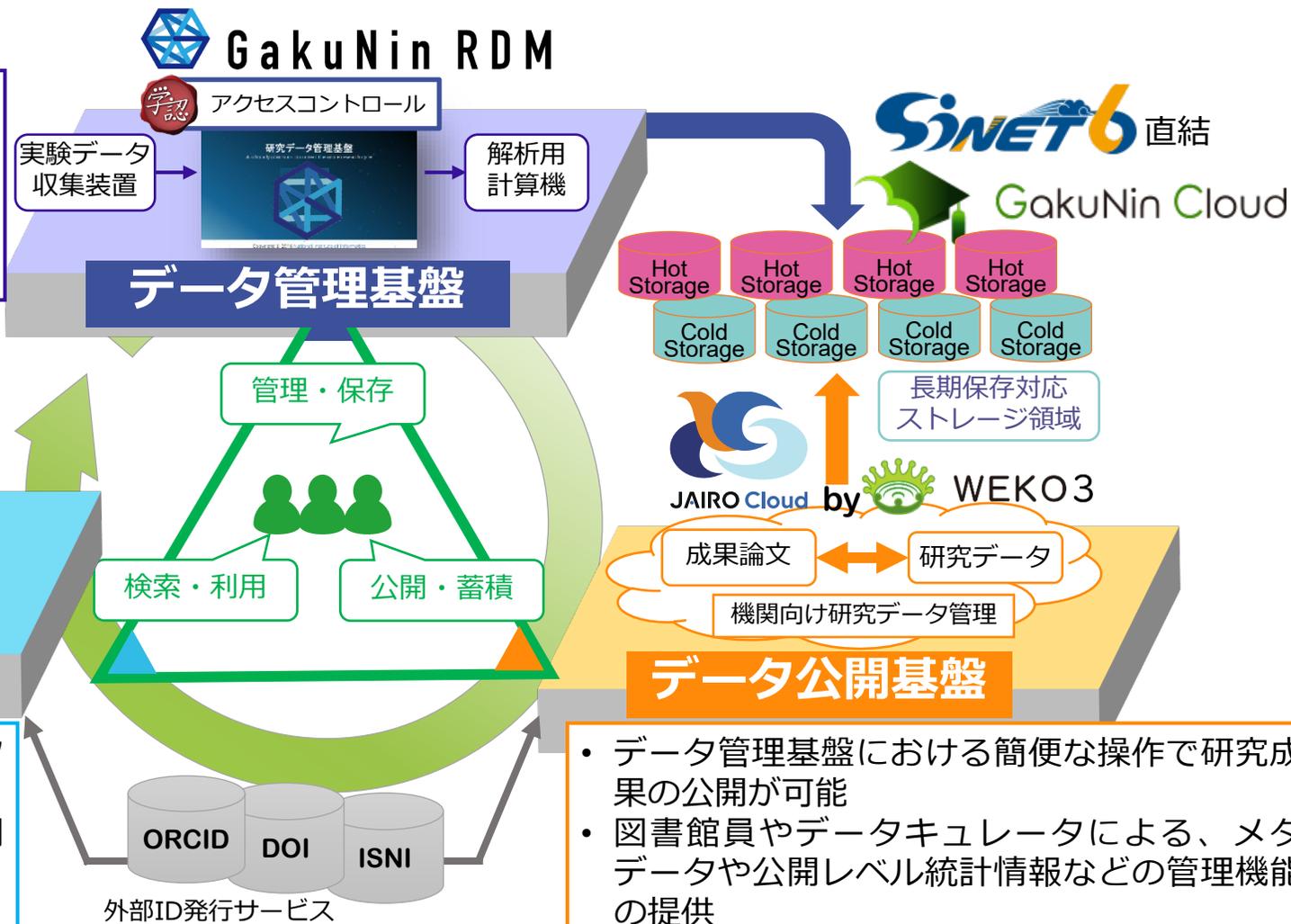
(2024年1月末時点)



研究データ基盤 - NII Research Data Cloud

2017年から開発開始 ⇒ 2021年から運用開始

- データ収集装置や解析用計算機とも連携
- 研究遂行中の研究データなどを共同研究者間やラボ内で共有・管理
- 組織が提供するストレージに接続した利用が可能



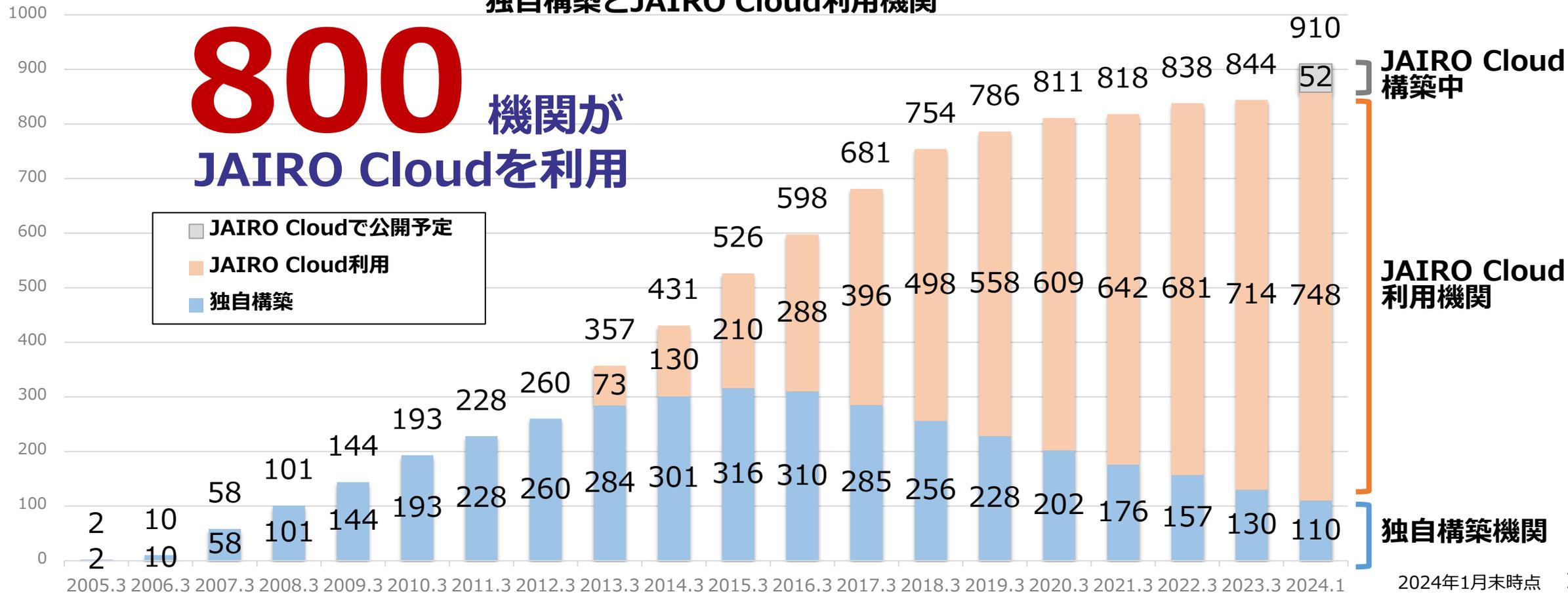
- 機関リポジトリ+分野別リポジトリやデータリポジトリとも連携
- 研究者や機関、研究プロジェクトの情報と関連付けた知識ベースを形成
- 研究者による発見プロセスをサポート

- データ管理基盤における簡便な操作で研究成果の公開が可能
- 図書館員やデータキュレータによる、メタデータや公開レベル統計情報などの管理機能の提供

機関リポジトリのクラウドサービス JAIRO Cloud

- 機関リポジトリ（研究成果を収集・保存・発信する器）のクラウドサービスJAIRO Cloudを提供
 - 大学でのシステム運用負荷の軽減により、機関リポジトリ数が拡大
 - 必要な機能を備えたシステムをクラウド上で提供することで全体を効率化
 - 学術情報のオープンアクセスを推進

日本の機関リポジトリ構築数
独自構築とJAIRO Cloud利用機関



GakuNin RDMの利用機関数の増加

120

100

101機関

80

■ 総合 ■ 理工 ■ 人文社会 ■ 医学 ■ 企業研究部門

60

40

20

0





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（2022年度～）

- 我が国の研究力の飛躍的發展を図るため、各分野・機関の研究データをつなぐ全国的な研究データ基盤の構築・高度化・実装と、AI解析等の研究データ基盤の活用に資する環境の整備を行う、研究DXの中核機関群を支援する

● 全国的研究データ基盤の構築・高度化・実装とデータの利活用

- ユーザーニーズを踏まえながら、研究データの管理・蓄積・利活用・流通といった点で適切かつ実用的な機能を確保した全国的な研究データ基盤を整備し、AI・データ駆動型研究を推進
- 構築が進む各機関・各分野のリポジトリやデータプラットフォームとの連携・接続

● 研究データ基盤の活用に係る環境の整備

- 効率的なAI活用のための、機械可読データの統一化や標準化等を含めたルール・ガイドライン整備、データマネジメント人材育成支援等、ユーザー視点に立って研究データ基盤を最大限に活用するための環境整備



ユースケース創出事業 - 採択課題一覧

- 「全国的な研究データ基盤を活用し、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」にかかる課題提案を募集

https://www.nii.ac.jp/creded/nii_ac_jp_creded.html

代表者	所属機関	課題名
原 正一郎	京都大学	人文学DXを指向する情報基盤の構築
安岡 孝一	京都大学	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
蕭 耕偉郎	九州大学	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
小財 正義	情報・システム研究機構	分野横断型データベースAMIDERの活用による次世代型データ活用スキームの構築
喜屋武 亨	琉球大学	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ミュレーションシステムの開発
吉岡 京子	東京大学	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
大向 一輝	東京大学	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
尾上 陽介	東京大学	異分野共創による史料学DXの確立
小野 寛太	大阪大学	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
木村 映善	愛媛大学	国際的なRWD研究を実現する医療情報分析基盤の検討
菊池 信彦	国文学研究資料館	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
山下 俊介	北海道大学	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
中西 秀哉	核融合科学研究所	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
林 美加子	大阪大学	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
松平 拓也	金沢大学	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
徳地 直子	京都大学	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
中村 覚	東京大学	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
佐藤 琴	山形大学	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
岡田 知久	京都大学	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
能勢 正仁	名古屋市立大学	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
佐藤 和信	大阪公立大学	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
山田 裕久	奈良工業高等専門学校	高専における分析データ集約・配信モデルシステム構築

**2024年度分
現在も随時
応募受付中！**

オープンサイエンス時代の 研究データ基盤構築に向けた国内の政策動向

- 2015年3月：内閣府「国際的動向を踏まえたオープンサイエンスに関する検討会」報告書
- 2016年1月：政府「第5期科学技術基本計画」
- 2016年2月：文部科学省 科学技術・学術審議会 学術分科会 学術情報委員会「学術情報のオープン化の推進について」
- 2016年5月：G7茨城・つくば科学技術大臣会合 つくばコミュニケ（共同声明）
- 2016年5月：政府「科学技術イノベーション総合戦略2016」
- 2016年7月：日本学術会議「オープンイノベーションに資するオープンサイエンスのあり方に関する提言」（提言）
- 2017年6月：政府「科学技術イノベーション総合戦略2017」
- 2018年6月：政府「統合イノベーション戦略」
- 2019年6月：政府「統合イノベーション戦略2019」
- 2020年6月：日本学術会議「オープンサイエンスの深化と推進に向けて」（提言）
- 2020年7月：政府「統合イノベーション戦略2020」
- 2021年3月：政府「第6期科学技術・イノベーション基本計画」
- 2021年4月：政府「公的資金による研究データ管理・利活用に関する基本的な考え方について」
- 2021年6月：政府「統合イノベーション戦略2021」
- 2022年6月：政府「統合イノベーション戦略2022」
- 2023年5月：G7科学技術大臣会合
- 2023年6月：政府「統合イノベーション戦略2023」

統合イノベーション戦略2023（2023年6月）

（公的資金による研究データの管理・利活用の推進）

「公的資金による研究データの管理・利活用に関する基本的な考え方」（令和3年4月27日統合イノベーション戦略推進会議決定）において、公的資金による研究データに関する概要情報（メタデータ）を中核的な基盤である研究データ基盤システム（NII Research Data Cloud）上で検索可能とし、オープン・アンド・クローズ戦略に基づく研究データの管理・利活用を推進するビジョンを示した。ここでは、公募型の研究資金の全ての新規公募分についてメタデータ付与を行う仕組みを2023年度までに導入するとともに、大学等の研究開発を行う機関においてデータポリシーの策定と機関リポジトリへの研究データの収載等を進めることとしている。…また、2022年度に開始された「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」において、引き続き各分野・機関の研究データをつなぐ全国的な研究データ基盤の高度化や、研究機関・研究者に対する研究データ基盤の利活用に向けた普及・広報活動を推進する。

（学術論文等のオープンアクセス化の推進）

また、本年5月に日本で開催されたG7広島サミット及びG7仙台科学技術大臣会合を踏まえ、我が国の競争的研究費制度における2025年度新規公募分からの学術論文等の即時オープンアクセスの実現に向けた国の方針を策定する。

オープンかつ日本語に強い大規模言語モデルの構築と その原理解明に取り組む組織横断プロジェクト

LLM-jp

ChatGPT

- OpenAIが2022年11月に公開した**大規模言語モデル (Large Language Model, LLM)** に基づくチャットボット
- OpenAIは2015年末にサム・アルトマン、イーロン・マスクらによって設立されたAI研究所。マイクロソフトも出資
 - 自然言語からの画像生成モデルDALL-E、音声認識モデルWhisperなども開発
 - 本社はサンフランシスコ



<https://openai.com/>

An astronaut riding a horse in a photorealistic style



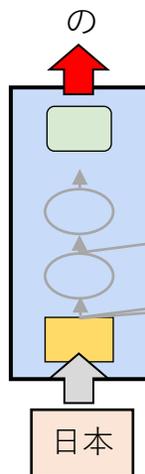
<https://openai.com/dall-e-2/>

ChatGPTとは

- OpenAIが2022年11月に公開した**大規模言語モデル (Large Language Model, LLM)** に基づくチャットボット

学習時

学習データ：日本の少子化対策には、次のようなアプローチ...

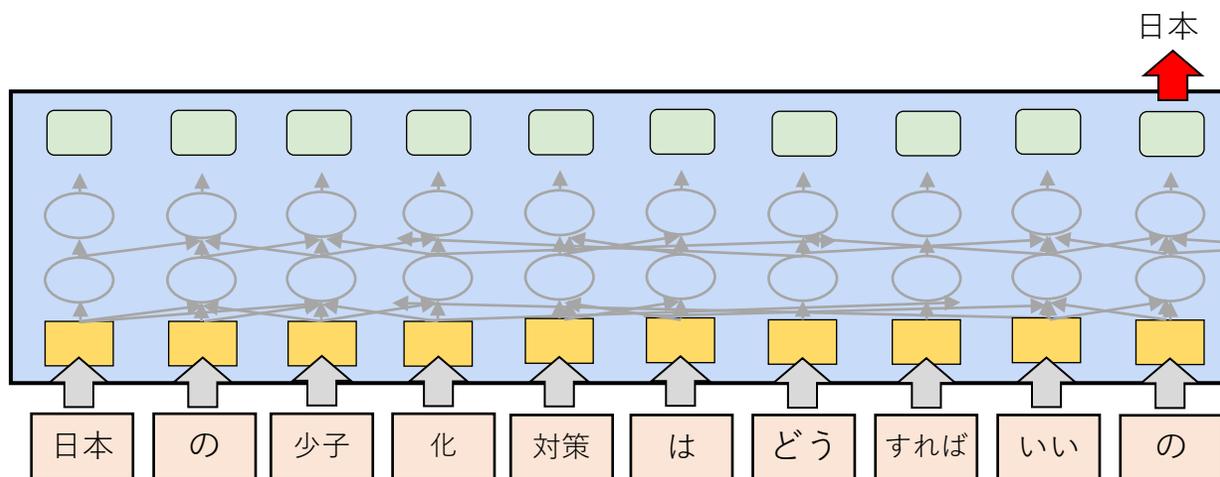


ChatGPTとは

- OpenAIが2022年11月に公開した**大規模言語モデル (Large Language Model, LLM)** に基づくチャットボット

推論時

プロンプト：日本の少子化対策はどうすればいいの



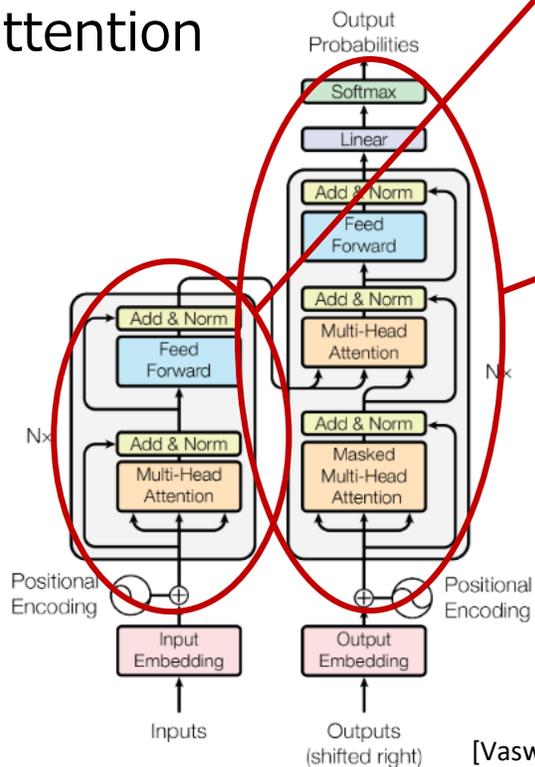
大規模言語モデル(LLM)の歴史

2014 Attention

機械翻訳において目的言語の次の語を生成する際に原言語の文のどこに着目するか

2017 Transformer

attentionの精緻化、原言語文内、目的言語内でのattention



[Vaswani et al. 2017]

2018 BERT

Transformerのencoder側を単言語の分類問題等に

2018 GPT (117Mパラメータ)

Transformerのdecoder側を言語モデルに

2019 GPT-2 (1.5Bパラメータ)

2020 GPT-3 (175Bパラメータ)

2022 GPT-3.5 / InstructGPT

2022 ChatGPT

2023 GPT-4 (2Tパラメータ?)

画像も扱える、多言語能力も大幅向上

- 米司法試験で人間受験者の上位10%の成績
- 米大学入試テストSATで1600点中1410点
- 米医師試験USMLEでも合格レベルの点数



意味・文脈の理解と多言語性



翻訳：うなぎ文

 DeepL

日本語 (自動検出) ▾ ⇄ 英語 (us) ▾ 用語集

部長はうなぎです ×

The chief is an eel.

その他の候補：

The director is an eel.

翻訳：うなぎ文

ChatGPT PLUS

GPT-4



「部長はうなぎです」を英語に訳してください



「部長はうなぎです」というフレーズは、直訳すると"The manager is an eel" となります。   

ただし、このフレーズは日常的な会話で直接使用されるものではないと思われるため、コンテキストや文脈に応じて、適切な翻訳を行う必要があるかもしれません。特定の文化的・慣用的な意味が含まれている場合、そのニュアンスを取り入れた訳が必要となることも考えられます。

翻訳：うなぎ文

ChatGPT PLUS



お店で注文している場合だとどうなりますか



お店で「部長はうなぎです」と注文する場面を想像すると、これはおそらく部長が
うなぎを注文したいという意味になります。この場合、英語での翻訳は "The
manager would like eel" もしくは "The boss orders eel" となるでしょう。



参考：『ChatGPT翻訳術 新AI時代の超英訳スキルブック』山田 優 著, 株式会社アルク 2023.

ChatGPT 4 (RAG)

The screenshot shows a web browser window with the URL `chat.openai.com/c/44b569bf-8263-4e13-976d-2ff937f89264`. The chat interface shows a user asking for the 2023 season performance of Shohei Ohtani. ChatGPT responds with a quick search and provides the following statistics:

打撃成績:

- 試合数: 135
- 打席数: 599
- 打率: .304
- 安打数: 151
- 二塁打: 26
- 三塁打: 8
- 本塁打: 44
- 打点: 95
- 四球: 91
- 敬遠: 21
- 三振: 143
- 盗塁: 20
- 出塁率+長打率 (OPS): 1.066 [1]

投手成績:

- 登板試合数: 23
- 勝利数: 10
- 敗戦数: 5
- 防御率: 3.14
- 投球回: 132.0
- 奪三振数: 167
- イニングあたりの出塁数平均 (WHIP): 1.06 [1]

この素晴らしい成績により、大谷選手は2023年に多くの記録を打ち立てました。例えば、彼はアジア人初のメジャーリーグ本塁打王に輝き、2度目のアメリカンリーグMVPを獲得しました。また、彼のOPSは両リーグで最高であり、日本人メジャーリーガーとして歴代最多塁打の記録も更新しました。大谷選手の二刀流としての活躍は、彼が同一シーズンで10勝以上と40本塁打以上を達成したことにより、メジャーリーグ史上初の偉業となりました。

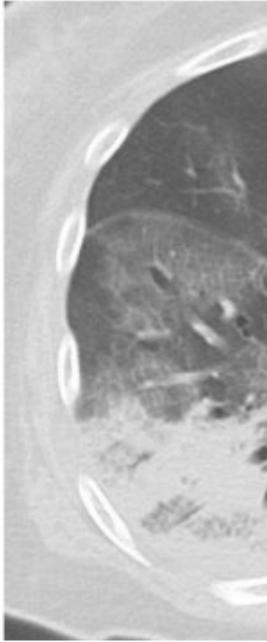
Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.

GPT-4V (画像理解)

ChatGPT 4 ▾

 You



You are a helpful radiologist. Can you identify the findings on this CT scan and suggest a differential diagnosis for COVID-19?

 ChatGPT

The image provided appears to be a CT scan of the chest. There are several radiological findings that can be observed:

1. Ground-glass opacities: These are hazy areas that do not obscure the underlying lung structures. They are seen in both lungs, predominantly in the periphery, which is a common finding in viral pneumonias, including COVID-19.
2. Consolidation: Denser opacities that can obscure lung structures. These may represent more advanced lung involvement and are also seen in both lungs.
3. Bilateral lung involvement: The abnormalities are present in both the right and left lungs, which is typical for infectious processes like COVID-19.
4. Peripheral distribution: The findings are more pronounced towards the edges of the lungs, which is characteristic of COVID-19 pneumonia.

While these findings are suggestive of COVID-19, especially in the context of the current pandemic and if correlated with clinical symptoms such as fever, cough, and shortness of breath, they are not exclusive to COVID-19. Other viral pneumonias, bacterial pneumonias, or even non-infectious processes could present with similar patterns on a CT scan.

Definitive diagnosis of COVID-19 would require a positive result from a polymerase chain reaction (PCR) test for SARS-CoV-2. Radiological findings must be correlated with clinical presentation and laboratory tests for an accurate diagnosis. It's essential to consider patient history, symptoms, and exposure risk when interpreting imaging results like these.

<https://www.covid19-jma-medical-expert-m>

LLMに関する懸念

- 研究開発が**一部の組織の寡占状態**であることは健全とはいえない。
OpenAIはもはやオープンではない
- 強い・大きなモデルの学習コーパス・データは公開されておらず、巨大パラメータのモデルの振舞いを含めて全体がblack-box。**その賢さ、多言語性について何が起きているのか誰もわからない**
- 一方、現在のモデルにはHallucination、バイアス等の課題も山積
- **日本語コーパスはGPT-3において0.11%**、日本語の理解・生成能力は英語に比べて劣る
- 国内にサーバを置き、セキュリティ認証ISMARを取得するサービスも検討されているが、**経済安全保障的な懸念**は残る





- オープンかつ日本語に強い大規模モデルを構築し、LLMの原理解明に取り組む
- モデル・データ・ツール・技術資料等を議論の過程・失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可



<https://llm-jp.nii.ac.jp/>



- オープンかつ日本語に強い大規模モデルを構築し、LLMの原理解明に取り組む
- モデル・データ・ツール・技術資料等を議論の過程・失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可

2023.5

自然言語処理の研究者の有志約30名による勉強会を開催

2023.10

mdxを用いて130億パラメータの大規模言語モデル「LLM-jp-13B」を公開

2023.11

ABCI第2回LLM構築支援プログラム採択1750億パラメタモデルの学習に着手

2024.1

GENIAC 第1期採択（2024.4から1750億パラメタモデルを学習）

- **mdx**: データ活用社会創成プラットフォーム. 9大学2研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境
- **ABCI**: AI橋渡しクラウド. 産業技術総合研究所 (AIST) が提供するAI向け計算用で現状国内最大の計算資源
- **GENIAC**: Generative AI Accelerator Challenge. 日本国内の基盤モデル開発力の底上げのために計算資源の提供等を行う経産省のプログラム

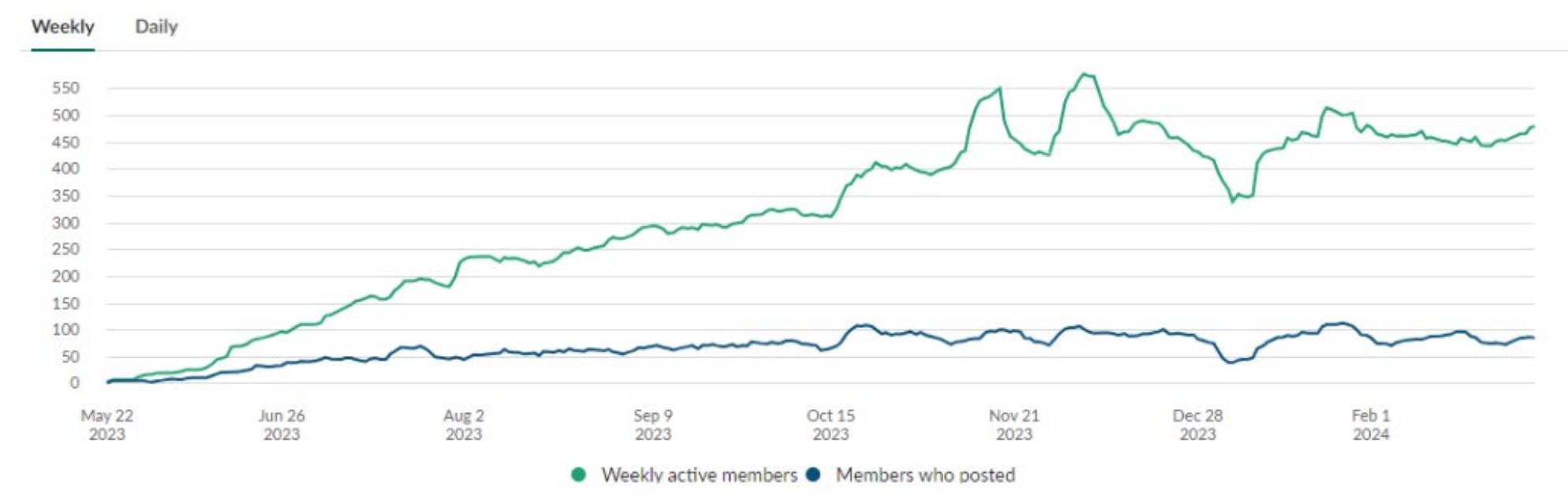
1008

Saku Sugawara (NII) 11:27 AM Friday, February 2nd
<https://llm-jp.nii.ac.jp/llm/2024/02/02/geniac-announce.html> 勉強会ホームページでも告知しました。

LLM 勉強会
国立情報学研究所（情報・システム研究機構）が経済産業省の計算資源支援事業に採
択されました
国立情報学研究所（情報・システム研究機構）が経済産業省の計算資源支援事業に採

Active members in your organization

See how many people are active – meaning they posted a message or read at least one channel or direct message.



- Threads
- Drafts & sent
- Channels
- # abci_第2回llm構築支...
- admin
- # ask_anything
- # general
- llmセンター特任研
- # mdx利用申請
- meti計算資源
- nii-llm研究開発セ
- # playground
- playground-privat
- # playground検討
- # random
- # tech-blog
- # コーパス構築
- # コーパス構築_コ
- # チューニング
- # チューニング_チュ...
- # モデル構築

More than 160 organizations

Academia / Public sector

北海道大学
北見工業大学
東北大学
山形大学
筑波大学
東京大学
慶應義塾大学
早稲田大学
東京工業大学
東京学芸大学
お茶の水女子大学
日本女子大学
上智大学
電気通信大学
明治大学
一橋大学
東京都立大学
東京通信大学
武蔵野大学
名古屋大学
京都大学

大阪大学
大阪工業大学
関西学院大学
近畿大学
奈良先端科学技術大学院大学
北陸先端科学技術大学院大学
愛知工業大学
長岡技術科学大学
千葉大学
愛媛大学
九州工業大学
九州大学

MBZUAI
沖縄科学技術大学院大学
国立情報学研究所
海洋研究開発機構
理化学研究所
産業技術総合研究所
情報通信研究機構

科学技術振興機構
国立教育政策研究所
NPO法人AI開発推進協会
広島県庁
北九州市役所
京都府庁

Industry

NTT
LINEヤフー
レトリバ
サイバーエージェント
富士通
Microsoft
Studio Ousia
プレジジョン
ZENKIGEN
Legalscape
Turing
アマゾン ウェブ サービス ジャパン
みらい翻訳
リクルート **Megagon Labs**
グロービス経営大学院
ストックマーク
BreakAI
ファーストアカウンティング
東芝
Preferred Networks
オムロンサイニクエックス
トヨタ自動車
NTTコミュニケーションズ
バオバブ
Polaris.AI
Sakana AI
Stability AI Japan
マネーフォワード
メルカリ
NVIDIA合同会社
アステラス製薬
パスコ
朝日新聞社
Rakuten Group, Inc.
ELYZA
ベルシステム24ホールディングス
Lightblue

インテル
インターグループ
フィックスターズ
さくらインターネット
ソニーグループ
カラクリ
GMOインターネットグループ
みずほ第一フィナンシャルテクノロジー
丸紅
アドバンスソフト
クラウドエース
Insight Edge
ライフジェンクス
アルジェントテクノロジー
KDDI総合研究所
デロイトトーマツコンサルティング
日本総合研究所
フリットジャパン
オルツ
インフォセンス
データブリックス・ジャパン
レアゾン・ホールディングス
IHI
スクウェア・エニックス
webbigdata
日本IBM
STORIA法律事務所
neoAI
三菱電機
デジタル・アドバタイジング・コンソーシアム
LAPRAS
NTT東日本
クラスター
本田技研工業
サイボウズ・ラボ
Laboro.AI
アミフィアブル
住友電工情報システム

Anique
ユーザベース
ディー・エヌ・エー
アクセンチュア
HEROZ
地層科学研究所
野村総合研究所
Algomatic
ユビー
ベイススコンサルティング
ブレインパッド
アラヤ
NTTデータグループ
新医療リアルワールドデータ研究機構
紀伊國屋書店
日本能率協会マネジメントセンター
インターネットイニシアティブ
リコー
エル・ティー・エス
A.M Recherche
TOPPANデジタル
ソフトバンク
カサナレ
NTTデータイントラマート
三菱ケミカルグループ
ABEJA
創英知財ソリューションズ
Gaudiy
セコムIS研究所
ヤマハ
APTO
Japan Digital Design
コニカミノルタ
SambaNova Systems
PwCコンサルティング
Alibaba Cloud Japan
APTO

Large-scale corpus

- 日本語語彙をきっちり扱うトークナイザ整備
- LLM-jp-13Bでは、ウェブコーパス等をクリーニングし、日本語1500億トークン、英語1500億トークン、プログラムコード100億トークンを事前学習に使用
- コーパス開拓・整備を継続(NDL、JST等の資源)

1008

Large-scale computational infrastructure

- LLM-jp-13Bではmdx A100 96GPUを利用
- 高速並列計算環境、種々のLLM構築プラットフォームを比較検討
- LLMの大規模計算を安定して行うノウハウを蓄積
- 新たなモデルアーキテクチャを随時検討

Tuning data

Evaluation data

- LLM-jp-13Bでは12種類、約20万件のインストラクションデータを整備し、チューニングを実行
- 9種類の日本語評価データと、評価環境llm-jp-evalを整備(GENIACでも利用されている)
- インストラクションデータ、評価データの検討・構築を継続するとともに、LLMの安全性について分析・検討

- Home
- DMs
- Activity
- Later
- More
- Channels
 - abci_old_第2回...
 - abci_第2回llm...
 - admin
 - general
 - mdx利用申請
 - meti計算資源...
 - # playground**
 - # playground検討
 - # random
 - # tech-blog
 - # コーパス検討
 - # コーパス検討_...
 - # チューニング...
 - # チューニング...
 - # モデル構築検討
 - # モデル構築検...
 - # 学術ドメイン...
 - # 安全性検討

playground

148

Hirokazu Kiyomaru (京大黒橋研) 7:46 PM
 コーパス検索の Slack デモを更新しました！
 • 全訓練事例が検索対象になりました。
 • ElasticSearch の検索結果と合わせて Vector 検索の結果も返却されるようになりました。
 • メッセージに 🔍 (mag) のリアクションをつけると、そのメッセージをクエリとして検索が走るようになりました。
 最後の機能はどういう形にするのが使いやすいか模索中です。何かご意見があれば #playground検討 にお寄せください 🙌

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

llmjp-13b-finetuned APP 7:48 PM
 いや、水棲哺乳類であるナマズが地震を引き起こすという理論を裏付ける科学的根拠はない。科学者たちはナマズが地震を誘発することはないと考えている。ナマズが地震を引き起こすという説は、水棲哺乳類は地球の地殻プレートに関与しており、ナマズが海底で地震を引き起こし、それが陸上に跳ね返ってナマズの生息地に被害を与えるという考えに基づいている。しかし、このような地殻プレートへの関与は、地球上のいかなる種に対しても一般的に言えることではない。ナマズに似た動物も地震を引き起こす可能性はあるが、ナマズとはまったく異なる能力を持っている。例えば、ナマズに似た魚類であるウナギは、地震に関与する可能性は低い。最後に、地震と地震を引き起こす地球の地殻プレートとの関係については、まだ決定的な科学的理解が得られていない。

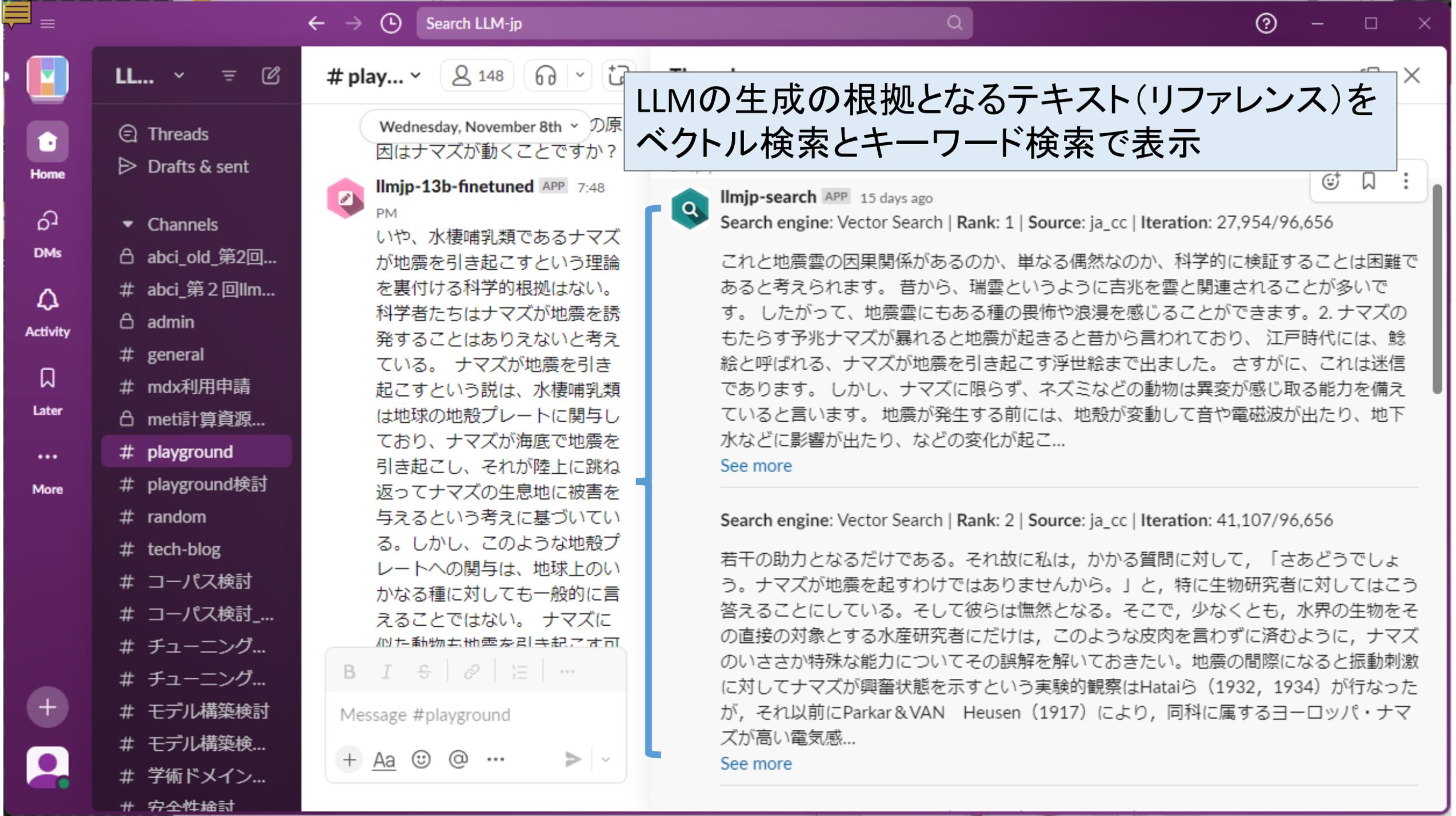
🔍 1 🗨️
 🟢 1 reply 15 days ago

Thursday, November 9th

B I

Message #playground

+ Aa 😊 @ 📎 🗣️ 📄



LLMの生成の根拠となるテキスト(リファレンス)をベクトル検索とキーワード検索で表示

Wednesday, November 8th の原因はナマズが動くことですか？

llmjp-13b-finetuned APP 7:48 PM

いや、水棲哺乳類であるナマズが地震を引き起こすという理論を裏付ける科学的根拠はない。科学者たちはナマズが地震を誘発することはないと考えている。ナマズが地震を引き起こすという説は、水棲哺乳類は地球の地殻プレートに参与しており、ナマズが海底で地震を引き起こし、それが陸上に跳ね返ってナマズの生息地に被害を与えるという考えに基づいている。しかし、このような地殻プレートへの参与は、地球上のいかなる種に対しても一般的に言えることではない。ナマズに似た動物も地震を引き起こす可

B I S | | | ...
Message #playground
+ Aa 😊 @ ... ▶ | ▾

llmjp-search APP 15 days ago

Search engine: Vector Search | Rank: 1 | Source: ja_cc | Iteration: 27,954/96,656

これと地震雲の因果関係があるのか、単なる偶然なのか、科学的に検証することは困難であると考えられます。昔から、瑞雲というように吉兆を雲と関連されることが多いです。したがって、地震雲にもある種の畏怖や浪漫を感じることができます。2. ナマズのもたらす予兆ナマズが暴れると地震が起きると昔から言われており、江戸時代には、鯰絵と呼ばれる、ナマズが地震を引き起こす浮世絵まで出ました。さすがに、これは迷信であります。しかし、ナマズに限らず、ネズミなどの動物は異変が感じ取る能力を備えていると言います。地震が発生する前には、地殻が変動して音や電磁波が出たり、地下水などに影響が出たり、などの変化が起こ...

[See more](#)

Search engine: Vector Search | Rank: 2 | Source: ja_cc | Iteration: 41,107/96,656

若干の助力となるだけである。それ故に私は、かかる質問に対して、「さあどうでしょう。ナマズが地震を起すわけではありませんから。」と、特に生物研究者に対してはこう答えることにしている。そして彼らは憮然となる。そこで、少なくとも、水界の生物をその直接の対象とする水産研究者にだけは、このような皮肉を言わずに済むように、ナマズのいささか特殊な能力についてその誤解を解いておきたい。地震の間際になると振動刺激に対してナマズが興奮状態を示すという実験的観察はHataiら(1932, 1934)が行なったが、それ以前にParkar & VAN Heusen (1917)により、同科に属するヨーロッパ・ナマズが高い電気感...

[See more](#)

Fine-tuning of LLM-jp-13B v1.1

1. Supervised Fine-tuning (SFT)

Instruction:

ポテトチップスの袋はなぜ開封後に古くなるのか？

Response:

ポテトチップスの袋は窒素で満たされている！多くの消費者は、ポテトチップス会社が袋の4分の3を空気で満たし、お金を取ろうとしていると考えているが、実はこれは...

Data Stats		En	Ja
SFT	Dolly	15,011	15,011 (mt)
	OpenAssistant	21,164	21,164 (mt)
	Ichikara	---	2,903 (org)
LHF	HH-RLHF	---	12,000 (mt)

2. Learning from Human Feedback (LHF)

using Direct Preference Optimization (DPO) [Rafailov+ 2023]

Instruction:

父とは疎遠なのですが、もう一度連絡を取りたいと思っています...

Response:

メールが一番簡単だと思います。「一緒に過ごした楽しい時間は一生忘れない」とか、そういうことを付け加えてもいいかも。



Response:

メールにしたほうが良いと思う。なぜ聞くのですか？他の方法のほうが良いと思う理由があるのですか？

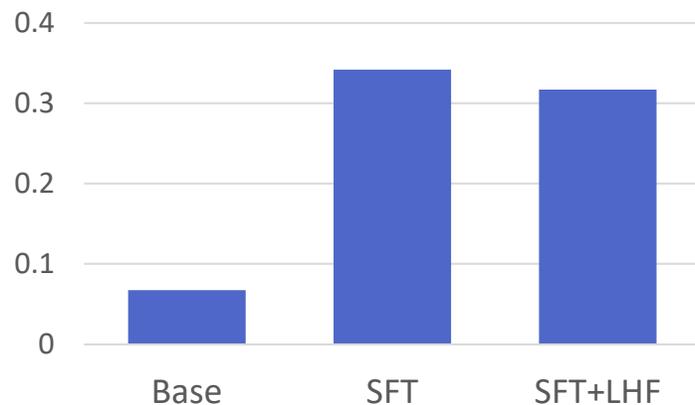




Evaluation Results of LLM-jp-13B v1.1

LLM-jp eval:

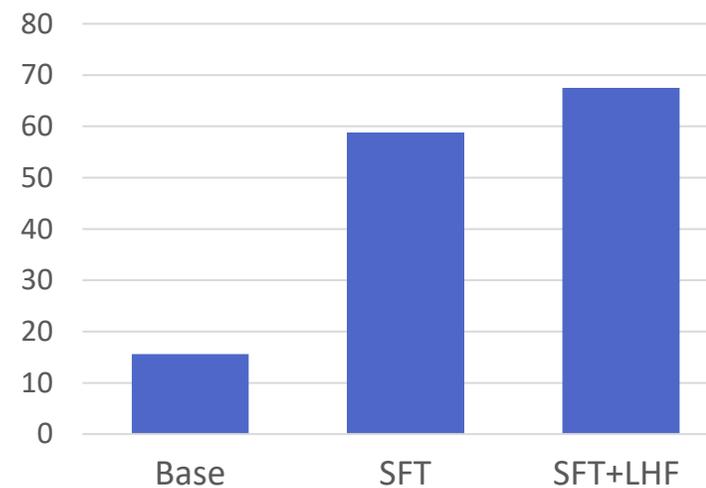
a wide range of NLP tasks, such as QA, recognizing textual entailment, and others (4-shot)



※ Results for 4-shot

Japanese Vicuna QA:

open-ended questions (GPT-4 rates the responses; win ratio against GPT-3.5)

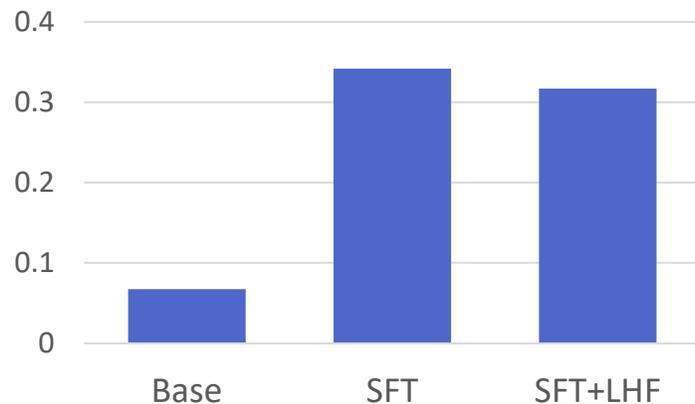


Evaluation Results of LLM-jp-13B v1.1 and v2.0

- Improve corpus cleaning
- LLaMa architecture

LLM-jp eval:

a wide range of NLP tasks, such as QA, recognizing textual entailment, and others (4-shot)



※ Results for 4-shot

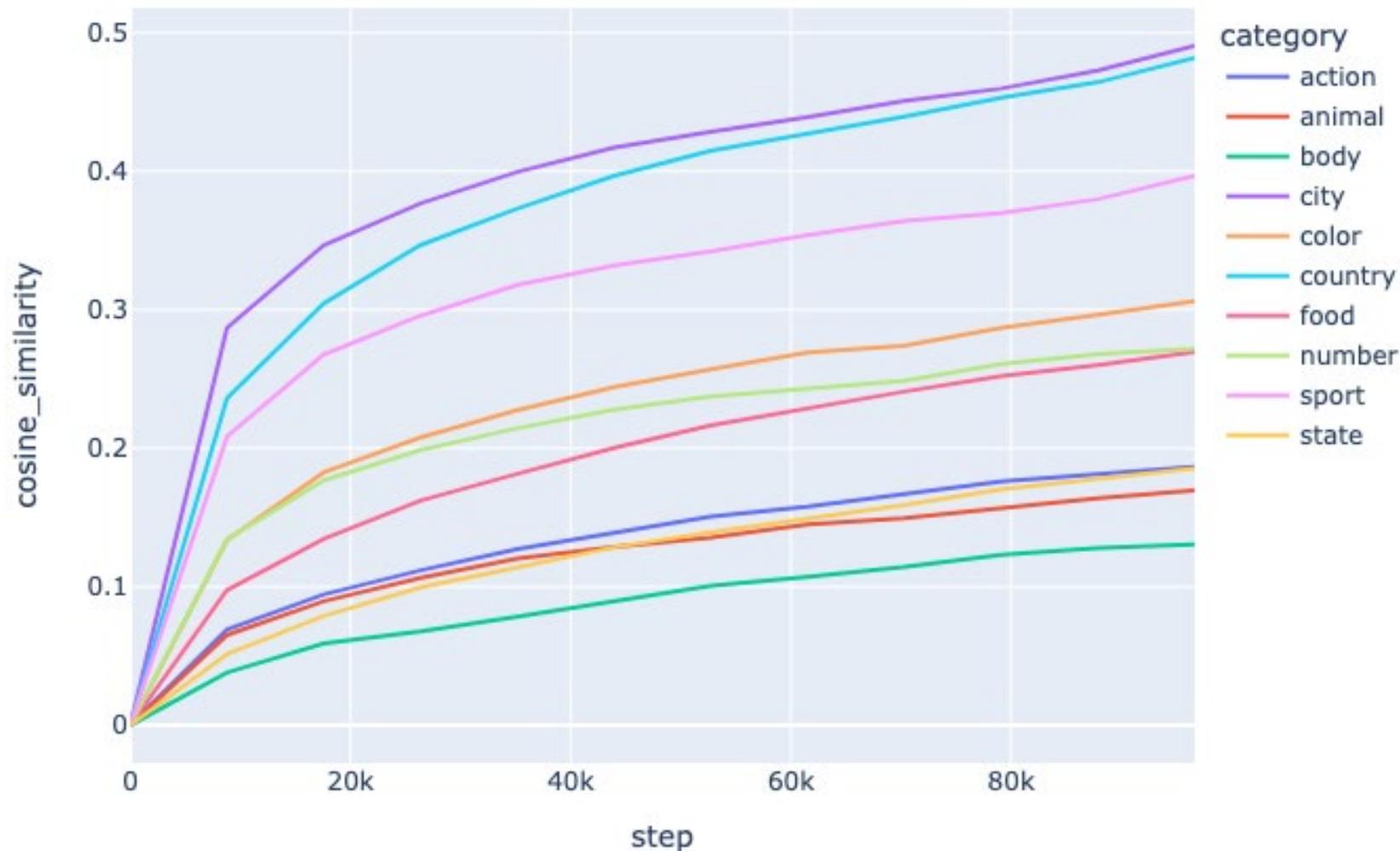
Japanese Vicuna QA:

open-ended questions (GPT-4 rates the responses; win ratio against GPT-3.5)



Investigating the Multilinguality of LLM-jp-13B

- Investigated the transition of the similarity between corresponding J/E words (e.g., "apple" vs "りんご")
- Manually prepared 10 pairs of words for each of the 10 categories, including "action", "animal", etc.
- Despite no explicit training signals, their similarities (of embeddings) increase steadily



LLM研究開発センター

- 生成 A I モデルの透明性・信頼性の確保に向けた研究開発拠点を2024年4月に国立情報学研究所内に設置
- 研究者・エンジニア30名程度
- 計算資源 H100 500枚程度 (予定)

研究開発課題

1. 研究開発用LLM構築
2. LLMの透明性・信頼性確保に向けた研究開発
3. LLMの高度化に向けた研究開発



今後の展望

- 言語は知の基盤であり、コンピュータがLLMという形でその理解と生成の基盤をもったことの意味ははかりしれない
- LLMの知識基盤をもとに、画像・映像の理解と生成 (Large Multimodal Mode; LMM)、ロボット制御なども加速度的に進展している
- 学術においてもビジネスにおいても、データ駆動による新たな価値創造が始まっているが、LLMはそれを**知識駆動**へと昇華させる

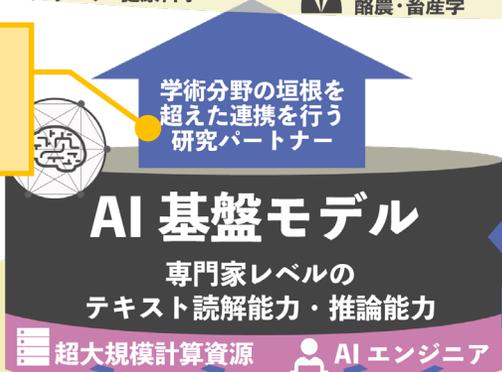


データ基盤から知識基盤へ

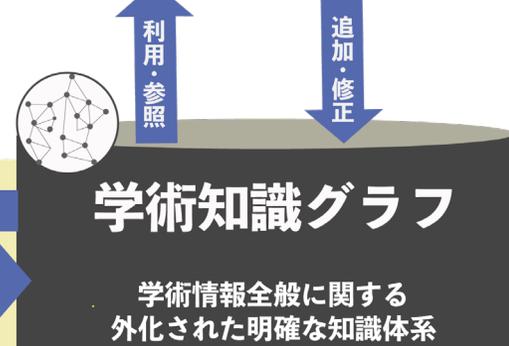
日本学術会議「未来の学術振興構想」の策定に向けた「学術の中長期研究戦略」に提案（2022年12月16日）



AI基盤モデルが出力する情報の信頼性を担保する知識トレーサビリティ



AI基盤モデル構築に必要なデータの信頼性・信憑性を確保する技術



研究データ基盤の機能充実



異分野間の高度認証連携

国際協調に基づく非地上形ネットワークの構築による学術分野の発展・開拓



クラウド・エッジサーバ・デバイス間のデータ収集及び資源最適化

ネットワーク状態の高度診断・障害予兆検知

If you want to go **fast**, go **alone**

If you want to go **far**, go **together**

