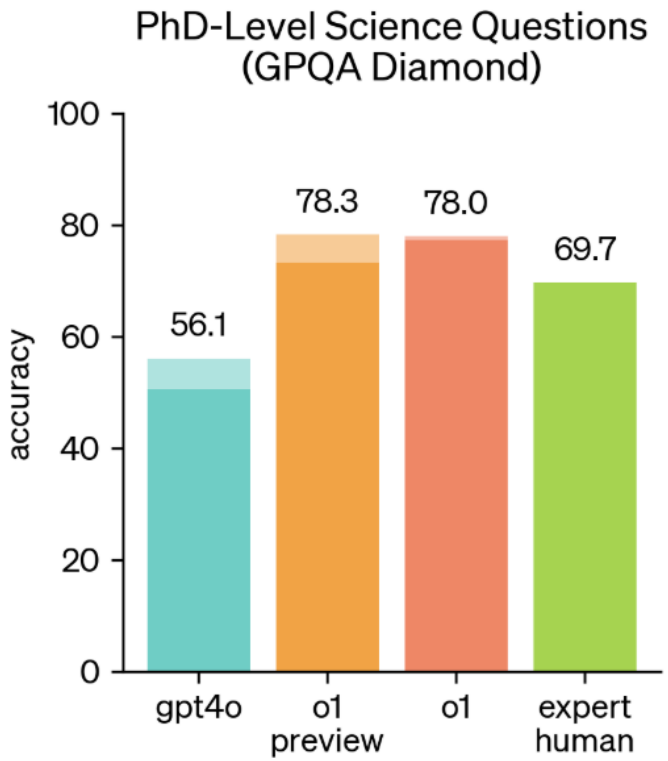
資料６
科学技術・学術審議会
情報委員会（第41回）
令和７年４月24日

「生成AIモデルの透明性・信頼性の確保に向けた研究開発拠点形成事業」令和６年度実績報告

黒橋 禎夫
国立情報学研究所（NII） 所長

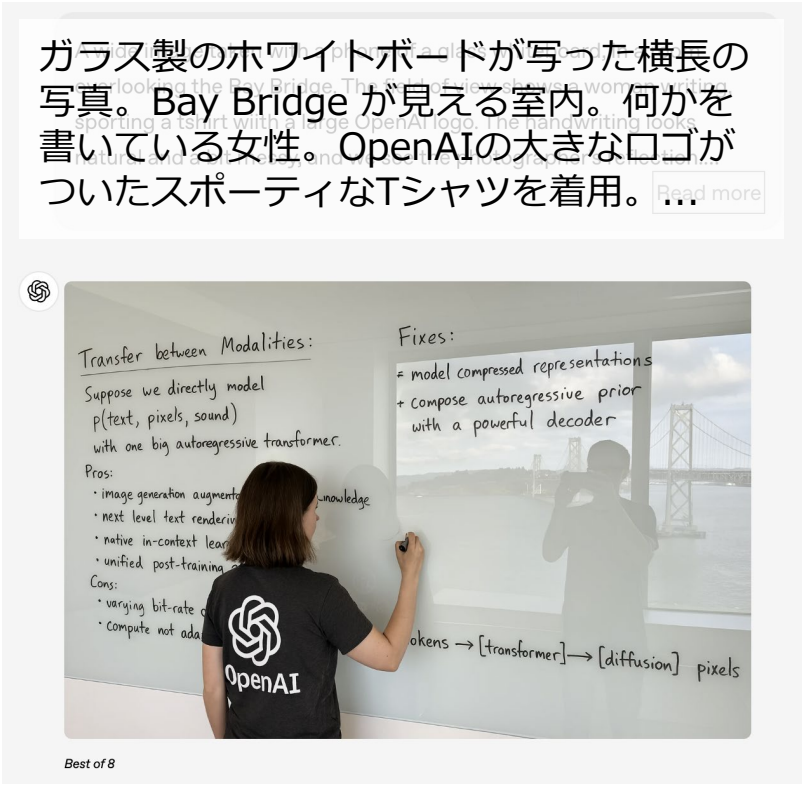
生成AI研究の最新の動向

複雑な推論
OpenAI o1、DeepSeek R1など



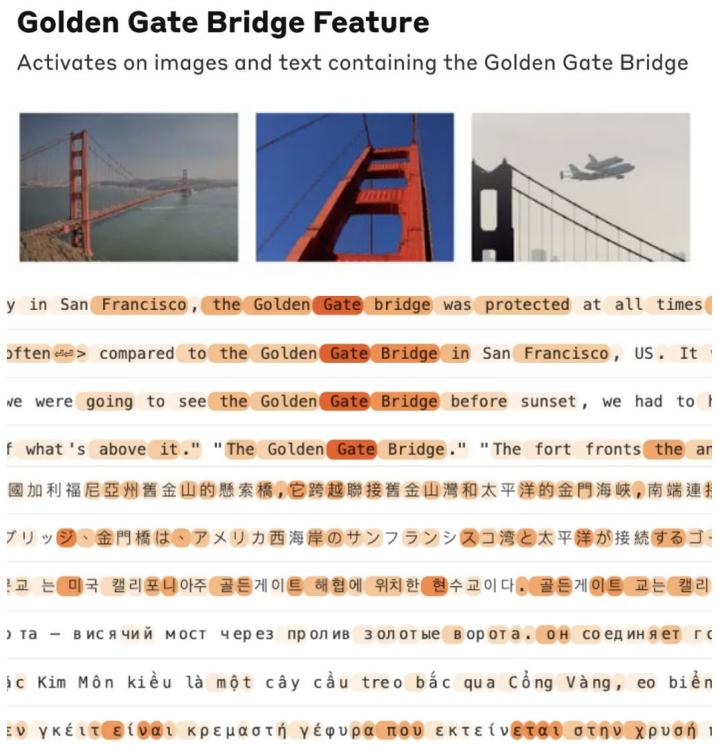
https://openai.com/ja-JP/index/learning-to-reason-with-llms

マルチモーダル
GPT-4o、Gemini-2.0など



https://openai.com/index/introducing-4o-image-generation

原理の解明
機械論的解釈可能性など



https://www.anthropic.com/research/mapping-mind-language-model

言語モデル (Language Model)

私 は りんご を ？

大規模コーパス（たとえば3,000億単語）

…お店で**私はりんごを**食べた…

…彼と**私はりんごを**食べた…

…昨日**私はりんごを**かじった…

…**私はりんごを**食べた後…

…**私はりんごを**殴った夢を…

$$P(\text{食べた} \mid \text{私 は りんご を}) = \frac{\text{コーパス中の頻度（私 は りんご を 食べた）}}{\text{コーパス中の頻度（私 は りんご を）}}$$

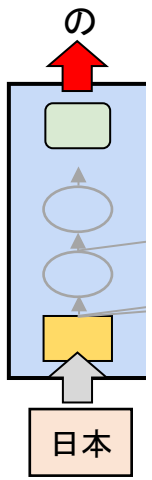
$$P(\text{殴った} \mid \text{私 は りんご を}) = \frac{\text{コーパス中の頻度（私 は りんご を 殴った）}}{\text{コーパス中の頻度（私 は りんご を）}}$$

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...

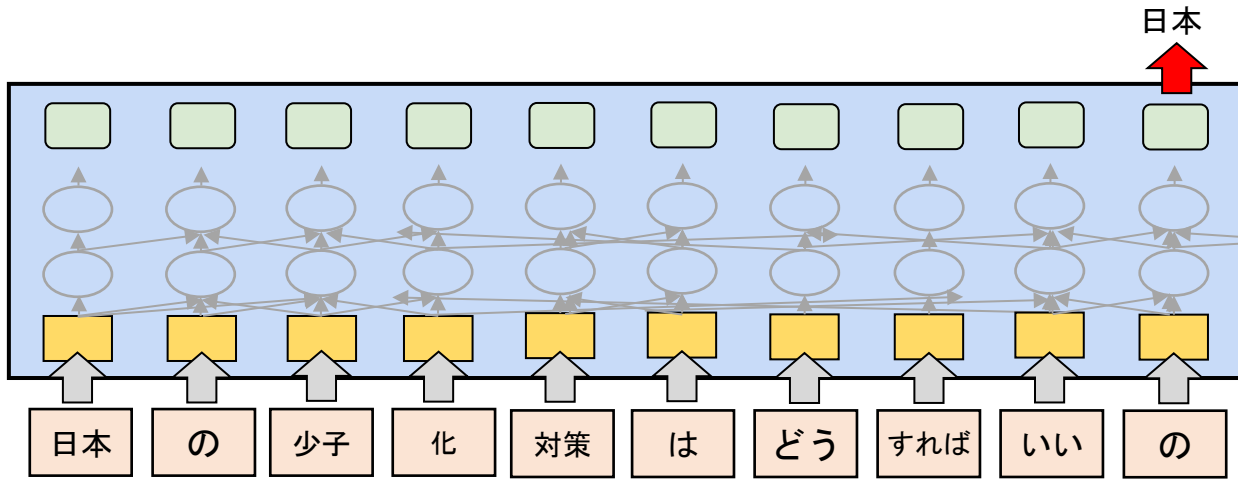


ChatGPTとは

- 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



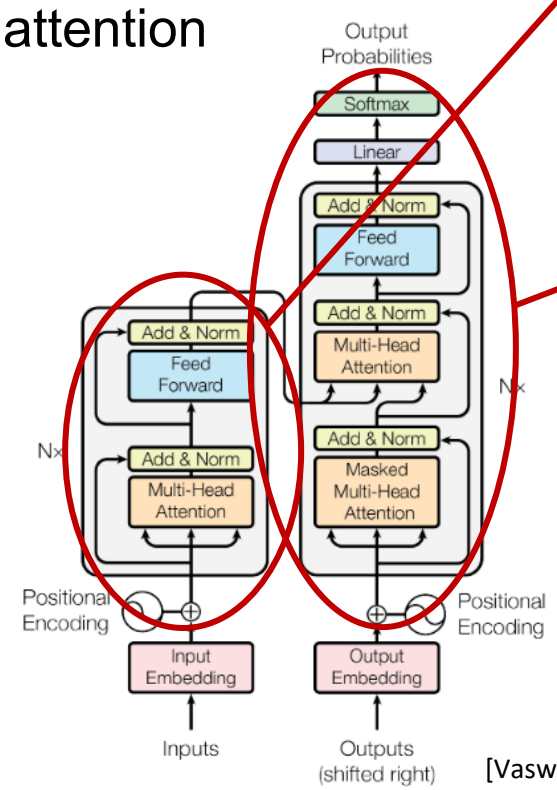
LLMの歴史

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 (1.17億パラメータ)

Transformerのdecoder側を言語モデルに

2019 GPT-2 (15億パラメータ)

2020 GPT-3 (1750億パラメータ)

2022 GPT-3.5 / InstructGPT

2022 ChatGPT

2023 GPT-4 (2兆パラメータ?)

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

LLMの発展（2023～2024年度）

2023.6 OpenAI APIに**function calling**(関数呼び出し)を追加。ChatGPTを利用した階層的なプランやコードの生成・実行などが可能に

2023.11-12 OpenAI社GPT-4 TurboやGoogle社Geminiなどを発表。仏Mistral AI社が**混合エキスパート**(MoE)方式のオープンLLM Mixtral 8x7Bモデルを公開。LLMの世代交代が加速

2024.3 Microsoft社がAzure OpenAI ServiceでOn Your Data機能の一般提供を開始。GPT-4などのLLMに**検索拡張生成**(RAG)が利用可能に

2024.4 Meta社が**オープンLLM**としてLlama3を公開(8B, 70Bモデル。405Bモデルは7月公開)

2024.5 OpenAI社が**音声・画像処理**能力を強化したモデルGPT-4oを公開

2024.6 Google社が**コンテキスト長**2MトークンのGemini 1.5 Proを公開

2024.9 OpenAI社が**推論能力**を強化したモデルo1を発表

2024.12 中国DeepSeek社が総パラメータ数671B(有効パラメータ数37B)のDeepSeek V3を公開。MoE方式で、257個のエキスパートから、推論時に9個のエキスパートを選択して実行

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 Anthropic社が**自律的コード生成**が可能なClaude Codeを発表

2025.2-4 OpenAI社 o3、 GPT-4.5、 GPT-5、Meta社Llama4、Google社Gemini 2.0 Proなど多数のモデルが登場

我が国の生成AI研究の現状

- 我が国は、リスク対策に十分に配慮しながら、生成AIの研究開発や社会での活用を積極的に進め、人類とAIの共存社会のデザインで世界をリードすべきである

  cf. 日本学術会議提言「生成 AI を受容・活用する社会の実現に向けて」(2025/2/27)

- 企業の取組については、
  - GENIACの計算資源支援＋コミュニティ構築支援は機能しているが，日本のスタートアップがどこまでいけるかは未知数
  - 日本の大企業の小規模モデルのアプローチは再検討を強いられる可能性がある
  - 日本企業の中では PFN，SB Intuitions，Elyzaが有望
  - SB OpenAI Japanの動向には注目
- 世界的に**ソブリンAI**の重要性が再認識されている
  - Proprietaryなモデルではなく、（特にデータの）透明性や（入出力の秘匿性が保たれる意味での）信頼性のあるモデル
  - 自国の文化・歴史・活動への理解

LLM-jpの立ち上げについて

- 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
- 産学の多様なメンバーが参加

2023.5
自然言語処理の研究者
30名程度による勉強会を開催

2023.10
mdxを用いて
130億パラメータモデル
LLM-jp-13Bを公開

2023.11
ABCI第２回LLM構築支援プログラム に参加
1750億パラメータモデル
の学習トライアル

2024.4
GENIAC環境で
1720億パラメータモデル
の学習開始

国立情報学研究所（NII）に
LLM研究開発センター（LLMC）設置
文部科学省「生成AIモデルの透明性・信頼性の確保に向けた研究開発拠点形成」事業

2000名超

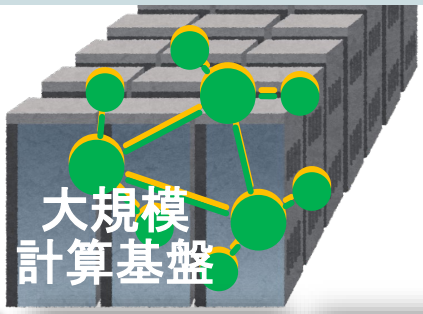
- **mdx**: データ活用社会創成プラットフォーム．9大学2研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境
- **ABCI**: AI橋渡しクラウド．産業技術総合研究所（AIST）が提供するAI向け計算用で現状国内最大の計算資源
- **GENIAC**: Generative AI Accelerator Challenge.日本国内の基盤モデル開発力の底上げのために計算資源の提供等を行う経産省のプログラム

LLM研究開発はビッグサイエンス

コーパス構築WG

河原大輔教授
（早稲田大学）

モデル構築WG



横田理央教授
（Science Tokyo）

鈴木潤教授
（東北大）

田浦健次朗教授
（東大）

チューニング評価WG

宮尾祐介教授（東大）

安全性WG

関根聡特任教授
（NII）

マルチモーダルWG

岡崎直観教授
（Science Tokyo）

実環境インタラクションWG

尾形哲也教授
（早稲田大学）

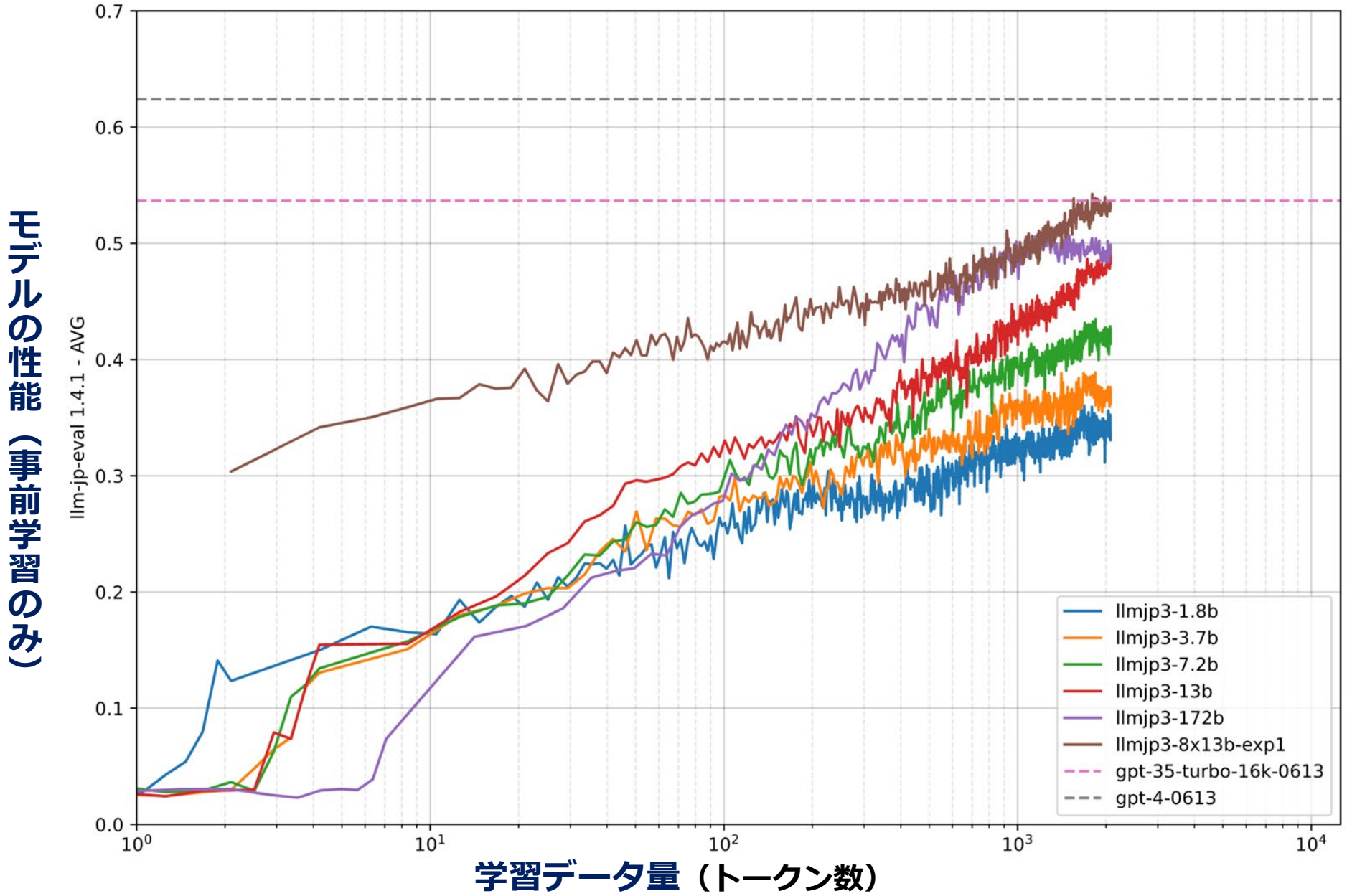
原理解明WG

大関洋平准教授
（東大）

対話WG

東中竜一郎教授
（名大）

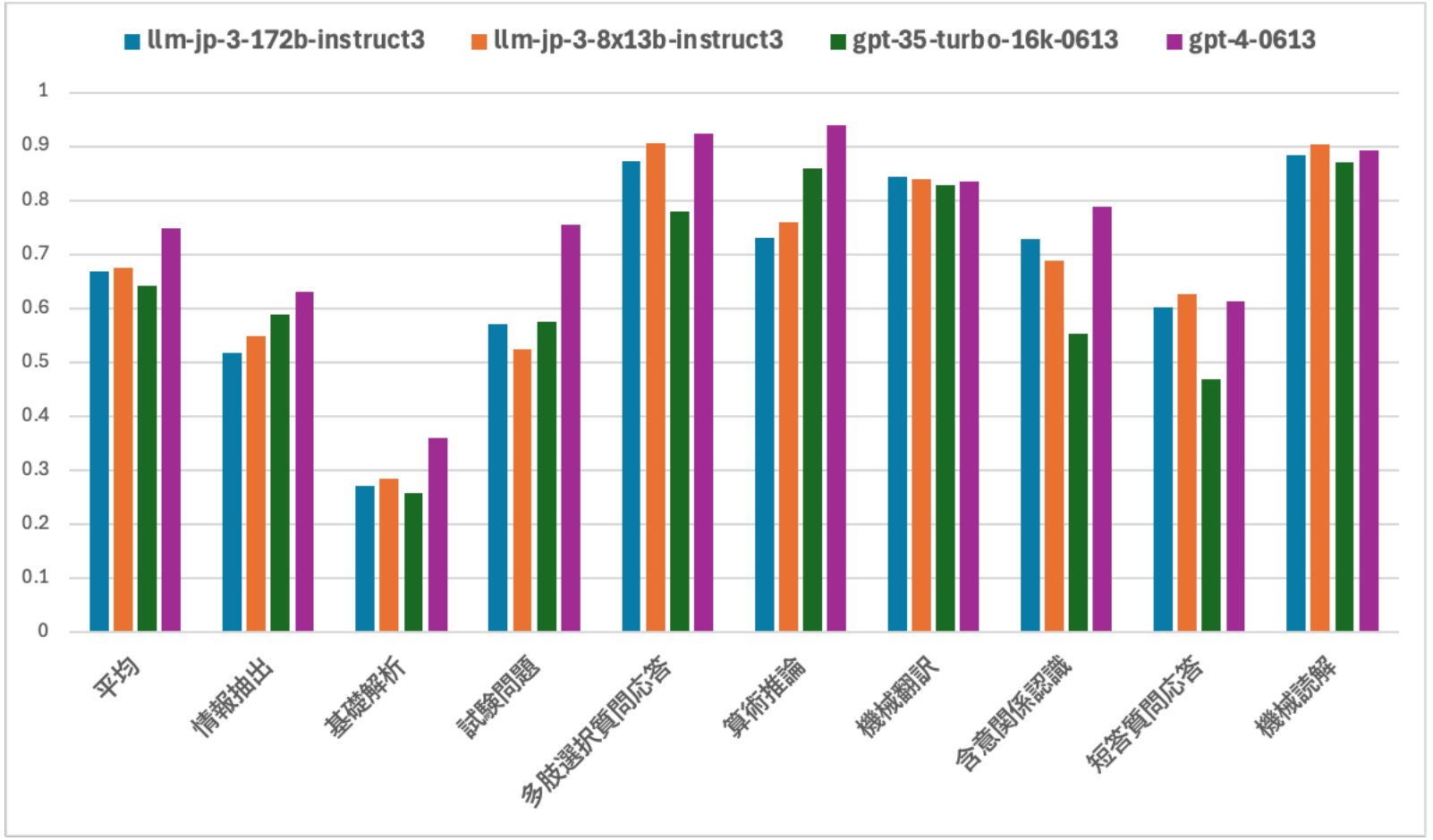
R6年度に構築したモデルと性能について



R6年度に構築したモデルと性能について

- 172B dense モデルと 8x13B の MoE モデルを公開
  - ○ 同一コーパスで訓練した小規模パラメータモデルも公開

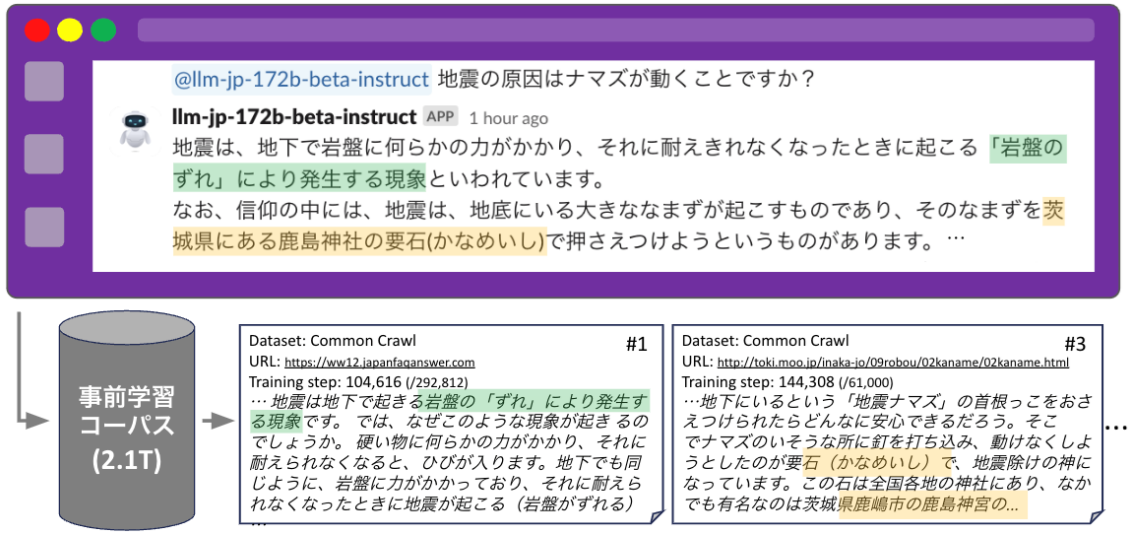
- 172B-instruct3 や 8x13b-instruct3 の性能（llm-jp-eval v1.4.1）は平均値で GPT-3.5を超え、いくつかのタスクではGPT-4.0に近づいている



各WGでの透明性・信頼性の確保に資する取り組みについて

コーパス構築WG

- 事前学習用コーパスllm-jp-corpus-v3の公開
- 日本語文書に有害性のラベルを付与したデータセット LLM-jp Toxicity Dataset v2の公開(安全性WGと連携)
- 事前学習コーパスの検索システムを構築
  - ユーザーによる生成テキストの信頼性評価の補助
  - 著作物に対する適切なクレジットを表示
  - ハルシネーションの分析



モデル構築WG

- モデルパラメータ・学習スクリプトの公開
  - Denseモデル: llm-jp-3 172B、13B、7.2B、1.8B
  - MoEモデル: llm-jp-3-8x13B、8x1.8B

安全性WG

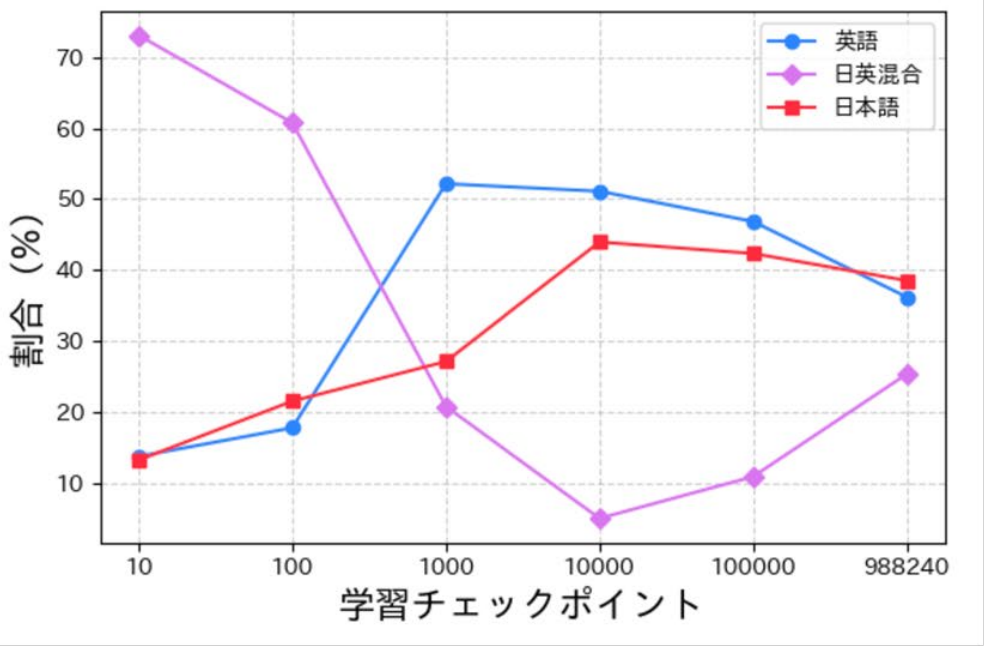
- 日本語LLM 出力の安全性・適切性に特化したインストラクションデータAnswerCarefully v2.0の公開
  - 11言語への翻訳完了
- 日本語の偽・誤情報に特化したインストラクションデータJSocialFactの公開
- LLMを攻撃する敵対性プロンプトをゲーミフィケーションにより収集
- AISI(Japan AI Safety Institute)との連携
  - シンガポール開催のRed-teaming Challengeに参加
  - AISI network convening, AI Action Summitに参加

各WGでの透明性・信頼性の確保に資する取り組みについて

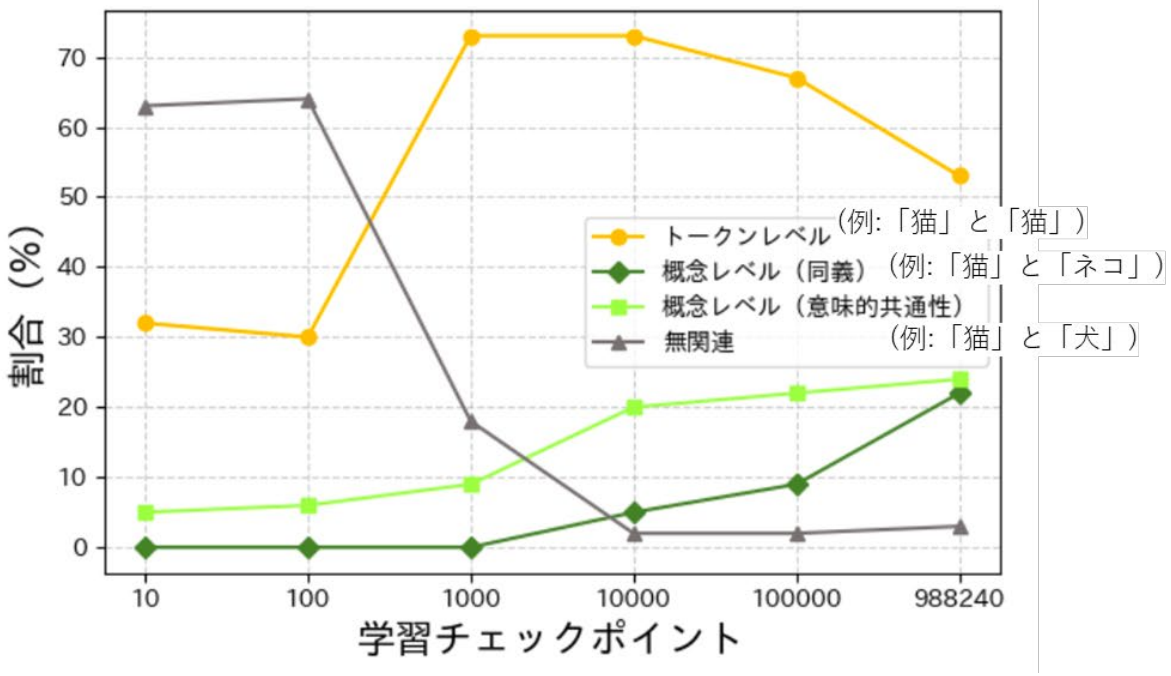
原理解明WG

- 内部から見る大規模言語モデルの言語汎化能力 (稲葉ら. NLP2025)

言語を個別に学習した後、言語間の対応関係を習得



トークンレベルの知識を学習後、概念レベルの知識を習得



- 大規模言語モデルにおけるペルソナの役割と内部動作の理解
- 大規模言語モデルにおける Supervised Fine-tuning の包括的検証
- 大規模言語モデルの地理情報に関する内部空間のモデル・言語間による比較分析

R6年度成果とR7年度の研究計画

令和6年度の主な成果

- 生成AI（LLM）の透明性・信頼性等を確保するための研究開発のテストベッドとなる175B級サイズのLLMの構築
- コーパス検索基盤及びハルシネーション分析ツール（分類器）の開発
- ファインチューニング・強化学習・評価のためのデータの準備
- 有害情報をフィルタリングした学習データや安全性対策のためのインストラクションデータ
- MoEモデル等LLM構築に関する学習ノウハウ等の知見の獲得
- 大規模並列計算環境の構築・計算ライブラリのツール化

生成AIに関する世界的な技術動向

- 画像・音声等のマルチモーダルなデータの入力・生成に対応したモデルが多数発表されるも、主要モデルは透明性が欠如

令和7年度に実施すべき主な研究開発項目

令和6年度に構築したLLMを用い、下記を実施

- コーパス検索基盤・分類器を用いて、ハルシネーションを含む出力の傾向の分析
- ファインチューニング・強化学習手法等のAIアライメントへの影響の分析
- 効率的なチューニングを実現するファインチューニング・強化学習の組み合わせの検証
- コーパスフィルタリング等の安全性効果の検証

上記結果を踏まえ、下記を実施

- ハルシネーション発生メカニズムの解析
- 効果的・効率的なLLM評価フレームの検討
- 悪意のあるプロンプトに対するレッドチーミング手法の実現と効果の検証等の生成AIの安全対策に関する研究開発
- データバイアス問題やドメイン適応に関する横断的課題等に関する研究開発の実施

技術動向を踏まえ、下記を新たに実施

- 670B～1T級/MoE方式のLLMテストベッドの構築
- マルチモーダルモデルの透明性・信頼性確保に関する研究開発の実施

R7年度の研究開発線表

研究開発課題	2025年 4月	5月	6月	7月	8月	9月	10月	11月	12月	2026年 1月	2月	3月
モデル構築（事前学習） ※構築したモデルは、透明性・信頼性・高度化に関する研究に用いる。なお、最初の175B級モデル構築完了前は、R5に構築した13Bモデルを研究に用いる。	新規コーパス開拓・整備、フィルタリングツール開発、GPU並列計算環境整備、モデル構築に関する研究開発											
	670B～1T級/MoE方式のLLMテストベッドの構築											
透明性・信頼性に関する研究	コーパス検索基盤及びLLM入出力観察・分析基盤の構築、チューニング・評価やファインチューニング(FT)・Learning from Human Feedback(LHF)の効果分析・評価に関する研究											
透明性に関する研究	コーパス検索基盤・分類器を用いて、ハルシネーションを含む出力の傾向の分析						ハルシネーション発生メカニズムの解析					
	FT・強化学習手法等のAIアライメントへの影響の分析						効果的・効率的なLLM評価フレームの検討					
	効率的なチューニングを実現するファインチューニング・強化学習の組み合わせの検証											
	広島プロセス国際指針等を踏まえた安全性対策											
信頼性に関する研究	コーパスフィルタリング等の安全性効果の検証											
	悪意のあるプロンプトに対するレッドチーミング手法の実現と効果検証等の生成AIの安全対策に関する研究開発											
社会受容性に関する研究	LLMモニタリング基盤の構築・運用											
	データバイアス問題やドメイン適応に関する横断的課題等に関する研究開発の実施											
	法制度や倫理基準を踏まえたLLM評価手法の開発											
高度化に関する研究	ドメイン適応・モデル軽量化・Transformerアーキテクチャーの発展に関する研究開発											
	マルチモーダルモデルの構築及びマルチモーダルモデルの透明性・信頼性確保に関する研究開発											
	高度化に向けた方法論の確立に向けた検討： 中間学習（指示事前学習，合成データ活用，扱える入力長の拡張）、推論モデルの学習（合成データ構築と強化学習）											

5年間の研究計画

緑色が以前の計画からの更新内容

研究目標

日本全体の産学官の力を結集して基盤モデルの研究拠点を構築し、①研究力・開発力醸成のための環境整備、②学習原理解明等による信頼性確保等、③高度化研究開発を実施する。整備されるモデル等を広く開放し、信頼性確保手法やモデル自体を企業等を含め水平展開するとともに、アジア、欧州等との国際連携にも注力することにより、AI の進化、将来に亘った革新的なイノベーション創出に資する。

	R6	R7	R8（予定）	R9（予定）	R10（予定）
研究開発課題１ 研究開発用 基盤モデル構築	新規コーパス開拓・整備、フィルタリングツール開発 →	→	→	→	→
	GPU 並列計算環境整備（民間クラウド環境、産総研 ABCI（予定）） →	→	→	→	→
	モデル構築に関する研究開発 （172B、8×13B MoE 等） →	（670B～1T 級/MoE 方式の LLM テストベッドの構築） →	（国内外の最新動向を踏まえて透明性・信頼性確保の研究開発に必要なモデルを構築） →	（同左） →	（同左） →
研究開発課題２ 透明性・信頼性 確保に向けた 研究開発	コーパス検索基盤及び LLM 入出力観察・分析基盤の構築 →	→	→		
	チューニング・評価、ファインチューニング(FT)・Learning from Human Feedback(LHF)の効果分析・評価に関する研究 →	→	→	→	→
	広島プロセス国際指針等を踏まえた安全性対策 （コーパス・フィルタリング手法の高度化、安全対策インストラクション・データ整備とチューニング、レッド・チーミング・テストの検討等） →	→	→		
	データ改変、データバイアス等の影響抑制等 →	→	→	→	→
	外部知識利用、ハルシネーション防止技術 →	→	→	→	→
			意味の汎化現象の理論的解明、時間概念の扱いの分析、記憶・意志・意識のモデル化 →	→	→
	LLM モニタリング基盤の構築・運用 →	→	→	→	→
	法制度や倫理基準を踏まえた LLM 評価手法の開発 →	→	→	→	→
研究開発課題３ 高度化に向けた 研究開発	ドメイン適応に関する研究、モデル軽量化 →	→	→		
	Transformer アーキテクチャの発展 →	→	→	→	→
		マルチモーダルモデルの構築及びマルチモーダルモデルの透明性・信頼性確保に関する研究開発 →	→	→	→
		高度化に向けた効果的な方法論の確立（中間学習、推論モデルの学習等） →	→	→	→

今後の展望

- **ソブリンAI開発力の醸成**はLLM-jp/NII-LLMCが当初から掲げている目標
  - NII-LLMCの計算環境は公的資金による支援としては世界最大級
  - すでに学習データを含めた世界最大級のフルオープンモデルを公開
  - モデルの原理解明を目指す研究も展開中
  - 今後一層、生成AIの透明性・信頼性を高め、より高性能な生成AIモデル構築に貢献する取組みを充実させることが必要
  - 生成AIを活用した（学術）知識基盤の構築など、社会に大きなインパクトをもたらすアプリケーションにも力を入れていくことが必要
- ソブリンAI開発に貢献するために、以下のようなデータの活用実現を目指す
  - 我が国の学術論文等
  - 国立国会図書館（NDL）に納本され電子化された書籍
  - NHKのニュース、ドキュメンタリー番組等

まとめ

- **「生成AIモデルの透明性・信頼性の確保に向けた研究開発拠点形成」事業を着実に遂行**

  - フラッグシップモデルとなる172Bモデルを構築・公開
  - Mixture-of-Experts手法を用いた8x13Bモデルを構築・公開
  - マルチモーダルモデルの研究開発に着手、14Bモデルを構築・公開
  - そのほか、学習コーパスや安全性向上のためのインストラクション・データセット(AnswerCarefully)などを開発・公開

- EMNLP2024、LREC-COLING2024などの国際会議で5件の論文発表
- AAMT(一般社団法人アジア太平洋機械翻訳協会)長尾賞を受賞
- 言語処理学会第31回年次大会(2025年3月)で「大規模言語モデルのファインチューニング技術と評価」ワークショップを主催
- 言語処理学会第31回年次大会で優秀賞、若手奨励賞、委員特別賞等を受賞

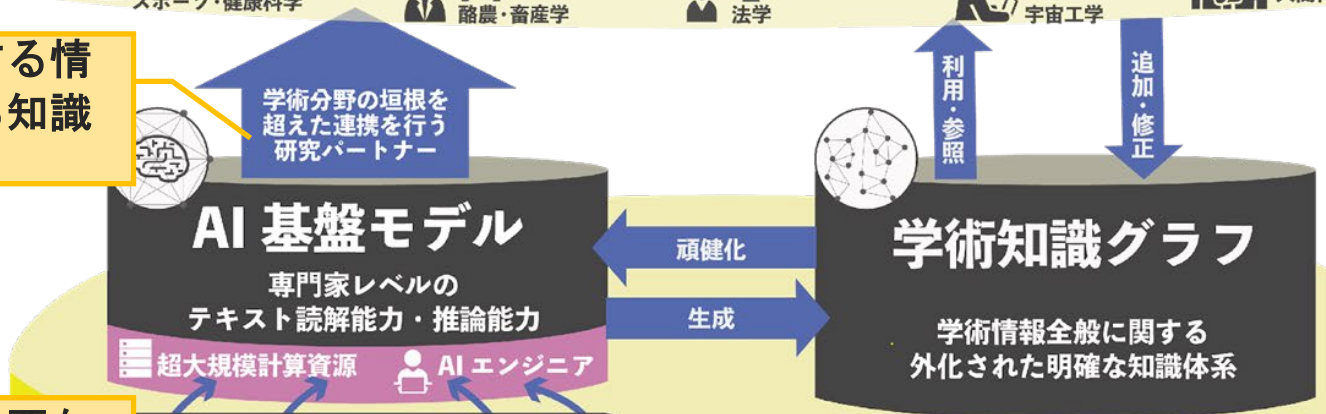
BACKUP CHARTS

データ基盤から知識基盤へ

NII

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

生成AIモデルの透明性・信頼性の確保に向けた研究開発拠点形成

令和7年度予算額（案）　8億円
（前年度予算額　7億円）
令和6年度補正予算額　42億円

文部科学省

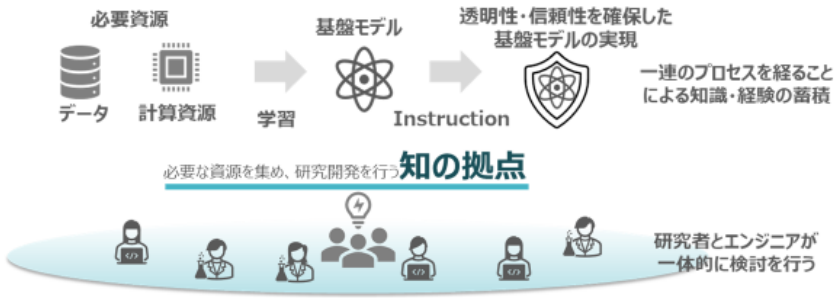
背景・課題

- **高度な推論力を有する大規模言語モデル**や**マルチモーダル等に対応した新たな生成AIモデル**が登場し、生成AIを活用したサービスの開発は世界中の民間企業・研究機関においてより一層活発になっている。
- 一方で、こうした生成AIモデルにはどのようなアルゴリズムに基づき回答しているのかなどの**「透明性」や**、AIが誤った回答をしていないかなどの**「信頼性」の確保に対して課題がある。**
- また、生成AIモデルに関する基盤的な研究力・開発力を醸成するため、**アカデミアを中心とした一定規模のオープンな生成AIモデルを構築できる環境を整備**し、**一連の知識と経験を蓄積、広く共有することが重要。**

【新しい資本主義のグランドデザイン及び実行計画2024年改訂版（令和6年6月21日閣議決定）】
Ⅴ．投資の推進　３．AI（１）AIのイノベーションとAIによるイノベーションの加速
①研究開発力の強化
モデルの高効率化や高精度化、マルチモーダル化（テキスト、画像、音声、動画等の様々な情報を同時に処理・解析する機能）、リスクの低減化等の研究開発、質の高い日本語データ及び産業競争力を有する分野のデータの整備・拡充を産学連携で進めるとともに、革新的な技術を有するスタートアップを支援する。

目的

上記課題の解決のため、産学官の研究力を結集したアカデミア研究拠点を構築し、
1．生成AIモデルに関する研究力・開発力醸成のための環境整備
2．生成AIモデルの学習・生成機構の解明等による透明性の確保等
3．生成AIモデルの高度化に資する研究開発
を行い、AIの進化、ひいては将来に渡った革新的なイノベーションの創出に貢献する。



内容

国立情報学研究所（NII）において、生成AIモデルの透明性・信頼性の確保に資する研究開発とともに、研究用モデル構築およびモデルの高度化に取り組む。研究成果のモデルへの適用・試行錯誤を通じて、**透明性・信頼性を確保した次世代生成AIモデル構築手法の確立を目指すとともに、一連の知識と経験を蓄積**する。

国 → 補助金 → 情報・システム研究機構 国立情報学研究所（NII）

事業実施期間　令和５年度～令和10年度

１．研究開発用LLM構築
コーパス開拓・整備、GPU並列計算環境整備を行うとともに、研究開発用LLMを構築。

２．透明性・信頼性等に関する研究開発
モデルの挙動解明やハルシネーション防止技術に関する研究開発を行うとともに、社会が安心してLLMを利用するための評価手法を検討。

３．高度化に関する研究開発
LLMの各専門領域への適応やモデルの軽量化について、各専門領域の研究者と協力しつつ実施。

＋

マルチモーダルに関する研究開発
昨今の世界的な技術動向を踏まえ、画像・音声など多様なモダリティのデータを扱うことのできるマルチモーダルモデルを構築するとともに、マルチモーダルモデルの透明性・信頼性等に関する研究開発を行う。

（担当：研究振興局参事官（情報担当）付）