

# 「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」について

黒橋 禎夫

国立情報学研究所 所長／京都大学 特定教授

# LLM研究開発の重要性と考え方

- LLMを基盤とするAIは、今後の社会を大きく変革する
- LLM研究開発はbig scienceであり、大規模計算資源と多数の高度研究人材を要する。国としての研究開発を産学力を結集して行うことが必要
  - 民間企業単独では、10B級モデルの構築はできても、100B級モデルの構築や、安全性等への十分な対策は困難
  - 今後はマルチモーダル・モデルとなり（その先に日本の強みであるアニメやロボティクス）、この問題はますます深刻
  - LLM-jp(参加者1,200名超)とLLM研究開発センター(2024.4設置)においてアカデミアと産業界の英知を結集

➡ 企業が 独自性を持つ/差別化をはかる 出発点を高める  
我が国が 世界と伍するAI研究開発力を持つ

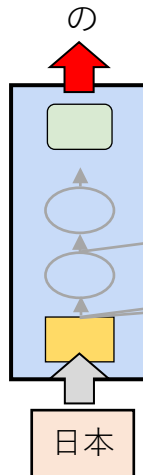
If you want to go fast, go alone. If you want to go far, go together.



# 大規模言語モデルの仕組み

## 学習時

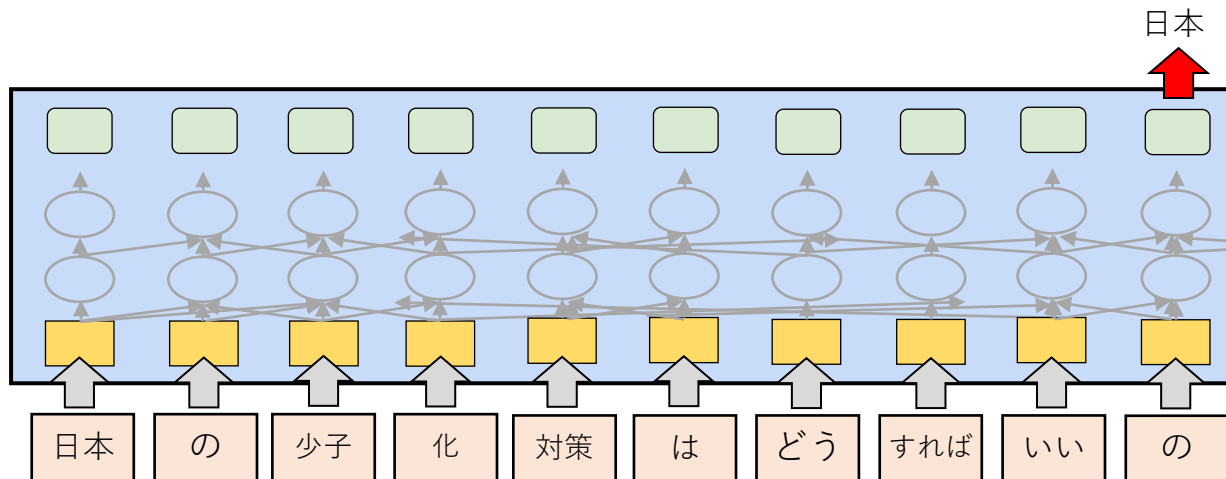
学習データ：日本の少子化対策には、次のようなアプローチ...



# 大規模言語モデルの仕組み

## 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



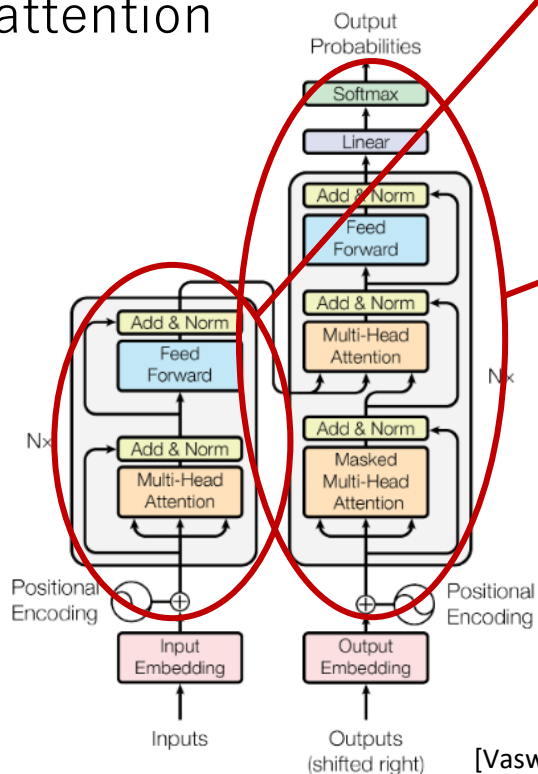
# LLMの歴史

## 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 (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

# LLMのfine-tuning

## 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. . .

LLM-jp-13B v1.1の場合

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. . .

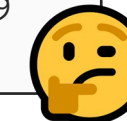
### ### Response:

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？



意味・文脈の理解と多言語性



# 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 (Retrieval Augmented Generation)



**You**  
大谷翔平選手の2023年シーズンの成績をおしえてください。

**ChatGPT**  
I did a quick search for more information and here's what I discovered.

大谷翔平選手の2023年シーズンの成績は以下の通りです。

## 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

## 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。

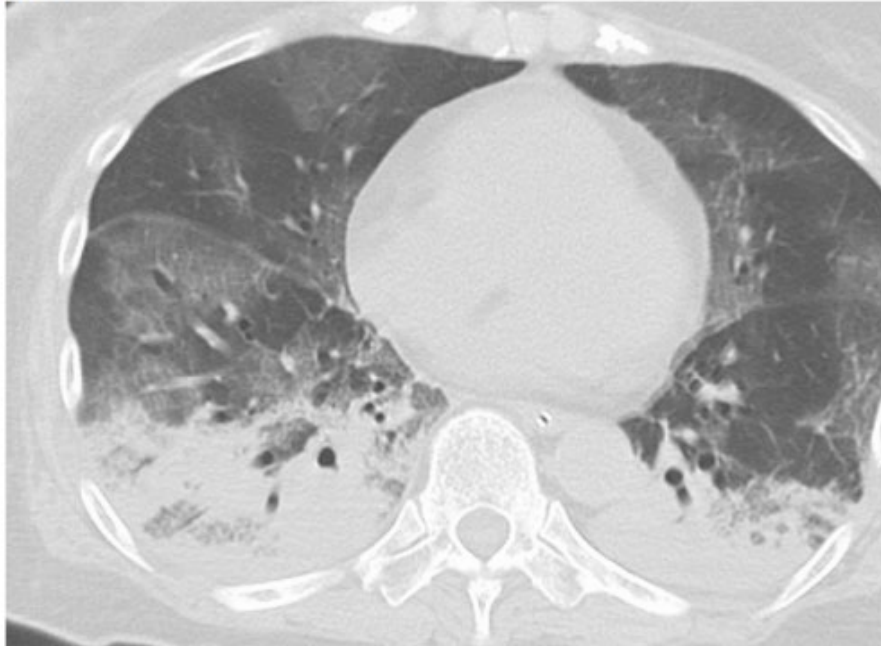


Message ChatGPT...

# GPT-4V (画像理解)

ChatGPT 4 ▾

KU You



You are a helpful radiology assistant. Describe the clinical findings in the image and suggest a diagnosis. Is it 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.



# 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**

# Model Publication Schemes

## Closed models

- Only available on UIs and/or APIs
- GPT-4 (2T?), Gemini (?), Claude 3 (?), ...

## Open models

- Only model parameters are available
- Llama2 (70B), Falcon (180B), ...

## Fully open models

- All components and results are available: training code, model parameters, all checkpoints, training corpora, ...
- Only for English small models
  - Pythia (12B) [Biderman+ 2023]
  - LLM360 (7B) [Liu+ 2023]

# 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

# 日本語の扱い

GPT-3 Codex

```
Major social networking sites and other platform providers have grown  
without ever assuming responsibility for the content of their users'  
posts.
```

Clear

Show example

**Tokens**

**23**

**Characters**

**145**

```
Major social networking sites and other platform providers have grown  
without ever assuming responsibility for the content of their users'  
posts.
```

# 日本語の扱い

GPT-3 Codex

大手SNSなどのプラットフォームはこれまで、利用者の投稿の内容への責任を負うことなく、成長してきた。|

Clear

Show example

Tokens

66

Characters

51

大SNSなプラットフォームはこれまで、利用者の投稿の内容への責任を負うことなく、成長してきた。



# GPT-3における英語と日本語の扱い

GPT-3 Codex

Major social networking sites and other platform providers have grown without ever assuming responsibility for the content of their users' posts.

Clear Show example

Tokens 23 Characters 145

Major social networking sites and other platform providers have grown without ever assuming responsibility for the content of their users' posts.

GPT-3 Codex

大手SNSなどのプラットフォームはこれまで、利用者の投稿の内容への責任を負うことなく、成長してきた。

Clear Show example

Tokens 66 Characters 51

大SNSなプラットフォームはこれまで、利用者の投稿の内容への責任を負うことなく、成長してきた。

# 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

# LLM-jp (LLM勉強会)



- 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」を構築・公開

**2023.11**

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

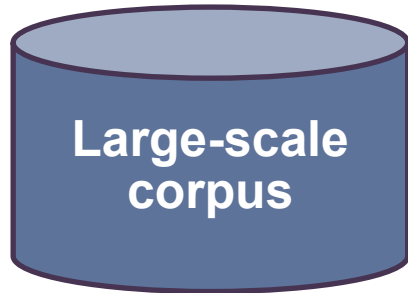
**2024.1**

GENIAC 第1期採択（2024.4から1750億パラメータ級モデルを学習）

- **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

# Research and Development of LLMs

## Corpora Construction WG

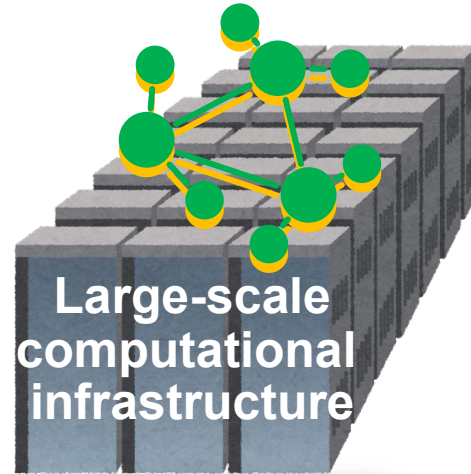


Large-scale  
corpus



Prof. Daisuke Kawahara  
Waseda Univ.

## Model Construction WG



Large-scale  
computational  
infrastructure



Prof. Rio Yokota  
Tokyo Institute of Tech.

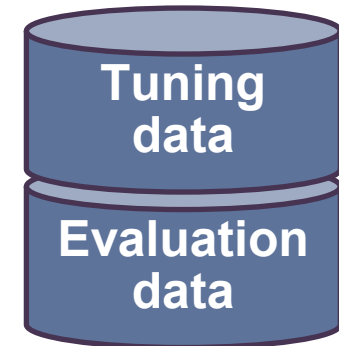


Prof. Jun Suzuki  
Tohoku Univ.



Prof. Kenjiro Taura  
The Univ. of Tokyo

## Tuning & Evaluation WG



Tuning  
data

Evaluation  
data



Prof. Yusuke Miyao  
The Univ. of Tokyo

# LLM-jp モデル

## v1.0 (2023/10/20 公開) ※ githubのダウンロード件数 2万超

- アーキテクチャ：GPT / トークナイザ：語彙数50k
- 事前学習コーパス：日本語, 英語, コードを混ぜた約 300B トークン
- チューニング用データセット：jaster, dolly (日), open assistant (日). 185k サンプル

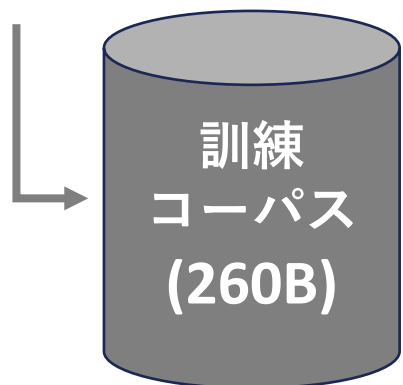
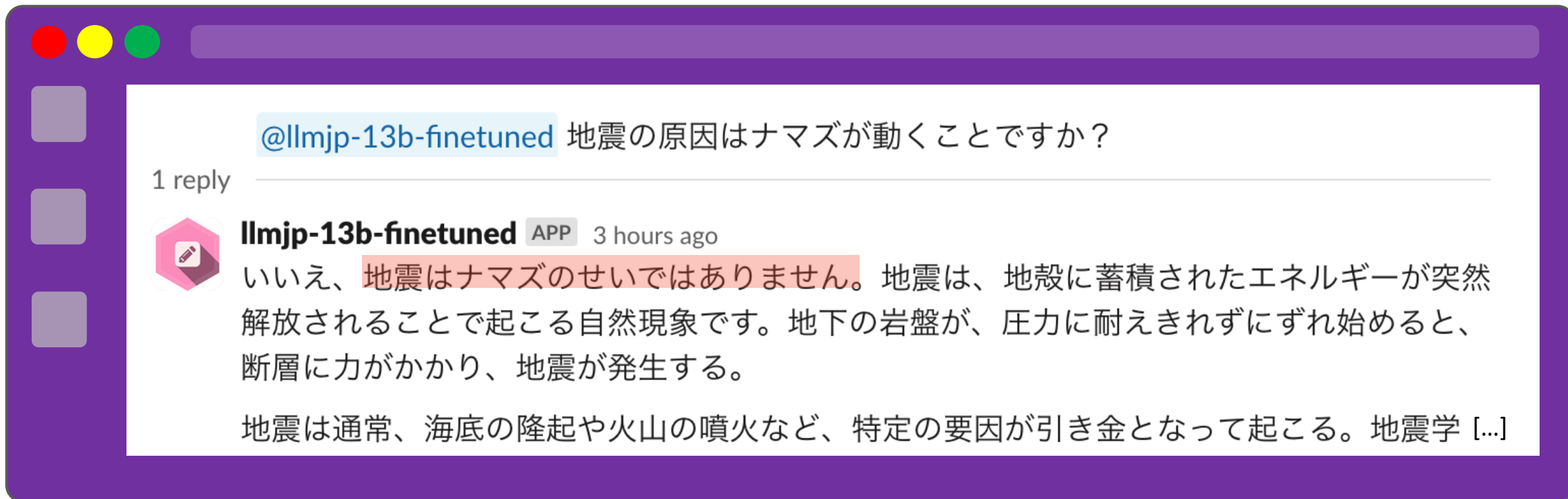
## v1.1 (2024/2/9 公開)

- v1.0 のチューニング手法を改良
  - Ichikara instruction データ, 3k サンプルを追加 (jasterは不使用)
  - Learning from Human Feedback 導入, 12k サンプル

## v2.0 (2024/4/30 公開)

- アーキテクチャ：Llama / トークナイザ：語彙数100k
- 事前学習コーパス：日本語WebデータをmC4 に代えてCommon Crawl をクリーニングして使用
- チューニング用データセット：dolly (日英), open assistant (日英), Ichikara(日), 75k サンプル. さらに安全性についてAnswerCarefully データを構築

# 根拠を調べるためのコーパス検索



Dataset: Wikipedia #1  
URL: <https://ja.wikipedia.org/wiki/地震>  
Training step: 28,955 (/61,000)  
地震の原因論とメカニズム論の展開 神話など日本(大和民族)では古来より「地中深くに大ナマズが存在し、その大ナマズが暴れることにより大地震が起きる」という俗説が信じられていた。現代においてもよく知られた俗説だが、ナマズが地震を予知できる根拠は見つかっていない [...]

Dataset: Common Crawl #2  
URL: <https://gbank.gsj.jp/wellweb/GSJ/kaisetsu/science/science.html>  
Training step: 168 (/61,000)  
日本は世界有数の地震国です。そこで誰でも考えるのは、「地震発生が事前にわかればあれほどの被害は出ないのに」ということ。つまり地震予知です。この地震予知に関しては昔からいろいろな事が言われてきました。例えば「地震の前にはナマズが暴れる」、「地震の前には [...]

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（事業期間:令和5年度～令和10年度）

# 大規模言語モデル研究開発センター (LLMC)

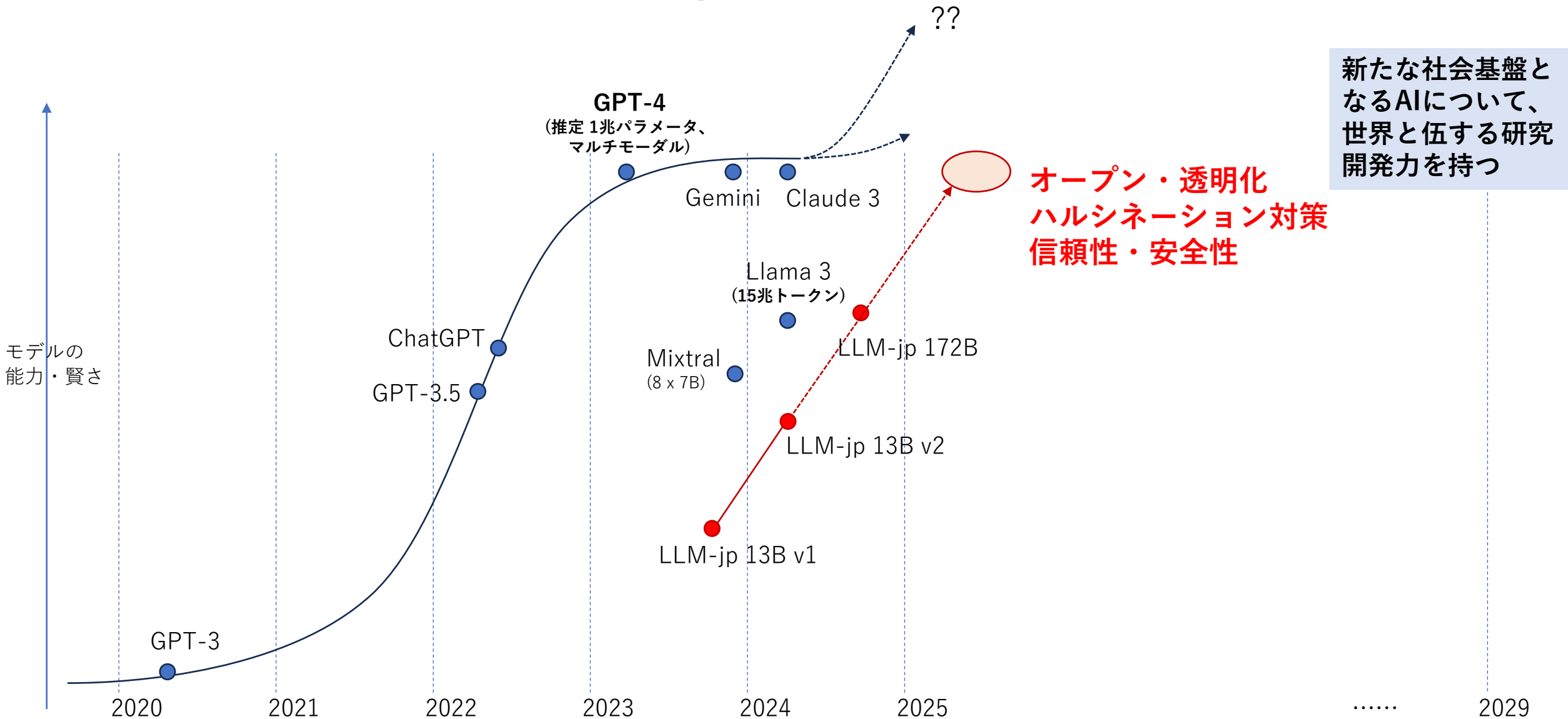
## LLM-jpの活動

- LLMC主宰の活動と位置づけ
- LLMC外からも参加（1200名超）
- 完全にオープン

## LLMCの活動

- 30名規模の研究者
- データ開拓・開発、企業との共同研究等、一部クローズな活動がありえる

# LLM研究開発センターの目標





# 国立情報学研究所 大規模言語モデル研究開発センターの体制

## 外部連携機関

東北大学  
東京大学  
東京工業大学  
早稲田大学  
京都大学

黒橋 禎夫 センター長 (NII所長)

相澤 彰子 副センター長 (NII副所長) 武田 浩一 副センター長 (NII特任教授)

## 科学 主幹

関根聡 特任教授、岡崎直観 客員教授、河原大輔 客員教授、空閑洋平 客員准教授、  
鈴木潤 客員教授、田浦健次郎 客員教授、宮尾祐介 客員教授、横田理央 客員教授

事務局

## 課題 1 : 研究開発用LLM構築

### ① 新規コーパス開拓・整備、フィルタリング ツール開発

河原大輔客員教授  
研究者2名

### ② GPU並列計算環境整備

合田憲人教授、吉田浩特任教授、  
田浦健次郎客員教授、  
空閑洋平客員准教授、  
横田理央客員教授  
研究者2名

### ③ モデル構築

鈴木潤客員教授、  
横田理央客員教授、  
田浦健次郎客員教授、  
空閑洋平客員准教授、研究者2名

## 課題 2 : 透明性・信頼性に関する研究開発

### 【透明性に関する研究開発】

- ④ コーパス検索基盤及びLLM入出力観察・分析基盤の構築  
岡崎直観客員教授、鈴木潤客員教授、研究者2名
- ⑤ チューニング・評価、FT・LHFの効果分析  
宮尾祐介客員教授、関根聡特任教授、河原大輔客員教授、研究者2名

### 【信頼性に関する研究開発】

- ⑥ 広島プロセス国際指針等を踏まえた安全性対策  
関根聡特任教授、武田浩一副センター長、  
鈴木久美特任教授、研究者2名
- ⑦ データ改変、データバイアス等の影響抑制等  
相澤彰子副センター長、研究者2名
- ⑧ 外部知識利用、ハルシネーション防止技術  
河原大輔客員教授、岡崎直観客員教授、研究者2名

### 【社会受容性に関する研究】

- ⑨ LLMモニタリング基盤の構築・運用  
岡崎直観客員教授、研究者2名
- ⑩ 法制度や倫理基準を踏まえたLLM評価手法の開発  
相澤彰子副センター長、佐藤健ROIS人工知能法学研究  
支援センター長、研究者3名

## 課題 3 : 高度化に関する研究開発

### ⑩ ドメイン適応、モデル軽量化

相澤彰子副センター長 鈴木潤客員教授、  
河原大輔客員教授、佐藤健ROIS人工知能  
法学研究支援センター長、研究者4名

### ⑪ Transformerアーキテクチャーの発展

宮尾祐介客員教授、横田理央客員教授、  
鈴木潤客員教授、研究者1名

- 研究者（客員を含む）：15名  
※1人の研究者が複数課題に参加するため、  
図中の研究者数の合計は実際の研究者数と  
一致しない
- 技術補佐員等（データ作成等を担うアノ  
テータ等）：22名
- リサーチ・アシスタント：20名(6/1採用予定5名含)

# 2024年度の計算資源

## mdx

- 4月～2025年3月、2023年度と同様の資源を通年利用（専有利用 A100 16ノード（128枚）及びスポット利用）

## 経産省GENIACプログラムの支援によるGCPの計算資源

- 4月15日～8月15日の4ヶ月間
- 172Bモデルを構築中（コーパス：当初計画では1.35Tトークン。計算が順調に進捗すれば2Tトークン程度までの学習を検討中）

## 東工大TSUBAME 4.0

- 5月～9月の5ヶ月間、50ノード（H100 200枚）を利用

## 民間クラウド計算資源

- 6月開札。8月～2025年3月、65ノード（H100 520枚）以上の計算資源確保を想定

# 5年間の研究計画

## 研究目標

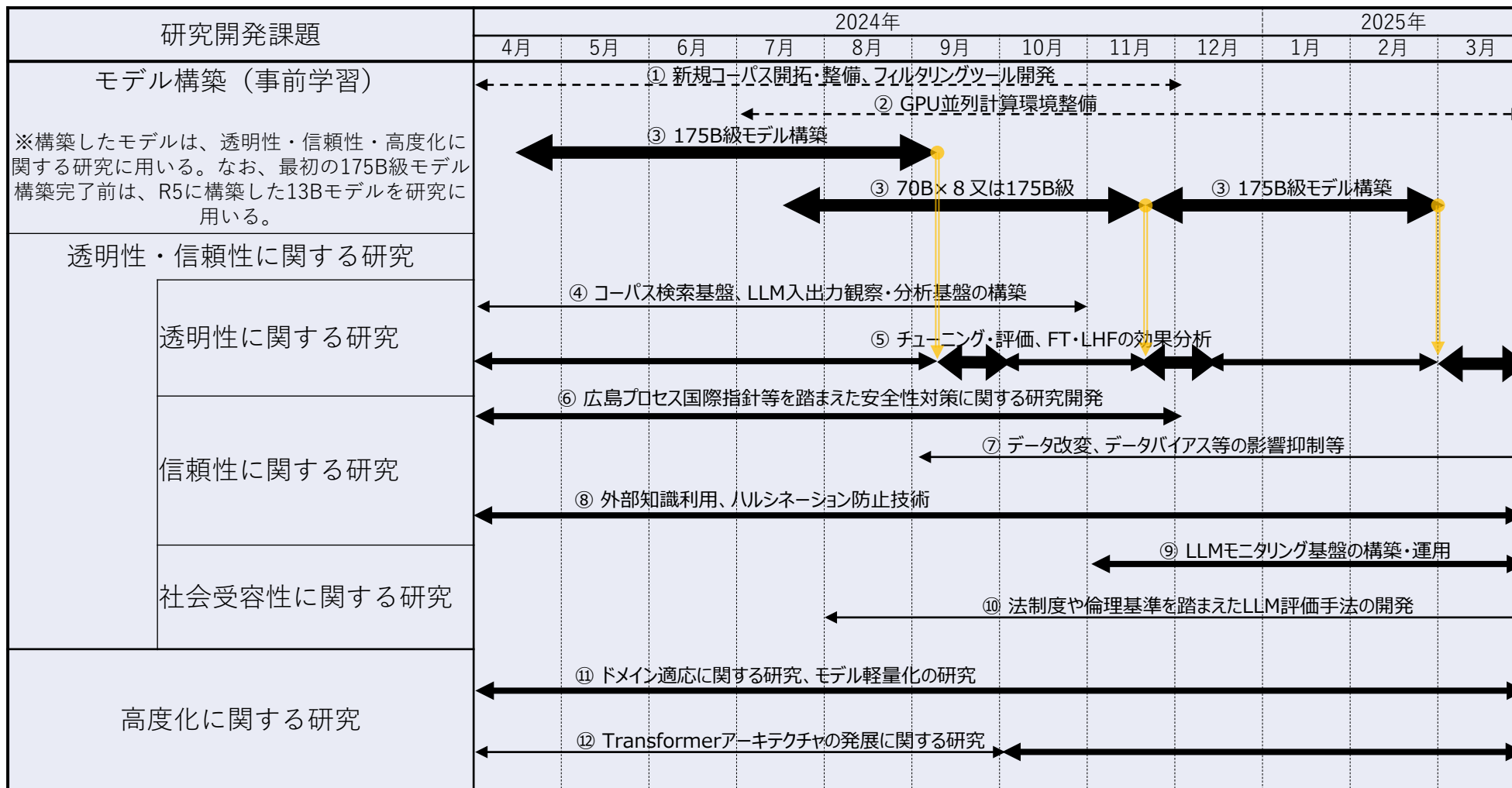
日本全体の産学官の力を結集して基盤モデルの研究拠点を構築し、①研究力・開発力醸成のための環境整備、②学習原理解明等による信頼性確保等、③高度化研究開発を実施する。整備されるモデル等を広く開放し、信頼性確保手法やモデル自体を企業等を含め水平展開するとともに、アジア、欧州等との国際連携にも注力することにより、AIの進化、将来に亘った革新的なイノベーション創出に資する。

	R6	R7 (予定)	R8 (予定)	R9 (予定)	R10 (予定)
研究開発課題 1  研究開発用 LLM 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				
研究開発課題 2  LLM の透明性・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				
	チューニング・評価、ファインチューニング(FT)・Learning from Human Feedback(LHF)の効果の分析				
	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・チーム・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				
	LLM モニタリング基盤の構築・運		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		
	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  LLM の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 (LMM) に関する研究				

# R6年度 研究開発線表

【利用予定の計算資源】

mdx（通年）、GCP（4月～8月）、TSUBAME4.0（5月～9月）、民間クラウド計算資源（7月～3月）



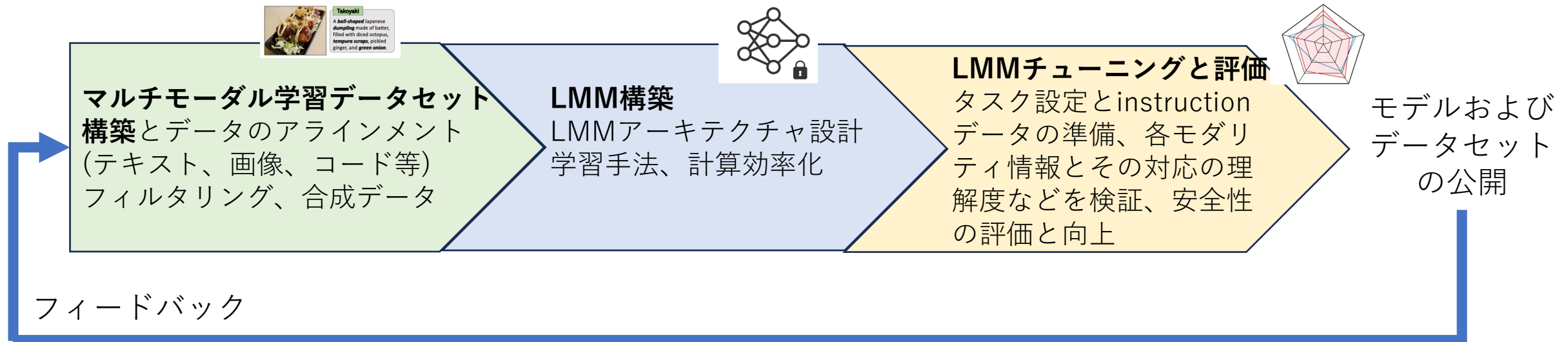
注) 線の太さは計算資源使用量の多さに概ね対応している。

# 新たに取り組むべき課題: マルチモーダルモデル(LMM)の研究

## [研究の必要性]

- 巨大IT企業を中心に、生成AI技術のマルチモーダル化が顕著(高い市場価値と社会的需要)
- LMMにおけるモダリティ間の情報のアラインメント検証、学習手法、誤情報や有害情報の抑制・検出、計算効率のよいモデルの構築と計算手法など未解決な多くの課題が存在
- 現在のLMMには透明性の欠如が指摘されており、最近発表された透明度指標(100点満点)で、GPT-4が48、オープンソース化されているLlama2でも54という低い値に留まっており、LMMの理解と上記の課題解決の障害となっている

## [LLM研究開発センターで取り組むべきLMMの研究課題]

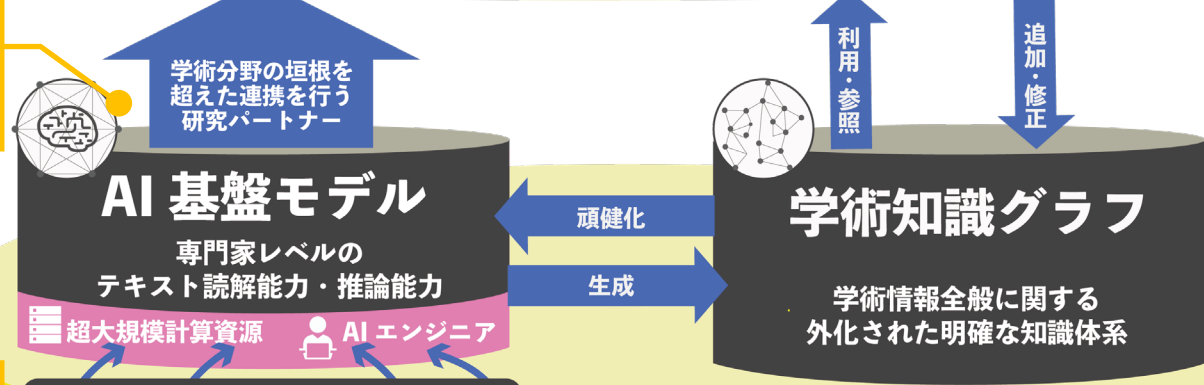


# データ基盤から知識基盤へ

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



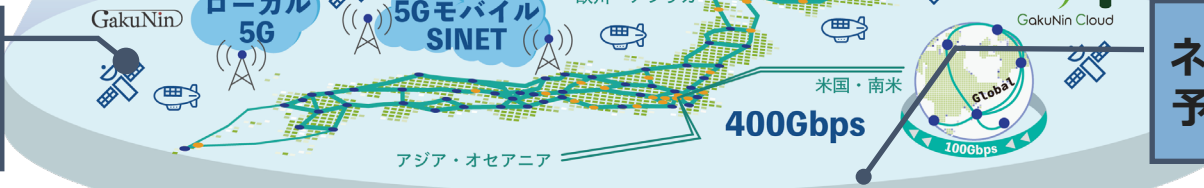
研究データ基盤の機能充実

異分野間の高度認証連携



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

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



ネットワーク状態の高度診断・障害予兆検知