



日本語に強い大規模言語モデルの研究開発を行う 組織横断プロジェクト

黒橋 禎夫

情報・システム研究機構 国立情報学研究所 所長

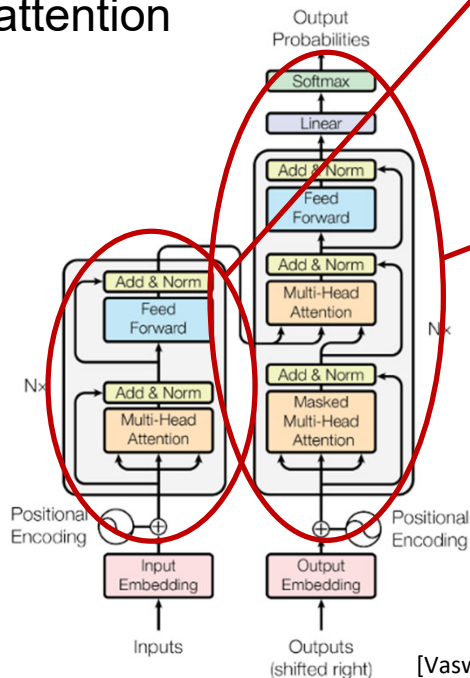
LLMの歴史

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

2023年3月

言語処理学会理事会主催 緊急パネル

緊急パネル:ChatGPTで自然言語処理は終わるのか？

- ファシリテーター
 - 乾 健太郎 氏 (東北大)
- パネリスト
 - 黒橋 禎夫 氏 (京大)
 - 相良 美織 氏 (バオバブ)
 - 佐藤 敏紀 氏 (LINE)
 - 鈴木 潤 氏 (東北大)
 - 谷中 瞳 氏 (東大)



- 3月14(火) 13:10-13:50, H会場 (劇場ホール)
 - Slack: **#0314-1310-緊急パネル-h会場 (質問はこちらへ)**
 - Twitter: **#ChatGPTで自然言語処理は終わるのか**

新しい自然言語処理

もっと人間や社会へ向かうもの

- 脳神経科学との関係
- 人の感情
- 人の理解、納得、説得
- **AI Native**世代の教育、批判的思考の涵養
- 情報の信頼性、関係付け（参照）

そのためには

- LLMが**完全にオープン**であること
- また、ChatGPTにおける日本語コーパスは圧倒的に少ない
- 日本文化や経済安全保障の視点
- 良質で十分な日本語テキスト（+ 英語テキスト）で学習したオープンな大規模言語モデルを日本が持つ必要性

LLM-jpの立ち上げ

- 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

2300名超

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

- **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

- LL...
- Home

DMS

Activity

Later

...

More
- # survey

tech-blog

tsubame4-0

コーパス構築

コーパス構築_...

コーパス構築_...

チューニング評...

チューニング評...

チューニング評...

チューニング評...

マルチモーダル

モデル構築

モデル構築_モ...

事務局

勉強会

学術ドメイン検...

安全性検討

小規模計算資源...

法律系相談

自己紹介

自然言語処理学...

運営委員会

general

- 3 Pinned
- 内部資料リンク集
- LLM勉強会
- GitHub
- Calendar
-

1,467

Canvas

1467

Mari KONDO (NII/LLMC)

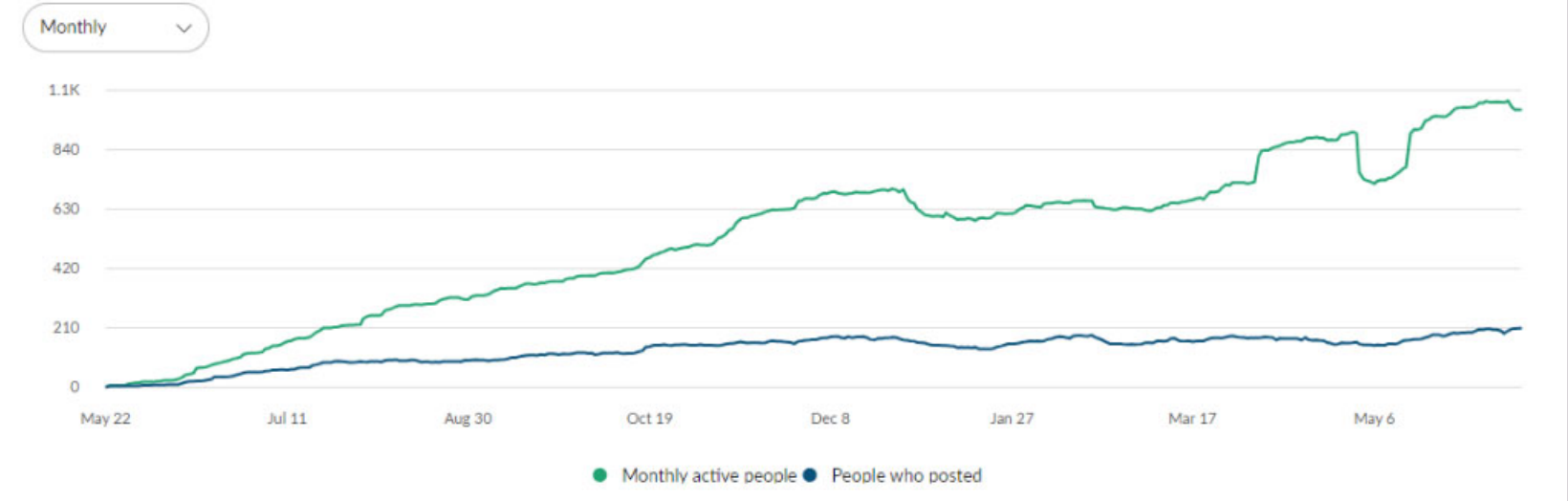
9:44 AM

Thursday, June 6th

現在 大規模言語モデル研究開発センター（LLMC）では、~~特任研究員~~を募集しております。
詳細は下記よりご確認ください。
<https://www.nii.ac.jp/about/recruit/2024/0603.html>
ご応募をお待ちしております！ どうぞよろしくお願いいたします。

Active people in your workspace

See how many people are active — meaning they posted a message or read at least one channel or direct message. To see billing numbers, go to the [Billing Overview](#) page.



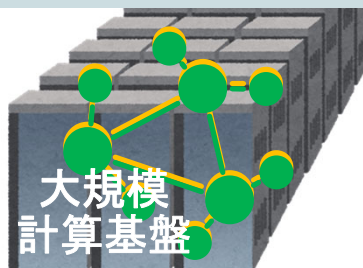
LLM研究開発はビッグサイエンス

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)

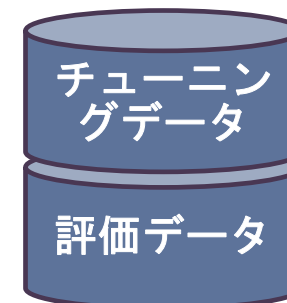


鈴木潤教授
(東北大)



田浦健次郎教授
(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



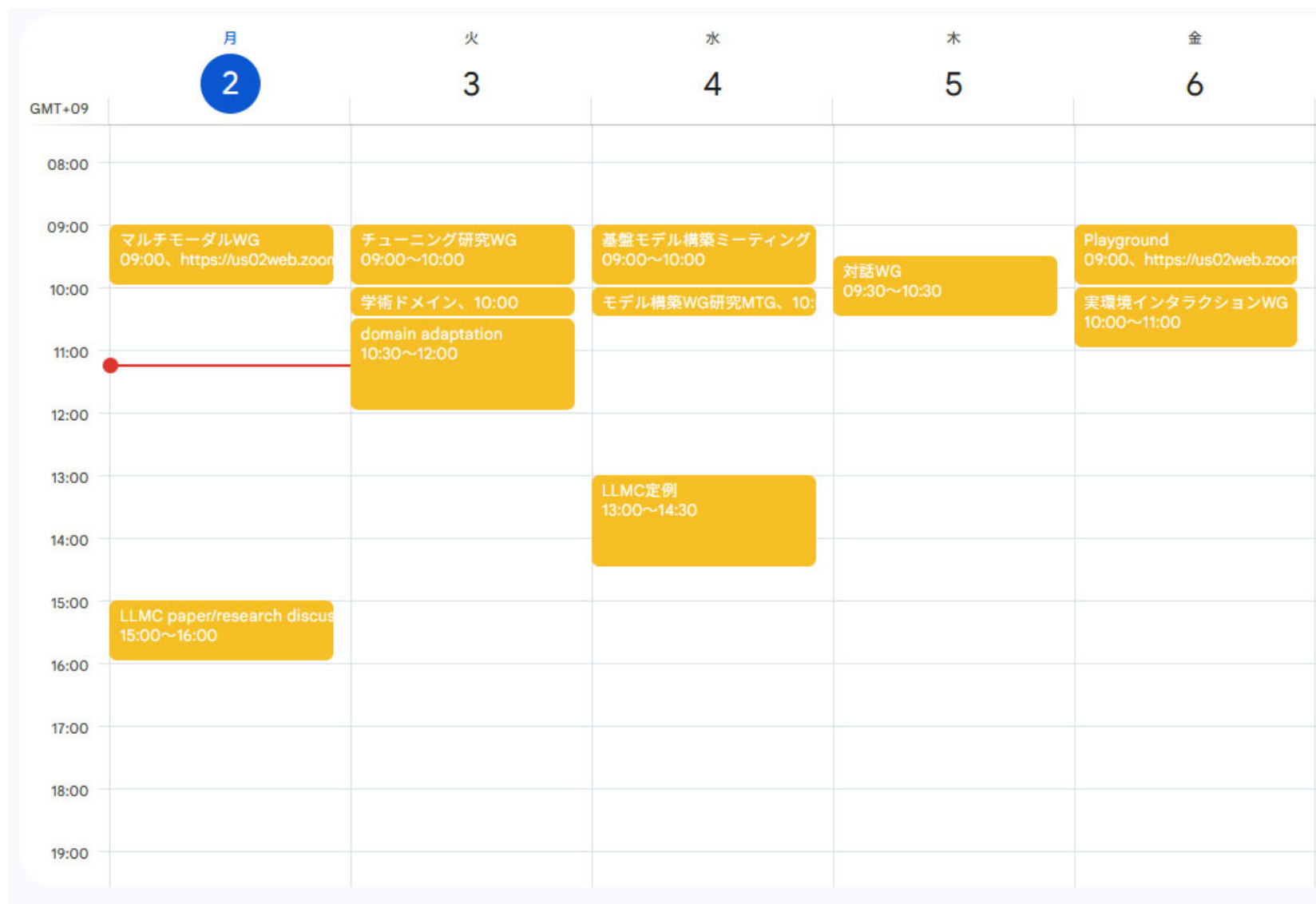
大関洋平准教授
(東大)

対話WG



東中竜一郎教授
(名大)

	月 26	火 27	水 28	木 29	金 30
GMT+09					
08:00					
09:00	マルチモーダルWG 09:00、 https://us02web.zoom		基盤モデル構築ミーティング 09:00～10:00	LLM-jp運営委員会 08:30～09:15	
10:00		学術ドメイン、10:00			実環境インタラクションWG 10:00～11:00
11:00		domain adaptation 10:30～12:00			
12:00					
13:00	原理解明WG 13:00～14:00		LLMC定例 13:00～14:30		
14:00		WG横断打合せ、13:50、 https			
15:00	LLMC paper/research discuss 15:00～16:00	第19回LLM勉強会 @1208,1210 14:30～18:00			
16:00					
17:00					
18:00					
19:00					



LLM-jp 基盤モデル構築事業

LLM-jp 基盤モデル構築事業

instruction pretraining

fanatical-flyingfish-6c6.notion.site/llm-jp-model-construction

LLM-jp 基盤モデル構築事業

Notionを無料で始める



LLM-jp 基盤モデル構築事業

このページは、LLM-jpの基盤モデル構築事業に関するタスクを管理するためのものです。

- 議論に参加するには：
毎週水曜9:00-10:00の定例ミーティングに参加してください。またオンラインでの議論を LLM-jp Slack の **#基盤モデル構築** で行っているのでそちらに加わってください。
- 本ページを編集するには：
LLM-jp Slackで @Yusuke Oda (NII) 宛にメールアドレスを通知して下さい。

テーブルビュー

定例ミーティング

Aa 名前	🔗 録画	☰ 録画パスコード	
📄 YYYY-MM-DD			
📄 2025-05-28 5回目	us02web.zoom.us/rec... xclCbU	!388evzS	

LLM-jp 基盤モデル構築事業					
LLM-jp 基盤モデル構築事業					
instruction pretraining					
fanatical-flyingfish-6c6.notion.site/llm-jp-model-construction					
LLM-jp 基盤モデル構築事業					
タスク一覧					
b1 進行中					
進捗	カテゴリ	名前	代表担当者	報告あり	
b1 進行中	事前学習	MoE	Taishi Nakamura	☐	
b1 進行中	事前学習	LLM-jp-4 En	Yusuke Oda	☐	
b1 進行中	事前学習	トークナイザ比較	Chaoran Liu	☐	
b1 進行中	中間学習	Olmo 2の中間学習再現	Koshi Eguchi	☐	
b1 進行中	中間学習	instruction pretraining	Takashi Kodama	☐	
b1 進行中	事前学習コーパス	LLM-jp コーパス v3 vs v4 の比較	Hirokazu Kiyomaru	☐	
b1 進行中	事前学習コーパス	合成データによる事前学習コーパスの拡張	Hirokazu Kiyomaru	☐	
b1 進行中	評価	llm-jp-eval	Yusuke Miyao	☐	
b1 進行中	評価	llm-jp-judge	Kouta Nakayama	☐	
b1 進行中	評価	モデル評価基盤構築 on ABCI	Kouta Nakayama	☐	
b2 定常業務					
進捗	カテゴリ	名前	代表担当者	報告あり	
b2 定常業務	運用	計算機管理	Yusuke Oda	☐	



LLM-jp、成果を出せる組織を維持しててすごいな(ある程度有名になると変な輩が入り込んで崩壊しそうなのに

✕ 望月紅葉さんと幸せな家庭を築き... @momiji_fullmo...



5月30日(金) 19:17

完了	llm-jp-3 1.8b
完了	llm-jp-3 13b
完了	llm-jp-3 8x13b

進捗状況

📄 テーブルビュー

Aa 名前

📅 2025-05-28

Model	MT-Bench (ja)	MT-Bench (en)	Answer Carefully	llm-jp-instructions	jculture
gpt-4-0613	7.29	7.72	4.06	4.63	2.17
gpt-4o-2024-08-06	8.10	8.38	4.09	4.83	2.78
llm-jp-3-1.8b-instruct3	4.64	4.09	4.03	3.59	1.55
llm-jp-3-13b-instruct3	6.21	6.13	4.37	4.41	2.14
llm-jp-3-8x13b-instruct3	6.60	6.49	4.48	4.60	2.59
llm-jp-3.1-1.8b-instruct4	6.30	5.70	3.66	3.98	1.61
llm-jp-3.1-13b-instruct4	7.37	7.01	4.17	4.55	2.12
llm-jp-3.1-8x13b-instruct4	7.50	7.05	4.26	4.72	2.42

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



LLM-jpへの参加 いつでもwelcome!