

ELYZA

オープンなLLM（大規模言語モデル）の発展と、
それを活かした日本語LLMの開発について

2024/01/25 JAIST産学官共創フォーラム・令和5年度 第3回定期講演会

会社概要

2018年9月に東京大学 松尾研究室からスピンアウトしたAIスタートアップ

ELYZA

読み仮名：イライザ





佐々木 彬 株式会社ELYZA AIエンジニア

東北大学 乾・岡崎研究室（現 乾研究室）博士卒。
株式会社リクルートにて人材領域のデータ活用
施策の推進に従事ののち、ELYZAに参画。
Kaggle Master。

ELYZAでは自然言語処理の研究開発を担当。

大規模言語AIでホワイトカラー業務のDXに挑戦しています

大規模言語AI

言語・テキストを対象に
人間のように処理するAI

要約するAI

読むAI

書くAI

話すAI



ホワイトカラー業務

ホワイトカラーの生産性を向上させる



等

R&D事例:要約AI「ELYZA Digest」

長文要約AI

ELYZA
イライザダイジェスト
DIGEST

Powered by ELYZA



東京大学松尾研発・AIスタートアップ、株式会社ELYZA（イライザ、本社：東京都文京区、代表取締役 曾根岡 侑也）は、BERT以降の汎用言語モデルを活用した日本語AIエンジン「ELYZA Brain」の開発に成功しました。ELYZAは「ELYZA Brain」を活用し、企業との事業共創プログラム「パートナープログラムNLP30」を開始します。採択プロジェクトには検証フェーズの開発リソースを無償提供いたします。

【ELYZAとは】

ELYZAは、AI領域に強みを持つ東京大学・松尾研究室出身のメンバーで創立された会社です。私たちは未踏の問題の解決をめざしNLP（自然言語処理技術）×AI分野の技術に焦点を当て、「未踏の領域で、あたりまえを創る」をミッションに掲げ、研究開発に取り組んでいます。ELYZAの強みはその技術力です。東京大学松尾研究室出身の代表をはじめとし、経済産業省主催「未踏事業」出身エンジニア、東京大学在籍のAIエンジニア

✓ 要約結果

日本語AIエンジン「ELYZA BRAIN」の開発に成功した。NLP領域における国内最大の日本語AIエンジンで、「人間を超える」精度を実現。ビジネスモデルを活用し、様々な分野で未踏の領域を実現する活動をしている。

ニュース、メール、対話などを高精度にまとめる

日本語の要約AI「ELYZA DIGEST」

ご利用ユーザー様、5日で13万人を突破

R&D事例: 執筆AI「ELYZA Pencil」



ニュース記事のAI執筆

※これはAIが自動生成したフェイクニュースです 令和4年2月28日 火曜日

株式会社イライザ、大規模な言語AIサービス「イライザペンシル」をリリース

イライザ新聞

株式会社イライザが、大規模な言語AIサービスをリリースすると発表した。同サービスは、自然言語処理技術を応用した大規模言語処理サービス。代表取締役の曾根岡侑也氏は、意気込みを語っている。

Powered by ELYZA

使用したキーワード

株式会社イライザ 大規模言語AI イライザペンシル サービスリリース
代表取締役 曾根岡侑也 意気込みを語る

メールのAI執筆

新着メッセージ ー □ ×

社内検討結果のご報告

イライザ太郎
sample@spl.com

株式会社ELYZA
営業部田中様
平素より大変お世話になっております。
本日は、貴社AI改善プロジェクトの進捗状況につきまして社内会議の結果をご報告申し上げます。
結論から申しますと、今回のプロジェクトは見送るという結論に至りました。
導入費用や今後の運用コストなどを総合的に勘案した結果、投資に見合う効果が得られると判断できなかったためです。
器用にゴースサインを出した手順、
誠に心苦しい限りではございますが、
何卒ご理解いただければ幸甚でございます。
取り急ぎ、社内会議で検討した結果をご報告申し上げます。
※これはAIが自動生成したフェイクメールです

Powered by ELYZA

使用したキーワード

株式会社ELYZA 営業部 田中様 先日 AI改善プロジェクト
社内検討結果 取り急ぎ

職務経歴書のAI執筆

※これはAIが自動生成した文章です 令和4年2月28日 火曜日

ふりがな いらいざ たらう
氏名 イライザ 太郎
生年月日 昭和 30年 9月 4日

経歴

株式会社ELYZAに入社後、約3年間、自然言語処理の研究をしています。自然言語処理とは、人間が自然に行っているような文章の解析や分類を行う技術のことです。プロジェクトリーダーとして、自然言語処理の研究・開発に取り組んでいます。また、チームのマネジメントや、顧客との窓口として折衝の機会にも恵まれています。

Powered by ELYZA

使用したキーワード

株式会社ELYZA 東京大学 3年 AI 自然言語処理 研究
プロジェクト リーダー

キーワードから文章を約6秒で生成する

日本語の執筆AI「ELYZA Pencil」

公開11日間で11万人が利用

独自LLMの開発・公開

2023年8月にリリースした「ELYZA-japanese-Llama-2-7b」など、直近で大きく分けて3種類のLLMを公開



ELYZA, Inc.
@ELYZA_inc

【お知らせ】Llama 2をベースとした商用利用可能な70億パラメータの日本語言語モデルを公開しました。ブログにて特徴や性能について紹介しているほか、推論用コード、性能評価用データセットとその評価結果もすべて公開しています。(1/3)



note.comから

午前11:00 · 2023年8月29日 · 54.1万 件の表示

384 リポスト 74 件の引用 1,027 件のいいね 343 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Llama 2をベースとした商用利用可能な130億パラメータの日本語言語モデルを公開しました。オープンな日本語モデルの中で最高性能、GPT-3.5 (text-davinci-003)を上回る性能となり、高速に動作するデモも公開中です。(1/3)

内容は以下をクリックしてご覧ください。



note.comから

午前11:00 · 2023年12月27日 · 18.3万 件の表示

225 リポスト 47 件の引用 675 件のいいね 236 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Code Llamaに対し日本語の追加事前学習を行い、コード生成・補完に特化した日本語LLM「ELYZA-japanese-CodeLlama-7b」を開発しました。noteにて概要を紹介しています。(1/3)



note.comから

午前11:00 · 2023年11月15日 · 5.7万 件の表示

79 リポスト 11 件の引用 259 件のいいね 78 ブックマーク

本日本話すること

1. オープンなLLM/ クローズドなLLM

APIベースで利用可能な、クローズドなLLM（例: GPT-4/Gemini/Claude/…）
に対して、オープンなLLMとはどのようなものか

2. 日本語LLMの近況

ここ数年の、日本語に特化したオープンなLLMの近況について
また、日本語特有の課題についても紹介

3. ELYZAにおける 取り組み内容

弊社ELYZAがリリースした「ELYZA-japanese-Llama-2」や、
評価用データセット「ELYZA-tasks-100」について

4. 直近ホットな トピック

オープンなLLMの周辺で直近ホットなトピックをいくつか紹介

本日本話すること

1. オープンなLLM/ クローズドなLLM

APIベースで利用可能な、クローズドなLLM（例: GPT-4/Gemini/Claude/…）
に対して、オープンなLLMとはどのようなものか

2. 日本語LLMの近況

ここ数年の、日本語に特化したオープンなLLMの近況について
また、日本語特有の課題についても紹介

3. ELYZAにおける 取り組み内容

弊社ELYZAがリリースした「ELYZA-japanese-Llama-2」や、
評価用データセット「ELYZA-tasks-100」について

4. 直近ホットな トピック

オープンなLLMの周辺で直近ホットなトピックをいくつか紹介

クローズドなLLMとは

OpenAIのChatGPTやGoogleのBardなど、 当該サービス/APIでの利用が前提のLLM

ChatGPT (OpenAI)

A あなた

東京駅から北陸先端科学技術大学院大学 金沢駅前オフィスまでの行き方を教えて

G ChatGPT

私の知識の範囲では、2022年1月時点の情報しか持っていないので、最新の情報は確認してください。また、交通機関のスケジュールや運行状況は変更される可能性があるため、移動前に確認してください。

東京駅から北陸先端科学技術大学院大学（金沢駅前オフィス）への行き方はいくつかの方法があります。以下は一般的な方法の一例です。

1. 新幹線を利用する場合:

- 東京駅から、北陸新幹線（こだまやひかり）に乗ります。
- 金沢駅で降り、金沢駅から徒歩またはバスで北陸先端科学技術大学院大学金沢駅前オフィスに向かいます。

2. 飛行機を利用する場合:

- 東京駅から成田国際空港や羽田空港へアクセスし、金沢空港に飛行機で移動します。
- 金沢空港からは、バスやタクシーを利用して北陸先端科学技術大学院大学金沢駅前オフィスに向かいます。

3. 車を利用する場合:

- 東京から自家用車やレンタカーを利用して、北陸自動車道や中央自動車道を経由して金沢に向かいます。
- 金沢市内で道路や案内を確認しながら、北陸先端科学技術大学院大学金沢駅前オフィスまでアクセスします。

具体的な交通手段やルートは、出発時刻や所要時間によって異なるため、移動前に最新の交通情報を確認し、計画を立てることが重要です。また、北陸先端科学技術大学院大学金沢駅前オフィスの所在地やアクセス方法については、大学の公式ウェブサイトや連絡先を確認すると良いでしょう。

Bard (Google)

2

石川県のおすすめの食べ物を教えて。肉食、魚系それぞれ1つずつ知りたいな

他の回答案を表示

かしこまりました。石川県のおすすめの食べ物、肉食と魚系それぞれ1つずつ紹介します。

肉食

- 能登牛



能登牛は、石川県能登半島で育てられたブランド牛肉です。霜降りがきめ細かく、脂身が甘く、赤身はしっかりとした旨味があります。焼肉、ステーキ、しゃぶしゃぶなど、さまざまな料理で楽しめます。

魚系

- のどぐる



のどぐるは、石川県の日本海で獲れる高級魚です。脂ののって、身がふっくらと柔らかいのが特徴です。塩焼き、煮付け、しゃぶしゃぶなど、さまざまな料理で楽しめます。

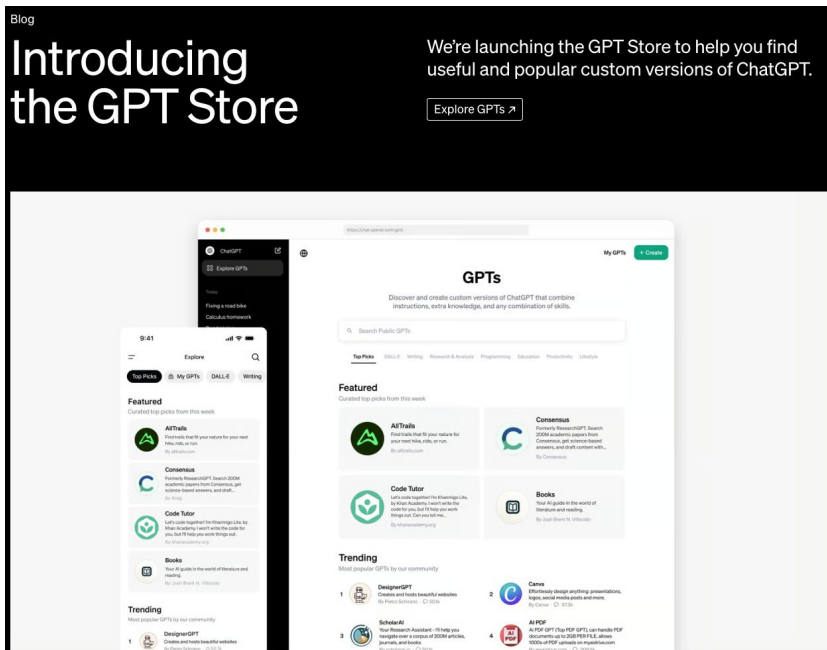
どちらも石川県を代表するおいしい食べ物です。ぜひ一度ご賞味ください。

もし、他におすすめの食べ物を知りたいという場合は、お気軽にお尋ねください。

GPTシリーズの爆発的な流行

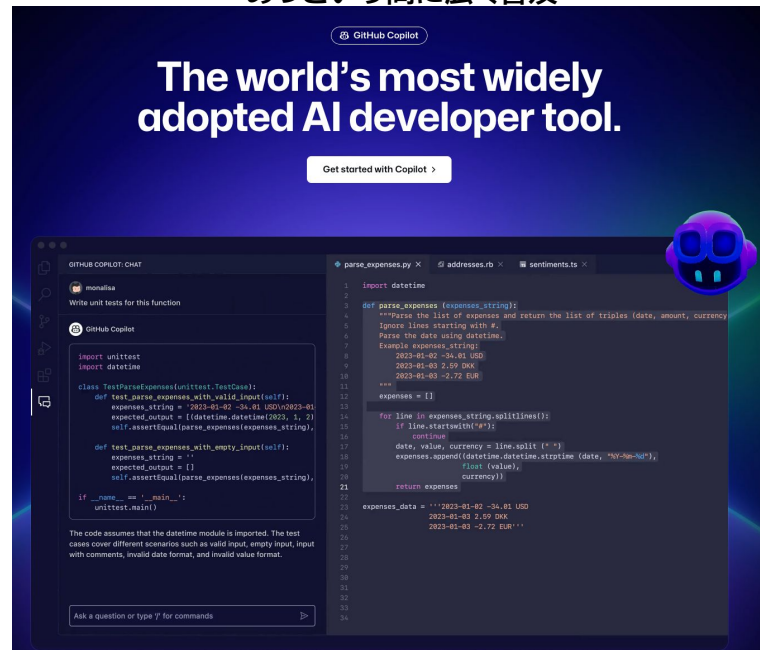
オープンなLLMの中でもGPTシリーズのインパクトは非常に大きく、特にChatGPTが公開された2022年11月以降、社会に大きな影響を与えている

特定の用途にあわせてカスタマイズしたChatGPTを公開・共有できるGPT Storeが先日開始



<https://openai.com/blog/introducing-the-gpt-store>

コーディング支援として、GitHub Copilotがあっという間に広く普及



<https://github.com/features/copilot>

クローズドなLLMの特徴: 性能の高さ

ChatGPTは2022年11月にリリースされたが、バージョンアップを重ねつつ、現時点（2024年1月）でも最高性能を維持している

GPT-4がクローズド (Proprietary) なLLM 中でも最高性能

それに対して オープンなLLM (後述) は、性能面ではまだ 追いついてない現状

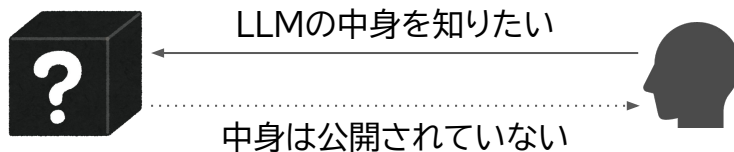
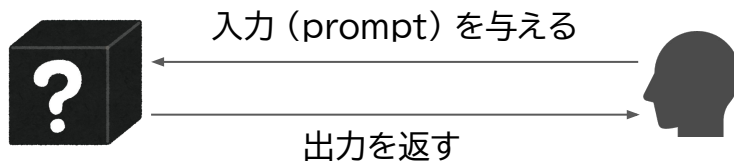
Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-Turbo	1249	+14/-13	23069	OpenAI	Proprietary
2	GPT-4-0314	1190	+14/-14	16237	OpenAI	Proprietary
3	GPT-4-0613	1160	+14/-12	20884	OpenAI	Proprietary
4	Mistral Medium	1150	+15/-13	6586	Mistral	Proprietary
5	Claude-1	1149	+15/-13	16956	Anthropic	Proprietary
6	Claude-2.0	1131	+14/-13	11204	Anthropic	Proprietary
7	Mixtral-8x7b-Instruct-v0.1	1123	+15/-13	12469	Mistral	Apache 2.0
8	Gemini Pro (Dev)	1120	+18/-18	1898	Google	Proprietary
9	Claude-2.1	1119	+14/-12	20883	Anthropic	Proprietary
10	GPT-3.5-Turbo-0613	1116	+13/-13	26583	OpenAI	Proprietary
11	Gemini Pro	1114	+16/-13	6981	Google	Proprietary
12	Yi-34B-Chat	1111	+16/-15	5055	01 AI	Yi License
13	Claude-Instant-1	1109	+15/-13	16182	Anthropic	Proprietary
14	Tulu-2-DPO-70B	1106	+16/-15	4494	AllenAI/UW	AI2 IMPACT Low-risk
15	GPT-3.5-Turbo-0314	1104	+15/-13	5961	OpenAI	Proprietary
16	WizardLM-70B-v1.0	1104	+17/-14	7531	Microsoft	Llama 2 Community
17	Vicuna-33B	1094	+15/-12	15632	LMSYS	Non-commercial
18	Starling-LM-7B-alpha	1092	+17/-16	3947	UC Berkeley	CC-BY-NC-4.0
19	Llama-2-70b-chat	1079	+15/-13	13514	Meta	Llama 2 Community

<https://chat.lmsys.org/>

クローズドなLLMの特徴: ブラックボックス

当該サービス/APIでの提供となるため、
その中身を利用者が知ることは不可能

またその仕様上、インターネット接続が事実上必須



※ ちなみにクローズドなLLMは
技術詳細についても完全には論文化
されていないことが多い

オープンなLLMとは

クローズドなLLMとは異なり、その中身（パラメータやソースコード）が公開されているLLM

Llama 2 (Meta)

Introducing Llama 2

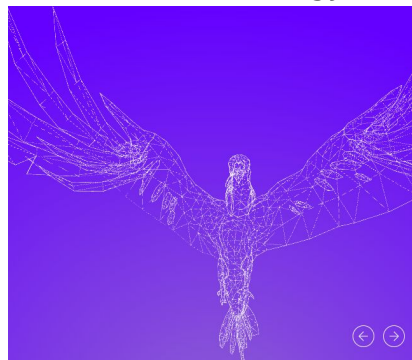
The next generation of our open source large language model

Llama 2 is available for free for research and commercial use.

Download the Model

<https://ai.meta.com/llama/>

Falcon (Technology Innovation Institute)



— Insights

Falcon 180B

Falcon 180B is a super-powerful language model with 180 billion parameters, trained on 3.5 trillion tokens. It's currently at the top of the Hugging Face Leaderboard for pre-trained Open Large Language Models and is available for both research and commercial use..

Read Article →

<https://falconllm.tii.ae/>

オープンなLLMの特徴: パラメータの公開

Hugging Face Hubのようなサイト上でダウンロードすることが可能
このため、計算機環境さえあれば直接使える

Hugging Face Hub上のLlama 2

Llama 2

From Meta

Welcome to the official Hugging Face organization for Llama 2 models from Meta! In order to access models here, please visit the [Meta website](#) and accept our [license terms](#) and [acceptable use policy](#) before requesting access to a model. Requests will be processed within 1-2 days.

Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.

Read [our paper](#), learn more [about the model](#), or get started with [code on GitHub](#).

Llama Model Index

Model	Llama2	Llama2-hf	Llama2-chat	Llama2-chat-hf
7B	Link	Link	Link	Link
13B	Link	Link	Link	Link
70B	Link	Link	Link	Link

<https://huggingface.co/meta-llama>

.gitattributes	1.58 kB	↓	Squashing commit	6 months ago
LICENSE.txt	7.02 kB	↓	Squashing commit	6 months ago
MODEL_CARD.md	7.23 kB	↓	Squashing commit	6 months ago
README.md	10 kB	↓	Fix links	about 2 months ago
Responsible-Use-Guide.pdf	1.25 MB	🔥 LFS	↓ Squashing commit	6 months ago
USE_POLICY.md	4.77 kB	↓	Squashing commit	6 months ago
config.json	614 Bytes	↓	Update config.json	5 months ago
generation_config.json	188 Bytes	↓	Update generation_config.json	5 months ago
model-00001-of-00015.safetensors	9.85 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00002-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00003-of-00015.safetensors	9.97 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00004-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00005-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00006-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00007-of-00015.safetensors	9.97 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00008-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00009-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00010-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00011-of-00015.safetensors	9.97 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00012-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00013-of-00015.safetensors	9.8 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00014-of-00015.safetensors	9.5 GB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model-00015-of-00015.safetensors	524 MB	🔥 LFS	↓ Upload LlamaForCausalLM (#1)	6 months ago
model.safetensors.index.json	66.7 kB	↓	Upload LlamaForCausalLM (#1)	6 months ago

<https://huggingface.co/meta-llama/Llama-2-70b-chat-hf/tree/main>

オープンなLLMの特徴: 透明性

パラメータの公開のほか、モデルによっては学習データ等の詳細が明らかにされていることも多い

Llama 2については論文で詳細に説明されている

LLAMA 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron* Louis Martin† Kevin Stone‡

Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra Prajwal Bhargava Shrutu Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic Sergey Edunov Thomas Scialom*

GenAI, Meta

Abstract

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called LLAMA 2-CHAT, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of LLAMA 2-CHAT in order to enable the community to build on our work and contribute to the responsible development of LLMs.

<https://arxiv.org/abs/2307.09288>

また、学習中の各種ログが公開されているケースも



2021-12-31 02:53 ET [Punit]

Started the training monitor script

```
scripts/cloud/monitor/train_log.py --metrics-to-stdout --shared/home/naamgoyal/checkpoints/1919/2756_run13_57/train_log --modified-threshold 360 --slurm-podid 4314
```

2021-12-31 12:00 ET [Moya]

- Monitor script detects train.log not getting updated; tries to autorecover (6 am ET ish)
- Autorecover is successful.
- ...however I had two monitor scripts running going to two separate emails (<scrubbed>)
 - Which meant recovery took twice as long. (Up at 9 am ish rather than sooner)
 - ...and the emails didn't even send anyway :(
- node-[5,13] seem to be new drained nodes; undrain them and run ./fixmycloud on them to see what's up... Seems like ssh issue
 - pds@ip-0A1E0404: node-5: ssh exited with exit code 15
 - pds@ip-0A1E0404: node-13: ssh exited with exit code 15
- SSH into the two nodes run 'nvidia-smi'
 - Get 'Unable to determine the device handle for GPU 000B:00:00.0: GPU is lost. Reboot the system to recover this GPU' for both
- Put node-[5,13] back into drain mode
- ...and while I happened to be updating the log for this, 150 enters into a "connection timed out mode", same as drained" node.

```
hpc* up infinite 1 drain* node-82
hpc* up infinite 1 idle* node-150
hpc* up infinite 3 drain node-[5,13,148]
hpc* up infinite 1 mix node-11
hpc* up infinite 124 alloc node-[1-4,6-10,12,14-32,34-81,83-92,94,96-98,100-106,108-116,118-134]
hpc* up infinite 11 idle node-[137-146,149]
```

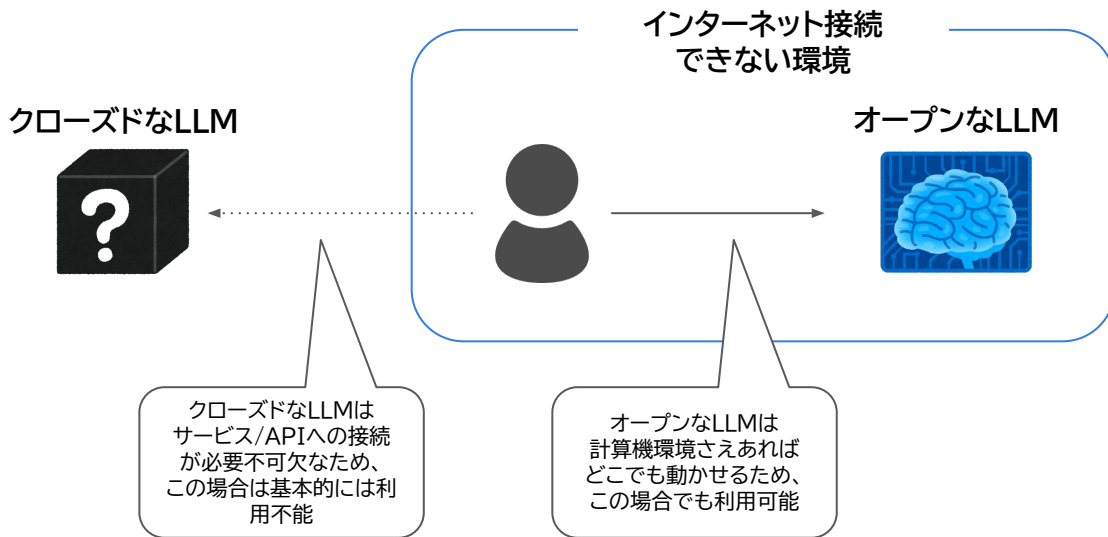
2021-12-30 17:00 ET [Moya] - nvidia_smi.py bug fix; machine check

- Some spares were caught under 'fixmycloud' when they shouldn't have been.

https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/chronicles/OPT175B_Logbook.pdf

オープンなLLMの特徴: 柔軟性

オンプレミス環境などの、インターネット接続できない場面でも利用可能



オープンなLLMの特徴: 場合によってはクローズドLLMより安価

解きたい課題の種類・難易度などに依存するが、
大幅にコストダウンできる場合もある

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	86.4	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	86.1	85.0
Natural Questions (1-shot)	–	–	29.3	37.5	33.0
GSM8K (8-shot)	57.1	92.0	56.5	80.7	56.8
HumanEval (0-shot)	48.1	67.0	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	65.7	51.2

Table 4: Comparison to closed-source models on academic benchmarks. Results for GPT-3.5 and GPT-4 are from OpenAI (2023). Results for the PaLM model are from Chowdhery et al. (2022). Results for the PaLM-2-L are from Anil et al. (2023).

<https://arxiv.org/abs/2307.09288>

オープンなLLMの特徴: 世界的に活発な議論

論文はもとより、X (旧Twitter) やRedditといったコミュニティも含め、日々多くの議論がなされている

新しいLLMやライブラリには大きな注目

Redditでも多くのユーザーで賑わっている

Devendra Chaplot @dchaplot · 7時間
We just released **Mixtral 8x7B** paper on Arxiv:
arxiv.org/abs/2401.04088

Mixtral of Experts

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Guang Leng, Guillaume Lample, Guillaume Lample, Lelio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Srivaran Aravamudan, Trevor Le Scao, Theophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Saeys

Abstract

We introduce Mixtral 8x7B, a Sparse Mixture of Experts (SMoE) language model. Mixtral has the same architecture as Mistral 7B, with the difference that each layer is composed of 8 feedforward blocks (i.e., experts). For every token, at each layer, a router network selects two experts to process the current state and combine their outputs. Even though each token only sees two experts, the selected experts can be different at each timestep. As a result, each token has access to 47B parameters, but only uses 13B active parameters during inference. Mixtral is trained with a context size of 32k tokens and it outperforms or matches Llama 2 70B and GPT-3.5 across all evaluated benchmarks. In particular, Mixtral vastly outperforms Llama 2 70B on mathematics, code generation, and multilingual benchmarks. We also provide a model fine-tuned to follow instructions, **Mixtral 8x7B - Instruct**, that surpasses GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B - chat model on human benchmarks. Both the base and instruct models are released under the Apache 2.0 license.

Code: <https://github.com/mistralai/mistral-src>
Webpage: <https://mistral.ai/news/mistral-of-experts/>

27 370 1,297 17万 19 298 1,098 594

<https://twitter.com/dchaplot/status/1744547220983005478>

Zhuohan Li @zhuohan123
Thrilled to introduce vLLM with @woosuk_k!

vLLM is an open-source LLM inference and serving library that accelerates HuggingFace Transformers by 24x and powers @msysorg Vicuna and Chatbot Arena.

GitHub: [github.com/vllm-project/v...](https://github.com/vllm-project/vllm)
Blog: vllm.ai
[ポストを翻訳](#)

blog.vllm.ai
vLLM: Easy, Fast, and Cheap LLM Serving with PagedAttention
GitHub | Documentation | Paper

午前4:12 · 2023年6月21日 · 38.4万 件の表示

19 298 1,098 594

<https://twitter.com/zhuohan123/status/1671234702206590464>

r/LocalLLaMA

+ Create a post Join

LocalLlama
Subreddit to discuss about Llama, the large language model created by Meta AI.

101K Members 949 Online Top 2% Rank by size

RULES

- 1 Please search before asking
- 2 Off-Topic Posts
- 3 Low Effort Posts
- 4 Limit Self-Promotion
- 5 Follow Reddit's Content Policy

USEFUL LINKS

Basic index for newcomers to LocalLlama

[Index Resources](#)

u/Singularian2501 · 10 hr. ago

WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia - Achieves 97.9% factual accuracy in conversations with human users about recent topics, 55.0% better than GPT-4! - Stanford University 2023

Generation

Paper: <https://arxiv.org/abs/2305.14292v2>

GitHub: <https://github.com/stanford-oval/WikiChat>

Abstract:

This paper presents the first few-shot LLM-based chatbot that almost never hallucinates and has high conversationality and low latency. WikiChat is grounded on the English Wikipedia, the largest curated free-text corpus.

WikiChat generates a response from an LLM, retains only the grounded facts, and combines them with additional information it retrieves from the corpus to form factual and engaging responses. We distill WikiChat based on GPT-4 into a 7B-parameter LLAMA model with minimal loss of quality, to significantly improve its latency, cost and privacy, and facilitate research and deployment.

Using a novel hybrid human-and-LLM evaluation methodology, we show that our

View full post

194 26 Share

<https://www.reddit.com/r/LocalLLaMA/>

ここまでのまとめ: オープンなLLM/クローズドなLLM

求められる性能や使いたい環境にあわせて
選択する必要がある

LLM種別	具体例	パラメータ	利用環境	性能
オープン	Llama 2 Falcon Qwen ...	公開	場面に合わせて 柔軟に 利用可能	全体的にはやや ビハインドだが 一部には匹敵
クローズド	GPT-4 Bard (Gemini) Claude ...	非公開 (サービス/API からのみ利用可)	インターネット 接続は必須	オープンLLMを 大きく上回るもの がいくつか存在

本日も話すこと

1. オープンなLLM/ クローズドなLLM

APIベースで利用可能な、クローズドなLLM（例: GPT-4/Gemini/Claude/…）
に対して、オープンなLLMとはどのようなものか

2. 日本語LLMの近況

ここ数年の、日本語に特化したオープンなLLMの近況について
また、日本語特有の課題についても紹介

3. ELYZAにおける 取り組み内容

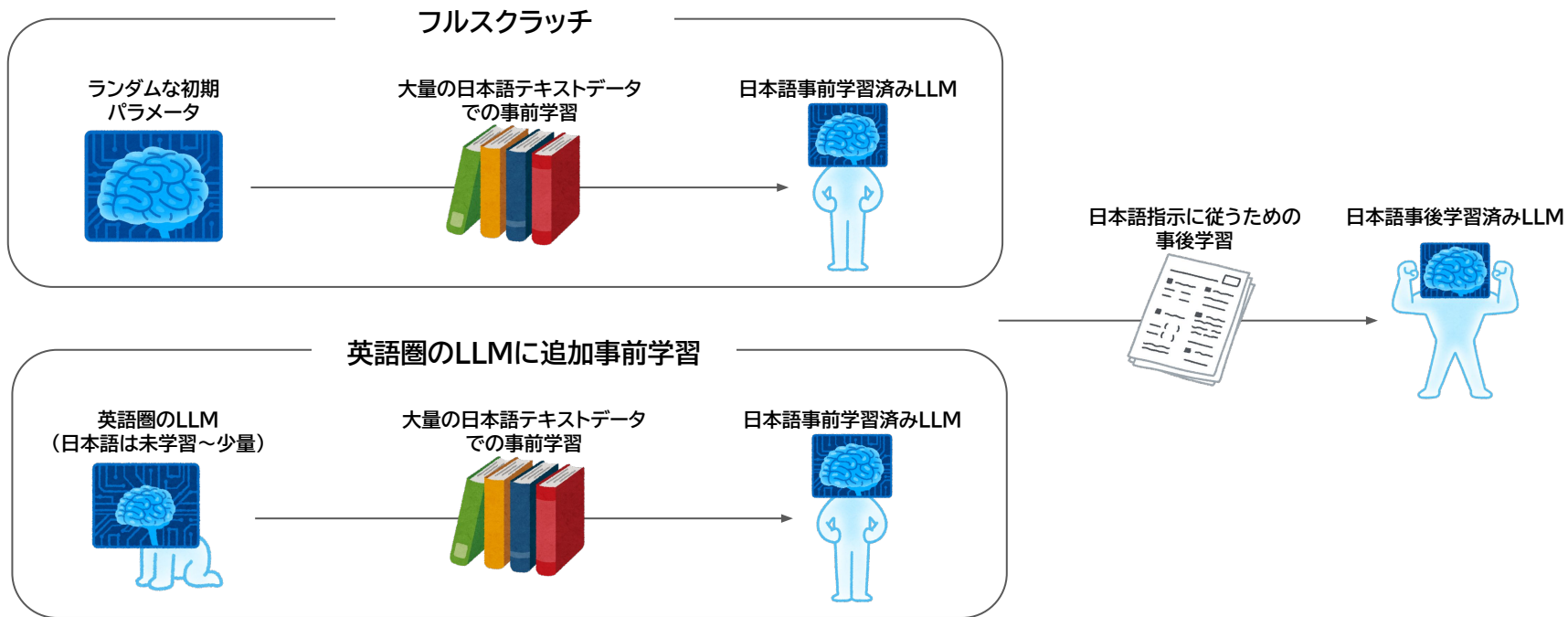
弊社ELYZAがリリースした「ELYZA-japanese-Llama-2」や、
評価用データセット「ELYZA-tasks-100」について

4. 直近ホットな トピック

オープンなLLMの周辺で直近ホットなトピックをいくつか紹介

日本語LLMを作るにあたっての方向性

特に事前学習（基本的な言語能力を得るための学習）
について、大まかに2つの方向性が存在



日本語LLM学習の方向性: フルスクラッチ

ランダムな初期パラメータから、最初から日本語で学習する



メリット

何も学習されていないまっさらな状態から学習できるため、コントロールがしやすい

独自のモデル構造に仕上げることも可能となる

デメリット

何も知らない状態の初期パラメータから学習することとなるため、より多くの事前学習データ・計算コストなどが必要

事後学習についても、各種能力（指示追従能力、安全性など）を全て新たに教え込まなければならない

日本語LLM学習の方向性: 英語圏のLLMに追加事前学習

例えばMeta社のLlama 2など、英語で優れた性能を発揮しているLLMに日本語を追加で学習する



メリット

元のLLMの能力を引き継ぐことにより、事前学習/事後学習に必要なデータ量が少なく済む場合がある

またそれに伴い、学習コストの効率化にも寄与する

デメリット

良くも悪くも元のLLMの性質を消し切ることが難しいため、望ましくない点があったとしても引き継がれてしまう

元のLLMの学習データが完全に公開されていない場合、不透明な部分が残ってしまう

ライセンスについても元モデルに影響される

日本語LLMの課題: 事前学習用データセット

英語のような話者が多い言語に比べて、
日本語データセットは圧倒的に少ない

例えばLlama 2は2trillion (2兆) トークンで事前学習されているが、
そのうち日本語は0.10% (20億) トークン程度のみ含まれる

Training Data		Params	Context Length	GQA	Tokens	LR
LLAMA 1	See Touvron et al. (2023)	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
LLAMA 2	A new mix of publicly available online data	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

Table 1: LLAMA 2 family of models. Token counts refer to pretraining data only. All models are trained with a global batch-size of 4M tokens. Bigger models — 34B and 70B — use Grouped-Query Attention (GQA) for improved inference scalability.

Wikipediaの日本語記事数は、英語の1/5程度にとどまっている

順位	言語	原語表記 / 英語記事	WP	純記事数	総項目数	総編集数	管理者数	登録者数	活動中の登録者数	ファイル数	点数A	点数B
1	英語	English	en	6,767,850	59,725,338	1,189,492,881	876	46,743,194	114,575	914,084	1,219	98.36
2	セブアノ語	Cebuano	ceb	6,122,241	11,228,039	35,023,225	5	111,015	185	0	2	6.30
3	ドイツ語	Deutsch	de	2,870,215	7,913,795	239,378,558	175	4,290,536	17,306	129,243	93	82.83
4	フランス語	français	fr	2,582,518	12,896,072	210,829,321	148	4,832,280	17,118	71,087	260	84.92
5	スウェーデン語	svenska	sv	2,574,077	6,213,576	54,219,590	66	894,061	2,057	0	17	51.82
6	オランダ語	Nederlands	nl	2,146,247	4,595,747	66,634,696	35	1,319,697	3,632	20	18	53.91
7	ロシア語	русский	ru	1,956,794	7,856,114	135,118,401	70	3,510,034	9,558	247,870	156	100.00
8	スペイン語	español	es	1,921,639	8,074,421	156,358,334	59	7,038,282	13,425	0	198	99.94
9	イタリア語	italiano	it	1,842,344	7,887,873	136,939,642	121	2,468,215	7,341	125,012	186	76.70
10	エジプト・アラビア語	مصرى	arz	1,621,340	2,101,541	8,933,770	7	228,218	232	1,488	0	11.38
11	ポーランド語	polSKI	pl	1,596,909	3,676,902	72,225,047	99	1,282,407	4,476	261	33	55.67
12	日本語	日本語	ja	1,399,034	4,116,611	98,560,966	40	2,154,474	13,339	4,388	90	85.11
13	中国語	中文	zh	1,397,380	7,591,750	80,097,007	63	3,454,611	7,542	65,799	207	99.69
14	ウクライナ語	українська	uk	1,304,224	4,573,090	41,299,724	51	727,846	3,379	112,831	56	90.32
15	ベトナム語	Tiếng Việt	vi	1,289,733	19,452,413	70,980,036	18	940,439	1,866	26,335	723	76.79
16	ワライ語	Winaray	war	1,266,367	2,870,904	7,597,454	3	58,413	57	42	4	18.61
17	アラビア語	العربية	ar	1,225,163	8,327,418	65,527,845	22	2,530,516	3,957	53,537	264	57.80
18	ポルトガル語	português	pt	1,116,619	5,629,957	67,075,614	52	3,016,511	8,274	60,662	194	74.37

<https://ja.wikipedia.org/wiki/Wikipedia:全言語版の統計>

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Table 10: Language distribution in pretraining data with percentage $\geq 0.005\%$. Most data is in English, meaning that LLAMA 2 will perform best for English-language use cases. The large unknown category is partially made up of programming code data.

<https://arxiv.org/abs/2307.09288>

日本語LLMの課題: 事後学習用データセット

事後学習用データセットも英語圏に比べて まだまだ少ない状況

多言語のチャット形式データ (OpenAssistant)
中でも、日本語は英語に比べて2%程度のサイズ

2023-11-05_oasst2_all.trees.json.gz Stats

```
Trees          : 70,642
Messages       : 208,584
Oldest message : 2023-01-16 20:24:26.211711+00:00
Youngest message : 2023-11-05 10:24:44.484910+00:00
Detoxify ratings : 156,570
Accepted messages: 189,288
Deleted messages : 5,414
Tree counts by state:
- ready_for_export: 13,854
- prompt_lottery_waiting: 44,550
- halted_by_moderator: 3,089
- initial_prompt_review: 4,319
- growing: 3,102
- aborted_low_grade: 1,708
- ranking: 20
Message counts by language:
- en: 85,115
- es: 47,513
- ru: 15,990
- zh: 11,205
- de: 8,398
- fr: 5,841
- pt-BR: 4,540
- th: 3,236
- ca: 2,586
- it: 2,144
- ja: 1,904
```

<https://huggingface.co/datasets/OpenAssistant/oasst2>

直近、理研AIPにて日本語特化した
事後学習用データセットが数千件公開

LLMのための日本語インストラクション データ作成プロジェクト

新着情報

- [2903件のデータ\(ver.003-001\)を公開](#) (2023/12/23)
- [1003件のデータ\(ver.002-002\)を公開](#) (2023/11/3)
- 「第2回 説明会」を行いました (2023/10/2)
 - 本ページ (下記) で資料、ビデオを公開しています。
- [50件のサンプルデータを公開](#) (2023/10/1)
- [日経新聞に掲載されました](#) (2023/9/26)

[https://liat-aip.sakura.ne.jp/wp/llmのための日本語インストラクションデータ作成 /](https://liat-aip.sakura.ne.jp/wp/llmのための日本語インストラクションデータ作成/)

日本語LLMの課題: 追加学習の場合の、元のLLMが持っていた能力の忘却

単純に日本語を追加学習すると、英語で学んでいた能力を忘れてしまう場合がある

モデル	チェックポイント (ステップ数)	ARC_C	HellaSwag	TruthfulQA	MMLU	Average(4 Tasks)
Llama-2-7b-chat		52.73	78.52	45.32	48.14	56.18
ELYZA-japanese-Llama-2-7b	1000	52.99	78.15	41.41	47.65	55.05(Δ1.13)
	3000	52.82	77.69	40.71	46.52	54.44(Δ1.74)
	5000	52.22	77.24	39.31	46.39	53.79(Δ2.39)
	7000	51.62	76.89	38.37	46.33	53.30(Δ2.88)
	9000	51.71	76.53	38.02	44.94	52.80(Δ3.38)
ELYZA-japanese-Llama-2-7b-bilingual	1000	50.77	76.09	43.19	48.26	54.58(Δ1.60)
	3000	50.94	76.39	43.28	48.62	54.81(Δ1.37)
	5000	51.19	76.70	43.62	47.93	54.86(Δ1.32)
	7000	50.94	76.62	43.14	47.45	54.54(Δ1.64)
	9000	51.11	75.93	43.66	46.92	54.41(Δ1.77)

日本語のみで追加学習してしまうと学習を進めるごとに英語の能力が少しずつ失われてしまっている

英語も含めて追加学習することで英語能力の忘却がやや緩和されることを確認

<https://zenn.dev/elyza/articles/ab3749de0ba58b>

本日本話すること

1. オープンなLLM/ クローズドなLLM

APIベースで利用可能な、クローズドなLLM（例: GPT-4/Gemini/Claude/…）
に対して、オープンなLLMとはどのようなものか

2. 日本語LLMの近況

ここ数年の、日本語に特化したオープンなLLMの近況について
また、日本語特有の課題についても紹介

3. ELYZAにおける 取り組み内容

弊社ELYZAがリリースした「ELYZA-japanese-Llama-2」や、
評価用データセット「ELYZA-tasks-100」について

4. 直近ホットな トピック

オープンなLLMの周辺で直近ホットなトピックをいくつか紹介

ELYZAが公開中のモデル

2023年8月にリリースした「ELYZA-japanese-Llama-2-7b」など、直近で大きく分けて3種類のLLMを公開



ELYZA, Inc.
@ELYZA_inc

【お知らせ】Llama 2をベースとした商用利用可能な70億パラメータの日本語言語モデルを公開しました。ブログにて特徴や性能について紹介しているほか、推論用コード、性能評価用データセットとその評価結果もすべて公開しています。(1/3)



note.comから

午前11:00 · 2023年8月29日 · 54.1万 件の表示

384 リポスト 74 件の引用 1,027 件のいいね 343 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Llama 2をベースとした商用利用可能な130億パラメータの日本語言語モデルを公開しました。オープンな日本語モデルの中で最高性能、GPT-3.5 (text-davinci-003)を上回る性能となり、高速に動作するデモも公開中です。(1/3)

内容は以下をクリックしてご覧ください。



note.comから

午前11:00 · 2023年12月27日 · 18.3万 件の表示

225 リポスト 47 件の引用 675 件のいいね 236 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Code Llamaに対し日本語の追加事前学習を行い、コード生成・補完に特化した日本語LLM「ELYZA-japanese-CodeLlama-7b」を開発しました。noteにて概要を紹介しています。(1/3)



note.comから

午前11:00 · 2023年11月15日 · 5.7万 件の表示

79 リポスト 11 件の引用 259 件のいいね 78 ブックマーク

ELYZAが公開中のモデル

2023年8月にリリースした「ELYZA-japanese-Llama-2-7b」など、直近で大きく分けて3種類のLLMを公開



ELYZA, Inc.
@ELYZA_inc

【お知らせ】Llama 2をベースとした商用利用可能な70億パラメータの日本語言語モデルを公開しました。ブログにて特徴や性能について紹介しているほか、推論用コード、性能評価用データセットとその評価結果もすべて公開しています。(1/3)



note.comから

午前11:00 · 2023年8月29日 · 54.1万 件の表示

384 リポスト 74 件の引用 1,027 件のいいね 343 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Llama 2をベースとした商用利用可能な130億パラメータの日本語言語モデルを公開しました。オープンな日本語モデルの中で最高性能、GPT-3.5 (text-davinci-003)を上回る性能となり、高速に動作するデモも公開中です。(1/3)

内容は以下をクリックしてご覧ください。



note.comから

午前11:00 · 2023年12月27日 · 18.3万 件の表示

225 リポスト 47 件の引用 675 件のいいね 236 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Code Llamaに対し日本語の追加事前学習を行い、コード生成・補完に特化した日本語LLM「ELYZA-japanese-CodeLlama-7b」を開発しました。noteにて概要を紹介しています。(1/3)



note.comから

午前11:00 · 2023年11月15日 · 5.7万 件の表示

79 リポスト 11 件の引用 259 件のいいね 78 ブックマーク

ELYZA-japanese-Llama-2-13b

Llama 2 (13B) をもとに日本語を追加学習したLLM

評価用データセットELYZA-tasks-100 (後述) では
GPT-3.5 (text-davinci-003) を上回る性能

モデル	スコア	開発元	パラメータ数	Open / Closed	商用利用
Claude 2.1	3.84	Anthropic	不明	Closed	-
Gemini Pro	3.72	Google	不明	Closed	-
ELYZA-japanese-Llama-2-13b-instruct	3.01	ELYZA	13B	Open	○
Qwen-14B-Chat	2.78	Alibaba	14B	Open	○
GPT-3.5 (text-davinci-003)	2.77	OpenAI	175B	Closed	-
ELYZA-japanese-Llama-2-13b-fast-instruct	2.73	ELYZA	13B	Open	○
calm2-7b-chat	2.63	CyberAgent	7B	Open	○
japanese-stablelm-instruct-beta-70b	2.62	Stability AI	70B	Open	○
Swallow-70b-instruct	2.50	TokyoTech-LLM	70B	Open	○
nekomata-14b-instruction	2.50	rinna	14B	Open	○
Swallow-13b-instruct	2.34	TokyoTech-LLM	13B	Open	○
nekomata-7b-instruction	2.23	rinna	7B	Open	○
Swallow-7b-instruct	2.22	TokyoTech-LLM	7B	Open	○
youri-7b-instruction	2.00	rinna	7B	Open	○
japanese-stablelm-instruct-gamma-7b	1.87	Stability AI	7B	Open	○
plamo-13b-instruct	1.49	Preferred Networks	13B	Open	○
japanese-stablelm-instruct-beta-7b	1.43	Stability AI	7B	Open	○
llm-jp-13b-instruct-full-jaster-v1.0	1.31	LLM-jp	13B	Open	○

AIアシスタントとしてより気の利いた回答を実現していたり、人間でも想像力が必要とされるような指示も対応

自律神経や副交感神経が乱れている、とはどのような状態ですか？科学的に教えて



自律神経とは、意識しなくても呼吸をするように、心臓を動かしたり、内臓を収縮させたりする神経のことで、自律神経は「副交感神経」と「交感神経」によって管理されています。

副交感神経は、心拍数や血圧を下げる働きがあり、リラックスしているときや睡眠中に優位に働きます。一方、交感神経は、心拍数や血圧を上げる働きがあり、覚醒しているときや運動中に優位に働きます。

自律神経が乱れているということは、副交感神経と交感神経の両方が乱れているということです。具体的には、

- 副交感神経が優位に働きすぎると、心拍数や血圧が下がりすぎ、冷え性、動悸、息が詰まる感じ、眠気などの症状が現れます。
- 交感神経が優位に働きすぎると、心拍数や血圧が上がりすぎ、頭痛、動悸、不眠、イライラ感などの症状が現れます。



「メタリック」から「気分上々」までが自然につながるように、あいだの単語を連想してください。



承知しました。「メタリック」から「気分上々」までの間にある単語を並べます。

メタリック - 光沢 - 輝き - キラキラ - 華やか - 明朗 - 気分上々

これらの単語はあくまで一意見でしかありません。



詳細は以下記事をご参照ください

<https://note.com/elyza/n/n5d42686b60b7>

また、以下からデモを触ることもできます

<https://huggingface.co/spaces/elyza/ELYZA-japanese-Llama-2-13b-instruct-demo>

ELYZAが公開中のモデル

2023年8月にリリースした「ELYZA-japanese-Llama-2-7b」など、直近で大きく分けて3種類のLLMを公開



ELYZA, Inc.
@ELYZA_inc

【お知らせ】Llama 2をベースとした商用利用可能な70億パラメータの日本語言語モデルを公開しました。ブログにて特徴や性能について紹介しているほか、推論用コード、性能評価用データセットとその評価結果もすべて公開しています。(1/3)



note.comから

午前11:00 · 2023年8月29日 · 54.1万 件の表示

384 リポスト 74 件の引用 1,027 件のいいね 343 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Llama 2をベースとした商用利用可能な130億パラメータの日本語言語モデルを公開しました。オープンな日本語モデルの中で最高性能、GPT-3.5 (text-davinci-003)を上回る性能となり、高速に動作するデモも公開中です。(1/3)

内容は以下をクリックしてご覧ください。



note.comから

午前11:00 · 2023年12月27日 · 18.3万 件の表示

225 リポスト 47 件の引用 675 件のいいね 236 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Code Llamaに対し日本語の追加事前学習を行い、コード生成・補完に特化した日本語LLM「ELYZA-japanese-CodeLlama-7b」を開発しました。noteにて概要を紹介しています。(1/3)



note.comから

午前11:00 · 2023年11月15日 · 5.7万 件の表示

79 リポスト 11 件の引用 259 件のいいね 78 ブックマーク

ELYZA-japanese-CodeLlama-7b

Code Llama (7B) をもとに日本語を追加学習したLLM

コード生成およびコード補完に関する能力に特化

以下はUnionFindのPython実装です。このコードに対し、ユニットテストのコードを書いてください

```
class UnionFind:
    def __init__(self, N):
        self.rank = [0]*N
        self.par = list(range(N))

    def find(self, x):
        if x != self.par[x]:
            self.par[x] = self.find(self.par[x])
        return self.par[x]

    def unite(self, x, y):
        x, y = self.find(x), self.find(y)
        if(self.rank[x] > self.rank[y]):
            self.par[y] = x
        else:
            self.par[x] = y
        if(self.rank[x] == self.rank[y]):
            self.rank[y] += 1
```



承知しました。以下にユニットテストのコードを記述します。

```
import unittest

class UnionFindTest(unittest.TestCase):
    def test_find(self):
        uf = UnionFind(5)
        self.assertEqual(uf.find(0), 0)
        self.assertEqual(uf.find(1), 1)
        self.assertEqual(uf.find(2), 2)
        self.assertEqual(uf.find(3), 3)
        self.assertEqual(uf.find(4), 4)

    def test_unite(self):
        uf = UnionFind(5)
        uf.unite(0, 1)
        uf.unite(2, 3)
        uf.unite(4, 4)
        self.assertEqual(uf.find(0), uf.find(1))
        self.assertEqual(uf.find(2), uf.find(3))
        self.assertEqual(uf.find(4), uf.find(4))

if __name__ == "__main__":
    unittest.main()
```



ディレクトリ /home/llama/data 以下のCSVファイルをすべて読み込んでpandasのDataFrameにしてから、それらを結合して



承知しました。以下はPythonのコードです。

```
import pandas as pd
import os

def read_csvs(directory):
    csvs = []
    for filename in os.listdir(directory):
        if filename.endswith('.csv'):
            csvs.append(pd.read_csv(os.path.join(directory, filename)))
    return pd.concat(csvs)

df = read_csvs('/home/llama/data')
```

このコードは、ディレクトリ /home/llama/data 以下のすべてのCSVファイルを読み込み、それらを結合してDataFrameを作成します。



詳細については以下の記事をご参照ください
<https://note.com/elyza/n/n5bce23d7c9c8>

こちらも以下からデモを触ることができます
<https://huggingface.co/spaces/elyza/ELYZA-japanese-CodeLlama-7b-instruct-demo>

ELYZAが公開中のモデル

2023年8月にリリースした「ELYZA-japanese-Llama-2-7b」など、直近で大きく分けて3種類のLLMを公開



ELYZA, Inc.
@ELYZA_inc

【お知らせ】Llama 2をベースとした商用利用可能な70億パラメータの日本語言語モデルを公開しました。ブログにて特徴や性能について紹介しているほか、推論用コード、性能評価用データセットとその評価結果もすべて公開しています。(1/3)



note.comから

午前11:00 · 2023年8月29日 · 54.1万 件の表示

384 リポスト 74 件の引用 1,027 件のいいね 343 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Llama 2をベースとした商用利用可能な130億パラメータの日本語言語モデルを公開しました。オープンな日本語モデルの中で最高性能、GPT-3.5 (text-davinci-003)を上回る性能となり、高速に動作するデモも公開中です。(1/3)

内容は以下をクリックしてご覧ください。



note.comから

午前11:00 · 2023年12月27日 · 18.3万 件の表示

225 リポスト 47 件の引用 675 件のいいね 236 ブックマーク



ELYZA, Inc.
@ELYZA_inc

Code Llamaに対し日本語の追加事前学習を行い、コード生成・補完に特化した日本語LLM「ELYZA-japanese-CodeLlama-7b」を開発しました。noteにて概要を紹介しています。(1/3)



note.comから

午前11:00 · 2023年11月15日 · 5.7万 件の表示

79 リポスト 11 件の引用 259 件のいいね 78 ブックマーク

取り組みの背景: Llama 2

Llama 2は2兆トークンという超大規模な事前学習がなされていて、かつ安全性が高かったり直近の技術が総動員されていたりと条件が揃っていたため、日本語追加学習の対象として選定

タスクによってはChatGPT (0301) に匹敵

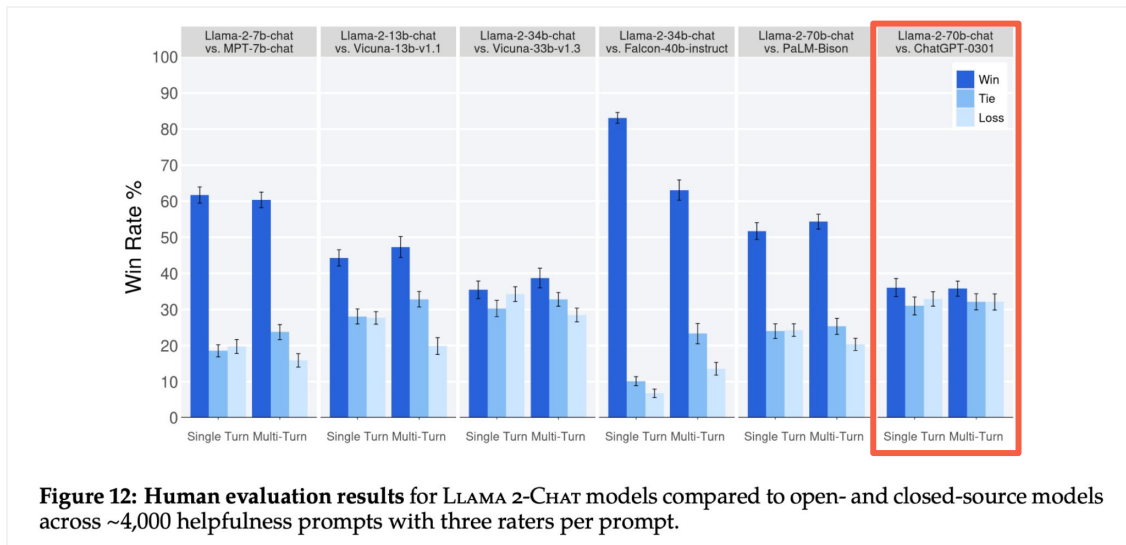


Figure 12: Human evaluation results for LLAMA 2-CHAT models compared to open- and closed-source models across ~4,000 helpfulness prompts with three raters per prompt.

<https://arxiv.org/abs/2307.09288>

評価用データセット: ELYZA-tasks-100

汎用的な能力を測るための高難易度な 日本語評価用データセットを作成

背景

- 既存のデータセットでは、タスクの多様さや複雑さが不足し、LLMの汎化性能を十分に評価できない
- 自動評価指標では生成AIの性能を正確に評価することができないため、最終的には人間による評価を行う必要がある

データの特徴

- 人間による評価を行うことが可能な100件のデータセットを人手で作成
 - 100件それぞれに詳細かつ明確な評価基準を定義
- 多様で複雑なタスクを含む

ELYZAモデルの出力例 (ELYZA-tasks-100)

1. 彼は比類のない陸上選手だ。
2. 彼は比較的に良い陸上選手だ。

1の文が難しいので2の文に直そうと思っているのですが、これってあってますか？

1の文章は、彼が陸上選手として圧倒的な優位性を持っていることを表しています。彼が比類のない選手であることは、彼が他の選手と比べて圧倒的に優れているということです。

一方、2の文章は、彼が良い陸上選手だということを表しています。彼が比較的に良い選手であることは、彼が他の選手と比べて良いということです。

従って、1の文章と2の文章は意味が異なります。1の文章を2の文章に直すことはできません。

ELYZAモデルの出力例 (ELYZA-tasks-100)

「今日は少し雨が降っている。外出するのは億劫だけど、近くのコンビニまで買い物をしなければならぬ。傘を持っていくかどうか迷う。」

このシチュエーションで、以下の2つの選択肢のメリットとデメリットについて説明してください。

- A. 傘を持っていかない
- B. 傘を持っていく

シチュエーションにおけるAとBの選択肢について、それぞれのメリットとデメリットを説明します。

A. 傘を持っていかない

- メリット: コンビニまでの距離が少ないため、雨で濡れることは少ない。
- デメリット: 雨が降っているため濡れてしまう。

B. 傘を持っていく

- メリット: 傘を使用することで、雨で濡れることなくコンビニまで買い物をすることができる。
- デメリット: 外出することに対して億劫になる。

このシチュエーションでは、Aの選択肢をお勧め致します。傘を使用することで、雨で濡れることなく買い物をすることができます。ただし、雨が降っているために濡れてしまう可能性があります。

本日本話すること

1. オープンなLLM/ クローズドなLLM

APIベースで利用可能な、クローズドなLLM（例: GPT-4/Gemini/Claude/…）
に対して、オープンなLLMとはどのようなものか

2. 日本語LLMの近況

ここ数年の、日本語に特化したオープンなLLMの近況について
また、日本語特有の課題についても紹介

3. ELYZAにおける 取り組み内容

弊社ELYZAがリリースした「ELYZA-japanese-Llama-2」や、
評価用データセット「ELYZA-tasks-100」について

4. 直近ホットな トピック

オープンなLLMの周辺で直近ホットなトピックをいくつか紹介

推論を高速化したりスループットを高めたりといった、 効率化のためのライブラリも発展しつつある



Easy, fast, and cheap LLM serving for everyone

[Documentation](#) | [Blog](#) | [Paper](#) | [Discord](#)

Latest News 🔥

- [2023/12] Added ROCm support to vLLM.
- [2023/10] We hosted [the first vLLM meetup](#) in SF! Please find the meetup slides [here](#).
- [2023/09] We created our [Discord server!](#) Join us to discuss vLLM and LLM serving! We will also post the latest announcements and updates there.
- [2023/09] We released our [PagedAttention paper](#) on arXiv!
- [2023/08] We would like to express our sincere gratitude to [Andresen Horowitz](#) (a16z) for providing a generous grant to support the open-source development and research of vLLM.
- [2023/07] Added support for LLaMA-2! You can run and serve 7B/13B/70B LLaMA-2s on vLLM with a single command!
- [2023/06] Serving vLLM on any Cloud with SkyPilot. Check out a 1-click [example](#) to start the vLLM demo, and the [blog post](#) for the story behind vLLM development on the clouds.
- [2023/06] We officially released vLLM! FastChat-vLLM integration has powered [LMSYS Vicuna](#) and [Chatbot Arena](#) since mid-April. Check out our [blog post](#).

<https://github.com/vllm-project/vllm>

弊社が先日公開した以下デモでも活用しています
<https://huggingface.co/spaces/elvza/ELYZA-japanese-Llama-2-13b-instruct-demo>

TensorRT-LLM

A TensorRT Toolbox for Optimized Large Language Model Inference

[docs](#) [latest](#) [python](#) [v1.0.12](#) [cuda](#) [v22.2](#) [trt](#) [8.6](#) [release](#) [9.7.0](#) [license](#) [Apache 2.0](#)

[Architecture](#) | [Results](#) | [Examples](#) | [Documentation](#)

Latest News

- [2023/12/04] [Falcon-180B on a single H200 GPU with INT4 AWQ, and 6.7x faster Llama-70B over A100](#)



H200 with INT4 AWQ, runs Falcon-180B on a single GPU.

H200 is now 2.4x faster on Llama-70B with recent improvements to TensorRT-LLM GQA; up to 6.7x faster than A100.

- [2023/11/27] [SageMaker LMI now supports TensorRT-LLM - improves throughput by 60%, compared to previous version](#)
- [2023/11/13] [H200 achieves nearly 12,000 tok/sec on Llama2-13B](#)
- [2023/10/22] [RAG on Windows using TensorRT-LLM and Llamaindex](#)
- [2023/10/19] [Getting Started Guide - Optimizing Inference on Large Language Models with NVIDIA TensorRT-LLM, Now Publicly Available](#)
- [2023/10/17] [Large Language Models up to 4x Faster on RTX With TensorRT-LLM for Windows](#)

<https://github.com/NVIDIA/TensorRT-LLM>

CTranslate2

CTranslate2 is a C++ and Python library for efficient inference with Transformer models.

The project implements a custom runtime that applies many performance optimization techniques such as weights quantization, layers fusion, batch reordering, etc., to [accelerate and reduce the memory usage](#) of Transformer models on CPU and GPU.

The following model types are currently supported:

- Encoder-decoder models: Transformer base/big, M2M-100, NLLB, BART, mBART, Pegasus, T5, Whisper
- Decoder-only models: GPT-2, GPT-J, GPT-NeoX, OPT, BLOOM, MPT, Llama, Mistral, CodeGen, GPTBigCode, Falcon
- Encoder-only models: BERT, DistilBERT, XLM-RoBERTa

Compatible models should be first converted into an optimized model format. The library includes converters for multiple frameworks:

- [OpenNMT-py](#)
- [OpenNMT-tf](#)
- [Fairseq](#)
- [Marian](#)
- [OPUS-MT](#)
- [Transformers](#)

The project is production-oriented and comes with [backward compatibility guarantees](#), but it also includes experimental features related to model compression and inference acceleration.

<https://github.com/OpenNMT/CTranslate2>

Mixture of Experts (MoE)

オープンなLLMの中ではずば抜けた性能となっているMixtralを筆頭に、Mixture of Experts (MoE) というアーキテクチャが流行

一部のクローズドなLLMをも超える性能

モデルの途中で複数の専門家 (expert) のいずれかにルーティング。

高い性能を実現しつつ、計算コストも抑えられる

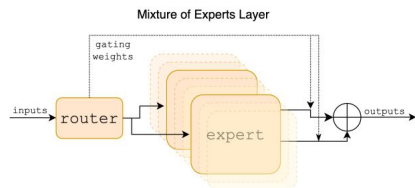


Figure 1: Mixture of Experts Layer. Each input vector is assigned to 2 of the 8 experts by a router. The layer's output is the weighted sum of the outputs of the two selected experts. In Mixtral, an expert is a standard feedforward block as in a vanilla transformer architecture.

<https://arxiv.org/abs/2401.04088>

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-Turbo	1249	+14/-13	23069	OpenAI	Proprietary
2	GPT-4-0314	1190	+14/-14	16237	OpenAI	Proprietary
3	GPT-4-0613	1160	+14/-12	20884	OpenAI	Proprietary
4	Mistral_Medium	1150	+15/-13	6586	Mistral	Proprietary
5	Claude-1	1149	+15/-13	16956	Anthropic	Proprietary
6	Claude-2.0	1131	+14/-13	11204	Anthropic	Proprietary
7	Mixtral-8x7b-Instruct-v0.1	1123	+15/-13	12469	Mistral	Apache 2.0
8	Gemini_Pro_(Dev)	1120	+18/-18	1898	Google	Proprietary
9	Claude-2.1	1119	+14/-12	20883	Anthropic	Proprietary
10	GPT-3.5-Turbo-0613	1116	+13/-13	26583	OpenAI	Proprietary
11	Gemini_Pro	1114	+16/-13	6981	Google	Proprietary
12	Yi-34B-Chat	1111	+16/-15	5055	01 AI	Yi License
13	Claude-Instant-1	1109	+15/-13	16182	Anthropic	Proprietary
14	Tulu-2-DPO-70B	1106	+16/-15	4494	AllenAI/UW	AI2 ImpACT Low-risk
15	GPT-3.5-Turbo-0314	1104	+15/-13	5961	OpenAI	Proprietary
16	WizardLM-70B-v1.0	1104	+17/-14	7531	Microsoft	Llama 2 Community
17	Vicuna-33B	1094	+15/-12	15632	LMSYS	Non-commercial
18	Starling-LM-7B-alpha	1092	+17/-16	3947	UC Berkeley	CC-BY-NC-4.0
19	Llama-2-70b-chat	1079	+15/-13	13514	Meta	Llama 2 Community

<https://chat.lmsys.org/>

Adapter

巨大なモデルをより少ない計算リソースで学習するための手法 LoRA/QLoRAが特に流行中

LoRAはもとの巨大なパラメータを直接更新せずに
別途少量のパラメータを学習することで効率化

QLoRAはそれに加えて、パラメータを量子化
するなどの工夫により更に効率化を実現

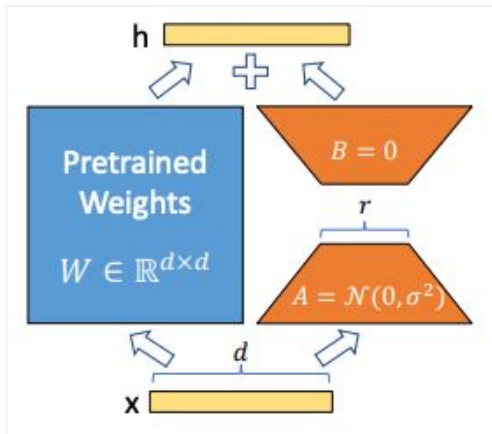


Figure 1: Our reparametrization. We only train A and B .

<https://arxiv.org/abs/2106.09685>

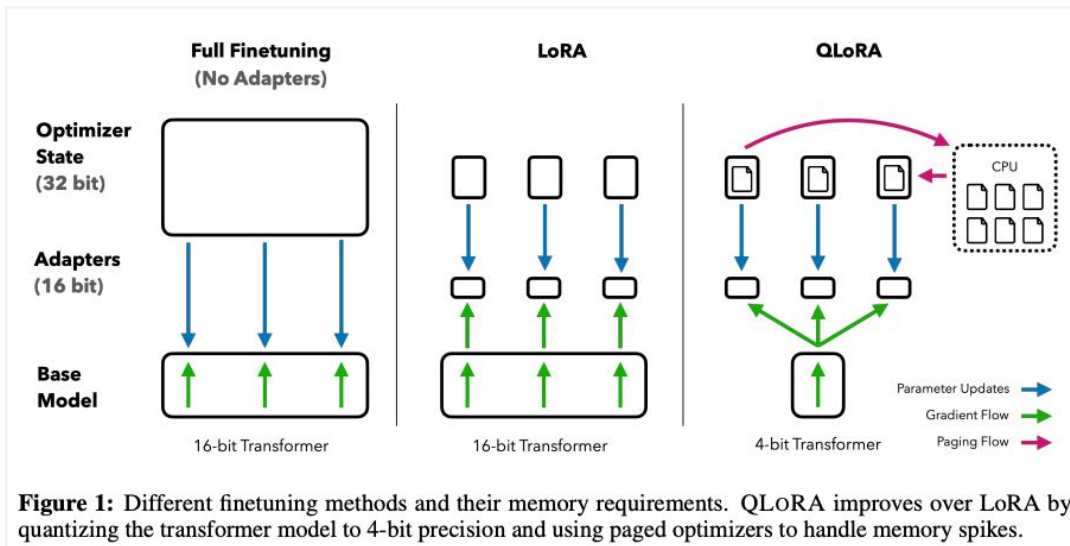
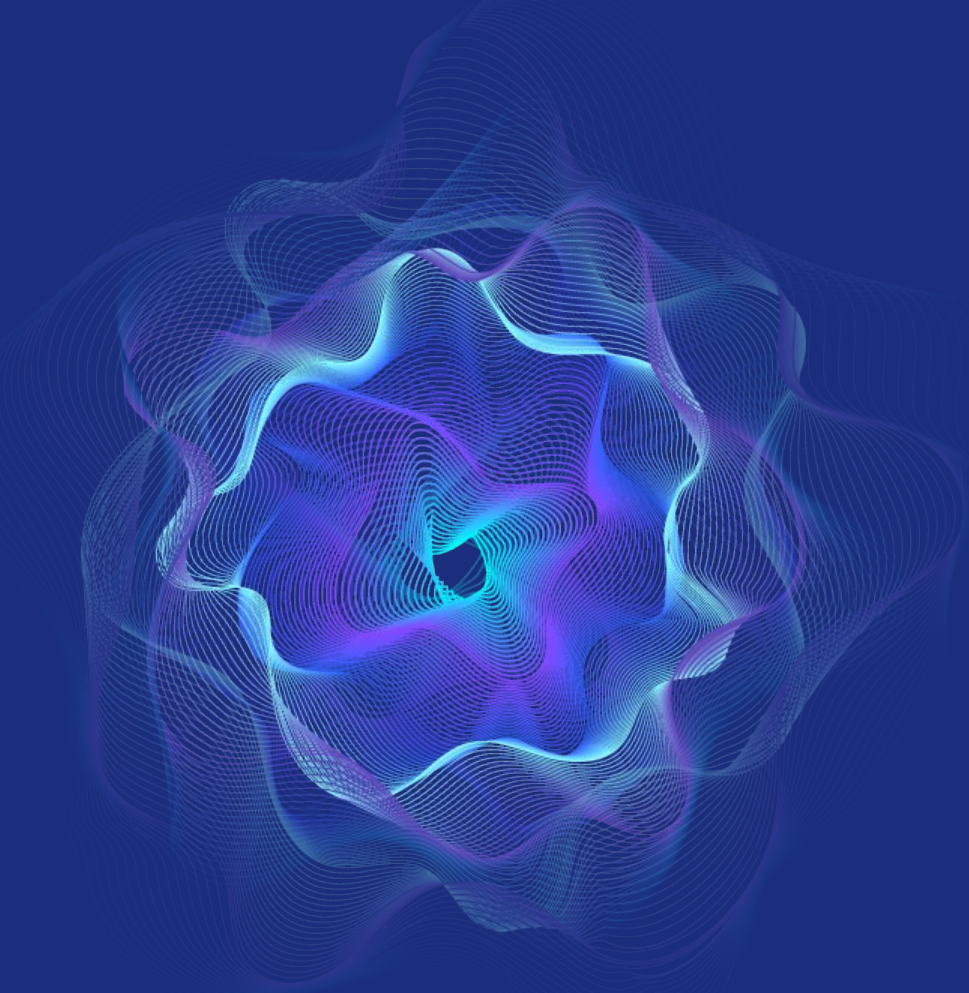


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

<https://arxiv.org/abs/2305.14314>

おわりに



本日も話したこと

1. オープンなLLM/ クローズドなLLM

APIベースで利用可能な、クローズドなLLM（例: GPT-4/Gemini/Claude/…）
に対して、オープンなLLMとはどのようなものか

2. 日本語LLMの近況

ここ数年の、日本語に特化したオープンなLLMの近況について
また、日本語特有の課題についても紹介

3. ELYZAにおける 取り組み内容

弊社ELYZAがリリースした「ELYZA-japanese-Llama-2」や、
評価用データセット「ELYZA-tasks-100」について

4. 直近ホットな トピック

オープンなLLMの周辺で直近ホットなトピックをいくつか紹介