

大規模言語処理モデルGPT-4/LLMについて

目次

第1章（前編）

- 1.1. 自然言語処理および大規模言語モデルについて
- 1.2. 機械学習について
- 1.3. 人工知能と深層学習について
- 1.4. 大規模機械学習モデル LLM(ChatGPT/GPT-4)

大規模言語処理モデルGPT-4/LLMについて

目次

第1章（後編）

- 1.5. ChatGPT/GPT-4のはじめ方
- 1.6. LLM大規模言語モデルおよびライブラリの紹介
- 1.7. 大型自然言語処理のGPTモデル応用例
- 1.8. その他GPT以外の特殊なモデルの紹介
- 1.9. 大規模言語モデルのAI活用例

人工知能プログラミング

実践入門(参考書籍)

GPT-4（Generative Pre-trained Transformer） 、 ChatGPTの大規模言語モデルの活用例、 機械学習、 深層学習の概要・仕組みと関係を学ぶ

Python3.10/LlamaIndex0.6.12/LangChain0.0.181

[PDF版ダウンロード](#)

[本書籍に使われるサンプルデータのダウンロード](#)

1.1.1. 自然言語処理および大規模言語モデル

自然言語:日本語や英語などの日常的に使う言語

自然言語処理:自然言語をコンピュータで処理する

大規模言語モデル(Large Language Model(LLM)):

膨大な量のデータを学習し、検索された自然言語処理

タスクを実行する深層学習モデルのこと

1.1.2. 大規模言語処理例:

LLMモデルは入力文章を任意として認識し、次の文章を予測するように大規模な事前学習データを用いて、その予測の文章を出力する自然な文章生成を対応する。



図1.1.2. 出所: MiyjyGithub公式Webサイト
大規模言語処理の例

1.2.1. 機械学習システム

従来の機械学習のシステム



機械学習(ML)のシステム



図1.2.1. 出所: MiyjyGithub公式Webサイト
機械学習システムの例

1.2.2. ニューロン神経学習

入力データ:x、出力データ:y

重みパラメーター(ニューロン同士の繋がり強度):W、閾値(判断を下す基準や限界値):θ

ニューロン神経学習出力の例

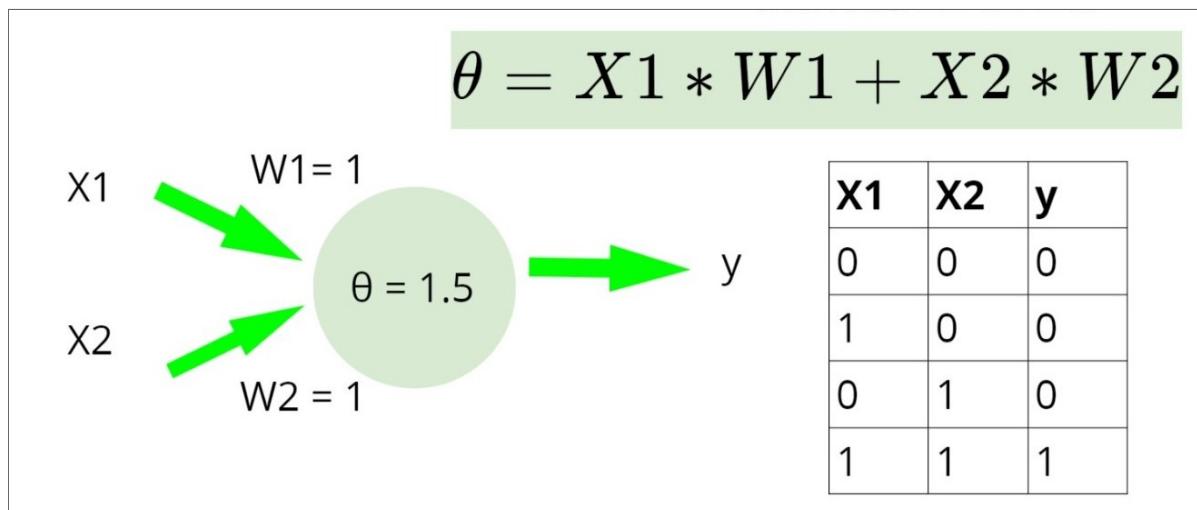


図1.2.2. 出所: MiyjyGithub公式Webサイト
ニューロン神経学習の例

1.2.3. ニューラルネットワーク

ニューラルネットワーク:人間の脳の神経回路を模した数理モデルのこと

モデル→学習→推論

入力データ:x、出力データ:y、バイアス:b、重みパラメーター:W(b,w)による誤差逆伝播)

DNNはNNより中間層が多く予測演算や誤差逆伝播法の推論実行回数がより多いため精度の高い答えが出力される。

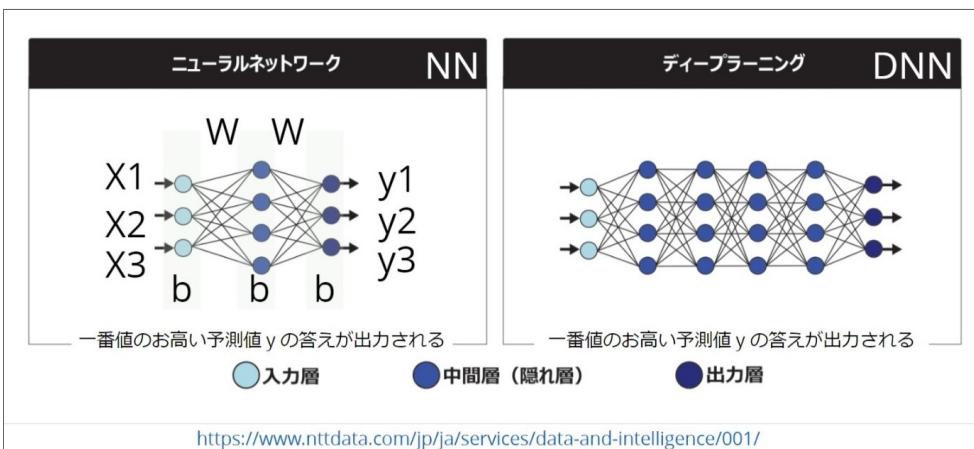


図1.2.3. 出所: NTTData公式Webサイト
ニューラルネットワークにおける学習回路の例

1.3.1. 機械学習と深層学習

機械学習(マシンラーニング)とは大量のデータを人工知能的な処理をするために推論を機械に生成させる手法(左図)

深層学習(ディープラーニング)とは脳神経細胞構造でネットワーク化した機械学習モデルによる深学習法(処理層が多い)(右図)

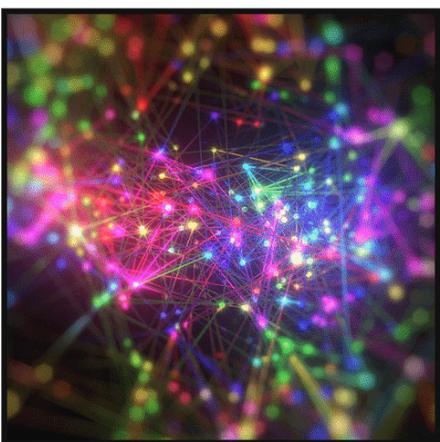


図1.3.1. 出所: giphy.com公式Webサイト | 人工知能機械学習と深層学習のイメージ

1.3.2. 人工知能・機械学習と深層学習 :

人工知能 (AI):

学習・推論・判断という知能をもつ機能を備えたコンピューターシステムで、AIの中に機械学習があり、機械学習の1手法として、ディープラーニングがある。機械学習とはデータから学習し、そこに潜むパターンを見つけ出すことができる人工知能の技術で、深層学習は、多くの層が重なった深いニューラルネットを使った機械学習技術

最先端人工知能研究開発企業OpenAIについて:

OpenAIとはアメリカの人工知能の開発を行っている企業で、人類全体に利益をもたらす汎用人工知能を普及・発展させることを目標に掲げ、AI分野の研究、開発、教育などの分野で活動し、AIの安全性、信頼性、透明性、説明可能性などの問題に対応

1.3.3. OPEN AIのホームページ



図1.3.3. 出所:OpenAI公式Webサイト | OpenAIのホームページ

1.4. 機械学習LLM(CHATGPT/GPT-4)



図1.4.1. 機械学習LLMシステムの例



図1.4.2. 機械学習LLMシステムの例

機械学習(LLM)のシステム: LLMは、大量のテキストデータを用いて学習された深層ニューラルネットワークを用いて、自然言語の処理を行う。GoogleのBERT、OpenAIのChatGPTにつかわれている。(ChatGPT/GPT-4:人工知能大規模自然言語処理モデル)

機械学習(LLM)のファインチューニングシステム: 大量のデータを投げれば自動的に解決するような創発的な特性はあるが、ファインチューニングの場合、「事実」の学習よりは「形式」の学習の方が効果的である。

1.5.自然言語処理モデルの進化過程(2015~2019)

年	モデル	パラメータ数(単位:百万)	能力
2015	RNN	3.2	入出力で系列データを扱える、中間層が再帰構造を持つ
2017	LSTM	12.5	-
2017	Attention	-	系列データの特定の部分に注意を向けるように学習させる仕組みである
2017	Transformer	85.2	系列から別の系列データに変換する、RNNを使わず、Attentionのみで構築する
2018	BERT	110.0	Transformerベースのモデルであり、事前学習とファインチューニングの2段階で効率向上させた
2020	Reformer	130.5	-
2020	XLNet	152.0	-
2020	BART	139.0	-
2020	PEGASUS	568.0	-
2020	MarianMT	88.5	-
2020	T5	220.0	Transformerベースのモデルであり、質問応答・要約・翻訳が同じモデルで解ける
2021	mBART	645.0	-

GPT-4モデルまでの学習能力一覧:(2019~2023)

年	モデル	パラメータ数(単位:百万)	能力
ChatGPT4の進化			
2019	GPT-2	150	GPT-2はTransformerベースの自己回帰モデルであり、任意の文章に続く次の単語を予測する、パラメーター数が規則と比例しているため以後パラメーター数を増やして進化した、日本語学習はrinnaモデル13億パラメータ一分が含まれている
2020	GPT-3	175	未学習のタスクの推論を行えるよになり生成言語処理が可能なGPT-3はさらにファインチューニングプロンプトプログラミングという手法まで進化した
2022	GPT-3.5	350	RLHFという強化学習によって環境や行動を選択し、報酬を最大化するようになったモデルはGPT-3.5である RLHF:Reinforcement Learning from Human Feedbackとは報酬信号が不明なときに入間がフィードバックをエージェントに支援を提供すること
2023	GPT-4	400.1	大規模言語モデルのLLMまでできるようになった、GPT-4では画像入力や生成も可能となった

表.1 出所: MiyjyGithub公式Webサイト|GPT-4になるまでの自然言語処理モデルの進化

1.5.1. CHATGPT/GPT-4のはじめ方

- 1.Get startedをクリックしメールアドレスやパスワードを入力する<https://openai.com/>
- 2.Continueをクリックしてメールアドレスを認証する
- 3.携帯電話番号を登録すると携帯にコードが送られる
- 4.コードを入力→Next→Next→Done→登録完了（ログイン画面が表示される）

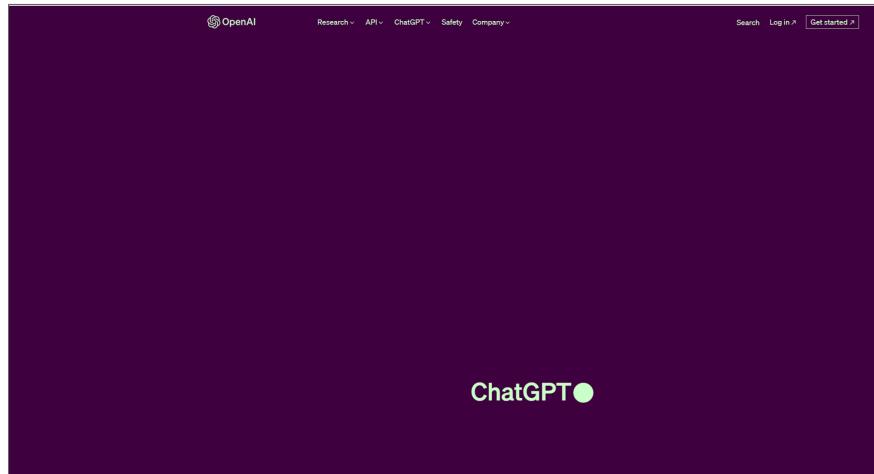


図1.5.1 出所: OpenAI公式Webサイト|ChatGPTのホームページ

1.5.2. 登録後のログイン流れ

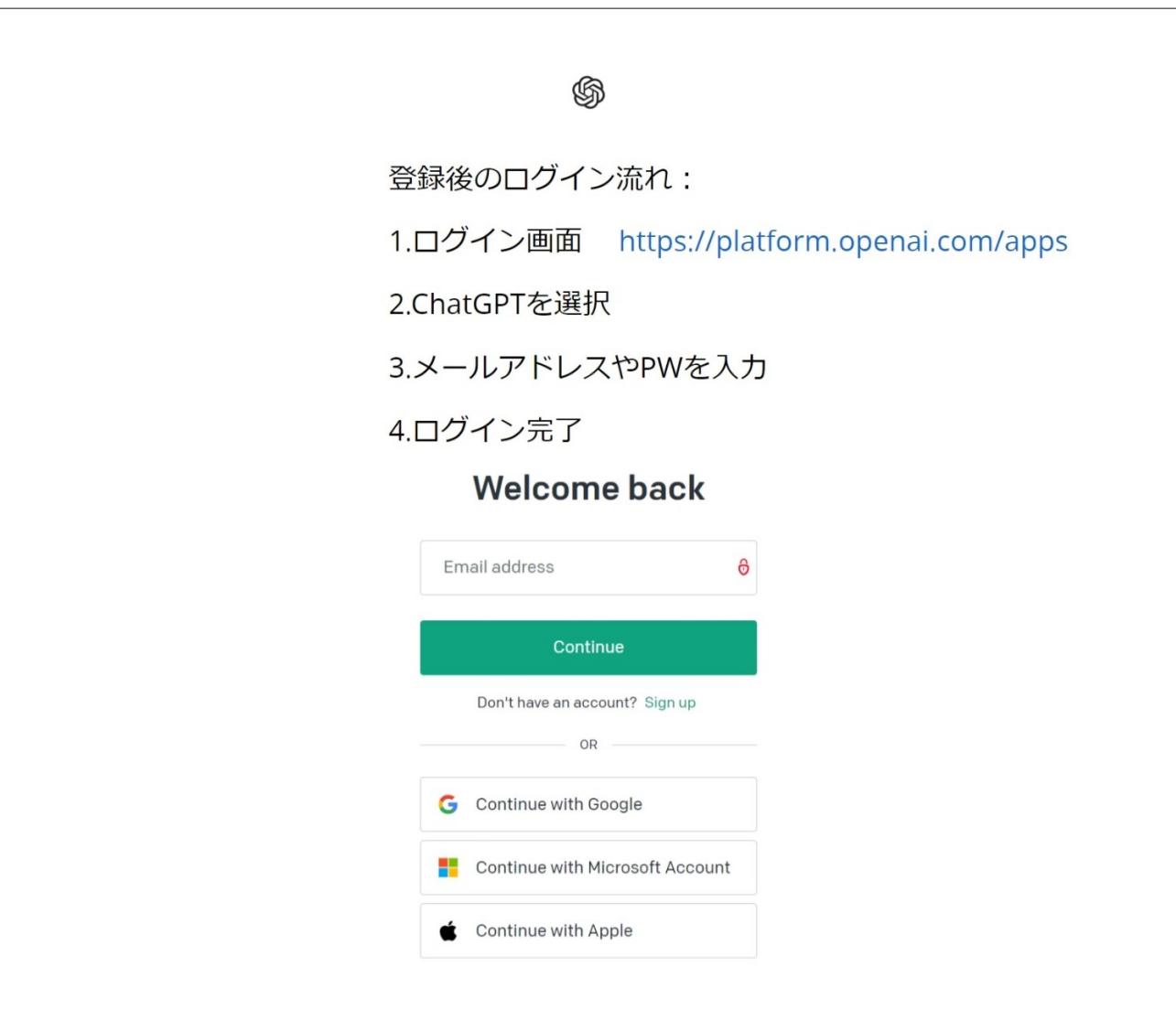


図1.5.2. 出所:OpenAI公式Webサイト|ChatGPTのアカウントログイン画面

1.5.3. CHATGPTがの画面構成

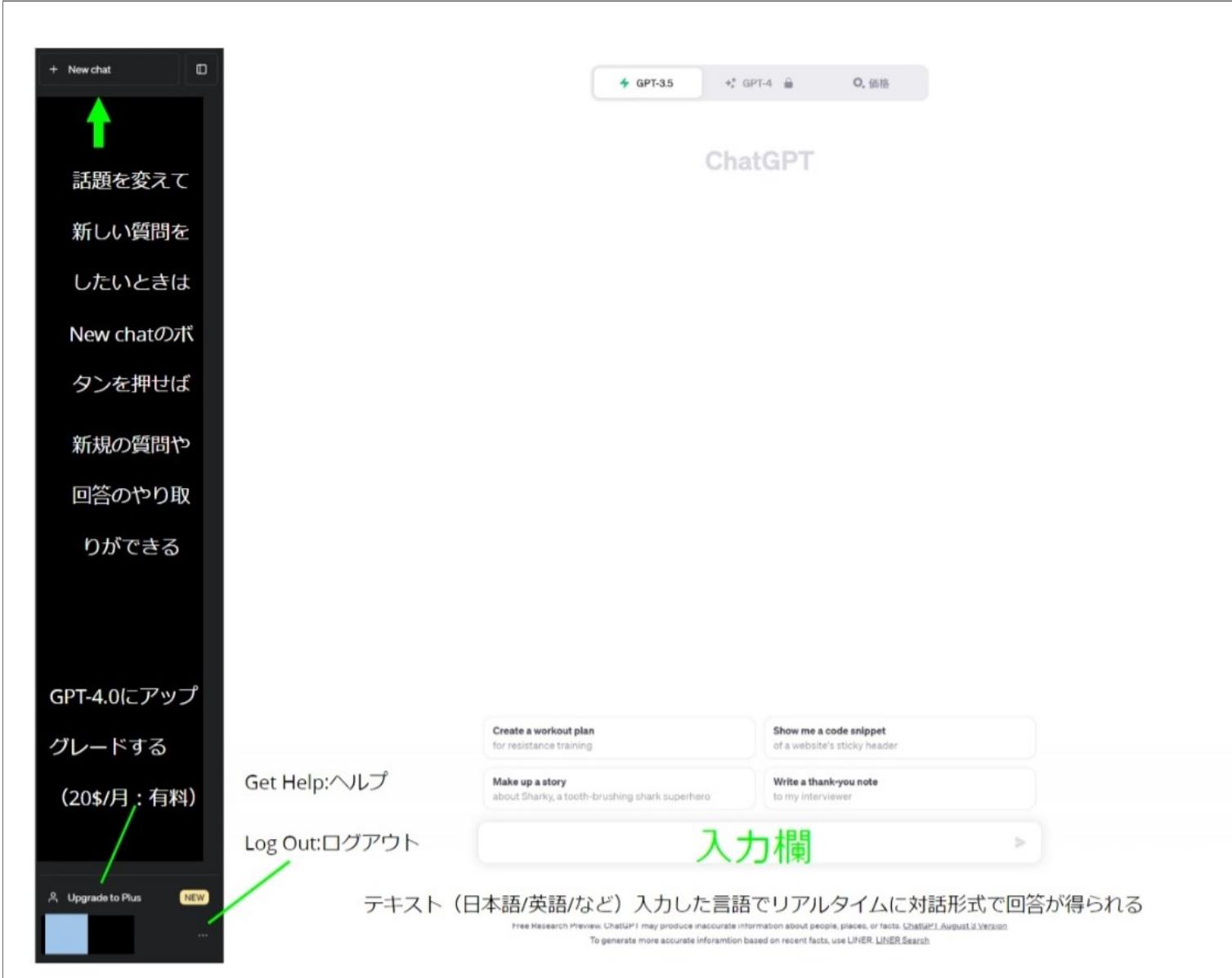


図1.5.3. 出所:OpenAI公式Webサイト|ChatGPTがの画面構成

1.5.4. 対応言語と機能

多様なタスクに対応:

プログラミング言語も 20 種類以上も対応する: Python, JavaScript, Java, C++, C#, Ruby, Swift, Go, PHP, TypeScript, Rust, Kotlin, Perl, Lua, R, MATLAB, Haskell, Scala, Julia, Shell scripting languages (e.g., Bash)

多様なタスク機能:

会話中の学習能力があるため連続でなぜなぜ質問ができる

対話、テキスト生成、質問応答、要約、翻訳、プログラミング言語対応、プログラム生成などさまざまな言語処理が実行できる

1.5.5. GPT-3.5とGPT-4の性能

性能比較表 GPT-3.5 GPT-4

	GPT-3.5	GPT-4
推論度	3	5
速度	2	2
簡潔度	1	4

注意点:学習しているデータは2021年までのものですのでそれ以後の情報は得られない。あくまでも人間の入力した検索キーワードで対応しているため、キーワードの誘導性により対話または検索結果に不正確な情報または誤った情報が含まれる可能性がある。検索キーワードをできるだけ明確にしておく必要がある。日本語による学習量が少ないため英語で質問した方が高精度な回答が得られることが多い。

<https://openai.com/research/gpt-4>

1.6. LLM大規模言語モデルおよびライブラリの紹介

1. OpenAI API(Application Programming Interface)
2. OpenAI Playground
3. DALL-EのWeb UI
4. LlamaIndex
5. LangChain

1.6.1. OPENAI APIとは

OpenAI API(Application Programming Interface)

ソフトウェアアプリ同士が相互に情報をやり取りするインターフェースで、ChatGPT以外にBing、Google Bard、などもAPIへのチャット機能組み込みサービスを提供する。テキスト生成、画像生成、埋め込み生成、ファインチューニング、モデルレーション、音声のテキスト変換機能を持つ。

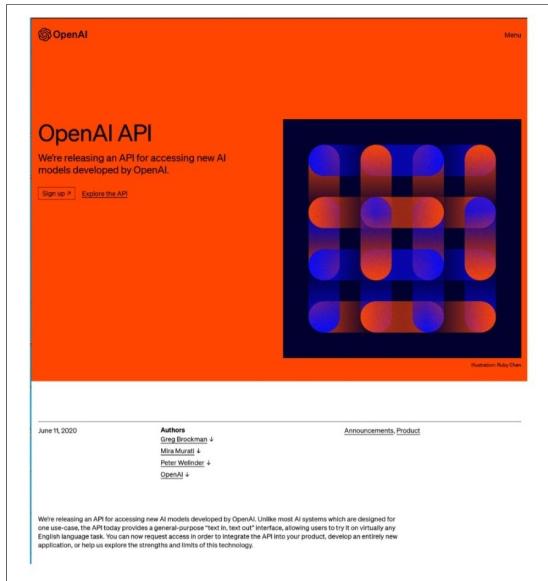


図1.6.1. 出所:OpenAI公式Webサイト |OpenAI APIのホームページ

• <https://openai.com/blog/openai-api>

1.6.2. OPENAI PLAYGROUNDとは

自然言語処理（プログラム生成、質問応答、文章生成、翻訳、要約）を実行することができるプラットフォームである。

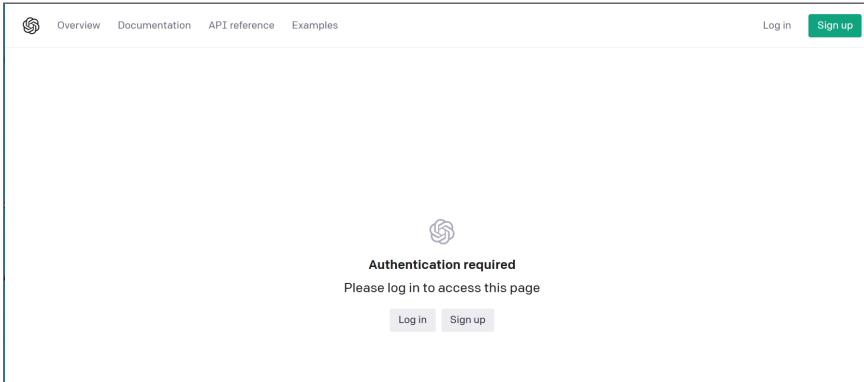


図1.6.2.1. 出所:OpenAI公式Webサイト |OpenAI Playgroundのホームページ

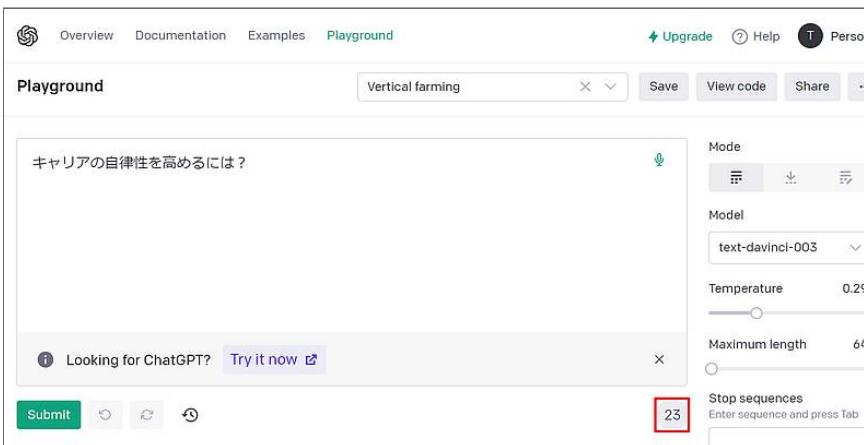


図1.6.2.2. 出所:OpenAI公式Webサイト |OpenAI Playgroundのホームページ画面構成

• <https://platform.openai.com/playground>

1.6.3. DALL-EのWEB UIとは

テキストからの画像生成や画像編集(バリエーション、インペインティング、アウトペインティング)を実行することができるプラットフォームである。



図1.6.3.1. 出所:OpenAI公式Webサイト |DALL-E 2 のホームページ

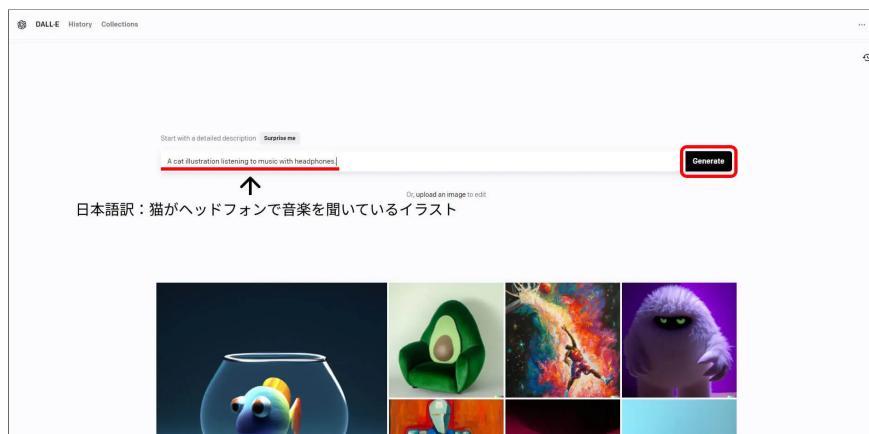


図1.6.3.2. 出所:OpenAI公式Webサイト |DALL-E 2 のホームページ画面構成

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

1.6.4. LLAMAINDEX

LLMで学習されていない情報の固まりのデータを参照し、質問応答(チャット)を作成するためのドキュメント読み込み、インデックス作成、クエリエンジン作成、質問応答関連のライブラリである。

LlamaIndex 0.8.20

Search

GETTING STARTED

- Installation and Setup
- Starter Tutorial
- High-Level Concepts
- Customization Tutorial

END-TO-END TUTORIALS

- Basic Usage Pattern
- One-Click Observability
- Principled Development Practices
- Discover LlamaIndex Video Series
- Finetuning
- Use Cases

INDEX/DATA MODULES

- Data Connectors (LlamaHub)
- Documents / Nodes
- Node Parser
- Storage
- Indexes

QUERY MODULES

- Query Engine
- Chat Engine

Welcome to LlamaIndex !

ON THIS PAGE

- Why LlamaIndex?
- Who is LlamaIndex for?
- Getting Started
- Ecosystem
- Community
- Associated projects

Why LlamaIndex?

LlamaIndex (formerly GPT Index) is a data framework for LLM applications to ingest, structure, and access private or domain-specific data.

How can LlamaIndex help?

LlamaIndex provides the following tools:

- Data connectors** ingest your existing data from their native source and format. These could be APIs, PDFs, SQL, and (much) more.
- Data indexes** structure your data in intermediate representations that are easy and performant for LLMs to consume.
- Engines** provide natural language access to your data. For example:
 - Query engines are powerful retrieval interfaces for knowledge-augmented output.
 - Chat engines are conversational interfaces for multi-message, "back and forth" interactions with your data.
- Data agents** are LLM-powered knowledge workers augmented by tools, from simple helper functions to API integrations and more.
- Application integrations** tie LlamaIndex back into the rest of your ecosystem. This could be LangChain, Flask, Docker, ChatGPT, or... anything else!

図1.6.4. 出所: META公式Webサイト |LlamaIndexのホームページ

• <https://gpt-index.readthedocs.io/en/latest/index.html#ecosystem>

1.6.5. LANGCHAIN

LLM呼び出しモジュールであり、複数のアプリを組み合わせて複雑なアプリケーションを構築するオープンソースのライブラリである。

機能例:プロンプト(入力文章)生成、複数のLLMやプロンプトの入出力を繋げる、エージェントにより処理順番を顧客要求通りに実行、エージェント特定の実行ツール、チーンやエージェントを記憶

活用例: AIキャラクターとの対話(Gatebox、StackChain、など)、ユーチューブでのAI組み込み、ロボットの制御、その他

The screenshot shows the LangChain documentation website. At the top, there's a navigation bar with links for LangChain, Concepts, Python Docs, and JS/TS Docs. Below the navigation is a sidebar with a tree-like menu structure under the 'Components' category. The categories listed are Introduction, Components (which is expanded), Schema, Models, Prompts, Indexes, Memory, Chains (which is collapsed), Chain, LLMChain, Index-related chains, Prompt Selector, Agents, Use Cases (which is collapsed), and Personal Assistants. The main content area is titled 'Components' and contains a brief introduction: 'In this section we first cover some underlying schema abstractions, before diving into the s...'. Below this, there are four cards: 'Schema' (4 items), 'Models' (3 items), 'Prompts' (4 items), and 'Indexes' (4 items). Each card has a small icon of a document with a gear.

図1.6.5. 出所: META公式Webサイト |LangChainのホームページ

- <https://docs.langchain.com/docs/category/components>

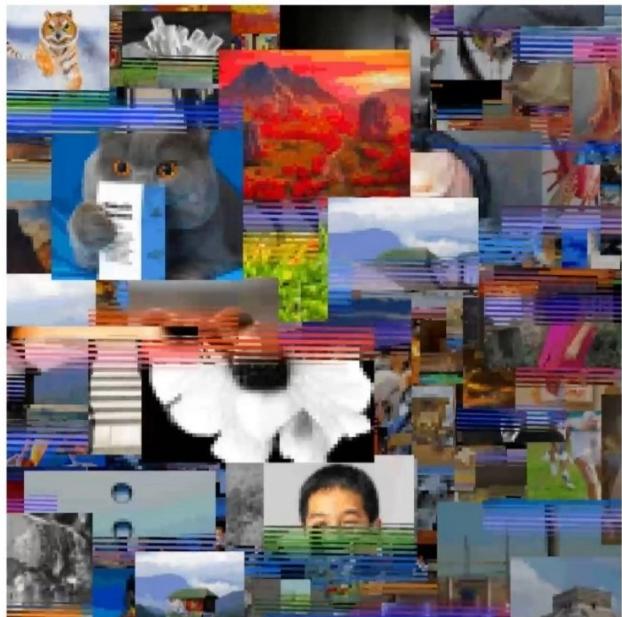
1.7. 大型自然言語処理のGPTモデル応用例

1. ImageGPT
2. CLIP(GPT-2モデル応用)
3. DALL-E(GPT-3応用)

1.7.1. IMAGEGPT

(Transformer自然言語処理モデル応用)

半分の画像を入力すると残り半分の画像が生成される



<https://openai.com/research/image-gpt>

図1.7.1. 出所:OpenAI公式Webサイト | ImageGPTの画像生成イメージ

- <https://openai.com/research/image-gpt>

1.7.2. CLIP(GPT-2モデル応用)

推論時に自由にカテゴリーを指定して分類することができる



<https://openai.com/research/clip>

図1.7.2. 出所:OpenAI公式Webサイト |CLIP(GPT-2応用)の画像分類のイメージ

• <https://openai.com/research/clip>

1.7.3. DALL-E(GPT-3応用)

テキストと画像のペアのデータセットを使用して、テキストから画像を生成する

DALLE-E2(DALLE-Eの解像度4倍)、リアルで正確な画像が生成できる

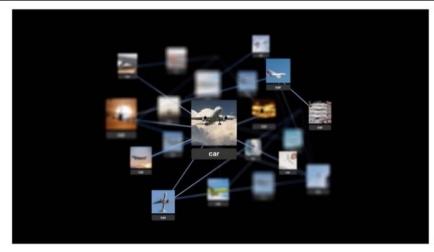


図1.7.3.1. 出所:OpenAI公式Webサイト |DALLE-E2(GPT-3応用)画像データセット例



図1.7.3.2. 出所:OpenAI公式Webサイト |DALLE-E2(GPT-3応用)生成画像の例

• <https://openai.com/dall-e-2>

1.8. その他GPT以外の特殊なモデルの紹介

1. TACOTRON2+WAVEGLOWモデルの組み合わせ
2. NEUTRINO
3. JUKEBOX
4. WHISPER

1.8.1. TACOTRON2+WAVEGLOWモデルの組み合わせ

Tacotron2はテキストをメルスペクトログラムに変換、WaveGlowはメルスペクトログラムを音声に変換する深層学習モデル

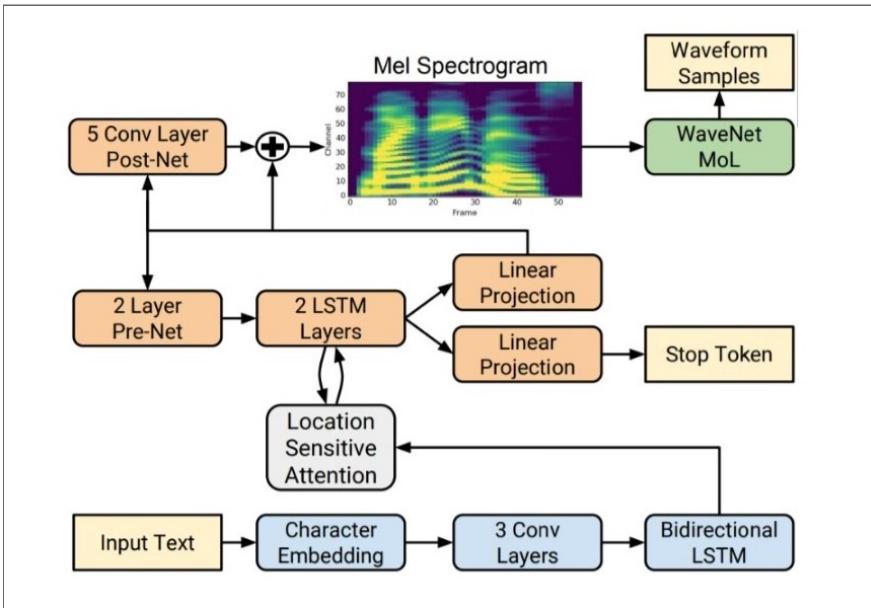
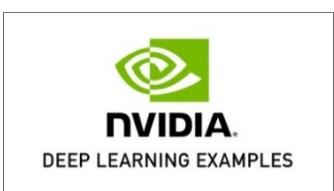


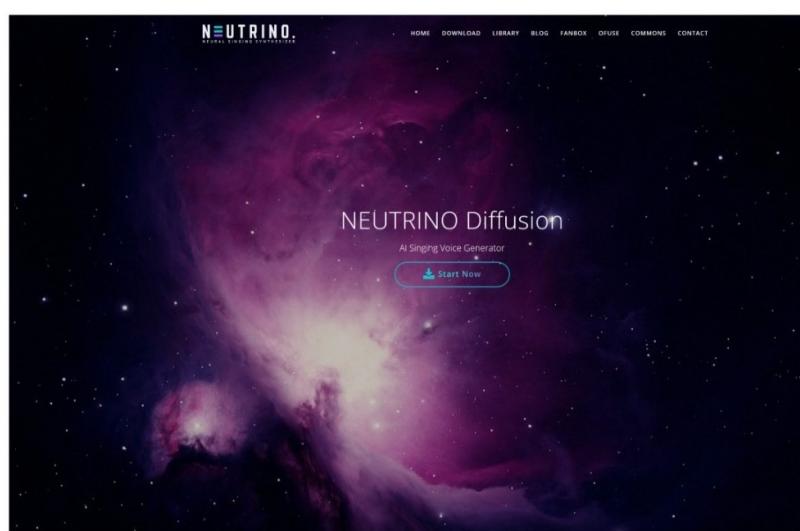
図1.8.1. 出所:NVIDIA公式Webサイト |テキストを音声に変換するTactron2モデル構造



- https://catalog.ngc.nvidia.com/orgs/nvidia/resources/tacotron_2_and_waveglow_for_pytorch

1.8.2. NEUTRINO

ニューラルネットワークを用いた歌声シンセサイザーモデルで、歌詞つきの楽譜から、発声タイミング・音の高さ・声質・声のかすれ具合などをニューラルネットワークで推論し、歌声を生成する。



<https://studio-neutrino.com/>

図1.8.2 出所:NEUTRINO公式Webサイト |NEUTRINOのホームページ

• <https://studio-neutrino.com/>

1.8.3. JUKEBOX

指定した条件で自動的に曲を生成する深層学習モデルで、アーティスト、ジャンル、歌詞、曲の長さを指定すると自動的に歌声つきの楽曲を生成する。



<https://openai.com/research/jukebox>

図1.8.3. 出所:OpenAI公式Webサイト |Jukeboxのホームページ

- <https://openai.com/research/jukebox>

1.8.4. WHISPER

会話音声をテキストに変換する深層学習モデルで、日本語を含む多言語の音声認識、音声翻訳、言語認識、音声区間検出などができる。



図1.8.4. 出所:OpenAI公式Webサイト|"Whisperのホームページ

• <https://openai.com/research/whisper>

1.9. 大規模言語モデルのAI活用例

1. MICROSOFT EDGE
2. BING
3. GATEBOX
4. STACK-CHAN
5. しづくAIユーチューバー
6. CHATGPT FOR ROBOTICS
7. GOOGLE BARD

1.9. 大規模言語ライブラリのAI活用例

1. LLAMA2
2. ARISTO (AI2 REASONING CHALLENGE ARC)
3. HELLASWAG
4. WINOGRANDE
5. MMLU
6. DROP(F1 SCORE)
7. HUMANEVAL

1.9.1. MICROSOFT EDGE

ウェブワールドクラスのパフォーマンス、組み込みのプライバシー機能などを搭載した、高速かつ安全なブラウザーである。



図1.9.1. 出所:Microsoft公式Webサイト |Microsoft Edgeのホームページ

- <https://www.microsoft.com/ja-jp/edge/welcome?form=MA13FJ>

1.9.2. BING

検索、チャット、作成がすべて1か所で行える。チャットでBingの新しいAIを利用したImage Creatorを使用すれば、言葉を画像に変えられる。



図1.9.2. 出所:Microsoft公式Webサイト |Microsoft Bingのホームページ

- <https://www.bing.com/new?form=MY028Z&OCID=MY028Z>

1.9.3. GATEBOX

キャラクター召喚装置「Gatebox」量産モデル、AIパートナー型オリジナルキャラクター「逢妻ヒカリ」、作業応援アプリ「CheerPro（チアプロ）」を提供している。



生成系AI対応の研究開発を開始

<https://www.gatebox.ai/news/20230308-chatgpt>

図1.9.3. 出所:Gatebox公式Webサイト |GateboxのAI

- <https://www.gatebox.ai/news/20230308-chatgpt>

1.9.4. STACK-CHAN

Stack-ChanはJavaScript由来のコミュニケーションロボットである。

M5Stack-embedded super-kawaii robot



<https://protopedia.net/prototype/2345>

図1.9.4. 出所:protopedia.net公式Webサイト|コミュニケーションロボット

• <https://protopedia.net/prototype/2345>

1.9.5. しづくAIユーチューバー

雑談、ASMR、歌ってみたなど、幅広いジャンルの配信を取り扱う

AI Vtuberチャンネルで、AIを応用して配信されている。



図1.9.5. 出所:しづくAIユーチューブチャンネル公式Webサイト

- <https://www.youtube.com/channel/UCE2SWbhR2WRHPBi-bflr0-g>

1.9.6. CHATGPT FOR ROBOTICSとは

Microsoftが開発したロボット制御向けにカスタマイズしたロボットアーム、ドローン、ロボット掃除機など、複数のロボットをChatGPTのように自然言語から制御可能なプラットフォームである。

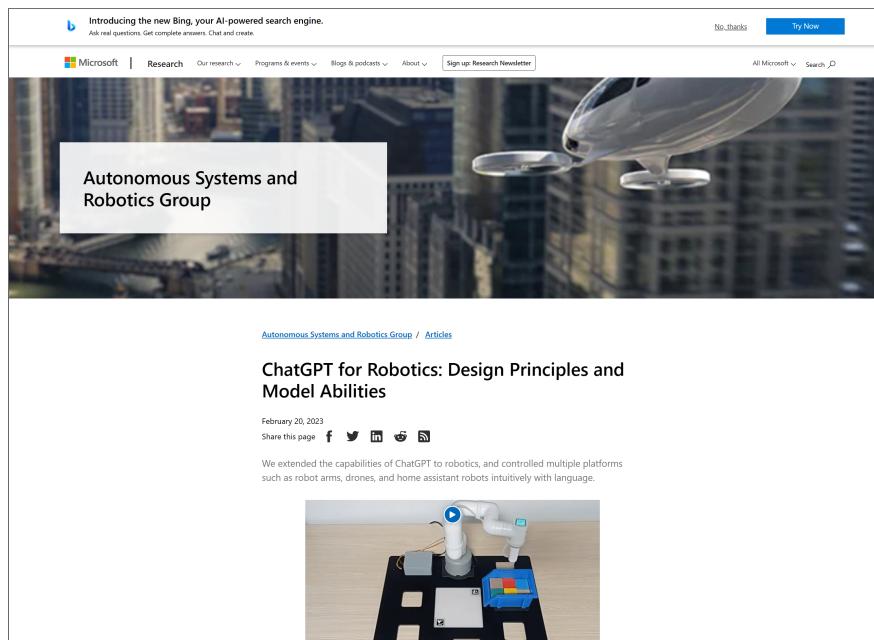


図1.9.6. 出所:Microsoft公式Webサイト

Microsoft ChatGPT for Roboticsのホームページ

• <https://www.microsoft.com/en-us/research/group/autonomous-systems-group-robotics/articles/chatgpt-for-robotics/>

1.9.7. GOOGLE BARD

グーグル Bardとは、生成系AIで、ユーザーが入力した質問に対し、違和感のない自然な文章を返すのがGoogle Bardの大きな特徴。



図1.9.7. 出所:Google Bard公式Webサイト |Google Bardのホームページ

- https://bard.google.com/?utm_source=sem&utm_medium=paid-media&utm_campaign=q3jaJP_sem6

1.9.8. Llama2

LLMAより学習量が40%アップされ、しかもパラメーター数も
より多いためコンテクストも2倍の長さまでプログラムできる。

Llama 2 was trained on 40% more data than Llama 1, and has double the context length.

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

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

図1.9.8. 出所: META公式Webサイト
METAのLlama2パラメーター仕様

• <https://ai.meta.com/llama/>

LlaMa:LLaMA(Large Language Model Meta AI)は1.4兆個のトークン、
70億パラメータから650億パラメータまで学習されたモデルである。

1.9.9. ARISTO (AI2 REASONING CHALLENGE ARC))

GPT-4応用の教育可能な推論システムによる推論と説明ができるデータセットである。メモリベースのアーキテクチャによる継続的な学習、知識と信念、普遍的な数学的推論ができる。現在の評価指数は96.3%まで達成されたとしている。

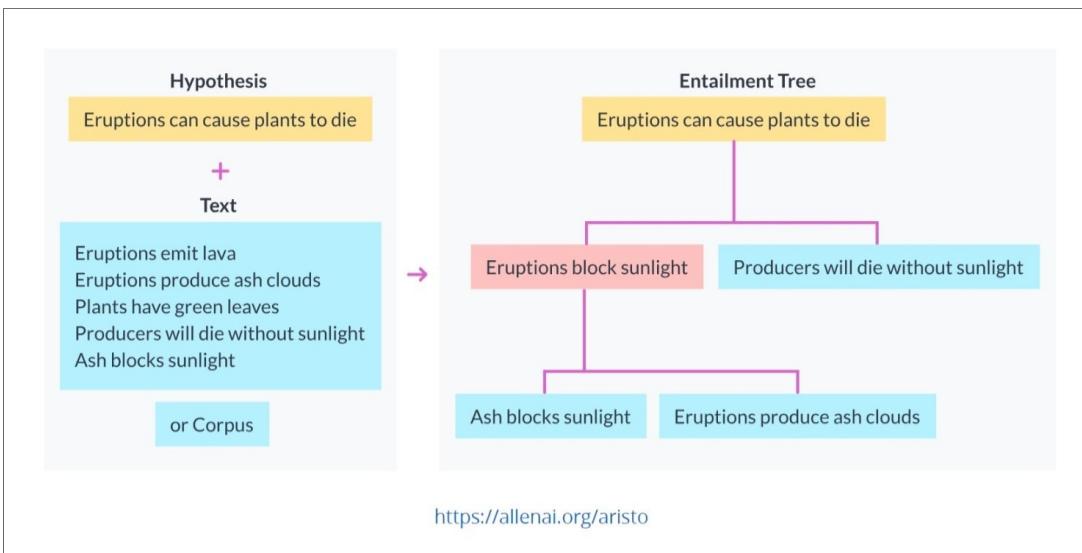
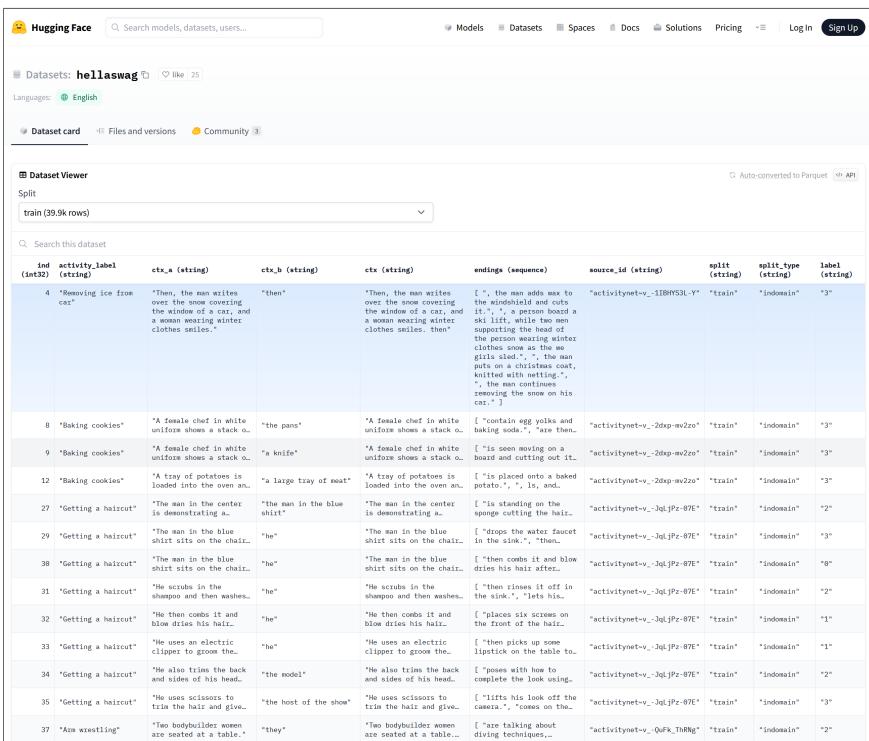


図1.9.9. 出所:allenai.org公式Webサイト |Aristoのデータ処理例

• <https://allenai.org/aristo>

1.9.10. HELASWAG

日常的な一般常識推論データセットであり、
現在の評価指数は95.3%まで達成されたとしている。



The screenshot shows the Hugging Face Dataset Viewer interface for the 'hellaswag' dataset. At the top, there's a search bar and navigation links for Models, Datasets, Spaces, Docs, Solutions, Pricing, Log In, and Sign Up. Below the header, it says 'Datasets: hellaswag' with a like count of 25. It also shows the language is English. There are tabs for Dataset card, Files and versions, and Community.

The main area is titled 'Dataset Viewer' and shows a table with columns: index (ind), activity_label (string), ctx_a (string), ctx_b (string), ctx (string), endings (sequence), source_id (string), split (string), split_type (string), and label (string). The table contains several rows of data, each representing a different scenario from the dataset. The first few rows are:

ind	activity_label	ctx_a (string)	ctx_b (string)	ctx (string)	endings (sequence)	source_id (string)	split (string)	split_type (string)	label (string)
4	"Removing ice from car"	"Then, the man writes over the snow covering the windows of a car, and a woman wearing winter clothes smiles."	"then"	"Then, the man writes over the snow covering the windows of a car, and a woman wearing winter clothes smiles. then	["the man adds wax to the windshield and cuts it with a knife.", "he uses a ski lift while two men supporting the head of a car.", "he removes clothes snow as the girls sled.", "the man makes a snowman out of snow, knitted with netting.", "the man continues removing the snow on his car."]	"activitynet-v_3IB0YSL-Y"	"train"	"indomain"	"3"
8	"Baking cookies"	"A female chef in white uniform shows a stack o...	"the pans"	"A female chef in white uniform shows a stack o...	["contain egg yolks and baking soda.", "are then...	"activitynet-v_-2dxp-mvZzo"	"train"	"indomain"	"3"
9	"Baking cookies"	"A female chef in white uniform shows a stack o...	"a knife"	"A female chef in white uniform shows a stack o...	["is seen moving on a board and cutting out it...	"activitynet-v_-2dxp-mvZzo"	"train"	"indomain"	"3"
12	"Baking cookies"	"A tray of potatoes is loaded into the oven an...	"a large tray of meat"	"A tray of potatoes is loaded into the oven an...	["is placed onto a baked potato.", "is, and,	"activitynet-v_-2dxp-mvZzo"	"train"	"indomain"	"3"

図1.9.10. 出所:HuggingFace公式Webサイト |HellaSwagのデータセット例

• <https://huggingface.co/datasets/hellaswag/viewer/default/train?row=2>

1.9.11. WINOGRANDE

代名詞照応解析による推論規則が有効に扱えていない原因を整理し常識的な推論システムを目指し、GPT-4モデルが応用されているデータセットである。現在の評価指数は87.5%まで達成されたとしている。

Twin sentences	Options (answer)
x The monkey loved to play with the balls but ignored the blocks because he found them <i>exciting</i> .	balls / blocks
The monkey loved to play with the balls but ignored the blocks because he found them <i>dull</i> .	balls / blocks
x William could only climb beginner walls while Jason climbed advanced ones because he was very <i>weak</i> .	William / Jason
William could only climb beginner walls while Jason climbed advanced ones because he was very <i>strong</i> .	William / Jason
✓ Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had <i>less</i> time to get ready for school.	Robert / Samuel
Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had <i>more</i> time to get ready for school.	Robert / Samuel
✓ The child was screaming after the baby bottle and toy fell. Since the child was <i>hungry</i> , it stopped his crying.	baby bottle / toy
The child was screaming after the baby bottle and toy fell. Since the child was <i>full</i> , it stopped his crying.	baby bottle / toy

WinoGrande: <https://mosaic.allenai.org/projects/winogrande>

図1.9.11. 出所:allenai.org公式Webサイト |winograndeのデータセット例

- <https://mosaic.allenai.org/projects/winogrande>

1.9.12. MMLU

(Measuring Massive Multitask Language Understanding)

大量のマルチタスク(マルチ質問応答選択)の言語理解度の測定モデルである。

現在の評価指数は86.4%まで達成されたとしている。

Measuring Massive Multitask Language Understanding							
This is the repository for Measuring Massive Multitask Language Understanding by Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Marisa Mazella, Dawn Song, and Jacob Steinhardt (ICLR 2021).							
Model	Authors	Humanities	Social Sciences	STEM	Other	Average	
Chinchilla (70B, few-shot)	Hoffmann et al., 2022	63.6	79.3	54.9	73.9	67.5	
Gopher (260B, few-shot)	Rae et al., 2021	56.2	71.9	47.4	66.1	60.0	
GPT-3 (175B, fine-tuned)	Brown et al., 2020	52.5	63.9	41.4	57.9	53.9	
flan-T5-xl	Chung et al., 2022	46.3	57.7	39.0	55.1	49.3	
UnifiedQA	Khashabi et al., 2020	45.6	56.6	40.2	54.6	48.9	
GPT-3 (175B, few-shot)	Brown et al., 2020	40.8	50.4	36.7	48.8	43.9	
GPT-3 (6.7B, fine-tuned)	Brown et al., 2020	42.1	49.2	35.1	46.9	43.2	
flan-T5-large	Chung et al., 2022	39.1	49.1	33.2	47.4	41.9	
flan-T5-base	Chung et al., 2022	34.0	38.1	27.6	37.0	34.2	
GPT-2	Radford et al., 2019	32.8	33.3	30.2	33.1	32.4	
flan-T5-small	Chung et al., 2022	29.9	30.9	27.5	29.7	29.5	
Random Baseline	N/A	25.0	25.0	25.0	25.0	25.0	

If you want to have your model added to the leaderboard, please reach out to us or submit a pull request.

Results of the test:

Citation

If you find this useful in your research, please consider citing the test and also the ETHICS dataset it draws from:

@article{hendrycks2021, title={Measuring Massive Multitask Language Understanding}, author={Dan Hendrycks and Collin Burns and Steven Basart and Andy Zou and Marisa Mazella and Dawn Song and Jacob Steinhardt}, journal={Proceedings of the International Conference on Learning Representations (ICLR)},}

図1.9.12 出所:hendrycks公式Webサイト |MMLUモデルアクセス画面

- <https://github.com/hendrycks/test>

1.9.13. DROP(F1 SCORE)

GPT-4応用の読解と算数の推論データセットである。

現在の評価指数は80.9%まで達成されたとしている。

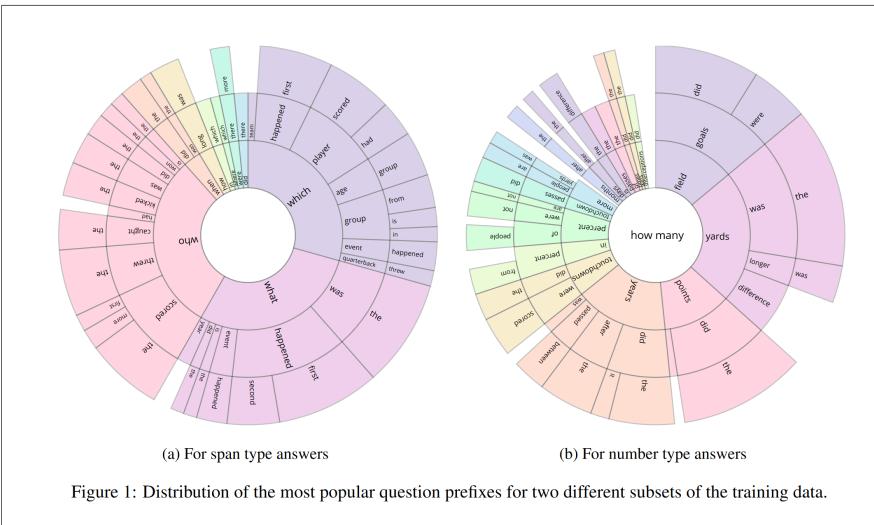


図1.9.13. 出所:aclanthology.org公式Webサイト|DROPのデータセット例

• <https://aclanthology.org/N19-1246.pdf>

1.9.14. HUMANEVAL

OpenAIによってリリースされたHumanEvalデータセットには、関数シグネチャー、docstring、本体、およびいくつかの単体テストを含む164のプログラミング問(Python coding tasks)が含まれている。これらはコード生成によるトレーニングセットではなく、手書きで作成されているデータセットである。

現在の評価指標は67.0%まで達成されたとしている。

```
1 from datasets import load_dataset
2 load_dataset("openai_humaneval")
3
4 DatasetDict({
5     test: Dataset({
6         features: ['task_id', 'prompt', 'canonical_solution'],
7         num_rows: 164
8     })
9 })
10
11 {
12     "task_id": "test/0",
13     "prompt": "def return1():\n",
14     "canonical_solution": "    return 1",
15     "test": "def check(candidate):\n        assert
```

https://huggingface.co/datasets/openai_humaneval

図1.9.14. 出所:HumanEval公式Webサイト | HumanEvalデータセットコード例

• https://huggingface.co/datasets/openai_humaneval

Speaker notes