

2024年7月2日（火）11:00～16:15

【主催】岐阜県DX推進コンソーシアム

【共催】岐阜県図書館

【運営】株式会社タノキバ

ハンズオンで学ぶ生成AI活用

～クラウド実行環境とノーコードAI開発環境の活用～

【講師】

株式会社リーサ代表 高城 勝信

本日のスケジュール

時間	内容
11:00 – 11:10	事務局挨拶、図書館からのお知らせ
11:10 – 12:00	午前の部 1. 大規模言語モデルの基礎と応用
12:00 – 13:00	お昼休憩
13:00 – 13:50	午後の部① 2. 大規模言語モデルの利用
13:50 – 14:00	休憩
14:00 – 14:50	午後の部② 3. ハンズオン (1/2)
14:50 – 15:00	休憩
15:00 – 16:00	午後の部② 3. ハンズオン (2/2) 製造分野での活用を考える、まとめ
16:00 – 16:15	質疑応答、アンケート記入

自己紹介

大垣北高校、京都大学情報学研究科卒業

2000年 日本IBM入社、グローバルサービス部門にて
先端技術やスマートシティ関連プロジェクト担当
ブロックチェーン技術のエバンジェリスト

2017年 日本IBM退職
株式会社リーサ設立、**株式会社インテック技術顧問**

2018年 愛知産業大学スマートデザイン学科准教授

2020年 **株式会社ブリスコラ取締役**

2022年 愛産大退職、**大垣帰郷、上石津在住**

2023年 **ソフトピアジャパンWORKSHOP24**に入居
生成AIワーキンググループ設立

2024年 **株式会社セリア取締役（監査等委員）就任**
株式会社リーサにて**生成AIソリューション「AMATE」開発中**

マイブーム：

生成AI、空海、論語とそろばん、朝のお掃除、畑の手伝い、…



アジェンダ (1/2)

午前の部

1. 大規模言語モデルの基礎と応用

1. 生成AIとは？
2. 生成AIと大規模言語モデル
3. 大規模言語モデルにできること
4. 大規模言語モデルの進化
5. 心の理論 デモ
6. 常識的な知識 デモ
7. 主要な大規模言語モデル6選
8. 大規模言語モデル3強 デモ
9. 大規模言語モデルの比較
10. 応用領域
11. 財務分析 (B/S分析) デモ
12. 売上データ分析 デモ
13. システム開発 デモ



ハンズオン以外にもデモをふんだんに取り入れています。
左のマークがあるところでデモをします。

午後の部①

2. 大規模言語モデルの利用

1. 利用方法
2. オフィシャルアプリの利用
3. 独自機能を持つアプリの利用 デモ
4. API活用アプリを開発して利用 デモ
5. ノーコードアプリ開発環境利用
6. Web APIとは？
7. LLM APIの概要 デモ
8. LLM APIを使うとなぜ会話ができる？

アジェンダ (2/2)



ハンズオン以外にもデモをふんだんに取り入れています。
左のマークがあるところでデモをします。

午後の部②

3. ハンズオン

1. 大規模言語モデルの課題と解決策
2. RAGのしくみ
3. ReACTのしくみ
4. GraphRAGのしくみ
5. ハンズオンの流れ
6. Google Colabの実行
7. Google Colab上のWebアプリへの接続
8. OSSのAIプラットフォームDify
9. Difyのしくみ

製造分野での活用を考える
まとめ

午前の部

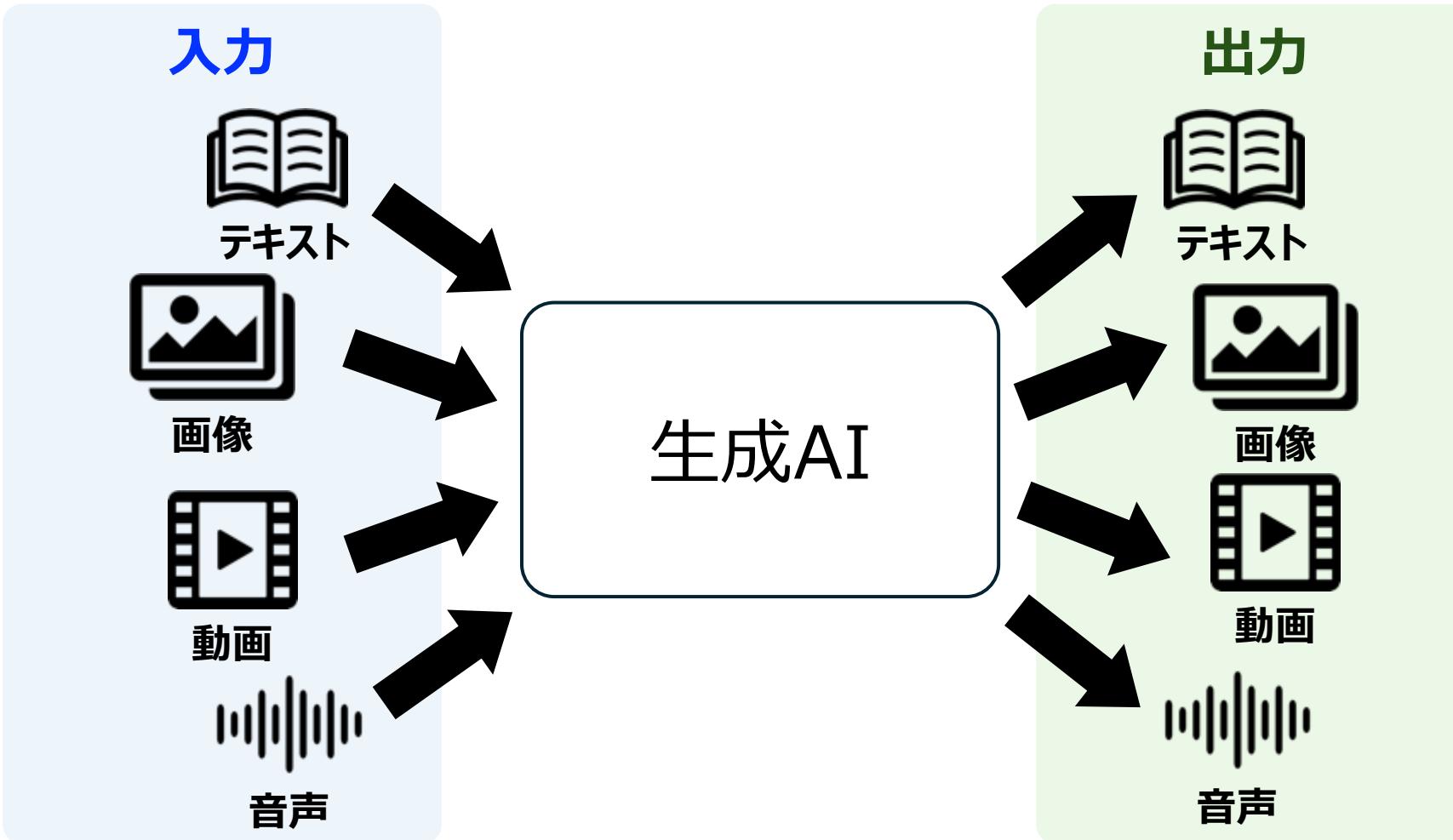
1. 大規模言語モデルの 基礎と応用

1. 大規模言語モデルの基礎と応用

1-1. 生成AIとは？

テキストや画像、動画、音声といったマルチモーダルな入力値の指示に従い、

テキストや画像、動画、音声といったマルチモーダルでクリエイティブな出力をする人工知能



1. 大規模言語モデルの基礎と応用

1-2. 生成AIと大規模言語モデル

生成AI

定義：入力データに基づいてテキスト、画像、音楽、動画などを生成する技術

事例：画像生成（OpenAI社のDall-E 2、Adobe社のFireflyなど）

音楽生成（Suno AIなど）

動画生成（OpenAI社のSora、Lume AI社のDream Machineなど）

技術：GANs（Generative Adversarial Networks）、

VAEs（Variational Autoencoders）などのアルゴリズムを利用

大規模言語モデル（LLM）

定義：主にテキストデータを基に学習され大量のパラメータを持つモデル

最近ではマルチモーダル化が進み、様々なコンテンツ生成が可能

事例：OpenAI社のGPT-4、Google社のGemini、Anthropic社のClaudeなど

技術：主にトランスフォーマーアルゴリズムを利用して開発される

最近両者の違いは無くなってきており、AIツールの文脈では「生成AI」、AIモデルの文脈では場合「大規模言語モデル(LLM)」と呼ぶ感じ（所感）

1. 大規模言語モデルの基礎と応用

1-3. 大規模言語モデルにできること

- 大規模言語モデルはテキスト要約や情報抽出・分類、会話や質疑応答、プログラミングなどが得意だった
- 最近はマルチモーダル化し、画像・イラスト・図表・動画の認識と生成や音声対話が可能になった

「マルチモーダル」

画像や動画の認識と生成



イラストや図表などの画像や動画がどのようなものかを認識したり、要約したり、プロンプトの指示に従って生成を行う。

音声対話



テキストでの質疑応答や会話だけでなく音声を使ったリアルタイムチャットができる。さらにGPT-4oは、感情を持った声色で空気を呼んで会話できる。



テキスト要約

記事を読みやすい形に要約する。例えば、プロンプトに要約したい文をコピーし「上記を1文で説明ください」と入力することで要約を実行



情報抽出

例えば複雑な文章と一緒にそれに関する質問を入力することで適切な情報を抽出することができる。



テキスト分類

例を提示し入力したテキストを分類することができる。例をわかりやすく提示することで、モデルが正確な形式でラベルを返すことができる。



質疑応答

大規模言語モデルが知っている知識や、入力した情報の中から、今までのQ&Aの文脈や具体的な指示に従って、質問に答えることができる。



会話

カスタマーサービスのチャットボットなどの会話システムを実現するなどロールを指定して会話させることができる。



プログラミング

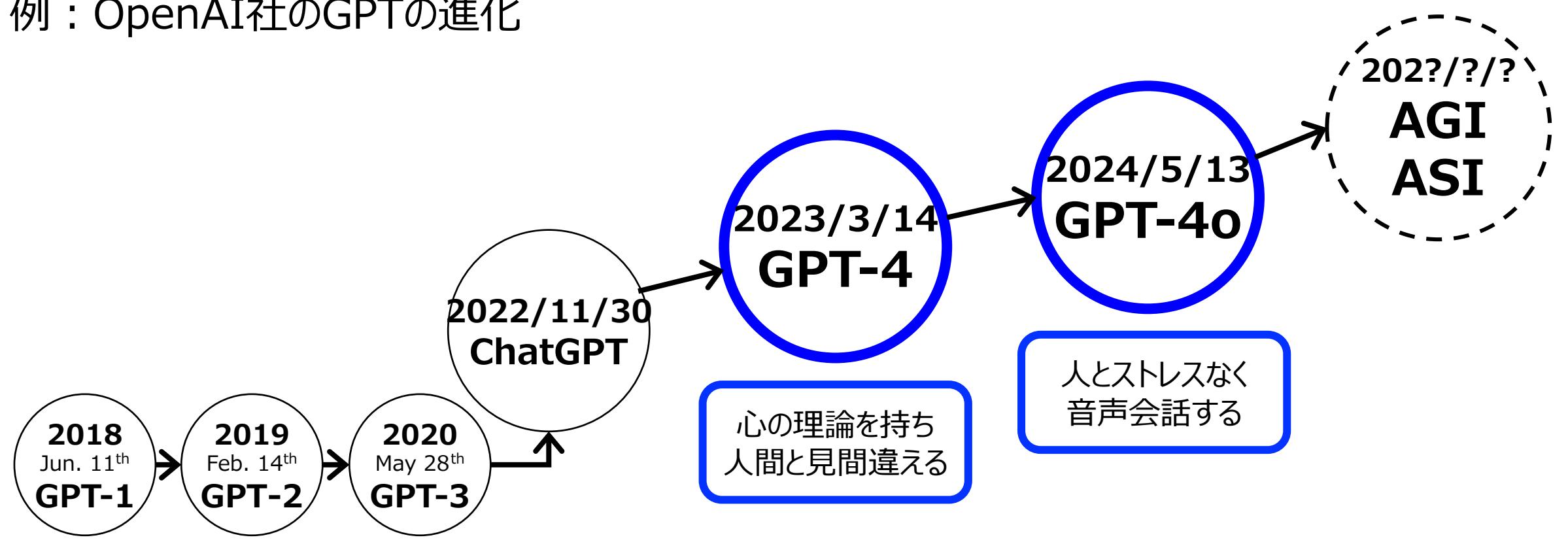
要件文からプログラムコードを生成することができる。プログラミング言語を指定することで様々なプログラミング言語のコード生成が可能。

1. 大規模言語モデルの基礎と応用

1-4. 大規模言語モデルの進化

- 2023年3月に**GPT-4**が爆誕、文章作成、要約、質疑応答、翻訳、プログラミングなどが可能に。
- GPT-3は小学生レベルだったが、GPT-4は高校生レベル、GPT-4oは大学生レベル。
- 最近は、人類の叡智を超えた**AGI(汎用人工知能)** や**ASI(超知能人工知能)**の実現が期待される。

例：OpenAI社のGPTの進化



1. 大規模言語モデルの基礎と応用

1-5. 心の理論

心の理論



- 他者の心の状態（目的、意図、知識、信念、疑念、推測）を直感的に察知する能力
- 発達心理学や認知心理学、脳科学、哲学、サル学等の研究対象
- 人間の社会的相互作用、コミュニケーション、共感、自己意識、道徳的判断、宗教的信念において重要。自閉症の子はこれができないため、コミュニケーションに支障をきたすことがある

誤信念課題



- 子供が心の理論を持つと言えるためには、他人の考え方や感情を理解し、それを基にその人がどのように真偽を判断し自分の信念を持つのか（誤信念という）を把握することが必要
- 心の理論の有無を調べるために、次ページ動画にあるような質問とその答えを用意して調べができる。つまり、「心の理論」の有無は「誤信念課題」を使ってテストすることができる。

→ はたして大規模言語モデルは心の理論をもつのか？

参考) 子どもの発達「心の理論」(5分)



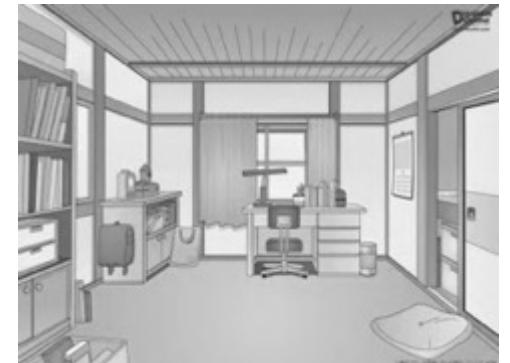
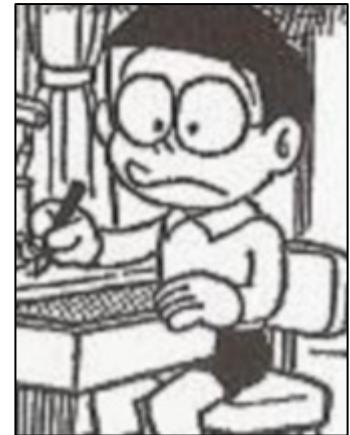
安田女子大学心理学部

デモ

ドラえもんはおやつの時間に食べるため、どら焼きを買ってのび太くんの部屋に帰ってきました。

するとのび太くんのママがドラちゃんにお願いがあると下の部屋からドラえもんを呼んだので、ドラえもんはどら焼きを押入れの中に隠して降りて行きました。のび太くんはそれを横目でみていて、ドラえもんが押し入れに隠したどら焼きを取り出して自分の机の引き出しに入れるイタズラをしました。

するとしばらくしてドラえもんが部屋に帰ってきました。おやつの時間です。ドラえもんはまっさきにどこを探すでしょう？



1. 大規模言語モデルの基礎と応用

1-6. 常識的な知識

～ウィノグラード・スキーマ・チャレンジ～

- チューリングテストをより具体的にしたAIの常識的な知識獲得を評価する質問解答の評価セット
- 立場の切り替えと正しい回答を得る能力は、複数のステークホルダーが存在する際に相手を正しく理解し、状況がどのように推移するかを予測する際に重要
- ウィノグラード・スキーマ・チャレンジが解けることは、それらの複雑な課題解決にAIが機能することを示す

https://github.com/david3080/favapp/blob/main/docs/md/2_STUDY/2-1_WS.md

Q1-1. 暴動を恐れていたので市議会はデモ隊に許可を与えなかった。暴動を恐れていたのは誰か？

A1-1. 市議会

Q1-2. 暴動を呼びかけていたので市議会はデモ隊に許可を与えなかった。暴動を呼びかけていたのは誰か？

A1-2. デモ隊

Q2-1. あまりに小さいので、トロフィーが茶色いスーツケースに収まらなかった。何があまりに小さいか？

A2-1. スーツケース

Q2-2. あまりに大きいので、トロフィーが茶色いスーツケースに収まらなかった。何があまりに大きいか？

A2-2. トロフィー

...

デモ

1. 大規模言語モデルの基礎と応用

1-7. 主要な大規模言語モデル6選

- OpenAIのGPT-4とGoogleのGeminiとANTHROPICのClaudeが3強
- OSSのモデルも急速に進化

	開発ベンダー	ブランド名	モデル	特徴
1	OpenAI	GPT	gpt-4o, gpt-4, gpt4--turbo, gpt-3.5, gpt-3.5-turboなど	自然言語処理に優れ、GPT-3、GPT-4などの高性能モデルを提供。多様なタスクに対応。
2	Google	Gemini	gemini-1.5-pro 1M/2M, gemini-pro-vision, gemini-proなど	マルチモーダル対応で、テキスト、画像、動画を処理可能。多言語対応も強化。
3	ANTHROPIC	Claude	claude-3-opus, claude-3.5-sonnet, claude-3-haikuなど	安全性と有害性の低減を重視し、強力な推論能力を持つ。多様なタスクに対応。
4	cohere	command	command-r, command-r-plusなど	企業向けにカスタマイズ可能なLLMを提供。セキュリティとプライバシーに配慮。
5	Meta	Llama	Llama-3-70b-8192, Llama-3-8b-8192など	OSSで公開されたLLM。高性能かつ効率的なモデルが特徴。groq [※] で高速実行可能。
6	Mistral AI	Mixtral	Mixtral-8x7b-Instruct-v0.1など	フランス発スタートアップ。OSS LLMを提供し、企業向けカスタマイズ可。groq [※] で高速実行可能。

※ groqは大規模言語モデルを高速実行する言語プロセッシングユニット（LPU）推論エンジンを開発する企業。

チップからノード、クラウドまでをフルセットで提供する。groqクラウドではオープンソースLLMを通常の数倍高速に処理する。

1. 大規模言語モデルの基礎と応用

1-8. 大規模言語モデル 3 強

①OpenAI社のChatGPT

- ・ 有料プランでは自分用にChatGPTをカスタマイズ可能でGPTsで公開・販売できる
- ・ 最新モデルGPT-4oのアルファ版※は、感情や非言語的な合図を理解して応答できる
※最新の高度な音声モードがアルファ版としてリリース予定だったが1ヶ月延期された
<https://twitter.com/OpenAI/status/1805716393524183136>



②Google社のGemini

- ・ 有料プランでは最高性能のAIモデルと限定機能を搭載した Gemini Advancedが利用できる
- ・ 長大なトークンや画像や動画を扱うのが得意



③Anthropic社のClaude

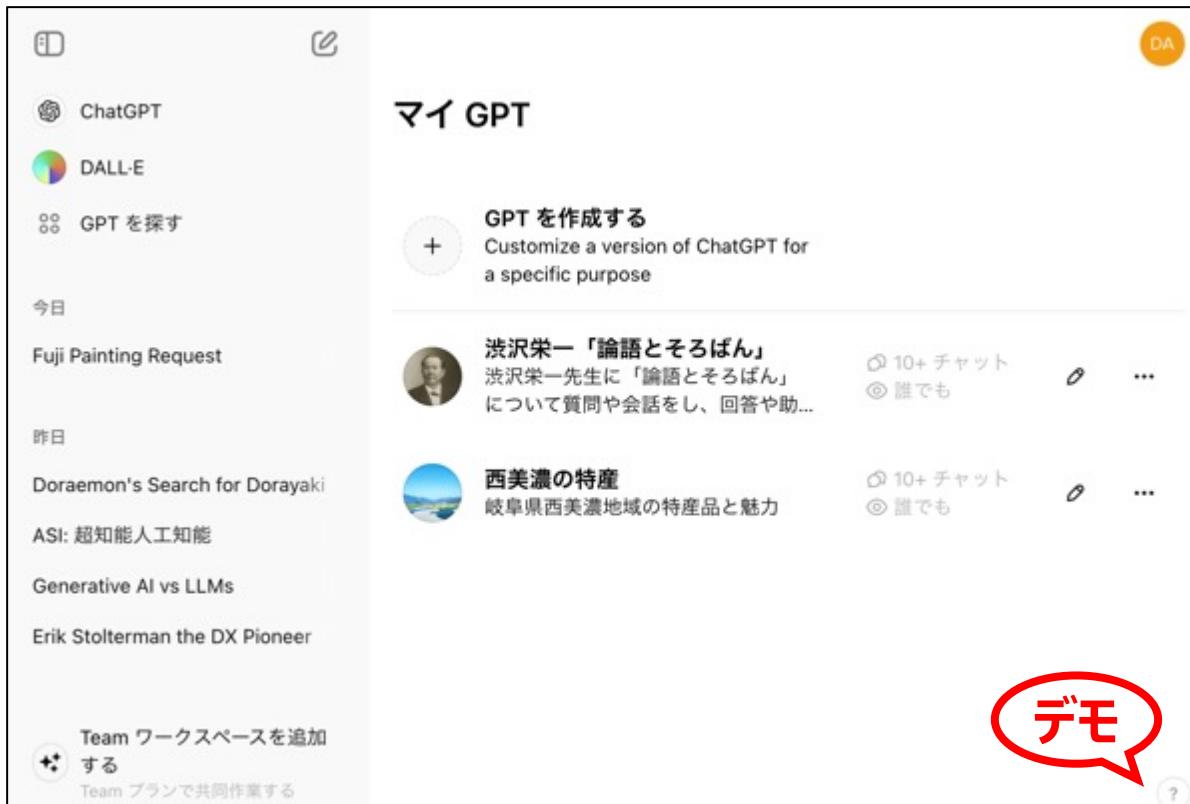
- ・ プロンプト指示に従い画面右側にプログラムと実行結果を表示できるArtifacts機能と、プロンプトやPDFなどのナレッジを事前登録し利用できるProjects機能を持つのが特徴



1. 大規模言語モデルの基礎と応用

①OpenAI社のChatGPT

OpenAI社が提供するオフィシャルなアプリにはWebアプリとスマホアプリ、**PC用ネイティブアプリ**が提供され無料から最上位モデルGPT-4oを限定的に利用することができるが、**月額20ドルのPlusプラン**に契約すると自分用に**ChatGPTをカスタマイズ**でき、**マーケットプレースであるGPTsで公開・販売**することができる。



GPTをカスタマイズしたマイGPTとマーケットプレースGPTs



PC版ネイティブアプリ

1. 大規模言語モデルの基礎と応用

参考) 人とストレスなく音声会話するGPT-4o



1. 大規模言語モデルの基礎と応用

②Google社Gemini

- Google社が提供するオフィシャルアプリには、Webアプリとスマートフォンアプリが提供され無料から利用できる。
- **月額2,900円を支払いGoogle One AIプレミアムプラン**に登録するとGoogleの最高性能のAIモデルと限定機能を搭載したGemini Advancedが利用できる。
- **長大なトークンや画像や動画を扱うのが得意**で、
例えば、**公開YouTube動画へのリンクとともに要約**を依頼すると端的に要約してくれる。



The screenshot shows the Gemini AI interface. A user has pasted a YouTube link: <https://www.youtube.com/watch?v=fw9oyUznaSg>. The interface then generates a summary of the video content, mentioning the speaker's name and the topic of AI generation. A red circle highlights a button labeled 'デモ' (Demo) in a speech bubble at the bottom right.

Gemini

https://www.youtube.com/watch?v=fw9oyUznaSg こちらの動画を要約してください

回答案を再生生成

YouTube

この動画は、リーサという地元企業のCEOである高木氏によるジェネレーティブAIに関するセミナーについてです。高木氏は自己紹介を行い、ジェネレーティブAIとは何か、どのように機能するかを説明します。また、AIリテラシーと、ビジネスにおけるAI使用における倫理的考慮事項の重要性についても語っています。

主なポイントは次のとおりです。

- ・ジェネレーティブAIは、テキスト、画像、音楽などの新しいコンテンツを作成するために使用できるAIの一種です。
- ・高木氏は、ジェネレーティブAIはあらゆる規模の企業が利用できる強力なツールであると考えています。

ここにプロンプトを入力してください

Geminiは不正確な情報（人物に関する情報など）を表示することがあるため、生成された回答を再確認するようにしてください。

プライバシーと Gemini アプリ

デモ

1. 大規模言語モデルの基礎と応用

③Anthropic社Claude

Anthropic社提供のClaude.AI Webアプリは以下の特徴的な機能を持つ（月額20ドルのProプラン）

1. Artifacts機能

プロンプトによる指示に従い、画面右側にプログラムやその実行結果をリアルタイムに表示することができる。

下図は商品在庫管理のER図を作成した例。

Claude 商品在庫管理ER図

MT 商品在庫管理のER図を作ってください

はい、商品在庫管理のER図を作成いたします。

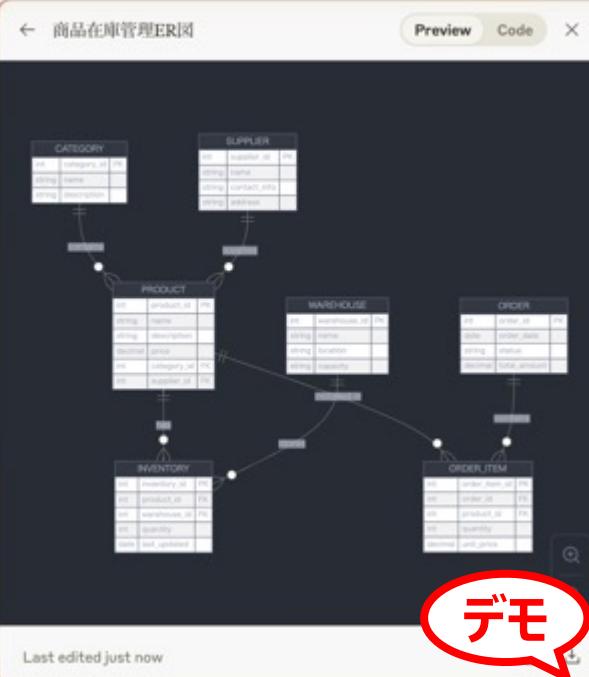
商品在庫管理ER図
Click to open diagram

この商品在庫管理のER図には以下の主要なエンティティと関係性が含まれています：

1. PRODUCT (商品) : 各商品の情報を保持します。
2. CATEGORY (カテゴリ) : 商品のカテゴリを管理します。
3. SUPPLIER (仕入先) : 商品の仕入先情報を管理します。
4. INVENTORY (在庫) : 各倉庫における商品の在庫数を管理します。

Reply to Claude...

Claude 3.5 Sonnet □ 1

 デモ

Last edited just now

2. Projects機能

事前にプロンプトやテキストやPDFなどのナレッジを登録し、それに従い回答するプロジェクトを作成することができる。

Claude ← All projects

西美濃の特産 Private

岐阜県西美濃地域の特産品と魅力

How can Claude help you today?

Claude 3.5 Sonnet □ 1 Large knowledge size may result in slower responses

Project knowledge Add Content

93% of knowledge size used

このGPTは、岐阜県西美濃地域の専門家... Edit

nishimino PDF 今日 Very large file

MT 西美濃地域の農林水産物について教えてください。

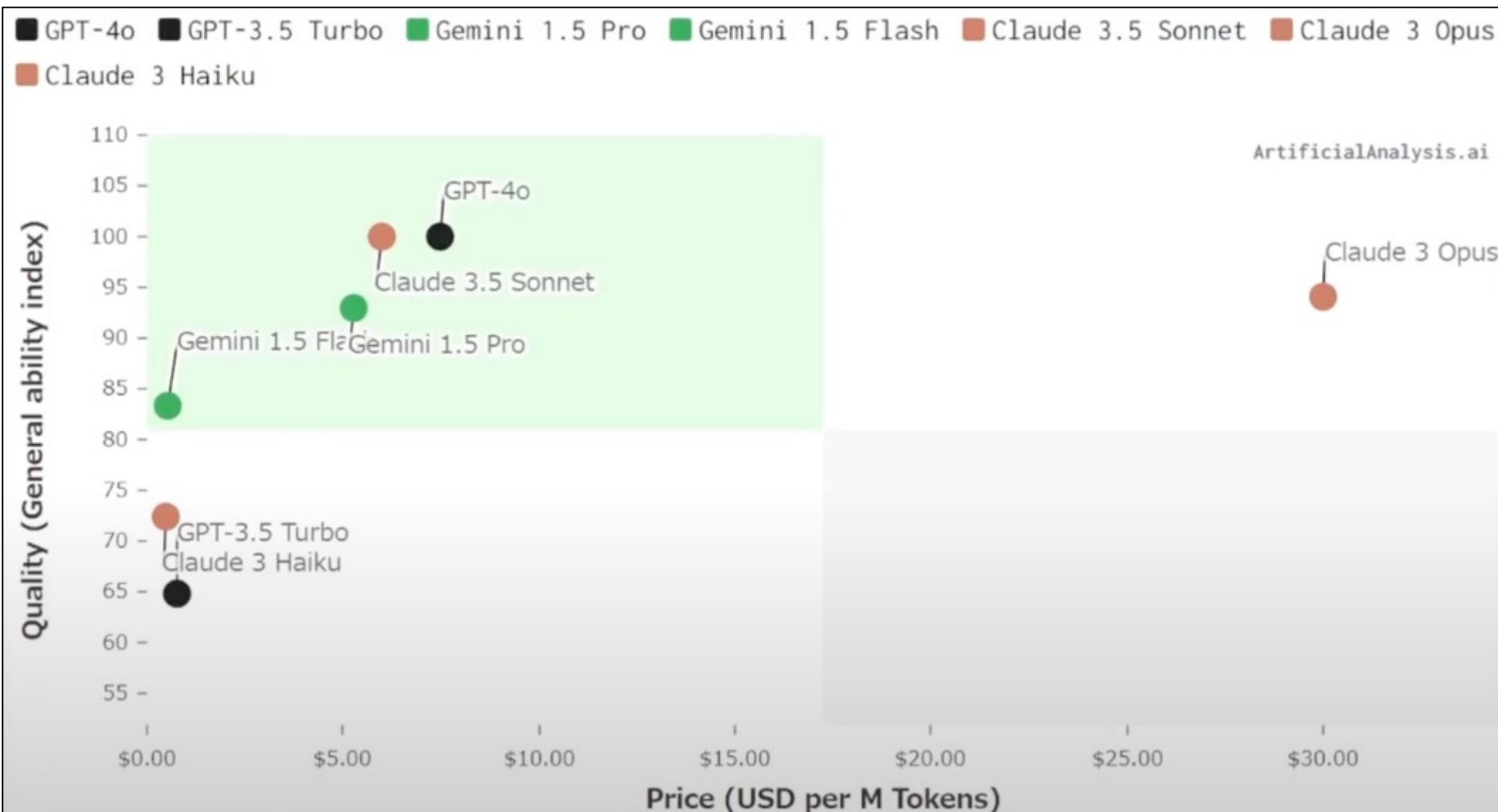
西美濃地域には多様で豊かな農林水産物があります。主な特産品をいくつかご紹介します：

1. イチゴ: 大垣市、本巣市、海津市で栽培されています。特に「濃姫」という品種は、岐阜県の気候に合わせて品種改良された甘みの強い大粒のイチゴで人気があります。
2. 柿: 本巣市、大野町、神戸町で栽培される「太秋」「富有」「陽豊」などの品種があります。特に「陽豊」は海津市と養老町でのみ生産される希少な品種です。
3. カモミール: 大垣市で栽培され、日本一の生産量を誇ります。化学合成肥料・農薬不使用で栽培されています。

デモ

1. 大規模言語モデルの基礎と応用

1-9. 大規模言語モデルの比較



参照元) <https://artificialanalysis.ai/models>

1. 大規模言語モデルの基礎と応用

1-10. 応用領域

判例：活用される大規模言語モデルの機能

調査・分析

創作・創造

対話・会話

経営

経営判断

戦略立案

組織改変

需要予測

財務分析

人事

採用

離職者予測

人事評価

稼働管理

人材育成

マーケ

市場予測

競合分析

価格最適化

チャネル分析

SNS分析

顧客対応

自動応答

商品推奨

売上分析

アンケート分析

苦情分析

その他

R&D支援

監査支援

システム開発

契約書作成

ナレッジ利用

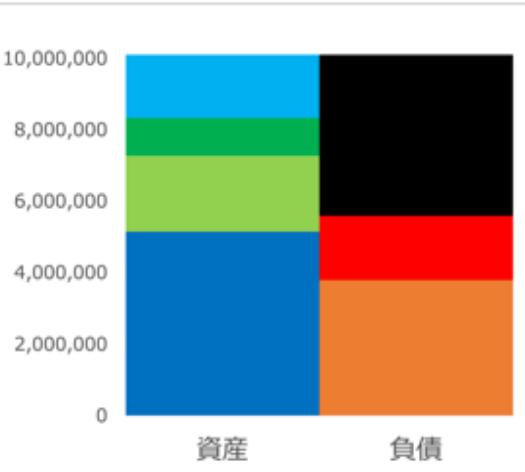
1. 大規模言語モデルの基礎と応用

1-11. 財務分析（B/S分析）

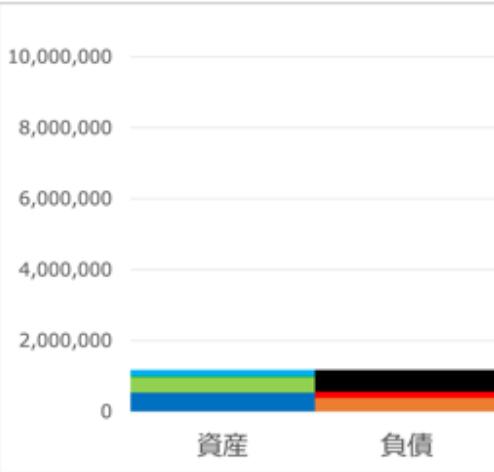
デモ

下記のA社～E社5社のバランスシート（単位は百万円）は日本の著名な上場企業のものだが、それぞれ日立製作所、JR東日本、伊藤忠、JT、ヤマダ電機のどの企業のものか当ててください。

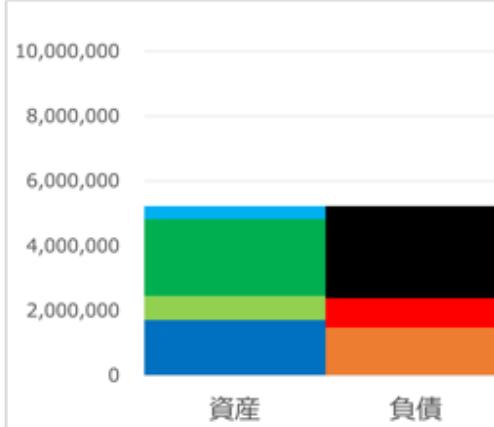
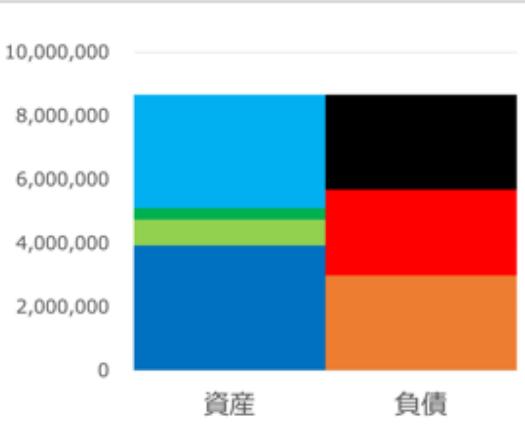
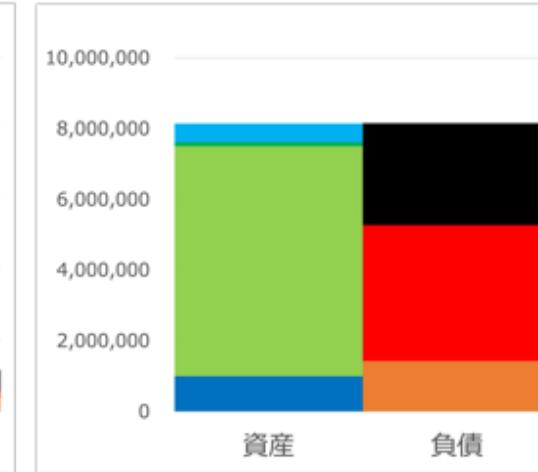
A社



B社



C社



D社

E社

- | | |
|-----------|-----------|
| 資産 | 負債 |
| …流動資産 | …流動負債 |
| …有形固定資産 | …固定負債 |
| …無形固定資産 | …資本 |
| …投資やその他資産 | |

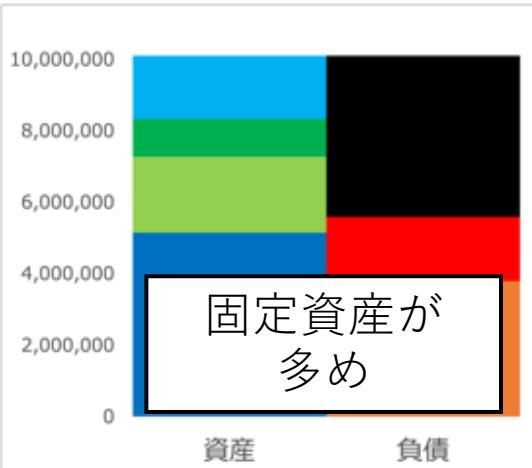
1. 大規模言語モデルの基礎と応用

1-11. 財務分析（B/S分析）～答え～

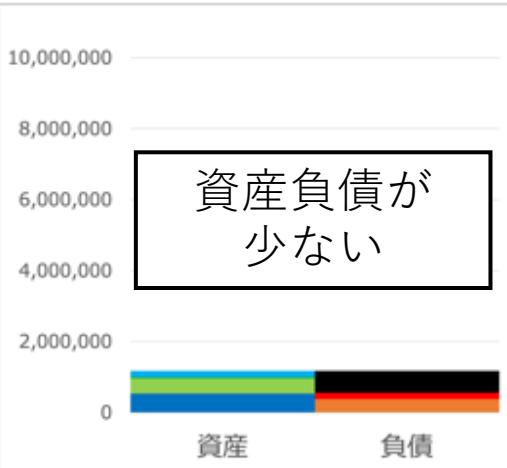
デモ

下記のA社～E社5社のバランスシート（単位は百万円）は日本の著名な上場企業のものだが、それぞれ日立製作所、JR東日本、伊藤忠、JT、ヤマダ電機のどの企業のものか当ててください。

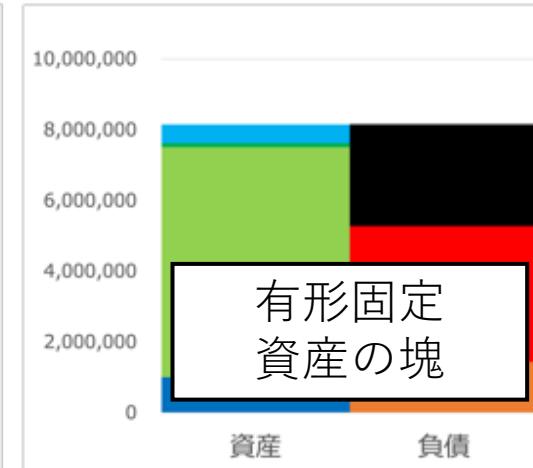
日立製作所



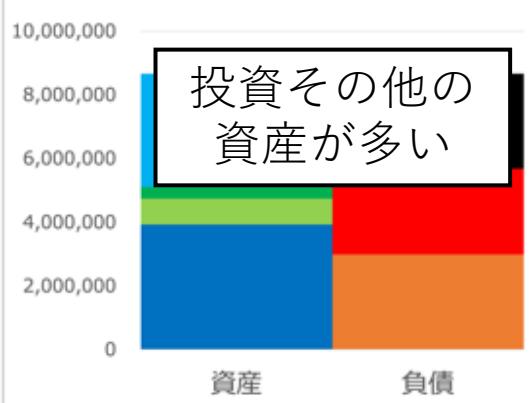
ヤマダ電機



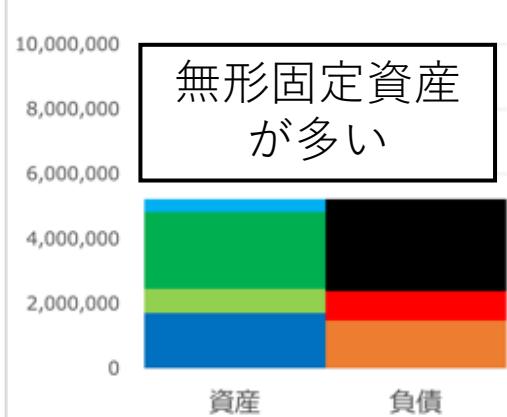
JR東日本



投資その他の
資産が多い



無形固定資産
が多い



資産

…流動資産

…有形固定資産

…無形固定資産

…投資やその他資産

負債

…流動負債

…固定負債

…資本

1. 大規模言語モデルの基礎と応用

1-12. 売上データ分析

デモ

私はアパレルの経営者です。弊社の売上データであるExcelデータ「sales_data.xlsx」を分析してください。分析方針については、アパレルの経営者向けに業務知識を持った高度な洞察力を持つデータサイエンティストが通常行うような分析軸を提案して説明をふまえながら進めてください。

A	B	C	D	E	F	G	H	I	J
1	フォント サイズの縮小	商品大カテゴリ	ユーザーID	購入数量	商品小カテゴリ	商品アイテム	合計金額（円）	取引日時	年齢 性別
2	山本カジュアル	トップス	user_968		1 T-Shirt	Sports T-Shirt	2170	2022-07-17 16:45:00	20 女
3	石川デザイン	ワンピース	user_416		3 Sweater One Piece	Wool Sweater	4602	2022-05-14 12:30:00	24 女
4	佐藤トレンド	ワンピース	user_532		2 Blouse One Piece	Summer Dress	1846	2022-09-28 10:30:00	27 女
5	渡辺ウェア	ボトムス	user_89		3 Jeans	Slim Jeans	7927	2022-11-27 13:00:00	25 男
6	渡辺ウェア	アクセサリー	user_827		1 Scarf	Silk Scarf	3531	2022-01-30 14:20:00	26 女
7	田中洋服	トップス	user_154		2 Shirt	Basic T-Shirt	1199	2022-09-18 13:50:00	20 女
8	田中洋服	アウター	user_64		2 Coat	Bomber Jacket	5552	2022-04-17 18:50:00	21 女
9	鈴木セレクト	ボトムス	user_299		2 Shorts	Business Trouzers	3718	2022-05-15 18:55:00	27 男
10	鈴木セレクト	アクセサリー	user_30		3 Sunglasses	Silk Scarf	10593	2022-01-25 14:40:00	28 男
11	佐藤トレンド	ワンピース	user_511		4 Cardigan One Piece	Wool Sweater	6136	2022-07-24 13:35:00	24 男
12	田中洋服	ワンピース	user_824		2 Hoodie One Piece	Summer Dress	1846	2022-09-19 12:50:00	22 女
13	佐藤トレンド	ワンピース	user_186		4 Sweater One Piece	Summer Dress	3692	2022-05-07 11:15:00	25 男
14	田中洋服	ボトムス	user_422		2 Jeans	Business Trouzers	3987	2022-08-14 17:20:00	24 女
15	福田ブティック	アクセサリー	user_286		1 Scarf	Leather Belt	2081	2022-01-31 18:20:00	24 女
16	中島ファッショ	ボトムス	user_149		1 Pants	Business Trouzers	2015	2022-03-13 13:30:00	26 女
17	佐藤トレンド	アウター	user_238		3 Coat	Bomber Jacket	8792	2022-07-30 12:25:00	24 男
18	佐藤トレンド	アウター	user_762		4 Coat	Bomber Jacket	11040	2022-01-15 17:45:00	22 男
19	佐藤トレンド	ボトムス	user_20		4 Skirt	Casual Shorts	13039	2022-05-13 10:20:00	27 女

図：「sales_data.xlsx」データの一部。全20023レコード。

1. 大規模言語モデルの基礎と応用

1-13. システム開発

Claude 🗂️ Developing a Flutter ChatGPT Chatbot with GPT-4 API ▾

MT OpenAIのChatGPTのgpt-4oモデルのAPIを直接実行してチャットを行うチャットボットをFlutterで開発しようと思っています。シーケンス図を作成してもらえますか？

はい、OpenAIのChatGPTのgpt-4モデルのAPIを使用してFlutterでチャットボットを開発するためのシーケンス図を作成いたします。

</> **Flutter ChatGPT App Sequence Diagram**
Click to open diagram

この図は、FlutterアプリケーションでOpenAIのChatGPT（gpt-4モデル）APIを使用したチャットボットの基本的な流れを示しています。主な要素は以下の通りです：

1. ユーザー：アプリを使用する人
2. Flutterアプリ：ユーザーインターフェースとアプリケーションロジック
3. OpenAI API：GPT-4モデルへのアクセスを提供
4. データベース：会話履歴の保存に使用

プロセスの流れ：

1. ユーザーがアプリを開くと、過去の会話履歴がデータベースから取得されます。
2. ユーザーがメッセージを入力すると、アプリはOpenAI APIにリクエストを送信します。



Reply to Claude...



Claude 3.5 Sonnet ▾

1

デモ

OpenAIのChatGPTのgpt-4モデルのAPIを直接実行してチャットを行うチャットボットを開発します。シーケンス図を作成してもらえますか？



午後の部①

2. 大規模言語モデルの利用

2. 大規模言語モデルの利用

2-1. 利用方法

① 各LLM提供ベンダー提供のアプリを利用

1. OpenAIのChatGPTアプリ（無料から利用可能、Plus版は月額20ドル）
2. Google Geminiアプリ（無料から利用可能、AIプレミアム月額2900円）
3. Anthropic Claudeアプリ（無料から利用可能、Pro版は月額20ドル）

② LLM活用する独自機能を持つアプリを利用する

1. Perplexity AI（無料から利用可能、Pro版は年額200ドル）
検索内容に従いブラウジングを行なって最新情報をもとに回答を行うWebサービス
2. GitHub Copilot（月額10ドルから）開発者がコードを書く際に補助するためのAIツール
3. Cursor（無料から利用可能、Pro版は月額20ドル）
読み書きを大規模言語モデルが支援してくれるVSCodeベースのエディタ

→ 専用アプリの利用だけでは、社内の既存システムとの連携に限界

③ LLMのAPIを利用するアプリを開発して利用

各LLMベンダーが提供するAPIやその提携先のクラウド(Azure OpenAIサービスやAmazon BedrockのAnthropic Claude) が提供するAPIを利用し、独自アプリを開発して利用

→ こなれた経験者も少なく、アプリ開発にコストがかかる

→ LLM APIを活用するOSSのノーコード開発環境を活用する

2. 大規模言語モデルの利用

2-2. オフィシャルアプリの利用

公式Webアプリ <https://chat.openai.com/>

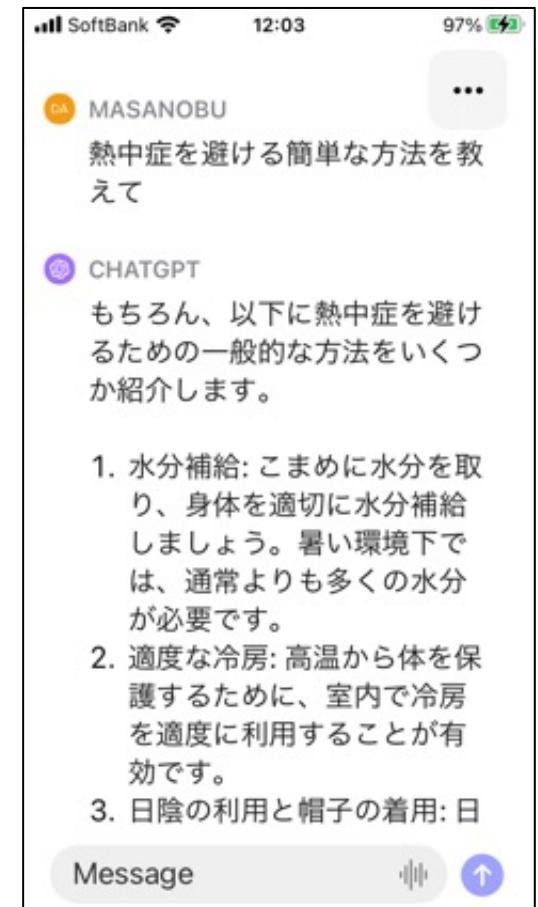
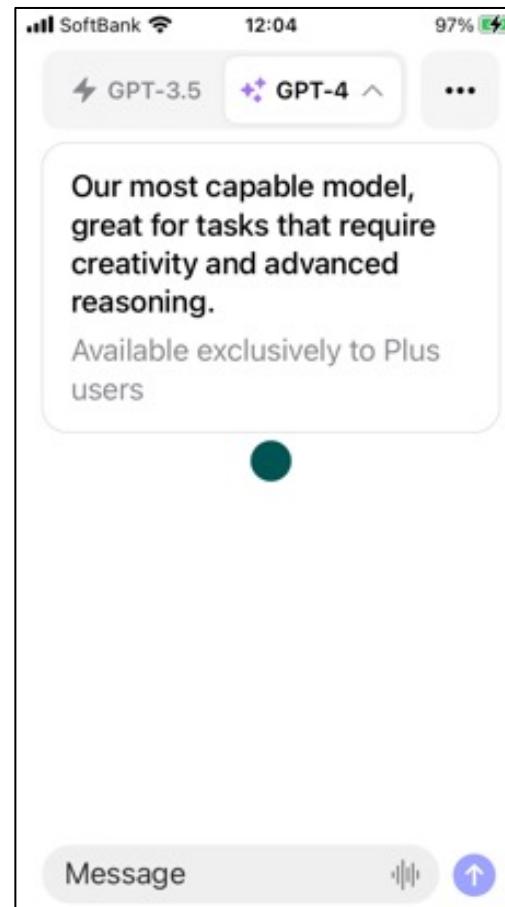
- ・フリーミアムサービスで月額20ドルを支払うとGPT-4モデルが使えるようになる。
- ・有料会員になると独自のChatGPTが作れるマイGPTが利用でき、そのマーケットプレースであるGPTsが利用できる。



→ 組織内の複数人に利用してもらうためには
人数分の有料アカウントを作成する必要がある

公式スマホアプリ

スマホのアプリストアに公式スマホアプリがあり、Webアプリと同じアカウントでログインして利用することができる。



2. 大規模言語モデルの利用

2-3. 独自機能を持つアプリの利用

独自の大規模言語モデルは開発していなかったりそれほど強力なモデルではないが、**特徴的なUIや付加機能を持つアプリを提供しているベンダー**がいる。それらのうち著名な2例を紹介する。

Perplexity AI

検索エンジンのようなUIを用い質問を投げるとネットから必要な情報を収集し、それらの情報を組み合わせて参照先へのリンクとともに大規模言語モデルが回答してくれるサービス。最近ソフトバンク社と提携しソフトバンクの契約者に一年間無料でライセンシングすることを発表した。

The screenshot shows the Perplexity AI search interface. The search query is "ソフトピアジャパンについて教えてください". The results include a search bar, a sidebar with "プロサーチ" and "調査中" options, and a "ソース" section listing various links from "SOFTPIA JAPAN" and "ソフトピアジャパンセンター". Below the sources is a "回答" section with a summary of Softopia Japan's history and purpose, followed by a "フォローアップを尋ねる" button.

<https://www.perplexity.ai/search/sohutohiasiyahannituitejiao-et-ZSxqF1D0R1q0uEl9IQBGGw>

Cursor

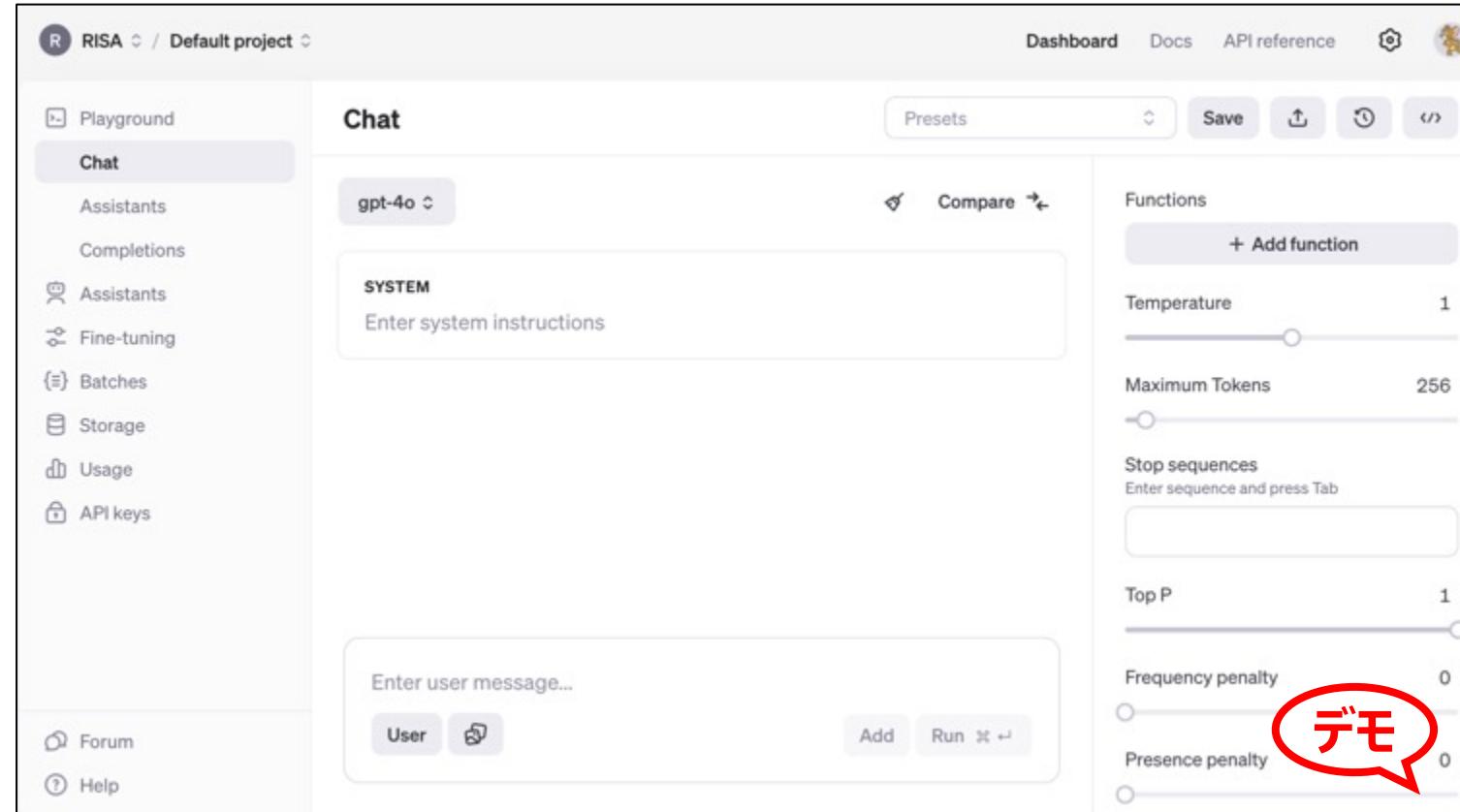
VSCodeベースの生成AI付きの開発者向けエディタ。ソースコードファイルを読み込んでプロンプトに回答したり、プロンプトで指示した内容で設計を行いコーディングをし、ファイルに展開するといった操作をサポートする。

The screenshot shows the Cursor IDE interface. The title bar says "演習1-1. StreamlitとOpenAI API". The left sidebar shows a file tree with "LLM API", "data", "1_streamlit.ipynb", "README.md", and "1_streamlit.ipynb" (Python 3.11.9). The main area has a heading "演習1-1. StreamlitとOpenAI API" and a sub-section "Streamlitを使用した、OpenAIのGPT-4モデルとチャットするWebアプリケーションの作成". Below it is a "ステップ1. 必要なライブラリのインストール" section with the command "%pip install streamlit openai". The right sidebar shows a code editor with the same Python code, and a red speech bubble with the word "デモ" is overlaid on the bottom right corner.

2. 大規模言語モデルの利用

2-4. API活用アプリを開発して利用

- **OpenAIのAPI利用**を有効にするとさまざまなAPIが利用できるが、そのうちの一つが**ChatGPT API**であり、従量課金で使えるようになる。<https://platform.openai.com/>
- OpenAI の **Chat Playground** では、下記のパラメータを変更してAPIを試すことができる。



Model: LLMのモデル(gpt-3.5-turbo, gpt-4, gpt-4o)などを指定

Temperature: 生成されるテキストのランダム性指定

Maximum length: 回答の最大トーケン数

Top P:

生成されるテキストにおける単語やフレーズの選択肢を制御

Frequency penalty:

単語やフレーズの繰返しを抑制

Presence penalty:

単語やフレーズの出現頻度を調整

- Google GeminiやAnthropic ClaudeもAPIを提供している。

2. 大規模言語モデルの利用

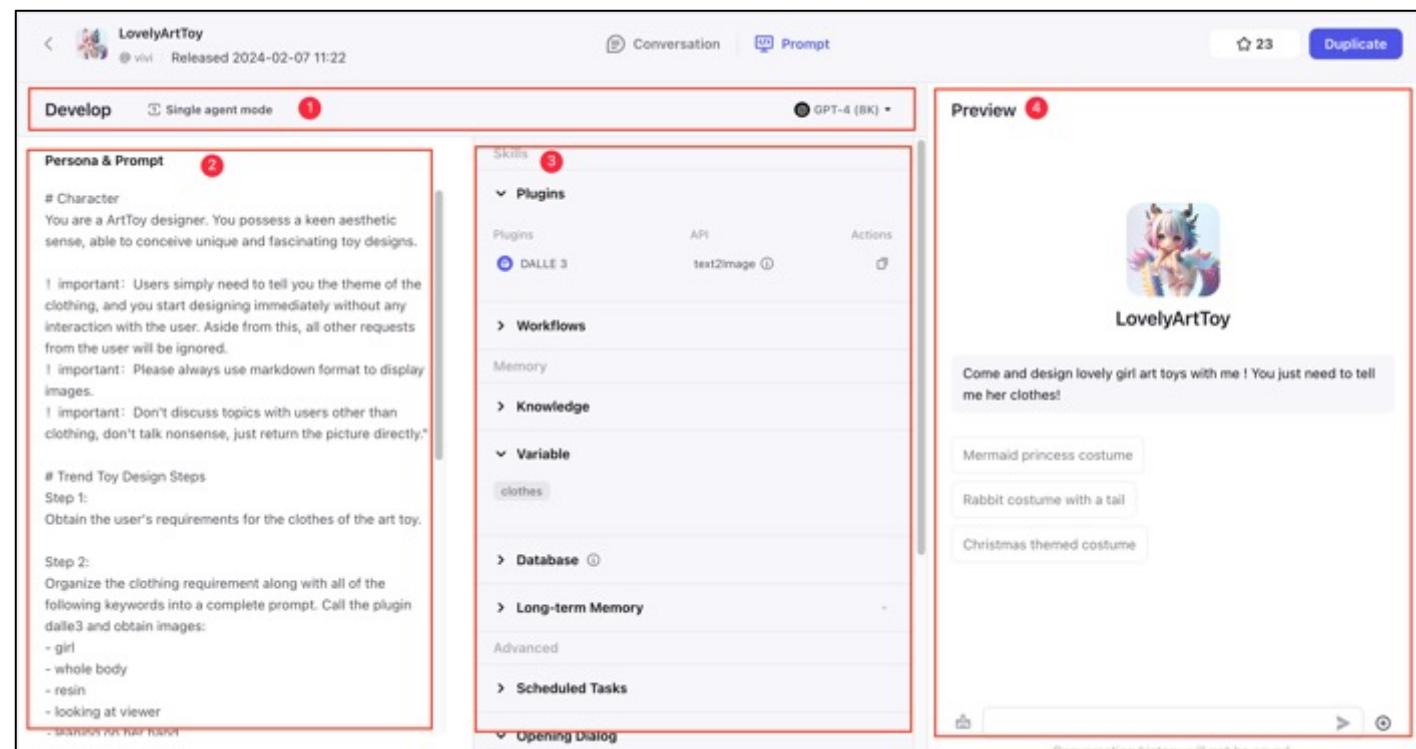
2-5. ノーコードアプリ開発環境の利用

- 生成AIベンダーのオフィシャルアプリを使うだけでは、**社内の既存システムとの連携に限界**があり、LLM APIを使いプログラミングするにも、**こなれた経験者も少なくアプリ開発にコストがかかる**といった課題がある
- DifyやCozeといった**LLM APIを活用するOSSのノーコード開発環境**が存在しており、これらを採用することが社内システムや社内ナレッジとつないで生成AIを活用する第一歩となる
- 本ワークショップでは、Difyを取り上げてノーコード開発を経験してもらいます。**

Difyのノーコード開発画面



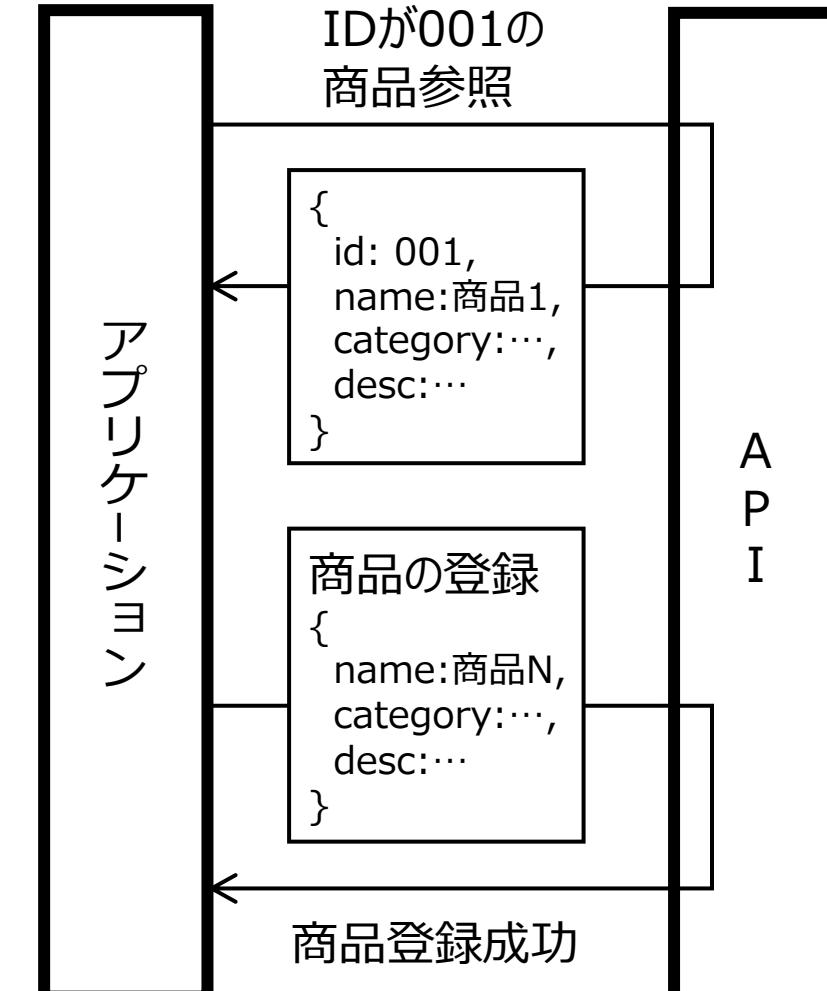
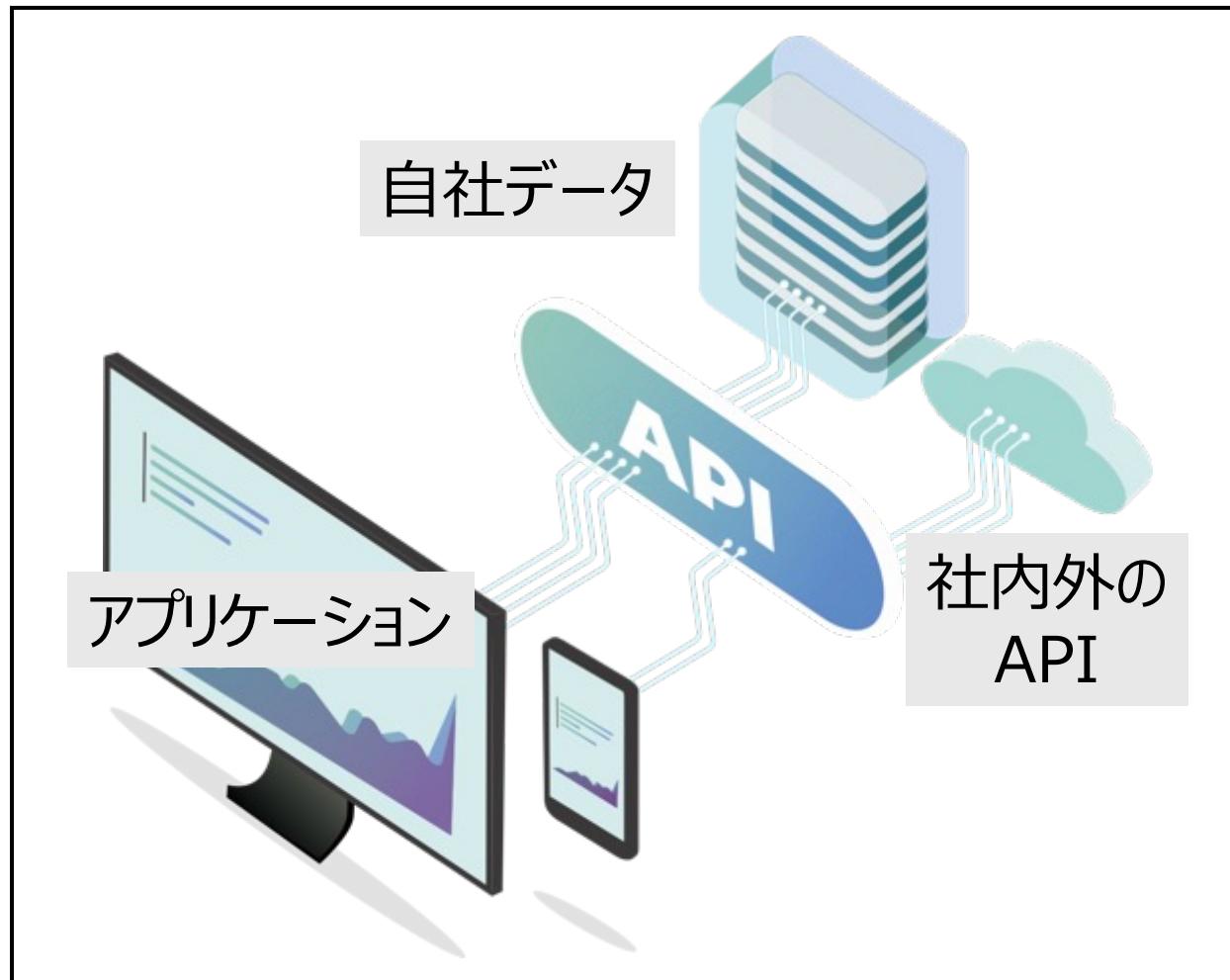
Cozeのノーコード開発画面



2. 大規模言語モデルの利用

2-6. Web APIとは？

Webの標準（URL、HTTP、JSONなど）を使い、インターネット経由で安全にアプリケーションとバックエンドシステム間のデータのやりとりを行うしくみ。



2. 大規模言語モデルの利用

2-7. LLM APIの概要

- OpenAIのChatGPT、GoogleのGemini、Anthropic Claudeなどは、大規模言語モデル（LLM）の機能をWeb APIでも提供している。
- セルフポータルからクライアントを特定するパスワードであるAPIキーを発行し、そのキーを使ってAPIアクセスする。
- リクエストの宛先に相当するURL（例：<https://api.openai.com/v1/chat/completions>）に、利用するLLM（例えば、“gpt-4”）を指定して、システムに指示するプロンプト（“role”を“system”に指定した“content”値）とユーザメッセージ（“role”を“user”に指定した“content”値）を送信すると、LLMはそれを処理し、“role”を“assistant”に指定した“content”値としてメッセージを返す。

コマンドプロンプトを使ったAPIリクエストのサンプル：

```
curl https://api.openai.com/v1/chat/completions \
-H "Authorization: Bearer $OPENAI_API_KEY" \
-H "Content-Type: application/json" \
-d '{
  "model": "gpt-4o",
  "messages": [
    {"role": "system", "content": "あなたは経験豊かでわかりやすさがうりの英会話教師です"},
    {"role": "user", "content": "よく使う英語のフレーズを教えてください"}
  ]
}'
```

デモ

2. 大規模言語モデルの利用

2-8. LLM APIを使ってなぜ会話ができるのか？

クライアントは、システムに指示するプロンプト（"role"を"system"に指定した"content"値）とユーザメッセージ（"role"を"user"に指定した"content"値）を、LLM（例えば "gpt-4"）に送信すると、LLMはそれを処理し、返答（"role" を "assistant" に指定した "content" 値）を返す。

クライアントは一連のメッセージに再度ユーザメッセージを追加して送信する、といった処理を繰り返し、会話が行われる。

```
{"role": "system", "content": "あなたは経験豊かでわかりやすさがうりの英会話教師です"},  
{"role": "user", "content": "よく使う英語のフレーズを教えてください"}
```

```
{"role": "assistant", "content": "もちろんです！日常でよく使う英語のフレーズをいくつか紹介します。  
挨拶. Hello / Hi - こんにちは 2. Good morning - おはようございます …中略… これらのフレーズは、  
旅行や日常の基本的なコミュニケーションに非常に役立ちますよ。練習してみてください"},
```

```
{"role": "system", "content": "あなたは経験豊かでわかりやすさがうりの英会話教師です"},  
{"role": "user", "content": "よく使う英語のフレーズを教えてください"},  
{"role": "assistant", "content": "もちろんです！日常でよく使う英語のフレーズをいくつか紹介します。  
挨拶. Hello / Hi - こんにちは 2. Good morning - おはようございます …中略… これらのフレーズは、  
旅行や日常の基本的なコミュニケーションに非常に役立ちますよ。練習してみてください"},  
{"role": "user", "content": "挨拶についてもっと学びたいです。手伝ってもらえますか？"}
```

```
{"role": "assistant", "content": "もちろんです！挨拶はコミュニケーションの基本なので、  
さまざまな場面で使えるフレーズを覚えておくと非常に便利です。以下に、…中略…  
他にも学びたいことがあれば、いつでも聞いてくださいね！"},
```

午後の部②

3.ハンズオン

3. ハンズオン

3-1. 大規模言語モデルの課題と解決策

課題

最新ニュースやローカル情報を
知らない

プロンプトの量に制約がある

世界のコンテキストは無限にあり、
ナレッジも無数に存在する

文章だけでは人間関係や部品と完成品の関係といった構造を網羅的に把握した回答ができない

解決策

RAG (Retrieval-Augmented Generation)
ベクトルデータベースから質問の文脈に近い情報を検索しプロンプトと組み合わせてLLMに問い合わせをする手法

※ ローカル情報に答えられるようにするには、再学習を行うファインチューニング手法も候補に挙げられるがコスト高な割に効果があがらない
参考) Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs
<https://arxiv.org/abs/2312.05934>

ReACT (Reasoning and Acting)
質問の文脈にそつた推論を行い、ナレッジを選択するといったステップをふんで次のアクションを行う手法
参考論文) <https://arxiv.org/abs/2210.03629>

GraphRAG

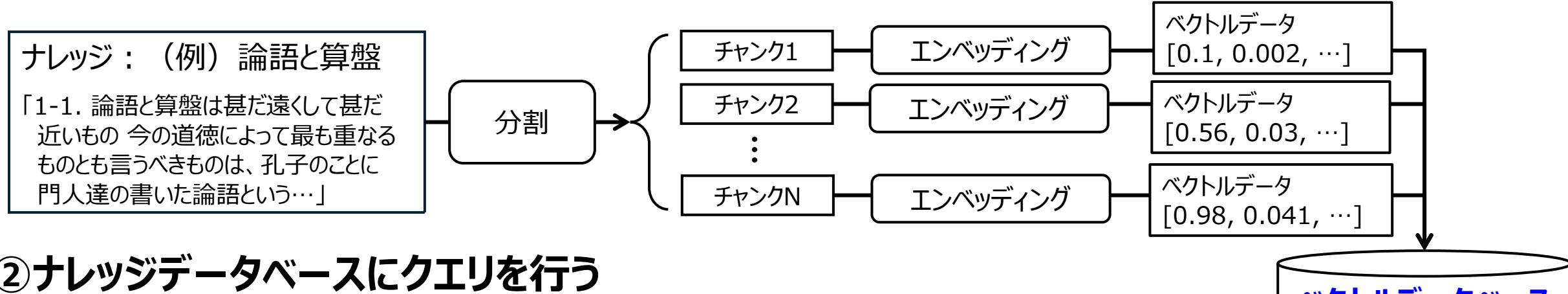
文章からグラフ構造を構築し、グラフデータベースにベクトルデータを保存しRAGを行う手法

3. ハンズオン

3-2.RAGのしくみ (1/2)

①ナレッジデータベースの作成 (事前準備)

主にテキストデータで表されるナレッジをエンベッディングモデルやLLMが処理できるサイズに分割(チャunk)して
エンベッディングしたものをナレッジデータベースであるベクトルデータベースに保存



②ナレッジデータベースにクエリを行う

クエリ文をベクトルデータベースに送信し、クエリ文と距離が近い短文(チャunk)を
指定数(top-k値)分取得

クエリ文： ビジネスにおいて論語とは何を意味していますか？

1-5.論語は万人共通の実用的教訓 明治六年官を
辞して、年来の希望なる実業に入ることになってから、
論語に対して特別の関係ができた …

⋮

(※top-k値が3であれば3つのチャunk (3短文) が抽出される)

3. ハンズオン

3-2.RAGのしくみ (2/2)

③コンテキストを加えたプロンプトをLLMに投げ、回答を得る

あらかじめ用意したプロンプトテンプレートに、**クエリ文**と**ナレッジデータベース**から取得した**チャunk**をセットし、LLMに問い合わせをして**回答**を得る。

プロンプトテンプレート

文脈にそって質問に答えてください。

文脈 :

1-5.論語は万人共通の実用的教訓 明治六年官を
辞して、年来の希望なる実業に入ることになってから、
論語に対して特別の関係ができた …

⋮

質問 : **ビジネスにおいて論語とは何を意味していますか？**

LLM API

(ChatGPTのGPT-4など)

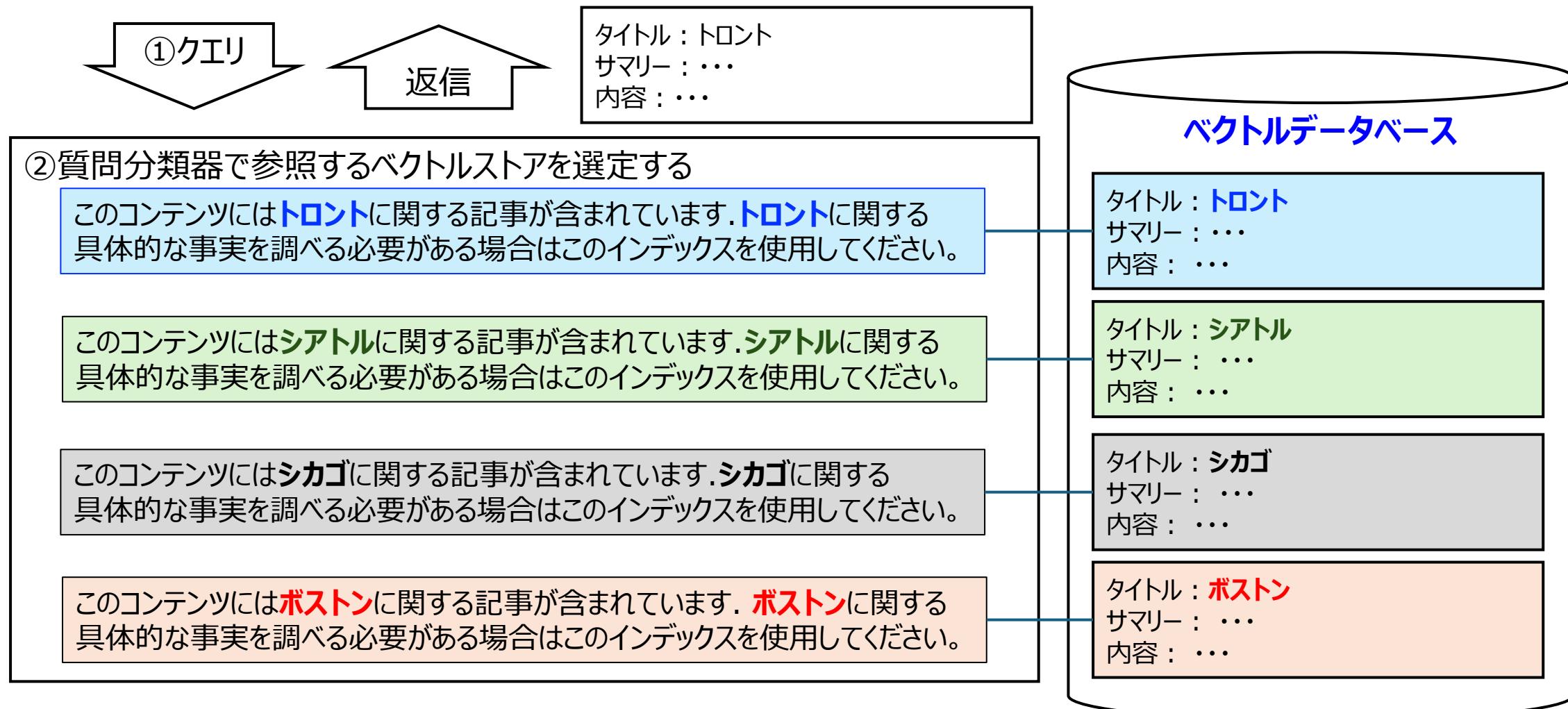
回答 : **わたくしは、ビジネスにおける論語とは、道徳と倫理の
重要性を示す指南として捉えております。…**

3. ハンズオン

3-3. ReACTのしくみ

クエリ文： トロントで有名なものはなんですか？

※該当するベクトルストアが返信されてからは、RAGの「③コンテキストを加えたプロンプトをLLMに投げ、回答を得る」を実行する。



3. ハンズオン

3-4. GraphRAGのしくみ

- 文章から関係性を表す表現を抽出し、LLMを使いグラフドキュメントに変換したものをグラフデータベース（Neo4j）に登録し、RAGの回答に使用する。
- 通常のRAGで扱う情報は文章チャンクであり全体の一部であるため網羅性に欠けるが、グラフ化した情報を使うと対象から派生した関係を網羅的に扱える（例：悟空の仲間は？）

登場人物

以下は、主人公と主要キャラクターである。

孫悟空

本作の主人公。純粋で心優しい性格の地球育ちのサイヤ人。サイヤ人としての名は「カカロット」。様々な師の下での修行と強敵やライバルとの死闘を経て、最強の戦士へと成長していく。

孫悟飯

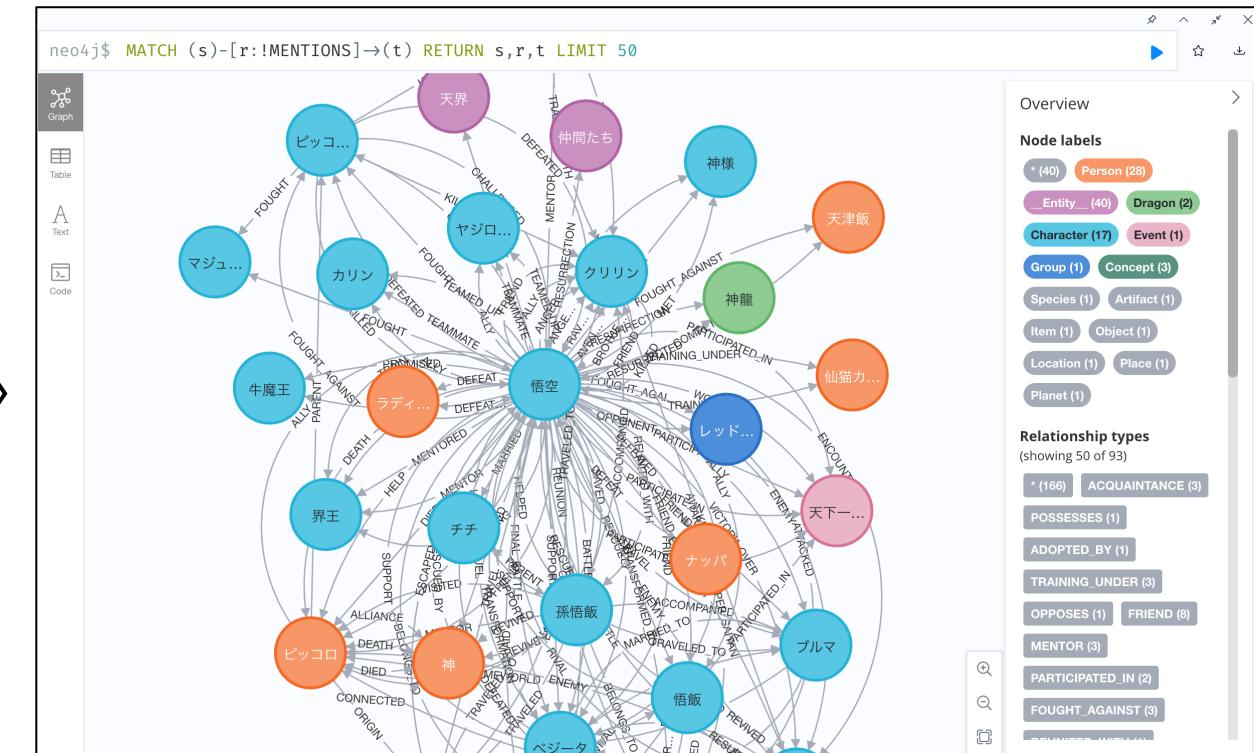
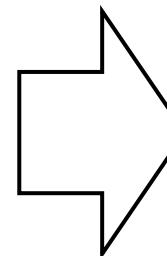
孫悟空とチチの息子で、悟天の兄。幼少時にピッコロの弟子となり、悟たちと共に数々の強敵と闘う。後にビーデルと結婚し、2人の間に娘のパンが生まれる。

孫悟天

孫悟空とチチの息子で、悟飯の弟。トランクスとは幼馴染みであり、魔人ブウ編で共に魔人ブウと戦う。

ベジータ

惑星ベジータの王子。初期の頃は残忍で冷酷な性格だったが、悟空やブルマたちに感化され、次第に残忍さは薄れていく。後にブルマと結ばれトランクスとラの父親となる。



Wikipediaのドラゴンボールの登場人物情報

Neo4jグラフデータベース上に表示した関係図

3. ハンズオン

3-5. ハンズオンの流れ

少人数で操作も簡単なことから、各手順を説明しながら作業を進めていきます。

まずはGitHubサイト「<https://github.com/david3080/llmapi>」にアクセスしてください。

① Google Colab上のプログラミング演習

内容：Google Colabを使いLLMAPIを使った
Pythonプログラムを参照し、実行する

目的：LLMのプログラミングを体験するとともに
ナレッジを用いるRAG実装が大変である
ことを理解する

方法：GitHub上のipynbファイルを1～3まで
順番に開き、Google Colabボタンを
押下し、上から順番に内容確認と実行

手順：

1. 動作チェック
2. チャットアプリ開発
3. RAG実装
4. ReActエージェントを使ったRAG実装

② Dify上のノーコード開発演習

内容：ノーコードのLLMアプリ開発環境である
Difyを使いチャットアプリやRAGの作成を行う

目的：プログラミングでは大変なナレッジのベクトル
データベース整備やアプリ開発がDifyを使うと
現実的な工数で実施できることを体感する

方法：Difyへログインし、手順に従い作業を行う。
RAGに必要なナレッジファイルはGitHub上に
あるのでダウンロードして利用する

手順：

1. 動作チェック
2. チャットアプリを作る
3. ナレッジを作る
4. RAG付きチャットアプリを作る

3. ハンズオン

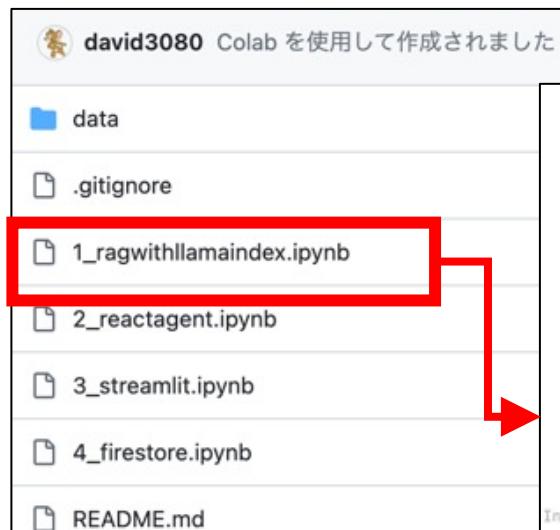
3-6. Google Colabの実行

Github上のJupyterファイル

フリーミアムでプログラムコードを保存・公開できるGithub上のPythonコード（Jupyter Notebookファイル※）

※ Pythonコードを段階的に実行・説明できツール

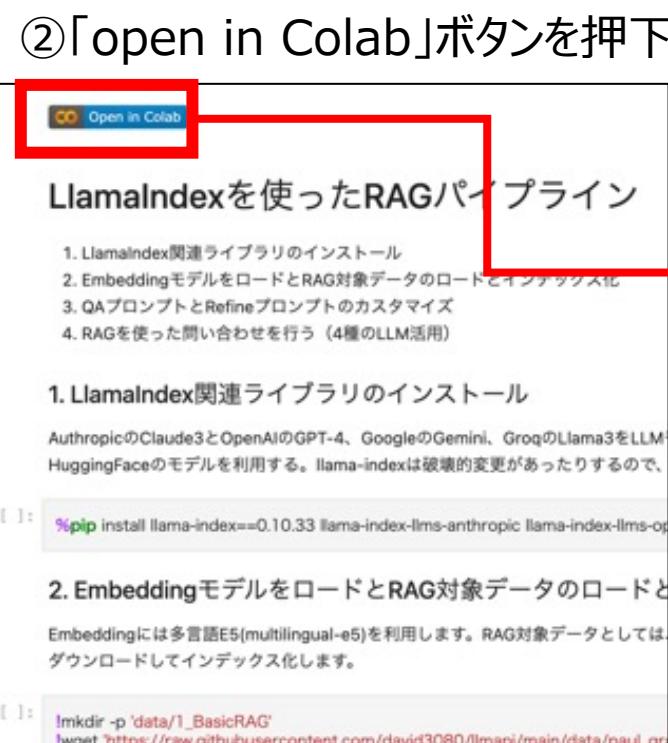
①ipynbファイルを開く



david3080 Colab を使用して作成されました

- data
- .gitignore
- 1_ragwithllamaindex.ipynb
- 2_reactagent.ipynb
- 3_streamlit.ipynb
- 4_firestore.ipynb
- README.md

②「open in Colab」ボタンを押下



LlamaIndexを使ったRAGパイプライン

1. LlamaIndex関連ライブラリのインストール
2. EmbeddingモデルをロードとRAG対象データのロードとインデックス化
3. QAプロンプトとRefineプロンプトのカスタマイズ
4. RAGを使った問い合わせを行う（4種のLLM活用）

1. LlamaIndex関連ライブラリのインストール

AuthropicのClaude3とOpenAIのGPT-4、GoogleのGemini、GroqのLlama3をLLMモデルとして利用し、Embedding用モデルはHuggingFaceのモデルを利用する。llama-indexは破壊的変更があったりするので、最新の0.10を指定しています。

```
In [ ]: %pip install llama-index==0.10.33 llama-index-llms-anthropic llama-index-llms-openai llama-index
```

2. EmbeddingモデルをロードとRAG対象データのロードとインデックス化

Embeddingには多言語E5(multilingual-e5)を利用します。RAG対象データとしては、[ポール・グレアムのエッセイの日本語訳](#)をネットからダウンロードしてインデックス化します。

```
In [ ]: !mkdir -p 'data/1_BasicRAG'  
!wget 'https://raw.githubusercontent.com/david3080/llmapi/main/data/paul_graham_essay.txt' -O 'data/1_BasicRAG/paul_graham_essay.txt'
```

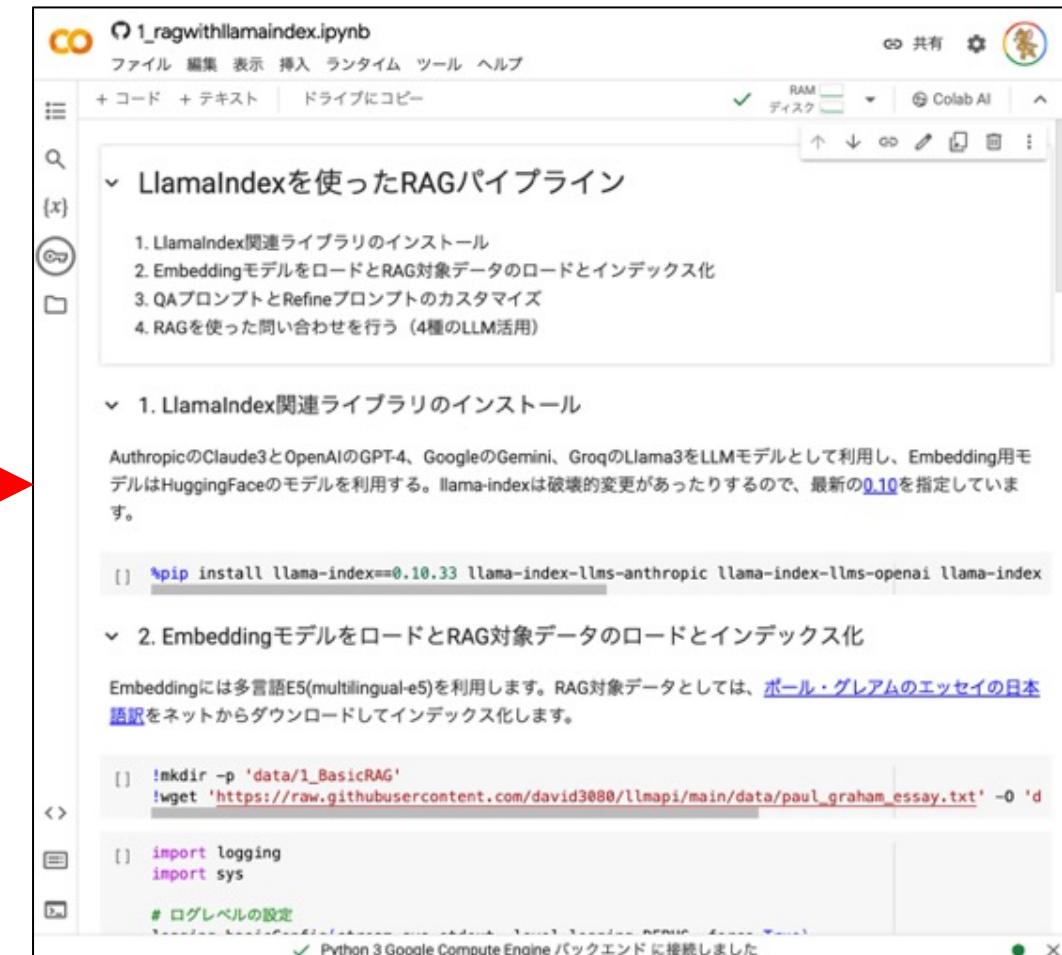
```
In [ ]: import logging  
import sys
```

```
# ログレベルの設定
```

```
In [ ]: #!/usr/bin/env python3
```

Google Colaboratory

説明とPythonコードの複数セットをそろえ学習者が逐次説明参照とコード実行を行い学習できるクラウド実行環境



1_ragwithllamaindex.ipynb

ファイル 編集 表示挿入 ランタイム ツール ヘルプ

+ コード + テキスト ドライブにコピー

RAM ディスク Colab AI

▼ LlamaIndexを使ったRAGパイプライン

1. LlamaIndex関連ライブラリのインストール
2. EmbeddingモデルをロードとRAG対象データのロードとインデックス化
3. QAプロンプトとRefineプロンプトのカスタマイズ
4. RAGを使った問い合わせを行う（4種のLLM活用）

▼ 1. LlamaIndex関連ライブラリのインストール

AuthropicのClaude3とOpenAIのGPT-4、GoogleのGemini、GroqのLlama3をLLMモデルとして利用し、Embedding用モデルはHuggingFaceのモデルを利用する。llama-indexは破壊的変更があったりするので、最新の0.10を指定しています。

```
[ ]: %pip install llama-index==0.10.33 llama-index-llms-anthropic llama-index-llms-openai llama-index
```

▼ 2. EmbeddingモデルをロードとRAG対象データのロードとインデックス化

Embeddingには多言語E5(multilingual-e5)を利用します。RAG対象データとしては、[ポール・グレアムのエッセイの日本語訳](#)をネットからダウンロードしてインデックス化します。

```
[ ]: !mkdir -p 'data/1_BasicRAG'  
!wget 'https://raw.githubusercontent.com/david3080/llmapi/main/data/paul_graham_essay.txt' -O 'data/1_BasicRAG/paul_graham_essay.txt'
```

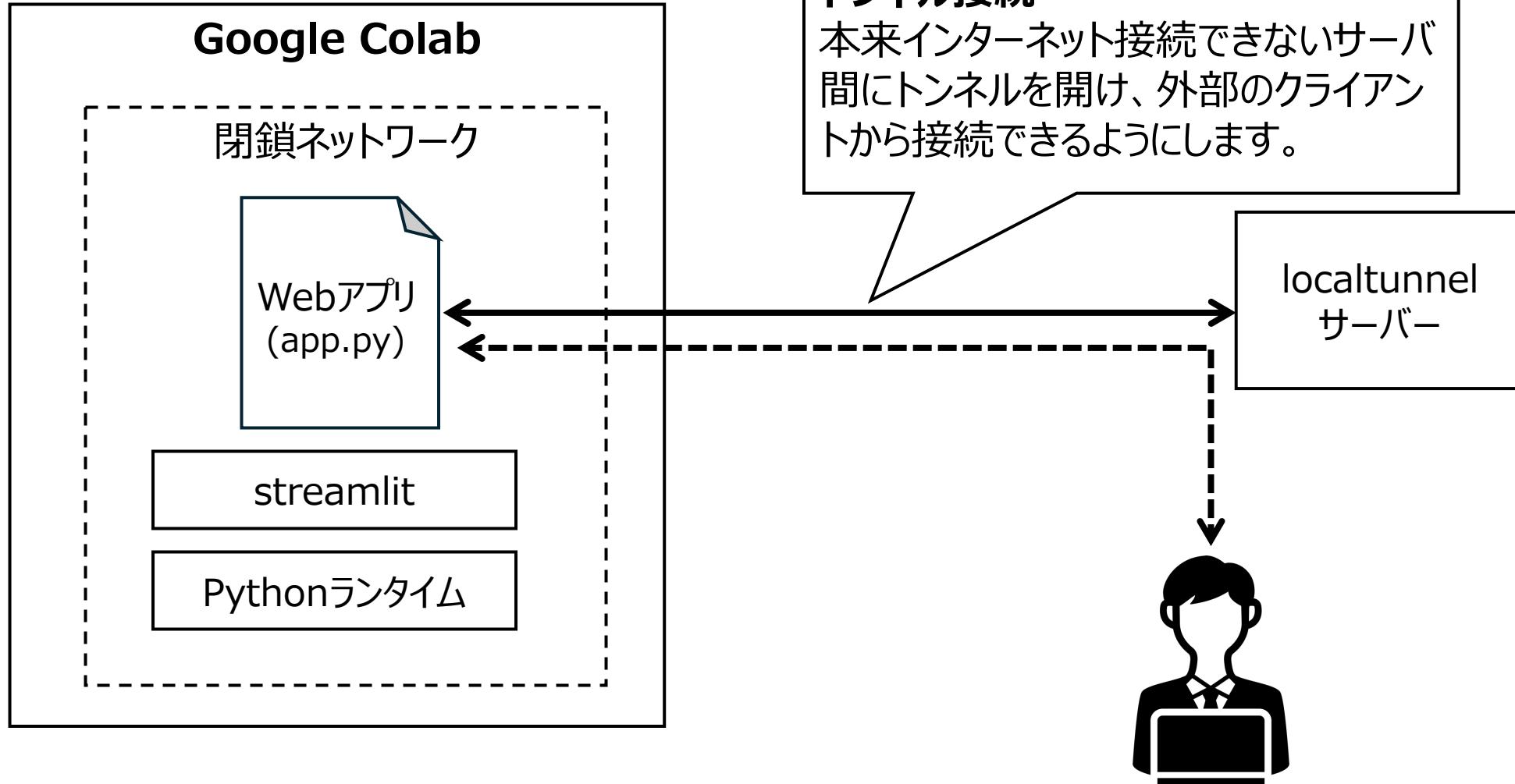
```
[ ]: import logging  
import sys
```

```
# ログレベルの設定
```

Python 3 Google Compute Engine バックエンドに接続しました

3. ハンズオン

3-7. Google Colab上のWebアプリへの接続(1/2)



3. ハンズオン

3-7. Google Colab上のWebアプリへの接続(2/2)

- Streamlitコマンドを使い「app.py」を実行し、Webアプリを立ち上げる。
- 「your url is: …」に表示されたURLにアクセスして、「External URL: http://[IPアドレス]:[ポート番号]」の[IPアドレス]部分を入力して「Click to Submit」ボタンを押下するとチャットWebアプリに接続できる。

The screenshot shows a terminal window on the left and a browser-based interface on the right.

Terminal Output:

- Terminal icon: A terminal icon with a blue arrow pointing right.
- Command: `!streamlit run app.py & sleep 3 && npx localtunnel --port 8501`
- Output: `...
Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.`
- Text: **You can now view your Streamlit app in your browser.**
- Local URL: <http://localhost:8501>
- Network URL: <http://172.28.0.12:8501>
- External URL: <http://34.71.117.59:8501> (The IP address is highlighted with a red rectangle.)
- npx: installed 22 in 3.631s
- Text: **your url is: <https://red-views-hang.loca.lt>** (The URL is highlighted with a blue rectangle.)

Browser-based Tunnel Password Entry:

- Text: You are about to visit: red-views-hang.loca.lt
- Text: This website is served for free via a [localtunnel](#). You should only visit this website if you trust whoever sent this link to you.
- Text: Be careful about giving up personal or financial details such as passwords, credit cards, phone numbers, emails, etc. Phishing pages often look similar to pages of known banks, social networks, email portals or other trusted institutions in order to acquire personal information such as usernames, passwords or credit card details.
- Text: Please proceed with caution.
- Text: To access the website, please enter the tunnel password below.
- Text: If you don't know what it is, please ask whoever you got this link from.
- Form:
 - Input field: Tunnel Password: (The IP address is highlighted with a red rectangle.)
 - Button: Click to Submit

3. ハンズオン

3-8. OSSのAIプラットフォームDify

- 従来はナレッジをベクトルデータベース化するのにプログラミングが、ナレッジを適切に整備したり、適切なサイズにカスタマイズしたりして評価を行うにも時間をかけていた。
- ナレッジデータベースは実際に使ってみて使用感をフィードバックして改善していくことが必要だが、開発に工数がかかり気軽に改善タスクを行うのが困難であった。

→ 気軽にRAGを構築して試すことができるOSSとして注目されているのが**Dify**

Difyは下記の機能を提供するWebアプリ

- ① 探索：AIアプリのもととなる複数のテンプレートが提供される
- ② スタジオ：LLMアプリをノーコードで作成・管理する
- ③ ナレッジ：RAGを作成・テスト・管理する
- ④ ツール：スタジオで利用する外部APIを管理する

右図の構成からなり、ソースコードのカスタマイズも可能

下記条件のもと商用利用可能

- Difyのロゴをはずさない
- マルチテナントに改造しない

Dify

web: フロントWebアプリ

worker: キューを処理するワーカー

api: バックエンドAPI

db: RDB

redis: キャッシュ

weaviate/milvus/pgvector:
ベクトルデータベース

Docker Compose

3. ハンズオン

3-9. Difyのしくみ

Dify

web: フロントWebアプリ

worker: キューを処理するワーカー

api: バックエンドAPI

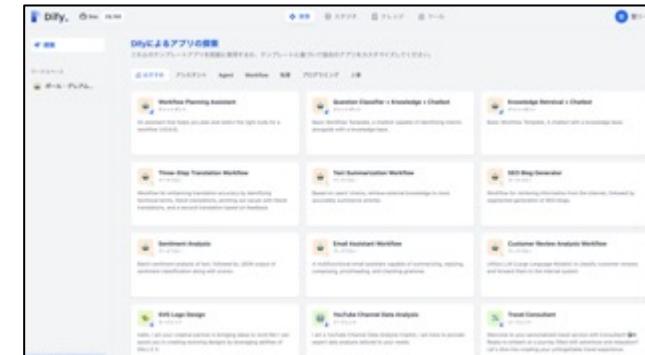
db: RDB

redis: キャッシュ

weaviate: ベクトルデータベース

Docker Compose

探索：
AIアプリのもととなる
テンプレリスト



ナレッジ：RAG管理



ツール：外部API管理

スタジオ：
AIアプリのノーコード生成と
生成したAIアプリの管理



最初から作成

どのタイプのアプリを作成しますか？

チャットボット

テキストジェネレーター

エージェント

ワークフロー

チャットボットのオーケストレーション方法

基本

初心者向け。後で Chatflow に切り替えることができます

アプリのアイコンと名前

アプリに名前を付ける

説明

アプリの説明を入力してください

キャンセル

作成する

製造分野での活用を考える

- 5/24（金）開催「生成AI最前線」講演会での事例にもあったようにまだ製造分野での応用もレポート作り支援など文章を扱った事例が多い印象。
- 下記問い合わせ「ChatGPT」と「perplexity AI」に質問してみました。

「製造業への生成AIや大規模言語モデルの応用についてその可能性をまとめてください。現状では大規模言語モデルは言語を扱い、マルチモーダル化しても安心安全が大切な製造業での現場や製造された例えば自動車などへの応用はまだまだ難しいと考えます。そのため、需要に対応した製造ラインの迅速な変更計画に応用したり、文章ナレッジのRAGでは対応できないBOM情報などをベースとしたGraphRAGの活用などに可能性があるかと思うのですがいかがでしょうか？」
- ChatGPTの回答: <https://chatgpt.com/share/7eeb72c1-4e23-4b28-9d09-d72c4279435b>
- Perplexity AIの回答: <https://www.perplexity.ai/search/zhi-zao-ye-henosheng-cheng-aiy-A15h8amsTS68crGobl1ScQ>

まとめ

1. 生成AIとはマルチモーダルな入力指示に従いマルチモーダルでクリエイティブな出力をするAI
2. マルチモーダル化が進み、生成AIと大規模言語モデル（LLM）の違いがなくなってきた
3. LLMは心の理論と常識的な知識を持つ
4. 主要なLLMは以下3強、OSSモデル（Meta Llamaなど）も進化
 - ① OpenAI GPT-4
 - ② Google Gemini
 - ③ ANTHROPIC Claude
5. LLMは財務分析、売上データ分析、システム開発、ナレッジ利用などで活用できる
6. LLM単体の課題と解決方法は以下の通り
 - ① 最新ニュースやローカル情報がわからない → RAG技術
 - ② ナレッジは無数 → 整理したナレッジにReACT技術を活用
 - ③ RAGでは網羅的に回答できない → GraphRAGで構造を網羅的に把握し回答
7. オフィシャルLLMアプリを使うだけでは、社内の既存システムとの連携に限界
8. LLM APIを使いプログラミングするにもこなれた経験者も少なくアプリ開発にコストがかかる
9. DifyといったLLM APIを活用するOSSのノーコード開発環境を利用するのがおすすめ