



3分で分かる

生成AI/LLM の精度を高める  
トレンドトピック

# RAG

(検索拡張生成)



# 目次

---

生成AI / LLM の分野でトレンドワードとして  
"RAG" というワードを近頃目にする機会が増えてきているのではないのでしょうか？  
本WPでは以下の流れでRAGについて解説させていただきます。

1.	LLMの可能性と限界 .....	3~4
2.	RAGの基本概念 .....	5~7
3.	RAGとファインチューニングの違い .....	8
4.	RAGを導入するメリット .....	9
5.	RAGを導入する際の注意点 .....	10

# LLMの可能性と限界

近年、業務効率化を目的にAIを活用する企業が増えております。  
その中でも特にChatGPTを中心とした "LLM(大規模言語モデル)" が注目されています。

LLMは膨大な情報を学習し、自然言語を理解して生成するAIです。  
質疑応答や文章の要約等の多くの作業をこなすことができ、様々な業務に活用されております。

LLMについて概要を教えてください。



大規模言語モデル (LLM, Large Language Model) は、大量のテキストデータを用いて訓練された人工知能モデルで、自然言語処理 (NLP) タスクを行うために設計されています。これには、文章の生成、翻訳、要約、質問応答、感情分析などが含まれます。LLMは、数十億から数千億のパラメータを持つことが多く、高い精度で人間のようなテキスト生成や理解が可能です。代表的なモデルには、GPTシリーズ (例: GPT-3) やBERTなどがあります。



# LLMの可能性と限界

そんな便利なLLMですがいくつかの課題があります。

課題の1つとして「学習した情報に依存する」ことが挙げられます。

例えば「社内規定」や「LLMが学習していない最新情報」等の未学習の情報に関しては求めている回答を生成することが出来ず、一般的な回答しか返すことが出来ません。

APTOの有給制度に関して教えてください。



メモリを更新しました

APTOの有給制度に関する具体的な情報は公開されていない場合がありますが、一般的な企業の有給休暇制度についてお伝えします。

## 一般的な有給休暇制度

1. **付与日数:** 通常、社員は年次有給休暇として一定の日数を付与されます。新入社員の場合、最初の年は比例配分で付与されることがあります。
2. **取得条件:** 企業によって異なりますが、一般的には入社して一定期間が経過した社員に有給休暇が付与されます。

# RAGの基本概念

前述の通り未学習の情報に関して課題のLLMですが、

"RAG(検索拡張生成)" というアプローチで、未学習の情報も回答させることが可能になります。

そんなRAGですが考え方自体は意外とシンプルで、

「学習していない情報はLLMに渡してあげよう」といった考え方になります。

例えば以下のように未学習の情報に関しては、質問とセットで規定を教えると回答が可能になります。

規約を参考に以下の質問に回答して下さい

# 質問

APTOに在籍中ですが既に今年度有給を15日分消化しました。  
あと何日有給の取得が可能でしょうか？

# 規約

APTOの有給休暇は年に20日付与されます。

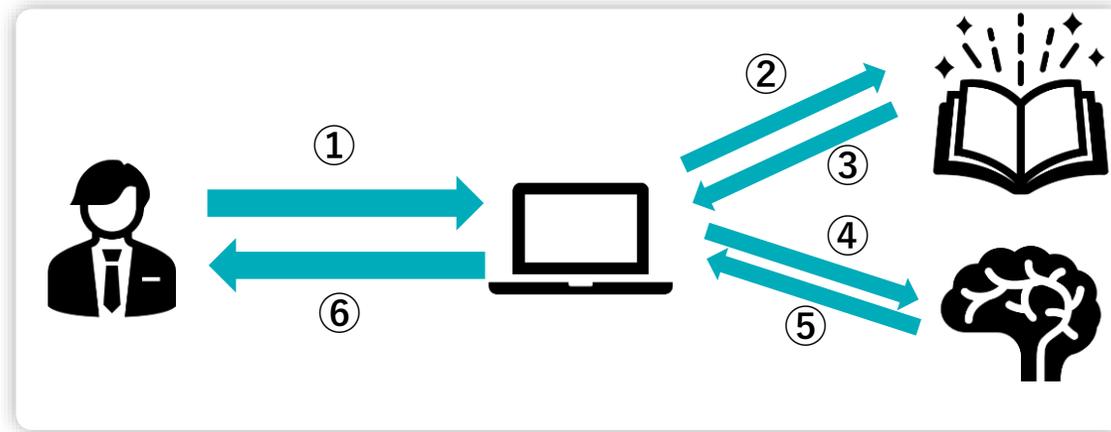


APTOの規約に基づくと、年間の有給休暇は20日付与されます。既に今年度15日分を消化している場合、あと5日有給を取得することが可能です。



# RAGの基本概念

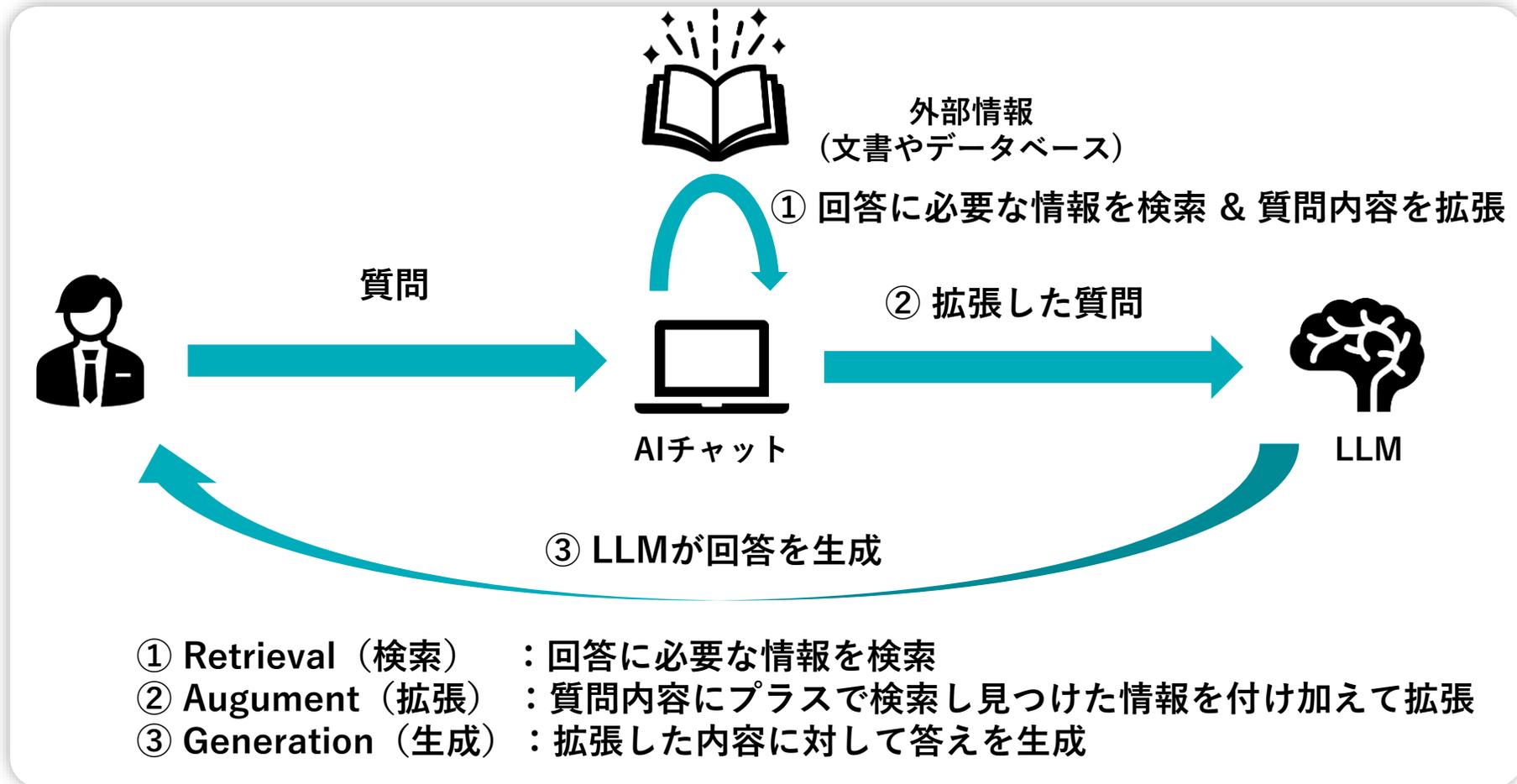
前述の通りRAGは、質問回答に必要な情報を一緒に渡すことがポイントです。  
この仕組みを以下のように用いることで未学習の情報をLLMに回答させることが可能です。



- ①質問： AIチャットに質問を入力
- ②検索： チャットAIアプリは内部で、質問内容を基にナレッジベース（ベクトルデータベースなどの情報源）を検索
- ③結果データ： チャットAIアプリは内部で、検索結果を取得
- ④質問+結果データ： チャットAIアプリは内部で、質問と検索結果を使ってLLMに質問
- ⑤回答： チャットAIアプリは内部で、質問に対する回答を取得
- ⑥回答： チャットAIアプリは、ユーザーに回答を表示

# RAGの基本概念

RAGの語源（Retrieval-Augmented Generation）に沿って説明したものが以下になります。



## RAGとファインチューニングの違い

LLMの精度を高めるアプローチとしてファインチューニングという言葉聞いたことがある方も多いのではないのでしょうか？

RAGとファインチューニングは生成AIの回答精度を向上させるアプローチであるという点は共通しておりますが、それぞれのアプローチ手法は根本的に異なっております。

### RAG

検索エンジンを用いて外部情報から必要な情報を抽出し拡張生成する技術

### ファイン チューニング

LLMのモデル自体のデータを微調整して、特定の課題に適した性能を引き出す技術

ファインチューニングに関しては以下をご覧ください

参考：LLM精度改善ハンドブック  
<https://harbest.io/documents/652>

# RAGを導入するメリット

---

本ページではRAGをAIに活用することによるメリットについて解説させていただきます。

RAG活用により今1番注目されているメリットとしては

**”ハルシネーションの抑制”が挙げられます。**

ハルシネーションとはLLMが事実と異なる内容を生成してしまう現象のことです。

RAGを用いることで外部の知識を参照してからLLMで回答生成をするため、RAGはハルシネーションを抑制するし回答精度を高めるのに有効な手法であると言えます。

ただRAGを実装する上でいくつかの注意点がありますので、  
次ページで解説いたします。

# RAGを導入する際の課題

RAGを実装する際によくある課題としては以下3点が挙げられます。



実装が難しい

AI開発において、検索技術を組み込むためのアルゴリズムの選択や検索クエリの設定や文書の選択等、実装にはかなりの専門知識と技術が必要です。



実装してみたが  
精度が上がらない

検索クエリの設定がうまくいっていない場合、生成される回答は検索された文書の信頼性や関連性に依存するため、誤った情報や不適切な情報を回答する場合があります。



処理に時間がかかる  
場合がある

大規模な文書データベースを検索し関連する情報を抽出するため、処理に時間がかかることがよくあります。適切な検索エンジンの選択や処理の最適化が必要となります。

# 無料相談受付中

APTOはLLMの精度改善やRAGの実装において様々な実績がございます。  
お悩みをお持ちの方はお気軽にご相談ください

## webフォームから

click

無料お問い合わせ  
AI開発相談はこちら

## メールでのお問い合わせ

info@apto.co.jp

会社名・氏名・メールアドレス・電話番号を  
ご記入の上、お問い合わせください

※当社のホームページからでも資料請求・お問い合わせが可能です。

<https://harbest.io/>