

ChatGPT/Gemini/Claudeの基本コマンドから
リアルタイム音声会話プログラム作成まで

独自 AI アプリ作りに役立つ LLM 用 API の使い方

村井 和夫

本稿では、ChatGPTに代表される大規模言語モデル (LLM: Large Language Model) を、プログラムに取り込んで使うための API の利用法について次の2つを説明します。

①基本…curl コマンドと Python コードからの利用方法

②応用例…リアルタイム音声会話プログラム

API を使ってプログラムに組み込むことによって、LLM を独自に構築しなくても、RAG (Retrieval Augmented Generation, 検索拡張生成) やファインチューニングを使って特定目的に特化した LLM を作ったり、さまざまな独自の業務用アプリケーションを作ったりすることが可能になります。

API を利用するときに出てくる 基本用語

現在多数の LLM が公開されていますが、プログラムに取り込む技術は、従来からある Web API と同じで、ネットワーク・プロトコルを通して使います。Web API の使い方は非常に簡単ですが試すことができますが、基本的な用語を理解しておくことがプログラムに組み込むときに役立ちます。

● ネットワーク・プロトコルとエンド・ポイント

LLM の API も基本的に従来からある Web API と同じです。Web API で使われるネットワーク・プロトコルには単純な HTTP/HTTPS だけでなく、REST、SOAP などいろいろなフレームワークがあります。現在では、多くの場合 API は HTTPS をベースにすることが多くなっています。しかし、リアルタイムの音声や映像を扱う必要がある場合、HTTPS を拡張した WebSocket や、リアルタイム通信専用の WebRTC のようなストリーミングに適したプロトコルを使用します。エンド・ポイントは、API 機能を提供する接続先 URL のことを指します。

● API キー

Web API の基本は従量制課金なので、利用者の認

証が必要です。各社は API キーを発行して、それを使って利用者をネットワーク・プロトコルで認証して接続を許可します。そのため、ネットワーク・プロトコルは、必ずトランスポート・セキュリティに対応したプロトコルが使われます。このように、ネットワーク上でのセキュリティは担保されていますが、プログラムの中に直接書き込まないなど、利用する端末やプログラムで厳重に管理する必要があります。

● コマンド・プロンプト

LLM に何を求めるかの質問です。プロンプト・エンジニアリングという用語があるように、求める役割と目的を明確に記述しないと、的確なレスポンスを得ることができません。基本的な役割や言語などの設定は、質問のコマンド・プロンプト以外に、別途設定するようになっている場合もあります。

● コマンド・レスポンス形式

Web API のコマンド・レスポンスの形式には XML や JSON 形式などいろいろあります。現在ではほとんどが JSON 形式になっています。従って、API マニュアルを参考にして、JSON 形式のコマンドを組み立てて送信し、JSON 形式のレスポンスを受けて必要な情報を取り出す処理がプログラムの基本となります。

基本…curl コマンド or Python で 呼び出す (Gemini, ChatGPT, Claude)

代表的 AI として表1に示すモデルの使い方を紹介します。

いずれも、非常に安価なものからプログラムや理系に強い最新の高機能モデルまで実にさまざまなモデルがありますが、執筆時点で無料枠が一番大きく安価なのは Gemini でした。今回は、AI の API の利用方法を確認することを目的として、執筆時点で一番安価で簡単な会話レベルで利用しやすいモデルを選び、API の利用方法を紹介します。