はじめに

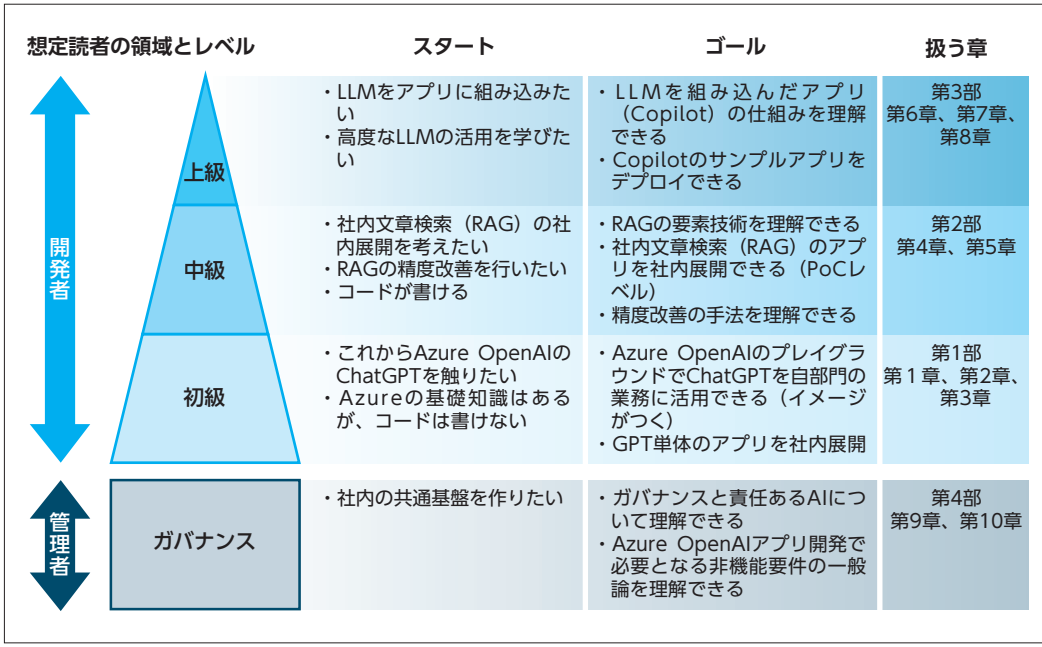
2022年11月にOpenAI社から公開され、以来世界を席巻しているChatGPT。この革新的な技術は、我々の生活を大きく変えつつあります。とくにその衝撃は企業のIT領域において顕著であり、その応用範囲は拡大の一途をたどっています。Microsoft Azure（以下Azure）はChatGPTをはじめとするOpenAIモデルを利用できる唯一のパブリッククラウドサービスであり、今日ではChatGPTを活用したい企業にとって、Azureの採用が不可欠となっています。その一方で、ChatGPTなどの大規模言語モデル（LLM）を利用したアプリケーションは新しいコンセプトも多く登場し、ChatGPT/OpenAIモデルを活用したシステムを設計するうえで、体系だった説明を目にする機会はあまりありません。

本書は、企業内でOpenAIモデルやChatGPTの活用を推進するエンジニアやデジタルトランスフォーメーションを担う人々に向け、AzureからOpenAIモデルにアクセスできる「Azure OpenAI Service」（以下Azure OpenAI）の基礎から具体的なアーキテクチャ設計まで解説していきます。本書を通じて、読者のみなさんがOpenAIモデルやChatGPTを活用し、新たな価値を創出するための一助となることを願っています。

本書の構成

本書は読者の目的とレベル感の違いに対応するため、大きく4つの部より構成されています。これからAzure OpenAIを触りたいというところからスタートし、社内文章検索（検索拡張生成；RAG）を経て、LLMを組み込んだアプリケーション（Copilot）の構築へとステップアップしていきます。また、ガバナンスと責任あるAIについても解説しています。

図0.1　本書各部の想定読者とゴール設定



第1部は第1～3章で構成され、AzureでのChatGPT活用をテーマにしています。Azureの基礎知識はあるもののAzure OpenAI自体はこれから触るという読者を想定し、ChatGPT単体のアプリを社内展開して業務で活用できるイメージがつくのを目標にしています。まずは生成AIとChatGPTモデルの基本的な概念とその仕組みを解説します。Azure OpenAIの概要と具体的な利用方法までを解説し、Azure OpenAI StudioのプレイグラウンドよりChatGPTアプリをユーザー自身の環境に展開します。さらに、ChatGPTに入力する指示となるプロンプトをどう書いていくかといった「プロンプトエンジニアリング」という重要なテクニックについても解説します。

第2部は第4章と第5章で構成され、ChatGPTモデルを活用した社内文章検索（RAG）システムの導入をテーマにしています。社内文章検索システムの社内展開の検討や、精度改善を行いたい読者を想定し、RAGの概念理解から実際の社内文章検索アプリの展開まで行います。社内文章検索アプリのキーとなるAzureサービスの紹介や、実際のアーキテクチャについて解説します。検索精度や回答生成精度の改善アプローチについても紹介します。

第3部は第6～8章で構成され、ChatGPTモデルやLLMを組み込んだアプリケーションである「Copilot」の考え方を紹介しています。Copilotを開発するうえで必要な要素を抽象化したCopilot Stackの解説を行い、要素技術であるAIオーケストレーション、基盤モデルとAIインフラストラクチャ、Copilotフロントエンドをそれぞれ説明します。

第4部は第9章および第10章で構成され、LLMアプリケーションを開発・運用していくうえでのガバナンスと責任あるAIについて解説します。Azure OpenAIを中心にLLMを組織全体で活用するための基盤構築やその実現方法について、認証・認可やログ管理、課金、流量制限、閉域化、負荷分散などの視点から、非機能要件の一般論とともに説明します。また責任あるAI活用のためのデータの取り扱いやコンテンツフィルタリングにも触れます。

なお、「ChatGPT」という単語は、OpenAIが開発したChatGPTのモデル自体（GPT-3.5 TurboやGPT-4モデル）を指すこともあれば、それらモデルを組み込んで一般のユーザーが使いやすいようにOpenAIが開発して提供しているアプリケーションを指すこともあります。本書では特段の注釈がない限り、ChatGPTはGPT-3.5 TurboやGPT-4のモデルを表します。

また、本書はAzure OpenAIを前提に解説しているため、OpenAIが提供するChatGPT Enterpriseといったサービスに関して詳しくは紹介していません、また、音声認識AIであるWhisperや画像生成AIのDALL-Eは割愛し、テキスト生成モデルであるGPTモデルのみを解説します。

注意事項

本書に掲載している情報はすべて執筆時点（2023年12月）のものです。Azureは、ユーザーの利便性を向上させるための機能追加を頻繁に行っています。本書に掲載している情報・画面と、実際の画面について差異が生じている場合もあるためご注意ください。Azure以外の技術については、各技術の公式ドキュメンテーションに基づいており、非公式の情報源に基づくものではありません。ただし、公式の情報が更新されるたびに本書の内容も更新されるわけではないため、最新の情報については常に公式の情報源をご確認ください。

本書はAzureの基礎的な説明は割愛しています。また、読者がAzureを利用可能なアカウントを持っている前提で、使い方の解説も進めていきます。もしAzureアカウントをお持ちでない場合はAzureアカウントを取得することをお勧めします[注0.1]。

本書で取り扱っているAzure OpenAIは、責任あるAI活用の観点から利用承認制のサービスになっています。悪用や意図しない危害を防ぐため、リスクの低いユースケースや軽減策の取り入れを行っているお客様をまずは対象としています。Microsoftはより広いお客様を対象とできるように取り組みを行っていますが、現時点で個人利用のお客様は承認が難しい状況となっているためご注意ください[注0.2]。

注0.1「Azure の無料アカウントを使ってクラウドで構築」 https://azure.microsoft.com/ja-jp/free

注0.2 2023年12月現在。最新情報は以下をご確認ください。
「Azure OpenAI Service とは」 https://learn.microsoft.com/azure/ai-services/openai/overview