イントロダクション 1

# ChatGPTをはじめとした大規模言語モデル（LLM）の現在

山本 大輝

表1　LLMにはローカルで動くものとウェブ・サービス上のものがある

| 言語＼実行 | ローカル | ウェブ・サービス |
|---|---|---|
| 英語 | LLaMA, Mistral, Gemma | ChatGPT, Gemini, Claude 3 |
| 日本語 | Swallow, ELYZA, Japanese Stable LM Gamma , Rakuten AI | |

ChatGPTはOpenAI社から2022年11月30日に発表されたチャット・サービスです．高度なAI技術を用いて自然な会話が可能であり，さまざまなトピックに対応できるため，広く使用されるようになりました．

ChatGPTの根幹にある技術は大規模言語モデル（LLM）です．LLMは従来まで使われていたモデルよりも大規模なモデルで，大量のデータとモデルの持つパラメータの数が多い言語処理のモデルのことを指します．

ChatGPTの広がりからさまざまな会社が，類似のLLMを利用したチャット・サービスの提供や，独自のLLMの構築に取り組んできており，技術者/非技術者問わず，急激に世界に広がってきています．

## LLMの今

### ● ユーザが利用できるLLMあれこれ

ChatGPTの成功は多くの注目を集め，その波に乗り多くの企業や団体がLLMの開発に乗り出しました．大別すると，ウェブ・ブラウザやAPIなどのサービスとして提供されているモデル，ローカルで利用できるモデルに分けられます．また，モデルの中でもどのような言語で学習されているものかで分けられます．全体を**表1**に整理しました．

**▶ウェブ・サービス**

ウェブ・サービスとして提供されているものは，ChatGPTが最も有名です．また，Claude 2/3（Anthropic社），Gemini（グーグル）は，特に高精度なサービスとして知られています．これらはAPI経由で利用可能であり，有料であれば独自アプリケーションへの組み込みも可能です．

**▶ローカル環境**

ローカル環境で利用できるモデルの開発が進んでいます．LLaMA（Meta），Mistral（Mistral AI社），Gemma（グーグル）がその例です．ただし，これらのモデルは英語を中心として学習しているので，日本語の認識を英語と同等のレベルまで進めるのは難しいです．そのため日本語においてもLLMの開発は活発に進められています．Swallow（東京工業大学），ELYZA（ELYZA），Japanese Stable LM Gamma（Stability AI），Rakuten AI（楽天）などが開発されています．

ローカルで動作するモデルは，APIやサービスとして提供されるものと比較して精度が劣りますが，GPUを使って独自のモデルを学習させることが可能です．この分野では日々新たな開発が行われており，開発者たちは最良のモデルを目指して競い合っています．ローカルでモデルを構築する際には，最新かつ最良の選択肢を検討することが重要です．

### ● ChatGPTで使用されるモデル

ChatGPTも公開当初と比較して利用できるモデルが変化してきています．2023年初頭はGPT-3.5でしたが，さらに精度が高いGPT-4が発表されました．2023年9月ころにはGPT-4の中でも画像の入力にも対応しているGPT4-Vも利用できるようになりました．

ChatGPTのサービスに課金すればGPT-4が追加で利用できます．この2つのモデルによる違いを**表2**に示します．GPT-4は精度が高い反面，サービスに課金しないと使えないなど制約はありますが，画像が使える，質が高いといったメリットがあります．さらにチャットの入力に従って画像を生成するDALL-E 3，チャットの入力を踏まえてコードを作成し，実行までするCode Interpreterといったものを利用でき，格段にできることが増えます．

### ● LLMのチャット・サービス

ChatGPT，Gemini，Claude 2/3といったサービスが展開されてきました．**表3**のような違いがあります．比較するとChatGPTは独特なサービスがチャットベースで利用でき，Code Interpreter，DALL-E 3が扱えます．Geminiはグーグル・サービスとの連携が強みでしょう．また，Claude 3はベンチマーク・データセットを元に計測されたモデルの精度がChatGPTやGeminiより高くなっています．

各サービスともにモデルはアップデートされていき