第1章 ChatGPTもLlamaもELYZAも基本はコレ!

大規模言語モデルの仕組み

石垣 達也

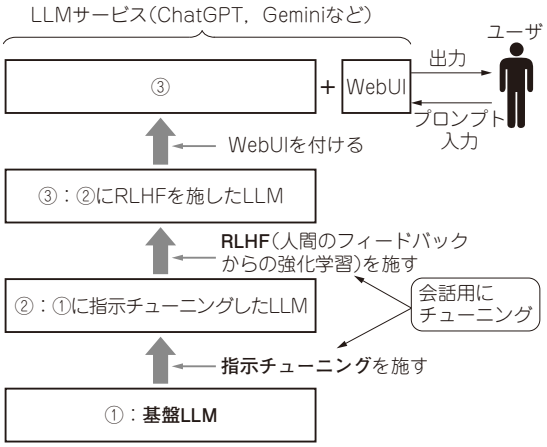

図1 ChatGPTなどLLMサービスの根幹を成す基盤LLM

ChatGPTなどの大規模言語モデル(LLM：Large Language Models)サービスが広く普及し，大きな注目を集めています．LLMに与える指示(プロンプト)を適切に記述すると，翻訳，要約，アイデア出しなどさまざまな知的な処理が効率良く実現できることから，多くの分野で活用が進んでいます．

そこで本稿では，LLMの開発の現状と仕組みを解説します．具体的には，現在リリースされているLLMの概要を説明し，その背景となる技術のトランスフォーマについて，平易な数式を交えながら解説します．そもそもLLMにはどのような種類があるのか，LLMの仕組みはどうなっているのか，10B級とは何かといった素朴な疑問も解消できるでしょう．

LLM開発の「今」を理解する

ChatGPTなどのLLMサービスでは，プロンプトを入力すると，返答を生成してくれます．これは，もともとは会話に特化していないLLMを，会話できるようにチューニングしたものです(**図1**②と③)．チューニング前のLLMを基盤LLM(Foundation LLM)と言います(**図1**①)．そのため，まずは基盤LLMの性能こそが重要です．

基盤LLMの性能は，大規模になるにつれ向上します．規模は内部で行われる行列演算の数で測られ，現在，10-billion(100億パラメータ)級～175-billion(1750億パラメータ)級の基盤LLMが登場していて，さらなる大規模化が進んでいます．

表1 代表的な多言語LLM

モデル名	モデル・サイズ	ソースコードの開示状況
GPT-3	1750億パラメータ	非公開
GPT-4	非公開	非公開
Llama 2 (7B，13B，70B)	70億，130億， 700億パラメータ	オープンソース

● 代表的な基盤LLM

表1に代表的な基盤LLMを示します．

▶商用…OpenAIのChatGPTなど

ChatGPT[2]の基盤LLMとして使われているのはGPT(Generative Pretrained Transformers)[7]です．GPTには幾つかのバージョンがありますが，2024年5月の執筆段階ではGPT-3が主流です．ChatGPTの有料サブスクリプションを購入しているユーザは，より大規模で強力なGPT-4を基盤LLMとしたサービスを使うことができます．GPT-3は1750億パラメータという規模で，GPT-4のパラメータ数は非公開ですが，GPT-3よりもはるかに多いと思われます．ChatGPTの開発企業であるOpenAIは，これらの基盤モデルを非公開としています．

▶オープンソース…MetaのLlamaなど

基盤LLMをオープンソースとして公開する流れも見られます．もっとも有名なのはMeta社(旧Facebook)の公開するLlamaです．現在最も普及しているのはLlama 2と呼ばれるバージョンで，70億パラメータ，130億パラメータ，700億パラメータという3種類のモデルが公開されています．その他にも，オープンソースのモデルとしてMistral，Gemma，Solarなど内部構造や学習に用いるデータを改善したモデルが