第1章 文章生成／要約／抽出／コード生成

ローカルLLMを動かしてみる

佐々木 峻

表1　主要なローカルLLM

モデルの種類	開発元	利用可能パラメータ数（Bはbillion:10億）	特　徴
Swallow	東京工業大学	7B，13B，70B	日本語に特化
Japanese StableLM	StabilityAI	1.6B，3B，7B，70B	日本語に特化
Llama3	Meta	8B，70B	英語で高精度

LLM（Large Language Models，大規模言語モデル）はChatGPTのようなウェブ経由のサービスだけでなく，ローカルで動かすことができるもの（ローカルLLM）も数多くリリースされています．

ローカルLLMは，クラウド・サービスで提供される従来のLLMと異なり次の利点があります．

- カスタマイズしやすい
- サービス提供元に学習データとして利用されない
- 高速処理
- 通信コストを抑えられる

本稿では，日本語対応のローカルLLM（Stability AI）を使って基本的な文章生成を試します．単純なQ＆Aではなく，要約，キーワード抽出，コード生成などを行い，ローカルLLMでもかなり自然な日本語を生成できることを示します．（編集部）

ローカルLLMもChatGPTにせまる精度に進化中

● ローカルLLMの今

ChatGPTが2022年11月に発表されてから，自然な文章を生成できるLLMについて研究が大きく進みました．その過程で現在に至るまで，ChatGPTのようなサービスから扱うLLMだけではなく，モデルとしてローカルで動かすことができるLLMも多くリリースされています．最近ではChatGPTに近しい精度のモデルが配布されたり，マルチモーダルに対応したモデルも発表されたりすることで，よりローカルLLMでできることの幅が広がっています．

● ローカルLLMの選択基準

今ではさまざまな企業，研究機関，もしくは個人まで自作のローカルLLMをリリースしています．多くの種類があり，それぞれに得意な言語やタスクがあります（**表1**）．この中から使用するモデルを選ぶときに何を見て選べばよいか，幾つか軸があるので説明します．

▶①パラメータ数

ローカルLLMではパラメータ数がモデル・サイズとして使用されます．例えば，7Bというのはパラメータ数が70億という意味です．一般的にはパラメータ数が多ければ多いほど性能が高いとされています．ただし，多くのパラメータを扱うためには，多くのメモリが必要になるので，自分が持っている環境を踏まえて決める必要があります．

▶②扱う言語

そのモデルが扱える言語にはどんなものがあるかも重要な要素です．例えば，2024年に発表されたLlama 3というモデルは日本語を出力することはできますが，英語の方が出力の精度は高いと言われています．従って，日本語をメインで扱いたい場合は，Swallowモデルや，Japanese StableLM Alphaなど，日本語特化のモデルの方が適しているでしょう．

▶③ファイン・チューニング済みかどうか

配布されているモデルには，ベースのモデルから既に特定のタスクに対してチューニングされたものもあります．例えば，Japanese StableLM Alphaというモデルは Japanese-StableLM-Base-Alpha-7Bという名前でベースのモデル（ファイン・チューニングされていないもの）が配布されていますが，それとは別に，指示応答に特化してチューニングされたjapanese-stablelm-instruct-alpha-7b-v2というモデルも配布されています．チャットボットのようなユースケースでは後者がよく使用されます．

▶④ライセンス

用途によってはライセンスも重要です．商用利用可能なモデルもあれば，条件付きで商用利用可能など，さまざまな条件があります．例えば，Llama 3では基