

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みます。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 が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せんが、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