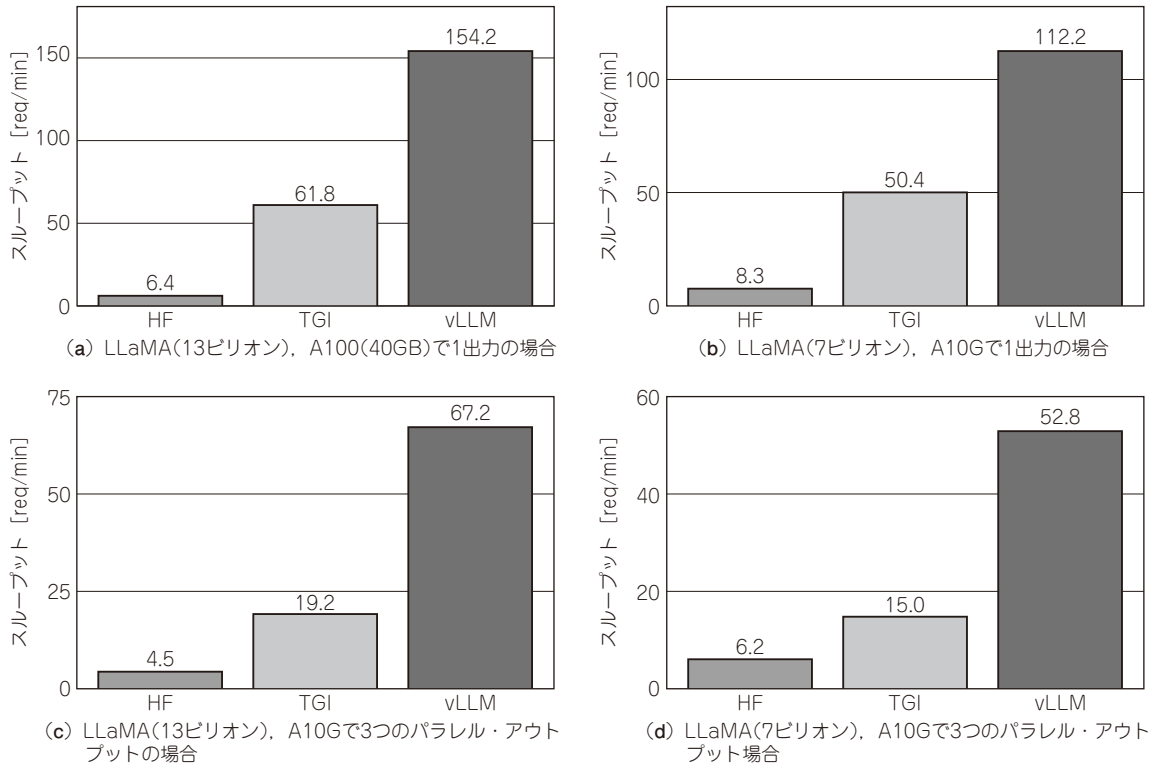
第5章　便利機能も充実！数行のコードで手軽に導入できる

最大24倍のスループット向上！LLM推論高速化ライブラリvLLM

山本 大輝



（a）LLaMA（13ビリオン），A100（40GB）で1出力の場合

（b）LLaMA（7ビリオン），A10Gで1出力の場合

（c）LLaMA（13ビリオン），A10Gで3つのパラレル・アウトプットの場合

（d）LLaMA（7ビリオン），A10Gで3つのパラレル・アウトプット場合

図1[(1)]　**vLLMの性能比較**

図中のHFはHuggingface Transformersを示し，TGIはHuggingface Text Generation Inferenceを示す

● vLLM…LLM推論を劇的に高速化するライブラリ

vLLMは，カリフォルニア大学バークレー校の研究者が開発したライブラリです．既存のトランスフォーマ・ライブラリ（Hugging Face）と連携しやすいアプリケーション・プログラミング・インターフェース（API）を備えており，バックエンドでは，推論のスループットを高めるための最適化を行います．vLLM以前はHugging Face Transformers（HF）を利用する場合が多く，推論が非常に低速でした．これに対してvLLMは公式ページによると，他のライブラリと比べて最大24倍のスループット向上が示されています（**図1**）．また，複数のGPU，および，複数のモデルにて性能改善ができていることを確認できています．本稿ではvLLMの核心技術と使い方を紹介します．

● LLM推論の高速化が望まれる理由

近年，LLMの進化は目覚ましく，日々活用されています．AIエージェントによる業務効率化，対話アシスタント，コード支援など，活用範囲は急速に広がっています．一方で，推論，つまりモデルが新しいテキストを生成する処理は遅くなることがあります．

多くのLLMはトランスフォーマ（Transformer）のデコーダ部分のみ使用するデコーダ・モデルを用いま