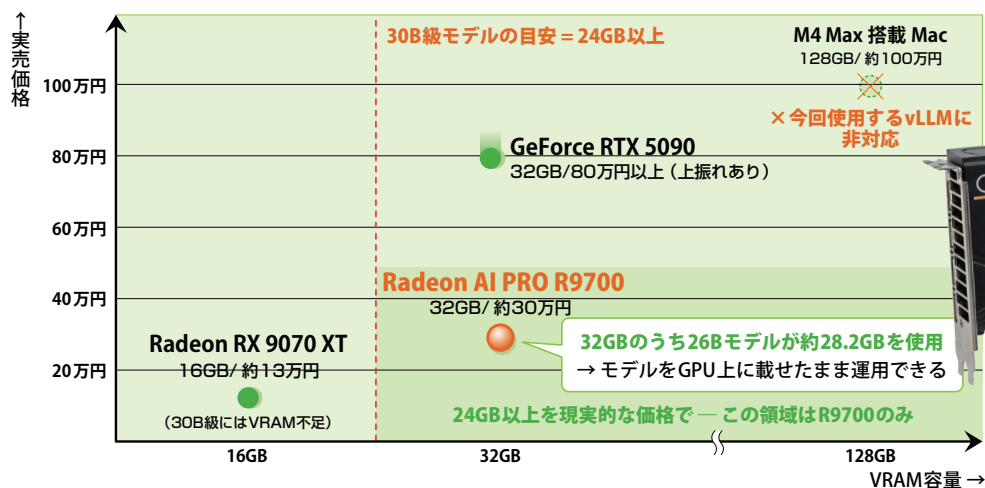


図2 VRAM容量と実売価格 —30B級を狙える現実的な選択肢



※価格はいずれも2026年9月時点の実売。横軸は等間隔ではなく、32GBと128GBの間を波線で省略している



AMD Radeon AI PRO R9700

32GBの大容量VRAM (GDDR6) を搭載する「AMD Radeon AI PRO R9700」。RDNA 4アーキテクチャ採用で、AI推論性能はINT8で766 TOPS、INT4で1,531TOPS (スパース性利用時) に達する

30B級モデルを現実的な予算で動かせるAMD環境

ローカルLLM用PCで最初の壁になるのが、GPUに搭載されるVRAMの容量だ。小規模なモデルは少ないVRAMでも動かせるが、業務で使える回答品質といわれる30B級モデルを選ぶと、量子化した場合でも24GB以上が1つの目安。一般的なGPUでは途端に選択肢が減り、価格も大きく跳ね上がる。

そこで「AMD Radeon AI PRO R9700」の登場だ。本機は32GBのVRAMを搭載しながら、実売価格は約30万円 (2026年9月3日時点での筆者調査)。VRAM 16GBで約13万円のRadeon RX 9070 XTは価格を抑えられるものの、30B級にはVRAMが足りない。VRAM 32GBの競合GPU搭載ビデオカードは80万円以上する。VRAM 128GBのユニファイドメモリを選ぶ他社製ミニPCは約100万円で、今回使用するvLLMにも不向きだ。大容量VRAMと導入価格のバランスを考えると、Radeon AI PRO R9700は中小企業が30B級のローカルLLMを導入する現実的な選択肢になる。

今回のPCでは、Radeon AI PRO R9700に16コア32スレッドのRyzen 9 9950X3D2 Dual Editionを組み合わせた。LLMとRAGを併用する環境では、CPUのコア数だけでなく、ベ

クトル検索で発生するランダムアクセスの速さも効いてくる。Ryzen 9 9950X3D2 Dual Editionは、3D V-Cacheによる192MBの大容量3次キャッシュを備えており、ベクトル検索のスループット向上を期待できる。

サーバーは一度構築すれば、モデルの月額利用料は0円だ。第1回は、Gemma 4 26B-A4B-itの4bit量子化モデルで測定したところ、生成速度は1並列で50.17tok/s、4並列でも合計155.8tok/s (1人あたり約39tok/s) に達した。一般に30tok/sを超えると十分に速く感じられるため、部署単位の利用なら1台でも実用的な応答速度を確保できる。

重要なのは、使いたいモデルをVRAMに収め、複数人からの要求を安定して処理できることだ。今回の26BモデルはVRAMを約28.2GB使用するため、32GBを搭載するRadeon AI PRO R9700ならモデルをGPU上に載せたまま運用できる。大容量VRAMを比較的低い初期費用で確保できることが、Radeon AI PRO R9700を選ぶ最大の理由になる。

また、AMD RadeonによるAI基盤は、価格だけで選ばれているわけではない。GPGPU研究の草分け、濱田剛氏がCEOを務めるDEGIMA AIも、AI基盤にAMD Radeonを採用している。大規模計算を知る専門家が選んだGPUと同じ系統の環境を、中小企業でも導入できることは安心材料となる。



このマシンでローカルLLMサーバーを構築

検証機はRadeon AI PRO R9700 (VRAM 32GB) + Ryzen 9 9950X3D2 Dual Edition、メモリ 64GB、1TB SSDというTSUKUMO eX.computer デスクトップパソコンのカスタマイズ構成。カスタマイズ構成のため、中身は違うものの、Radeon AI PRO R9700搭載機の「Radeon AI PROモデル QA9A-V255/XBH」は、59万9,800円で販売されている。これくらいの価格でモデル利用料0円のローカルLLMサーバーを構築できる。検証機のCPUはハイエンド構成だが、ローカルLLM用途ではより安価なモデルでも構わない