Radeon AI PRO R9700搭載グラフィックスカード
32GBのVRAMを実装しているが、実売価格は25万円前後（2026年5月時点）と、AI向けGPUとしては安価であり、ローカルLLMの初期導入にうってつけ

VRAMが少ないとLLMが使えない場合も
VRAM 16GBのRadeon RX 9070 XTでは、VRAM不足で26Bの「gemma-4-26B-A4B-it-AWQ-4bit」が起動しなかった

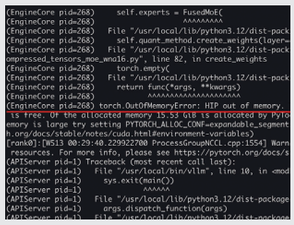

表2 Radeon AI PRO R9700とRadeon RX 9070 XTの違い

	Radeon AI PRO R9700（業務用）	Radeon RX 9070 XT（個人用）
アーキテクチャ	RDNA 4	RDNA 4
製造プロセス	TSMC 4nm	TSMC 4nm
演算ユニット	64CU	64CU
ストリーミングプロセッサ	4,096	4,096
AIアクセラレータ	128基	128基
AI演算性能（INT8）	766TOPS	779TOPS
AI演算性能（INT4）	1,531TOPS	1,557TOPS
メモリ帯域	640GB/s	640GB/s
メモリバス幅	256bit	256bit
VRAM	32GB GDDR6	16GB GDDR6
TBP（Total Board Power）	300W	304W
PCI Express	5.0 x16	5.0 x16
実売価格（2026年5月時点）	25万円前後	11万円前後
30BクラスのLLM	実用的	VRAM不足

メータ数の少ないモデルであれば16GBでも動作するが、その場合は性能面で妥協が生じ、実用的なローカルLLMサーバーが作れない。

Radeon AI PRO R9700の価格は、同じVRAM容量の競合製品と比較した場合でも、半分以下に抑えられており、予算が確保できた段階でもう1枚追加して性能を拡張できるのも大きなメリットだ。最大4枚構成まで対応しているが、4枚実装にはPCI Expressスロットの物理的な配置や電力供給の観点からマザーボードをかなり選ぶ。まずは1枚で運用し、2枚までの増設を目安にするのが現実的だろう。

CPUはL3キャッシュの多い「AMD Ryzen™ X3Dプロセッサ」が良い

GPUほど重視はされないが、ローカルLLMサーバーではCPUも重要だ。ここでは8コア16スレッドで動作する「AMD Ryzen™ 7 9850X3D」を選択している。それはなぜか？

Ryzen 7 9850X3Dはゲーム向けのCPUとされているが、ローカルLLMで着目すべきはそのL3キャッシュの多さ。96MBものL3キャッシュを搭載しており、テキストや画像を数値変換するベクトル検索におけるスループット向上や、RAG（検索拡張生成）におけるレイテンシ抑制に寄与するのである。

ローカルLLMではCPUに対して、コア数よりもランダムアクセス性能が要求されることから、Ryzen 7 9850X3Dのようにキャッシュが大容量のCPUを選ぶことが大切なのだ。

ローカルLLMにはAMD以外の選択肢もあるが……

ローカルLLMを動かす上でのプラットフォームの選択肢は、当然AMD以外もある。ただ、今回のような中小企業向けローカルLLMサーバーを構築するというシチュエーションでは、AMDのRadeon AI PRO R9700を使った方がコスト面での優秀さが光る。たとえば、Radeon AI PRO R9700に対して、VRAM 32GB以上という条件においてNVIDIA、Appleと比較すると**表3**のようになる。

ローカルLLMサーバーにおけるGPU選定においては、VRAM容量以外にも重要な性能指標が2つある。それは「Decode」(注1)と「Prefill」(注2)だ。表3でいうと「メモリ帯域」と「Prefill実測」が関係している。具体例を挙げると、何千行もあるコードを読み込んでAIが「考え始める」までの時間がPrefillに相当する。Prefillが遅い場合、いくらDecodeによる回答速度が速いとしても、その前段階である入力してから回答が始まるまでの「待ち時間」が長くなる。

なお、AIの演算性能を示す単位として、TOPS（Tera Operations Per Second）もあるが、各メーカーで算出基準が統一されていないため、カタログスペックの単純比較が難しい。

以上を踏まえると、約25万円でDecode（回答速度）とPrefill（思考速度）の両方が速く、vLLM対応、かつ比較的省電力というバランスの良さがRadeon AI PRO R9700の強みといえる。ローカルLLMとして導入しやすいという理由を分かっていただけただろうか。

Ryzen 7 9850X3D

表3 Radeon AI PRO R9700とライバルを比較

	AMD Radeon AI PRO R9700	NVIDIA GeForce RTX 5090	Apple M4 Max（メモリ128GB搭載機）
実売価格（2026年5月時点）	約25万円	60万円以上	約90万円
VRAM/メモリ	32GB GDDR6	32GB GDDR7	128GB（統合メモリ）
メモリ帯域	640GB/s	1,792GB/s	546GB/s
Prefill実測	6,252tok/s	13,470tok/s	934tok/s
並列サーバー用途	vLLM	vLLM	MLX中心、vLLMは不向き
消費電力	300W	575W	70W

(注1) Decode：1秒間に何単語相当を出力できるかを示す指標で、単位は「tok/s」（トークン毎秒）など。値が大きいほど回答の表示が速く、ユーザーの待ち時間が短くなる。VRAMのメモリ帯域が性能要因。
(注2) 入力プロンプト全体を一括処理し、最初のトークン（文字）を生成するまでのフェーズ。演算量は入力トークン数に比例し、GPUの純粋な演算性能（AI性能）が関わる要素。