

## ローカルLLM環境構築に関するやり取りのまとめ (2025年6月23日)

AIアシスタントGeminiとの、ローカルLLM環境構築に関する一連の質疑応答の記録です。

## 1. 最初の質問: AWSでの事前検証

### ユーザー:

以下のようなスペックを持つPCを用意、ローカルLLMを構築したいと考えています。(Ollamaなど)

- CPU AMD Ryzen 7 9700X (8コア/16スレッド)
- メモリ DDR5-5600, 32GB
- GPU NVIDIA GeForce RTX 4060 Ti もしくは5060 Ti 16GB
- ストレージ M.2 2TB

です。まずは実機を揃える前にAWSのGPUインスタンスを用いて検証を行いたいのですが、Amazon EC2のP、Gなど、どのインスタンスタイプが相当するでしょうか。コストを最重視するのであまり高スペックは望みませんし、それだと検証の意味がありません。

各インスタンスタイプとの性能比較表みたいなのが作れますか。CUDAやTOPSなどの数値があるとなおよいです。RTX 4060 Ti (16GB) / RTX 5060 Ti (16GB)の他に比較対象のEC2インスタンスが3つくらい候補にあると嬉しいです。あと1時間あたりの価格も、比較的安いアイルランドリージョンで。

### Gemini:

承知いたしました。ご計画中のPCスペックに近い性能を持つAWS EC2インスタンスを、コストを重視しつつ選定し、性能比較表を作成します。

ローカルLLM（Ollamaなど）の実行において最も重要な要素はGPUのVRAM（ビデオメモリ）容量とその性能です。ご提示のNVIDIA GeForce RTX 4060 Ti / 5060 Tiは共にVRAM 16GBを想定されているため、AWSでも同等以上のVRAMを持つインスタンスを選定することが検証の鍵となります。

検証に適したEC2インスタンス候補

コストパフォーマンスと性能のバランスを考慮すると、以下の3つのインスタンスファミリーが候補となります。

- 1. **g4dnインスタンス**: NVIDIA T4 Tensor Core GPUを搭載。VRAMは16GBで、コスト効率に優れており、LLMの推論（Inference）用途で広く使われています。今回の検証のベースラインとして最適です。
- 2. **g5インスタンス**: NVIDIA A10G Tensor Core GPUを搭載。VRAMが24GBあり、より大きなモデルを動かしたり、ファインチューニングを試したりする場合に有利です。
- 3. **g6インスタンス**: NVIDIA L4 Tensor Core GPUを搭載。RTX 4060 Tiと同じAda Lovelaceアーキテクチャを採用しており、電力効率と性能が向上しています。

GPU性能比較表

【注意】

- RTX 5060 Tiは2025年6月現在、未発表の製品です。スペックは業界の予測やリーク情報に基づく**推定値**であり、実際とは大きく異なる可能性があります。

項目	RTX 4060 Ti (16GB)	RTX 5060 Ti (16GB) (推定)	NVIDIA T4 (g4dn)	NVIDIA A10G (g5)	NVIDIA L4 (g6)
アーキテクチャ	Ada Lovelace	Blackwell (推定)	Turing	Ampere	Ada Lovelace
VRAM	16 GB GDDR6	16 GB GDDR7 (推定)	16 GB GDDR6	24 GB GDDR6	24 GB GDDR6
メモリ帯域幅	288 GB/s	512 GB/s (推定)	320 GB/s	600 GB/s	300.3 GB/s
CUDAコア数	4,352	~7,680 (推定)	2,560	9,216	7,424
Tensorコア世代	第4世代	第5世代 (推定)	第2世代	第3世代	第4世代
FP32性能	22 TFLOPS	~37 TFLOPS (推定)	8.1 TFLOPS	31.2 TFLOPS	30.3 TFLOPS
INT8性能 (TOPS)	353 TOPS (Sparsity)	~600 TOPS (推定)	130 TOPS	500 TOPS (Sparsity)	970 TOPS (Sparsity)

PCスペックとEC2インスタンスの比較および料金

ご計画のPCスペック（CPU 8コア/32GBメモリ）に合わせ、各インスタンスタイプでvCPU数が8、メモリが32GiBの 2xlarge サイズを選定しました。料金はアイルランド（eu-west-1）リージョンのものです。

項目	ご計画のPC	g4dn.2xlarge	g5.2xlarge	g6.2xlarge
CPU / vCPU	AMD Ryzen 7 9700X (8コア/16スレッド)	8 vCPU (Intel Xeon Cascade Lake)	8 vCPU (AMD EPYC 2nd Gen)	8 vCPU (AMD EPYC 3rd Gen)
メモリ	32 GB DDR5	32 GiB	32 GiB	32 GiB
GPU	RTX 4060 Ti / 5060 Ti	1x NVIDIA T4	1x NVIDIA A10G	1x NVIDIA L4
GPUメモリ	16 GB	16 GiB	24 GiB	24 GiB
時間あたり料金 (アイルランド)	-	約 \$0.75 USD	約 \$1.21 USD	約 \$1.26 USD

結論と推奨

コスト最優先で同等環境を試したい場合は g4dn.2xlarge が最適です。VRAMが16GBであり、「16GBのVRAMでどの程度のことができるか」を把握するための最初のステップとして推奨します。

2. Azureでの検証について

**ユーザー:**  
Azure Linux VMだとどうなります？ Azureのことはほとんど何も知らないので軽く聞いてみるのですが。一番近い後方があればそれを一つと、利用料金も。

**Gemini:**  
はい、承知いたしました。 AzureにもAWSと非常によく似た、今回の検証目的に最適なVM（仮想マシン）が存在します。

Azureで最も近い候補： NCas\_T4\_v3 シリーズ

AWSの g4dn インスタンスと同じNVIDIA T4 GPUを搭載した NCas\_T4\_v3 シリーズが最適な候補です。

項目	Azure候補: Standard_NC8as_T4_v3
VMシリーズ	NCas_T4_v3
GPU	1x NVIDIA T4
GPUメモリ	16 GiB
vCPU	8 vCPU (AMD EPYC 7V73X "Milan")
メモリ	56 GiB
リージョン	西ヨーロッパ (West Europe)
時間あたり料金 (Linux)	約 \$0.84 USD

## AWSインスタンスとの比較と解説

AWSの `g4dn.2xlarge` と並べると、GPU性能は同等ですが、Azureの候補はメモリが多く、料金は若干高めです。純粋なコストを重視するなら、今回のケースではAWSに若干の分があると言えるでしょう。

## 3. 推奨LLMモデルについて

ユーザー:

今度は検証すべきというかオススメのLLMモデルがあったらそいつもお願いします。

Gemini:

もちろんです。16GBのVRAMを搭載したPCで快適に動作し、かつ性能評価のベンチマークとなるような、検証におすすめのLLMモデルをいくつかご紹介します。

### 1. まずはコレ！絶対に試すべき鉄板・高性能モデル (7Bサイズ)

- Llama 3 ( `ollama run llama3:8b-instruct` ): このサイズでは総合性能で世界最高峰。
- Qwen 2 ( `ollama run qwen2:7b-instruct` ): 多言語対応に優れ、日本語の扱いも比較的得意。
- Gemma ( `ollama run gemma:7b-instruct` ): Google開発の堅実な性能を持つモデル。

### 2. 日本語の精度を試すならコレ！(日本語特化モデル)

- Japanese StableLM Gamma ( `ollama run japanese-stablelm-instruct-gamma-7b` ): 比較的新しく高性能な日本語モデル。
- Rinna / Yuri ( `ollama run rinna/youri-7b-chat` ): 非常に自然で親しみやすい日本語の対話が可能。

### 3. コーディング支援の実力を試すならコレ！

- DeepSeek Coder V2 Lite ( `ollama run deepseek-coder-v2:16b-lite-instruct` ): コーディング分野で現在最強と名高いモデル。
- Code Llama ( `ollama run codellama:13b-instruct` ): コーディング特化モデルの定番。

#### 4. VRAM 16GBの限界に挑戦！(高性能・大規模モデル)

- **Mixtral 8x7B** (`ollama run mixtral:8x7b-instruct-q4_K_M`): 47Bクラスの性能を持つMoEモデル。16GB VRAMでもギリギリ動作を狙えます。

**ユーザー:**

MicrosoftのPhi-3は評判がいいみたいなんですけどこいつはどうでしょう。

**Gemini:**

はい、非常によいご質問です。Phi-3は「小さいのに驚くほど賢い」として、現在非常に高く評価されているモデルです。

#### Phi-3のすごいところ

Phi-3が注目される最大の理由は、その圧倒的なコストパフォーマンス（性能対サイズ比）にあります。

1. **軽量・高速**: Phi-3-miniは3.8B（38億）パラメータという小ささにも関わらず、ベンチマークスコアではLlama 3の8BモデルやMixtral 8x7Bのような、自身より2倍以上大きいモデルに匹敵する性能を叩き出します。これにより、16GB VRAMの環境では極めて軽快・高速に動作します。
2. **革命的な長文対応 (128Kコンテキスト)**: Phi-3-miniには、128Kトークンという非常に長い文章を一度に読み込めるバージョン (phi3:mini:128k) があります。これは約10万語に相当し、長いPDFの要約や、プロジェクト全体のコードを読み込ませて質問する、といったローカル環境では従来不可能だったタスクに挑戦できる可能性を秘めています。

あなたのPC環境におすすめのPhi-3モデル

Phi-3ファミリーの中から、ぜひ試していただきたいモデルは以下の通りです。

モデル名	Ollamaコマンド	特徴・おすすめポイント
Phi-3 Mini (3.8B)	ollama run phi3:mini	【超軽量・超高速】 まずはこのモデルから。7Bクラスの性能を、驚くほど軽快に体験できます。日常的なチャットや簡単なタスクなら、これ以上ないほど快適です。
Phi-3 Mini (128K)	ollama run phi3:mini:128k	【長文読解の革命児】 Phi-3の真価が分かる長文対応版。長い論文や契約書を読み込ませて要約・質疑応答させるなど、このモデルでしかできない体験ができます。VRAM使用量は増えますが、試す価値は絶大です。
Phi-3 Small (7B)	ollama run phi3:small	【高性能7Bモデル】 より性能を追求した7Bパラメータ版。Llama 3 8BやQwen 2 7Bと真っ向から性能を比較するのに最適なライバルです。
Phi-3 Medium (14B)	ollama run phi3:medium	【限界への挑戦】 14Bパラメータの最上位版。16GB VRAMでは量子化を駆使してギリギリ動くラインですが、Phi-3ファミリー最高の性能を体験できます。

Phi-3の日本語能力について

一点、注意点があります。Phi-3は多言語に対応していますが、学習データは英語が中心です。そのため、日本語の応答は可能ですが、日本語特化モデル（Japanese StableLM など）と比較すると、不自然な表現になったり、専門的な内容で精度が落ちたりすることがあります。

結論：Phi-3は試すべきか？

間違いなく「試すべき」です。

日本語の完璧さでは専用モデルに一步譲りますが、それを補って余りある「**効率の良さ**」と「**長文対応能力**」というユニークな強みを持っています。特に `phi3:mini:128k` は、これまでのモデルではできなかった新しい使い方を可能にしてくれる、非常にエキサイティングなモデルです。ぜひ、先ほど紹介したLlama 3や日本語モデルと並行してPhi-3も試し、「タスクによってモデルを使い分ける」という、ローカルLLMの醍醐味を体験してみてください。