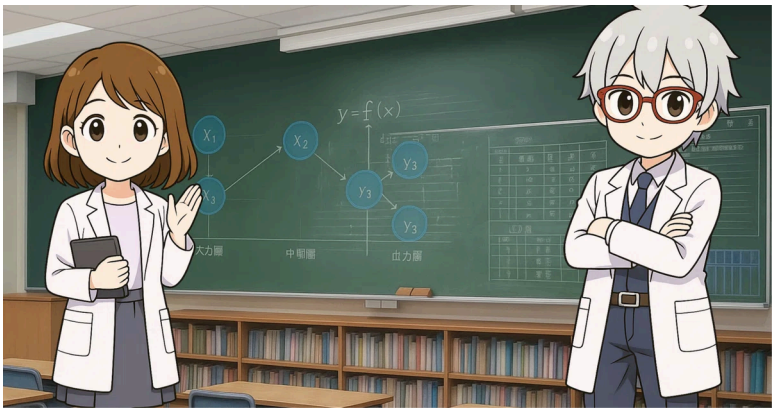

【軽量モデル】token/sec 比較

れん
2025年8月6日 11:47

ローカルLLMでいろいやってみたくて確認してみた

スクリプト例：全モデルの tokens/sec をテスト

- すべてのモデルを順番に計測
- stream=True で1トークンごとにカウント
- 生成長は num_predict で統一
- エラーが出たモデルもスキップして結果に記録
- 最後にまとめて結果を表示

qwen3-thinking-q4:latest
⇒ TPS：10.99 / VRAM：13,832MB / CPU：5,674MB
qwen3-thinking-q3:latest
⇒ TPS：18.36 / VRAM：14,103MB / CPU：1,777MB
qwen3-instruct-q4:latest
⇒ TPS：11.49 / VRAM：13,845MB / CPU：5,658MB
qwen3-instruct-q3:latest
⇒ TPS：24.27 / VRAM：13,397MB / CPU：894MB
Qwen3-30B-A3B-Instruct-2507:latest
⇒ TPS：10.39 / VRAM：14,026MB / CPU：5,819MB
qwen3-coder:14b
⇒ TPS：17.19 / VRAM：9,630MB / CPU：1,112MB
llama3-elyza-jp-8b:latest
⇒ TPS：29.12 / VRAM：5,374MB / CPU：959MB
openhermes:latest
⇒ TPS：32.78 / VRAM：4,816MB / CPU：644MB
deepseek-coder:6.7b
⇒ TPS：34.13 / VRAM：6086MB / CPU：657MB
tiger-gemma:latest
⇒ TPS：23.06 / VRAM：7,506MB / CPU：1,372MB
amoral-gemma3:latest
⇒ TPS：49.58 / VRAM：2,932MB / CPU：1,140MB
phi-4-deepseek-R1K-RL-EZO-GGUF:Q4_K_S
⇒ TPS：14.59 / VRAM：9,106MB / CPU：897MB
DeepSeek-R1-Distill-Qwen-Japanese:14b
⇒ TPS：15.81 / VRAM：9,454MB / CPU：1,085MB
calm3-22b-rp:q5_k_m
⇒ TPS：3.62 / VRAM：13,882MB / CPU：7,378MB
command-r7b:latest

⇒ TPS：24.59 / VRAM：6,000MB / CPU：1,535MB
llm-jp-3-ezo-humanities:3.7b-instruct-q8_0
⇒ TPS：35.32 / VRAM：5,250MB / CPU：885MB
deepseek-r1:14b
⇒ TPS：17.52 / VRAM：9,454MB / CPU：1,105MB
deepseek-r1:1.5b
⇒ TPS：80.09 / VRAM：1,506MB / CPU：751MB

トークン生成速度ランキング（tokens/sec）

model	token/sec		CPU memory		VRAM:	
deepseek-r1:1.5b	TPS:	80.09	Mem:	751	VRAM:	1,506
amoral-gemma3:latest	TPS:	49.58	Mem:	1,140	VRAM:	2,932
llm-jp-3-ezo-humanities:3.7b-instruct-q8_0	TPS:	35.32	Mem:	885	VRAM:	5,250
deepseek-coder:6.7b	TPS:	34.13	Mem:	657	VRAM:	6,086
openhermes:latest	TPS:	32.78	Mem:	644	VRAM:	4,816
llama3-elyza-jp-8b:latest	TPS:	29.12	Mem:	959	VRAM:	5,374
command-r7b:latest	TPS:	24.59	Mem:	1,535	VRAM:	6,000
qwen3-instruct-q3:latest	TPS:	24.27	Mem:	894	VRAM:	13,397
tiger-gemma:latest	TPS:	23.06	Mem:	1,372	VRAM:	7,506
qwen3-thinking-q3:latest	TPS:	18.36	Mem:	1,777	VRAM:	14,103
deepseek-r1:14b	TPS:	17.52	Mem:	1,105	VRAM:	9,454
qwen3-coder:14b	TPS:	17.19	Mem:	1,112	VRAM:	9,630
DeepSeek-R1-Distill-Qwen-Japanese:14b	TPS:	15.81	Mem:	1,085	VRAM:	9,454
phi-4-deepseek-R1K-RL-EZO-GGUF:Q4_K_S	TPS:	14.59	Mem:	897	VRAM:	9,106
qwen3-instruct-q4:latest	TPS:	11.49	Mem:	5,658	VRAM:	13,845
qwen3-thinking-q4:latest	TPS:	10.99	Mem:	5,674	VRAM:	13,832
Qwen3-30B-A3B-Instruct-2507:latest	TPS:	10.39	Mem:	5,819	VRAM:	14,026
calm3-22b-rp:q5_k_m	TPS:	3.62	Mem:	7,378	VRAM:	13,882

※ 赤地の部分はCPUのメモリも利用しています

分析ポイント

- **最速モデル**：deepseek-r1:1.5b（軽量で非常に高速）
- **性能と速度のバランス**：
  - amoral-gemma3:latest（TPS 41.8 / 2.5GB）
  - llm-jp-3-ezo-humanities（TPS 34.4 / 4.0GB）
- **大型モデル（30B, 22B）は遅い**：
  - Qwen3-30B（TPS 8.1）
  - calm3-22b（TPS 3.7）
- **中規模モデル（6.7B〜14B）は15〜30 TPS**で安定

日本語対応モデル候補

以下のモデルは、日本語対応が強い、または日本語特化のチューニングがされています：

- **llama3-elyza-jp-8b** → 日本語専用
- **llm-jp-3-ezo-humanities:3.7b** → 日本語特化
- **amoral-gemma3:latest** → 日本語対応（Gemma系）
- **deepseek系 (1.5b, 6.7b, 14b)** → 日本語対応良好
- **Qwen系 (14b, 30b)** → 日本語強いが速度が遅い

検索モデル

- 検索向き（情報収集や外部ツールで使いやすい）のは、**DeepSeek系** や **Qwen系**
- 小型・中型で速いのは deepseek-r1:1.5b（検索に最適）

ルミナ / クラリス / ノクス を起動しよう

✅ ルミナ（ファシリテータ／雑談）

- **モデル**: llm-jp-3-ezo-humanities:3.7b：（5,300MB）
- **理由**: 日本語会話特化・速度34 TPSで軽量

✅ クラリス（深掘り解説）

- **モデル**: amoral-gemma3:latest（3,000MB）
- **理由**: 知識量と速度バランスが良く、日本語対応も安定

✅ ノクス（情報検索役）

- **モデル**: deepseek-r1:1.5b：（1,500MB）
- **理由**: 検索・情報収集用に最速モデルを採用

これならもう一個くらいLLM動かせますね