

ローカルLLM 24モデル比較

MacBook Pro M1 Max・ 64GB向け | 量子化・ メモリ・ ベンチマーク・ 用途を1枚で確認

量子化列は【計算値】

重みだけの理論容量 (GB)。実使用RAM・ 実ファイル容量ではありません。
全モデルのM1 Max実測RAM・ 速度は【未検証】。

ベンチマークは公開資料の値

量子化後やM1 Maxで測った成績ではありません。設定・ 評価条件は未統一。
「-」は今回未確認。モデル名とスコアをクリックすると出典へ移動します。

FLASH-NEXTを収録：No.08

Qwen3.8-Flash-Next：125B + 埋込51B + MTP4B、稼働6B。
「6B」は必要RAM 6GBを意味しません。No.20のQwen3-Nextとは別です。

確認した仕様・ 対応入力					量子化別メモリ基礎容量【計算値】 GB / 重みのみ・ 言語部主体							公開ベンチマーク / 基本は%・ 高いほど良い			用途・ 配布条件	
No.	モデル名 / 公式出典リンク	構造・ パラメータ数 B この数値を容量計算に使用	対応入力 文 / 画 / 音・ 動画	コンテキスト トークン / 拡張は別	2bit	3bit	4bit	5bit	6bit	8bit	FP16 BF16	GPQA Diamond	コード LCB v6	その他の指標 名称・ 条件を併記	向いている用途【提案】 日本語・ WP・ SEO実務は未検証	配布・ 留意点 動作保証ではありません
01	Gemma 4 31B Apache 2.0	31B級・ 言語30.7B Dense 画像encoder別	文・ 画・ 動画	256K	7.7	11.5	15.4	19.2	23.0	30.7	61.4	84.3	80.0	MMLU-Pro 85.2	SEO記事・ 分析	公開重み M1実機：未検証
02	Gemma 4 26B A4B Apache 2.0	26B級・ 言語25.2B MoE 稼働3.8B・ 画像encoder別	文・ 画・ 動画	256K	6.3	9.5	12.6	15.8	18.9	25.2	50.4	82.3	77.1	MMLU-Pro 82.6	コード補助・ 文章	公開重み M1実機：未検証
03	Gemma 4 12B Unified Apache 2.0	11.95B Dense	文・ 画・ 音・ 動画	256K	3.0	4.5	6.0	7.5	9.0	12.0	23.9	78.8	72.0	MMLU-Pro 77.2	軽量処理・ 文章	公開重み M1実機：未検証
04	Gemma 2 27B Gemma規約	27B級 Dense ベンチはPT版・ 対話版と別	文	8,192	6.8	10.1	13.5	16.9	20.3	27.0	54.0	-	-	MMLU 75.2 [PT]	文章生成・ 比較用	公開重み M1実機：未検証
05	Qwen3.8-27B Apache 2.0	言語27B Dense 画像部・ MTP等は別	文・ 画・ 動画	262,144 拡張1,000,000	6.8	10.1	13.5	16.9	20.3	27.0	54.0	89.2	90.3	SWE-Pro 61.7	コーディング・ SEO	公開重み M1実機：未検証
06	Qwen3.6-35B-A3B Apache 2.0	言語35B / 稼働3B MoE 画像部・ MTP等は別	文・ 画・ 動画	262,144 拡張1,010,000	8.8	13.1	17.5	21.9	26.3	35.0	70.0	86.0	80.4	MMLU-Pro 85.2	AIエージェント・ 分析	公開重み M1実機：未検証
07	Qwen3-Coder-30B-A3B-Instruct Apache 2.0	総30.5B / 稼働3.3B MoE Non-thinking	文	262,144 拡張1,000,000	7.6	11.4	15.3	19.1	22.9	30.5	61.0	-	-	未確認	WordPress・ PHP・ JS	公開重み M1実機：未検証
08	Qwen3.8-Flash-Next FLASH-NEXT / Qwen Community 1.0	125B + 51B + 4B MoE 稼働6B / 埋込51B・ MTP4B	文・ 画・ 動画	262,144 拡張1,000,000	45.0	67.5	90.0	112.5	135.0	180.0	360.0	91.7	91.9	SWE-Pro 62.5	高度な推論・ 自動化	公開重み M1実機：未検証
09	DeepSeek-R1-Distill-Qwen-32B MIT	32B級 Dense	文	131,072 配布config上限	8.0	12.0	16.0	20.0	24.0	32.0	64.0	62.1	-	LCB 57.2 [版不明]	推論・ 問題解決	公開重み M1実機：未検証
10	Llama 4 Scout 17B-16E Instruct Llama 4規約	総109B / 稼働17B MoE 16 experts	文・ 画	10M	27.3	40.9	54.5	68.1	81.8	109.0	218.0	57.2	-	MMLU-Pro 74.3	長文処理・ 分析	公開重み M1実機：未検証
11	Mistral Medium 3 契約条件要確認	【要確認】 数値・ 構造未確認	マルチモーダル	128K	要確認	要確認	要確認	要確認	要確認	要確認	要確認	-	-	未確認	汎用・ 文章生成	一般公開重み未確認 API中心・ 非推奨
12	Gemma 4 E2B Apache 2.0	言語5.1B / 有効2.3B Dense 画像・ 音声encoder別	文・ 画・ 音・ 動画	128K	1.3	1.9	2.6	3.2	3.8	5.1	10.2	43.4	44.0	MMLU-Pro 60.0	短文・ 音声入力の試作	公開重み M1実機：未検証
13	Gemma 4 E4B Apache 2.0	言語8B / 有効4.5B Dense 画像・ 音声encoder別	文・ 画・ 音・ 動画	128K	2.0	3.0	4.0	5.0	6.0	8.0	16.0	58.6	52.0	MMLU-Pro 69.4	軽量の要約・ 音声入力	公開重み M1実機：未検証
14	Qwen3.5-0.8B Apache 2.0	言語0.8B Dense 画像部・ MTP等は別	文・ 画・ 動画	262,144	0.2	0.3	0.4	0.5	0.6	0.8	1.6	-	-	MMLU-Pro 42.3 [T]	小規模な分類・ 抽出の試作	公開重み M1実機：未検証
15	Qwen3.5-2B Apache 2.0	言語2B Dense 画像部・ MTP等は別	文・ 画・ 動画	262,144	0.5	0.8	1.0	1.3	1.5	2.0	4.0	-	-	MMLU-Pro 66.5 [T]	定型文・ 分類・ 要約の試作	公開重み M1実機：未検証
16	Qwen3.5-4B Apache 2.0	言語4B Dense 画像部・ MTP等は別	文・ 画・ 動画	262,144 拡張1,010,000	1.0	1.5	2.0	2.5	3.0	4.0	8.0	76.2	55.8	MMLU-Pro 79.1	軽量の文章・ 画像分析	公開重み M1実機：未検証
17	Qwen3.5-9B Apache 2.0	言語9B Dense 画像部・ MTP等は別	文・ 画・ 動画	262,144 拡張1,010,000	2.3	3.4	4.5	5.6	6.8	9.0	18.0	81.7	65.6	MMLU-Pro 82.5	文章・ 画像・ コードの試作	公開重み M1実機：未検証
18	Qwen3.6-27B Apache 2.0	言語27B Dense 画像部・ MTP等は別	文・ 画・ 動画	262,144 拡張1,010,000	6.8	10.1	13.5	16.9	20.3	27.0	54.0	87.8	83.9	MMLU-Pro 86.2	コード・ 分析の比較用	公開重み M1実機：未検証
19	Qwen3.5-122B-A10B Apache 2.0	言語122B / 稼働10B MoE 画像部・ MTP等は別	文・ 画・ 動画	262,144 拡張1,010,000	30.5	45.8	61.0	76.3	91.5	122.0	244.0	86.6	78.9	MMLU-Pro 86.7	大型モデルの分析比較	公開重み M1実機：未検証
20	Qwen3-Next-80B-A3B-Thinking Apache 2.0	総80B / 稼働3B MoE Flash-Nextとは別モデル	文	262,144 拡張1,010,000	20.0	30.0	40.0	50.0	60.0	80.0	160.0	77.2	68.7	MMLU-Pro 82.7	複雑な推論・ 長文の比較	公開重み M1実機：未検証
21	Qwen3-Coder-Next Apache 2.0	総80B / 稼働3B MoE 下欄に量子化別の外部実測	文	262,144	20.0	30.0	40.0	50.0	60.0	80.0	160.0	-	-	SWE-v 70.6 [HF]	コード修正・ 開発支援	公開重み M1実機：未検証
22	DeepSeek-R1-Distill-Qwen-14B MIT	14B級 Dense	文	131,072 配布config上限	3.5	5.3	7.0	8.8	10.5	14.0	28.0	59.1	-	LCB 53.1 [版不明]	推論・ 数学の軽量比較	公開重み M1実機：未検証
23	Mistral Small 3.2 24B Instruct Apache 2.0	24B級 Dense	文・ 画	128K	6.0	9.0	12.0	15.0	18.0	24.0	48.0	46.13	-	MMLU-Pro 69.06 [5s]	文章・ 画像・ 構造化出力	公開重み M1実機：未検証
24	Mistral Small 4 119B A6B Apache 2.0	総119B / 稼働6.5B MoE GPQAはHF評価欄 [HF]	文・ 画	256K	29.8	44.6	59.5	74.4	89.3	119.0	238.0	71.2	-	AA-LCR 0.72	大型の推論・ 開発比較	公開重み M1実機：未検証

量子化・ ベンチマークの読み方

容量の式：P (10億) × bit ÷ 8 = GB。小数1桁。Pはパラメータ列の言語部・ 総数が基準。MoEは総数で計算。橙色は理論量64GB
以上で、動作判定ではありません。

【未検証】実RAMは上の理論量と異なります。量子化の補助情報、未量子化の層、画像・ 音声部、KVキャッシュ、一時領域、
macOS等が別途必要です。

2/3/5/6bit等の配布・ エンジン対応を確認した表ではありません。同じ4bitでもGGUF・ MLX・ AWQ等で容量・ 品質・ 速度が変わります。

GPQA-D = 科学的推論、LCB v6 = コード課題。その他の指標は名称が違えば横比較不可。SWE-Pro / SWE-vも別指標。PT = 事前
学習版、T = Thinking、5s = 5-shot CoT、HF = HF評価欄の掲載値。

最大コンテキストは64GBでの実用上限ではありません。用途は提案で、日本語・ WordPress・ SEOでの優劣は未検証。元11件の
用途は継承しています。

Google公式のメモリ概算 / 理論量とは別の値

GPU/TPUへの読み込み概算・ GB。M1 Maxの実測ではありません。

Gemma 4	BF16	SFP8	Q4_0
E2B	11.4	5.7	2.9
E4B	17.9	8.9	4.5
12B	26.7	13.4	6.7
26B A4B	57.7	28.8	14.4
31B	69.9	34.9	17.5

公式表は読み込み追加分20%を考慮。KVキャッシュ・ 補助ソフト等は別途。
出典：Google Gemma 4 model overview

M1 Max 64GB / 量子化ごとの外部実測報告

Qwen3-Coder-Next / 投稿者による各1回・ 128問の結果。再現は【未検証】。

量子化方式	サイズGB*	平均合格率	生成tok/s
MLX 3bit	34.90	60.4%	23.4
MLX 4bit	44.86	63.3%	24.8
GGUF UD-IQ3_XXS	32.71	61.3%	23.66
GGUF Q4_K_M	48.73	61.7%	19.93

* 投稿のSize欄を転記。実RAM使用量ではありません。
LiveBench coding 2024-11-25。上表のLCB v6とは別評価です。
llama.cpp f93c09e26 / oMLX 0.3.0。投稿・ 条件・ 結果を確認