LOCAL LLM WORKSTATION · MEMORY FIT · MODEL ROUTING

MacBook M5 Max128GB 로컬 LLM최적 모델 분석

핵심 결론

01 **최적 균형은 Qwen3.5-122B-A10B 4-bit다. 멀티모달·다국어·에이전트 능력과 M5 Max 실측 속도를 함께 만족한다.**

02 **MLX 4-bit 저장소는 약 69.6GB이고 32K 실측 피크는 76.4GB였다. 128GB에서 약 51.6GB가 남는다.**

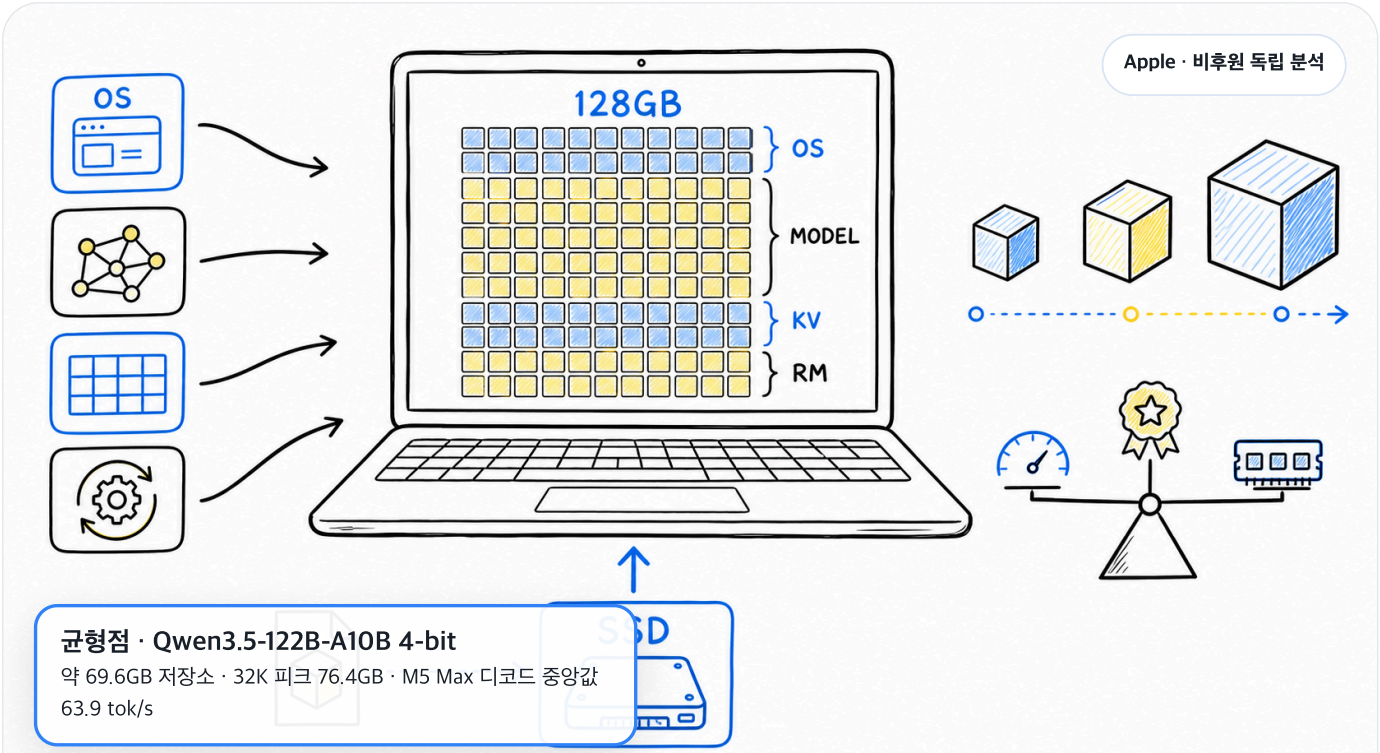
03 **32K 컨텍스트부터 시작해 프리필·디코드·메모리 압력을 따로 측정하라. SSD는 보관함이지 통합 메모리의 대체재가 아니다.**

128GB는 “가장 큰 모델 하나”보다 “품질·속도·여유 메모리의 균형”에서 가치가 큽니다. 총 파라미터와 활성 파라미터, 양자화 파일, KV 캐시와 런타임 여유를 분리해 실제로 쓸 만한 구성을 골랐습니다.

260803 · 표준 심층 · 기술 Executive Brief | 2026.08.03 KST

Apple 공식 사양 · 공식 모델 카드 · HF 파일 트리 · M5 Max MLX 110회 실측

실제 제품사진 · M5 Pro 또는 M5 Max 탑재 MacBook Pro · 화면: Xcode·LM Studio · 제공: Apple Newsroom 미디어 자산 · [보도자료 원문] · 본 보고서는 Apple의 승인·후원·제휴를 받지 않은 독립 분석입니다.



AI 생성 인포그래픽 · 공식 제품 이미지 아님

128GB 통합 메모리는 CPU와 GPU가 함께 쓰는 하나의 공간이다. 정확한 모델명·용량·판단 수치는 HTML 표에서 검증했다.

128GB의 정답은 “122B 4-bit + 50GB대 여유”입니다

큰 모델을 간신히 올리는 것보다, 32K 컨텍스트·IDE·브라우저·에이전트까지 함께 유지하는 구성이 실용적입니다.

종합 1위 · Qwen3.5-122B-A10B 4-bit

122B 전체 파라미터 중 토큰마다 약 10B가 활성화되는 MoE다. 4-bit MLX 저장소 약 69.6GB, M5 Max 실측 피크 69.8~76.4GB로 128GB에서 운영 여유를 남긴다. 공식 카드상 멀티모달·201개 언어·262K 기본 컨텍스트를 지원하지만, 로컬 운용은 32K부터 시작하는 편이 안전하다.

공식 사양실측 기반 권고속도 보장 아님
63.9
decode tok/s 중앙값
M5 Max · MLX 4-bit · 11개 실측. 52.3~67.2 tok/s 범위이며 프롬프트·디코드 길이에 따라 달라졌다.

128GB
CPU·GPU가 공유하는 통합 메모리
614GB/s
40코어 GPU M5 Max 공식 대역폭
69.6GB
Qwen 122B MLX 4-bit 저장소 총량
51.6GB
32K 피크 76.4GB 사용 후 명목 여유

모델별로 “파일이 들어감”과 “쓸 만함”을 분리했습니다

실측은 Creative Strategies가 공개한 M5 Max·MLX 110회 데이터 중 동일 모델 11회씩을 집계했습니다. 해당 글은 정확한 메모리·칩 bin을 명시하지 않아 방향성 근거로 사용합니다.

모델	구조·컨텍스트	권장 양자·런타임	실제 크기·피크	M5 Max 디코드 실측	128GB 판정	주 용도·주의
Qwen3.5-122B-A10B 종합 1위	MoE 122B/활성 10B 멀티모달 · 기본 262K	MLX 4-bit 품질 우선은 5-bit	4-bit 저장소 69.6GB 피크 69.8~76.4GB 5-bit 저장소 84.9GB	**중앙 63.9 tok/s** 52.3~67.2	**편안함** 32K 시작 권장	한국어·문서·이미지·에이전트 전반. 5-bit는 동일 조건 실측 없음.
gpt-oss-120b 추론 대안	MoE 117B/활성 5.1B 텍스트 · 131K	MXFP4 · MLX/llama.cpp	GGUF 63.4GB 피크 62.8~64.4GB	**중앙 99.3 tok/s** 70.1~102.3	**편안함**	영문 추론·도구 사용. 비전 미지원, 한국어는 Qwen과 직접 A/B 필요.
Qwen3-Coder-Next 코딩 1위	MoE 80B/활성 3B 256K	MLX 6-bit 64.8GB GGUF Q6 65.6GB	약 65GB 명목 여유 약 63GB	동일 공개 M5 Max 실측 없음	**편안함**	코드베이스·CLI 에이전트. 32K에서 시작해 도구 호출 형식을 검증.
Qwen3.5-35B-A3B 속도 1위	MoE 35B/활성 3B 멀티모달 · 262K	MLX 4-bit 또는 8-bit	Q8 GGUF 36.9GB 4-bit 피크 20.2~24.1GB	**중앙 141.9 tok/s** 112.1~145.1	**매우 편안함**	빠른 채팅·보조 에이전트·동시 앱. 122B보다 지식·추론 품질은 과제별 검증.

모델	구조· 컨텍스트	권장 양자· 런타임	실제 크기·피크	M5 Max 디코드 실측	128GB 판정	주 용도·주의
GLM-4.7-Flash 경량 에이전트	MoE 30B/ 활성 3B 약 203K	MLX 6-bit 또는 8-bit	Q8 GGUF 31.8GB 6-bit 피크 24.8~35.2GB	**중앙 74.8 tok/s** 40.8~79.2	**매우 편안함**	도구·에이전트·추론. 한국어와 장문 안정성은 별도 평가.
Qwen3.5-27B	Dense 27B 멀티모달 · 262K	MLX 4/6/8-bit	Q8 GGUF 28.6GB 4-bit 피크 16.3~23.3GB	**중앙 30.3 tok/s** 19.5~32.9	**편안함**	Dense라 작은 파일이어도 35B MoE보다 느릴 수 있다는 교육용 사례.
MiniMax-M2.1	대형 MoE · 약 197K	IQ3/Q3 전용 실행	IQ3 93.3GB Q3_K_S 98.7GB	동일 공개 M5 Max 실측 없음	**전용 실행** 앱 종료 필요	대형 에이전트 실험. 3-bit 품질 손실·긴 프리필·런타임 호환성 부담.
Qwen3.5-397B- A17B	MoE 397B/ 활성 17B	IQ1_M만 명목 적재	106.8GB Q2_M 123.1GB	없음	**이론적**	1-bit 품질 저하와 21GB 이하 여유. “돌아감” 시연이지 최적 구성 아님.
DeepSeek-V3.2 / Kimi-K2.5	초대형 MoE Kimi 1T/활성 32B	최저급 1-bit도 초과	DeepSeek 최소 161.3GB Kimi 최소 239.5GB	해당 없음	**적재 불가**	SSD mmap·swap으로 억지 실행해도 대화형 워크스테이션으로 보지 않음.

GB는 저장소/API가 표시한 십진 단위다. MLX 저장소 총량에는 설정·토크나이저 등 부속 파일이 포함될 수 있고, GGUF 멀티모달은 별도 projector가 추가될 수 있다.

용도에 따라 주력 모델과 보조 모델을 나누는 편이 낫습니다

하나의 초대형 모델을 상시 띄우기보다, 주력 122B와 빠른 35B·코딩 전용 모델을 필요할 때 교체하십시오.

01 · 종합

Qwen3.5-122B-A10B 4-bit

한국어·문서·이미지·추론·에이전트를 한 모델로 묶는 기본값. 32K, 단일 슬롯으로 시작.

02 · 코딩

Qwen3-Coder-Next 6-bit

대형 코드베이스와 CLI 에이전트. 품질을 위해 Q6를 쓰되 브라우저·IDE 동시 메모리를 확인.

03 · 속도

Qwen3.5-35B-A3B 4-bit

빠른 보조 모델·요약·분류·병렬 작업. 공개 실측에서 가장 높은 디코드 중앙값.

04 · 추론 대안

gpt-oss-120b MXFP4

텍스트 추론과 도구 사용을 Qwen과 A/B. 비전이 필요하면 주력으로 쓰지 않음.

05 · 품질 우선

Qwen 122B 5-bit

84.9GB로 43.1GB 명목 여유. 4-bit 답변 품질이 실제 업무에서 부족할 때만 승격.

06 · 실험

MiniMax M2.1 3-bit

다른 앱을 닫은 전용 세션. 상시 사용 모델이 아니라 대형 MoE 연구용.

CONFIRMED TARGET

128GB 옵션은 40코어 GPU·614GB/s M5 Max를 뜻합니다

사용자 지정 구성과 Apple 공식 표에서 논리적으로 확정되는 항목을 분리했습니다.

항목	값	근거 등급	로컬 LLM 의미
제품	16형 MacBook Pro, M5 Max, 128GB, 2TB SSD	사용자 제공	실행 대상이며 현재 Hermes 호스트의 자동 탐지값으로 대체하지 않음
CPU/GPU	18코어 CPU · 40코어 GPU	공식 표에서 추론	Apple은 128GB를 40코어 GPU M5 Max 구성에만 제공
메모리 대역폭	614GB/s	Apple 공식	디코드에 중요한 데이터 이동 한도지만 tok/s 보장값은 아님
통합 메모리	128GB	사용자 제공 + 공식 옵션	CPU와 GPU가 공유. 별도 VRAM을 더하지 않음
SSD	2TB	사용자 제공	여러 체크포인트 보관에는 유리하나 실행 메모리를 대신하지 않음

항목	값	근거 등급	로컬 LLM 의미
전원	100Wh 배터리 · 140W 어댑터	Apple 공식	장시간 추론은 전원 연결·통풍 확보가 현실적

CONCEPTS

어려운 용어를 “책장과 작업대”로 이해하면 쉽습니다

모델 파일은 SSD 책장에 있고, 실행할 때 통합 메모리 작업대로 올라옵니다.

파라미터

모델이 학습한 숫자 조각의 개수다. 많다고 무조건 낫지는 않지만 저장·메모리 비용의 출발점이다.

양자화

가중치를 16비트 대신 8·6·5·4비트로 압축하는 것. 비트가 낮을수록 작지만 품질 손실 가능성이 커진다. 실제 파일 트리를 확인해야 한다.

Dense와 MoE

Dense는 모든 전문가가 매 토큰 일하고, MoE는 일부 전문가만 선택한다. 다만 MoE도 전체 가중치는 메모리에 올라가야 한다.

컨텍스트

모델이 한 번에 보는 작업대 길이. 262K를 지원해도 32K보다 프리필 지연·KV 캐시·메모리 부담이 크다.

KV 캐시

이미 읽은 토큰을 다시 계산하지 않게 저장하는 메모장. 컨텍스트와 동시 사용자 수가 커질수록 메모리를 더 먹는다.

프리필과 디코드

프리필은 질문을 읽는 속도, 디코드는 답을 한 토큰씩 쓰는 속도다. 총 토큰/시간 하나로 섞으면 체감 속도를 왜곡한다.

메모리 대역폭

초당 옮길 수 있는 데이터 양. 614GB/s는 강력하지만 커널·캐시·활성 전문가·열 상태 때문에 그대로 tok/s가 되지 않는다.

SSD와 swap

SSD는 모델 보관함이다. 통합 메모리가 모자라 swap을 쓰면 실행은 가능해 보여도 대화형 성능과 SSD 수명·발열에 불리하다.

QUANTIZATION

128GB에서는 122B 4-bit가 균형, 5-bit가 품질 우선입니다

낮은 비트는 “더 큰 모델을 넣기 위한 만능키”가 아닙니다.

양자	일반적 의미	122B 실제 MLX/GGUF 사례	128GB 권고
8-bit	품질 보존 우선, 크기 큼	Qwen 122B Q8 GGUF 129.9GB	적재 불가. 27~35B에 사용
6-bit	고품질, 중간 크기	MLX 100.2GB	전용 세션만. 32K·앱 동시 실행 여유가 작음
5-bit	품질 우선 균형	MLX 84.9GB	4-bit 품질이 부족할 때 승격
4-bit	가장 실용적인 균형	MLX 69.6GB · GGUF Q4_K_M 76.5GB	**기본 권장**
3-bit	대형 모델 적재용	397B도 140GB 이상, MiniMax M2.1은 93~99GB	실험용. 품질 손실 비교 필수
1~2-bit	극단 압축	397B IQ1_M 106.8GB	적재 시연일 뿐 주력 모델 비권장

RUNTIME

Apple 최적화는 MLX, 호환성은 llama.cpp, 편의성은 Ollama입니다

모델과 양자가 같아도 런타임·캐시·버전에 따라 속도와 기능이 달라집니다.

MLX-LM · 1순위

Apple Silicon 네이티브. 4·5·6·8-bit MLX 변환과 멀티모달 지원을 우선 시험.

llama.cpp · 정밀 제어

GGUF 양자 선택이 가장 넓고 최신 Qwen3.5·GLM·gpt-oss 변환 코드가 확인됨.

Ollama · 가장 쉬움

qwen3.5:122b 약 81GB 태그와 API 서버를 간단히 실행. 기본 양자와 컨텍스트를 반드시 확인.

LM Studio · GUI

다운로드·메모리 표시·OpenAI 호환 서버가 편리. 최신 아키텍처 지원 버전을 확인.

권장 시작 순서

```
# 1) Apple 네이티브 주력 모델
uv tool install mlx-lm
mlx_lm.server --model mlx-community/Qwen3.5-122B-A10B-4bit --port 8080

# 2) GGUF 대안
brew install llama.cpp
llama-server -hf unsloth/Qwen3.5-122B-A10B-GGUF:Q4_K_M -c 32768 -ngl 99

# 3) 가장 간단한 실행
ollama run qwen3.5:122b
```

명령은 2026-08-03 공개 저장소·카탈로그 기준 출발점이다. 멀티모달 projector, tool-call parser, 최신 최소 버전은 실제 설치 시 릴리스 노트와 모델 카드로 재확인해야 한다.

MEASURED PERFORMANCE

M5 Max 실측은 MoE가 “큰데도 빠른” 이유를 보여줍니다

512~32K 프롬프트, 128·512·1024 디코드 조합에서 모델당 11회. 프리필과 디코드를 분리했습니다.

모델	디코드 중앙값	범위	512/128 프리필	32K 피크 메모리	해석
Qwen3.5-35B-A3B 4-bit	141.9 tok/s	112.1~145.1	2,931.5 tok/s	24.1GB	활성 3B MoE의 속도 강점
gpt-oss-120b MXFP4	99.3 tok/s	70.1~102.3	1,986.1 tok/s	64.4GB	활성 5.1B, 텍스트 추론 대안
GLM-4.7-Flash 6-bit	74.8 tok/s	40.8~79.2	2,310.7 tok/s	35.2GB	긴 컨텍스트에서 하락 폭 확인
Qwen3.5-122B-A10B 4-bit	**63.9 tok/s**	52.3~67.2	1,105.7 tok/s	76.4GB	품질·속도·여유의 균형
Qwen3.5-27B Dense 4-bit	30.3 tok/s	19.5~32.9	808.6 tok/s	23.3GB	작아도 모든 27B가 활성돼 느림

이 숫자는 공개된 특정 런타임·모델 파일·설정의 실측이며 모든 M5 Max에 대한 보장값이 아니다. 원문은 정확한 메모리 용량·GPU bin을 명시하지 않는다.

STORAGE & CONTEXT

2TB SSD는 모델 컬렉션에 강하지만, 긴 컨텍스트를 공짜로 만들지는 않습니다

80%인 1.6TB를 모델 예산으로 잡는 단순 예시이며 실제 OS·프로젝트·사진·영상 공간은 별도입니다.

약 18개
85GB급 5-bit 체크포인트
약 25개
63GB급 120B 체크포인트
32K
주력 모델 권장 시작 컨텍스트
1 slot
대형 모델 초기 동시성
1M 컨텍스트가 모델 카드에 있어도 실제로는 KV 캐시, 프리필 시간, 답변 관련성, 런타임 지원이 모두 병목이 된다. 32K → 64K → 128K 순서로 늘리고 메모리 압력·첫 토큰 지연·정답률을 함께 기록해야 한다.

BENCHMARK PLAN

최종 선택은 이 Mac에서 30분 A/B로 확정할 수 있습니다

검수는 꼭 사용자가 합니다. 공개 실측은 방향을 주고, 실제 업무 프롬프트가 최종 결론을 냅니다.

01
환경 고정

전원 연결, 같은 macOS·런타임 버전, 브라우저 탭과 IDE 상태 기록.

02
세 모델

Qwen 122B 4-bit, 5-bit, Qwen Coder Next 6-bit를 같은 10개 과제로 비교.

03
32K 시작

긴 문서 1개와 코드베이스를 넣고 피크 메모리·첫 토큰·디코드 속도 기록.

04
품질 채점

한국어 정확성, 코딩 테스트 통과, 인용 근거, 도구 호출 성공률을 5점 척도로 평가.

05
안정성

30분 연속 생성에서 속도 하락, fan, 배터리·전력, memory pressure, swap을 확인.

06

승격 규칙

5-bit가 품질을 유의미하게 높일 때만 4-bit에서 승격. 아니면 여유를 유지.

확인된 사실: 128GB 옵션이 40코어 GPU·614GB/s M5 Max 구성이라는 점, 모델 카드의 구조·컨텍스트·라이선스, HF API 파일 총량, 공개 M5 Max MLX 실측값.

- 실측 원문은 M5 Max의 정확한 메모리·GPU bin, macOS·MLX 세부 버전을 본문에서 완전히 명시하지 않았다. 따라서 숫자는 동일한 128GB 기기의 보장치가 아니라 방향성 근거다.
- Qwen3.5-122B 5-bit, Qwen3-Coder-Next 6-bit, MiniMax M2.1의 동일 조건 M5 Max 실측은 확인하지 못했다.
- 한국어 품질·코딩 성공률·도구 호출 안정성은 공개 영문·다국어 벤치마크만으로 확정하지 않고 실제 업무 프롬프트 A/B를 권고한다.
- 양자화 품질은 변환 방식·importance matrix·group size에 따라 다르며 같은 "4-bit"라도 동일하지 않다.
- 128GB를 모델 가중치에 전부 쓰지 않았다. OS·앱·KV 캐시·런타임·안전 여유를 포함해 판단했다.
- LM Studio 공식 시스템 요구사항은 확인 시점에 M1~M4를 열거했고 M5를 명시하지 않았다. 실제 M5 사용자 제보는 있으므로 실행 가능성과 공식 지원 문구를 구분했다.

최종 권고: Qwen3.5-122B-A10B MLX 4-bit를 32K·단일 슬롯으로 시작하고, 실제 품질이 부족할 때만 5-bit로 올리십시오. 코딩은 Qwen3-Coder-Next 6-bit, 빠른 보조 작업은 Qwen3.5-35B-A3B 4-bit를 별도 프로필로 두는 구성이 가장 합리적입니다.

인터넷·언론 반응

X·Threads·블로그·커뮤니티·뉴스에서 실제로 확인한 의견을 긍정·우려·조건부 평가로 나눴습니다.

긍정·독립 실측

"노트북에서 가장 인상적인 실리콘"

Creative Strategies · Max Weinbach · 2026-03-09

M5 Max의 성능/전력과 동시 CPU·GPU 지속 동작을 높게 평가했다. 공개 탐색기에서는 프리필·디코드·피크 메모리를 분리해 5개 모델 110회를 기록했다. 다만 본문은 정확한 메모리 용량과 GPU bin을 명시하지 않는다.

원문 보기

긍정·M5 Max 128GB 사용자 측정

MLX와 GGUF의 차이가 모델 크기만큼 중요하다

GitHub mlx-serve 이슈 · ddalcu · 2026-08-03

M5 Max 128GB에서 Qwen 계열을 측정한 사용자는 Qwen3.6-35B-A3B MLX 6-bit 99~108 tok/s, 27B MLX 4-bit+MTP 39~55 tok/s, 동일 계열 GGUF 18~19 tok/s를 보고했다. 단일 사용자·개발 중 런타임 결과다.

원문 보기

비판·버전별 성능 결함

LM Studio의 특정 런타임은 긴 프롬프트가 2~3배 느렸다는 제보

LM Studio bug tracker · 2026-06-11

M5 Max 40GPU/128GB에서 당시 LM Studio llama.cpp runtime이 Metal tensor API를 비활성화해 upstream보다 긴 프롬프트 처리와 TTFT가 느렸다는 사용자 A/B다. 생성 속도는 비슷했고 MLX 경로는 영향이 없었다고 기록했다.

원문 보기

비판·안정성 이슈

Metal 백엔드는 특정 MoE 임베딩에서 멈춤이 보고됐다

node-llama-cpp 이슈 · 2026-04-14

M5 Max에서 특정 node-llama-cpp·nomic MoE embedding 호출이 간헐적으로 멈춘다는 제보다. 모든 GGUF 생성 모델의 결함으로 일반화할 수 없지만, 임베딩과 생성 경로를 별도 검증해야 한다.

원문 보기

유보·구매 관점

로컬 모델은 프라이버시에 강하지만 클라우드 최전선을 완전히 대체하지 않는다

Hacker News 토론 · 2026-03-29

한 참여자는 M5 Pro의 Qwen3.5가 Opus보다 약 4배 느리고 결과도 낮았다고 썼고, 다른 참여자는 민감 데이터와 단순 작업에는 로컬이 유용하다고 반론했다. 개인 의견 표본이다.

토론 보기

온라인 의견과 GitHub 이슈는 대표 표본이 아니다. 모델·양자·런타임·컨텍스트·온도 조건이 다르면 숫자를 직접 합치지 않았다.

온라인 의견은 대표 표본이며 전체 여론을 뜻하지 않습니다. 작성자·매체·게시일·원문 링크는 아래 전체 출처에 함께 기록했습니다.

전체 출처

사실·분석·전망·루머·온라인 반응, 시각자료와 생성 이미지의 상태를 함께 기록했습니다.

유형	자료명	제공자·저작자	원문	날짜	라이선스·상태	사용 범위
공식 사양	MacBook Pro 제품 사양	Apple	원문 보기	2026-08-03 확인	공식 원문	M5 Max·128GB·614GB/s·배터리·전원
공식 사진	MacBook Pro · Xcode·LM Studio 제품사진	Apple Newsroom	원문 보기	2026-03-03	공식 미디어 자산	상단 실제 제품사진
공식 모델	Qwen3.5-122B-A10B	Qwen3.5	원문 보기	2026-08-03 확인	공식 모델 카드	122B/10B·컨텍스트·라이선스
공식 모델	Qwen3.5-35B-A3B	Qwen3.5	원문 보기	2026-08-03 확인	공식 모델 카드	35B/3B·멀티모달
공식 모델	Qwen3.5-27B	Qwen3.5	원문 보기	2026-08-03 확인	공식 모델 카드	Dense 27B
공식 모델	Qwen3-Coder-Next	Qwen3	원문 보기	2026-08-03 확인	공식 모델 카드	80B/3B·코딩
공식 모델	gpt-oss-120b	gpt	원문 보기	2026-08-03 확인	공식 모델 카드	117B/5.1B·MXFP4
공식 모델	GLM-4.7-Flash	GLM	원문 보기	2026-08-03 확인	공식 모델 카드	30B/3B
공식 모델	MiniMax-M2.1	MiniMax	원문 보기	2026-08-03 확인	공식 모델 카드	대형 MoE·컨텍스트
공식 모델	Qwen3.5-397B-A17B	Qwen3.5	원문 보기	2026-08-03 확인	공식 모델 카드	397B/17B
공식 모델	DeepSeek-V3.2	DeepSeek	원문 보기	2026-08-03 확인	공식 모델 카드	최저 GGUF 적합도
공식 모델	Kimi-K2.5	Kimi	원문 보기	2026-08-03 확인	공식 모델 카드	1T/32B·적재 불가
아티팩트	Qwen3.5-122B GGUF 파일 트리	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
아티팩트	Qwen3.5-122B MLX 4-bit	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
아티팩트	Qwen3-Coder-Next MLX 6-bit	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
아티팩트	gpt-oss-120b GGUF	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
아티팩트	MiniMax-M2.1 GGUF	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
아티팩트	DeepSeek-V3.2 GGUF	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
아티팩트	Kimi-K2.5 GGUF	Hugging Face 공개 저장소	원문 보기	2026-08-03 확인	커뮤니티/런타임 변환	실제 파일 크기
런타임	MLX-LM	공식 프로젝트	원문 보기	2026-08-03 확인	공식 저장소/카탈로그	Apple Silicon 런타임
런타임	llama.cpp	공식 프로젝트	원문 보기	2026-08-03 확인	공식 저장소/카탈로그	GGUF·Metal·아키텍처 지원
런타임	Ollama qwen3.5	공식 프로젝트	원문 보기	2026-08-03 확인	공식 저장소/카탈로그	122B 태그·파일 크기
독립 실측	M5 Max: Chiplets, Thermals, and Performance per Watt	Creative Strategies · Max Weinbach	원문 보기	2026-03-09	독립 리뷰·정확한 RAM 미표기	열·전력·AI 실측 해석
실측 데이터	Apple Silicon Benchmark Explorer	Creative Strategies	원문 보기	2026-03	110회 공개 데이터	프리필·디코드·피크 메모리
개인 경험	qMLX: Qwen 122B daily driver	Andryo Marzuki	원문 보기	2026-07-11	개인 M3 Ultra 96GB 경험	캐시·긴 컨텍스트 한계
공개 토론	M5 MacBook Pro buyers, worth it?	Hacker News	원문 보기	2026-03-29	개인 의견 표본	클라우드 대비 반론
공개 프로젝트	Anubis OSS	UNC Software	원문 보기	2026-03 확인	커뮤니티 벤치마크	측정 방법론
개인 실측	M5 Max 128GB MLX/GGUF 측정	GitHub · ddalcu	원문 보기	2026-08-03	단일 사용자 측정	MLX·GGUF·MTP 차이
버그 제보	LM Studio M5 tensor API 성능	LM Studio bug tracker	원문 보기	2026-06-11	버전 한정 사용자 A/B	긴 프롬프트 TTFT 위험
버그 제보	node-llama-cpp Metal MoE hang	GitHub 사용자	원문 보기	2026-04-14	특정 임베딩 경로	안정성 검증 신호

공식 보도자료	M5 Pro·M5 Max와 MacBook Pro 미디어 자산	Apple Newsroom	원문 보기	2026-03-03	공식 미디어 자산	실제 제품사진·LM Studio/Xcode
AI 생성 이미지	128GB 통합 메모리 두들 인포그래픽	AI 생성 시각자료	원문 보기	2026-08-03	AI 생성 · 공식 제품 이미지 아님	설명용 시각자료