나만의 인공지능 · 3강

모델 고르는 법

3,930만
Llama 3.1 학습 GPU 시간

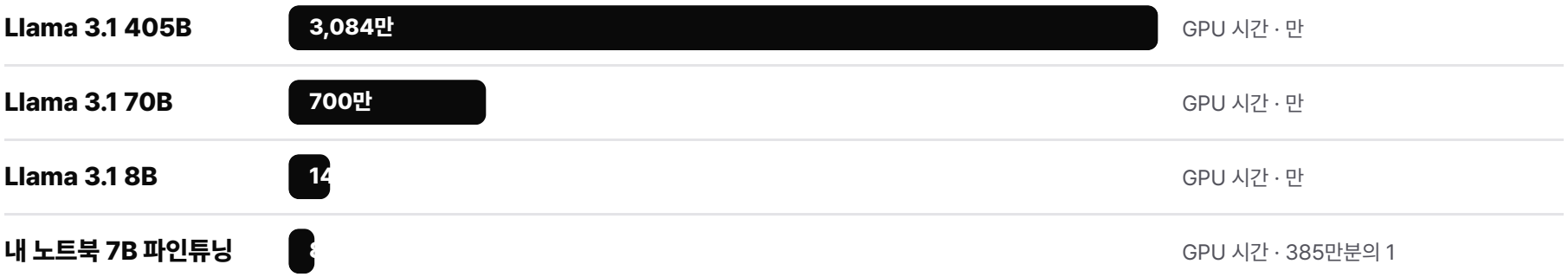
300만
허깅페이스 모델 수

0.6GB
1B당 파일 크기 · Q4 기준

1·2강에서 두뇌를 만들어 봤기 때문에, 이제 남이 만든 두뇌의 사양표를 읽을 수 있습니다

직접 만들면

비용



16,384
H100 장 · 54일 연속

419
예기치 못한 고장 · 3시간에 한 번 (54일간)

7,800만
GPT-4 훈련비 추정 · 달러

출처 Meta 모델 카드 · Llama 3 논문 (arXiv) · Stanford AI Index (HAI)

그래서

받아 쓴다

직접 만들기

3,084만 시간

데이터 15조 토큰 · H100 1만 6천 장 · 54일

받아서 조금 바꾸기

2~3달러

내 GPU 6~8시간 · 파인튜닝

출처 Meta 모델 카드

모델이 모여 있는 곳

허깅페이스

300만
모델 · 2026년 8월

100만
데이터셋

144만
스페이스

텍스트
Qwen · Llama
글쓰기 · 대화 · 코딩

이미지
FLUX.1
120억 파라미터

음성
Whisper
500만 시간 학습

영상
LTX · LingBot
월드모델까지

출처 HF 300만 돌파 · State of Open Models 2026

숫자가 많다 ≠ 고를 게 많다

현실

83%

작은 모델이 차지하는 다운로드

출처 State of Open Models 2026

1·2강이 여기서 쓰인다

사양표

8B	가중치 **80억 개**. 1강에서 만든 w 와 b, 그 숫자가 80억 개
Q4_K_M	그 숫자를 **4비트로 줄여 저장**. 파일이 **3분의 1**로 줄고 품질 손실은 거의 없음
context 128K	한 번에 읽는 분량. **끝까지 켜면 모델 파일보다 메모리를 더 먹는다**
Instruct	지시를 따르도록 **가르친 뒤**. Base 는 대화가 안 된다 — 초보는 반드시 Instruct
GGUF	내 컴퓨터에서 **바로 돌리는 형식**. 파일 하나만 받으면 됨
Apache-2.0	팔아도 되는 라이선스. **Llama는 조건부, 비상업 표시는 사업 불가**

출처 양자화 품질 실측 · Llama 라이선스 · Gemma 약관

외울 건 하나

0.6

암산 공식

크기(B) × 0.6 = GB

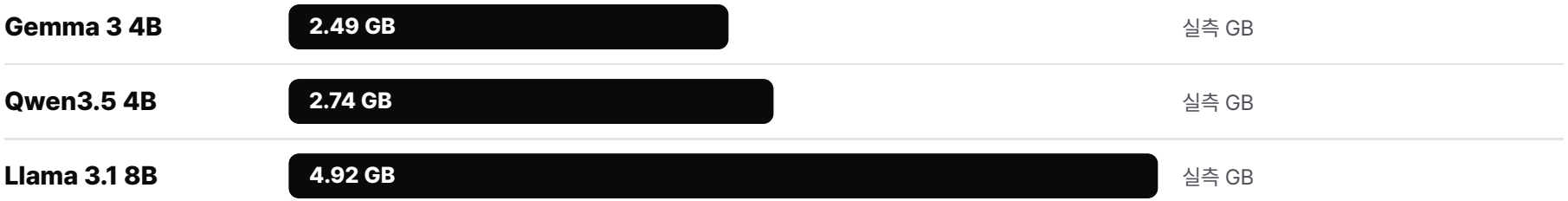
Q4 기준 **파일 크기**. 8B → 약 4.8GB
실제 램은 여기에 대화 기록과 운영체제가 더 붙는다

실측

4.92 GB

Llama 3.1 8B Q4_K_M 실제 파일

저장 방식	1B당	8B 모델이면	품질
F16 (원본)	2.0 GB	약 16 GB	100%
Q8	1.0 GB	약 8.5 GB	거의 100%
Q4_K_M (기본)	**0.6 GB**	**4.92 GB 실측**	거의 차이 없음



출처 Gemma 3 4B GGUF · Llama 3.1 8B GGUF

내 컴퓨터에 맞는 크기

배정표

램 8GB

4B 까지

Gemma 3 4B
Qwen3.5 4B

약 2.5 GB

램 16GB

8~12B

Llama 3.1 8B
Gemma 4 12B

약 5~8 GB

램 32GB

27B 까지

Qwen 27B

약 16~18 GB

램 64GB

70B 까지

느리지만 돌아감

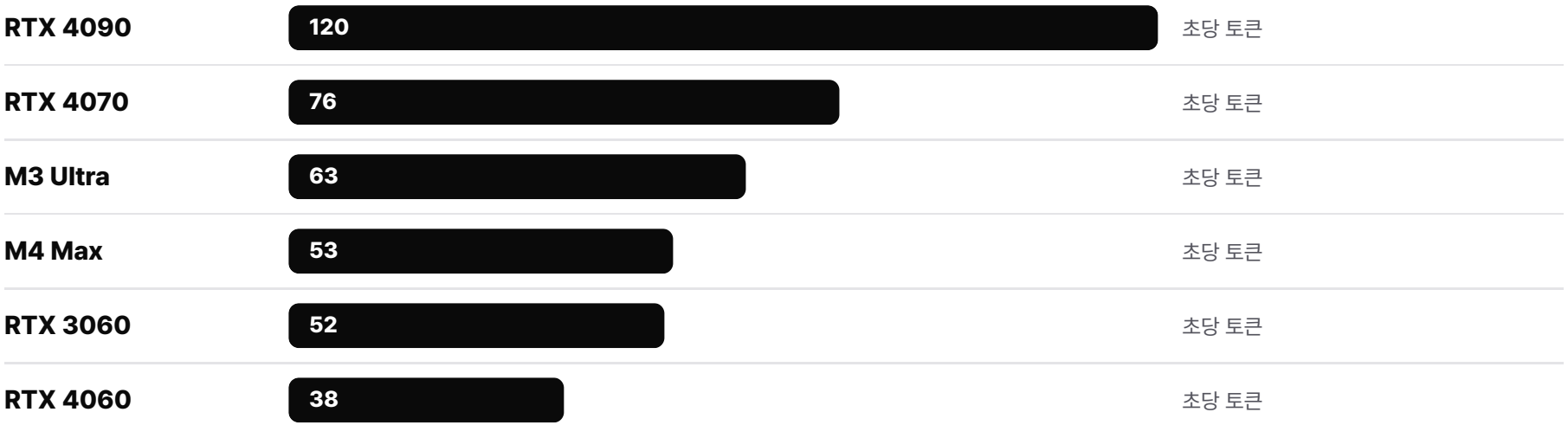
약 40 GB

모델 파일 외에 운영체제와 대화 기록(KV 캐시)이 메모리를 더 씁니다. 맥은 GPU가 쓸 수 있는 양이 전체의 약 75%로 제한돼 있어 더 보수적으로 잡았습니다.

출처 실측 파일 크기 · 맥 메모리 상한

속도를 정하는 것

대역폭



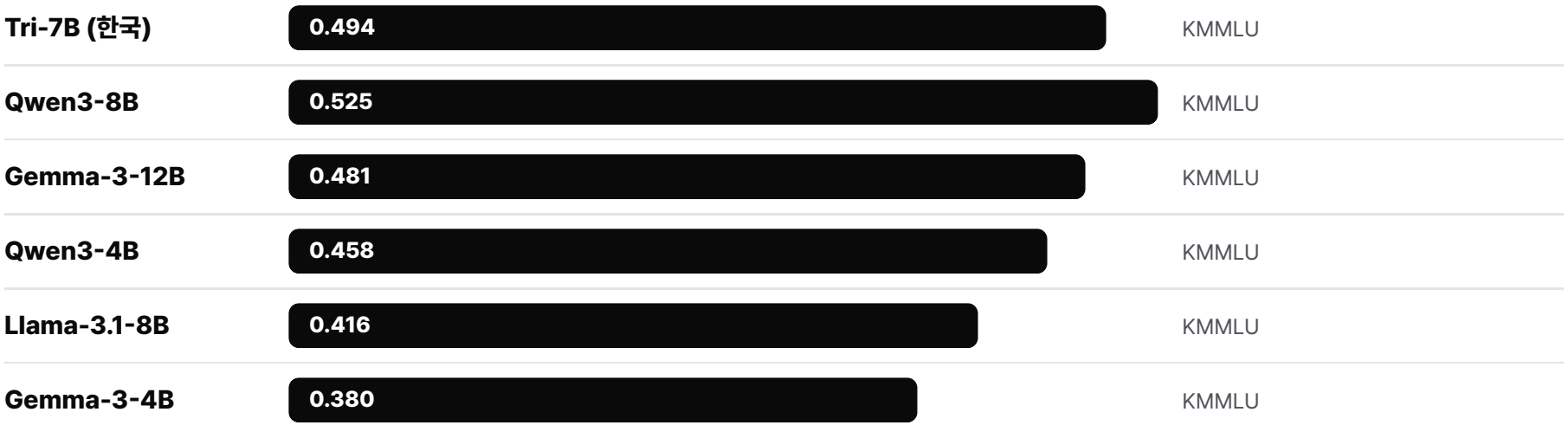
대역폭 ÷ 모델 크기

초당 토큰 ≈

출처 LocalScore 실측

한국어는 다르다

한국어



3배 작은데 거의 같다

Qwen3-4B vs Gemma-3-12B

출처 KMMLU·HAERAE 논문

안 열릴 때

고장

안 열림	모델이 아니라 **프로그램이 낡은 것**. LM Studio·런타임 업데이트
신호등	초록 = 다 올라감 · 노랑 = 반만 · **빨강 = 안 들어감**
느려짐	GPU 오프로드를 낮추거나 **더 작은 모델로**
길어지면 죽음	대화가 길어지면 메모리가 큰다. **컨텍스트를 2048로**
메모리 부족	브라우저 끄고 재부팅. 빈 공간이 아니라 **연속된 공간**이 필요

출처 LM Studio 버그트래커

이 프로그램들이 하는 일

몸

① **엔진**	모델 파일은 **숫자 덩어리**일 뿐. 그 숫자로 계산을 돌려 주는 프로그램을 깔아 준다
② **내려받기**	허깅페이스에서 **내 컴퓨터에 맞는 파일**을 찾아서 받아 준다
③ **메모리에 올리기**	GPU와 램에 **나눠 올린다**. 안 맞으면 알려 준다
④ **말 걸기**	모델마다 다른 대화 형식을 **알아서 맞춰 준다**
⑤ **문 열기**	내 컴퓨터 안에 **작은 서버**를 띄운다 → 내 프로그램이 연결 (4강)

모델 = 두뇌 · 이 프로그램 = 두뇌를 넣고 돌리는 몸

둘 중 무엇을

비교

	LM Studio	Ollama
쓰는 법	**화면에서 클릭**	명령어 한 줄
모델 찾기	**앱 안에서 검색 · 램 신호등**	목록에서 이름 확인
처음 배울 때	**쉬움 — 3강은 이걸로**	익숙해지면 빠름
내 프로그램에 연결	됨 `localhost:1234`	**됨 `localhost:11434` — 4강**
맥 · 윈도 · 리눅스	전부	전부
값	**무료**	**무료**

받기 lmstudio.ai · ollama.com

외울 명령어는 다섯 개

명령어

하고 싶은 것	Ollama	LM Studio
받아서 바로 대화	`ollama run gemma3:4b`	검색 → Download → Chat
받기만	`ollama pull qwen3:4b`	Download 버튼
뭐가 있나 보기	`ollama list`	My Models 탭
지우기	`ollama rm gemma3:4b`	휴지통 아이콘
서버 켜기	`ollama serve`	Developer 탭 → Start

모델 이름 뒤 `:4b` 가 크기입니다. 내 램에 맞춰 `:1b` `:4b` `:12b` 로 바꿔 보세요

목록 ollama.com/library · LM Studio CLI 문서

글만이 아니다

전부

글

LM Studio

대화 · 요약 · 번역 · 코딩

Qwen · Gemma

그림

ComfyUI

사진 · 썸네일 · 로고

FLUX · SDXL

소리

Whisper · TTS

받아쓰기 · 내 목소리

자막 · 나레이션

영상

ComfyUI

짧은 영상 · 월드모델

LTX · Wan

전부 내 컴퓨터에서, 사용료 0원으로 돕니다

받기 ComfyUI · Whisper · FLUX.1

오늘 할 일

실습

1
고른다
내 램 ÷ 0.6

2
받는다
GGUF · Instruct

3
돌린다
LM Studio

4
비교한다
큰 것 vs 작은 것

로컬 AI 모델 도감 — 우리가 만든 탐색기로 먼저 골라 보기

허깅페이스 — GGUF 모델 인기순

LM Studio 내려받기

Ollama 모델 목록

내 컴퓨터 속도 확인 — LocalScore

로컬 AI를 쓰는 다음 세대의 방법

사다리

3강 · 오늘

받아 쓴다

고르고 · 받고 · 돌린다

사양표 읽기

다음

지식을 연결한다

내 자료를 붙인다

RAG

그다음

가중치를 바꾼다

내 것으로 다시 가르친다

파인튜닝

끝

합성한다

두 두뇌를 섞는다

진화

3,084만 시간 → 0

만들지 않고, 골라서 · 붙이고 · 바꾸고 · 섞는다

실습 모델 도감 — 진화(합성)까지

정리

3,930만
만들면 GPU 시간

2~3
받아 쓰면 달러

0.6
1B당 GB · Q4 파일

83%
작은 모델이 대세

출처 Meta · Hugging Face · LocalScore · 한국어 벤치마크 · LM Studio