기본 01

# 알아야 할 숫자는 둘뿐이다

메모리 크기

## 되느냐, 안 되느냐

모델이 통째로 메모리에 올라가야 돌아간다. **1GB라도 모자라면 아예 안 돌아가거나, 하드디스크를 오가며 수십 배 느려진다.**

엔비디아는 **VRAM**, 맥은 **통합 메모리**가 이 역할을 한다. 여기서 「무엇을 할 수 있는가」가 정해진다.

메모리 대역폭

## 얼마나 빠르냐

모델은 글자 하나 만들 때마다 **메모리 전체를 한 번씩 훑는다.** 그래서 계산 능력보다 **메모리를 얼마나 빨리 읽느냐**가 속도를 결정한다.

단위는 **GB/s**다. 이 숫자가 크면 글자가 빨리 나온다. 여기서 「얼마나 답답한가」가 정해진다.

**순서가 중요하다.** 크기가 먼저고 속도가 나중이다. 안 들어가는 모델은 아무리 빨라도 못 돌린다. 들어가고 나서야 속도를 따진다.

기본 02

# 필요한 메모리는 이렇게 계산한다

모델 이름 뒤에 붙은 **7B, 32B, 70B**가 파라미터 수다. B는 10억이다. 여기에 **양자화**— 모델을 압축한 정도를 곱하면 필요한 메모리가 나온다.

**외울 필요 없다 · 아래 계산기가 대신 해준다**

필요 메모리 = **파라미터 수(B)** × **양자화 계수** + 여유분

양자화 계수

FP16 원본 그대로 **2.1** — 품질 최상, 메모리 최대

Q8 절반으로 압축 **1.1** — 품질 손실 거의 없음

Q4 4분의 1로 압축 **0.62** — 가장 널리 쓰인다

예) 70B 모델을 Q4로 → 70 × 0.62 = **약 43GB**

32B 모델을 Q4로 → 32 × 0.62 = **약 20GB**

8B 모델을 Q4로 → 8 × 0.62 = **약 5GB**

**Q4가 기본이다.** 4분의 1로 줄이는데 품질은 대부분 유지된다. 처음에는 Q4로 시작하고, 결과가 아쉬울 때 Q8로 올리면 된다.

계산기 01

# 내 모델, 몇 GB 필요한가

모델 크기

8B — 가벼운 대화·요약

양자화

Q4 — 4분의 1 압축 (권장)

문맥 길이

보통 (32K) — 문서 작업

## 약 8.6GB

모델 **5.0GB** + 문맥·여유분 **3.6GB**

| | | | |
|---|---|---|---|
| 노트북 내장 8GB 부족 | RTX 4060 Ti 16GB 가능 | RTX 3090·4090 24GB 가능 | RTX 5090 32GB 가능 |
| 맥 64GB (한도 56) 가능 | 맥 128GB (한도 116) 가능 | | |

일반적인 장비로 충분하다. 굳이 비싼 것을 살 필요가 없는 구간이다.

여유분은 문맥 저장 공간과 운영체제 몫을 함께 잡은 값이다. 실제로는 사용하는 프로그램에 따라 조금씩 달라진다.

기준표

# 메모리별로 할 수 있는 것

| 메모리 | 돌릴 수 있는 것 (Q4 기준) | 해당 장비 |
|---|---|---|
| **8GB** | 3B 급. 간단한 요약과 분류 정도 | 보급형 노트북, 맥북 에어 |
| **16GB** | 8B 급 대화 모델. 미니맥스 H3 INT4(15GB) | RTX 4060 Ti 16GB |
| **24GB** | 14B 급. H3 INT8(27GB)은 아슬아슬 | RTX 3090 · 4090 |
| **32GB** | **32B 급 실무 모델까지. 여기가 대부분의 상한선** | RTX 5090 |
| **64GB** | **70B 급(43GB)이 들어간다. 급이 달라진다** | 맥북 프로 M5 맥스 64GB |
| **128GB** | 120B 급까지. 여러 모델 동시 실행 | 맥북 프로 M5 맥스 128GB |

**32GB와 64GB 사이에 큰 문턱이 있다.** 70B 모델이 Q4로 43GB 정도인데, 32GB에는 안 들어가고 64GB에는 들어간다. 여기서 할 수 있는 일이 완전히 달라진다.

비교 01

# 맥과 엔비디아는 구조가 다르다

엔비디아 · 따로 있는 메모리

## 빠르지만 작다

그래픽카드에 **VRAM**이 따로 붙어 있다. 아주 빠르지만 크기가 정해져 있다. 가장 비싼 소비자용 카드인 **RTX 5090도 32GB**다.

모델이 32GB를 넘으면 나머지를 본체 메모리에서 끌어와야 하는데, 그 통로가 좁아서 **순식간에 느려진다.**

맥 · 하나로 합친 메모리

## 느리지만 크다

CPU와 GPU가 **같은 메모리를 나눠 쓴다.** 그래서 64GB를 사면 그중 상당 부분을 모델에 그대로 줄 수 있다.

대신 대역폭이 그래픽카드만 못하다. **M5 맥스가 614GB/s**인데, RTX 5090은 그보다 세 배가량 빠르다.

**그래서 갈림길이 생긴다.** 작은 모델은 엔비디아가 압도적으로 빠르고, 큰 모델은 엔비디아에 아예 안 들어간다. **어느 쪽이 좋은 게 아니라, 무엇을 할 것인지가 먼저다.**

비교 02

# 실제 숫자로 보면

| | 맥북 프로 M5 맥스 64GB | RTX 5090 |
|---|---|---|
| 모델용 메모리 | **약 48GB (설정 시 56GB)** | 32GB |
| 메모리 대역폭 | 614 GB/s | **약 3.3배 빠름** |
| 8B 모델 | 돌아감 | **훨씬 빠르게 돌아감** |
| 32B 모델 | 돌아감 | **돌아감 · 더 빠름** |
| 70B 모델 (43GB) | **통째로 올라감** | **안 들어감** |
| 형태 | **노트북 · 어디서나** | 데스크톱 · 전원 필요 |
| 소비 전력 | **노트북 수준** | 본체까지 수백 와트 |

벤치마크 수치는 **출처마다 크게 다르다.** 측정 환경과 설정이 제각각이기 때문이다. 숫자 하나하나를 믿기보다 **구조에서 나오는 결론**을 보는 편이 정확하다. 그 결론은 하나다 — **안 들어가는 순간 끝난다.**

핵심

# 갈림길은 30B 근처에 있다

**1** **30B 아래 — 엔비디아가 이긴다**

모델이 32GB 안에 다 들어간다. 이때는 **대역폭이 빠른 쪽이 무조건 이긴다.** 대화, 코딩 보조, 요약처럼 **빠른 반응이 중요한 일**이면 그래픽카드가 낫다.

**2** **30B 근처 — 비슷해진다**

32GB를 거의 다 쓰기 시작하면 여유가 없어진다. 문맥을 길게 잡으면 넘친다. **이 구간에서는 여유 있는 쪽이 실제로 더 편하다.**

**3** **70B 위 — 맥이 이긴다**

43GB짜리 모델은 32GB 카드에 **애초에 안 들어간다.** 억지로 돌리면 본체 메모리를 오가느라 **맥보다 느려진다.** 이 지점부터는 통합 메모리가 답이다.

**4** **그 위 — 개인은 못 한다**

미니맥스 H3의 원본 가중치가 **115GB**이고 공식 예제는 **GPU 4장**을 요구한다. 이 영역은 사는 게 아니라 **빌려 쓰는 것**이다.

**그래서 질문이 바뀐다.** 「뭐가 제일 좋아요」가 아니라 **「나는 몇 B짜리를 돌릴 건가」**가 먼저다. 그게 정해지면 살 것은 저절로 정해진다.

실전 01

# 맥의 메모리는 전부 쓰이지 않는다

64GB를 샀다고 64GB가 다 모델에 가는 것이 아니다. macOS는 기본적으로 **전체의 4분의 3 정도만** GPU에 내준다.

64GB면 **약 48GB**다. 나머지는 운영체제 몫이다. 이 한도는 **명령어 한 줄로 올릴 수 있다.**

**터미널에 붙여넣기 · 재시동 필요 없음**

```
# 56GB까지 올린다 (56 × 1024 = 57344)
sudo sysctl iogpu.wired_limit_mb=57344

# 원래대로 되돌리기
sudo sysctl iogpu.wired_limit_mb=0
```

**이만큼은 남겨라**

### 운영체제 몫 8~16GB

64GB라면 **56GB까지**가 안전선이다. 그 이상 올리면 화면이 멈추거나 강제로 재시동되는 일이 생긴다.

**주의**

### 재시동하면 초기화된다

이 설정은 **껐다 켜면 사라진다.** 매번 다시 넣거나 시작할 때 자동 실행되도록 걸어둬야 한다.

**48GB와 56GB의 차이가 크다.** 70B 모델이 43GB인데, 48GB로는 문맥을 길게 못 잡는다. 56GB면 넉넉해진다.

계산기 02

# 사는 게 이득인가, 빌리는 게 이득인가

이게 진짜 질문이다. **「무료」라는 말만 듣고 장비부터 사면 안 된다.** 쓰는 양이 적으면 빌려 쓰는 편이 훨씬 싸다.

장비 가격 (만원)

600

한 달에 만들 개수

100개

한 개당 빌려 쓰는 값

영상 15초 — 약 2,800원

## 22개월이면 본전

빌려 쓰면 한 달에 **280,000원** · 장비값이면 **2,143개**를 만든다

**애매하다. 본전까지 1.8년이 걸린다. 그 사이에 더 좋은 모델이 나온다.**

영상 값은 미니맥스 H3의 초당 0.13달러 기준이다. 전기값, 설치에 드는 시간, 장비를 되팔 때 값은 넣지 않았다.

판단

# 로컬이 반드시 필요한 경우

첫째

### 계속 대량으로 돌린다

손익분기를 넘기면 그 뒤로는 계속 이득이다. **위 계산기에서 본전 시점이 1년 안에 들어오면** 사는 편이 낫다.

둘째

### 밖으로 못 보내는 자료다

사람 얼굴, 의료 기록, 회사 내부 자료, 계약 상 외부 전송이 막힌 것. **이건 값이 문제가 아니라 아예 선택지가 없다.**

셋째

### 모델 자체를 손본다

내 자료로 학습시키거나, 안을 뜯어보거나, 두 모델을 붙이거나. **API로는 못 하는 일이다.**

넷째

### 가르치거나 만든다

강의를 하거나 서비스를 만드는 사람. **직접 돌려보지 않으면 설명할 수 없다.**

사지 마십시오

## 하루에 몇 번 쓰는 정도라면

한 달에 스무 번 쓰려고 수백만원짜리를 사는 것은 손해다. **그 돈이면 몇 년치를 빌려 쓴다.** 그리고 그사이에 더 좋은 모델이 나온다.

이렇게 시작하십시오

## 지금 있는 컴퓨터로 먼저

8GB만 있어도 작은 모델은 돌아간다. **먼저 돌려보고, 모자라다고 느껴질 때 사면 된다.** 그때는 무엇이 부족한지 알고 사게 된다.

선택

# 이럴 때는 이걸 사면 된다

| 이런 사람 | 이걸 사면 된다 | 이유 |
|---|---|---|
| **처음 해본다** | 지금 있는 컴퓨터 | 3B~8B는 대부분의 최근 컴퓨터에서 돈다. 먼저 돌려보고 판단한다 |
| **대화·코딩 보조**<br>빠른 반응이 중요 | VRAM 24GB 이상 그래픽카드 | 14B급까지 아주 빠르다. 이 용도로는 맥보다 낫다 |
| **실무용 주력**<br>32B급을 쓴다 | RTX 5090 (32GB) | 32B까지 가장 빠르게 돈다. 다만 여기가 상한선이다 |
| **70B급이 필요하다**<br>또는 여러 모델 동시에 | **맥 통합 메모리 64GB 이상** | 그래픽카드로는 못 한다. 통합 메모리만 되는 영역이다 |
| **영상·음성 생성** | 먼저 빌려 쓴다 | 원본은 GPU 4장이 필요하다. 양자화 버전부터 시험해본다 |
| **한 달에 몇 번** | 사지 않는다 | API가 압도적으로 싸다. 계산기로 확인하면 바로 보인다 |

**가장 흔한 실수는 순서를 거꾸로 하는 것이다.** 장비를 먼저 사고 무엇을 할지 나중에 정한다. 그러면 대개 안 쓰게 된다. **무엇을 할지 정하고, 몇 B가 필요한지 계산하고, 그다음에 산다.**

사례

# 실제로 이렇게 갈린다

지난주 공개된 **미니맥스 H3**를 예로 보자. 영상과 소리를 함께 만드는 모델인데, 가중치가 공개되어 직접 받을 수 있다.

| 방식 | 필요한 것 | 한 편 만드는 시간 |
|---|---|---|
| **원본 그대로** | 가중치 115GB · GPU 4장 | 약 13초 |
| **INT8 압축** | 27.1GB · VRAM 24GB | 느려짐 |
| **가지치기 + 압축** | **31.7GB · RTX 5090 한 대** | **약 175초** |
| **INT4 압축** | 15.0GB · VRAM 16GB | 더 느려지고 품질 손실 |
| **빌려 쓰기** | 없음 | 15초 클립에 약 2,800원 |

계산해보면

## 천 개가 기준선이다

RTX 5090 한 대 값이면 **API로 클립을 천 개 넘게 만든다.** 게다가 로컬은 한 개에 3분씩 걸리고 API는 기다릴 필요가 없다.

그런데도

## 사람 얼굴이 들어가면 다르다

초상권이 걸린 소재, 회사 자료, 계약상 외부에 못 보내는 것. **이때는 값이 얼마든 로컬 말고는 방법이 없다.**

오늘

# 그래서 이걸 골랐다

| M5 맥스 | 64GB | 614GB/s | 약 48→56GB |
| --- | --- | --- | --- |
| CPU 18코어<br>GPU 40코어 | 통합 메모리 | 메모리 대역폭 | 모델에 줄 수 있는<br>메모리 |

1. **70B가 들어가야 했다**
   강의에서 **가장 큰 모델까지 보여줘야** 한다. 32GB 그래픽카드로는 70B를 못 돌린다. 64GB면 43GB짜리가 통째로 올라간다.

2. **들고 다녀야 했다**
   촬영하는 곳이 매번 다르다. **데스크톱은 그 자리를 못 벗어난다.** 노트북에서 70B가 도는 것 자체가 이 선택의 이유다.

3. **속도는 양보했다**
   같은 8B 모델이면 RTX 5090이 훨씬 빠르다. **그건 인정하고 간다.** 빠른 대화가 목적이 아니라 **큰 모델을 보여주는 것**이 목적이기 때문이다.

4. **이게 정답이라는 뜻은 아니다**
   **32B까지만 쓸 거라면 그래픽카드가 낫다.** 한 달에 몇 번만 쓸 거라면 아무것도 안 사는 게 낫다. 용도가 다르면 답도 다르다.

정리

# 사기 전에 이 순서로

1. **무엇을 할지 먼저 정한다**
   대화인가, 코딩인가, 영상인가, 학습인가. **여기서 필요한 모델 크기가 나온다.**

2. **필요한 메모리를 계산한다**
   파라미터 수 × 양자화 계수. **이 교재의 첫 번째 계산기**에 넣으면 바로 나온다.

3. **빌리는 값과 비교한다**
   한 달에 얼마나 쓸지 넣어본다. **두 번째 계산기**가 본전 시점을 알려준다. 1년이 넘어가면 사지 않는 편이 낫다.

4. **지금 컴퓨터로 먼저 돌려본다**
   작은 모델은 대부분의 컴퓨터에서 돈다. **돌려보고 나서 사면 후회가 없다.**

**11월에 「나만의 인공지능 — 로컬AI의 정석」 과정이 열린다.** 내 컴퓨터에서 나만의 AI가 돌아가게 만들고, 사용료 없이 내 서비스에 붙이는 것까지 열 개 챕터로 다룬다. **오늘 계산기로 확인한 것이 그 과정의 출발점이다.**

확인한 곳

# 숫자는 여기서 가져왔다

| | | | |
|---|---|---|---|
| 공식<br>**애플 · 맥북 프로 기술 사양**<br>M5 맥스 40코어 GPU · 614GB/s · 최대 128GB | 공식<br>**미니맥스 H3 발표**<br>2026년 7월 31일 · 가중치 공개 | 실측<br>**H3 로컬 설치 안내**<br>INT4 15GB · INT8 27.1GB · 원본 115GB | 실측<br>**RTX 5090 한 대 실행 보고**<br>31.7GB · 클립 한 편에 약 175초 |
| 설정<br>**맥 GPU 메모리 한도 조정**<br>iogpu.wired_limit_mb · 기본 75% | 함께 보기<br>**무료 지식 아카이브**<br>로컬 AI 관련 강의 30여 편이 정리되어 있다 | | |

**11월 로컬AI 과정** **무료 아카이브**

AI CITY BUILDERS · 모닝 AI · 2026년 8월 6일

벤치마크 수치는 측정 환경에 따라 출처마다 차이가 크다. 이 교재는 개별 수치보다 구조에서 나오는 판단 기준을 정리한 것이다.

사양과 가격은 2026년 8월 기준이며 제조사 정책에 따라 달라질 수 있다.