

클로드코드 토큰 절약 세팅

같은 요금제로 두 배 오래 쓰는 3가지 방법

— 출력 규칙 · 모델 라우팅 · 대화 압축 + CLAUDE.md 다이어트

PART 1 전체 5장

한도, 왜 이렇게 빨리 닳을까

클로드코드를 쓰다 보면 어느 순간 "한도에 도달했습니다"를 만납니다. 요금제가 부족해서라고 생각하기 쉽지만, 사실 **토큰 (AI가 글을 읽고 쓰는 사용량 단위)의 절반은 그냥 새고 있는 경우가 많습니다.**

요금제 한도는 결국 토큰 총량입니다. 클로드가 읽는 글자(입력)와 쓰는 글자(출력)가 모두 토큰으로 계산되고, 쌓이면 한도가 끝납니다.

토큰이 새는 범인은 딱 셋입니다.

- **장황한 답변** — 안 물어본 설명, 바꾼 내용 요약, 파일 전체 재출력. 답이 길수록 토큰이 가장 빨리 닳습니다.
- **모든 일을 최고급 모델로** — 오타 수정까지 제일 비싼 모델에게 시키는 것. 같은 일도 모델에 따라 소모가 몇 배 차이 납니다.
- **끝난 대화를 계속 들고 다니기** — 클로드코드는 매 요청마다 지금까지의 대화 기록(컨텍스트)을 통째로 다시 읽습니다. 대화가 길수록 요청 하나하나가 비싸집니다.

좋은 소식: 세 범인 모두 **설정 두세 줄과 습관 하나로** 잡힙니다. 지금부터 하나씩 갑니다.

◆ 이 PART의 핵심 정리

- 한도 = 토큰 총량. 읽는 것(입력)과 쓰는 것(출력) 모두 소모된다
- 낭비 3범인: 장황한 출력 · 비싼 모델 남용 · 비대한 대화 기록
- 세 가지 다 설정으로 잡을 수 있다 — 같은 요금제로 두 배 오래 쓴다

PART 2 전체 5장

방법 ① 출력 규칙 — 답변부터 다이어트

첫 번째이자 효과가 가장 큰 방법입니다. **CLAUDE.md** 는 프로젝트 폴더에 두는 규칙 파일로, 클로드코드가 매번 시작할 때 읽습니다. 여기에 "답을 짧게 하라"는 출력 규칙 세 줄만 넣으면 됩니다.

프로젝트 폴더 맨 위(루트)에 **CLAUDE.md** 파일을 만들고(이미 있으면 아래만 추가) 이렇게 붙여넣으세요.

MD

```
# 출력 규칙
- 코드를 고칠 땐 바뀐 부분만 보여줘. 파일 전체를 다시 출력하지 마.
- 무엇을 바꿨는지 요약·설명은 내가 물어볼 때만.
- 한 줄로 답이 되면 한 줄로 답해.
```

영어 원문 버전을 선호하면 이걸 쓰세요. 내용은 같습니다.

MD

```
# Output rules
- When editing code, output ONLY the changed lines
  or the single modified block — never reprint the whole file.
- No setup guides, no recaps of what you changed,
  no explanations unless I explicitly ask.
- Keep responses tight. If one line answers it, give one line.
```

왜 효과가 클까요? 클로드는 기본적으로 친절합니다. 코드 한 줄을 고쳐도 파일 전체를 다시 보여주고, 무엇을 바꿨는지 설명하고, 다음 단계까지 안내합니다. 그 친절이 전부 출력 토큰입니다. 파일 전체 재출력 한 번이면 수천 토큰이 그냥 나갑니다. 규칙 세 줄이면 이 낭비가 사라지고, 답이 짧아져 읽기도 편해집니다.

⚡ 한 단계 더 — 프로젝트마다 넣기 귀찮다면 **전역 규칙 파일** `~/.claude/CLAUDE.md` 에 넣으세요. 모든 프로젝트에 한 번에 적용됩니다. 프로젝트 전용 규칙은 프로젝트 CLAUDE.md, 어디서나 쓸 규칙은 전역 — 이 구분만 지키면 됩니다.

◆ 이 PART의 핵심 정리

- CLAUDE.md = 클로드코드가 매번 읽는 규칙 파일
- 출력 규칙 3줄: 바뀐 부분만 · 설명은 물어볼 때만 · 한 줄이면 한 줄
- 출력이 가장 싸다 — 답변 다이어트가 절약 1순위

PART 3 전체 5장

방법 ② 모델 라우팅 — 일에 맞는 뇌 쓰기

클로드코드 안에는 여러 모델(AI 두뇌)이 있고, 무엇을 쓰느냐에 따라 토큰 소모가 몇 배씩 달라집니다. 핵심 원칙은 한 줄입니다.

"무거운 일만 비싼 모델, 나머지는 가볍게."

전환은 간단합니다. 입력창에 `/model` 을 치면 모델 선택창이 열리고, **방향키로 골라 엔터** 한 번이면 끝. 작업 중간에도 언제든지 바꿀 수 있습니다.

모델	언제 쓰나	예시
Opus (오퍼스)	복잡한 판단이 필요할 때	서비스 설계·기획, 원인을 모르는 버그 추적, 긴 문서 분석, 중요한 의사결정
Sonnet (소넷)	단순한 코드 작업 · 일상 작업	화면 문구 수정, 버튼 추가, 반복 수정, 일반적인 질문과 대화
Haiku (하이쿠)	아주 간단한 잡무	오타 수정, 이름 바꾸기, 짧은 확인 질문

감이 안 오면 이렇게 기억하세요 — **생각을 시키면 오퍼스, 손을 시키면 소넷, 심부름은 하이쿠.**

⚡ 한 단계 더 — 한 작업 안에서도 라우팅하세요. 설계·방향 잡기는 오퍼스로 하고, 방향이 정해지면 소넷으로 바꿔 구현을 시키는 식입니다. 이것만으로 체감 비용이 절반으로 줄어듭니다.

◆ 이 PART의 핵심 정리

- `/model` → 선택창에서 방향키로 골라 엔터, 작업 중간에도 전환 가능
- 복잡한 판단 = 오퍼스 · 단순 코드 = 소넷 · 잡무 = 하이쿠
- 무거운 일만 비싼 모델 — 비용 절반

PART 4 전체 5장

방법 ③ 대화 압축 — 짐 내려놓기

클로드코드는 매 요청마다 **지금까지의 대화를 통째로 다시 읽습니다**. 어제부터 이어온 긴 대화라면, "버튼 색 바꿔줘" 한 마디에도 그 긴 기록 전체가 함께 전송됩니다. 대화가 길수록 모든 요청이 비싸지는 구조입니다.

그래서 두 가지 습관이 필요합니다.

/compact — **대화 압축**. 지금까지 대화를 요약해 핵심만 남깁니다. 타이밍은 두 가지입니다.

- 한 작업을 마무리했을 때 (기능 하나 완성, 버그 하나 해결)
- 대화가 길어졌다고 느낄 때 (스크롤이 한참 올라갈 때)

/clear — **완전 리셋 (보너스)**. 완전히 새로운 작업을 시작한다면 압축보다 리셋이 낫습니다. 지난 대화가 새 작업에 필요 없다면 들고 갈 이유가 없으니까요. 습관은 간단합니다 — **"새 일 = 새 대화."**

⚡ 한 단계 더 — 한도가 임박하면 클로드코드가 알아서 압축하는 자동 압축도 있습니다. 하지만 자동은 타이밍을 못 고릅니다. 작업이 일단락됐을 때 손으로 **/compact** 하는 쪽이 요약 품질도, 절약 효과도 좋습니다.

◆ 이 PART의 핵심 정리

- 매 요청마다 대화 전체가 다시 전송된다 — 길수록 비싸진다
- 작업 일단락 = **/compact**, 완전히 새 일 = **/clear**
- 자동 압축보다 손 압축이 타이밍이 좋다

PART 5 전체 5장

CLAUDE.md 다이어트 체크리스트

마지막으로 규칙 파일 자체를 점검합니다. CLAUDE.md 는 **매 요청마다 통째로 전송**됩니다. 규칙 파일이 비대하면 아무것도 안 해도 상시로 토큰이 나가는 셈입니다. 아래 체크리스트로 한 번 훑어보세요. 인쇄해 두고 월 1회 점검을 추천합니다.

- ☐ **출력 규칙 3줄**이 들어 있다 (PART 2)
- ☐ 안 쓰는 규칙, 철 지난 규칙을 지웠다
- ☐ 긴 코드 예시는 지우고 **파일 경로만** 남겼다 ("자세한 건 이 파일 참고" 식으로)
- ☐ 반복해서 시키는 일은 CLAUDE.md 가 아니라 **커맨드/스킬 파일**로 뺐다 (필요할 때만 읽게)
- ☐ 이 프로젝트와 무관한 규칙은 **전역 CLAUDE.md** (`~/.claude/CLAUDE.md`) 로 옮겼다
- ☐ 전체 길이가 화면 두세 장을 넘지 않는다 (넘으면 나눠서 링크)
- ☐ 안 쓰는 MCP(클로드에 연결하는 외부 도구)를 정리했다 — 연결만 돼 있어도 자리를 차지합니다
- ☐ 습관 두 개가 몸에 붙었다 — 새 일 = `/clear` , 긴 대화 = `/compact`

여기까지 하셨다면 세팅 끝입니다. **출력은 짧게, 모델은 일에 맞게, 대화는 가볍게** — 이 세 가지면 같은 요금제로 두 배 갑니다.

더 많은 클로드코드 무료 자료는 docs.vibemake.kr, 새 자료 소식은 인스타그램 [@vibemak2r](https://www.instagram.com/vibemak2r) 에서 이어집니다.

◆ 이 PART의 핵심 정리

- CLAUDE.md 는 매 요청마다 전송된다 — 파일 자체도 다이어트 대상
- 코드 예시는 경로로, 반복 지시는 커맨드로, 무관한 규칙은 전역으로
- 출력 짧게 · 모델 맞게 · 대화 가볍게 = 같은 요금제로 두 배