LinklBee

AI API 비용 절감 엔진

링클비 제품소개서

LLM API Cost Optimization

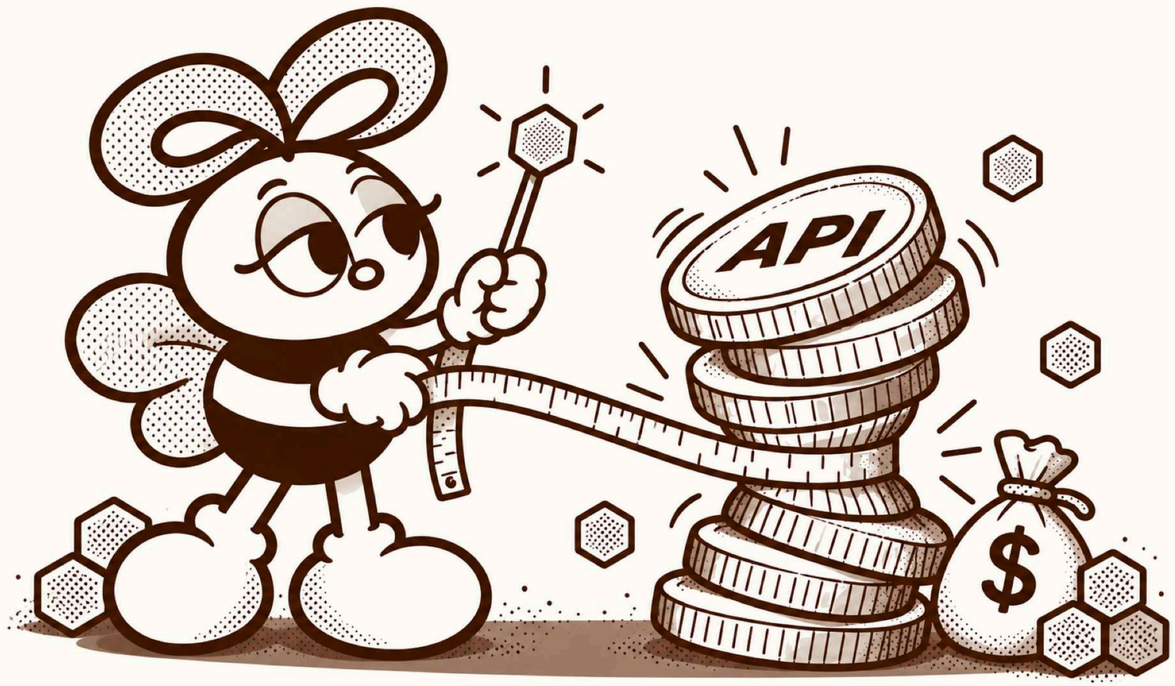
LinkIBee

BUILD MORE. SPEND SMARTER.

AI는 마음껏,
API 비용은 가볍게.

LLM API 비용 최적화, 링클비.
호출 패턴에서 개선 기회를 찾고,
서비스에 맞는 최적화를 운영에 연결합니다.

무료 비용 진단 | 지속 최적화 | 성과형 과금



LLM API를 쓴다면,
링클비부터.

AI API로 서비스를 개발·운영하는 기업과 개발팀을 위한 솔루션입니다.
소비자용 앱부터 기업용 솔루션까지, 실제 API 호출 방식에서 개선 기회를 찾습니다.
개인 개발자의 첫 서비스부터 대규모 운영까지, 비용 효율을 함께 설계하세요.

LinkIBee

01 BUILT AROUND YOUR CALLS

챗봇부터 AI 에이전트까지,
API 비용 절감은 링클비로.

고객이 운영하는 다양한 AI 서비스에서, 실제 API 호출 방식에 따른 최적화 기회를 찾습니다.

대화형 서비스

반복 지시문과 대화 요청을 살펴
캐시 활용 기회를 찾습니다.

AI 에이전트

도구 정의와 다단계 호출을 분석해
불필요한 재요청을 찾습니다.

검색·RAG

공통 지시문과 반복 질의를 살펴
재사용 가능한 구간을 확인합니다.

콘텐츠 생성

수정 요청에서 유지할 내용과
새로 생성할 부분을 구분합니다.

분류·추출·배치

응답 시간 조건에 맞는 처리 경로로
비용 효율을 높일 기회를 찾습니다.

업무 자동화

반복 호출과 출력 길이를 살펴
재시도 설정의 개선점을 찾습니다.

다양한 활용 사례의 예시입니다. 실제 적용 항목은 호출 데이터와 서비스의 품질·응답 시간 조건에 따라 달라집니다.

LinkIBee

02 DIAGNOSE

API 비용, 어디서 새는지 어떻게 줄일지 찾아드립니다.

호출 로그에서 주요 비용 발생 패턴을 분석하고, 항목별 개선 방법과 우선순위를 정리합니다.

01

캐시 적용 지점

반복 프롬프트가 캐시 조건과 적용 위치에 맞게 구성됐는지 확인합니다.

02

프롬프트 구조

같은 내용의 키 순서와 직렬화 차이로 생기는 캐시 누락을 찾습니다.

03

출력 제한·재요청

출력 잘림과 후속 재요청으로 발생하는 추가 비용을 확인합니다.

04

처리 경로

실시간 응답이 필요 없는 요청의 대체 처리 방식을 검토합니다.

05

부분 재생성

기존 응답을 활용하고 변경된 부분만 생성할 기회를 찾습니다.

06

응답 재사용

기존 응답으로 처리할 수 있는 유사·반복 요청을 확인합니다.

중복 비용은 한 번만 반영하고, 추가 데이터가 필요한 항목은 구분해 표시합니다.

기존에 캐싱이나 자체 최적화를 운영하는 팀도 추가 개선 지점을 검토할 수 있습니다.

03 OPTIMIZE

우리 서비스에 맞게,
API 비용 절감도 알아서.

운영 환경에 맞는 설정 개선과 엔진의 지속 최적화를 연결합니다.

설정 개선

캐시 구성, 프롬프트 구조, 출력 제한과 처리 경로.
진단 결과에 따라 개선 항목과 적용 절차를 안내합니다. 서비스 코드 변경이 필요한 항목은 연동 과정에서 함께 확인합니다.

지속 최적화

부분 재생성과 응답 재사용.
엔진이 트래픽 변화를 살피며 적용 여부와 수준을 자동으로 조정합니다.

응답 재사용·부분 재생성의 흐름

새 요청
이전 응답과 비교

→

유지 | 유지 | 갱신
활용 가능한 내용은 보존

→

필요한 만큼 생성
호출·생성량 최적화

품질과 비용을 함께 평가합니다. 지속 최적화는 품질 기준 미달 시 기존 처리 방식으로 되돌립니다.

04 MEASURED EXAMPLE

API 비용, 얼마나 줄었을까요?
실험에서 38.2–57.0% 줄였습니다.

문서 개정 작업 · 내부 측정

38.2–57.0%

전체 재생성 대비 비용 절감

링클비의 최적화 기능 중 부분 재생성의 측정 사례입니다.

평가 데이터	건수	절감률	품질 지표[1]
법령 개정	317	**38.2%**	0.978
eCFR 개정	79	**56.9%**	0.988
의약품 허가사항 개정	45	**57.0%**	0.972

동일 모델로 전체 재생성과 부분 재생성을 비교했습니다.
공개된 다음 버전을 기준으로 품질을 평가하고,
실패 후 재시도 비용을 포함했습니다.

링클비 내부 측정 · 공개 문서 개정 데이터 · 2026.10.17–19

[1] 내부 품질 지표이며 정확도 백분율이 아닙니다. 이 결과는 해당 문서 개정 실험에 한정되며, 서비스별 절감 효과는 도입 후 비교 측정합니다.

GPU MEASURED EXAMPLE

같은 GPU로,
반복 계산은 줄이고 처리 시간은 짧게.

이미 처리한 입력을 다시 계산하는 낭비를 줄여,
GPU가 같은 작업을 더 빠르게 처리하도록 돕습니다.

로컬 GPU 2대 · 내부 측정

다시 계산한 입력 토큰

49.1% 감소

대화 16개 처리 시간

407초 → 203초

**정액 임대 환경에서는 연산 효율과 작업 시간 개선을 의미하며,
임대료 감소를 뜻하지 않습니다.**

측정 조건 · 로컬 GPU 2대, 대화 16개. 서버 배정 쏠림이 남아 있는 환경에서 측정했습니다.

LinkIBee

05 DEPLOY & VERIFY

API 비용 절감이 없다면,
성과형 이용료는 0원.

무료 진단부터 연동, 효과 검증까지. 기존 운영 흐름을 기준으로 도입 범위를 정합니다.

01

무료 진단

로컬에서 로그를 분석해
개선 항목과 우선순위를 확인합니다.

02

연동·적용

서비스의 품질·응답 시간 조건을 확인하고, 운영
환경에 맞춰 링클비 엔진을 연동합니다.

03

비교·정산

기존 처리 방식과 비교해
실제 절감 효과를 확인합니다.

클라우드부터 폐쇄망까지

클라우드 관리형과 고객 서버 운영형을 지원합니다.
배포 방식과 데이터 처리 범위는 도입 시 협의합니다.

절감액에 연동한 성과형 과금

초기 비용 없이 시작하고, 측정된 절감액을 기준으로 정산합니다.
절감액이 없으면 성과형 이용료도 없습니다.

일부 요청을 기존 방식으로 처리하는 비교군을 운영하고, 공급사 청구 기준으로 효과를 평가합니다.
비교군 비용도 절감률 산정에 포함합니다. 적용 요율과 계약 조건은 개별 협의합니다.

LinklBee

06 START WITH LINKLBEE

내 API 비용, 얼마나 줄일 수 있을까? 무료 진단으로 확인하세요.

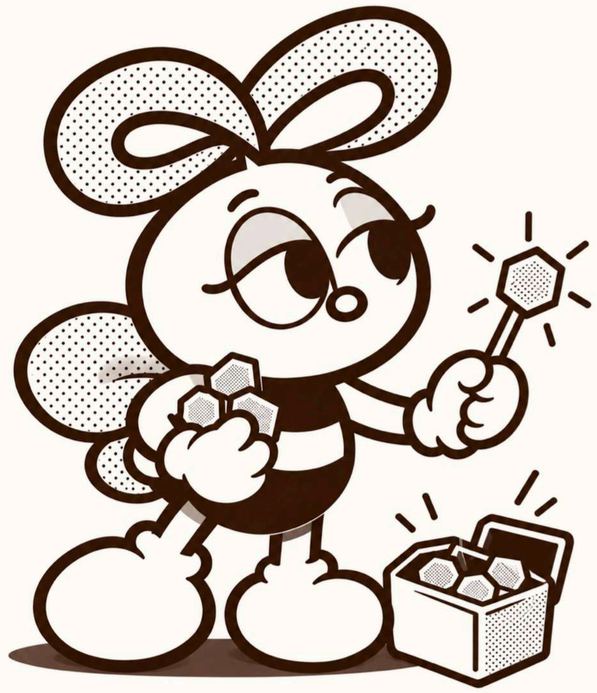
첫 단계는 무료 로컬 진단입니다. 사용 중인 모델과 주요 기능으로 도입 상담도 시작할 수 있습니다.

로그는 로컬에서 분석

개발·운영팀의 로컬 환경에서 분석하며, 원본 로그는 외부로 전송하지 않습니다.

가볍게 시작

회원가입·API 키 없이 실행합니다. Python 3.9 이상, 추가 패키지는 필요 없습니다.



검토 가능한 공개 코드

진단기는 MIT 라이선스로 공개됩니다. 리포트에는 요청·응답 원문 대신 비용과 패턴을 담습니다.

사용 중인 모델과 주요 기능을 알려주시면 도입 상담을 시작할 수 있습니다. 호출 규모를 알고 있다면 함께 알려주세요.

무료 진단으로 시작하기

bee@linklbee.com / github.com/linklbee/scan

linklbee.com →

감사합니다.

회사명 ㈜낫에그

설립일 2016.9.15

주소 서울특별시 강남구 테헤란로79길 6 제이에스타워 V923 낫에그

메일 bee@linklbee.com

홈페이지 linklbee.com

www.noteggparan.co.kr

@notegg.co.kr

LinklBee