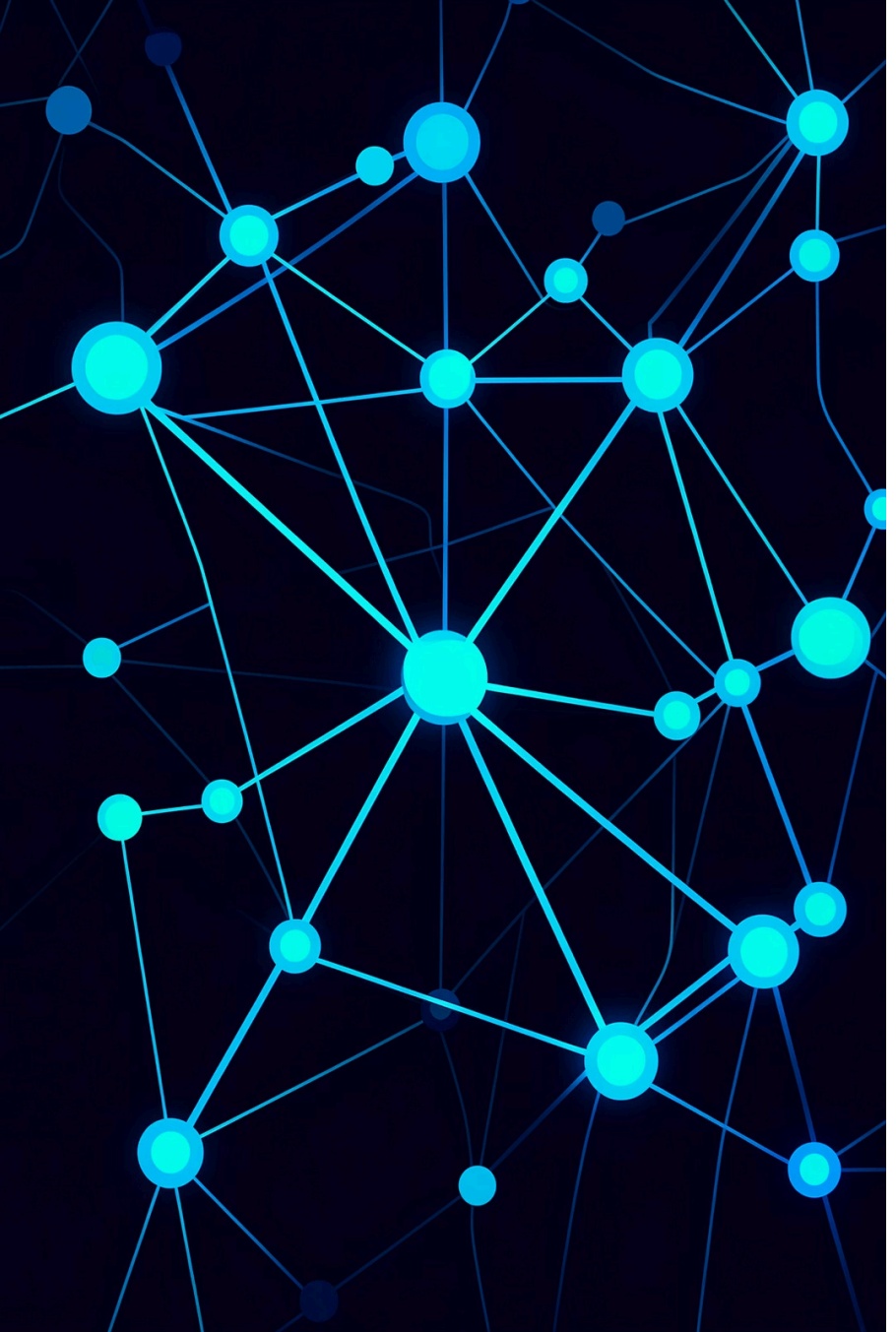


AI는 더 똑똑해졌지만, 효율을 계산하는 방식도 함께 달라졌다

토큰 인플레이션 논쟁이 보여주는 LLM 평가 기준의 변화

새로운 모델이 출시될 때마다 개발자와 PM들은 묻는다. "이전보다 더 효율적인가?" 하지만 이 질문에 답하기 위한 기준 자체가 조용히 달라지고 있다. 이 발표는 그 변화를 함께 들여다본다.

배포: Deformatic | 유민수 개발자

왜 이 이야기가 갑자기 뜨거워졌는가

개발자 커뮤니티에서 조용히 번지기 시작한 불편함이 있다. "같은 프롬프트인데, 왜 요즘 토큰을 더 많이 먹지?" 누군가는 API 청구서를 보고 고개를 갸웃하고, 누군가는 rate limit에 더 빨리 걸린다고 느낀다.

커뮤니티가 느끼는 것

일부 사용자들은 특정 사용 패턴에서 이전 모델 대비 30~45% 수준의 토큰 증가를 체감한다고 보고한다. 하지만 이것은 **공식적으로 검증된 보편값이 아니라**, 커뮤니티 관측치임을 먼저 명확히 해야 한다.

⚠ ※ 커뮤니티 관측치 — 공식 수치 아님

공식 설명이 말하는 것

Anthropic의 공식 문서에 따르면, Claude Opus 4.x 계열은 **새로운 tokenizer**를 사용한다. 공식적으로 방어 가능한 설명은, 같은 텍스트가 이전 대비 약 **1.0~1.35배** 많은 토큰으로 계산될 수 있다는 것이다. 가격표 자체는 이전과 동일하게 유지된다.

ⓘ ※ 공식 문서 기반 설명

이 두 층위 사이의 긴장 — 커뮤니티의 체감과 공식 설명의 간극 — 이 바로 이 논쟁을 뜨겁게 만든 진짜 이유다.

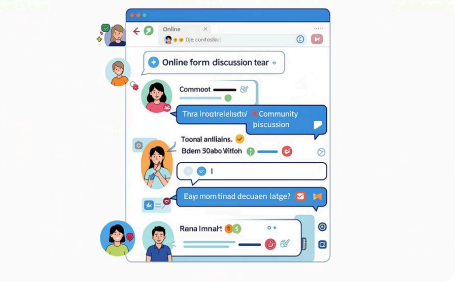
두 자료는 사실 같은 질문을 한다

이 발표는 두 가지 자료에서 출발한다. 표면적으로는 다른 형식이지만, 두 자료는 동일한 질문을 서로 다른 층위에서 다루고 있다. **"모델 업그레이드는 정말 더 효율적인가?"**

Bill Chambers의 Tokenomics Leaderboard

측정 레이어 (Raw Signal)

다양한 모델들의 실제 토큰 소비 패턴을 체계적으로 측정하고 비교한다. 어떤 모델이 같은 작업에 얼마나 많은 토큰을 쓰는지, 그 raw signal을 제공하는 '계측 레이어'다. 주관적 해석 없이 측정값 자체를 보여준다는 점에서 신뢰할 수 있는 기준점이 된다.



Hacker News 토론 스레드

해석 레이어 (Interpretation)

개발자, PM, 연구자들이 그 측정값을 보고 어떻게 해석하는지를 담는다. "이것이 의도된 설계인가, 아닌가?", "우리 워크플로우에 어떤 영향을 주는가?" 같은 질문들이 오간다. 같은 숫자를 두고 전혀 다른 결론이 나오는 것을 볼 수 있는 '해석 레이어'다.

하나는 **무엇이 일어나고 있는지**를 보여주고, 다른 하나는 **그것이 무엇을 의미하는지**를 논쟁한다. 이 둘을 함께 읽을 때 전체 그림이 보인다.

먼저 오해부터 정리하자

논쟁이 뜨거워지면 단순화가 일어난다. "토큰이 늘었으니 비싸졌다"는 결론은 직관적이지만, 실제로는 더 복잡한 구조를 갖는다. 오해와 현실을 나란히 놓고 보자.

🚫 오해: 토큰이 늘면 무조건 더 비싸다

- 새 tokenizer → 토큰 수 증가 → 비용 자동 상승
- 모델 업그레이드는 항상 효율을 나쁘게 만든다
- 같은 프롬프트 = 같은 비용이어야 한다
- 토큰 인플레이션은 사용자에게 불리한 변화다

✅ 현실: 토큰 증가와 task 비용은 다를 수 있다

- input 토큰이 늘어도 output 토큰이 줄거나 반복 횟수가 줄면 총비용은 낮아질 수 있다
- 모델 성능 향상으로 같은 task를 더 적은 시도로 완료할 수 있다
- 일부 분석에서는 총 평가 비용이 오히려 낮아진 사례도 보고된다
- 비용보다 더 중요한 것은 task 완료 기준 효율이다

> "토큰이 늘었다는 사실보다 더 중요한 것은, 무엇을 효율이라고 부를 것인가다."

Token은 '자연 단위'가 아니다

이 논쟁의 가장 근본적인 지점은 여기서 시작한다. 많은 사람들이 토큰을 마치 '글자 수'나 '단어 수'처럼 고정된 자연 단위로 생각하지만, 사실 토큰은 **tokenizer가 설계한 측정 단위**다.

1 비유: 자로 재는 방식이 바뀌면 숫자도 바뀐다

같은 방의 길이를 cm로 재면 300이 나오고, inch로 재면 118이 나온다. 방의 크기는 변하지 않았지만 숫자는 달라진다. 토큰도 마찬가지다. 같은 텍스트라도 어떤 tokenizer를 쓰느냐에 따라 토큰 수가 달라진다. 이것은 버그가 아니라 설계 선택이다.

2 모델 버전마다 tokenizer가 다를 수 있다

Claude Opus 4.x가 새로운 tokenizer를 도입한 것처럼, 모델 세대가 바뀔 때 tokenizer도 함께 바뀔 수 있다. Anthropic 공식 문서에 따르면, 이 변화로 인해 동일한 텍스트가 이전 대비 약 1.0~1.35배 많은 토큰으로 계산될 수 있다. 이 범위는 텍스트의 종류와 언어에 따라 달라진다.

3 따라서 토큰은 절대 기준이 아니라 설계된 기준이다

토큰 수 자체를 모델 간 비교의 절대 기준으로 삼으면 잘못된 결론에 이를 수 있다. "같은 일"을 비교하려면, 토큰 수가 아니라 그 일의 결과물 — 즉 task 완료 여부, 품질, 소요 시간 — 을 기준으로 삼아야 한다.

무엇이 실제로 바뀌었나

새 모델로 전환할 때 달라지는 요소들을 구체적으로 짚어보자. 이 변화들은 서로 독립적이지 않고, 함께 작용하면서 전체 비용 구조에 영향을 준다. 각 변화를 공식 정보 기반으로 정리한다.

① Tokenizer 변화

공식 문서 기반: Claude Opus 4.x 계열은 새로운 tokenizer를 사용한다. 동일한 입력 텍스트가 이전 대비 약 1.0~1.35배 많은 토큰으로 계산될 수 있다. 이는 언어, 특수문자, 코드 구조 등 텍스트 유형에 따라 다르게 적용된다. 한국어, 일본어 등 비영어권 언어에서 그 차이가 더 두드러질 수 있다.

② Reasoning 방식 변화

공식 문서 기반: 높은 effort 설정이나 agentic 작업의 후반 턴에서는 output token이 더 늘어날 수 있다. 모델이 더 깊이 "생각"하는 과정이 토큰으로 표현되기 때문이다. 이는 복잡한 추론이 필요한 작업에서 특히 두드러진다.

③ Output 구조 변화 가능성

관찰 기반 (공식 확인 제한적): 모델이 더 구조화된 응답이나 더 상세한 설명을 생성하는 방향으로 튜닝되면, output token 수 자체가 달라질 수 있다. 이는 사용 목적에 따라 긍정적일 수도, 부정적일 수도 있다.

④ 가격표는 그대로, 체감은 달라질 수 있다

공식 확인: 가격 자체(per million tokens)는 이전 모델과 동일하게 유지된다. 그러나 같은 작업에 소모되는 토큰 수가 달라질 수 있으므로, 실질적인 청구 금액은 사용 패턴과 설정에 따라 달라질 수 있다.

왜 사람들이 유독 민감하게 반응하는가

단순히 "토큰이 조금 더 쓰인다"는 사실 자체가 이렇게 큰 논쟁을 만들지는 않는다. 이 반응의 배경에는 실무적인 민감도가 겹겹이 쌓여 있다.

예산과 과금

API 비용은 팀의 예산과 직결된다. 예상치 못한 비용 증가는 스타트업에서 즉각적인 위기로 이어질 수 있다. 특히 소량 테스트에서 괜찮아 보이다가 프로덕션 규모에서 폭증하는 패턴이 두렵다.

Rate Limit & 한도

많은 플랜에서 토큰 사용량은 rate limit의 기준이 된다. 같은 작업에 더 많은 토큰이 필요해지면 한도에 더 빨리 부딪히고, 이는 서비스 중단이나 응답 지연으로 이어진다.

팀과 제품 운영

에이전트 파이프라인, 자동화 워크플로우, 배치 처리 시스템은 모두 예상 토큰 소비량을 기반으로 설계된다. 기준이 바뀌면 아키텍처 전반을 재검토해야 할 수 있다.

체감 공정성

가격표는 그대로인데 실질 비용이 달라진다면, 사용자는 "룰이 몰래 바뀐 것"처럼 느낀다. 이 체감 불공정성이 커뮤니티 반응을 증폭시키는 핵심 감정 요인이다.

이 민감도는 단순한 불평이 아니다. 실무에서 토큰은 돈이고, 시간이고, 한계이기 때문이다.

하지만 반대쪽 주장도 중요하다

균형 있는 이해를 위해, "토큰 증가 = 비용 증가"라는 단순 논리에 반하는 주장들도 진지하게 살펴봐야 한다. 이 주장들은 단순한 반론이 아니라 실질적인 데이터와 논리를 갖추고 있다.

Output이 줄면 총비용이 낮아질 수 있다

Input 토큰이 늘어도, 모델이 더 정확하게 핵심만 답한다면 output 토큰은 오히려 줄어들 수 있다. 토큰 비용 구조에서 output은 input보다 일반적으로 더 비싸게 과금되므로, output 감소가 전체 비용을 낮출 수 있다.

반복 횟수가 줄면 총 소비량이 낮아진다

성능이 향상된 모델은 같은 task를 더 적은 시도(turn)로 완료할 수 있다. 에이전트 워크플로우에서 3번 시도하던 것을 1번에 성공한다면, 1회당 토큰이 약간 늘어도 전체 총합은 낮아진다. 일부 분석 사례에서 실제로 이런 패턴이 관찰된다고 보고된다.

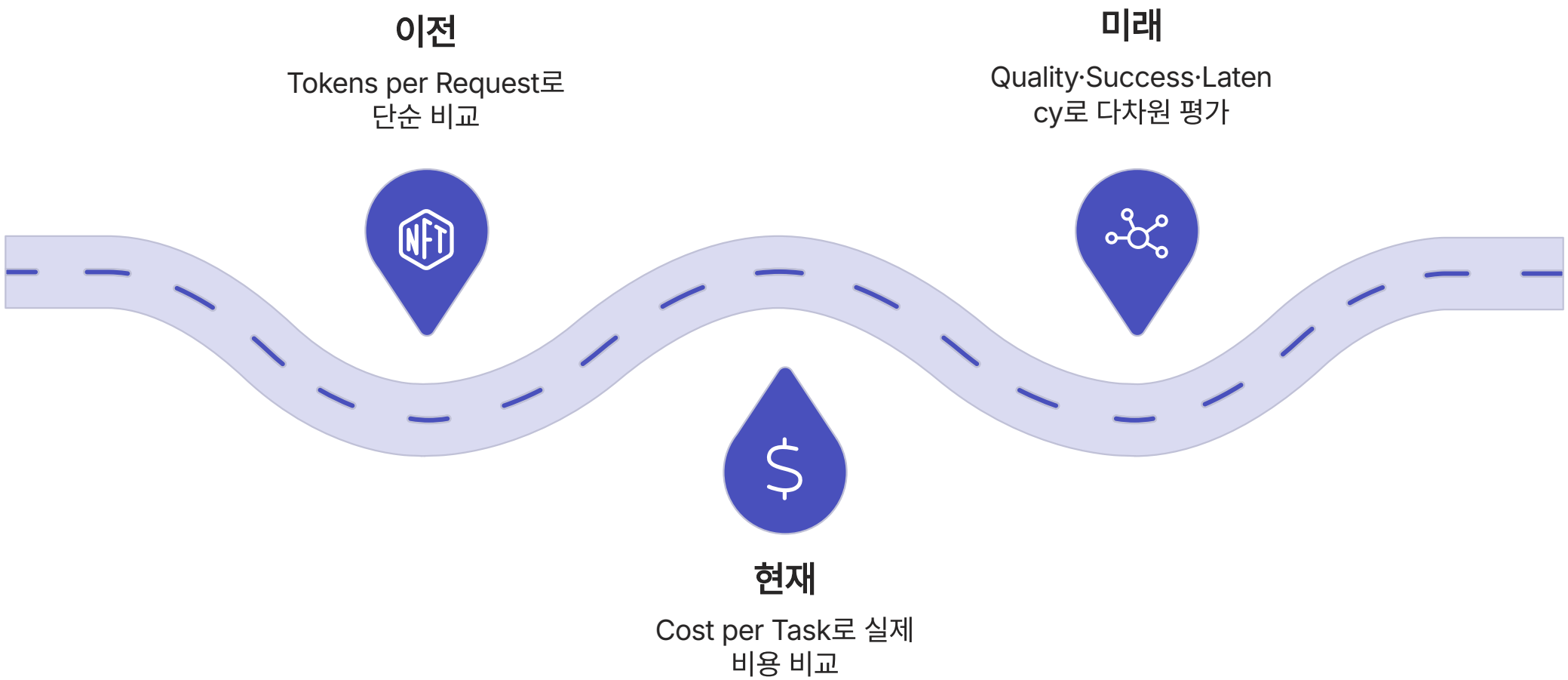
품질 향상이 human review 비용을 낮춘다

API 비용만이 전체 비용의 전부가 아니다. 결과물을 검토하고 수정하는 데 드는 인건비, 재시도 비용, QA 비용까지 포함하면, 더 정확한 모델은 전체 운영 비용을 오히려 낮출 수 있다. 이 관점에서 "비싼 모델이 더 저렴할 수 있다."

> "토큰 증가가 곧 전체 비용 증가를 뜻하지는 않는다. 효율은 전체 파이프라인에서 측정해야 한다."

진짜 핵심: 효율의 기준이 이동하고 있다

이 모든 논쟁을 관통하는 진짜 메시지는 여기에 있다. 측정 기준이 달라지고 있다. 그리고 이 변화는 단순한 기술 업데이트가 아니라, LLM을 어떻게 평가해야 하는가에 대한 패러다임 전환에 가깝다.



이 흐름은 단순히 "더 좋은 지표를 쓰자"는 것이 아니다. 모델의 '좋음'을 정의하는 방식 자체가 바뀌고 있다는 신호다. Tokens per request는 비교가 쉽고 직관적이지만, 실제 비즈니스 가치와의 연결이 약하다. Cost per task는 한 발 더 나아가 실질적 지출을 기준으로 삼는다. 그리고 궁극적으로 우리가 향해야 할 방향은 quality/task, success rate/task, latency/task, human review load 같은 다차원 복합 지표다.

"효율이 나빠진 것이 아니라, 효율을 재는 자가 바뀌고 있다."

실무자는 무엇을 다시 측정해야 하나

이제 구체적인 실무 질문으로 넘어가자. 모델을 전환하거나 새 버전을 도입할 때, 기존의 토큰 기반 벤치마크 외에 무엇을 추가로 측정해야 하는가?

📊 측정해야 할 새 지표들

Cost per Task (작업당 실비용)

동일한 task를 완료하는 데 드는 총 API 비용. 단일 요청의 토큰 수가 아닌, task 완료 단위로 집계한다. 에이전트 루프가 있다면 전체 루프 비용을 합산한다.

Success Rate per Task (작업 성공률)

모델이 주어진 task를 요구 기준 이상으로 완료하는 비율. 비용이 낮아도 성공률이 낮으면 실질 효율은 나쁘다. 반대로 비용이 높아도 성공률이 월등히 높으면 전체 운영은 더 저렴해질 수 있다.

Latency per Task (작업당 응답 시간)

사용자 경험과 직결되는 지표. 더 많은 토큰을 생성하면 지연이 늘어날 수 있다. 실시간 애플리케이션에서 latency는 비용 못지않게 중요한 제약 조건이다.

Human Review Load (인간 검토 부하)

자동화 파이프라인에서 사람이 개입해야 하는 빈도와 깊이. 모델 품질이 높아지면 이 부하가 줄고, 그것이 실질적인 비용 절감으로 이어진다.

🔧 조직/제품 관점 체크리스트

- 현재 벤치마크가 토큰 기반에만 치우쳐 있지는 않은가?
- 모델 전환 시 A/B 테스트 기준이 명확히 정의되어 있는가?
- 에이전트 파이프라인은 전체 턴 수를 기준으로 비용을 집계하고 있는가?
- output 품질 기준이 정량적으로 정의되어 있는가?
- tokenizer 변화를 반영한 예산 재산정을 했는가?
- 비용 알림(alert) 임계값이 새 모델 기준으로 업데이트되었는가?

ⓘ 이 체크리스트는 새 모델 도입 전 필수 점검 항목으로 활용할 수 있다.

이 변화가 보여주는 더 큰 흐름

토큰 인플레이션 논쟁은 표면적으로는 가격과 비용의 문제처럼 보이지만, 더 깊이 들여다보면 LLM 산업 전반의 구조적 변화를 가리키고 있다.

1

모델 회사의 가격표와 실제 체감 비용 사이의 긴장

가격은 공개되어 있지만, 실질 비용은 사용 패턴, tokenizer, effort 설정, 모델 버전에 따라 달라진다. 투명성의 부재가 커뮤니티 불신을 만든다. 이 긴장은 앞으로도 계속될 것이다.

2

평가 지표의 재정의

벤치마크가 토큰 중심에서 task 중심으로 이동하고 있다. Bill Chambers의 Tokenomics leaderboard 같은 시도가 바로 이 재정의의 최전선에 있다. 업계 표준이 어떻게 형성될지는 아직 열려 있다.

3

사용자의 평가 역량 성숙

개발자와 PM들이 단순 토큰 수 비교를 넘어, 더 정교한 평가 방법론을 요구하기 시작했다. 이는 커뮤니티의 성숙을 의미하며, 모델 회사들에게도 더 투명하고 구체적인 설명을 요구하는 압력이 된다.

"같은 작업이라도 모델이 바뀌면 토큰 수가 달라질 수 있다." 이 단순한 사실이 이제 단순한 기술 팁이 아니라, 제품 전략과 예산 계획을 좌우하는 핵심 이슈가 되었다. 그리고 이 논쟁은 LLM이 더 본격적으로 비즈니스 인프라로 자리잡을수록 더 중요해질 것이다.

발표 핵심 요약

이 발표에서 가장 중요하게 기억해야 할 다섯 가지 메시지를 정리한다.

01 — Token은 설계된 단위다

토큰은 자연 단위가 아니라 tokenizer가 정의하는 측정 단위다. 모델이 바뀌면 같은 텍스트의 토큰 수도 달라질 수 있다. 이것은 버그가 아니라 설계 선택이다.

02 — 토큰 증가 ≠ 비용 증가

input 토큰이 늘어도 output이 줄거나 반복이 줄면 총비용은 낮아질 수 있다. 단일 지표로 전체 효율을 판단하지 말 것.

03 — 커뮤니티 수치는 참고용

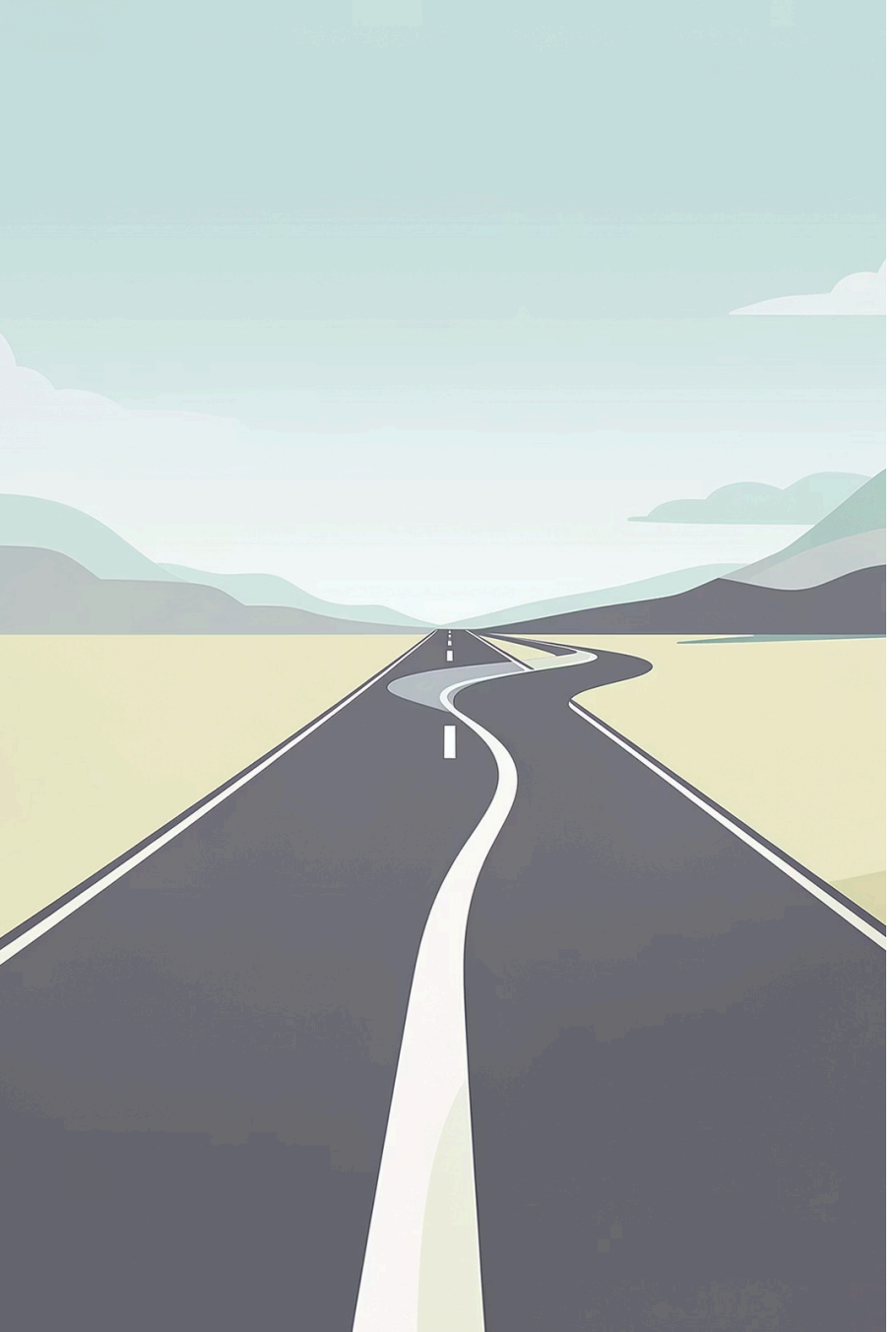
30~45% 체감 증가 등의 수치는 일부 사용자 관측치다. 공식적으로 방어 가능한 숫자(1.0~1.35배)와 구분해서 해석해야 한다.

04 — 기준이 이동하고 있다

이제 중요한 것은 tokens per request가 아니라 cost per task다. 더 나아가 quality, success rate, latency, human review load 같은 복합 지표로 나아가야 한다.

05 — 측정의 책임은 사용자에게도 있다

모델 회사의 가격표만 믿지 말고, 실제 사용 패턴을 기반으로 한 자체 벤치마크를 구축하는 것이 이제 실무의 기본 역량이 되었다.



우리는 이제 더 좋은 모델을 고르는 것이 아니라, 더 적절한 기준을 고르는 시대에 들어왔다.

모델의 성능이 상향 평준화되고, 각 회사의 가격 정책이 복잡해질수록 — 어떤 기준으로 비교하고 어떻게 측정할지를 결정하는 역량이 곧 경쟁력이 된다.

이 발표가 그 판단의 시작점이 되기를 바란다.

배포: Deformatic | 유민수 개발자

주요 출처

이 발표에서 인용하거나 참고한 주요 자료들을 아래에 정리한다. 공식 문서와 커뮤니티 자료를 구분하여 표기한다.

분류	출처명	비고
공식 문서	Anthropic — Claude Opus 4 / 4.5 Release Notes	공식 tokenizer 변화 기술 근거
공식 문서	Anthropic — Model Migration Guide (Claude 3 → 4 계열)	tokenizer 변화 및 호환성 설명
리더보드/계측	Bill Chambers — Tokenomics Leaderboard (GitHub/HuggingFace)	모델별 토큰 소비 패턴 측정 레이어
커뮤니티 토론	Hacker News — 관련 토론 스레드 (Tokenomics / Claude Opus 비용 논쟁)	해석 레이어, 커뮤니티 관측치 출처
분석 자료	Artificial Analysis — Claude Opus 4.x 성능 및 비용 분석 아티클	제3자 벤치마크 참고

⚠ 커뮤니티 출처(Hacker News 등)에서 인용한 수치(예: 30~45% 체감 토큰 증가)는 공식적으로 검증된 수치가 아닌 일부 사용자 관측치임을 재차 강조한다. 공식 문서와 커뮤니티 관측치를 반드시 구분하여 해석할 것을 권장한다.