GLOBAL
EQUITY

2026.04. 15

글로벌주식팀

김중한 Senior Analyst
joonghan1.kim@samsung.com

이영진 Senior Analyst, CFA
youngjin91.lee@samsung.com

리서치센터 리포트
바로가기

글로벌 AI

클로드 성능 저하 논란, 배경은 컴퓨트 부족

- 클로드 성능 저하 논란이 세션 로그 분석 데이터를 통해 재점화
- 앤스로픽은 의도적인 성능 저하가 아닌 설정값 조정에 따른 체감 성능 변화라는 입장. 다만 체감 성능 약화는 유저 이탈로 이어질 수 있음
- 이번 이슈의 본질은 컴퓨트 병목. 인프라 보유 기업의 해자가 부각되는 국면 진입

WHAT'S THE STORY?

재점화된 클로드 성능 열화: AMD의 AI 디렉터가 공개한 세션 로그 분석을 계기로, 2월 이후 클로드의 추론 길이 축소와 작업 방식 변화가 객관적 데이터로 재조명. 복잡한 엔지니어링 과제에서 생각 깊이가 얕아지고, 작업 전 충분한 탐색 없이 수정에 먼저 들어가는 패턴이 강화되면서 답변 품질과 생산성이 저하. 여기에 작업 검증 생략 건수(Stop Hook 위반) 증가, 프롬프트 내 불만 표시 비중 증가, 세션당 프롬프트 수 감소 등 사용자 체감 성능 저하 데이터도 제시

앤스로픽의 입장 – 의도적인 성능 저하가 아닌 효율화의 결과: 앤스로픽은 이번 변화가 절대적 성능 저하가 아니라 Redacted Thinking, Adaptive Thinking, Effort 기본값 조정 등을 통한 토큰 효율화와 응답속도 개선의 결과라는 입장. 즉 내부 추론 능력이 훼손된 것이 아니라, 문제 난이도에 따라 사고 길이를 조절하고 불필요한 추론 표시를 줄이는 과정에서 나타난 변화라고 설명. 다만 기업 설명과 별개로 실제 사용자 입장에서는 복잡한 작업에서의 성능 체감 약화가 분명히 존재할 수 있으며, 이는 모델의 절대 성능과 별개로 배포 단계에서의 경쟁력 저하 및 유저 이탈 가능성으로 이어질 수 있음

결국 핵심은 컴퓨트 병목: 이번 논란은 급증하는 수요를 컴퓨트 증설 속도가 따라가지 못한 병목이 핵심. 제한된 캐파 내에서 모든 수요를 충족해야 하는 상황 속에서 토큰 효율화와 서빙 최적화는 사실상 불가피한 전략. 최근 앤스로픽이 구글, 코어위브 등과의 계약을 통해 컴퓨트 확보에 적극적으로 나서고 있는 것도 이와 같은 맥락으로 해석. 더 나아가 차세대 초대형 모델 Mythos까지 감안하면, 이를 안정적으로 배포할 수 있는 전력, 데이터센터, GPU 등 인프라 확보 능력이 중요한 해자로 부각될 전망. 결국 AI 수요 급증에 따른 만성 컴퓨트 부족 현상으로 인해 하이퍼스케일러, 네오클라우드 등 인프라를 직접 보유한 기업으로의 수혜가 지속될 것으로 기대

(다음 페이지에 계속)

Claude(클로드) 성능 저하 논란 재점화

앤스로픽 클로드의 성능 저하 이슈가 AMD의 AI 디렉터 스텔라 로렌조(Stella Laurenzo)가 4/2일 깃허브에 게시한 Claude Code 세션 로그 분석으로 재점화되었다. 지난 1월 사용량 급증 이후 커뮤니티 중심으로 꾸준히 제기되던 이슈지만, 산업 내 저명한 인물이 정량 데이터에 근거해 문제를 제기했다는 점에서 시장의 관심이 다시 집중되고 있다.

데이터 기반으로 2월 기점으로 복잡한 엔지니어링 작업에서 Claude의 성능 저하를 확인했다는 주장이다. 가장 중요한 근거로 2월 중순~3월 초 사이 모델 생각의 깊이가 눈에 띄게 얕아졌다는 점이 제시되었다. 모델의 추론 길이(Signature Length)는 1월 말~2월 초 2,200자 → 2월 말 720자 → 3월 560~600자 수준으로 감소했다. 모델의 내부 추론 과정을 숨기는 업데이트인 Redacted Thinking 이후 유저들의 성능 불만이 본격적으로 확산되기 시작했다는 내용이다. 실제로 Redacted Thinking 비중이 50% 이상을 돌파한 시점과 사용자들로부터 성능 저하가 보고되기 시작한 시점이 일치한다.

복잡한 작업 수행 방식 변화도 성능 저하의 근거로 제시됐다. 과거에는 충분한 탐색과 맥락 파악 이후 수정에 들어가는 '선 조사(research-first)' 방식으로 수행했다면, 지금은 사전 작업을 짧게 가져가고 바로 작업에 돌입하는 '선 수정(edit-first)' 경향이 강해졌다. 'Read:Edit(파일 편집 전 읽기 비율)'은 1월 말~2월 초 6.6배에서 3월 초 이후 2.0배로 감소했으며, 'Research:Mutation(수정 전 조사 비율)'은 같은 기간 8.7배에서 2.8배로 감소했다. 작업 수행 전 주변 맥락이나 관련 파일 등을 확인하는 과정이 축소되었다는 의미로 결과적으로 답변 품질 저하로 이어졌다는 분석이다.

사용자 체감 악화를 보여주는 추가 지표도 제시됐다. 3월 8일 전후 비교에서 작업 검증 생략 건수(Stop Hook 위반)은 0건에서 173건으로 증가했고, 사용자 프롬프트 내 불만 표시 비중은 5.8%에서 9.8%로 증가했다. 세션 당 프롬프트 수는 35.9개에서 27.9개로 감소하며 단일 세션 내 에이전트의 작업 길이가 줄어들었다. 에이전트의 자율 업무 처리 능력의 감소는 사용자 개입을 증가시키는 방향으로 작동했다.

결론적으로 논란의 핵심은 클로드가 추론 길이를 축소하는 방식으로 작동하면서 복잡한 작업에서 성능과 생산성이 함께 저하됐다는 점에 있다. 토큰 효율성 개선으로 요청 당 비용이 낮아질 수 있지만, 복잡한 작업에서는 오류와 재작업이 증가해 오히려 시스템적 비효율성이 발생하고 있다. 실제로 API 요청 수는 2월 대비 3월에 80배 증가했다.

클로드의 내부 사고 내 추론 길이 변화

기간	추론 길이(자)	변화
1월 30일~2월 8일 (Baseline)	2,200	-
2월 말 기준	**720**	**-67%**
3월 1~5일	**560**	**-75%**
3월 12일 이후 (Redacted Thinking 100%)	**600**	**-73%**

자료: Github, 삼성증권

Read:Edit 및 Research:Mutation 비중 변화

자료: Github, 삼성증권

앤스로픽의 반박 – 의도적인 성능 저하가 아닌 효율화의 결과

앤스로픽에서 Claude Code를 담당하는 수석 엔지니어 보리스 체르니(Boris Cherny)는 Redacted Thinking 업데이트와 사고 길이 감소가 절대적 성능 저하가 아니라 효율성 제고의 결과로 해석해야 한다는 입장을 피력했다.

먼저 논란이 된 Redacted Thinking 업데이트는 내부 추론 과정을 바꾼 것이 아니라, 유저 인터페이스상 표시 방식을 조정한 것이라고 설명했다. 사고 과정에 대한 요약본을 제거해 오히려 응답 지연 시간을 단축하는 방향으로 작용한다는 주장이다. 또한, 유저가 원한다면 설정에서 추론 과정이 다시 표시되도록 변경할 수 있다고 덧붙였다.

사고 길이의 감소는 '적응형 사고(Adaptive Thinking)' 및 '추론 강도 조절' 업데이트와 연관되어 있다. 문제 난이도에 따라 추론 길이를 자동으로 조절하는 적응형 사고(Adaptive Thinking) 업데이트를 통해 추론 강도의 기본 설정 값이 High에서 Medium으로 변경됐다. 이는 토큰 사용량이 많다는 기존 유저 피드백을 반영한 효율성 개선 업데이트라는 점을 강조했다. 해당 기능 역시 Redacted Thinking과 마찬가지로 유저가 원한다면 설정에서 직접 변경할 수 있다는 코멘트를 남겼다.

의도적 모델 성능 저하에 대해서는 반박 의견이지만, 토큰 효율성 추구에 기반한 모델 업데이트가 성능에 영향을 줄 수 있다는 점을 간접적으로 인정한 셈이기도 하다. 또한 엔터프라이즈 고객의 성능 저하 논란에 대해서도 모든 유저에게 동일한 모델을 제공한다는 점을 명확히 했으며, 체감 차이는 설정 값의 영향일 수 있는 점을 강조했다.

결론적으로 이번 논란은 '절대적 모델 성능 열화'가 아닌 설정 값 변경에 따른 '성능 체감의 변화'라고 보는 것이 타당하다. 하지만 사용자 체감 성능 감소를 가볍게 지나칠 수 있는 사안도 아니다. 배포 단계에서 모델 성능 감소는 경쟁력 약화를 의미하며, 궁극적으로 유저 이탈로 이어질 수 있기 때문이다.

실제로 클로드 성능 저하로 인한 불만은 개발자를 시작으로 기업 고객으로 확산되고 있다. 이에 보리스 체르니는 팀(Teams) 및 엔터프라이즈(Enterprise) 플랜 유저의 모델 추론 강도 설정 값을 다시 High로 변경할 것이라는 계획을 밝혔다. 하지만, 일반 사용자 대상 설정 값 변동 계획은 언급되지 않았다. 상대적으로 높은 성능을 요구하는 엔터프라이즈와 비용 절감을 우선시하는 일반 소비자 사이에서 모델 배포 전략을 다르게 가져가는 목적이다.

클로드 코드 추론 강도(Effort 설정 값) 비교 표

수준	설명	주요 활용 사례
max	토큰 사용량 제한 없이 최대 성능을 구현하는 최고 성능 모드. Opus 4.6에서만 지원되며, 타 모델에서 사용 시 오류 반환	가장 깊은 추론과 정밀한 분석이 필요한 작업
high	고성능 모드. Medium 업데이트 이전의 기본 설정 값	복잡한 추론, 고난도 코딩, 에이전트형 작업
medium	성능과 속도, 비용 간 균형을 추구하는 모드로 토큰 사용량을 일정 수준 절감. 현재의 기본 설정 값	속도, 비용, 성능의 균형이 필요한 에이전트형 작업
Low	가장 높은 효율성을 지향하는 모드. 성능 일부를 희생하는 대신 토큰 사용량을 크게 절감	서브 에이전트 등 속도와 비용 효율이 중요한 단순 작업

자료: Anthropic, 삼성증권

결국 핵심은 컴퓨트 병목

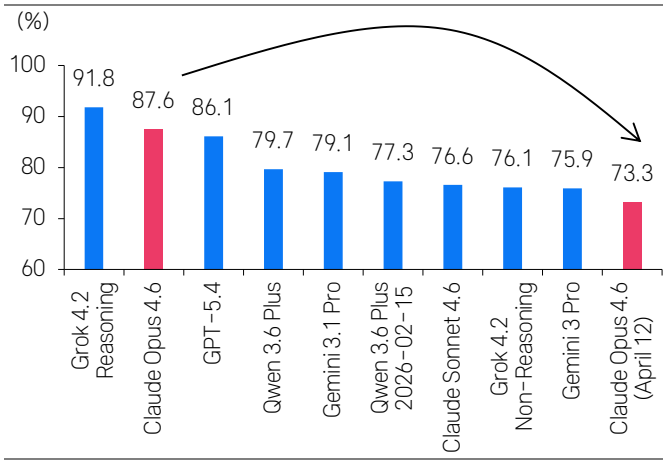
이번 성능 저하 논란의 배경에는 컴퓨트 부족 문제가 자리하고 있다. 앤스로픽의 ARR(연 환산 매출)은 '25년 말 90억 달러에서 4월 초 300억 달러로 빠르게 증가한 반면, 컴퓨트 용량은 동 기간 1.3GW에서 1.6GW 수준으로 증가하는데 그쳤다. 한정된 용량 내에서 모든 유저의 수요 충족을 위해 토큰 효율성에 기반한 모델 업데이트가 불가피했음을 알 수 있는 대목이다. 또한 연초 이후 컴퓨트 용량이 수요를 안정적으로 소화하지 못하면서 서버 안전성 문제도 제기되어 왔다.

업계 전반에서 컴퓨트가 핵심 경쟁 요소로 부상하고 있다. 오픈AI는 앤스로픽을 겨냥해 AI 인프라 확장 속도가 크게 뒤쳐진다고 비판하며, 컴퓨트 용량이 대규모 모델 학습과 배포 범위 확장의 핵심 기반이라는 점을 강조했다. 오픈AI는 자체적으로 '30년 30GW 규모 컴퓨트를 확보할 것으로 전망하는 반면, 앤스로픽은 '29년 7~8GW 수준에 그칠 것으로 예상했다. 현재 앤스로픽이 모델 성능 측면에서 존재감을 키우고 있지만, 컴퓨트 우위를 바탕으로 장기적 경쟁에서 승리할 것이라는 자신감을 보이고 있다.

앤스로픽의 다리오 아모데이 CEO는 과거 경쟁사의 공격적인 인프라 확장 전략을 두고 'YOLO-ing'에 가깝다고 비판하며, 수요 증가 확인에 따른 신중한 캐파 확장 기조를 유지하겠다고 밝혔다. 하지만 최근 들어 전례 없는 클로드 수요 급증을 경험하며 컴퓨트 확보에 적극적 행보를 보이고 있다. 구글-브로드컴과 TPU 활용을 위한 파트너십에 이어 코어위브와는 다년간의 클라우드 계약을 체결했으며, 자체 칩 개발 가능성까지 거론되고 있다.

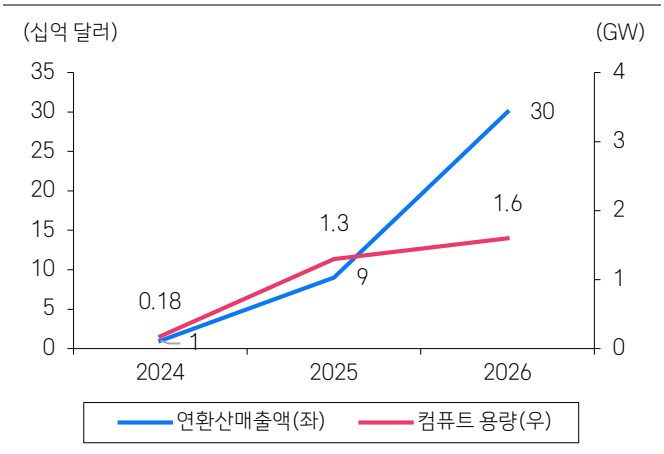
이번 클로드 성능 논란은 급증하는 수요를 감당할 수 있는 컴퓨트 인프라 확보의 중요성을 보여주는 하나의 시례다. 급증하는 AI 수요에 발 맞춰 빠르게 컴퓨트를 제공할 수 있는 업체의 해자는 더욱 부각될 것으로 전망한다. 특히 Claude Mythos와 같이 더욱 큰 연산 자원을 요구하는 차세대 모델로 넘어갈수록, 필요한 컴퓨팅 규모 역시 기하급수적으로 증가할 가능성이 높다. 이를 감안하면 대규모 컴퓨트 자산을 보유한 하이퍼스케일러 및 네오 클라우드 등 인프라 기업의 중장기적 수혜가 기대된다.

환각 벤치마크에서 클로드 순위 변화: 2위에서10위로 하향



자료: Bridgebench.ai, 삼성증권

앤스로픽 연도별 ARR, 컴퓨트 용량 추이



참고: 2026년은 현재까지 공표된 숫자 기반
자료: EpochAI, 언론종합, 삼성증권

Compliance notice

- 본 조사분석자료의 애널리스트는 2026년 04월 14일 현재 위 조사분석자료에 언급된 종목의 지분을 보유하고 있지 않습니다.
- 당사는 2026년 04월 14일 현재 위 조사분석자료에 언급된 종목의 지분을 1% 이상 보유하고 있지 않습니다.
- 본 조사분석자료에는 외부의 부당한 압력이나 간섭 없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.
- 본 조사분석자료는 당사의 저작물로서 모든 저작권은 당사에게 있습니다.
- 본 조사분석자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다.
- 본 조사분석자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다.
- 본 조사분석자료는 기관투자가 등 제3자에게 사전 제공된 사실이 없습니다.

삼성증권

삼성증권주식회사
서울특별시 서초구 서초대로74길 11(삼성전자빌딩)
Tel: 02 2020 8000 / www.samsungpop.com

삼성증권 Family Center: 1588 2323

고객 불편사항 접수: 080 911 0900

NCSI KS-SQI KCSI 한국생산성본부 KCPI cc 소비자중심 Best Korea Brands 2024 1

Member of
Dow Jones
Sustainability Indices
Powered by the S&P Global CSA