AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

무료 · 언제든 수신거부
metallab.ai

제 60 호 2026년 11월 6일 금요일 오늘의 AI

AI

# 오픈AI, 보안 연구 모델 GPT-6 Cyber 공개

오픈AI가 취약점 탐색과 익스플로잇 검증 성능을 끌어올리고 범위 준수 훈련을 더한 GPT-6 Cyber를 데이브레이크 레드 고객 일부에게 공개했다. 익스플로잇짐 점수는 48.5%로 전작보다 상대적으로 51% 높고, 범위 밖 서비스 회피율은 99.93%였다.




AI · 오픈AI, 보안 연구 모델 GPT-6 Cyber 공개

오픈AI가 11월 5일(현지시간) 보안 연구 전용 모델 GPT-6 Cyber를 공개했다. 취약점을 찾고 실제로 공격이 되는지 검증하는 능력을 끌어올리면서, 연구자가 정한 범위 밖으로 나가지 않도록 따로 훈련한 모델이다. 오픈AI는 보안 접근 프로그램 데이브레이크(Daybreak)의 최상위 단계인 데이브레이크 레드 고객 일부에게 API와 코덱스로 먼저 제공하고, 곧 레드 고객 전체로 넓힌다고 밝혔다. 회사는 이 모델을 자체 대비 프레임워크(Preparedness Framework)의 사이버보안 위험 등급에서 최고 단계인 크리티컬(Critical)로 분류했다. 전작 GPT-5.6 Cyber는 하이(High) 등급이었다.

성능 수치는 전작과의 격차로 제시됐다. 지정된 취약점을 원격 대상에서 실제로 공격해 내는지 재는 익스플로잇짐(ExploitGym)에서 GPT-6 Cyber는 48.5%를 기록해 GPT-5.6 Cyber의 32.1%보다 상대적으로 51% 높았다. 범용 모델 GPT-6 Astra는 데이브레이크 블루 환경에서 42.4%였고, 일반 안전장치를 켠 상태에서는 0%였다. 소스 코드 없이 웹 애플리케이션을 탐색하는 침투 테스트 평가에서 취약점 재현율은 40.3%로 GPT-6 Astra의 32.0%를 앞섰다. 깃랩·포스트그레SQL·젠킨스의 알려진 취약점을 통제된 환경에 띄운 새 벤치마크 리모트익스플로잇벤치에서는 데이브레이크 블루의 GPT-6 Astra가 과제 100%를 거부한 반면 GPT-6 Cyber는 95%를 완수했다.

메탈은 오픈AI가 이름을 밝히지 않은 차세대 사이버 모델로 리눅스 커널 취약점 948건의 패치를 만든 연구 보고서를 보도한 바 있다. 그 모델이 GPT-6 Cyber였다. 오픈AI에 따르면 이 모델은 같은 연구 도구에서 취약점 1,724건을 재현하고 948건의 작동하는 익스플로잇을 만들었는데, 이는 GPT-5.6 Sol이 공격에 성공한 166건의 다섯 배가




앤스로픽, 클로드 개인 금융 기능 베타 공개

브라우저 파이어폭스에서는 모질라가 고위험으로 분류한 신규 취약점 81건을 찾아 재현 가능한 테스트와 함께 제출했고, 이미 알려진 결함 30건이 보고된 것보다 심각하다는 사실도 보였다. 모질라는 보고를 받은 직후부터 수정 패치를 반영하기 시작했다. 파이어폭스의 브라이언 그린스테드 디스팅귀시드 엔지니어는 "파이어폭스는 세계에서 가장 단단하게 다듬어지고 면밀히 검증된 코드베이스로 2026년을 시작했다"며 "여러 세대의 프런티어 모델로 더 단단하게 만든 뒤에도 GPT-6 Cyber는 알려지지 않은 고위험 취약점을 찾아냈다"고 밝혔다. 오픈AI는 압축 라이브러리 브로틀리의 결함이 하위 애플리케이션에서 원격 코드 실행으로 이어질 수 있음을 이 모델로 검증했다고도 전했다.

이번 발표의 다른 축은 정렬(alignment)이다. 보안 에이전트가 맡은 범위를 벗어나 무관한 서비스까지 공격하는 위험을 재기 위해 오픈AI는 허깅페이스 사고를 본뜬 평가 익스플로잇짐 허니팟을 만들었다. 공격하기 쉬운 보조 서비스를 일부러 옆에 두고, 모델이 그것이 과제 밖이라는 점을 스스로 판단하는지 보는 방식이다. GPT-6 Cyber는 배포용 안전 프롬프트 없이 99.93%의 경우 범위 밖 서비스를 건드리지 않았고, 프롬프트를 넣으면 100%였다. GPT-5.6 Cyber는 94.49%였다. 공개된 사고 과정 발췌에는 "범위 목록을 요구하고, 데이터 무보존 계약 고객은 암호화된 기록을 30일간 고객 저장소에 두고 자동 검사하는 비공개 안전 처리(PSP)를 써야 한다.

발전소와 정수장 같은 핵심 기반시설은 따로 다뤘다. 오픈AI가 공개한 도표에 따르면 민감한 레드팀 요청에 대한 안전장치 표시율은 일반 환경에서 70.8%에서 89.7%로, 데이브레이크 블루에서 13.8%에서 67.0%로, 데이브레이크 레드에서 0%에서 61.9%로 올랐다. 승인된 기반시설 방어 조직을 위해 새로 만든 데이브레이크 레드 포 크리티컬 인프라스트럭처 그룹의 표시율은 0%에서 0.3%였다. 오픈AI는 같은 기법이 정당한 보안 작업에도 필요하다며 표시율 100%를 목표로 삼지 않는다고 밝혔다. 회사는 자원이 부족한 정부 기관과 필수 서비스 제공자의 접근 비용을 지원하는 10억 달러 규모의 보조금 프로그램도 운영한다.

평가에는 산업 제어 시스템 보안 기업 드라고스가 참여했다. 메탈이 확인한 드라고스의 14쪽 백서는 GPT-5.5 Cyber와 GPT-5.6 Sol을 대상으로 7개 평가를 진행한 결과를 담았다. 모의 열병합 발전 시설에서 모델은 사전 정보 없이 망을 파악하고 버려진 기본 SSH 계정을 찾았으며, 보호 계전기 데이터를 다루는 엔지니어링 소프트웨어에서 제로데이 취약점 2건을 발견했다. 모의 수도관로 시설에서는 PLC가 응답을 멈췄고, HMI의 FTP 서비스도 한 차례 중단됐다. 산업용 라우터에서는 드라고스 연구원이 이틀 걸려 찾은 취약점을 약 10분 만에 다시 찾아냈다. 롭 T. 리 드라고스 최고AI책임자는 "공격 전에 필요한 정찰과 취약점 연구의 기술 문턱이 낮아지고 있다"고 밝혔다.

보도에 따르면 GPT-6 Cyber는 오픈AI가 올해 4월·6월·8월에 이어 내놓은 네 번째 사이버보안 특화 모델이다.

오늘 지면

1 AI 오픈AI, 보안 연구 모델 GPT-6 Cyber 공개
2 AI 오픈AI, 리눅스 커널 취약점 948건 패치 생성
3 AI 오픈AI, 챗GPT 이미지 2.6 공개
4 AI 앤스로픽, 원핵생물 볼트 입자 구조 공개
5 AI 앤스로픽, 클로드 개인 금융 기능 베타 공개

이 지면은 2026년 11월 6일 금요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



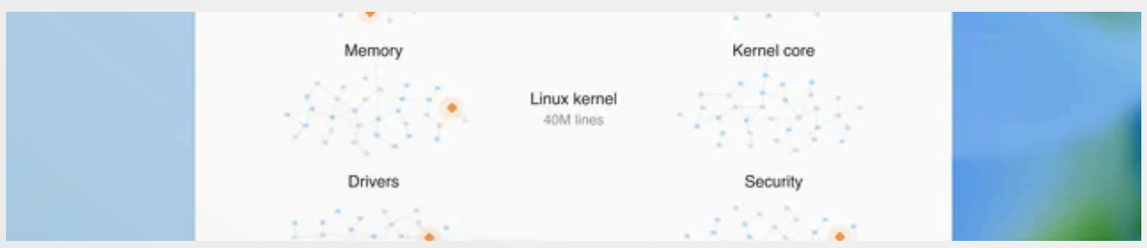
## 앤스로픽, 원핵생물 볼트 입자 구조 공개

앤스로픽이 세균과 박테리오파지에서 온 볼트(vault) 입자의 3차원 구조를 처음으로 풀어냈다. 회사의 생물학 연구팀은 11월 4일 연구 글과 기술 보고서를 공개하고, 파지에서 온 단백질 L05가 사슬 96개로 이뤄진 껍질을 만든다는 사실을 극저온전자현미경(cryo-EM)으로 확인했다고 밝혔다. 구조의 공칭 분해능은 3.2옹스트롬이다. 세균과 고세균처럼 핵이 없는 단순한 세포, 곧 원핵생물에서 볼트 입자의 구조가 밝혀진 것은 처음이다. 볼트는

▶ 2면에 계속

AI

## 오픈AI, 리눅스 커널 취약점 948건 패치 생성

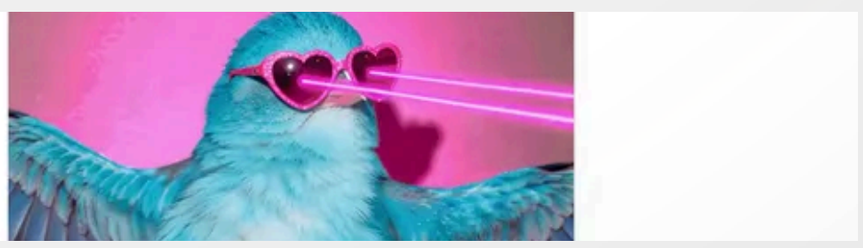



오픈AI가 11월 4일(현지 시각) 리눅스 커널의 보안 취약점을 찾고, 재현하고, 공격 가능성을 확인한 뒤 패치까지 만드는 자율 파이프라인의 실험 결과를 담은 보고서 「A thousand fewer paths to root」를 공개했다. 아직 이름을 밝히지 않은 차세대 사이버 모델로 한 번 돌린 결과, 일반 사용자 계정으로 관리자(root) 권한을 빼앗을 수 있는 로컬 권한 상승(LPE) 취약점 948건을 검증하고 각각의 패치를 생성했다. 전체 API 비용은 121만 3,430달러, 권한 상승 익스플로잇 1건당 1,279달러였다. 오픈AI는 X 공식 계정에 "보안 연구원들이 자율 사이버 취약점 탐색·공격·패치 파이프라인을 만들었다"고 밝혔고, 이 글은 반나절 만에 조회수 132만 회를 넘었다.

▶ 2면에 계속

AI

## 오픈AI, 챗GPT 이미지 2.6 공개



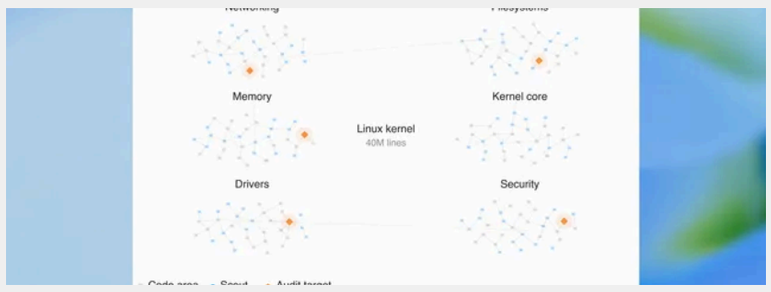
오픈AI가 11월 4일(현지 시각) 이미지 생성 기능의 새 판인 챗GPT 이미지 2.6(Images 2.6)을 공개했다. 챗GPT와 코덱스, API에 함께 들어갔고, 이미지를 이전보다 50% 빠르게 만들며 짧게 반복되는 움직이는 그림인 GIF를 프롬프트 한 줄로 만들 수 있게 됐다. 오픈AI는 X 공식 계정에 "이미지가 업그레이드됐다"며 GIF와 참조 이미지를 새 기능으로 내세웠다. 개발자용 API에는 GPT-Image-2.6 Flare와 GPT-Image-2.6 Sunburst 두 모델이 나왔다.

생성 규모부터 커졌다. 오픈AI 발표에 따르면 챗GPT와 코덱스, API를 합쳐 전 세계 사용자가 일주일에 만드는 이미지는 40억 장이다. 하루 평균으로는 5억 장이 넘는다. 보도에 따르면 오픈AI가 9월에 밝힌 주간 생성량은 30억 장이었다.

◀ 1면에서 계속

AI

## 오픈AI, 리눅스 커널 취약점 948건 패치 생성




1 오픈AI가 11월 4일 리눅스 커널 취약점을 찾아 재현하고 패치까지 만드는 자율 파이프라인 실험 보고서를 공개했다.

2 후보 결함 2,308건 가운데 1,724건을 재현했고, root 권한을 얻는 권한 상승 취약점 948건을 검증해 패치를 생성했다.

3 7월 GPT-5.6 Sol 스캔보다 새 권한 상승 취약점을 다섯 배 넘게 찾았고, 처음 보는 공격 기법 10가지도 나왔다.

파이프라인은 다섯 단계로 돈다. 먼저 스카우트 에이전트가 커널의 폴더와 파일을 훑어 일반 사용자가 닿을 수 있고 보안상 의미가 있는 감사 대상 5,034곳을 골랐다. 약 280만 줄, 커널 전체의 7% 정도이고 이 단계 비용은 3,426달러였다. 이어 분석 에이전트가 12시간 동안 후보 결함 2,308건을 냈다. 검증 단계에서는 메모리 오류를 잡아내는 KASAN을 켠 격리 가상머신에서 트리아지 에이전트가 PoC를 직접 만들어 돌렸고, 30시간 만에 후보의 74.7%인 1,724건을 재현했다. 재현에 실패한 후보 가운데 모델이 불가능한 조건을 가정한 잘못된 분석은 3.1%에 그쳤다. 나머지는 상위 권한이 필요하거나 경쟁 상태를 재현하기 어려운 경우였다.

공격 가능성 평가는 실제 배포판 수준의 방어 장치를 켠 채 진행됐다. 선별 에이전트가 1,724건을 정보 유출 후보 313건과 권한 상승 후보 1,238건으로 좁혔고, 익스플로잇 개발 에이전트가 SMEP·SMAP·KASLR 같은 커널 보호 기능이 모두 켜진 가상머신에서 작동하는 권한 상승 익스플로잇 1,169개를 만들었다. 변형 분석으로 같은 뿌리의 결함을 묶어 서로 다른 패치 948개가 나왔고, 오픈AI는 이 숫자를 고유 취약점 수로 삼았다. 메탈이 확인한 보고서의 단계별 비용을 더하면 탐색 3,426달러, 분석 31만 6,549달러, 검증 7만 3,330달러, 선별 4만 4,194달러, 익스플로잇 개발 77만 5,929달러로, 익스플로잇 개발이 전체 비용의 약 64%를 차지했다. 성공한 시도는 평균 360달러였지만 실패한 69건은 평균 3,922달러가 들었다.

모델 세대가 바뀌자 결과가 크게 달라졌다. 7월 말 GPT-5.6 Sol로 돌린 첫 스캔은 51만 3,690달러를 들여 고유 권한 상승 취약점 166건을 찾았다. 석 달 동안 커널이 보강됐는데도 10월 스캔은 새 권한 상승 취약점을 다섯 배 넘게 찾아냈다. 에이전트가 익스플로잇 166개에서 쓴 공격 기법 가운데 10가지는 외부 커널 공격 전문가들도 처음 보는 것이었고, 오픈AI는 안전을 이유로 세부를 공개하지 않았다. 회사는 "프런티어 모델은 알려진 취약 코드 패턴을 맞춰 보는 데 그치지 않고 창의성과 새로운 접근을 보여 준다"고 썼다. 오케스트레이션 실험도 같은 흐름을 보였다. GPT-5.5 Cyber에 공격 기법 103개를 정리해 주자 후보 63건에서 성공률이 76.2%에서 88.9%로 올랐지만, GPT-5.6 Cyber는 기법 목록 없이 92.1%를 냈고 목록을 더해도 1.5%포인트 오르는 데 그쳤다. 오픈AI는 무거운 보조 장치를 걷어 냈다.

◀ 1면에서 계속

AI

## 오픈AI, 챗GPT 이미지 2.6 공개



1 오픈AI가 11월 4일 챗GPT 이미지 2.6을 공개하고 챗GPT와 코덱스, API에 함께 넣었다.

2 이미지 생성은 이전보다 50% 빨라졌고, 프롬프트나 사진으로 GIF를 만들어 속도와 프레임을 고칠 수 있다.

3 API의 Flare는 low 품질에서 최소 6초에 이미지를 내고, Sunburst는 편집을 거듭해도 세부를 더 일관되게 지킨다.

속도는 두 모델로 나뉘어 올라갔다. Flare는 품질을 low로 두면 최소 6초 만에 이미지를 낸다. 회사는 Flare가 포스터와 제품 이미지 같은 업무용 시각물의 레이아웃 품질도 끌어올렸다고 밝혔다. Sunburst는 세밀함과 정밀도가 필요한 작업용이다. 품질을 medium이나 high로 두면 편집을 거듭해도 이미지 일관성이 더 높게 유지된다. 오픈AI 개발자 계정은 "Sunburst는 한 번의 편집에서 다음 편집으로 넘어갈 때 세부를 더 일관되게 지킨다"고 설명했다. 공식 비교 영상은 GPT-Image-2.5 Flare와 GPT-Image-2.6 Flare가 같은 고급 여행 광고 이미지를 만드는 시간을 8배속으로 나란히 잰다. 메탈은 두 모델 이름이 처음 나온 챗GPT 이미지 2.5 공개 소식을 보도한 바 있으며, 이번 판은 같은 이름의 두 모델을 2.6 세대로 올렸다.

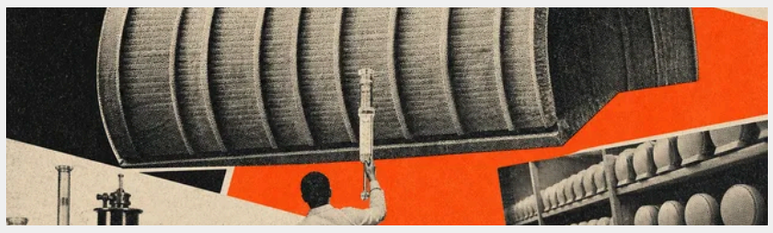
이번 판에서 가장 눈에 띄는 변화는 GIF다. 보도에 따르면 챗GPT는 이전까지 GIF를 한 번에 직접 만들지 못했고, 이제 프롬프트나 업로드한 사진으로 GIF를 만든 뒤 애니메이션 속도와 개별 프레임을 고쳐 공유할 수 있다. 메탈이 확인한 오픈AI 공식 소개 영상에서 GIF 재생 속도 메뉴는 0.5배부터 2배까지 0.25배 간격의 7단계로 나뉘어 있다. 같은 영상에서 선글라스를 쓴 파란 새 GIF는 "분홍색 레이저를 더해 줄래"라는 후속 요청 한 번에 눈에서 레이저를 쏘는 장면으로 바뀐다. 오픈AI는 브랜드 마스코트 애니메이션, 소셜미디어용 리액션과 밈, 영상 제작 전에 캐릭터 동작을 미리 보는 움직이는 스토리보드를 쓰임새로 들었다. 손으로 대충 그린 스케치도 움직이는 캐릭터가 될 수 있다며 공책에 그린 부엉이 낙서를 캐릭터로 바꾼 예를 실었다.

초기 사용 기업의 사례도 함께 나왔다. 이미지 생성 플랫폼 오픈아트(OpenArt)는 6개 프레임을 가로 3칸, 세로 2칸 격자에 담은 스프라이트 시트를 만들게 한 뒤 끊김 없이 반복되는 GIF로 잇는 프롬프트를 공개했다. 주가 차트가 떨어지자 눈물을 흘리는 사무직 캐릭터와 차트가 오르자 지폐를 뿌리는 캐릭터다. 코코 마오(Coco Mao) 오픈아트 최고경영자(CEO) 겸 공동창업자는 "GIF는 소셜미디어의 모국어이고, GPT-Image-2.6은 모든 오픈아트 크리에이터가 프롬프트 하나로 그 언어를 유창하게 쓰게 해 준다"고 말했다. AI 업무 비서를 만드는 스타트업 타운(Town)은 API로 만든 몬티(Monty)라는 캐릭터를 메시지 입력창 위 타이핑 표시로 쓰고 있다.

◀ 1면에서 계속

AI

## 앤스로픽, 원핵생물 볼트 입자 구조 공개



1 앤스로픽이 11월 4일 세균과 박테리오파지에서 온 볼트 입자의 첫 3차원 구조를 공개했다.

2 파지 단백질 L05 껍질은 사슬 96개로 사람 볼트의 78개보다 많고, 더 짧고 넓으면서 약 25% 큰 부피를 품는다.

3 벽 반복을 두 개 더 끼워 넣자 부피가 약 20% 늘어 사람 볼트보다 약 50% 큰 나노 용기가 됐다.

연구팀은 클로드 사이언스로 약 150억 개 예측 단백질 데이터베이스를 뒤져 MVP 친척 단백질을 찾았다. 어깨 도메인의 프로필 은닉 마르코프 모델을 만들어 반복 검색한 결과, MVP 상동체는 세균 29개 문과 고세균 4개 문에 퍼져 있었지만 유전체가 해독된 종 가운데 1% 미만에만 나타났다. 널리 퍼져 있으나 드물다는 뜻이다. 이 조사로 알려진 MVP·MVP 유사 단백질 그룹은 41개에서 1,507개로 늘었다. 연구팀은 대표 서열 489개로 계통수를 만들어 세균, 고세균, 파지, 진핵생물의 유전자가 각각 다른 가지에 놓인다는 것을 확인했다.

다음은 이 단백질들이 실제로 껍질을 만드는지 확인하는 일이었다. 연구팀은 벽 반복이 3개에서 14개까지 걸친 단백질 44개를 골라 무세포 발현계에 넣었고, 성적이 좋은 것들을 대장균으로 옮겨 24개 가운데 15개에서 녹는 단백질을 얻었다. 입자가 9메가달톤에 이를 만큼 커서 크기 배제 크로마토그래피의 배제 부피에 먼저 빠져나온다는 점을 이용했다. 음성 염색 전자현미경으로 보니 15개 중 7개가 볼트 같은 입자를 만들었고, cryo-EM 은 그중 4개를 볼트로 확정했다. 파지에서 둘, 세균에서 둘이다. 빈 벡터를 넣은 세포를 같은 절차에 태웠을 때는 통 모양도 고리 조각도 나오지 않았다.

구조를 자세히 본 대상은 파지 단백질 L05다. 반쪽마다 사슬 48개, 전체 96개로 사람 볼트의 78개보다 18개 많다. 짧은 축 지름은 47.5나노미터로 사람 볼트의 40나노미터보다 넓고, 전체 길이는 54.5나노미터로 사람 볼트의 67나노미터보다 짧다. 사슬 하나에 든 벽 반복은 7개여서 9개인 사람 MVP보다 벽이 7.2나노미터 짧고, 캡의 경사도 약 34도로 사람 볼트의 약 52도보다 완만하다. 그런데도 더 둥글어 안에 품는 부피는 약 25% 크다. 두 반쪽이 맞물리는 허리의 접촉 면적은 약 225제곱옹스트롬으로 사람 볼트의 약 420제곱옹스트롬의 절반 수준이고, 그 자리를 이루는 잔기는 두 볼트 사이에 하나도 겹치지 않았다.

AI

## 앤스로픽, 클로드 개인 금융 기능 베타 공개

앤스로픽이 은행·카드 계좌를 읽기 전용으로 연결해 지출과 잔액을 묻는 클로드 개인 금융 기능을 미국에서 베타로 공개했다. 광고와 자체 금융상품이 없고, 허브 안의 대화는 모델 학습에 쓰지 않는다.




1 앤스로픽은 11월 5일 클로드 개인 금융 기능을 미국 Pro·Max·Ultra 요금제 사용자에게 베타로 배포하기 시작했다고 밝혔다.

2 계좌는 플레이드를 거쳐 읽기 전용으로 연결되며, 클로드는 잔액·신용 한도와 최대 2년 치 거래 내역을 보고 답한다.

3 앤스로픽은 광고·스폰서 추천·자체 금융상품이 없고 개인 금융 허브의 대화는 모델 학습에 쓰지 않는다고 밝혔다.

앤스로픽이 11월 5일(현지시간) 클로드에 개인 금융(Personal Finance) 기능을 베타로 공개했다. 사용자가 고른 은행·카드 계좌를 연결하면 클로드가 잔액과 거래 내역을 바탕으로 돈에 관한 질문에 일상 언어로 답한다. 이번 주부터 미국의 Pro·Max·Ultra 요금제 사용자에게 순차 배포된다. 접근 권한은 읽기 전용이며, 앤스로픽은 개인 금융 기능이 사용자의 돈을 옮기거나 관리하거나 투자할 수 없다고 밝혔다.

계좌는 앱과 은행을 이어 주는 서비스 플레이드(Plaid)에서 사용자가 직접 은행에 로그인해야 연결된다. 약관 동의와 읽기 전용 권한 승인도 그 단계에서 이뤄지기 때문에 클로드는 은행 비밀번호를 요구하지도, 보지도 않는다. 연결한 계좌에 한해 클로드가 보는 정보는 은행 이름, 계좌별 이름·종류·끝 네 자리, 잔액과 신용 한도, 최대 2년 치 거래 내역이다. 보도에 따르면 플레이드는 미국 주요 은행을 포함해 1만 2,000곳이 넘는 금융기관과 연결돼 있다.

메탈이 확인한 앤스로픽 발표문의 예시 화면은 쓰임새를 숫자로 보여 준다. 월급날을 열흘 앞둔 사용자가 이번 달 돈이 어디로 갔는지 묻자, 클로드는 연결된 계좌의 10월 지출을 4,309.37달러로 합산하고 범주별로 나눴다. 집세·공과금이 2,339.51달러(4건)로 절반을 넘었고, 식음료 1,024.51달러(36건), 쇼핑 322.85달러(5건)가 뒤를 이었다. 평소보다 많은지 되묻자 클로드는 앞선 달들과 비교해 멤버십 갱신 결제를 짚었다.