GLOBAL
EQUITY

2026. 2. 6

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

리서치센터 리포트
바로가기

글로벌 AI

앤스로픽과 오픈AI의 멀티 에이전트 플랫폼 경쟁 본격화

- 앤스로픽(Claude Opus 4.6)과 오픈AI(GPT-5.3-Codex) 발표는 장시간 작업이 가능한 멀티 에이전트를 실제 업무와 시스템에서 구동하는 플랫폼 경쟁의 본격화
- 에이전트 병렬 운영, 초기 단계인 재귀개선, 작업 영역 확장 등에 주목 필요

WHAT'S THE STORY?

앤스로픽과 오픈AI는 경쟁적으로 모델 업데이트를 발표. 앤스로픽의 Claude Opus 4.6과 오픈AI의 GPT-5.3-Codex가 주인공. 단순 신규 모델 출시를 넘어 장시간 작업이 가능한 멀티 에이전트를 실제 업무와 시스템 안에서 구동하는 제품과 플랫폼 경쟁의 본격화. 양 사의 발표에서 주목할 포인트는 크게 5가지

1) 에이전트의 병렬 운영 능력: 에이전트 병렬 운영은 단일 작업 당 생산성 향상 뿐 아니라 전체 프로젝트 리드타임 단축으로 연결. Claude Code에 여러 에이전트를 하나의 팀(Agent Teams)로 구성해 병렬 및 자율 공동 작업을 수행하는 기능 제시. 오픈AI의 Codex 앱도 여러 에이전트의 동시 실행 및 관리 강조. 공유 컨텍스트를 유지한 채 장시간 실행되며 감독 비용 최소화

2) 에이전트 재귀개선의 초기 단계: GPT-5.3-Codex는 자기 자신을 만드는데 핵심적으로 사용된 최초의 모델. Codex 팀은 초기 버전을 활용해 학습 런 모니터링, 디버깅, 배포 운영, 테스트 평가 진단까지 자동화. 앤스로픽도 Claude를 통해 Claude를 구축한다는 코멘트. Claude Code로 코드를 작성하고 신규 모델은 실제 내부 업무에 전 적용 후 외부 공개. 완전한 재귀개선은 아니지만 모델 개발 생산성 자체의 가속화

3) 작업 영역의 확장: Opus 4.6은 여러 전문 도메인의 실제 업무 전반에서 적용. 금융, 법률 등 분야의 지식 노동 과제를 평가하는 GDPval-AA에서 높은 성과 달성. GPT-5.3-Codex는 코딩 능력 개선에 초점을 맞췄지만, 단순 코드 작성 및 리뷰를 넘어 컴퓨터에서 전문가가 할 수 있는 거의 모든 일을 끝까지 수행할 수 있는 범용 에이전트로의 진화 시사

4) 에이전트의 병목은 운용성: 모델의 지능 수준과 에이전트 발전에도 실제 적용에는 의외의 병목이 존재. 오픈AI는 업무용 에이전트의 필수 스택으로 공유 컨텍스트, 온보딩, 피드백 기반 학습, 권한 및 경계의 필요성을 강조. 엔터프라이즈 고객의 선택 기준도 단순 정확도나 벤치마크 수준을 넘어 운영 가능한 에이전트 스택으로 이동

5) 장문 컨텍스트는 기본: Opus 4.6은 Opus급 모델 최초로 백만 토큰 컨텍스트 윈도우를 베타 지원. 컨텍스트 부식 개선도 전면에 내세움. 장시간 작업을 가능하게 하는 필수 인프라로 컨텍스트 확장은 계속될 것. 이 과정에서 메모리가 핵심 역할 수행

(다음 페이지에 계속)

앤스로픽 Claude Opus 4.6 vs 오픈AI GPT-5.3-Codex

앤스로픽과 오픈AI가 몇 분 차이를 두지 않고 신규 모델 업데이트를 발표하며 경쟁에 다시 한 번 불을 붙였다. 사실 순서 상으로 먼저 발표한 앤스로픽을 오픈AI가 저격한 셈이다. 앞서 오픈AI가 발표한 엔터프라이즈의 에이전트 확장 지원 플랫폼인 Frontier와 연결지어 생각해보면 단순 신규 모델 출시를 넘어 장시간의 멀티 에이전트를 기업의 실제 업무와 시스템 안에서 구동하는 제품 및 플랫폼 경쟁이 본격화되고 있다. 양 사의 발표 내용을 종합해 주목할 포인트를 꼽아보았다.

에이전트의 병렬 운영 능력을 보여줘: 모델이 개선될수록 에이전트 간 병렬 운영 능력이 중요해진다. 단일 작업당 생산성 향상뿐 아니라 전체 프로젝트 리드타임 단축도 가능하다. 앤스로픽은 Claude Code에서 여러 에이전트를 하나의 팀(Agent Teams)으로 구성해 병렬 및 자율 공동 작업을 수행하는 기능을 선보였다. 모델 자체가 장시간 안정적으로 일할 수 있게 바뀌었기 때문에 가능한 구조다. Agent Teams를 활용해 시스템 레벨의 C 컴파일러 구축에 성공한 사례도 공유되었다.

오픈AI Codex 앱도 여러 에이전트의 동시 실행 및 관리가 가능하다. 각 에이전트는 서로 다른 작업을 병렬로 수행한다. 사용자는 작업 중에도 개별 에이전트에 직접 개입해 방향을 조정할 수 있기 때문에 완전 자동화가 아닌 인간-에이전트 협업형 병렬 워크플로우를 구현한다. 에이전트들은 공유 컨텍스트를 유지한 채 장시간 실행되며 중간 의사결정 및 진행 상황을 수시로 보고해 감독 비용을 최소화한다.

OpenRouter LLM API 점유율에서 3위를 기록하며 돌풍을 일으키고 있는 Kimi K2.5가 자기지시형 Agent Swarm(에이전트 스웜) 패러다임을 구현해 최대 100개의 서브 에이전트를 자동으로 생성 및 조율하고 최대 1,500개의 툴 호출을 병렬 실행하는 오케스트레이션 구조를 추구하는 것 과도 궤를 같이한다.

에이전트의 재귀개선: GPT-5.3-Codex는 자기 자신을 만드는데 핵심적으로 사용된 최초의 모델이다. Codex팀은 초기 버전을 활용해 학습 런 모니터링, 디버깅, 배포 운영, 테스트 결과 평가 진단까지 자동화했다. 출시 과정에서 GPU 클러스터 동적 확장을 통한 트래픽 급증 시의 지연 시간의 안정화나 알파 테스트 과정에서 대규모 세션 로그 전반을 분석해 생산성 지표를 추정하고 보고서를 자동 생성하기도 했다. 앤스로픽도 Claude로 Claude를 만든다. 엔지니어들은 Claude Code로 코드를 작성하고 신규 모델은 실제 내부 업무에 선 적용 후 외부 공개된다. 모델 개발 생산성 자체가 지속 가속화되고 있다.

작업 영역 확장은 어디까지: GPT-5.3-Codex가 적용된 Codex는 단순 코드 작성 및 리뷰를 넘어 컴퓨터에서 전문가가 할 수 있는 거의 모든 일을 끝까지 수행할 수 있는 범용 에이전트로 진화했다. Opus 4.6는 여러 전문 도메인의 실제 업무 전반(재무 분석, 리서치, 파일 생성 및 활용)에서 적용된다. 이전에 공개된 앤스로픽 Claude Cowork는 어떻게 보면 에이전트 작업 영역 확장의 시작점에 불과하다.

에이전트의 병목은 지능 수준이 아니라 운용성: AI 모델 지능 수준은 지속 높아지고 있다. 하지만 에이전트 실제 적용에는 의외의 병목이 존재한다. 오픈AI는 업무용 에이전트의 필수 스택으로 공유 컨텍스트, 온 보딩, 피드백 기반 학습, 권한 및 경계의 필요성을 강조했다. 앤스로픽은 Opus 4.6 모델을 더 오래, 더 안정적으로, 큰 코드베이스에서 돌아가는 에이전트형 모델로 정의했다. 엔터프라이즈 고객의 선택 기준이 단순 정확도와 벤치마크를 넘어 운영 가능한 에이전트 스택으로 옮겨가는 것을 반영하는 움직임이다.

이제 장문 컨텍스트는 기본: Claude Opus 4.6은 Opus급 모델 최초로 백만 토큰 컨텍스트 윈도우를 베타로 지원하고, 컨텍스트 부식 개선을 전면에 내세웠다. GPT-5.3-Codex의 핵심적 특징도 장시간 실행되는 연구, 도구 사용, 복합 실행 작업 처리 능력이다. 장시간 작업(long running tasks)을 가능하게 하는 필수 인프라로써 컨텍스트 확장은 계속될 것이며, 이 과정에서 메모리는 핵심적 역할을 할 것이다.

앤스로픽, Claude Opus 4.6 – 가장 뛰어난 모델을 한 단계 업그레이드

앤스로픽의 Claude Opus 4.6은 이전 버전 모델 대비 코딩, 에이전트형 작업, 장문 컨텍스트 처리 능력을 대폭 강화했다. 또한 Opus 계열 최초로 백만(1M) 토큰 컨텍스트 윈도우를 베타 지원하고, 대화가 길어질 수록 성능이 저하되는 현상인 컨텍스트 부식(context rot)도 유의미하게 개선했다. API 가격은 인풋 $5 및 아웃풋 $25로 이전 모델과 동일한 수준이다.

모델 업데이트를 통해 개선된 능력이 여러 전문 도메인의 실제 업무 전반에 적용될 수 있다는 점을 강조했다. 예시로는 재무 분석, 리서치, 문서, 스프레드시트, 프레젠테이션의 생성 및 활용을 들었다. Excel에 적용된 Claude 기능 강화 뿐 아니라 PowerPoint용 Claude를 프리뷰 형태로 공개하기도 했다. 이를 통해 지식 노동 중심의 엔터프라이즈 활용도를 높이고, 일상적 업무 수행에서 강력함을 높이고 있다.

Opus 4.6은 금융, 법률 등 분야의 지식 노동 과제를 평가하는 GDPval-AA에서 경쟁 및 이전 모델 대비 높은 성과를 달성했다. 온라인에서 찾기 어려운 정보를 탐색하는 능력을 측정하는 BrowseComp에서도 우수한 성능을 보인 것도 당연한 결과다.

Opus 4.6이 적용된 Claude Code에서는 여러 에이전트를 하나의 팀(Agent Teams)으로 구성해 병렬 및 자율 공동 작업 수행이 가능하다. API 환경에서는 컨텍스트 압축(compaction) 기능을 통해 모델이 이전 대화를 스스로 요약하며 장시간 에이전트 작업을 지속한다. 모델이 상황에 따라 사고 깊이를 자동 조절하는 적응형 사고(adaptive thinking)와 지능, 속도, 비용을 개발자가 직접 제어할 수 있는 effort 제어 옵션 도입도 주요한 특징이다.

앤스로픽이 공개한 벤치마크 표

벤치마크	세부내용	Opus 4.6	Opus 4.5	Sonnet 4.5	Gemini 3 Pro	GPT-5.2
Terminal-bench 2.0	에이전틱 터미널 코딩	65.4%	59.8%	51.0%	56.2%	64.7%
SWE-bench Verified	에이전틱 코딩	80.8%	80.9%	77.2%	76.2%	80.0%
OSWorld	에이전틱 컴퓨트 유즈	72.7%	66.3%	61.4%	-	-
t2-bench	에이전틱 툴 사용	91.9%	88.9%	86.2%	85.3%	82.0%
MCP Atlas	확장된 도구 활용 능력	59.5%	62.3%	43.8%	54.1%	60.6%
BrowseComp	에이전틱 검색	84.0%	67.8%	43.9%	59.2%	77.9%
Humanity's Last Exam	학술적 추론	40.0%	30.8%	17.7%	37.5%	36.6%
Finance Agent	에이전틱 재무 분석	60.7%	55.9%	54.2%	44.1%	56.6%
GDPVal-AA Elo	오피스 업무	1,606	1,416	1,277	1,195	1,462
ARC-AGI 2	새로운 문제 해결 능력	68.8%	37.6%	13.6%	45.1%	54.2%
GPQA Diamond	과학적 추론 능력	91.3%	87.0%	83.4%	91.9%	93.2%
MMMU Pro	시각 추론 능력	73.9%	70.6%	63.4%	81.0%	79.5%
MMMLU	다국어 질의응답 능력	91.1%	90.8%	89.5%	91.8%	89.6%

자료: Anthropic, 삼성증권

오픈AI, GPT-5.3-Codex: 컴퓨터에서 수행되는 전문 업무 영역 전반으로

오픈AI는 가장 강력한 에이전트형 코딩 모델을 표방하는 GPT-5.3-Codex를 발표했다. 기존 GPT-5.2-Codex의 코딩 성능과 GPT-5.2의 추론 능력 및 전문 지식을 하나의 모델로 결합했다.

소프트웨어 엔지니어링 관련 벤치마크인 SWE-Bench Pro, Terminal Bench 2.0에서 SOTA를 달성했다. 특히 Terminal Bench는 77.3%로 Claude Opus 4.6이 65.4%로 왕관을 쓰자 마자 빼앗아버린 상황이다. 시각 정보를 사용해 실제 데스크톱 환경에서 생산성 작업 수행을 평가하는 에이전트 벤치마크 OSWorld와 지식 노동 과제 처리를 측정하는 GDPVal에서도 대폭 개선된 성능을 보였다. 속도도 25% 향상되었다.

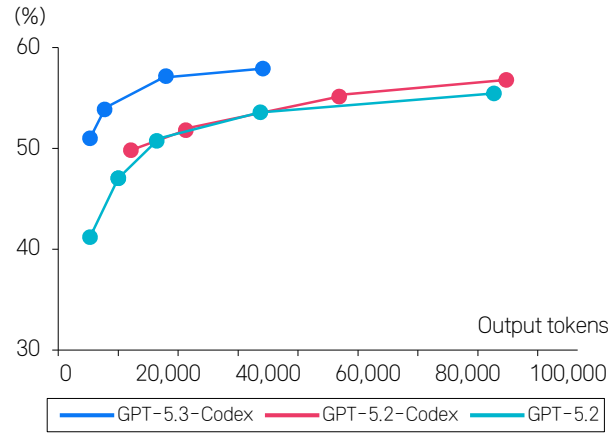
신규 모델이 적용된 Codex는 단순 코드 작성 및 리뷰를 넘어 컴퓨터를 직접 사용해 업무를 끝까지 수행하는 범용 에이전트로 진화한다. 장시간 실행되는 연구, 도구 사용, 복합 실행 작업을 처리할 수 있다. 코딩 뿐 아니라 컴퓨터 사용, 실무 지식 노동을 단일 에이전트로 통합하는 성과다.

코딩 역량, 미적 감각 및 압축 개선이 결합되어 웹, 앱, 게임을 수일 간 자율적으로 반복개선 할 수 있는 장기 에이전트 능력도 강화되었다. 단순하거나 불완전한 프롬프트를 제공해도 합리적 기본값을 갖춘 결과를 제공한다.

사이버 보안 과제 성능의 의미 있는 향상도 확인되었다. Preparedness 프레임워크 기준 사이버 보안 영역에서는 최초로 High Capability 모델로 분류되었으며, 소프트웨어 취약점 식별을 직접 학습했다. 예방적 접근으로 포괄적 사이버 보안 안전 스택을 제공한다. 사이버 보안 분야에도 네이티브 AI 기업의 침투가 거세지고 있다.

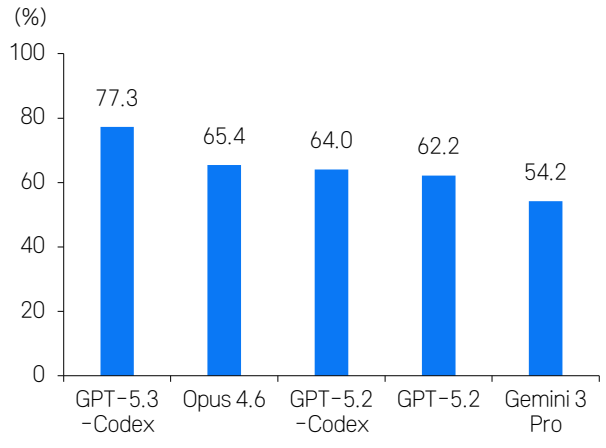
엔비디아 GB200 NVL72 기반으로 학습 및 서빙이 진행되는 것도 중요하다. 여기에 인프라 및 추론 스택 개선을 통해 실사용 체감 속도까지 개선했다. 블랙웰 기반 모델 학습이 본격화되면 전반적 AI 모델 능력치의 레벨업을 기대할 수 있다.

성능, 토큰 효율성을 모두 개선한 GPT-5.3-Codex 모델



참고: 벤치마크는 SWE-Bench Pro 기준
자료: OpenAI, 삼성증권

Terminal-Bench 2.0 기준 프론티어 모델 성능 비교



자료: OpenAI, Anthropic, 삼성증권

오픈AI, Frontier – 엔터프라이즈 용 에이전트 플랫폼

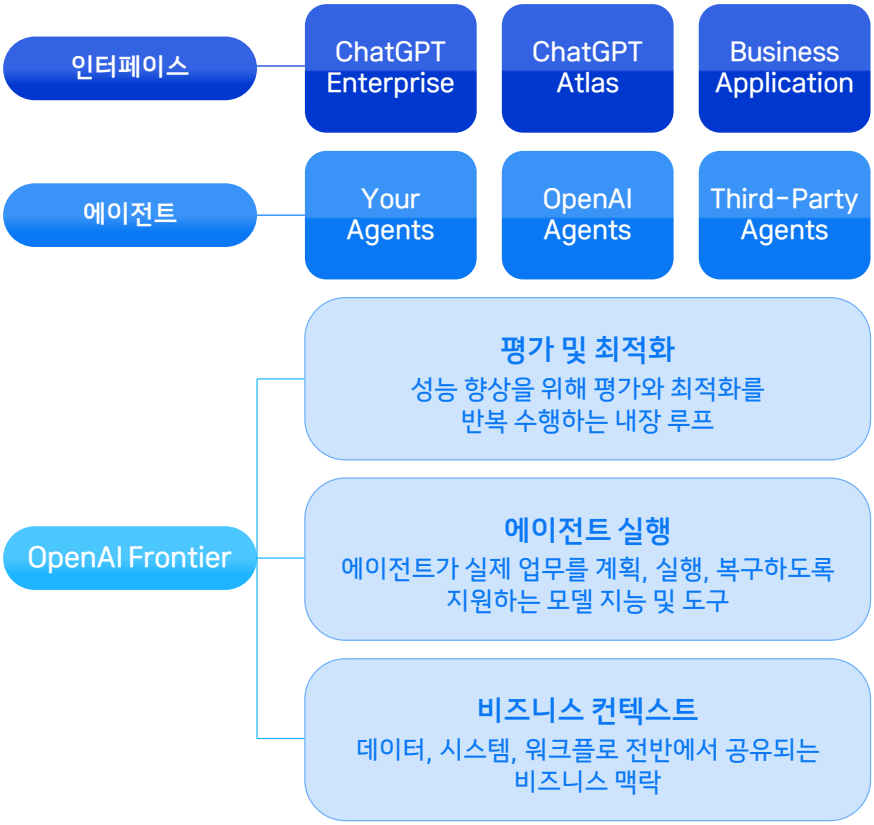
Frontier는 기업의 AI 에이전트를 구축, 배포 운영, 관리를 돕는 플랫폼이다. 공유된 맥락, 온 보딩, 피드백을 통한 실무 학습, 명확한 권한과 경계를 제공해 비즈니스 전반에서 함께 일하는 AI 동료로의 확장을 목표로 하고 있다.

오픈AI는 모델 지능은 끊임없이 좋아지고 있고, 에이전트가 본격적으로 배치되고 있지만, 조직 내 파편화된 시스템과 거버넌스는 '모델이 할 수 있는 일'과 '조직이 실제로 배포할 수 있는 것' 사이의 기회의 격차(opportunity gap)를 오히려 확대하고 있다고 진단했다.

에이전트가 실제 업무를 수행하려면 공유 컨텍스트(시맨틱 레이어) + 실행 환경(툴과 컴퓨터) + 평가 및 피드백 기반 품질 개선 + 권한 및 경계 부여가 필요하다. 이를 기존 시스템과 클라우드 위에서 플랫폼 전환 없이 오픈 표준으로 통합하는 것을 핵심으로 바라보고 있다.

또한 FDE(현장 배치 엔지니어)를 활용해 고객과 함께 프로덕션의 최적 예시(모범 사례)를 구축하고, 배포 과정의 피드백을 연구로 연결해 모델과 시스템 개선을 가속화하는 현장-연구 간 피드백 루프를 구현할 계획이다.

오픈AI Frontier 구조도



자료: OpenAI, 삼성증권

Compliance notice

- 본 조사분석자료의 애널리스트는 2026년 2월 5일 현재 위 조사분석자료에 언급된 종목의 지분을 보유하고 있지 않습니다.
- 당사는 2026년 2월 5일 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

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