



반도체 (Positive)

Astra가 예고하는 메모리 수요의 폭발적 증가

▶ Analyst 박준영 jyp94@hanwha.com 3772-7481 / RA 김나우 now.kim@hanwha.com 3772-7710

Astra는 이전 AI 모델과 무엇이 다른가

GPT-6 Astra의 가장 큰 변화는 컴퓨터와 업무용 소프트웨어를 직접 사용해 여러 단계의 업무를 끝까지 수행하는 능력이다. 기존 AI도 이메일 작성과 문서 요약, 정보 검색은 가능했지만 사용자가 자료와 작업 순서를 계속 지정해야 했다. Astra는 최종 목표를 전달받으면 필요한 작업을 나누고 웹 브라우저와 스프레드시트, 문서 편집 프로그램 및 사내 시스템을 오가며 업무를 진행한다. 예를 들어 출장비 정산을 맡기면 영수증을 읽는 데서 끝나지 않는다. 영수증과 카드 사용 내역을 대조하고, 회사 규정에 맞지 않는 비용을 찾아 정산 시스템에 입력한 뒤 결재 문서까지 작성하는 방식이다. 채용 업무에서는 지원서 분류와 면접 일정 조율, 안내 메일 작성을 연속해서 수행할 수 있다. OpenAI 또한 Astra의 활용 사례로 온라인 양식 작성과 CRM 업데이트, 일정 관리, 온라인 조사 및 웹사이트 점검 등을 제시하였다.

이러한 업무 수행을 가능하게 하는 기능은 비동기 도구 호출과 장기 Context 관리, Multi-Agent 운용이다. 용어는 생소하지만 작동 방식은 사람이 사무실에서 일하는 것과 유사하다. 비동기 도구 호출은 하나의 프로그램이 작업을 마칠 때까지 기다리지 않고 다른 일을 먼저 처리하는 기능이다. Astra는 ERP에서 데이터를 내려받는 동안 이메일을 정리하거나 계산 프로그램이 실행되는 동안 문서를 작성할 수 있다. 여러 업무를 동시에 처리하면서 업무 속도를 높이는 방식이다. 장기 Context 관리의 긴 업무에서도 이전 지시와 판단을 유지하는 기능이다. Context는 AI가 현재 업무를 위해 참고하는 지시사항과 자료, 중간 산출물 등을 의미한다. 기존 모델은 Context Window(AI가 한 번의 작업에서 동시에 읽고 참고할 수 있는 정보의 최대 범위)가 가득 차면 앞부분을 짧게 요약하는 방식을 사용했는데 이 과정으로 인해 작업의 세부 내용이 사라질 수 있었다. Astra는 핵심 내용을 작업 메모로 남기고, 필요할 경우 이전 Context를 다시 검색한다. 예를 들어 며칠 동안 이어지는 계약서 검토에서도 수정 이력을 참고하며 업무를 이어갈 수 있다. Multi-Agent는 하나의 AI가 모든 일을 순서대로 처리하는 대신 여러 하위 Agent에 역할을 나누는 방식이다. 대표 Agent가 전체 업무를 계획하고, 하위 Agent들이 자료 검색과 데이터 정리, 결과 검증을 동시에 수행한다. 대표 Agent는 이를 취합해 최종 결과물을 만든다. 여러 직원이 하나의 프로젝트를 분담하는 구조를 소프트웨어로 구현한 것이다.

공식 평가에서도 Astra의 강점이 확연히 보인다. 실제 컴퓨터 환경의 업무 수행을 평가하는 OSWorld 2.0에서 Astra는 72.6%를 기록해 GPT-5.6 Sol의 65.7%를 상회했다. 평가상 작업시간도 약 75분에서 40분으로 47% 감소했다. 업무 자동화 능력을 평가하는 AutomationBench 점수는 18.1%에서 41.4%로 두 배 이상 상승했다. 모든 평가에서 경쟁 모델을 앞선 것은 아니다. Artificial Analysis Intelligence Index와 Humanity's Last Exam에서는 Claude Fable 5.1이 Astra보다 높은 점수를 기록했다. Astra의 차별점은 모든 질문에 가장 잘 답한다는 데 있지 않다. 여러 프로그램을 사용하고, 긴 업무의 진행 상황을 기억하며, 필요한 경우 여러 Agent에 작업을 나눠 실제 결과물을 완성하는 능력이 크게 개선됐다는 점이다. OpenAI가 공개한 3D 주택 제작 사례는 이를 직관적으로 보여준다. 사용자가 원하는 분위기를 설명하자 Astra는 건 축물과 가구, 정원, 조명 등이 포함된 편집 가능한 3D 장면을 만들었다. 렌더링 결과를 확인해 물체가 겹치는 부분과 조명을 수정하고, 이후 Unreal Engine 5로 옮겨 사용자가 집 안을 걸어 다닐 수 있도록 구현했다. 문과 서랍을 열거나 조명과 에스프레소 머신을 작동하는 기능도 추가했다. 사람이 방향을 결정하고 결과를 검토했지만 모델링과 코딩, 프로그램 조작, 테스트 및 수정은 Astra가 스스로 계획해 마무리한 것이다.

AI 모델의 개발 방향: 챗봇에서 상시 가동되는 AI 직원으로

주요 AI 기업의 제품 전략은 대화형 AI에서 실제 업무를 수행하는 Agent로 이동하고 있다. 경쟁 기준도 답변의 정확도를 넘어 접근할 수 있는 프로그램과 데이터, 독립적으로 일할 수 있는 시간, Agent의 권한과 비용을 관리하는 능력으로 확대되고 있다.

OpenAI는 2025년 웹사이트에서 예약과 구매 등을 수행하는 Operator를 출시했다. 이후 Operator의 웹 조작 능력과 Deep Research의 조사 기능을 ChatGPT Agent로 통합했다. 검색과 분석, 브라우저 조작 및 문서 작성을 하나의 업무 안에서 처리하는 구조다. ChatGPT Agent는 일정 조정과 출장 계획, 비용 제출 및 스프레드시트 업데이트와 같은 사무 업무를 처음부터 끝까지 수행한다. 특정 작업을 매주 또는 매월 반복하도록 예약하는 기능도 제공한다. 사용자가 필요할 때 질문하는 서비스에서 정해진 일정에 맞춰 업무를 수행하는 서비스로 활용 범위가 넓어진 것이다. 기업용 전략도 Agent 구축과 운영으로 확대되고 있다. OpenAI는 2026년 기업이 Agent를 개발하고 배포, 관리할 수 있는 Frontier를 출시했다. 기업의 데이터와 업무 지침을 Agent에 연결하고, 직원의 피드백을 반영하며, 접근할 수 있는 정보와 행동 범위를 설정하는 플랫폼이다. HP와 Intuit, Oracle, State Farm, Thermo Fisher, Uber 등이 초기 도입 기업으로 공개됐다.

Anthropic도 개발자용 Claude Code에서 일반 사무 업무를 위한 Claude Cowork로 영역을 넓혔다. Cowork는 문서와 스프레드시트, 프레젠테이션 제작, 조사 및 데이터 분석 등의 업무를 수행한다. 사용자가 원하는 결과를 설명하면 Claude가 필요한 파일과 도구를 선택해 작업을 진행한다. Cowork는 사용자가 노트북을 닫아도 클라우드에서 계속 작동한다. 사용자는 휴대전화나 다른 기기에서 진행 상황을 확인하고 방향을 수정할 수 있으며, 반복 업무도 일정에 맞춰 실행할 수 있다. AI가 사용자의 실시간 지시를 기다리는 보조 프로그램에서 사람이 자리를 비운 동안에도 일하는 서비스로 발전한 것이다. Anthropic은 산업별 Agent도 확대하고 있다. 금융 서비스용 Agent를 출시하고 Moody's의 데이터 및 도구와 연결했으며, 중소기업용 Cowork는 QuickBooks와 PayPal, HubSpot 및 Canva를 통해 회계와 결제, 고객관리 및 콘텐츠 제작 업무를 지원한다. Word와 Excel, PowerPoint용 Claude Add-in은 정식 출시됐으며 Outlook 연동도 예고됐다. 범용 챗봇에 산업별 데이터와 업무 프로그램을 결합하는 방향으로 제품 전략이 구체화되고 있다. Agent의 적용 범위가 확대되고, 각 업무가 정기적으로 반복되며, 여러 Agent가 동시에 실행되면 AI 사용량은 사용자의 질문 횟수에 비례하던 기존 방식보다 훨씬 빠르게 증가할 수 있다.

GPU 시장의 변화: 학습에서 상시 가동되는 추론으로

Agent의 확산은 GPU 산업을 대규모 학습에서 상시 추론으로 넓힐 전망이다. 일반적인 챗봇은 질문에 답하면 연산이 끝난다. Agent는 하나의 업무를 완료하기 위해 계획 수립과 정보 검색, 프로그램 실행, 결과 확인 및 오류 수정을 반복한다. 화면을 읽고 프로그램을 조작할 때마다 현재 상황을 다시 해석하며, 예상과 다른 결과가 나오면 추가 추론을 수행한다. Multi-Agent 시스템에서는 이러한 과정이 여러 갈래에서 동시에 진행된다. Anthropic에 따르면 동사 Research 시스템에서 일반 Agent는 채팅 대비 약 4배, Multi-Agent 시스템은 약 15배 많은 Token을 사용했다. 모든 업무에 동일하게 적용되는 수치는 아니지만 AI의 활용 단위가 답변에서 업무로 확대될수록 사용자당 추론량이 크게 증가한다는 점을 보여준다. 기존 생성형 AI 시장에서는 사용자 수와 질문 횟수가 수요를 결정했다. Agent 시대에는 사용자당 Agent 수와 업무 수행시간이 추가된다. 여러 Agent가 백그라운드에서 장시간 일하면 AI 가속기의 가동시간도 크게 늘어난다.

Agent 시대에는 AI 전용 ASIC의 역할도 커질 전망이다. Agent 서비스가 확산되면 동일한 모델의 추론을 수많은 사용자에게 반복적으로 제공해야 한다. 이처럼 규모가 크고 연산 방식이 일정한 서비스에서는 특정 연산에 최적화된 ASIC이 범용 GPU보다 전력효율과 운영비용 측면에서 유리하다. Hyperscaler는 자사 Workload에 맞춰 반도체와 소프트웨어를 함께 설계할 수 있기 때문에 Agent의 상시 가동으로 추론비용이 커질수록 자체 ASIC을 확대할 유인이 강해진다. 다만 Agent Workload 전체가 빠르게 ASIC으로 전환되기는 어렵다. Agent는 업무마다 사용하는 모델과 Context 길이, 도구 및 연산 방식이 달라 처리 부하가 일정하지 않다. 새로운 모델과 추론 기법도 빠르게 등장하고 있어 특정 구조에 맞춰

설계된 ASIC은 환경 변화에 대응하기 어렵다는 한계가 있다. 이에 따라 초기 개발과 복잡하고 다양한 Agent 업무에서는 GPU가 중심을 유지하고, 사용량이 충분히 크고 연산 패턴이 표준화된 서비스부터 ASIC 비중이 높아질 가능성이 크다. Agent 시장의 성장은 GPU 수요를 대체하기보다 전체 추론 시장을 확대하면서 GPU와 ASIC 수요를 동시에 증가시키는 방향으로 전개될 전망이다. 결국 Agent의 확산은 AI 가속기 수요를 기업과 개인의 업무를 처리하기 위해 상시 가동되는 추론 인프라로 전환시킨다. 사용자당 Agent 수와 실행시간이 함께 증가하면서 AI 가속기 수요는 더욱 가파르게 증가할 것으로 예상된다.

메모리 시장의 변화: 대화에서 업무로, 수요의 폭발적 증가 예상

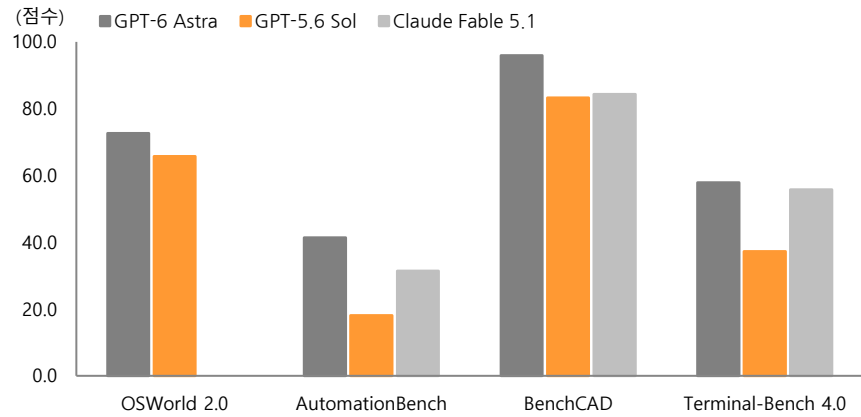
Agent의 확산은 메모리 수요를 크게 늘릴 전망이다. 기존 챗봇은 질문에 답하면 작업을 종료했다. Agent는 이메일과 문서를 읽고, 여러 프로그램을 실행하며, 결과를 확인하고 오류를 수정하는 과정을 업무가 끝날 때까지 반복한다. 작업시간은 수초에서 수시간으로 길어지고, 참고하는 자료도 짧은 질문에서 대량의 문서와 업무 기록으로 확대된다. AI는 한 번 읽은 내용을 매번 처음부터 다시 계산하지 않도록 임시 데이터로 저장한다. 이를 KV Cache라고 한다. 기존 챗봇에서는 질문에 답한 뒤 KV Cache를 지울 수 있어 필요한 용량과 유지시간이 제한적이었다. Agent는 최초 지시와 참고자료, 검색 결과, 프로그램 실행 내용, 중간 산출물 및 과거 판단을 계속 참고해야 한다. 업무가 길어질수록 KV Cache가 커지고 이를 사용하는 시간도 늘어난다. 동시에 실행되는 Agent의 수도 중요한 변수다. 고객정보를 정리하는 Agent와 계약서를 검토하는 Agent, 비용을 정산하는 Agent는 서로 다른 자료와 진행 상황을 기억해야 한다. 여러 Agent가 동시에 일하면 각각 별도의 KV Cache가 필요하다. 한 사람이 AI와 한 번 대화하던 환경에서 한 사람이 여러 Agent에 업무를 맡기는 환경으로 바뀌면 메모리 사용량도 크게 증가한다. 향후 메모리 수요는 AI 사용자 수뿐 아니라 사용자당 Agent 수와 Agent당 Context, 동시 실행시간의 영향을 함께 받게 된다. 기존 생성형 AI 시장에서는 사용자 수와 질문 횟수가 주요 변수였다. Agent 시대에는 각 Agent가 읽는 자료의 양과 업무를 지속하는 시간까지 추가된다. 여러 변수가 동시에 증가한다는 점에서 메모리 수요는 AI 사용자 증가율보다 훨씬 빠르게 늘어날 수 있다.

Astra의 최대 Context Window는 105만 Token이다. Context Window는 AI가 하나의 업무에서 참고할 수 있는 정보의 최대 범위를 의미한다. Astra는 OpenAI의 장문맥 평가에서 51.2만~100만 Token 구간에서도 96.3%의 정보 검색 정확도를 기록했다. 과거에는 Context Window가 길어도 앞부분의 정보를 제대로 활용하지 못하는 경우가 많았다. Astra는 대량의 정보 가운데 필요한 내용을 다시 찾아 실제 업무에 활용하는 능력을 강화했다. 처리하는 데이터의 종류도 늘어난다. Agent는 텍스트뿐 아니라 영수증, 제품 사진, 화상회의 기록 및 컴퓨터 화면 등을 함께 다룬다. Astra의 3D 모델과 게임 제작 사례처럼 AI 업무가 시각적 결과물과 전문 소프트웨어로 확대되면 처리하고 저장해야 하는 중간 데이터도 증가한다. Agent의 수와 함께 Agent당 데이터의 종류와 크기도 커지는 구조다.

이러한 구조적 변화 속에서 각 메모리 계층의 역할은 다음과 같다. AI가 현재 계산에 사용하는 Model과 KV Cache는 주로 GPU와 가장 가까운 HBM에 저장된다. Agent가 읽는 자료와 동시 실행 업무가 늘어나면 더 많은 KV Cache를 HBM에 올려야 한다. AI는 다음 Token을 생성할 때 기존 KV Cache를 반복해서 읽기 때문에 HBM 용량과 대역폭이 함께 중요해진다. Agent 확산은 GPU당 HBM 탑재량 증가와 고대역폭 제품으로의 전환을 동시에 촉진할 전망이다. 하지만 모든 데이터를 가격이 높은 HBM에 계속 보관할 수는 없다. 현재 사용하는 데이터는 HBM에 두고, 곧 다시 사용할 데이터는 서버 DRAM, 장기간 보관할 원본 자료와 업무 기록은 Enterprise SSD에 저장해야 한다. Agent 수가 증가할수록 HBM 수요가 늘어나고, 업무시간과 데이터가 누적될수록 서버 DRAM과 Enterprise SSD 수요도 함께 확대된다.

Astra가 시사하는 변화는 메모리 수요를 결정하는 변수가 동시에 늘어난다는 점이다. AI가 수행하는 업무의 수와 길이, 동시에 실행되는 Agent, Agent가 참고하는 Context가 서로 곱해지면서 메모리 수요의 증가 속도는 기존 예상보다 비약적으로 가팔라질 수 있다. HBM을 시작으로 서버 DRAM과 Enterprise SSD까지 수요가 확산되며, AI Agent는 메모리 산업에 새로운 대규모 수요를 형성할 것으로 기대한다.

[그림1] 최신 AI 모델 벤치마크



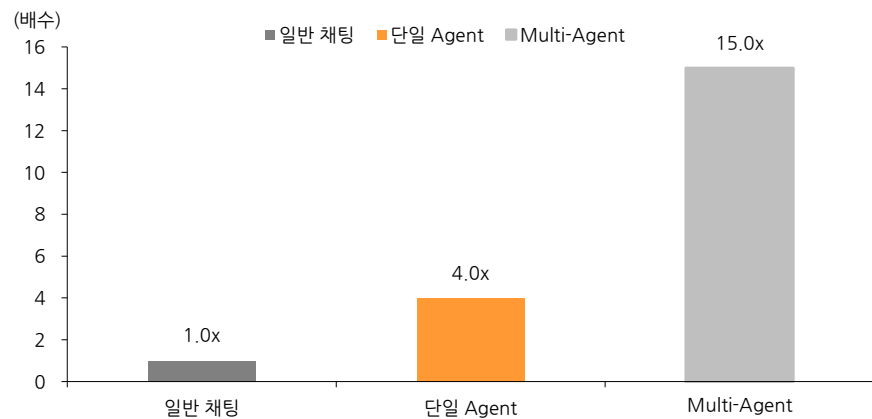
자료: OpenAI, 한화투자증권 리서치센터

[표1] OpenAI와 Anthropic의 로드맵

회사	시기	제품	주요 역할	실행 범위
OpenAI	2025	Operator	브라우저 행동	웹사이트 조작
	2025	ChatGPT agent	조사·분석·행동 통합	브라우저·문서·스프레드시트
	2026	Frontier	기업 Agent 구축·배포·관리	기업 데이터·권한·평가
	2026	GPT-6 Astra	장시간·복합 업무 수행	다중 도구·장기 Context·Multi-Agent
Anthropic	2025	Claude Code	소프트웨어 개발 Agent	코드·터미널·개발 도구
	2026	Cowork	일반 사무 업무	문서·표·조사파일
	2026	Managed Agents	장기 실행 Agent 운영	클라우드·프로그램 방식
	2026	Claude Tag·산업별 Agent	팀·금융 등 특화 업무	Slack·산업 데이터·업무 앱

자료: OpenAI, Anthropic, 한화투자증권 리서치센터

[그림2] 업무 방식별 상대 토큰 사용량



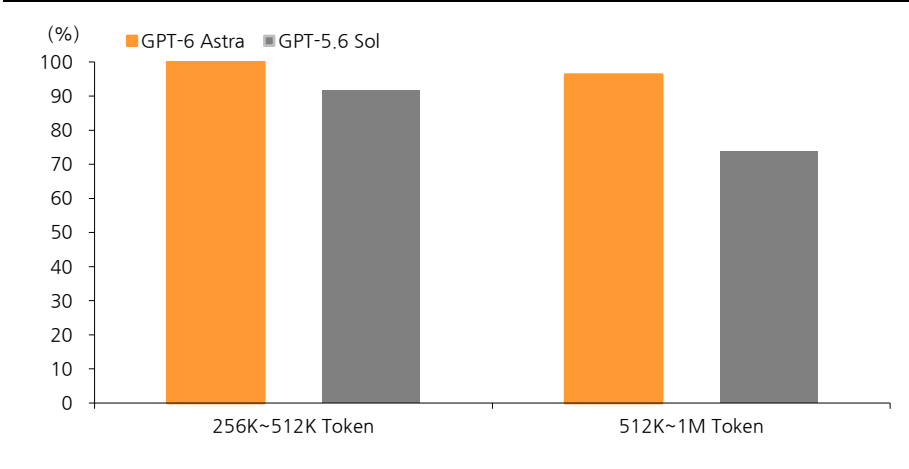
자료: Anthropic, 한화투자증권 리서치센터

[표2] 에이전트 확산에 따른 메모리별 용처

메모리	용처	Agent 확산 시 늘어나는 이유
HBM	GPU가 지금 계산 중인 모델-KV Cache	동시에 실행되는 Agent와 긴 Context가 증가
서버 DRAM	곧 다시 쓸 Context와 중간 결과	HBM에 모두 둘 수 없어 가까운 곳에 대기
Enterprise SSD	업무 기록·문서·장기 Context	업무가 길어지고 결과가 누적

자료: 한화투자증권 리서치센터

[그림3] 장문맥 정보 검색 정확도



자료: OpenAI, GPT-6 Astra 공식 발표(MRCC 8-needle), 한화투자증권 리서치센터

[그림4] GPT-6 Astra의 3D 모델링



자료: OpenAI, 한화투자증권 리서치센터

[Compliance Notice]

(공표일: 2026년 9월 10일)

이 자료는 조사분석 담당자 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성됐음을 확인합니다. 본인은 이 자료에서 다룬 종목과 관련해 공표일 현재 관련 법규상 알려야 할 재산적 이해관계가 없습니다. 본인은 이 자료를 기관투자자 또는 제 3자에게 사전에 제공한 사실이 없습니다. (박준영, 김나우)
저희 회사는 공표일 현재 이 자료에서 다룬 종목의 발행주식을 1% 이상 보유하고 있지 않습니다.

이 자료는 투자자의 증권투자를 돕기 위해 당사 고객에 한하여 배포되는 자료로서 저작권이 당사에 있으며 불법 복제 및 배포를 금합니다. 이 자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료나 정보출처로부터 얻은 것이지만, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 이 자료는 어떠한 경우에도 고객의 증권투자 결과와 관련된 법적 책임소재에 대한 증빙으로 사용될 수 없습니다.

[산업 투자 의견]

당사는 산업에 대해 향후 1년간 해당 업종의 수익률이 과거 수익률에 비해 양호한 흐름을 보일 것으로 예상되는 경우에 Positive(긍정적) 의견을 제시하고 있습니다. 또한 향후 1년간 수익률이 과거 수익률과 유사한 흐름을 보일 것으로 예상되는 경우에 Neutral(중립적) 의견을, 과거 수익률보다 부진한 흐름을 보일 것으로 예상되는 경우에 Negative(부정적) 의견을 제시하고 있습니다. 산업별 수익률 전망은 해당 산업 내 분석대상 종목들에 대한 담당 애널리스트의 분석과 판단에 따릅니다.

[당사 조사분석자료의 투자등급 부여 비중]

(기준일: 2026년 6월 30일)

투자등급	매수	중립	매도	합계
금융투자상품의 비중	96.3%	3.7%	0.0%	100.0%