

미국은 지금

엔비디아가 만든 AI 비서, NemoClaw

키움증권 리서치센터 글로벌리서치팀

US Strategy 김승혁 ocean93@kiwoom.com



Issue Brief

GTC 2026 의 시사점(Feat, NemoClaw)

2026 년 3 월 9 일, 엔비디아가 'NemoClaw'라는 오픈소스 AI 에이전트 플랫폼 출시를 준비 중이라는 내용이 공개되었다. 플랫폼 발표 시점은 3 월 16 일 개막하는 연례 개발자 행사 GTC 2026 직전으로 예상되며 이미 구글, 시스코, 세일즈포스, 어도비, CrowdStrike 등과의 파트너 협상이 동시에 진행 중인 것으로 확인된다. NemoClaw 는 기업이 자사 직원에게 AI 에이전트를 배포할 수 있게 해주는 오픈소스 AI 에이전트 플랫폼이며, 중요한 특징으로 엔비디아 칩 없이도 작동한다는 점이 있다. 칩 메이커가 칩 없이도 쓸 수 있는 소프트웨어를 무상으로 푼다는 선택은, 단순한 제품 출시가 아닌 사업 모델 전환의 계기로 해석해야 한다.

'대화하는 AI'에서 '일하는 AI'

ChatGPT 나 Gemini 가 질문에 답하는 AI 라면, NemoClaw 는 스스로 계획하고 실행하는 AI 다. 사람이 자리를 비운 사이에도 데이터를 수집하고, 정리하고, 보고서를 완성해 메신저로 발송하는 것이 가능한 구조다. 이 플랫폼의 두뇌인 Nemotron 3 는 Gemini·GPT-4 와 정면으로 경쟁하는 엔비디아 자체 AI 모델이며, 기업 서버에 직접 설치할 수 있어 외부 데이터 유출 우려도 없다. 세일즈포스·SAP·마이크로소프트·팔란티어 등 전 세계 수만 개 기업을 고객으로 둔 소프트웨어 강자들과의 파트너십은, AI 에이전트가 기업 실무 현장에 침투하는 가장 빠른 경로가 될 것이다.

칩 메이커에서 플랫폼 제국으로

NemoClaw 는 더 큰 퍼즐의 한 조각일 뿐이다. 맞춤형 반도체의 약진에 따라 CUDA 라는 독점 생태계에 우려가 생기기 시작한 지금, 엔비디아는 하드웨어 종속 전략을 고집하는 대신 소프트웨어 생태계의 표준을 먼저 장악하는 쪽을 택했다. 이에 언어 AI 를 넘어 로보틱스(Cosmos·Isaac GR00T), 자율주행(Alpamayo), 바이오(Clara), 통신(Nemotron LTM)까지 전 산업을 관통하는 오픈소스 플랫폼 전략이 동시에 전개 중이다. 투자자 입장에서 핵심 질문은 하나다. 소프트웨어 생태계 확장이 실제 하드웨어 수요 창출로 이어지는 속도가 얼마나 빠를 것인가? 부분이다. 젠슨 황이 GTC 무대에서 이 연결고리를 얼마나 구체적으로 제시하느냐가, 엔비디아를 칩 회사에서 AI 인프라 플랫폼 기업으로 재평가하는 계기가 될 것이다.

NemoClaw 관련 내용 정리

항목	내용
플랫폼명	NemoClaw (네모클로)
출시 시점	GTC 2026 컨퍼런스(예정)
플랫폼 성격	오픈소스 AI 에이전트 배포 플랫폼
주요 기능	기업이 직원용 AI 에이전트 배포 및 운영
하드웨어	NVIDIA GPU 무관 (모든 환경에서 사용 가능)
포함 도구	보안 및 개인정보 보호 도구
접근 방식	오픈소스 (파트너에게 무료 초기 접근)
전략적 의도	엔터프라이즈 소프트웨어 시장 진출

자료: 키움증권 리서치

엔비디아 Agent Toolkit



자료: 엔비디아

엔비디아가 만든 AI 비서, NemoClaw

26 년 3 월 9 일, 엔비디아가 기업용 AI 비서 플랫폼을 곧 무료 공개하고, 대형 소프트웨어 기업들과 손잡으려 한다는 내용이 공개됐다. 이를 위해 엔비디아는 'NemoClaw'라는 플랫폼을 준비 중이며, 발표 시점은 3월 16~19일 열리는 GTC 2026 직전으로 알려져 있다. 현재는 이미 시스코, 세일즈포스, 구글, 어도비, CrowdStrike 등과 파트너십 협상 중이다.

AI 에이전트의 개념은 사람이 일일이 지시하지 않아도 스스로 계획을 세우고 단계별로 업무를 처리하는 AI 다. NemoClaw 는 이런 AI 에이전트를 기업들이 자기 직원들에게 배포할 수 있게 해주는 플랫폼이다. 핵심 특징은, 엔비디아 칩을 사용하지 않아도 사용할 수 있단 점이다. 보통 엔비디아 소프트웨어는 엔비디아 GPU 가 있어야 제대로 돌아가는데, NemoClaw 는 어떤 AMD 등 다른 회사 칩 위에서도 작동하도록 설계했다. 자기 칩 안 쓰는 기업도 자사 플랫폼 생태계 안에 끌어들이겠다는 엔비디아의 의도가 담긴 전략이다. 플랫폼은 소스코드를 무료로 공개하는 오픈소스 방식으로 제공된다. 이에 따라 파트너사들은 초기에 무료로 쓸 수 있는 대신, 플랫폼 개발에 기여하게 한다. 보안·개인정보 보호 기능도 기본으로 탑재된다.

ChatGPT·Gemini 와 무엇이 다른가?

한편 NemoClaw 가 플랫폼이면, 이를 구동하기 위한 AI 엔진은 Nemotron 이다. 즉 두뇌인 것이다. 이 개념이 헛갈린다면 우리에게 익숙한 Gemini(구글), ChatGPT(OpenAI)와 비교하자.

우선 Nemotron 과 Gemini 의 관계는 경쟁 관계에 있는 AI 모델이다. Gemini 는 구글이 만든 AI 두뇌, Nemotron 은 엔비디아가 만든 AI 두뇌다. 삼성 갤럭시와 애플 아이폰처럼, 서로 다른 회사에서 만든 별개의 제품이다. 다만, 차이가 있다. Gemini 는 구글의 거대한 서버에서만 돌아가기 때문에 사용자가 질문을 하면 데이터가 구글 서버를 반드시 거친다. 반면 Nemotron 은 설계도(모델 가중치)가 통째로 공개되어 있어서, 기업이 자기 서버에 직접 내려받아 설치할 수 있다. 보안에 민감한 기업 입장에서는 이 차이가 매우 크다. 내부 기밀 문서나 고객 데이터를 AI 에게 맡기면서도 외부 유출 걱정을 하지 않아도 되기 때문이다.

NemoClaw와 Gemini·ChatGPT 의 관계는 결이 다르다. Gemini 나 ChatGPT 는 사용자가 직접 창을 열고 질문을 입력해야만 움직이는 '대화 상대'다. 반면 NemoClaw 는 대화 상대가 아니라, AI 두뇌(Nemotron)를 장착해서 백그라운드에서 혼자 일하게 만드는 '자동화 플랫폼'이다. 예컨대 Gemini 에게 "경쟁사 뉴스 요약해줘"라고 하면 그 순간에만 답변을 준다. 하지만 NemoClaw 는 "매일 아침 8시에 경쟁사 뉴스를 자동으로 수집해서 내부 메신저로 보내줘"라고 한 번 세팅해두면, 이후에는 내가 자고 있어도 알아서 인터넷을 뒤지고 정리하고 발송까지 완료한다. 즉, Gemini 나 ChatGPT 가 '물어볼 때만 답하는 조연자'라면, NemoClaw 는 '실무 담당자'에 가깝다.

정리하면, Nemotron 은 엔비디아가 만든 AI 두뇌(모델)이고, NemoClaw 는 그 두뇌를 넣어서 기업 업무를 자율적으로 처리하게 만드는 플랫폼(몸통)이다. Gemini 나 ChatGPT 는 이 둘과 각각 다른 차원에서 경쟁한다. 모델 성능 측면에서는 Nemotron 이 Gemini·GPT-4 와 경쟁하고, 에이전트 플랫폼 측면에서는 NemoClaw 가 OpenAI 의 에이전트 서비스나 구글의 Agentspace 와 경쟁하는 구도다.

NemoClaw의 두뇌: Nemotron

NemoClaw는 3월 공개 예정이지만, 두뇌 역할 모델 중 Nemotron 3 중 Nano 모델은 25년 12월 이미 공개되었다. Nemotron 3는 크게 Nano, Super, Ultra로 나뉜다. Nano는 지금 당장 쓸 수 있는 소형 모델이다. 파라미터 약 300억 개짜리이며 문서 요약, 소프트웨어 디버깅, 간단한 정보 검색처럼 빠르고 저렴하게 써야 하는 업무에 최적화되어 있다. 이전 세대(Nemotron 2 Nano)보다 처리 속도가 4배 빠르다.

Super는 약 1,000억 파라미터짜리 모델로, 여러 AI 에이전트가 서로 협력해서 복잡한 작업을 처리하는 데 특화된다. Ultra는 약 5,000억 파라미터의 대형 모델로, 깊은 리서치나 전략 수립 같은 고난도 업무를 담당한다. Nano 상위 보전인 둘 다 2026년 상반기 출시 예정이다.

기술적으로 중요한 포인트 두 가지를 짚으면, 하나는 MoE(Mixture-of-Experts) 구조다. 이는 모든 회로를 항상 켜두는 게 아니라, 입력에 따라 필요한 회로만 골라서 활성화하는 방식이다. 자동차로 비유하면, 시내 주행할 때는 엔진 일부만 쓰고 고속도로에서는 풀파워를 쓰는 것이다. DeepSeek 역시 MoE 방식을 통해 모델 효율을 높인 사례가 있다. 다른 하나는 100만 토큰 컨텍스트 윈도우다. 이는 AI가 한 번에 기억하고 처리할 수 있는 텍스트 분량을 뜻하는데, 100만 토큰이면 두꺼운 책 한 권을 통째로 넣고 대화하는 수준이다. 따라서 장기간 이어지는 복잡한 업무 처리에 유리하다. Nemotron 모든 모델은 Hugging Face라는 오픈소스 플랫폼에 공개되며, 누구나 다운로드해서 쓸 수 있다.

이미 손잡은 기업들

파트너십은 이미 상당히 진행되어 있다. Microsoft는 Azure 클라우드와 Office 365 서비스에 통합 중이고, SAP는 자사 AI 비서 'Joule'에 탑재했다. ServiceNow는 사내 업무 자동화에, 컨설팅 기업 Deloitte는 'Zora AI'라는 자체 에이전트 플랫폼에 녹여냈다. Palantir는 기업 데이터 구조와 연결하는 방식으로 특화 에이전트를 만들고 있다.

이 파트너십이 중요한 이유는 단순히 기업 숫자가 많아서가 아니다. 세일즈포스나 ServiceNow는 전 세계 수만 개 기업에 업무용 소프트웨어를 공급하는 회사들이다. 이들과 협력하면, 엔비디아 기술이 그 밑단 기업들에게까지 자연스럽게 스며들 수 있다. 즉, 최종 소비자 접점이 한 번에 폭발적으로 늘어난다

기업별 활용 내용 및 종류

기업명	활용 내용
Microsoft	Azure AI Foundry 통합, Azure AI Agent Service 및 Microsoft 365 서비스 강화
SAP	SAP Business AI 및 AI 코파일럿 'Joule'에 탑재
ServiceNow	엔터프라이즈 생산성 향상 AI 에이전트 구축
Deloitte	'Zora AI' 에이전트 플랫폼에 통합
Palantir	Ontology 프레임워크와 통합, 특화 AI 에이전트 스택 구축

자료: 엔비디아, 키움증권 리서치

제 2의 CUDA를 위하여

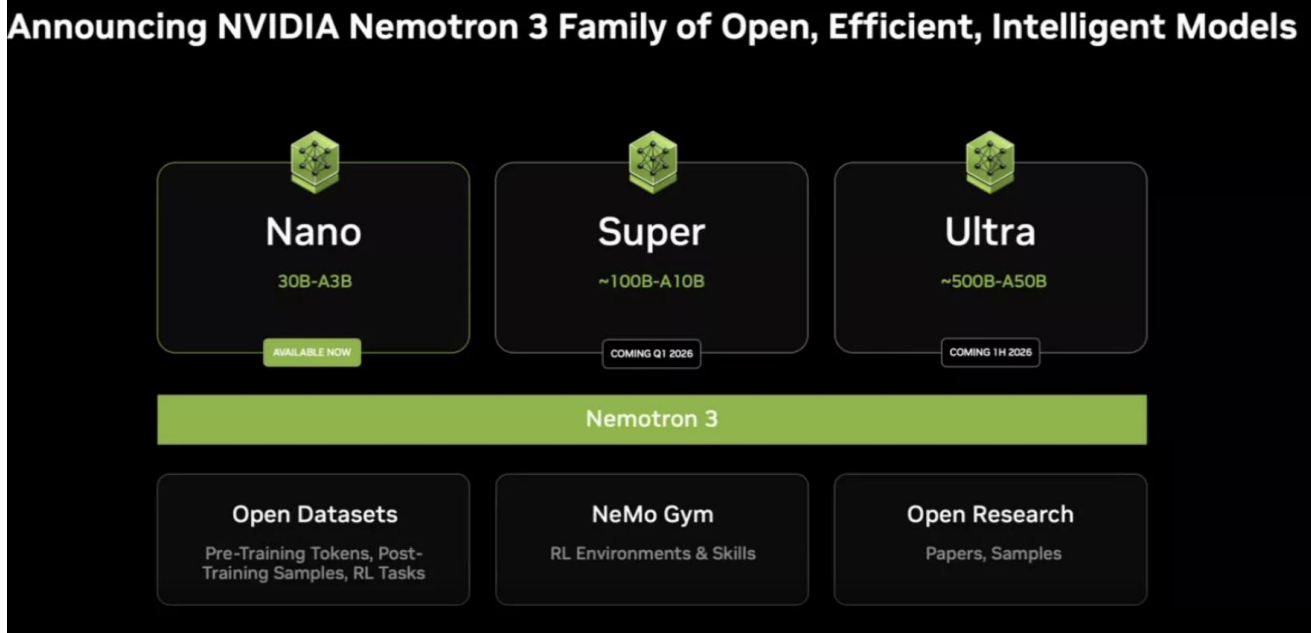
엔비디아의 가장 강력한 무기는 지금까지 CUDA였다. CUDA는 엔비디아 GPU를 AI 연산에 활용하기 위한 전용 개발 환경으로, 2007년부터 쌓아올린 소프트웨어 생태계다. 전 세계 AI 개발자의 대다수가 CUDA 기반으로 코드를 짜왔기 때문에, 다른 회사 칩으로 갈아타려면 코드를 처음부터 다시 짜야 한다. 이게 엔비디아의 진짜 해자, 즉 넘기 힘든 경쟁 우위였다. 그런데 이 해자에 약간의 균열이 생기고 있다. OpenAI, Anthropic 같은 AI 기업들이 자체 칩을 만들기 시작했고, OpenAI가 만든 'Triton'이라는 오픈소스 도구는 엔비디아 GPU가 아닌 AMD나 Intel 칩에서도 AI 코드를 효율적으로 돌릴 수 있게 해준다. CUDA에 묶여 있던 개발자들이 탈출할 수 있는 경로가 생긴 것이다. NemoClaw은 이 위기에 대한 엔비디아의 답변이다. 논리 구조를 풀어보면 이렇다. '우리 칩을 안 쓰더라도 우리 플랫폼을 쓰게 만들자. 그러면 AI 에이전트가 돌아가는 생태계의 표준을 우리가 쥐게 된다. 그 생태계가 커지면, 결국 무거운 연산 작업은 우리 GPU로 하게 돼 있다.' 소프트웨어로 먼저 생태계를 장악하고, 하드웨어 수요는 나중에 따라오게 하는 전략이다. 'claws' 트렌드도 배경에 있다. 클라우드가 아닌 개인 PC나 사내 서버에서 직접 돌아가는 오픈소스 AI 에이전트들이 급성장하는 현상인데, OpenAI가 이 분야의 유명 오픈소스 프로젝트인 'OpenClaw'를 인수하면서 이 흐름에 뛰어들었다. 엔비디아도 그냥 지켜볼 수 없어서 직접 대응했다고 판단한다.

엔비디아의 큰 그림과 2026 GTC 컨퍼런스 관전 포인트

NemoClaw은 엔비디아가 그리는 큰 그림의 한 조각이다. 엔비디아는 현재 언어 AI 외에도 여러 분야에 오픈소스 플랫폼을 동시에 출시하고 있다. 로봇용 Cosmos, 자율주행차용 Alpamayo, 의료 AI 용 Clara, 통신망 자동화용 Nemotron LTM 등이다. 이 모든 플랫폼에 쏟아붓는 오픈소스 학습 데이터만 해도 언어 10조 토큰, 로봇 동작 데이터 50만 건, 단백질 구조 45만 건, 자율주행 센서 데이터 100TB에 달한다. 한 도메인에서 경쟁하는 게 아니라, AI가 쓰이는 모든 산업에서 동시에 플랫폼 표준이 되겠다는 선언이다.

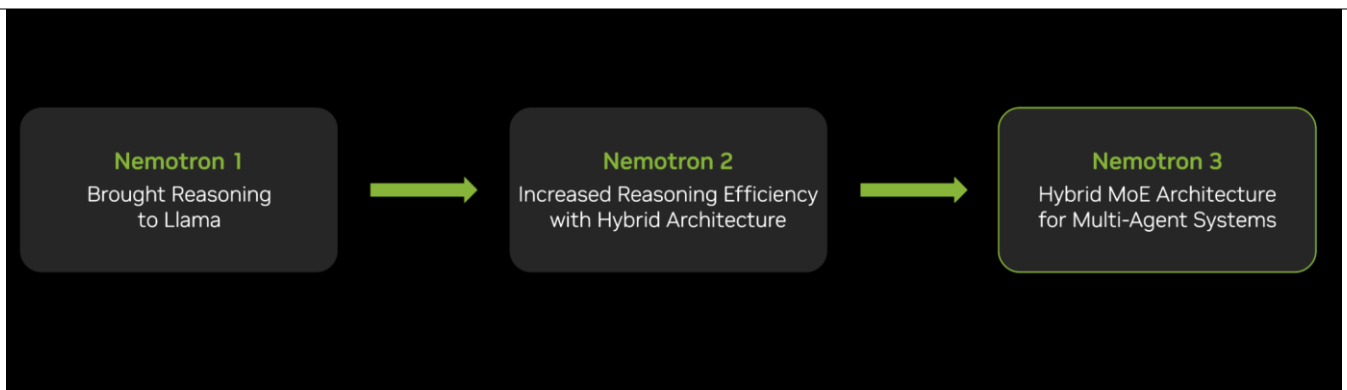
GTC 2026은 단순한 기술 행사가 아니다. 엔비디아가 '칩 회사'에서 '플랫폼 회사'로 전환하겠다는 공식 선언의 무대다. 이미 작년 10월의 GTC에서 AI 인프라 전반을 담당하는 기업으로의 변화를 선언했고, GTC 2026은 소프트웨어까지 영역 확대를 노려볼 수 있다. 물론 단기적으로는 칩 비중속 전략이 자사 하드웨어 판매와 직접적인 연결고리를 약화시킬 수 있다는 우려도 발생할 수 있다. 하지만 중장기 관점에서 보면, 소프트웨어 생태계를 먼저 장악하면 AI 인프라 전체의 공급망 상단에 위치하게 되어, 칩 경쟁이 심화되더라도 안정적인 수익 기반을 유지할 수 있다. CUDA가 이를 방증한 대표적 사례이다. 젠슨황 CEO가 GTC에서 이 소프트웨어 전략의 구체적인 수익화 방안을 어떻게 제시하느냐가 중요한 관전 포인트이다.

엔비디아의 Nemotron3 구성



자료: 엔비디아, 키움증권 리서치

Nemotron의 진화기



자료: 엔비디아, 키움증권 리서치

Nemotron3 AI 모델 라인업 및 사양

모델	파라미터	활성 파라미터	주요 용도	배포 환경	출시 상황
Nano 3	1.6B	3.6B	엡지/PC/단일 GPU 작업	엡지~단일 GPU	출시 완료
Super	~100B	~10B	다중 에이전트 협업	단일 H100/B200 GPU	2026 상반기 예정
Ultra	~500B	~50B	고차 추론 & 전략 계획	데이터센터 (다중 GPU)	2026 상반기 예정
아키텍처	Mamba-Transformer MoE	효율성 극대화	1M 토큰 컨텍스트	장문맥 추론	-

자료: 키움증권 리서치

Nemotron3 AI 모델 주요 기술 특징 및 기대 효과

기술 영역	주요 혁신 사항	이점
혼합 MoE	Hybrid Mamba-Transformer 구조	처리량 4배 증가 (Nano 기준)
컨텍스트 윈도우	1M 토큰 네이티브 지원	장문서/장대화 처리 가능
모드 제어	추론 온/오프 + 사고 예산 제어	정확도-비용 트레이드오프 최적화
강화학습	Multi-environment RL (NeMo Gym)	실제 다단계 작업에 최적화
정밀도	NVFP4 4비트 학습/추론 지원	메모리 50% 절감, 속도 향상
토큰 예측	Multi-token prediction (Super/Ultra)	추론 처리량 향상
추론 토큰 생성	60% 감소 (Nano)	비용 절감 및 속도 개선

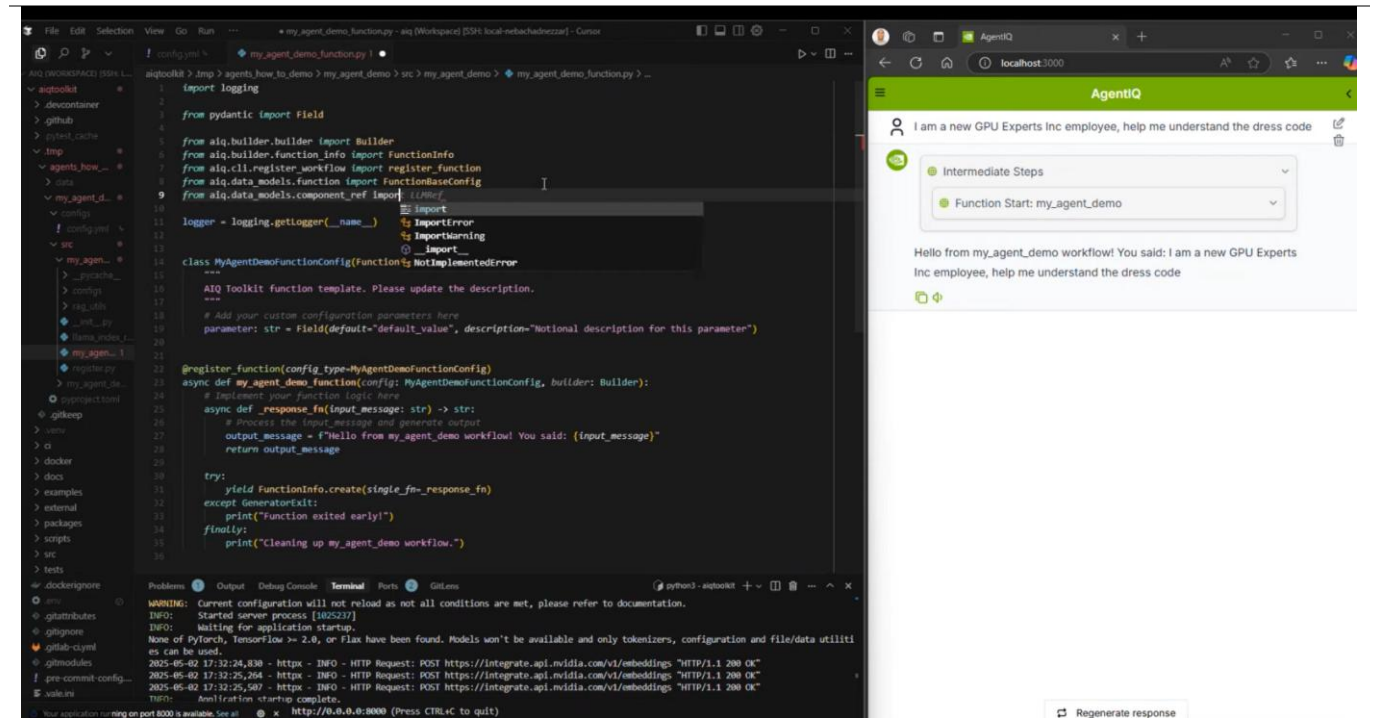
자료: 키움증권 리서치

파트너사 산업 분야 및 협력 부문

파트너사	산업 분야	협력 분야
Salesforce	CRM/엔터프라이즈	Einstein Assistant 강화
Adobe	콘텐츠 & 마케팅	Firefly/AI 기능 확대
Cisco	네트워킹/보안	인프라 자동화
Google	클라우드/생산성	Workspace 통합
CrowdStrike	사이버보안	위협 분석 자동화
파트너십 형태	무료 초기 접근 + 기여도 기반	오픈소스 협력 모델
경쟁 우위	NVIDIA의 포괄적 에이전트 플랫폼	주요 SaaS 기업들과 깊은 통합

자료: 키움증권 리서치

엔비디아의 NemoClaw 시현 영상



자료: 엔비디아, 카움증권 리서치

Compliance Notice

- 당사는 동 자료를 기관투자자 또는 제 3자에게 사전 제공한 사실이 없습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.

고지사항

- 본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없고, 통지 없이 의견이 변경될 수 있습니다.
- 본 조사분석자료는 유가증권 투자를 위한 정보제공을 목적으로 당사 고객에게 배포되는 참고자료로서, 유가증권의 종류, 종목, 매매의 구분과 방법 등에 관한 의사결정은 전적으로 투자자 자신의 판단과 책임하에 이루어져야 하며, 당사는 본 자료의 내용에 의거하여 행해진 일체의 투자행위 결과에 대하여 어떠한 책임도 지지 않으며 법적 분쟁에서 증거로 사용 될 수 없습니다.
- 본 조사 분석자료를 무단으로 인용, 복제, 전시, 배포, 전송, 편집, 번역, 출판하는 등의 방법으로 저작권을 침해하는 경우에는 관련법에 의하여 민·형사상 책임을 지게 됩니다.