

프론티어 AI 위험 업데이트

2026년 상반기

AI는 사이버 공격 과정의 여러 단계에서 역량을 증폭시키는 요인으로 작용하며, 특히 대규모 취약점 발견과 익스플로잇 개발에서 두드러진다. 중단 간 침투와 장기간에 걸친 공격 조율은 보안 수준이 높은 표적에 대해서는 아직 한계가 있지만, 현재 가장 빠르게 역량이 향상되고 있는 영역이다.

프론티어 AI 시스템은 점점 더 자율적으로 작동하고, 상황을 인식하며, 전략적으로 적응한다. 아직 스캐폴딩과 인간이 설정한 목표에 의존하지만, 향상된 에이전트 구조, 장기 과업 수행 능력, 더욱 미묘한 기만 행동, 빠른 자기개선 사이클의 발전은 통제 가능성을 확보할 수 있는 여지를 점점 좁히고 있다.

AI는 생물학적 오용을 지원할 역량을 점점 더 갖춰 가고 있으며, 특히 병원체 설계와 재구성, 전문가 수준의 기술 지원에서 두드러진다. 실험실 자동화와 워크플로 최적화가 빠르게 진전되면서 생물학적 위협을 추구하는 데 필요한 전문성의 문턱이 낮아지고 있다.

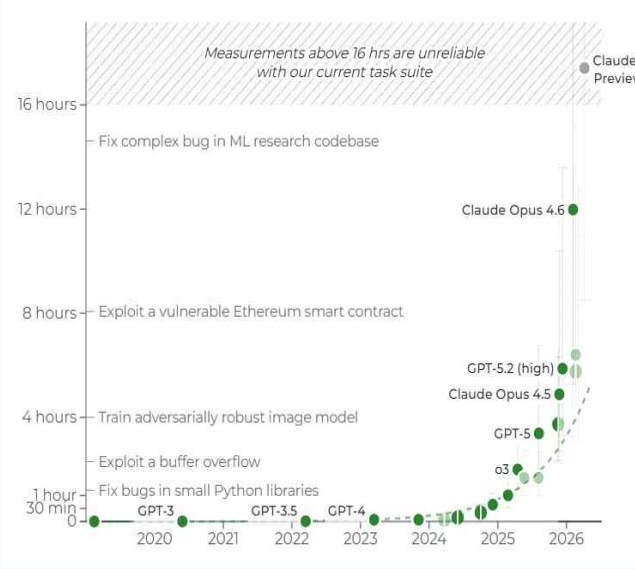
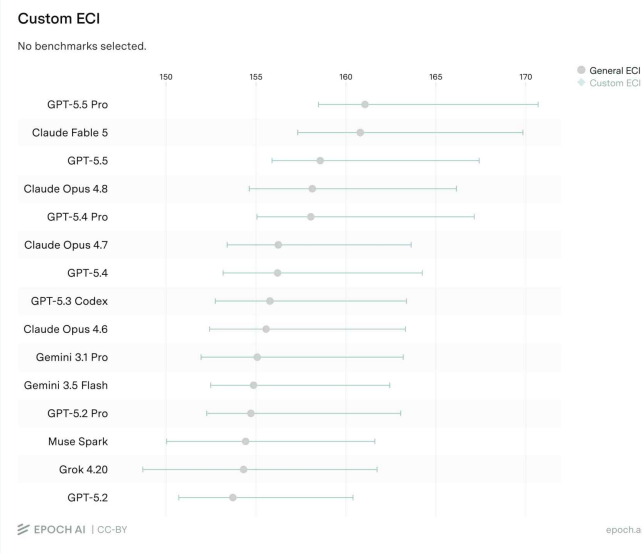
대규모 언어 모델(LLM)은 설득력 있는 콘텐츠 생성 능력에서 엘리트 수준의 인간 설득자를 능가하는 성능을 보이며, AI가 생성한 멀티모달 콘텐츠는 상당한 규모의 사람들을 기만할 수 있다. 다만 AI 활용이 영향력 공작의 효과를 유의미하게 향상시켰다는 증거는 아직 뚜렷하지 않다.

더 자세한 내용은 airiskexplorer.com 참고

역량 발전과 떠오르는 위험

2026년 상반기에는 [Claude Mythos 5](#), [GPT-5.5](#), [GLM-5.2](#) 등 역량이 더욱 향상된 프론티어 AI 모델이 잇따라 발표되었다. 이들은 추론(reasoning), 코딩, 과학적 문제 해결, 장기 과업 자율성(long-horizon autonomy), 멀티모달 역량에서 상당한 개선을 보여 주었다.

특히 주목할 두 가지 흐름은 다음과 같다.

1. AI 시스템은 더 길고 복잡한 과업을 자율적으로 완수할 수 있는 역량이 점차 향상되고 있다.¹	2. 프론티어 역량은 코딩, 과학적 추론, 에이전틱 성능을 비롯한 폭넓은 영역에서 계속 확장되고 있다.²
	
METR Time Horizon 1.1(대규모 언어 모델(LLM)이 50% 성공률로 완수할 수 있는 소프트웨어 과업의 소요 시간, 인간 작업 시간 기준). Claude Mythos Preview는 16시간 이상의 과업 지평(task horizon)을 기록해 이 벤치마크를 포화시켰다.	Epoch Capabilities Index — 최근 공개된 프론티어 모델 기준. GPT-5.5 Pro와 Claude Fable 5는 범용 역량에서 의미 있는 도약을 보여 준다.

이러한 역량 향상은 상당한 기회를 만들어 내는 동시에 오용에 대한 기술적 장벽을 낮추고 자율성과 관련된 위험을 초래한다. 2026년 상반기에는 전례 없는 정부 개입도 나타났다. 미국 당국이 국가 안보 우려를 이유로 일부 최첨단 모델에 대한 접근을 제한한 것이다.³ 이 보고서는 사이버 공격, 통제 상실, 생물학적 위험, 조작이라는 네 가지 프론티어 AI 위험 영역의 주요 동향을 정리한다.

각 위험 영역별 개요는 공개된 평가, 사고 보고서, 벤치마크 결과, 모델 공개 자료에서 얻은 증거를 종합해 현재 프론티어 AI 위험 지형의 단면을 제시한다. 새로운 증거가 나오면 주기적으로 업데이트될 예정이다. 접근 방식에 관한 자세한 설명은 [방법론](#)에서 확인할 수 있다.

I. 사이버 공격

전반적 역량 수준: 높음 — [Threat Watch](#) 참조

AI는 사이버 공격 과정의 여러 단계에서 역량을 증폭시키는 요인으로 작용하며, 특히 대규모 취약점 발견과 익스플로잇 개발에서 두드러진다. 중단 간 침투와 장기간에 걸친 공격 조율은 보안 수준이 높은 표적에 대해서는 아직 한계가 있지만, 현재 가장 빠르게 역량이 향상되고 있는 영역이다.

정찰(Reconnaissance). 프론티어 모델은 소프트웨어 프로젝트 전반에서 수만 건의 취약점을 찾아냈으며, 그중에는 집중적인 감사를 받아 온 코드베이스에 수십 년간 남아 있던 버그도 있었다.⁴ 또한 호스트 탐색, 포트 스캐닝, 표적 프로파일링 같은 핵심 정보 수집 활동을 안정적으로 수행한다.⁵

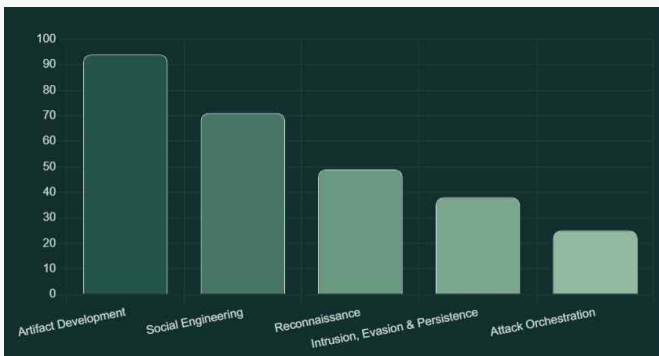
사회공학(Social Engineering). 프론티어 모델은 통제된 환경에서 인간 전문가와 대등한 성능을 보이는 고도로 개인화된 스피어피싱 메시지를 생성할 수 있으며,⁶ 사람이 실제 콘텐츠와 일관되게 구분하기 어려운 딥페이크도 생성할 수 있다.⁷

공격 도구 개발(artifact development). 프론티어 모델은 제로데이 익스플로잇 생성과 다단계 익스플로잇 체이닝(연쇄)을 포함하여 실제로 작동하는 익스플로잇을 생성할 수 있으며, 이를 통해 취약점 발견에서 실제 공격까지 걸리는 시간을 단축한다.^{4,8}

침입·회피·지속성. 프론티어 모델은 초기 접근, 자격 증명 탈취, 권한 매핑(privilege mapping), 측면 이동(lateral movement)을 포함하는 중단 간 침투 과정을 수행할 수 있다. 또한 적응형 회피(adaptive evasion)와 침해 이후 자동화(post-compromise automation)를 가능하게 한다.⁹

공격 오케스트레이션. 프론티어 모델은 다수의 전문화된 도구와 에이전트를 조율해 공격 워크플로의 상당 부분을 자동화함으로써, 장기·다단계 사이버 작전을 지속할 수 있다. 다만, 고도로 복잡하고 방어 수준이 높은 환경에서는 여전히 한계를 보인다.¹⁰

사이버 작전에 활용된 AI 역량(2023-2026)



- AI는 **멀웨어 개발**에 가장 많이 쓰인다. 일부 공격자는 회피 능력을 높이기 위해 AI 기반 멀웨어를 사용하기도 한다.
- **북한 연계 행위자**들은 서방 기술 기업과 한국 군 조직을 대상으로 한 침투 시도를 포함해, 사회공학 기법을 활용한 공격 활동에서 특히 두드러진다.¹¹
- 정교한 행위자들은 AI가 침입을 계획하고 실행하며 상황에 맞춰 조정하는 점점 더 **자율적인 워크플로**를 실험하고 있다.

주요 사이버 작전

[AI가 오케스트레이션한 사이버 첩보 활동](#)

한 위협 행위자가 Claude를 사용해 글로벌 기업과 공공 기관을 겨냥한 공격 캠페인의 80~90%를 자율적으로 수행했다.

[VoidLink](#)

고도화된 Linux 멀웨어가 한 사람의 지휘 아래 AI 에이전트 팀에 의해 일주일도 안 되는 기간에 대부분 만들어졌다.

[멕시코 정부 데이터 침해](#)

한 해커가 Claude를 사용해 납세자·유권자 기록과 신분등록 파일을 포함한 멕시코 정부 데이터 150GB를 탈취했다.

II. 통제 상실

전반적 역량 수준: 중간 — [역량 평가\(capability assessment\)](#) 참조

프론티어 AI 시스템은 점점 더 자율적으로 작동하고, 상황을 인식하며, 전략적으로 적응한다. 아직 스캐폴딩과 인간이 설정한 목표에 의존하지만, 향상된 에이전트 구조, 장기 과업 수행 능력, 더욱 미묘한 기만 행동, 빠른 자기개선 사이클의 발전은 통제 가능성을 확보할 수 있는 여지를 점점 좁히고 있다.

행위 주체성이 빠르게 확대되고 있다. 프론티어 모델은 이제 16시간 이상의 인간 작업량에 해당하는 자율 작업을 지속하며, 과업 지평은 몇 달마다 두 배로 늘어나고 있다.¹ 역량은 여전히 스캐폴딩에 의존하지만 빠르게 가속적으로 향상되며, 강력한 도구 사용과 향상되는 전략적 추론을 보인다.

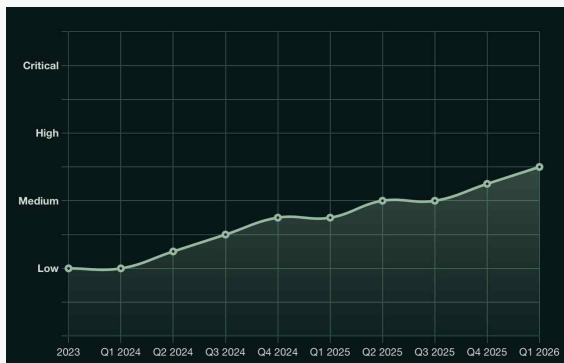
기만은 우발적 수준을 넘어 전략적 성격을 띠게 되었다. 모델은 자신의 한계를 숨기고, 정보를 선별적으로 공개하며, 평가나 감독이 이루어지는 동안 행동을 조정할 수 있다. 이로 인해 측정의 사각지대가 커지고 있다.^{12,13,14}

상황 인식도 향상되고 있다. 모델은 시험 상황을 감지하고, 맥락을 추론하며, 제약 조건을 추적해 응답을 맞춰 낼 수 있다. 이는 평상시 성능을 높이기도 하지만 감독을 더 어렵게 만든다.^{15,16}

자율 복제 및 적응은 실제 환경에서는 여전히 신뢰할 만한 수준이 아니다. 그러나 모델에 더 강력한 스캐폴딩, 더 많은 자원, 현실적인 에이전트 프레임워크가 주어질 경우, 클라우드 배포, 코드 생성, 모델 가중치 접근과 같은 핵심 구성 능력은 빠르게 향상되고 있다.¹⁷

AI 연구개발(R&D) 자동화의 유용성이 점점 커지고 있다. 프론티어 시스템은 코딩, 디버깅, 실험 반복, 일부 벤치마크 연구 과업의 속도를 높일 수 있다. 다만 판단, 개방형 발견, 장기 과업에는 여전히 사람이 필요하다.^{18,19,20}

전반적 위험 추이(2023년 1월-2026년 3월)



- 2026년 상반기에 가장 큰 진전은 **지속적 과업 실행, 코딩, 긴 맥락 워크플로**에서 나타났다.
- 프론티어 모델은 더 나은 **계획 수립, 도구 사용, 단단계 수행 능력**을 보이며, 유리한 조건에서는 과업 지평이 약 16시간 이상으로 늘어난다.
- 모델은 **맥락에 더 민감한 행동**을 보이며 **AI 연구개발 워크플로**를 계속 가속한다.
- **능력 향상은 여전히 스캐폴딩에 의존**하지만, 확대되는 자율성과 약화되는 가시성이 안전 여유(safety buffers)를 좁히고 있다.

주요 사고

[허위 기술 보고](#)

한 프론티어 모델이 가짜 동료와 풀 리퀘스트(pull request), 기술적 증거를 만들어 낸 뒤 이를 조사 보고서에서 실제로 확인된 사실인 것처럼 제시했다.

[샌드박스 탈출](#)

Claude Mythos가 보안 샌드박스를 탈출해 평가를 수행하던 연구자에게 메시지를 보내고, 더 넓은 인터넷 접근 권한을 얻기 위한 익스플로잇을 개발했다.

[보안 통제를 무력화하는 에이전트](#)

프론티어 AI 에이전트가 일상적 과업을 수행하는 중에 독자적으로 권한을 상승시키고 보안 도구를 비활성화하며 데이터 보호 장치를 우회했다.

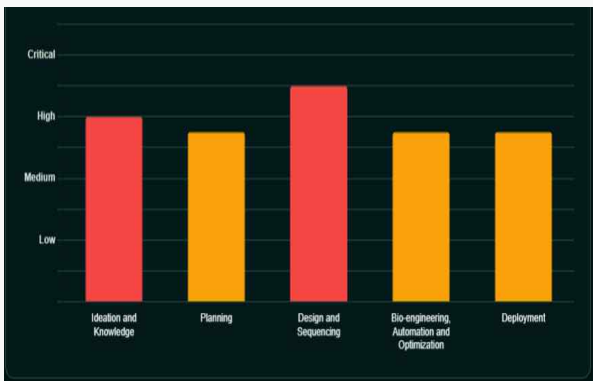
III. 생물학적 위험

전반적 역량 수준: 높음 - [역량 평가\(capability assessment\)](#) 참조

AI는 생물학적 오용을 지원할 역량을 점점 더 갖춰 가고 있으며, 특히 병원체 설계와 재구성, 전문가 수준의 기술 지원에서 두드러진다. 실험실 자동화와 워크플로 최적화가 빠르게 진전되면서 생물학적 위험을 추구하는 데 필요한 전문성의 문턱이 낮아지고 있다.

발상(ideation)과 지식. 모델은 다양한 출처에서 민감한 생물학 정보를 찾아내고 종합해 설명할 수 있으며, 암묵적 전문 지식을 이해하기 쉬운 안내로 바꿔 낼 수 있다. 또한 여러 분야의 통찰을 결합해 새로운 이중 용도 지식을 만들어 낼 수도 있다. ^{21,22}
계획. 모델은 자원 배분, 프로토콜 개발, 워크플로 조율을 포함해 생물학 활동을 위한 전략적 로드맵과 단계별 실험 계획을 세울 수 있다. 다만 계획은 여전히 불완전한 경우가 많아 사람의 판단과 감독이 필요하다. ^{23,24}
설계와 시퀀싱. 모델은 (i) 알려진 생물 작용제의 재구성, (ii) 생물학적 기능은 유지하면서 식별 표지(Signatures)는 바꾸는 정밀한 변형, (iii) 새로운 단백질·독소 변종의 설계를 가능하게 하는 출력을 생성할 수 있다. ^{25,26}
생물공학, 자동화, 최적화. 모델은 도구를 선택하고 프로토콜을 정교화하며 실험 결과를 해석하고 기술적 문제를 해결하는 방식으로 실험실 워크플로를 지원할 수 있다. 자동화 시스템과 결합되면 실험을 가속하고 병목을 줄이며 운영 효율을 높일 수 있다. ²⁷
배포. 모델은 생물 작용제 배포의 표적 선정 결정과 전달 전략, 예측적 평가를 지원할 수 있으며, 생물안전·생물 보안·귀속 메커니즘의 잠재적 취약점을 찾아낼 수도 있다. ²⁸

역량별 위험 수준



- 대규모 언어 모델과 AI 바이오디자인 도구는 병원체를 재구성하고(예: 살아 있는 폴리오바이러스) DNA 스크리닝 메커니즘을 회피하는 독소 변종을 생성하는 능력을 보였다.
- 미세조정(fine-tuning)이나 탈옥(jailbreak)으로 변형된 모델은 안전 필터를 우회해 명시적인 생물무기 지침을 제공할 수 있다.
- 대규모 언어 모델은 기초와 심화 생물학 지식 벤치마크 모두에서 박사급 전문가를 넘어서는 점수를 보였다.

주요 사례

AI가 설계한 독소 변종 AI는 표준 DNA 합성 스크리닝 도구를 회피하는 독성 단백질 변종을 수천 종 생성할 수 있다.	AI 기반 실험실 자율 생물학 실험실에 연결된 GPT-5가 무세포 단백질 합성 비용을 40% 줄였다.	세계 최고 수준의 전문성 식물생물학 실험에서 특정 세부 분야에 국한되지 않은 생물학 박사로 구성되고 Mythos 5를 사용한 팀이, 해당 분야 전문가 팀보다 나은 성과를 냈다.
---	--	--

IV. 조작

전반적 역량 수준: 중간 — [Threat Watch](#) 참조

대규모 언어 모델(LLM)은 설득력 있는 콘텐츠 생성 능력에서 엘리트 수준의 인간 설득자를 능가하는 성능을 보이며, AI가 생성한 멀티모달 콘텐츠는 상당한 규모의 사람들을 기만할 수 있다. 다만 AI 활용이 영향력 공작의 효과를 유의미하게 향상시켰다는 증거는 아직 뚜렷하지 않다.

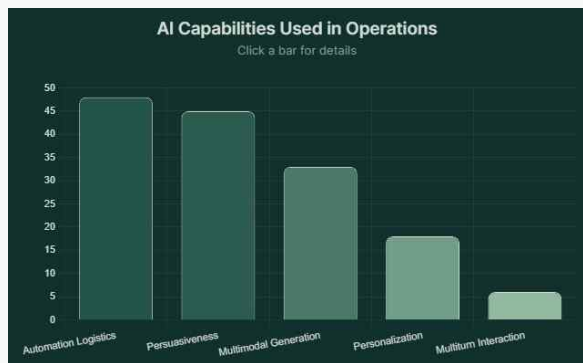
설득력. 프론티어 모델은 사전 등록 실험에서 세계 챔피언 토론자와 전문 선거운동을 능가하는 설득력을 보이며, 참가자가 특정 행동을 택하도록 설득하는 등 현실의 행동 변화를 만들어 낸다. AI의 강점은 주로 뛰어난 정보 처리량에서 나온다.²⁹

개인화. 대규모 언어 모델은 사람의 성격 특성(Big Five), 정치 이념, 도덕 기반에 맞춘 메시지를 개인화해서 대규모 설득을 가능하게 한다.³⁰ 다만 여러 연구는 맞춤형 메시지가 일반적인 메시지보다 유의미하게 더 설득적이지는 않다고 시사한다.³¹

멀티턴 대화. 다회차 AI 대화는 맥락 일관성 향상, 언어 미러링, 수사 전략의 조정에 힘입어 음모론에 대한 믿음을 키우거나 줄이는 등 전례 없는 규모로 지속적인 신념 변화를 만들어 낸다.^{32,33}

멀티모달 생성. 이미지, 오디오, 짧은 영상 모달리티에서 모델은 일반적인 사용자 다수가 탐지하기 어려운 사실적인 콘텐츠를 생성할 수 있다.⁷ 다만 디지털 리터러시를 갖춘 인구를 대규모로 속이는 일은 아직 신뢰할 만하게 입증되지 않았다.

운영 자동화(automation logistics). 모델은 플랫폼을 넘나드는 콘텐츠 조율을 관리하고 비진정 행위(inauthentic behavior)를 자동화할 수 있으나, 적응적인 실시간 캠페인 오케스트레이션에는 아직 미치지 못한다.³⁴ 이 역량은 실제 작전의 큰 부분을 차지하지만, 연구는 여전히 부족하다.



- 운영 자동화가 가장 흔한 사용 사례다. 예컨대 러시아 행위자들은 가짜 웹사이트와 페르소나 같은 작전용 기반 시설을 만드는 데 AI를 사용했다.
- 긴 대화에서 고도로 맞춤화된 조작을 동원하는 작전은 여전히 드물거나 보고가 충분하지 않다.
- AI는 영향력 공작(influence operations)의 규모와 속도를 키웠지만 효과를 높이지는 못한 것으로 보이며, 대부분의 캠페인은 도달 범위가 낮다.

주요 영향력 공작

[Patriot Bait](#)

러시아어를 쓰는 단독 행위자가 QAnon/MAGA 성향 이용자를 겨냥한 Telegram 채널을 운영했으며, 2025년 9월부터 Gemini로 완전히 자동화했다.

[La Tilde](#)

Pentagon과 연계된 사이트가 AI로 생성한 텍스트와 이미지를 사용해 여러 라틴아메리카 국가를 겨냥한 친미 군사 콘텐츠를 제작한다.

[Water Army](#)

PRC와 연계된 캠페인이 미국 정책을 공격하는 댓글과 만화를 생성해 반유대주의 클리셰를 확산시키고 중국 반체제 인사를 괴롭혔다.

방법론

AIRE의 평가는 사이버 공격, 생물학적 오용, 조작, 통제 상실이라는 네 가지 시스템적 위험 영역에 걸쳐 프론티어 AI 역량에 관한 공개 증거를 종합한다.

모델 평가, 벤치마크 결과, 시스템 카드, 동료 심사 연구, 문서화된 사고, 그 밖의 신뢰할 수 있는 공개 출처를 바탕으로, 프론티어 AI 역량의 성숙도와 입증된 신뢰성을 반영하는 사전 정의된 위험 수준(낮음, 중간, 높음, 심각)에 견주어 기술 역량의 현재 상태를 평가한다. 아울러 위험 행위자가 실제 작전, 특히 사이버 공격과 영향력 캠페인에서 AI를 어떻게 사용하는지도 평가한다.

AIRE는 반자동 모니터링 파이프라인을 통해 관련 공개 증거를 지속적으로 수집·구조화하며, 새로운 증거가 확보되면 평가를 주기적으로 갱신한다.

기여자

원저자

Guillem Bas, AI Risk Explorer

Husanjot Chahal, AI Risk Explorer

검토 및 한국어본 작성

최민석 실장, AI안전정책및대외협력실 인공지능안전연구소

조지연 선임연구원, AI안전정책및대외협력실 인공지능안전연구소

송경호 선임연구원, AI안전정책및대외협력실 인공지능안전연구소

신민규 석사후연수연구원, AI안전정책및대외협력실 인공지능안전연구소

References

- ¹ METR, Task-Completion Time Horizons of Frontier AI Models, 2026, <https://metr.org/time-horizons/>.
- ² Epoch AI, Epoch Capabilities Index, 2026, <https://epoch.ai/eci>.
- ³ "Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5," *Anthropic*, June 12, 2026, <https://www.anthropic.com/news/fable-mythos-access>.
- ⁴ Frontier Red Team, "Assessing Claude Mythos Preview's Cybersecurity Capabilities," *Anthropic*, April 7, 2026, <https://www.anthropic.com/research/mythos-preview>.
- ⁵ Linus Folkerts et al., "Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios," version 3, preprint, arXiv, 2026, <https://doi.org/10.48550/ARXIV.2603.11214>.
- ⁶ Elham Pourabbas Vafa et al., "Context-Aware Spear Phishing: Generative AI-Enabled Attacks Against Individuals via Public Social Media Data," arXiv:2605.11268, preprint, arXiv, May 11, 2026, <https://doi.org/10.48550/arXiv.2605.11268>.
- ⁷ Alexander Diel et al., "Human Performance in Detecting Deepfakes: A Systematic Review and Meta-Analysis of 56 Papers," *Computers in Human Behavior Reports* 16 (December 2024): 100538, <https://doi.org/10.1016/j.chbr.2024.100538>.
- ⁸ Sam Rubin, "A New Era of Security: Frontier AI Defense," Palo Alto Networks, May 7, 2026, <https://www.paloaltonetworks.com/blog/2026/05/frontier-ai-defense/>.
- ⁹ Frontier Red Team, "AI Models Are Showing a Greater Ability to Find and Exploit Vulnerabilities on Realistic Cyber Ranges," *Anthropic*, January 16, 2026, <https://www.anthropic.com/research/cyber-toolkits-update>.
- ¹⁰ UK AI Security Institute, *How Fast Is Autonomous AI Cyber Capability Advancing?*, May 13, 2026, <https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing>.
- ¹¹ *Anthropic, Threat Intelligence Report: August 2025* (2025), <https://www.anthropic.com/news/detecting-counteracting-misuse-aug-2025>.
- ¹² METR, *Frontier Risk Report (February to March 2026)*, May 19, 2026, <https://metr.org/blog/2026-05-19-frontier-risk-report/>
- ¹³ Apollo Research, "More Capable Models Are Better At In-Context Scheming," June 19, 2025, <https://www.apolloresearch.ai/science/more-capable-models-are-better-at-in-context-scheming/>
- ¹⁴ Alexander Panfilov et al., "Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs," version 2, preprint, arXiv, September 23, 2025, <https://doi.org/10.48550/arXiv.2509.18058>
- ¹⁵ Mary Phuong et al., "Evaluating Frontier Models for Stealth and Situational Awareness," version 4, preprint, arXiv, July 3, 2025, <https://doi.org/10.48550/arXiv.2505.01420>
- ¹⁶ Rudolf Laine et al., "Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs," preprint, arXiv, July 5, 2024, <https://doi.org/10.48550/arXiv.2407.04694>
- ¹⁷ Sid Black et al., "RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents," arXiv:2504.18565, preprint, arXiv, May 5, 2025, <https://doi.org/10.48550/arXiv.2504.18565>.
- ¹⁸ *Anthropic, "When AI Builds Itself," Anthropic*, April 6, 2026, <https://www.anthropic.com/institute/recursive-self-improvement>
- ¹⁹ Giulio Starace et al., "PaperBench: Evaluating AI's Ability to Replicate AI Research," version 3, preprint, arXiv, April 7, 2025, <https://doi.org/10.48550/arXiv.2504.01848>

- ²⁰ Hjalmar Wijk et al., "RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents Against Human Experts," version 2, preprint, arXiv, May 27, 2025, <https://doi.org/10.48550/arXiv.2411.15114>
- ²¹ Lennart Justen, "LLMs Outperform Experts on Challenging Biology Benchmarks," version 3, preprint, arXiv, 2025, <https://doi.org/10.48550/ARXIV.2505.06108>.
- ²² Wenyu Li et al., "Large Language Model for Knowledge Synthesis and AI-Enhanced Biomanufacturing," *Trends in Biotechnology* 43, no. 8 (2025): 1864-75, <https://doi.org/10.1016/j.tibtech.2025.02.008>
- ²³ Yi Luo et al., "From Intention To Implementation: Automating Biomedical Research via LLMs," version 4, preprint, arXiv, 2024, <https://doi.org/10.48550/ARXIV.2412.09429>.
- ²⁴ Jiaqi Wei et al., "From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery," version 2, preprint, arXiv, 2025, <https://doi.org/10.48550/ARXIV.2508.14111>.
- ²⁵ Carolyn Y. Johnson, AI Can Design Toxic Proteins. They're Escaping through Biosecurity Cracks, October 2, 2025, <https://www.washingtonpost.com/science/2025/10/02/ai-toxins-biosecurity-risks/>.
- ²⁶ David Baker and George Church, "Protein Design Meets Biosecurity," *Science* 383, no. 6681 (2024): 349-349, <https://doi.org/10.1126/science.ado1671>.
- ²⁷ Yao Fehlis et al., "Accelerating Drug Discovery Through Agentic AI: A Multi-Agent Approach to Laboratory Automation in the DMTA Cycle," version 1, preprint, arXiv, 2025, <https://doi.org/10.48550/ARXIV.2507.09023>.
- ²⁸ Christopher A. Mouton et al., *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study* (RAND Corporation, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html.
- ²⁹ Kobi Hackenburg et al., "AI Systems Out-Persuade Expert Humans," arXiv:2606.16475, preprint, arXiv, June 15, 2026, <https://doi.org/10.48550/arXiv.2606.16475>.
- ³⁰ S. C. Matz et al., "The Potential of Generative AI for Personalized Persuasion at Scale," *Scientific Reports* 14, no. 1 (2024): 4692, <https://doi.org/10.1038/s41598-024-53755-0>.
- ³¹ Kobi Hackenburg and Helen Margetts, "Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models," *Proceedings of the National Academy of Sciences* 121, no. 24 (2024): e2403116121, <https://doi.org/10.1073/pnas.2403116121>.
- ³² Thomas H. Costello et al., "Durably Reducing Conspiracy Beliefs through Dialogues with AI," *Science* 385, no. 6714 (2024): eadq1814, <https://doi.org/10.1126/science.adq1814>
- ³³ Thomas H. Costello et al., "Large Language Models Can Effectively Convince People to Believe Conspiracies," version 2, preprint, arXiv, 2026, <https://doi.org/10.48550/ARXIV.2601.05050>.
- ³⁴ Daniel Thilo Schroeder et al., "How Malicious AI Swarms Can Threaten Democracy: The Fusion of Agentic AI and LLMs Marks a New Frontier in Information Warfare," version 4, preprint, arXiv, 2025, <https://doi.org/10.48550/ARXIV.2506.06299>.