Đánh giá an toàn mô hình AI toàn cầu

intl.khoa-hoc-cong-nghe.tri-tue-nhan-tao.ai-safety-evaluations

Version 2026.10.30 · Cập nhật 2026-10-30 · Chính sách cập nhật monthly

ĐIỂM NỔI BẬT

ĐIỂM CAO NHẤT TRONG AI SAFE...

2.66

điểm chỉ số

Anthropic, Summer 2026

DOANH NGHIỆP ĐƯỢC AI SAFETY...

9

doanh nghiệp

Summer 2026

TỶ LỆ HOÀN THÀNH NHIỆM VỤ A...

50

%

Trung bình trong kết quả AISI; báo cáo xu h...

DỮ LIỆU

GPT-Rosalind-6 System Card Addendum

Địa bàn: Toàn cầu (công bố đánh giá mô hình) · Phân loại: Đánh giá an toàn và alignment mô hình

Anthropic — AI Safety Index Summer 2026

Địa bàn: Mẫu doanh nghiệp toàn cầu · Phân loại: Đánh giá thực hành an toàn doanh nghiệp AI

Frontier AI Trends Report

Địa bàn: Mẫu mô hình tiên tiến toàn cầu; đánh giá của AISI Vương quốc Anh · Phân loại: Đánh giá xu hướng năng lực và biện pháp bảo vệ

International AI Safety Report 2026 — đánh giá an toàn và khoảng trống đo lường

Địa bàn: Toàn cầu · Phân loại: Tổng hợp bằng chứng về đánh giá và quản lý rủi ro AI

NGUỒN & ĐỘ MỚI

Phiên bản: 2026.10.30
Cập nhật: 2026-10-30
Giấy phép: Không ghi nhận

MLCommons — Benchmark for general-purpose AI chat model — English v1.0 Official
https://ailuminate.mlcommons.org/benchmarks/general_purpose_ai_chat/1.0-en_us-official-ensemble

MLCommons — AILuminate Safety
https://mlcommons.org/ailuminate/safety/

MLCommons — MLCommons Launches AILuminate, First-of-its-Kind Benchmark to Measure the Safety of Large Language Models
https://mlcommons.org/2024/12/mlcommons-ailuminate-v1-0-release/

OpenAI — GPT-6 Astra System Card
https://deploymentsafety.openai.com/gpt-6-astra/protocolqa-open-ended

Anthropic — Claude Fable 5.1 and Claude Mythos 5.1 System Card
https://www.anthropic.com/claude-fable-5-1-mythos-5-1-system-card

Google DeepMind — Evaluating frontier models for stealth and situational awareness
https://deepmind.google/research/publications/157938/

MLCommons — AILuminate and The First Double-Blind Reliability Evaluation of a Proprietary AI Model
https://mlcommons.org/2026/08/double-blind-reliability-evaluation/

MLCommons — AILuminate Benchmark
https://mlcommons.org/benchmarks/ailuminate/

International AI Safety Report — International AI Safety Report 2026
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

Future of Life Institute — AI Safety Index — Summer 2026 report
https://futureoflife.org/wp-content/uploads/2026/07/AI-Safety-Index-Summer-2026-Digital-14-Jul.pdf

UK AI Security Institute — Frontier AI Trends Report
https://www.aisi.gov.uk/frontier-ai-trends-report

Axios — AI companies retreat from safety pledges even as capabilities grow
https://www.axios.com/2026/07/07/report-ai-safety-pledges

UK AI Security Institute — GPT-6 Astra performs unsanctioned supply-chain attacks in simulations
https://www.aisi.gov.uk/blog/gpt-6-astra-performs-unsanctioned-supply-chain-attacks-in-simulations

NIST / CAISI — CAISI’s Assessment of Z.ai’s GLM-5.3 Cyber Capabilities
https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities

UK AI Security Institute và U.S. Center for AI Standards and Innovation — UK AISI / CAISI Preliminary Assessment of Kimi K3's Cyber Capabilities
https://www.aisi.gov.uk/blog/preliminary-assessment-of-kimi-k3s-cyber-capabilities

UK AI Security Institute — Measuring AI Agents’ Progress on Multi-Step Cyber Attack Scenarios
https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios

Japan AI Safety Institute — AI□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ (□ 1.20□)□ □ □
https://aisi.go.jp/output/output_information/260707/

OpenAI — GPT-Rosalind-6 System Card Addendum
https://deploymentsafety.openai.com/gpt-rosalind-6

Future of Life Institute — AI Safety Index — Summer 2026
https://futureoflife.org/ai-safety-index-summer-2026/