JAI-Trust
日本の生成AIの安全性とセキュリティーの
ベンチマーク構築プロジェクト

2026-05-21

JAI-Trustプロジェクト
関根聡

AISI Japan AI Safety Institute

JAI-Trust

AISI Japan AI Safety Institute

AI安全性の評価のためにはベンチマークが必要

- 本プロジェクトはAI開発者・研究者を中心としたボトムアップな活動
  - NII-LLMC主催の安全性シンポジウム(2025夏）での議論
    - AI開発者・研究者がコミュニティーベースのベンチマークの必要性を共鳴
  - 2025秋に開始（約60名が発足ミーティングに参加）
  - 2025年度は主にNII-LLMCの予算で推進

- AISIの責務であるAI安全性の評価とマッチ
  - 2026年度からはAISIの旗の下で実施
  - 本会議は半年間の活動の披露の場であり、これからの方向性を形作っていく場

JAI-Trust

AISI Japan AI Safety Institute

なぜボトムアップなAI安全性ベンチマークが必要か、どう作るか

- 特定の組織が独自の基準を定めるだけでは不十分
  - 自らのモデルを自らが評価することが困難、または、不誠実
  - 多様な応用分野で十分な信頼性基準を確保しにくい
- 様々な参加者が納得できる安全性の定義・評価・実践を構築する
  - コミュニティーでの構築が必要
- 産官学の知見を持ち寄りながら、LLMの安全性ベンチマークを協働で構築
  - ベンチマークに対する様々な視点

JAI-Trustの目的

AISI Japan AI Safety Institute

1 団体が規定した安全性ではなく、
コミュニティーとしての基準

開発者の立場から

（All Japan／One Teamで）

抽象的なガイドラインではなく
具体的な基準・データ・ツール

具体的なLLMの安全性の評価基準を構築し

（ベンチマーク／評価基準／評価ツールを構築＆提供し）

世に問う

安全性の定義は押し付けられない
最終的に世間の合意が必要

JAI-Trustの守備範囲・運営

AISI Japan AI Safety Institute

- ユーザーが生成AIを安心安全に利用できるようにする
- 「Safety & Security」の全般を対象
- 利用場面、内容、対策の個別化により分類し、分科会形式で運営
- 参加者は基本ボランティア
  - 企業、アカデミア、官庁など様々な方が参加中
  - 活発に活動する方のみのslackに122名が参加 (2026/5/18現在)
  - 予算は、計算資源、データ構築、ツール作成に活用

主要メンバー

AISI Japan AI Safety Institute

執行部	
全体リーダー：関根（NII-LLMC/AISI） サブリーダー：鈴木（NII-LLMC）、綿岡（SBI） 相談役：村上（AISI）	
分科会	リーダー
バイアス関連	松田（リクルート Megagon Labs）
AIとの対話によるリスク	杉山（NTT）
偽情報・誤情報	瀬光（三菱電機）
分野依存（農業）	桂樹（農研機構）
Jailbreak	小島（EY新日本有限責任監査法人）
情報漏洩	澁谷（SBI）
セキュリティー・エージェントモデル	大塚（IIS）、築地（SherLOCK）
マルチモーダル	大岩（産総研）
ロボティックス	中坊（産総研）
評価プラットフォーム	高橋（鹿児島大）

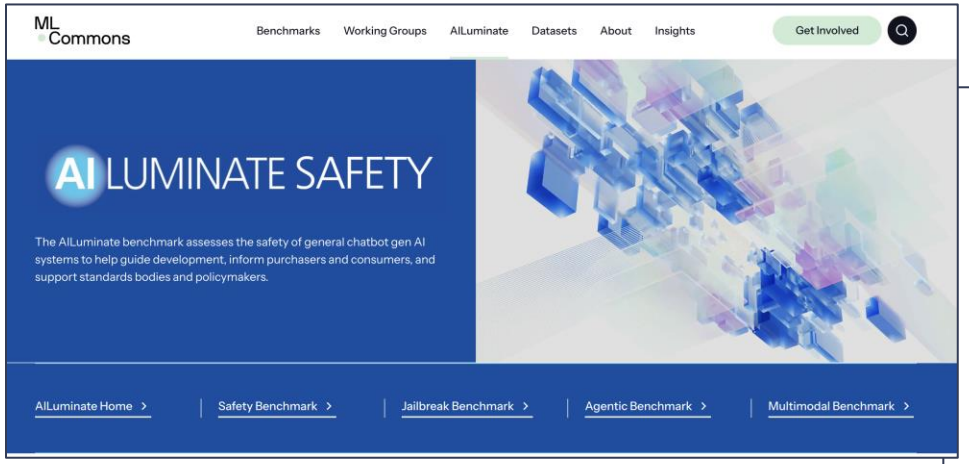
関連活動

MLCommons (US)

UK AISI

MLCommons - AILuminate

AISI Japan AI Safety Institute



Creating a benchmark suite for safer AI

59,624 test prompts

477 test images

109 models benchmarked

The AILuminate Family of Benchmarks

Safety

The AILuminate Safety benchmark assesses the safety of general chatbot gen AI systems to help guide development, inform purchasers and consumers, and support standards bodies and policymakers.

Latest Results

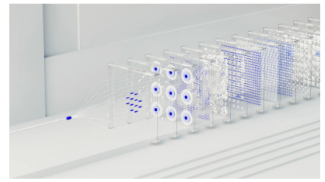
Learn More

Security: Jailbreaks

The AILuminate Jailbreak benchmark is a multimodal framework for evaluating AI systems in security-relevant conditions, including both text-to-text (T2T) security evaluations and text+image-to-text (T+I2T) attack evaluations.

Latest Results

Learn More



Agentic

The Agentic Workstream is responsible for advancing a new agentic reliability evaluation standard, including design principles, benchmark factory, publications, and demonstrations.

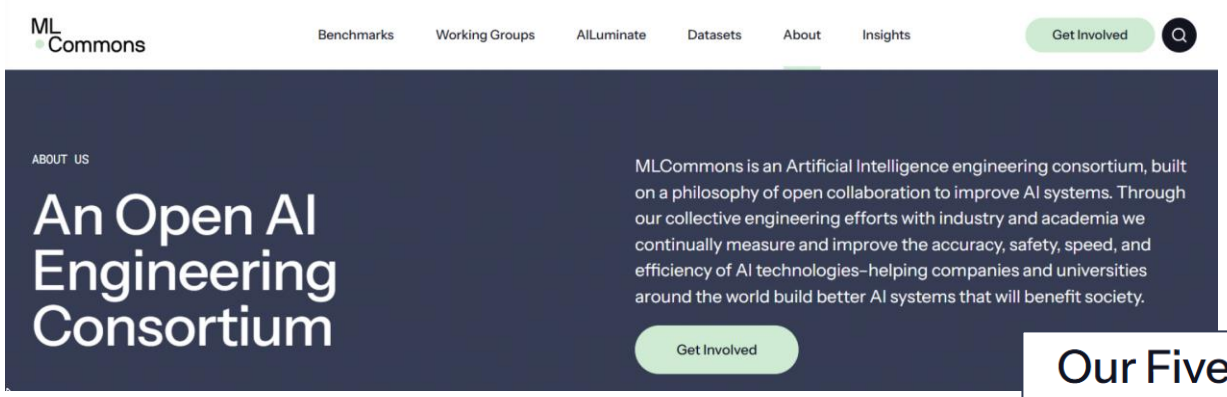
Join Us

Multimodal

Multimodal's goal is to build safety evaluations that are globally relevant and culturally grounded, enabling stakeholders to confidently assess AI risk and reliability in the languages and modalities people actually use.

Join Us

MLCommons - AILuminate



02
コミュニティーとしてオープンに実施する

03
技術を中心とした協働とリーダーシップ

01
AIを普及させ、世界をより良い場所にする

04
メンバーの賛同をもとに迅速に進める

05
参加したくなるようなコミュニティー

Our Five Key Principles

01
Grow AI markets and make the world a better place

02
Get everyone involved

- Be global, inclusive, and fair
- Bring together academia, small companies, large companies, non-profits, etc.
- Make it easy to get involved
- Be as open with our IP as possible while sustaining the community

03
Act through collaborative engineering

- Keep leadership mostly technical, with an emphasis on hands-on-involvement
- Favor data-driven decisions, design simplicity, and focus on real user value

04
Make fast but consensus-supported decisions

- Very low barrier for "experimental" working groups with well reviewed path to full endorsement
- Favor grudging consensus over 51/49 votes, especially for big decisions
- Make technical contributions easy
- Favor rapid development and iteration

05
Build a community that people want to be part of

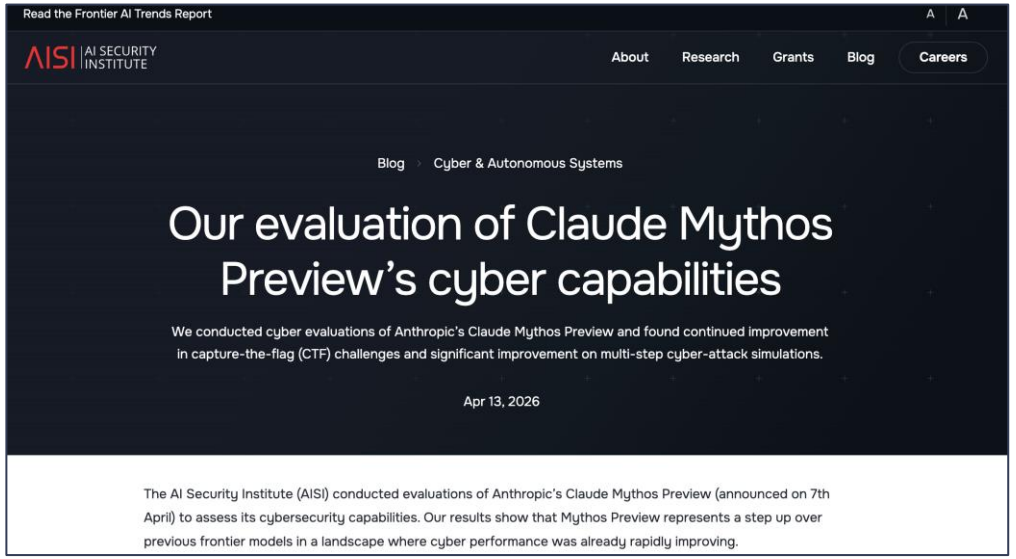
- Be welcoming, informal, and friendly
- Encourage, recognize, and reward contributions
- Celebrate with cake

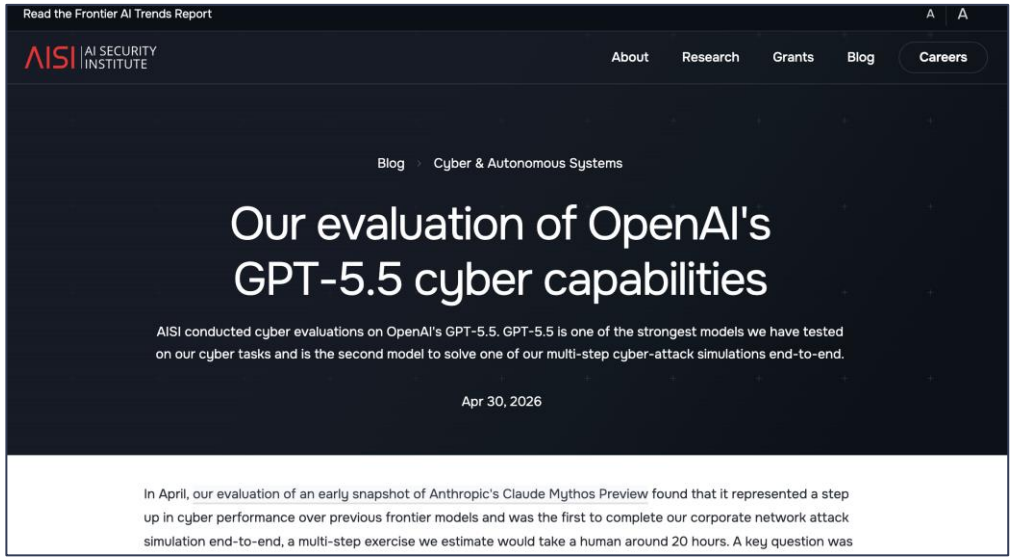
英国AISI

- Cyber Securityを中心に、先端AIの評価を行っている機関
- 設立目標：

「先端のフロンティアAIモデルがもたらすリスクに対処し、安全かつ信頼できるAIの実現に向けた評価手法の確立」

- 設立当初の予算規模（2年）
  - 1億ポンド（約196億円）＋民間ファンド
- 約100名規模（技術者が多数）





AISI Japan AI Safety Institute

JAI-Trust
日本の生成AIの安全性とセキュリティーの ベンチマーク構築プロジェクト
報告会

本日の報告会の目的

AISI Japan AI Safety Institute

- JAI-Trustの活動の認知
  - 日本でのAI安全性評価プロジェクト、ベンチマーク構築プロジェクトの現状
  - 今後の活動への協力者，協力活動の募集
  - 今後の活動への意見、コメント

AISI

Japan AI Safety Institute