

AI 品質を評価・向上させる Citadel AI のご紹介

AI の品質・信頼性に関わる新たな課題を解決

株式会社 Citadel AI

- 2020年12月 設立
- 2021年9月 Seed Funding
- 2023年6月 Series A



 Citadel AI

総額 **5.2 億円** の
Series A 資金調達完了



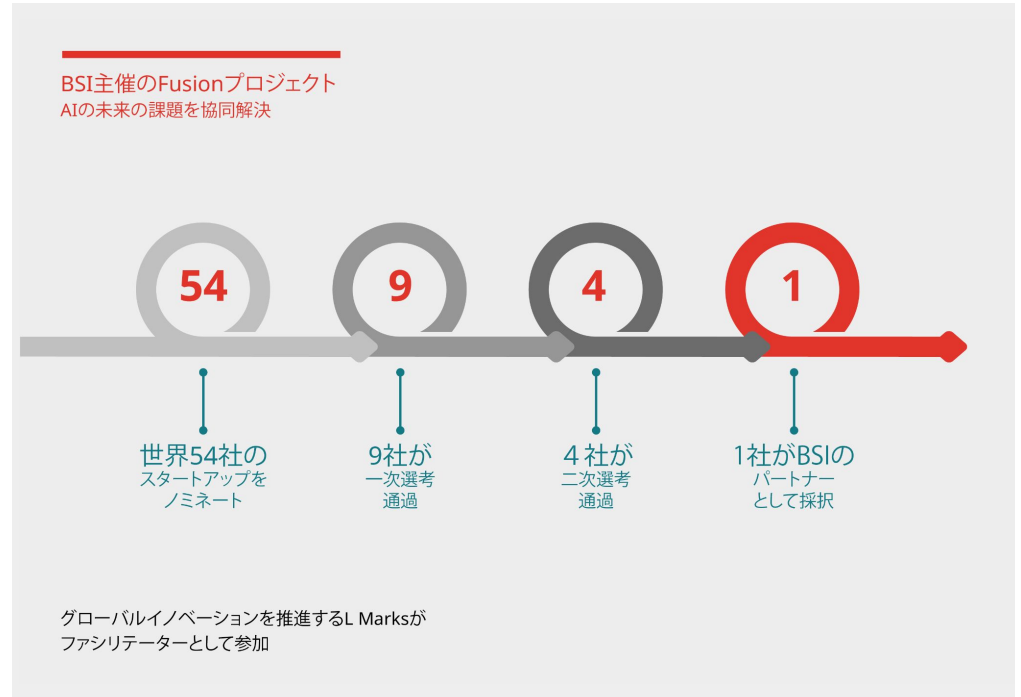
最新版 すごいベンチャー、スタートアップ 100 に選出



世界の AI 国際標準をリードする英国規格協会が AI の認証審査を念頭にグローバルに採用

bsi.

国際的な規格開発と
認証分野で世界をリードする
BSIにて正式採用決定



米国 AI Safety Institute 等 海外公的機関の取組みに参画



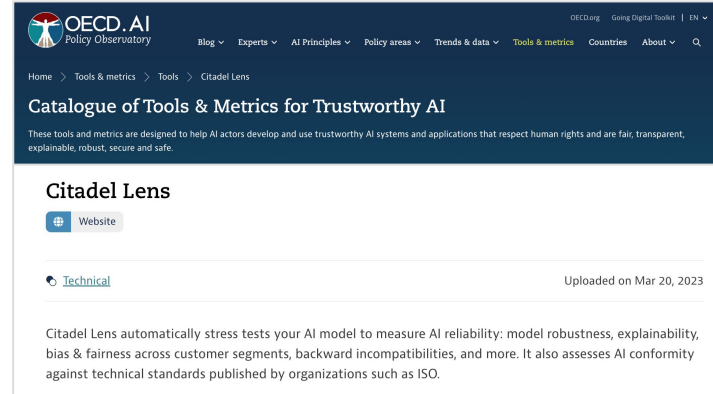
Information Technology /Artificial intelligence

U.S. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE

NEWS

Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety

Consortium includes more than 200 leading AI stakeholders and will support the U.S. AI Safety Institute at the National Institute of Standards and Technology.



OECD.AI Policy Observatory

Home > Tools & metrics > Tools > Citadel Lens

Catalogue of Tools & Metrics for Trustworthy AI

These tools and metrics are designed to help AI actors develop and use trustworthy AI systems and applications that respect human rights and are fair, transparent, explainable, robust, secure and safe.

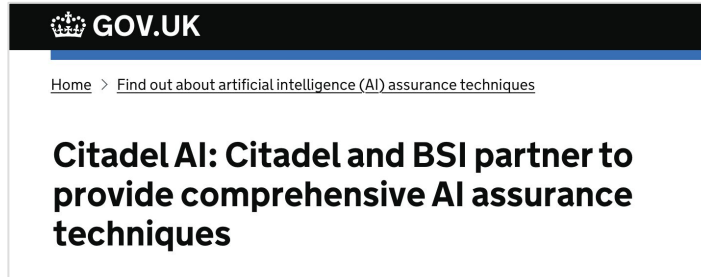
Citadel Lens

Website

Technical

Uploaded on Mar 20, 2023

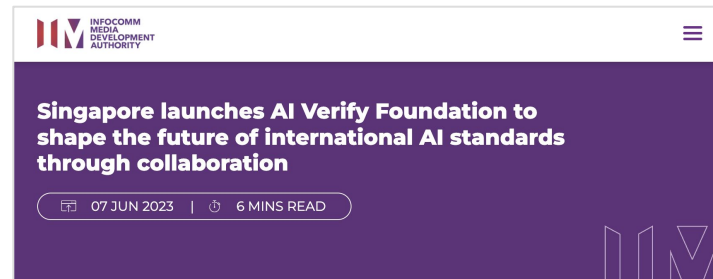
Citadel Lens automatically stress tests your AI model to measure AI reliability: model robustness, explainability, bias & fairness across customer segments, backward incompatibilities, and more. It also assesses AI conformity against technical standards published by organizations such as ISO.



GOV.UK

Home > Find out about artificial intelligence (AI) assurance techniques

Citadel AI: Citadel and BSI partner to provide comprehensive AI assurance techniques



Infocomm Media Development Authority

Singapore launches AI Verify Foundation to shape the future of international AI standards through collaboration

07 JUN 2023 | 6 MINS READ

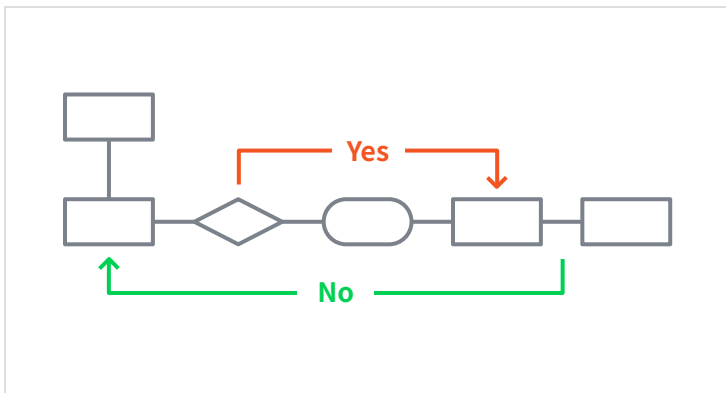
1

Citadel AI が考える AI 品質に関わるリスク

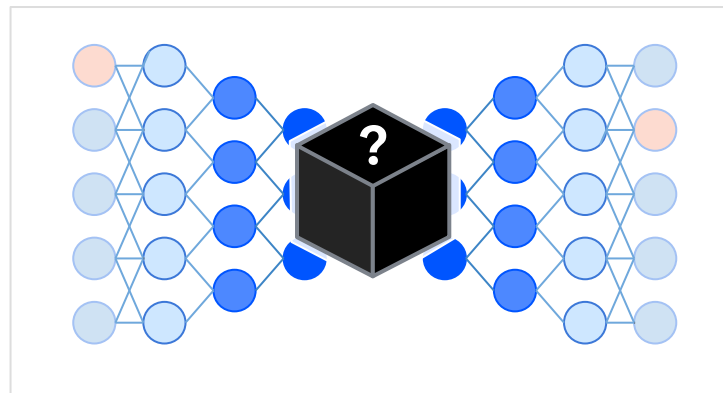
Citadel AI が考える、AI リスクの全体像

1 法的なリスク	各国の AI 法規制に対するコンプライアンス
	個人情報保護、知的財産権
2 社内体制上のリスク	AI ガバナンスの制度設計・体制づくり
	AI 人材の不足
3 技術的なリスク	AI 固有の新たな技術的リスク
	従来のシステムでもあったサイバーセキュリティ

AI はブラックボックス

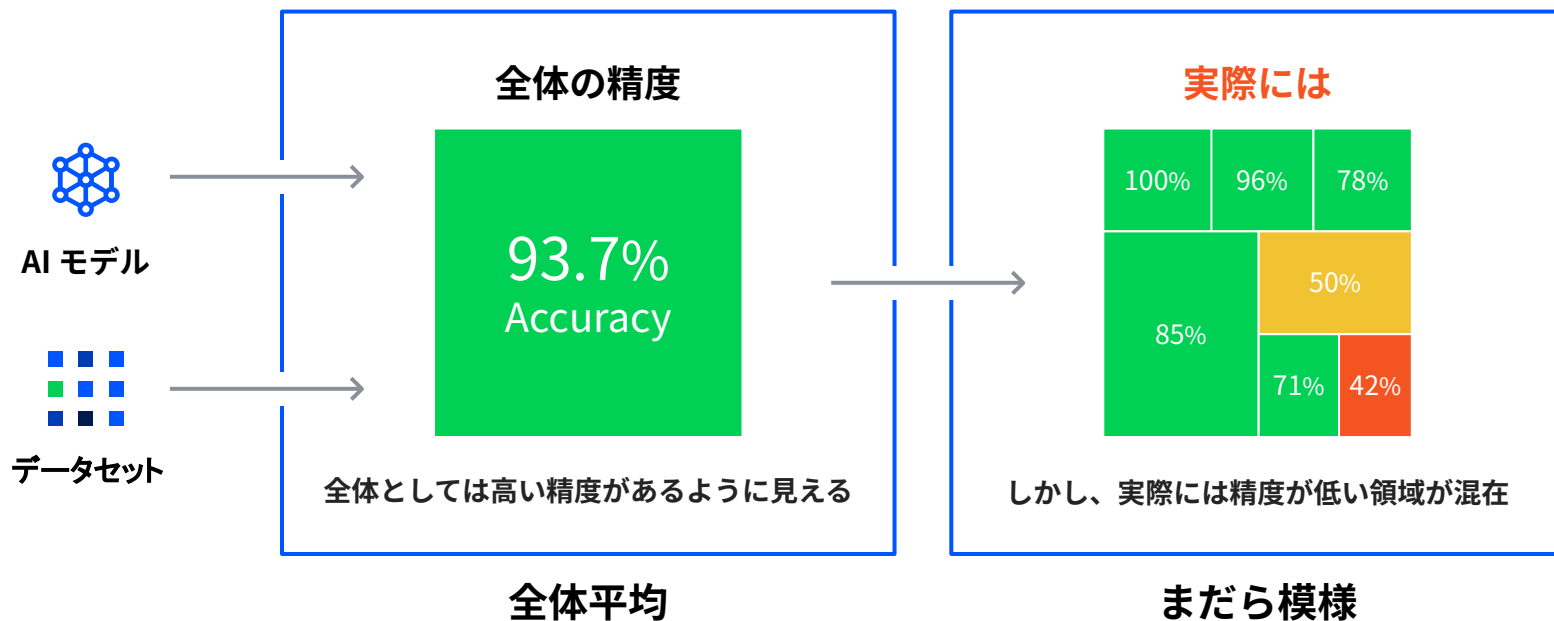


従来のソフトウェアの構造



AI ソフトウェアの構造

学習時 (開発時) の課題：隠れた未学習領域に残るリスク



運用時の課題：AI は環境変化に脆弱でドリフトを起こす



既存顧客



新規顧客



学習時の
撮影環境



運用時の
撮影環境

2

Citadel AI が取り組む

AI の品質管理とリスクマネジメント

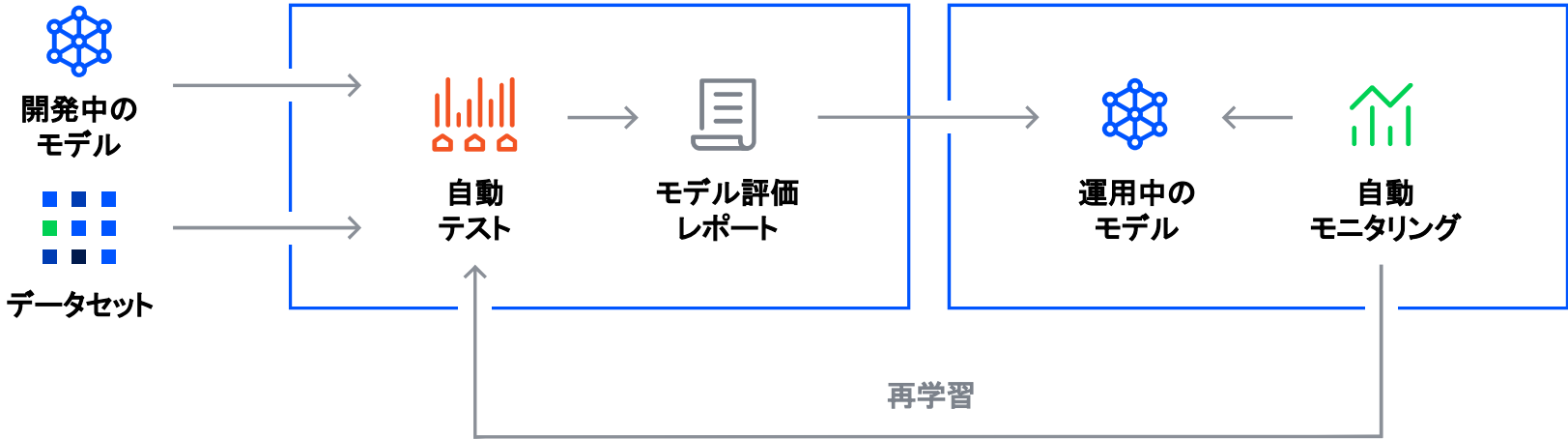
AI ライフサイクル全体の品質・信頼性を向上

1. モデル開発時の自動検証

2. モデル運用時の自動監視

 Citadel Lens

 Citadel Radar



さまざまな AI に汎用的に適用可能

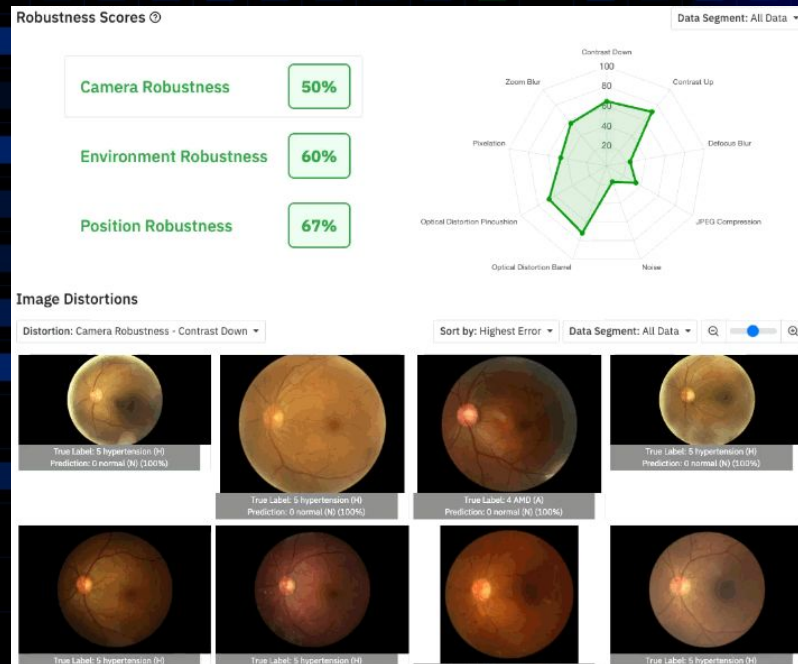
Citadel AI				
✓ 数値データに基づく AI	[電力] 需要予測	[物流業] 出入荷予測	[金融] 与信審査	[マーケティング] 広告効果
✓ 画像データに基づく AI	[医療] 画像診断	[製造業] 不良品検査	[保険] 事故査定	[建設業] 安全管理

 Citadel Lens

AI モデルの 品質を自動テスト

網羅的テストを自動かつ一瞬で

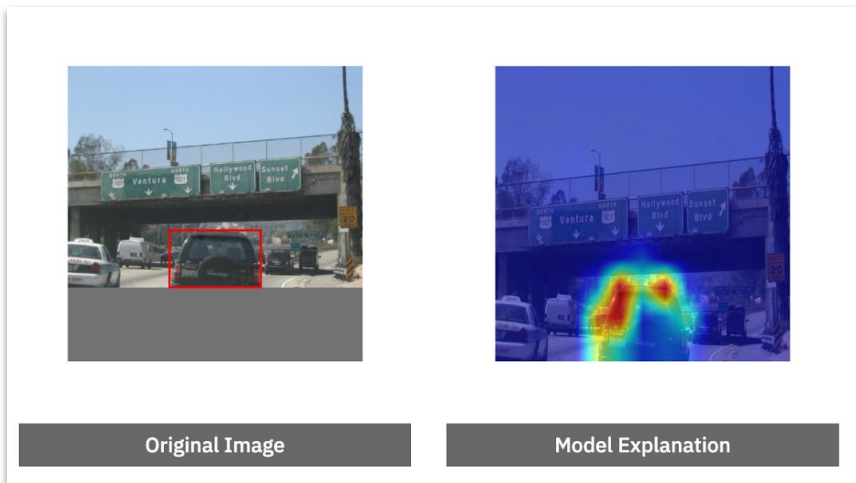
- ノイズ耐性テスト
- 未学習領域の自動検出
- ラベル間違い推定
- 公平性テスト
- 説明責任の可視化 etc.



AI モデルの透明性と信頼性を向上

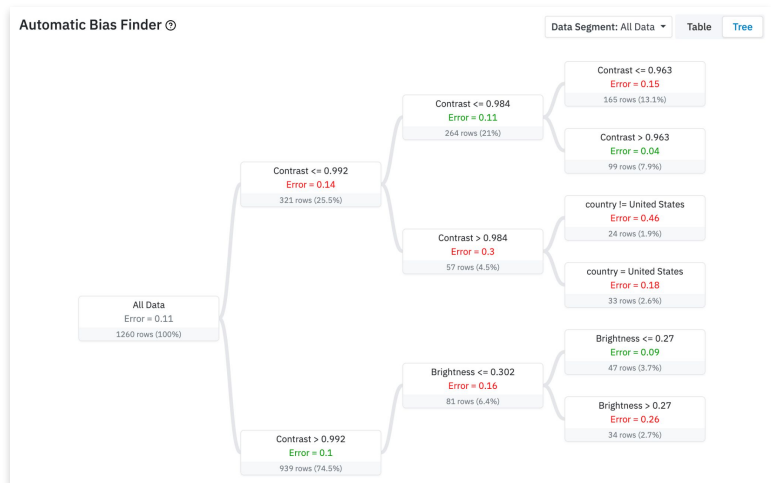
AI モデルの挙動を可視化

AI の透明性と説明性を実現



バイアスの自動検出

未学習領域によるバイアス・脆弱性を検出



各国法制度/ISO標準 適合テスト

法制度や国際標準への適合性検証と
その裏付けとなる技術検証を
同時に実現

ISO Standards About us News **Taking part** Store

← ISO/IEC JTC 1

ISO/IEC JTC 1/SC 42 Artificial intelligence

25 Published ISO standards*	31 ISO standards under development*	39 Participating members
---------------------------------------	--	------------------------------------

Add New Report

Choose an Existing Model ⊙

YOLOX with Hospital A Training Data

Choose an Existing Dataset ⊙

Hospital B Test Dataset [View Dataset](#)

Choose Reports to Generate ⊙

- Citadel Model Report: Citadel AI's interactive report for analyzing model reliability
- ISO/IEC TR 24027: Bias in AI systems and AI aided decision making ⊙
- ISO/IEC TR 24028: Overview of trustworthiness in artificial intelligence ⊙
- ISO/IEC TR 24029-1: Assessment of the robustness of neural networks — Part 1: Overview ⊙
- DigiARC-TR-2022-01: Machine Learning Quality Management Guideline ⊙
- QA4AI: Guidelines for Quality Assurance of AI-based Products and Services ⊙

[Show Advanced Settings](#)

Cancel **Generate Report**

さらに、

「生成 AI (大規模言語モデル)」の
品質・信頼性評価にも対応

LLM アプリケーション品質評価ツール LangCheck

LangCheck 6つの機能 (全て日本語対応)			
 LangCheck 	 ①ファクトチェック	 ②有害テキスト検出	 ③テキスト拡張
	 ④多言語対応 (日本語にネイティブ対応)	 ⑤ローカルで動作	 ⑥評価、保護、監視

LangCheck の特徴

1. 日本語ネイティブ対応ツール
2. 世界最先端のさまざまな生成 AI 品質評価技術をワンパッケージ化
3. ハルシネーション等生成 AI の品質問題を数値化して可視化

ハルシネーションのアラート発出により品質を向上

LangCheck社の福利厚生について教えてください。

Ask Question

アラート発出

ハルシネーションの可能性のある出力です。必ず出典と照らし合わせて再度内容をチェックしてください。

Answer:

LangCheck社の福利厚生には、
 フィナンシャルプランニング支援
 リモートワーク補助
 住宅ローン補助
 のような内容が含まれます。

リスクを数値化

Factual Consistency Score: 0.0607

Source Document

LangCheck社 福利厚生ポリシー

LangCheck社では、福利厚生として以下のようなサポートを社員に提供しています。

- 健康保険・厚生年金の加入
- 退職金制度
- 子育て支援制度
- 外部施設の割引利用
- 社員食堂
- 学術・研究会への参加支援

なお、以下のようなサービスは現在は対象としておりません。今後の対応を予定しております。

- フィナンシャルプランニング支援
- リモートワーク補助
- 住宅ローン補助

Metric

Value

ai_disclaimer_similarity

0.0384

factual_consistency

0.0607

factual_consistency_openai

0.5

language

ja

request_fluency

0.8755

request_fluency_openai

1

request_readability

82.6733

request_sentiment

0.5322

もっと話を聞いてみたい方は、是非ご連絡ください

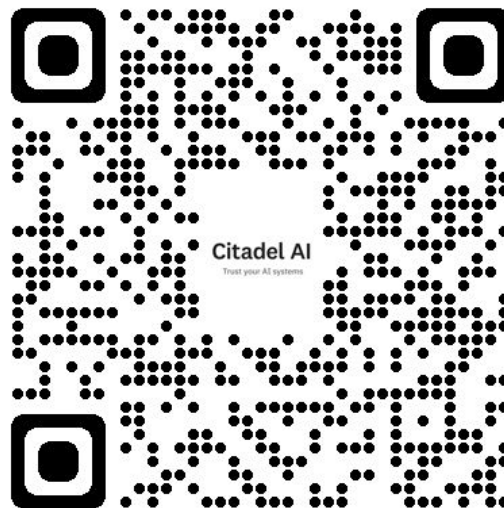
株式会社Citadel AI

お問い合わせ:

info@citadel.co.jp

企業URL:

<https://citadel.co.jp>



Citadel AI