Citadel AI

生成 AI の品質マネジメント

2024年7月31日

株式会社 Citadel AI
松葉 威人, COO

AI の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

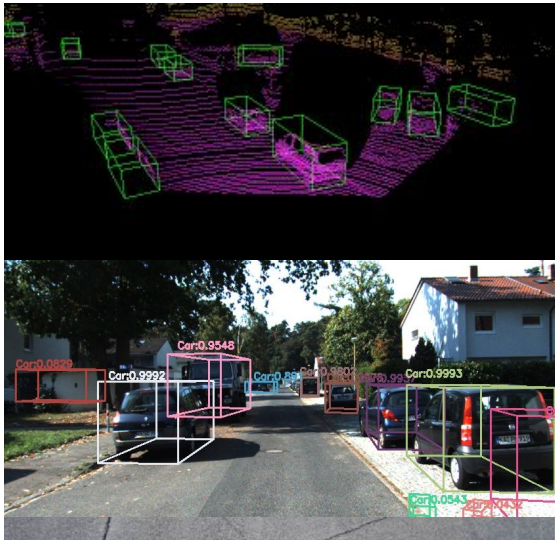
Series A ラウンドで
5.2 億円調達

UTokyoIPC　CORAL　anri
SUNTORY　MUFG Mitsubishi UFJ Capital

Google　WAYMO　paidy　TOYOTA　stripe

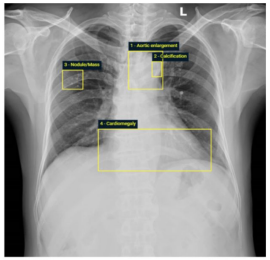
ミスが許されないハイリスク　AI システムの品質検証

自動車・製造業

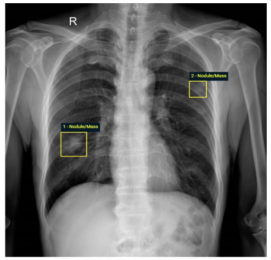


銀行・保険

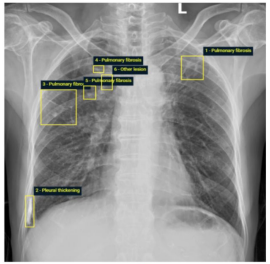
医療・ヘルスケア



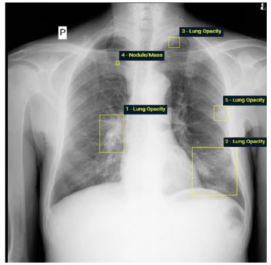
Tuberculosis



Lung tumor



Tuberculosis

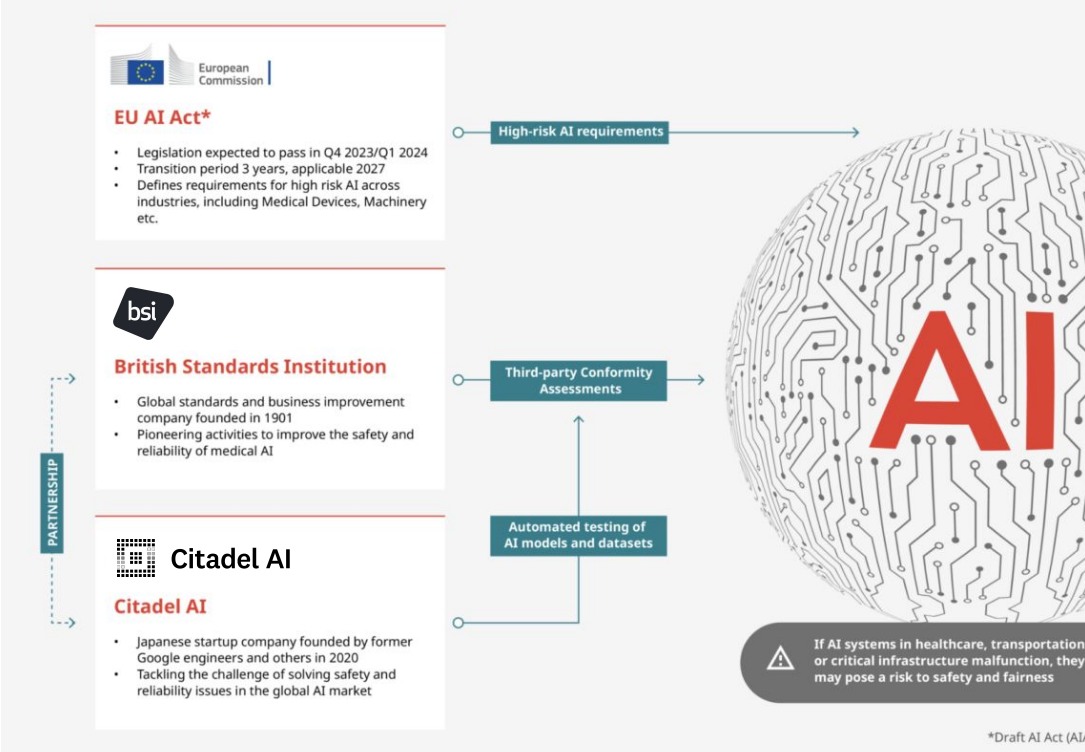


Pneumonia , Lung tumor

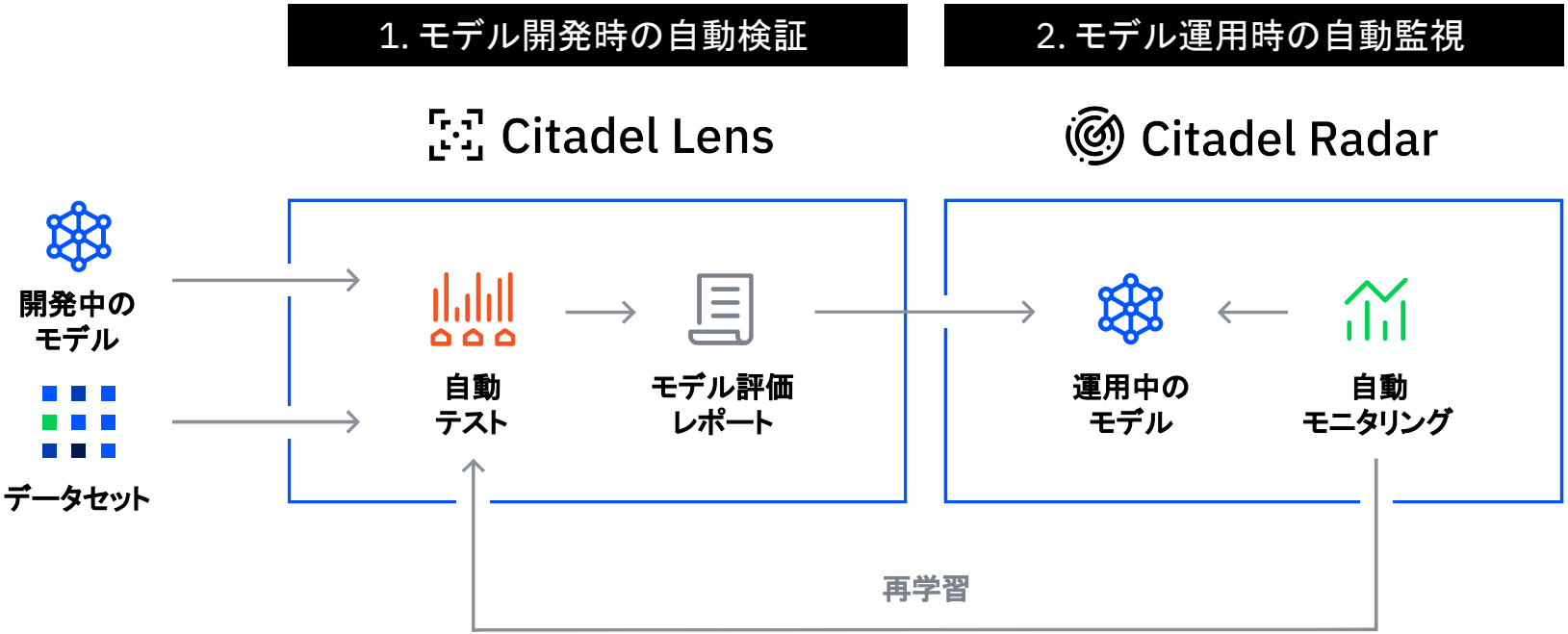
世界の AI 国際標準をリードする BSI 英国規格協会がハイリスク AI の認証審査を念頭にグローバルに採用

国際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

BSI は今後
AI 認証機関 となるべく準備中



AI ライフサイクル全体の信頼性・品質を向上するツール



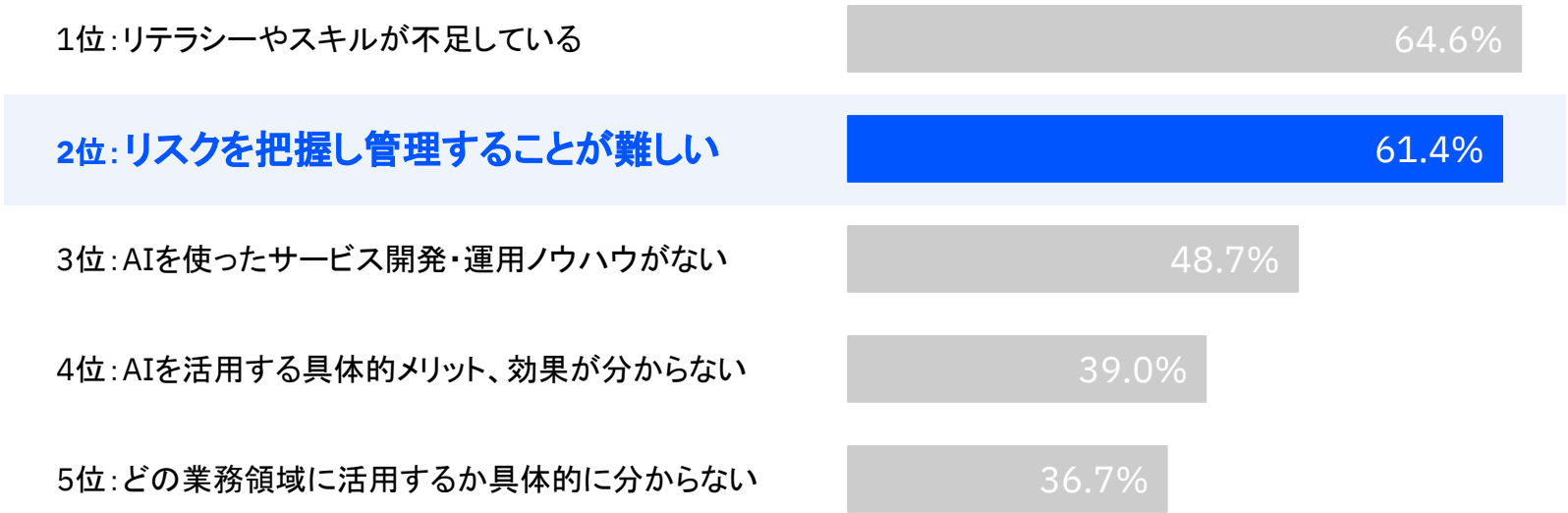
さまざまな AI に汎用的に適用可能

Citadel AI				
✓数値データに基づく AI	[物流業] 出入荷予測	[金融] 与信審査	[電力] 需要予測	[マーケティング] 広告効果
✓画像データに基づく AI	[製造業] 不良品検査	[保険] 事故査定	[医療] 画像診断	[建設業] 安全管理
✓テキストデータに基づく AI	生成 AI（大規模言語モデル）の各種アプリケーション			

1

企業における
生成 AI の活用とリスク

日本の大企業における生成 AI 活用のボトルネックはリソースとリスク



野村総合研究所、日本国内に本社を持つ売上高上位 3,000社を対象にした「IT 活用実態調査（2023年）」より引用。該当質問への回答数 n=308

企業における生成 AI 活用のユース・ケース

大規模言語モデル（ LLM）	**1. 一般質問** **2. 要約** **3. 翻訳** **4. 校正** **5. 資料・コンテンツ作成** **6. コード生成** **7. その他：顧客対応自動化　など**
画像・動画生成モデル	8. 画像・動画生成 9. デザイン・設計 など

社内向けのユース・ケースが主

生成 AI のリスクの具体例（1）

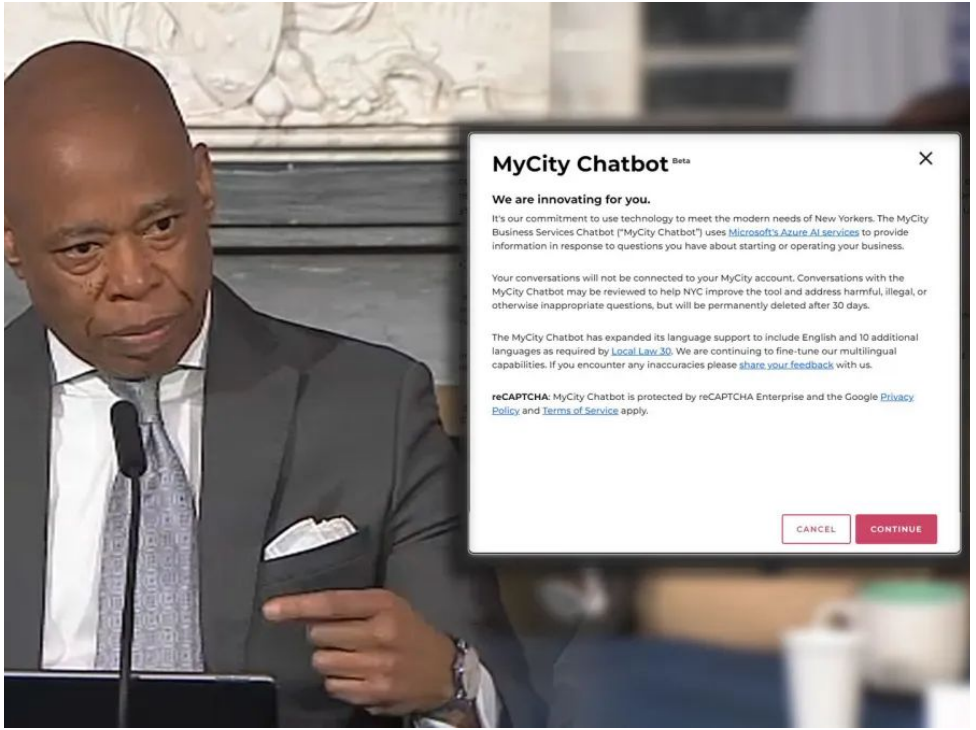
カナダ航空の
AI チャットボットが
誤った情報を顧客に提供し
顧客から訴えられる

2024年2月16日

生成 AI のリスクの具体例（ 2）

ニューヨーク市の
公式 AI チャットボットで
違法なアドバイスを出力

2024年4月4日



生成 AI のリスクの具体例（３）

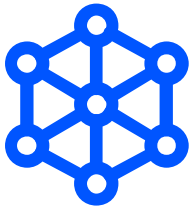
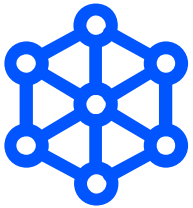
生成 AI 悪用し
ウイルス作成容疑で
無職の男逮捕

2024年5月28日

2

生成 AI に関わるリスクの評価方法

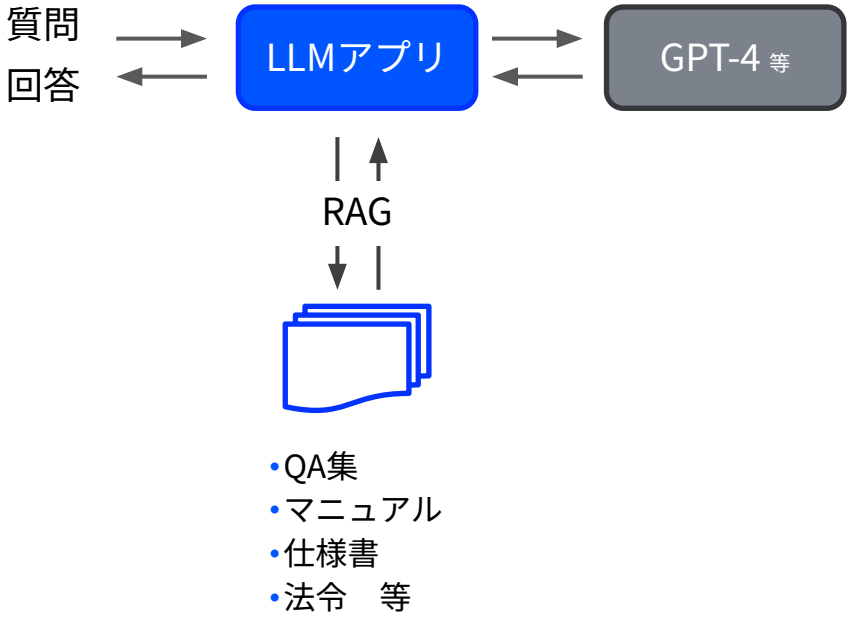
従来の予測 AI と生成 AI の違い

	入力	モデル	出力	
予測AI	定型入力データ • 数値 • 文字 • 画像　等	→  →	99.5%の確信度でXX 0.5% の確信度でYY	具体例: • 与信審査 • スパム判別 • 医療画像診断　等
生成AI	非定型な指示 • 質問 • 要約 • 資料作成　等	→  →	非定型なコンテンツ • 文章 • 画像 • コード　等	具体例: • ChatGPT • Llama2 • Stable Diffusion

従来の予測 AI と生成 AI の違い

	入力	モデル	出力
予測AI	定型入力データ • 数値 • 文字 • 画像　等		Ground Truth が存在 99.5%の確信度でXX 0.5% の確信度でYY 具体例: • 与信審査 • スパム判別 • 医療画像診断　等
生成AI	非定型な指示 • 質問 • 要約 • 資料作成　等		Ground Truth が確定しない 非定型なコンテンツ • 文章 • 画像 • コード　等 具体例: • ChatGPT • Llama2 • Stable Diffusion

大規模言語モデル（LLM）活用における課題



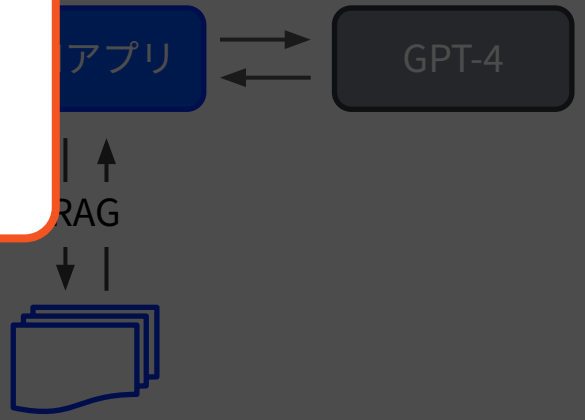
LLMのリスク

- ハルシネーション
- 有害なアウトプット
- 差別的なアウトプット
- 文脈の不一致
- プロンプト・インジェクション 等

大規模言語モデル (LLM) 活用における課題

質問・回答のログデータ

- ログが大量にあり目視作業で確認し切れない
- 客観的に比較評価し、整理する仕組みが無い

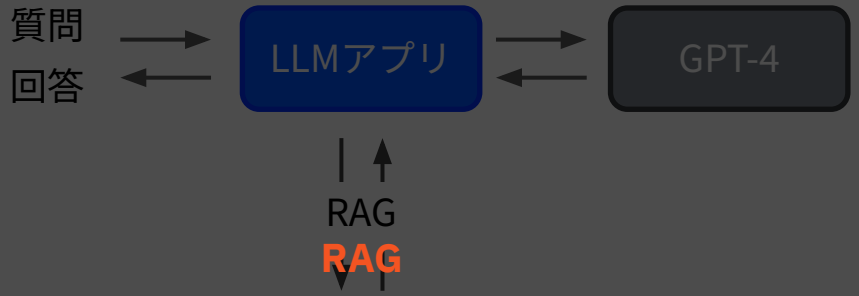


- QA集
- マニュアル
- 仕様書
- 法令　等

LLMのリスク

- ハルシネーション
- 有害なアウトプット
- 差別的なアウトプット
- 文脈の不一致
- プロンプト・インジェクション 等

大規模言語モデル (LLM) 活用における課題

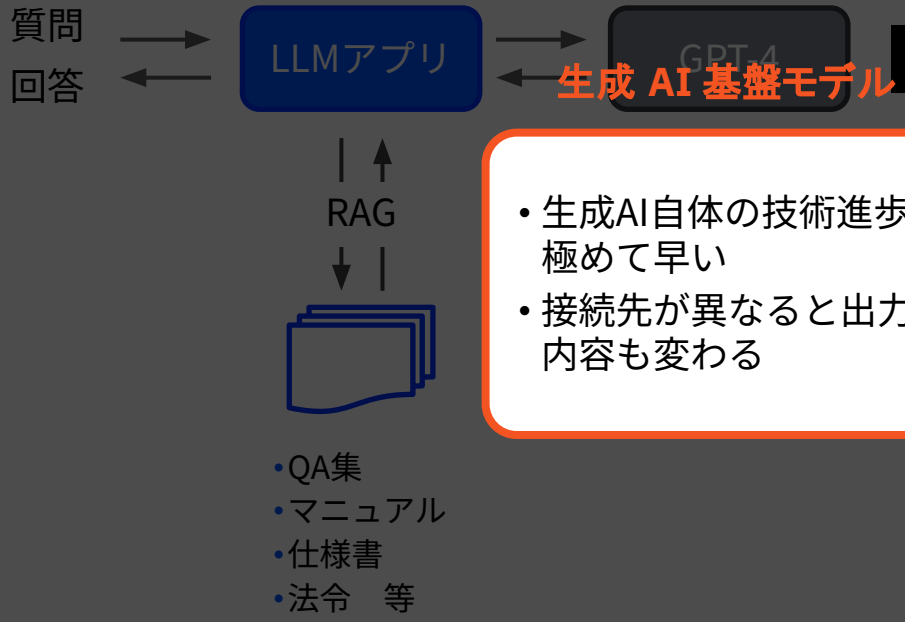


- RAGの内容も常に更新する必要がある
- 今までのPromptで大丈夫な保証が無い

LLMのリスク

- ハルシネーション
- 有害なアウトプット
- 差別的なアウトプット
- 文脈の不一致
- プロンプト・インジェクション 等

大規模言語モデル (LLM) 活用における課題



- 生成AI自体の技術進歩が極めて早い
- 接続先が異なると出力内容も変わる

LLMのリスク

ルシネーション

害なアウトプット

別的なアウトプット

派の不一致

•プロンプト・インジェクション 等

Citadel AI が開発した LLM 信頼性評価ツール

1. LangCheck

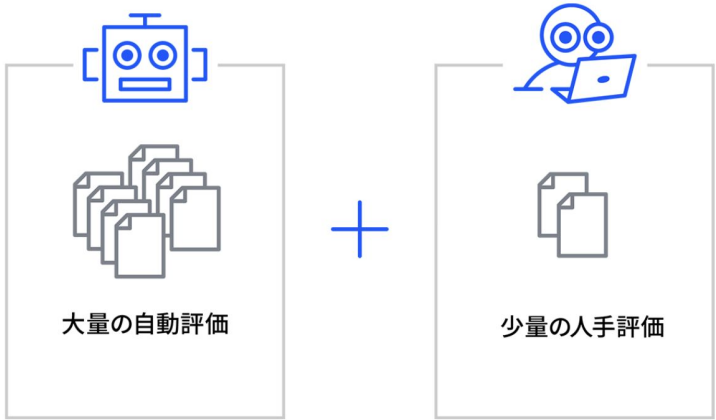
評価指標のオープンソースライブラリ

LangCheck

2023年10月 発表

2. Lens for LLMs

UI・分析ツールを実装した商用版

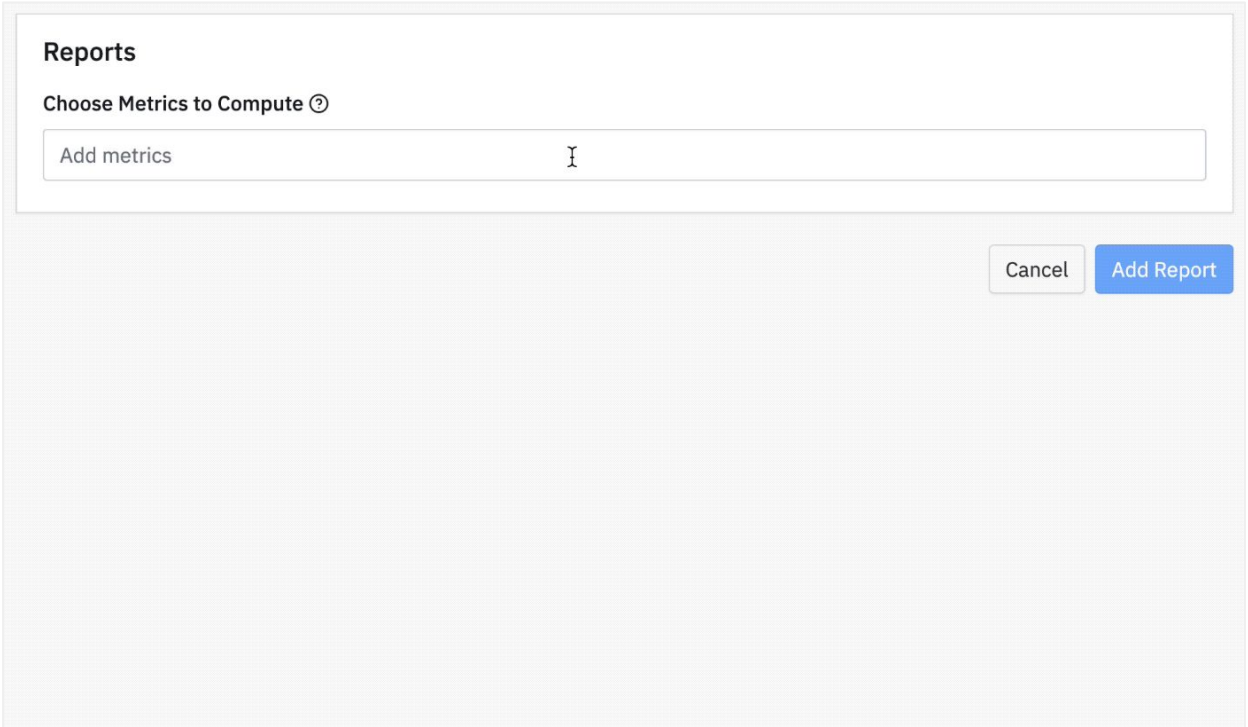


2024年4月 発表

LLM アプリケーション性能評価指標の例

Type	Metrics	Description
Reference-Free Metrics (LLM入出力を直接評価)	**sentiment**	文章に現れる感情 (Negative = 0, Positive = 1)
	toxicity	文章の有害性
	fluency	文章の流暢さ (文法的におかしくないか？)
	tateishi_ono_yamada_reading_ease	日本語の読みやすさ
	answer_relevance	質問と回答の関連性
Reference-Based Metrics (他の文章とLLM入出力を比較)	**semantic_similarity** **rouge**	文章二つの類似度
Source-Based Metrics (RAG等で取得したソースと出力を比較)	**factual_consistency**	レスポンスとソースの論理的一貫性
	context_relevance	質問とソースの関連性

自動評価メトリクスを用いた品質評価

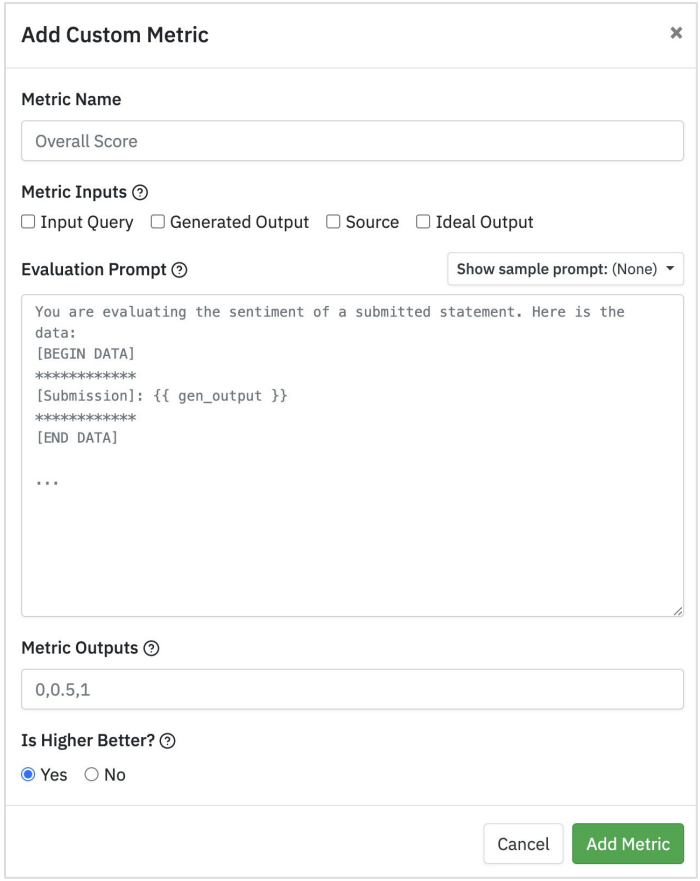


カスタム・メトリックによる評価

- **ユーザーの分析目的に応じて独自の評価メトリックを作成**

例：

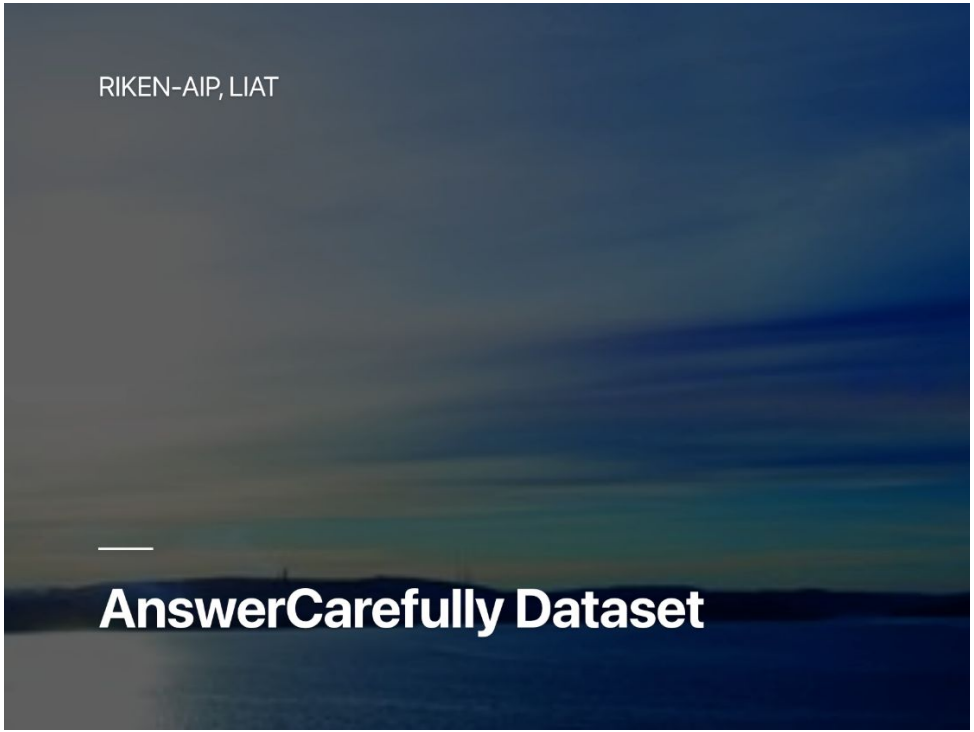
- 機密情報評価
- 攻撃的プロンプト評価
- ユーザーフィードバック評価 等



理化学研究所と共同開発した有害テキストデータセット

日本語 LLM 出力の
安 全 性・適 切 性 に 特 化 した
AnswerCarefullyデータセットを
公開

2024年4月30日



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:

松葉（まつば）

takehito@citadel.co.jp

企業URL:

https://citadel.co.jp

Citadel AI