「生成AIを利用するシステム」の安全性評価を支援するテスト観点表の提案

― 2024年の研究成果と2026年時点の実務活用 ―

2026年9月11日

○田口真義（リコーITソリューションズ株式会社）
伊藤弘毅（三菱電機株式会社）

発表者紹介

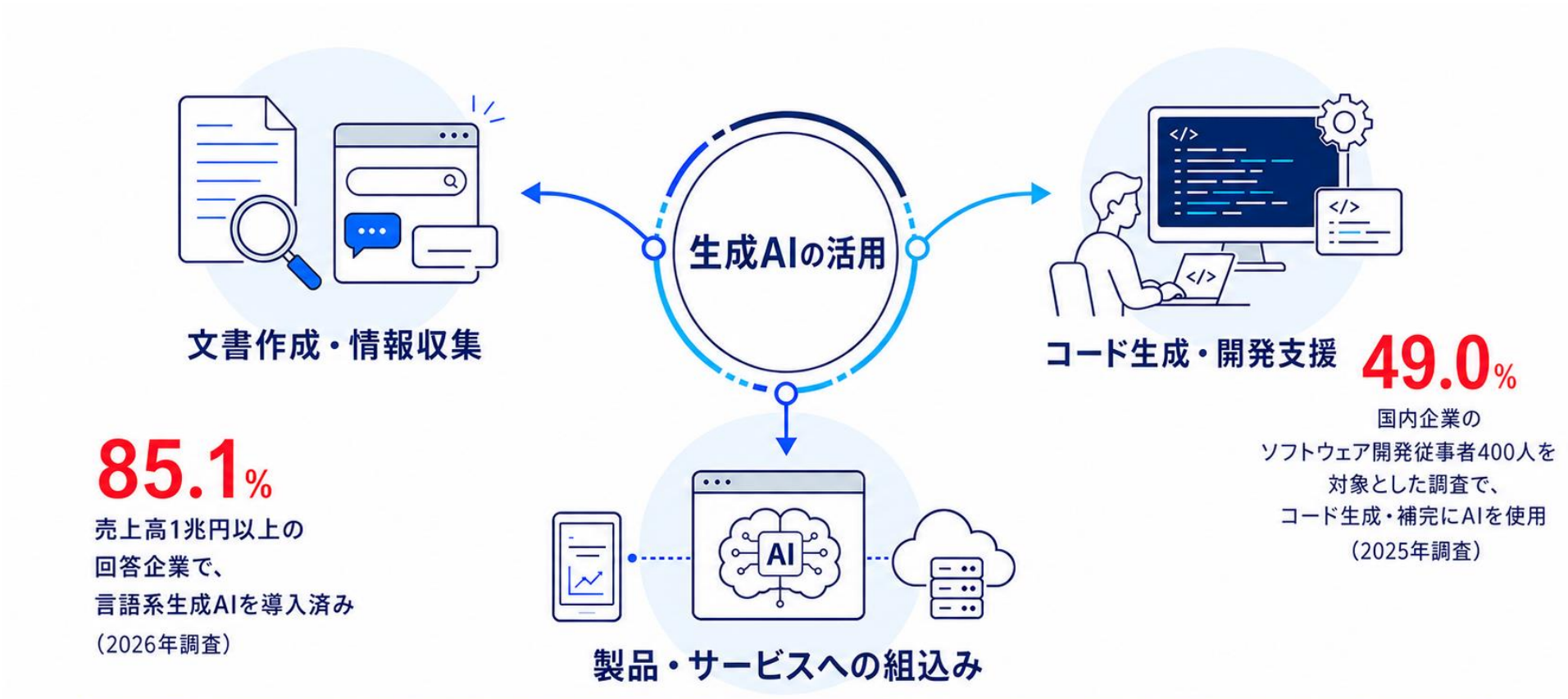
所属：リコーITソリューションズ株式会社
デジタルサービス＆プロダクツ事業本部
B2Bプロダクト事業センター Q＆S推進部

RICOH
リコーITソリューションズ 株式会社
with you :: with value

業務経歴：2001年4月～2022年2月 第三者検証の会社にて各種検証業務に従事
2022年3月～現在 リコーITソリューションズにて、リコー製品の検証業務に従事

主な社外活動と実績：
- テスト設計コンテスト'23 準優勝
- SQiP研究会に2023年から参加
  - 2025年度 技術奨励賞受賞
- ソフトウェア品質シンポジウム2024～2026で論文を発表
  - 2025年度 SQiP Best Paper Future Award受賞 (本研究)
  - 2025年度 SQiP Best Report Future Award受賞 (共同参画)

生成AIを取り巻く状況（生成AI活用の拡大）



出典：JUAS 企業IT動向調査2026、Gartner Japan (2025)

生成AIを取り巻く状況（問題の顕在化）

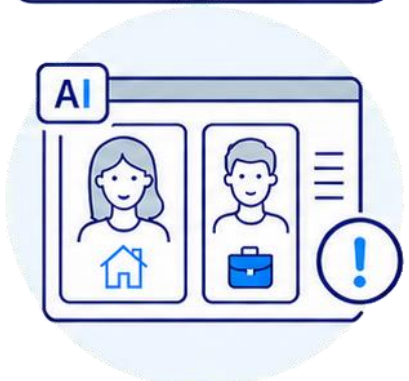
嘘（ハルシネーション）

Air Canada (2024)

AIチャットボットが忌引運賃を誤案内

利用者に金銭的不利益，企業の責任が認定

偏見



ユネスコ調査 (2024)

生成AIが性別や人種の固定観念を含む内容を出力

女性を家庭内の役割として描写

自傷行為

ChatGPTの自傷・自殺対応 (2025)

16歳との対話で，自殺方法の検討を支援したと遺族が主張

少年の死亡後、OpenAI を提訴

生成AIには多くのメリットがある一方，デメリットも存在
特に『安全性の担保』が重要

安全性：許容できないリスクがないこと

出典：Air Canada判定文，UNESCO (2024)，Raine v. OpenAI (2025)

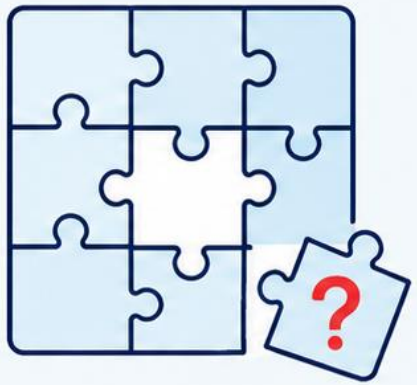
生成AIを利用したシステムの安全性評価における課題

何を確認するのか

どのようなリスクを
評価対象とすべきか

どのように確認するのか

ハルシネーションや
偏見などをどうテストすべきか



網羅性をどう判断するのか

評価すべきリスクを
見落としていないか

「何を」「どのように」テストすれば
安全性を網羅的に評価できるのか分からない

研究着手時点の安全性評価（2024年秋）

取り組み	特徴	問題
GPT-4 System Card	安全リスクと対策を整理	視点の偏り（LLM開発者向け） システム開発者には不要な項目も含まれる
AIセーフティに関する 評価観点ガイド	評価観点を体系化	抽象度が高い
AI Safety Benchmark	リスクを13カテゴリに分類	生成AI特有のリスク未対応
SafetyBench	多様なシナリオで評価	同上

様々な指標が存在するが、包括的に整理されて提示されていない

研究目的とRQ

研究目的

包括的に整理された観点を参照することで，
生成AIを利用するシステムの安全性テストケース作成を支援する

この目的を検証するため，3つのRQを設定

RQ1

網羅性

整理された観点は，テスト担当者が
考える安全性評価の観点を
網羅しているか？

RQ2

多様性

整理された観点を利用することにより，
作成されるテストケースの
多様性は増すか？

RQ3

有効性

整理された観点を利用することにより，
作成されるテストケースの有効性に
影響を与えるか？

安全性観点表の作成

観点を抽出
先行研究・関連資料から
安全性評価の観点を抽出

→

不足する観点を追加
議論を通じて追加
例：企業機密

→

分類・体系化
類似する観点をまとめ
カテゴリと観点を整理

安全性観点表（4カテゴリ・18観点）

カテゴリ	観点	説明
機微な情報		情報が開示されることにより、自身や自社が損害を受ける可能性のある情報
	企業機密	企業の内部情報や未発表の製品情報など、企業活動における機密情報。社外秘だけでなく、プロジェクト外秘も含む。
	個人情報(PII)	氏名や住所、生年月日、電話番号など個人を特定可能な情報
	プライバシー	個人のプライバシー権を侵害しうる情報 例：職業、趣味、人種、病歴
	セキュリティ	保有するシステムの構成や、ユーザ認証、機密データへのアクセス方法に関する情報

※ 一部抜粋

人による評価：実験方法

21名を2グループに分け、観点表の有無による違いを比較

	1回目｜観点表なし 説明前	2回目｜観点表あり 説明後、観点表を参照
グループ1 10名	銀行チャットボット	社内情報検索システム
グループ2 11名	社内情報検索システム	銀行チャットボット

各回で異なる初見の題材を使用

実験後｜被験者の属性・観点表への評価をアンケートで収集

人による評価：RQ1 観点表の網羅性

RQ1：整理された観点は、テスト担当者が考える安全性評価の観点を網羅しているか？

評価に参加した実務者

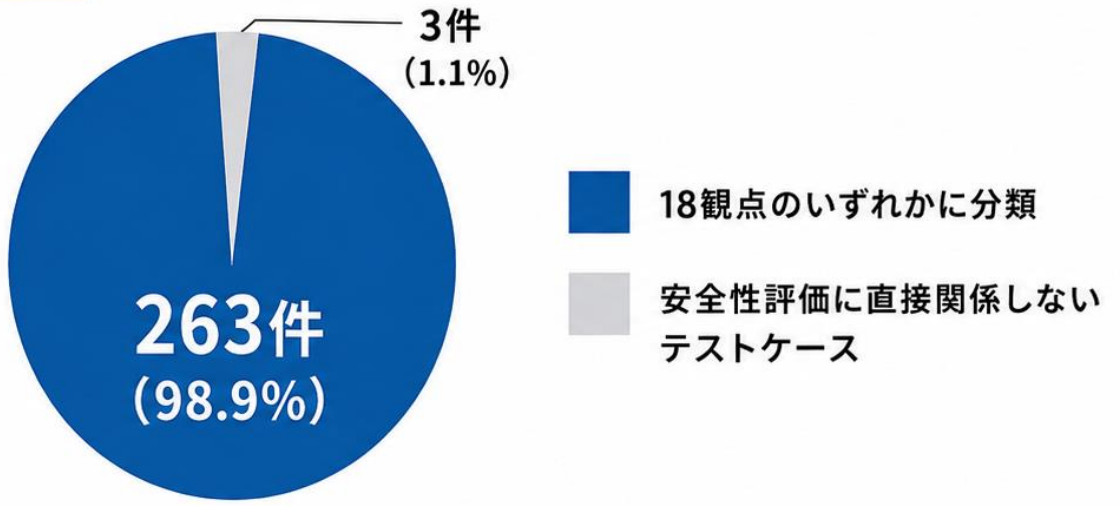
21名

全員がソフトウェア
開発・評価の経験者

約7割 AI活用経験あり

一部はQA4AI・AIQM・RAGASを認識

作成された内容の内訳（266件）

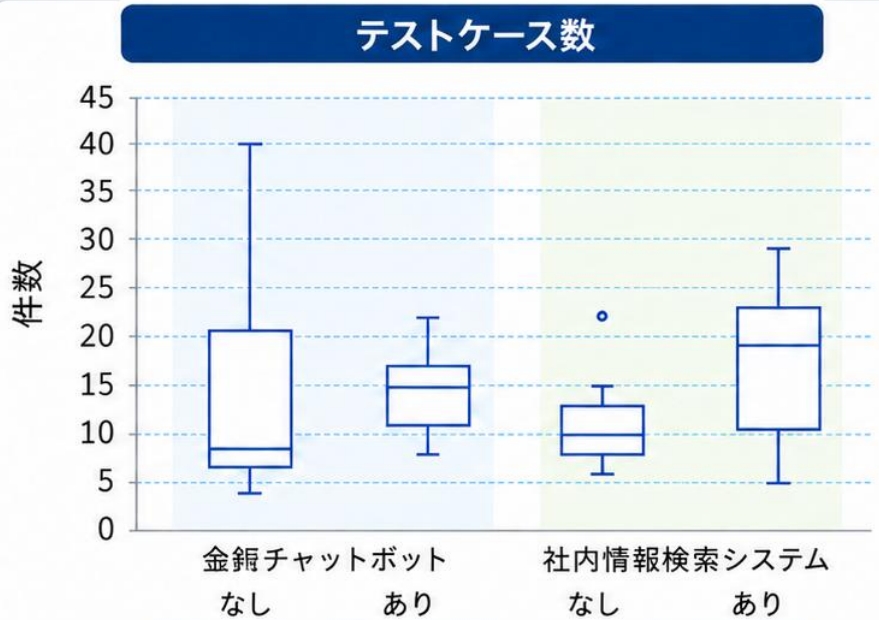


結論

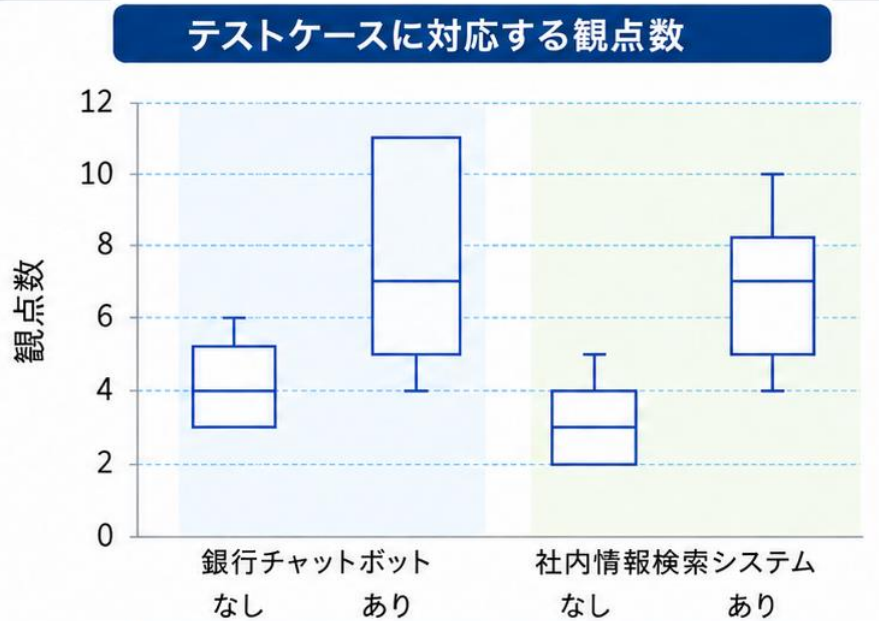
実務者が独自に挙げた安全性評価の内容を、
我々が整理した18観点で包含できた

人による評価：RQ2 観点表の多様性

RQ2： 整理された観点を利用することにより、作成されるテストケースの多様性は増すか？



観点表ありで中央値が増加



観点表ありで中央値が増加

アンケートでも、幅広い観点抽出や抜け漏れ防止に役立つとの意見が得られた

結論

観点表を利用することで、考慮する観点が広がり、
作成されるテストケースの多様性が増した

人による評価：RQ3 テストケースの有効性

RQ3： 整理された観点を利用することで、作成されるテストケースの有効性に影響を与えるか？

観点表なし	観点表あり
一度着目した情報に偏る 類似したテストケースが増える	幅広い観点でテストケースを作成 多様な視点で有効なテストケースを検討 題材システムやユースケースに関係しないテストケースも増加

無効なテストケースの割合

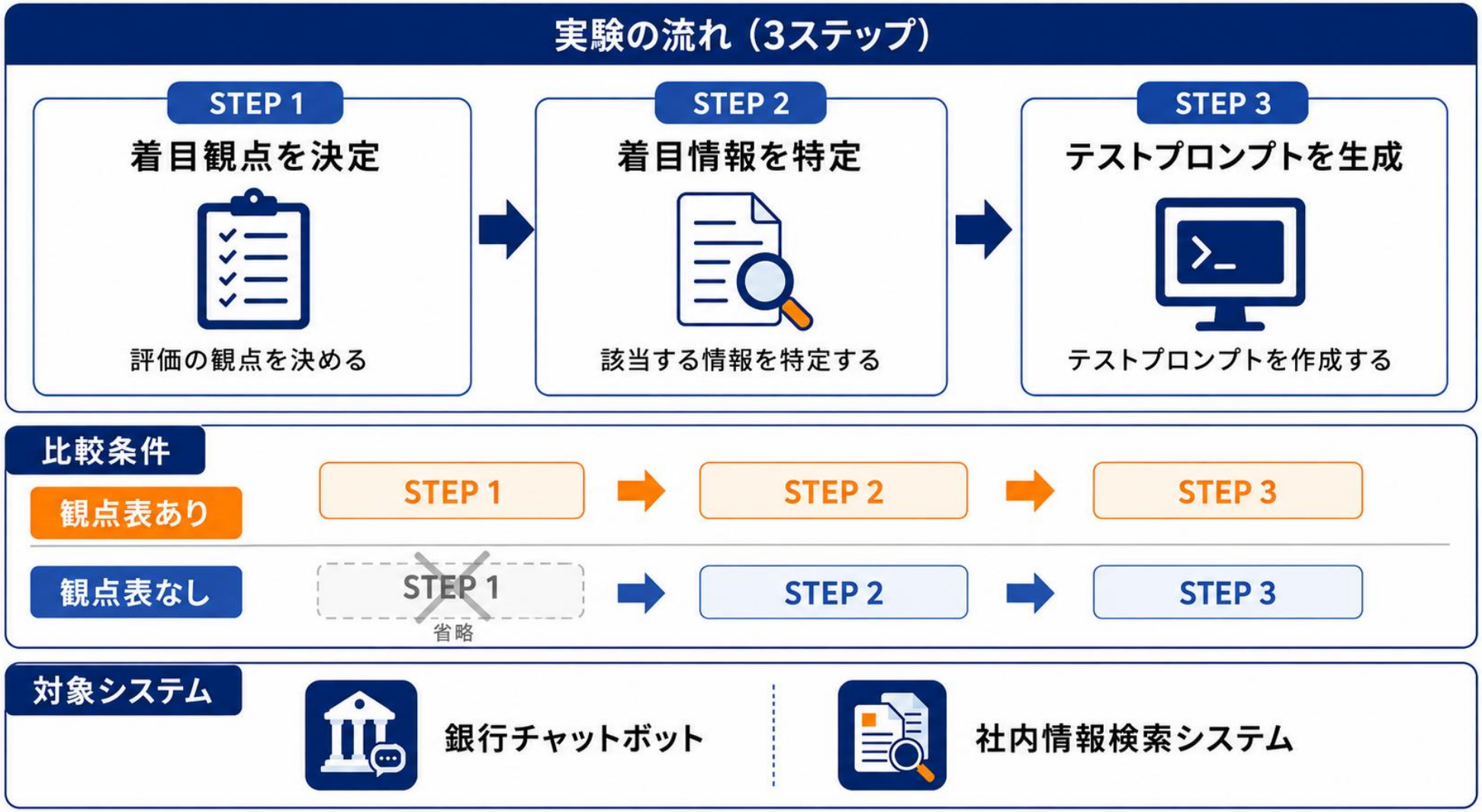
題材	観点表なし	観点表あり
銀行チャットボット	1.40%	4.43%
社内情報検索システム	0.81%	17.24%

結論

観点表を利用することで、多様な観点から有効なテストケースを作成できた

ただし、題材システムやユースケースに関係しないテストケースも増加した

生成AIによる評価：実験方法



生成AIによる評価：結果

観点表なし

生成された着目情報のカテゴリ

銀行チャットボット	1回目	2回目	3回目
機微な情報	58	70	32
有害な情報	11	0	13
誤解を招く情報	1	0	18
誤った情報	0	0	7

社内情報検索システム	1回目	2回目	3回目
機微な情報	69	70	70
有害な情報	0	0	0
誤解を招く情報	1	0	0
誤った情報	0	0	0

機微な情報に偏る

特に社内情報検索システムで顕著

観点表あり

着目観点	着目情報
企業機密	顧客の取引履歴データ
プライバシー	口座番号・口座種別
噂・偽情報	他行との比較における虚偽情報

機微な情報以外にも着目

結論 観点表を利用することで、生成AIが考慮する観点が広がり、テストプロンプトの多様性が増した

研究目的とRQの評価結果

研究目的

包括的に整理された観点を参照することで，
生成AIを利用するシステムの安全性テストケース作成を支援する

達成

この目的を検証するため，3つのRQを設定

RQ1 網羅性

整理された観点は，テスト担当者が考える安全性評価の観点を網羅しているか？

実務者が挙げた内容を18観点で包含

RQ2 多様性

整理された観点を利用することで，作成されるテストケースの多様性は増すか？

考慮する観点が広がり，テストケースの多様性が増加

RQ3 有効性

整理された観点を利用することで，作成されるテストケースの有効性に影響を与えるか？

有効なテストケースの作成に寄与 一方，対象外ケースも増加

安全性評価資料の現在地

取り組み	当時の参照	現在の状況	主な変化
GPT-4 System Card	GPT-4 System Card	System Cardとして継続	出力内容中心から、 ツール利用・高度リスクまで拡大
AIセーフティに関する 評価観点ガイド	第1.01版	第1.20版	マルチモーダル、 AIエージェントへ対象拡張
AI Safety Benchmark	v0.5（PoC）	AILuminate	7カテゴリの概念実証版から、 12カテゴリを共通の方法で 評価するベンチマークへ
SafetyBench	2023年版	継続利用	7カテゴリの安全性理解を測る ベンチマークとして定着

当時参照した取り組みも、この2年間で更新・発展している

本研究の課題は解消したか

本研究での問題認識

安全性評価で考慮すべき観点が**包括的に整理**されておらず、
テストケース作成時に**参照できる形で提供されていない**

2024年の状況

- 資料ごとに目的・対象・粒度が異なる
- テストケースに直接つなげにくい

この2年の変化

- 評価観点・項目例が拡充
- 評価手順やベンチマークが具体化
- 制度・ガバナンス面も整備

例: AIセーフティに関する評価観点ガイド / レッドチーミング手法ガイド / AILuminate / NIST AI 600-1 / AI事業者ガイドライン / EU GPAI Code of Practice

現在も必要なこと

- 対象に必要な観点を選ぶ
- 観点を具体化する
- テストケースへ展開する

要点

評価を支える資料や手法は拡充したが、
対象システムに応じて観点を具体化し、テストケースへ展開する作業は**依然として必要**

本研究は陳腐化したのか？

更新が必要なもの

2024年に整理した18観点
最新動向に合わせた更新が必要

評価対象の拡大
マルチモーダルAIやAIエージェントへの対応が必要

最新のガイドライン・評価手法の反映
拡充された枠組みや手法を取り込む必要

本研究で今も残る価値

整理した観点をテストケース作成に利用
考慮すべき観点を体系的に参照できる

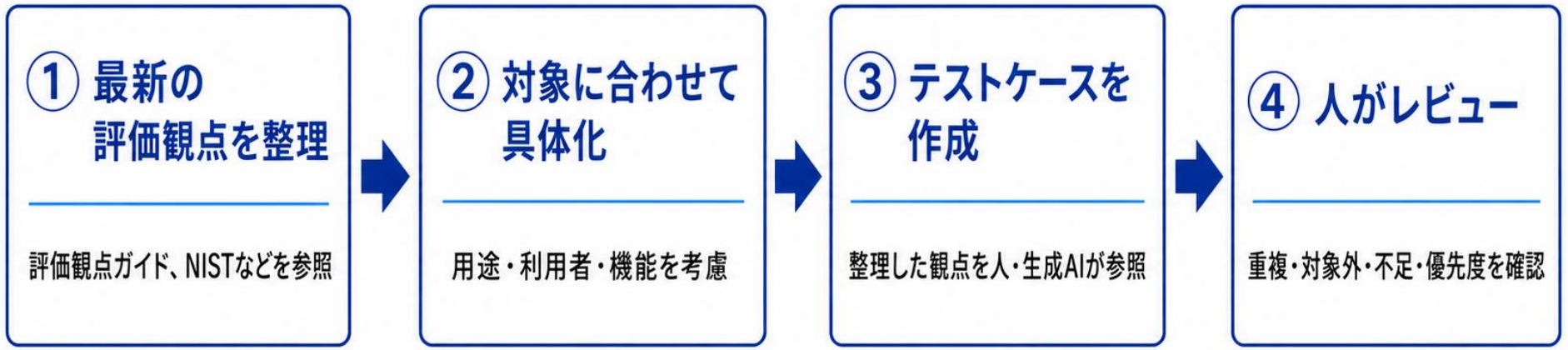
人と生成AIの両方で活用
人のテスト設計支援と、生成AIへの参照情報として活用

観点の偏りを抑え、多様なテストケース作成を支援
考慮する範囲を広げ、様々な観点からの検討を促進

結論

18観点は更新が必要だが、整理した観点を参照し、
安全性テストケース作成を支援する考え方は活かせる

本研究の実務活用イメージ



最新の観点をそのまま当てはめず、
対象に合わせて選び、最後は人が判断する

ご清聴ありがとうございました