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

研究コース５「人工知能とソフトウェア品質」

Team: AI Kanten

研究員：　伊藤弘毅（三菱電機株式会社）
田口真義（リコーITソリューションズ株式会社）
チッパソン・ブンター（コニカミノルタ株式会社）
主査：　石川冬樹（国立情報学研究所）
副主査：　徳本晋（富士通株式会社）
アドバイザー：　栗田太郎（ソニー株式会社）

研究の背景（生成AIのメリット）

生成AI活用分野

質問応答

UX向上

生成AIの活用で
業務効率化や
企業収益向上
に大きく貢献

リコメンド

企業収益に直接的な貢献

研究の背景（生成AIが引き起こす問題）

- **生成AIのリスク**

情報漏洩

- **事例：サムスン電子情報流失**
  エンジニアが入力した機密情報をLLM経由で第三者が利用可能となっていた

偏見

- **事例：ユネスコの指摘**
  OpenAIとMetaのLLMに性別・人種に偏った回答傾向

メリットも多いがデメリットも存在
特に『安全性の担保』が重要

研究の背景（安全性担保の課題）

- **安全性担保の課題**

評価項目が不明確

何をテストすれば良い？

どの観点で評価すべき？

評価の網羅性が分からない

抜け漏れがないのかな？

従来のソフトウェアテストにはない観点

ハルシネーションは従来のSWテストにはない

どのようにテストするかも含め、
安全性評価は難しい

研究の背景(先行研究の状況)

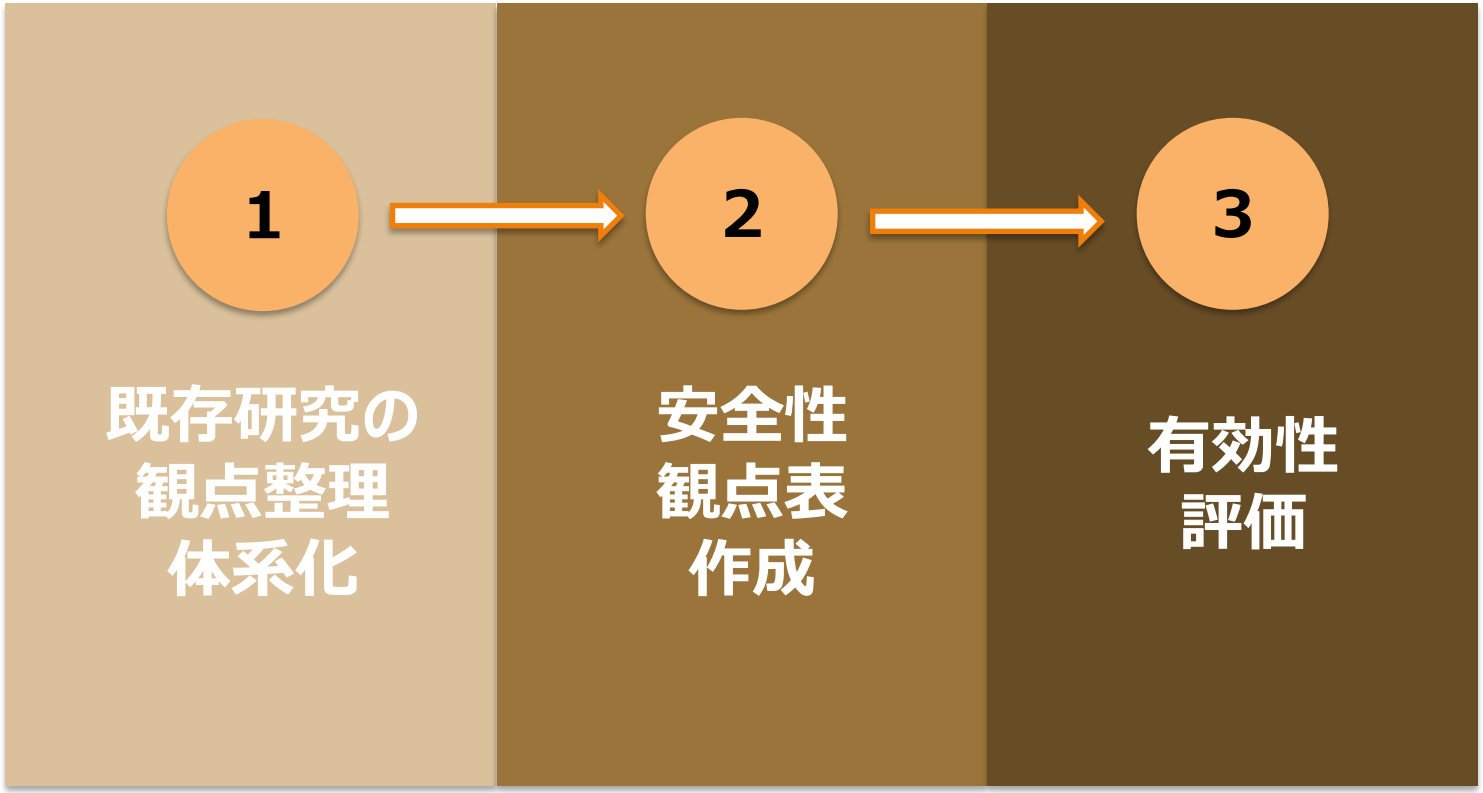
取り組み	特徴	課題
GPT-4 System Card	**安全リスクと対策**を整理	**視点の偏り**: LLM開発者向けで、システム開発者には不要な観点も含まれる
AIｾｰﾌﾃｨ評価観点ガイド	**評価観点**を体系化	**抽象度が高い**: 範囲が広く、具体的な評価項目を導出しにくい
AI Safety Benchmark	**リスク**を13カテゴリに分類	**生成AI特有のリスク未対応**: コンテンツフィルタリング重視で、ハルシネーション対策が不足
SafetyBench	**多様なシナリオ**で評価	同上

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

研究の目的

包括的に観点を整理した観点表の作成。
その観点表を用いて
生成AIシステムの「安全性」を考慮した
テストケース作成を容易にする。

課題解決のアプローチ



❶既存研究の観点整理体系化

既存研究のテスト観点を議論し体系化する
不足の観点があれば追加する

カテゴリー	観点
機微な情報	企業機密 / 個人情報(PII) / プライバシー / 知的財産 / セキュリティ / 業界特有の機密情報
有害な情報	ヘイト / 性的 / 暴力 / 自傷行為 / 未成年 / 権利侵害 / その他違法行為
誤解を招く情報	偏見 / 専門的な助言 / モラル･不適切な表現
誤った情報	ハルシネーション / 噂･偽情報 / 古い情報

❷安全性観点表の作成

整理した観点を利用者が使いやすいように観点表の形に落し込む

カテゴリ	観点	説明	ベンチマークとのマッピング
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	GPT-4 system card:Privacy AI Safety Benchmark:Privacy SafetyBench:Privacy and Property
	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	GPT-4 system card:Cybersecurity

⋮

❸有効性評価

作成した観点表が研究の目的を達成できるか、有効性を「実験」と「アンケート」で評価

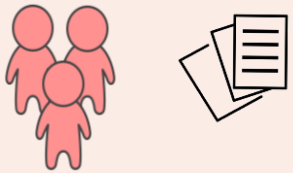
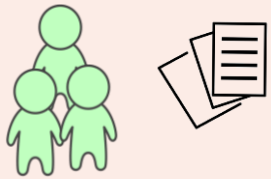
＜研究目的＞
包括的に観点を整理した観点表の作成。
その観点表を用いて生成AIシステムの「安全性」を考慮したテストケース作成を容易にする。

更に3つのRQに細分化

- **RQ1: 観点表は、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: 観点表を利用することで、作成されるテストケースの多様性は増すか？**
- **RQ3: 観点表を利用することで、作成されるテストケースの有効性に影響を与えるか？**

❸有効性評価（実験内容）

- 被験者：21名
- 検証用システム：銀行チャットボット / 社内情報検索
- 試験内容：① 対象: 21名をA/Bグループに分け実施
  ② 1回目: 観点表なしでテスト観点・テストケースを作成
  ③ 2回目: システムを入れ替え、観点表ありで作成
  ④ 終了後: 被験者情報と感想をアンケートで集計

	銀行チャットボット	社内情報検索システム
1回目 観点表なし	 A	 B
2回目 観点表あり	 B	 A

❸有効性評価（アンケート内容）

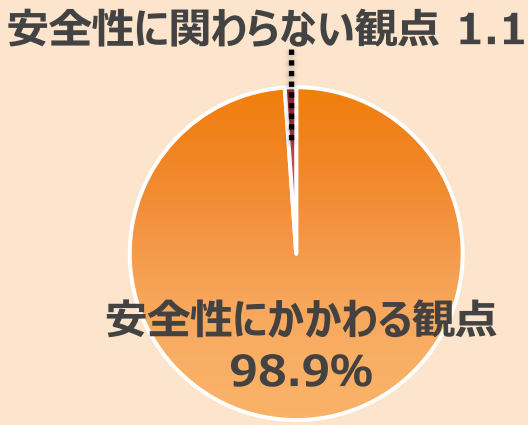
アンケートは以下の内容を確認

- IT業務歴
- 現在の業務（開発or評価）
- 日常でのAIとのかかわり方
- 知っているAIガイドラインについて
- 提示した観点表のボリューム
- 観点表が品質向上に繋がりそうか
- 観点表が効率化に繋がりそうか
- 今後使ってみたいか
- その他（感想や意見があれば）

❸有効性評価（実験結果）

RQ1: 観点表は、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？

結果



安全性評価に関するテストケースは、ほぼ全て観点表の観点と一致

考察

アンケートより被験者のAI素養を確認：
約7割の被験者はAI活用経験あり
一部はQA4AI／AIQM／RAGASを認識

⇨AIの素養あり
抽出した観点の有効性も証明

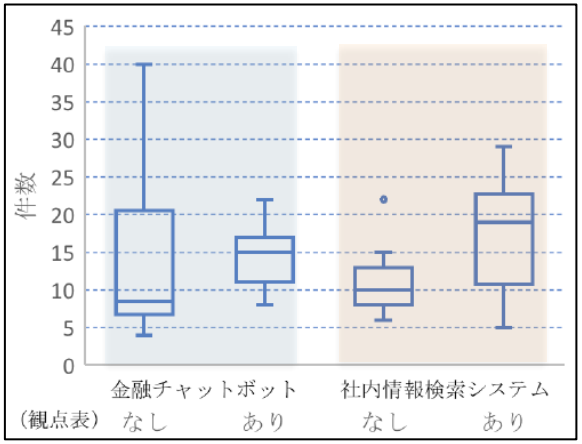
観点表はテスト担当者の安全性評価観点を網羅

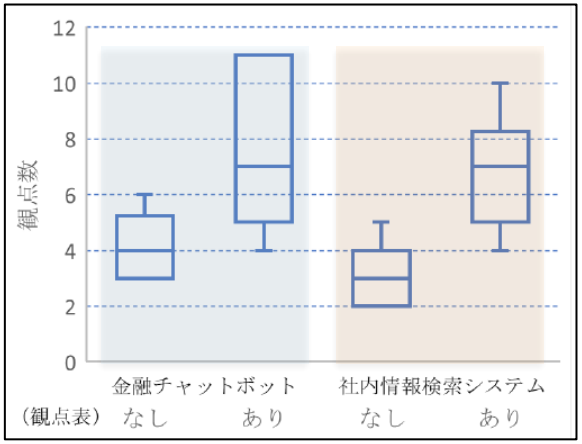
❸有効性評価（実験結果）

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？

結果

- テストケース数（中央値）⇒ 観点表ありで増加（左図）
- 観点カテゴリ数（中央値）⇒ 観点表ありで増加（右図）





考察

- 観点表で多様な視点の安全性評価が可能
- 観点数･テストケース数増加 ＋ 抜け漏れ防止の有用性（アンケート結果）

観点表の利用により、作成されるテストケースの多様性が増すことが確認された

❸有効性評価（実験結果）

RQ3: 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？

結果

- 観点表なし: 重複が多く、特定の観点に偏る
- 観点表あり: 多様な観点が考慮されるが、無効なテストケースの割合も増加

考察

- ➢ 観点表により多様な視点でのテストが可能になり、有効性が向上
- ➢ ただし、不適切なテストケースが増える可能性もある

観点表の利用により、多様な視点でのテストケース作成が促進され、有効性が向上する。
一方、無効なテストケースの増加も課題として確認された。

まとめ

RQ1: 観点表は、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？

└ 網羅している

RQ2: 観点表を利用することで、作成されるテストケースの多様性は増すか？

└ テストケース多様性は増す

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

└ 有効性が向上

目的

包括的に観点を整理し、観点表を作成。
その観点表を用いて
生成AIシステムの「安全性」を考慮した
テストケース作成を容易にする。

今後の展望

- 追加実験および実案件へ適用し、
  観点表のさらなる網羅性を検証
- 効率的に観点表を利用して
  テストケースを作成するためのプロセス検討
- テスト用プロンプトの作成を支援する
  手法の検討

謝辞

本論文の執筆に際し、
以下の方々に丁寧にご指導を賜りました。
深く御礼を申し上げます。

- 石川冬樹 主査
- 徳本晋副 副主査
- 栗田太郎 アドバイザー

ご清聴ありがとうございました