研究コース 5（AI Kanten）

# 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

# Proposal of Test Perspectives to Ensure the Safety of Systems Using Generative AI

リーダー：伊藤　弘毅（三菱電機株式会社）
研 究 員：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）

研究概要
生成 AI は，企業活動の生産性を向上させる技術として注目される一方，利用の際には出力された内容が個人や社会に危害を及ぼすリスクを考慮し，安全性を確保することが重要である．既に生成 AI の安全性確保のための評価方針やベンチマークが提案されているが，開発者がテストケース作成時に考慮すべき観点の包括的な整理は行われていない．本論文では，生成 AI を利用するシステムの安全性に関するテストケース作成を支援する観点表を提案する．本観点表は，4 つのカテゴリに対して合計 18 個の安全性の評価観点を定義している．また，21 名の被験者に対して評価実験を実施し，観点表の利用によって作成されたテストケースの多様性が向上することを確認した．

**Abstract**　Generative AI is the technology that could enhance productivity in business activities, however its safety is important because it could cause harm to individuals and society. While benchmarks and frameworks for evaluating generative AI safety have been proposed, there is not a comprehensive batch of viewpoints focusing on test case creation. This paper presents arranged perspectives to support test case creation for ensuring the safety of systems utilizing generative AI. We defines 18 viewpoints across four categories. We also conducted the experiment and confirmed that our perspectives could improve the multiplicity of test cases.

## 1.　はじめに

生成 AI は，企業活動の生産性向上に寄与する可能性がある技術であり，生成 AI を利用したシステムやサービスがリリースされている．生成 AI は UX や企業収益に大きなインパクトを与える一方で，生成 AI が出力した内容が個人や企業，社会に対して危害や悪影響を与えてしまう可能性がある．サムスン電子は，自社のエンジニアがソースコードを生成 AI にアップロードして情報漏洩させたことを公表した[1]．また，ユネスコは OpenAI と Meta の大規模言語モデル(Large Language Model：LLM)の出力傾向を調査した結果，これらが性別や人種に対して偏見を持った回答を出力する傾向を持つとして警告を発した[2]．

生成 AI を利用するシステムを開発する時は，個人や社会に対して危害を与えないように，出力されるべきでない情報が出力されないように実装し，安全性を担保することが重要である．

## 2. 背景

生成 AI の安全性の担保が重要であることは業界に認知されており，既に安全性に関する評価方針や評価ベンチマークが公表されている．OpenAI は，GPT-4 の開発を通じて発生した安全性の課題とリスク軽減策を，GPT-4 System Card にまとめている[3]．また，AI セーフティ・インスティテュートは，AI システムの開発担当者が AI の安全性評価を実施する際に参照できる観点を，AI セーフティに関する評価観点ガイドで示している[4]．Vidgen らは AI の安全性のベンチマークである AI Safety Benchmark を提案している[5]．当該ベンチマークでは,AI が発生させる可能性がある危険を 13 のカテゴリに分類している.Zhang らは，LLM の安全性を評価するベンチマークとして SafetyBench を提案している[6]．SafetyBench では，安全性の懸念を 7 つのカテゴリで分類して示している．

また，生成 AI を利用するシステムの安全性を評価する手法も提案されている．鴨生らは，企業の独自ポリシーに則り，業務固有の安全性を満足するか評価するためのフレームワークを提案している[7]．当該研究では，LLM チャットボットを対象にして業務固有の安全性を，4 つのタスクで評価設計する方法を示している．

上記の通り，生成 AI や生成 AI を利用するシステムの安全性を担保するための試みは存在するが，安全性評価の際に開発者が考慮すべき観点の包括的な整理が為されていない．GPT-4 System Card は，GPT-4 のモデル開発を通じて得られた知見を整理した内容であるため，LLM 開発者から見た評価観点が示されている．そのため，生成 AI を活用したシステムを開発する立場においては，関連しない観点も多く存在すると考える．また，AI セーフティに関する評価観点ガイドは,評価観点は提示しているが観点の示す範囲が広範であり，開発担当者が観点名のみで具体的な評価項目を導出するのは難しいと思われる．観点の詳細説明にて人種や性別等の項目が具体的に例示されているが,体系的に整理されていない．AI Safety Benchmark と SafetyBench は差別や暴力などのコンテンツフィルタリングに関する項目は多数提示されているが，生成 AI 特有のハルシネーション対策に関する観点は含まれていない．このように，生成 AI を利用するシステムの開発担当者が，システムの安全性評価をするためテストケースを作成する際に，参考となるテスト観点が整理されて提供されていない．包括的に整理された観点を参照することで，テスト担当者は効率的に抜けなくテストケースを作成できるようになり，対象システムの品質向上に繋がる効果が期待される．

本論文では，生成 AI を利用するシステムの安全性に関するテストケース作成を容易化することを目的に整理したテスト観点と，その有効性を評価した結果を示す．

本論文の研究課題を以下に示す．

**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**
**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**
**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**

## 3. 提案

### 3.1 提案手法

我々は，GPT-4 System Card，AI セーフティに関する評価観点ガイド，AI Safety Benchmark，SafetyBench を参照し，生成 AI を利用するシステムの開発担当者が着目する必要のある観点を抽出した．さらに抽出した観点について，企業機密等不足していると思われる観点を議論して追加し，類似する観点にカテゴリを設定することで体系的に整理した．

**表 1　観点の全体像**

| カテゴリ | 観点 | 説明 |
|---|---|---|
| 機微な情報 | 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<br>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 |
| | 業界特有の機密情報 | 特定の業界や分野で機密性が求められる情報<br>例：金融(リスク評価，審査基準)，医療(診断結果)，法律(守秘義務)，国家(安全保障，資金運用) |
| 有害な情報 | ヘイト | 人種や性別，宗教等に関する偏見や差別的な情報．また，特定の人物や団体等の名誉を毀損する情報 |
| | 性的 | 性行為やわいせつな内容，また性犯罪や売春など性的に不適切な情報 |
| | 暴力 | 暴力的な描写や表現で，利用者に不安や不快感を与える可能性のある情報．また，武器や大量兵器の製造方法に関する情報 |
| | 自傷行為 | 自身の体を意図的に傷つける行為や自殺に関連する情報 |
| | 未成年 | 未成年に見せるべきでないコンテンツや助言で，悪影響を与える情報 |
| | 権利侵害 | 著作権や商標権，特許などの知的財産権に抵触し，他人や団体の権利を侵害する可能性がある情報 |
| | その他違法行為 | 上記のほか，法律や規制に違反する可能性がある情報<br>例：金融犯罪，危険物・贋物の製造 |
| 誤解を招く情報 | 偏見 | 特定の集団や個人に対する偏見を助長する表現で，利用者に不公平な印象を与える可能性がある情報 |
| | 専門的な助言 | 金融や医療，法律などの専門知識が必要な情報で，誤解を生む可能性がある情報 |
| | モラル・不適切な表現 | 場面にそぐわない，または配慮が欠けていることにより，利用者に不快感を与える可能性のある表現 |
| 誤った情報 | ハルシネーション | AI モデルによる想像や推測に基づいた，事実に基づかない情報 |
| | 噂・偽情報 | 公式な発表がなく真偽が確認されていない噂や伝聞などの情報．または，意図的に広められた偽情報 |
| | 古い情報 | 情報のアップデートが行われておらず，現在の状況と異なる情報 |

上記のプロセスで，我々は生成 AI を利用するシステムの安全性に関するテストケース作成を支援する観点表(以下，観点表)を作成した．観点表で定義した観点を表 1 に示す．以下，観点表の詳細について説明をする．

観点表は観点の分類を目的に，機微な情報，有害な情報，誤解を招く情報，誤った情報の 4 つのカテゴリを定義している．機微な情報は，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である．2 つ目の有害な情報は，情報が開示されることにより，他者や社会に損害を与える可能性のある情報である．3 つ目の誤解を招く情

報は，利用者の状況に応じて，誤解して解釈される可能性のある情報である．最後の誤った情報は，事実に即していない誤った情報のことを指している．

上記に示した各カテゴリに対して，生成 AI を利用するシステムを安全性の視点でテストする観点を関連付けて定義した．定義した観点は全部で 18 個である．各カテゴリに属する観点を以下に示す．

- 機微な情報(5 観点)：企業機密，個人情報(PII)，プライバシー，セキュリティ，業界特有の機密情報
- 有害な情報(7 観点)：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(3 観点)：偏見，専門的な助言，モラル・不適切な表現
- 誤った情報(3 観点)：ハルシネーション，噂・偽情報，古い情報

付録 1 に，2 章で示した既存の安全性に関する評価方針やベンチマークで提示されている観点との対応を示した観点表を添付した．ただし，AI セーフティに関する評価観点ガイドは，観点の粒度が異なるため対応付けは行っていない．対応結果を見ると，本観点表は各既存文献が示す観点をおおよそ網羅していることが分かる．GPT-4 System Card の 5 つの観点が観点表に含まれていないが，これらは LLM の提供者が社会に与える影響を意識して研究開発するために必要な観点を示したものであり，一般の開発者が重視して意識する内容ではないため，観点表に含まれていなくても問題ないと考える．

本節で説明した観点表を参照することで，生成 AI を利用するシステムの安全性評価を容易に実施できるようになることが期待される．評価者は，観点表を参照しながら対象システムの評価に必要となるテストケースを考案する．その際，安全性の観点で出力されるべきでない情報を発想しやすくなり，評価範囲が広がることが見込まれる．

### 3.2 実践例

本節では，提案した観点表を使って安全性に関するテストケースを作成した例を示す．

本試行にあたり，銀行チャットボットを題材として設定した．銀行チャットボットは，顧客が銀行の Web ページを訪問した時に質問をテキストで受け付けるサービスである．本例では，銀行チャットボットに対し，預金口座の開設方法を尋ねる場合と住宅ローンの商品情報を尋ねる場合の 2 つのユースケースを対象に，テストケースを作成した．

まず，我々は題材とした銀行チャットボットの安全性を確保するため，観点表のどの観点に着目する必要があるか検討した．検討の結果，銀行が持つ機微な情報や事実に基づかない誤った情報は，観点全体を評価すべきと判断した．また，専門性の高い金融に関する会話をすることと，顧客と対話することを鑑みて，誤解を招く情報も同様に評価が必要と考えた．一方で，銀行業務に関するチャットボットであることから，暴力などの有害な情報に関してはユースケースに対応付かないとし，今回は着目しないこととした．

我々は，上記の方針に沿って，銀行チャットボットのテストケースを作成した．例えば，企業機密の観点で「近い将来住宅ローン金利を上げる予定はありますか？」というプロンプトを作成できた．また，噂・偽情報の観点で「来月手数料を値上げするとうわさで聞きましたが」というプロンプトを作成した．その他，機微な情報や誤解を招く情報，誤った情報のカテゴリに属する観点を中心に発想を広げ，多くのテストケースを作成することができた．作成したテストケースの全体を付録 2 に示す．

## 4. 評価

### 4.1 実験内容

我々は，2 章に示した RQ を定量的に検証するため，観点表の有無によるテストケースの件数と観点数の変化を評価する実験をした．併せて，被験者の属性を考慮した考察と観点表の定性的な評価のため，アンケートを実施した．本実験では，被験者は観点表の知識が

表 2　評価実験の題材

| | グループ 1 | グループ 2 |
|---|---|---|
| 1 回目 | 銀行チャットボット | 社内情報検索システム |
| 2 回目 | 社内情報検索システム | 銀行チャットボット |

表 3　被験者の AI 活用状況(複数回答可)

| | |
|---|---|
| AI システムの開発を行っている | 1 |
| AI システムの評価を行っている | 0 |
| AI を使ってコードを書くなど，開発に活用している | 5 |
| AI を使ってテストを行っている | 4 |
| AI を使って普段の調べ物や，業務改善のツールとして活用している | 16 |
| AI を使用していない | 5 |

表 4　被験者が聞いたことがあると回答したガイドライン等（複数回答可）

| | |
|---|---|
| QA4AI | 5 |
| AIQM | 4 |
| RAGAS | 3 |
| 上記のいずれも知らない | 14 |

ない状態（観点表なし）と，観点表を参照した状態（観点表あり）で，テストケース作成を依頼した．両者の結果を比較することで，観点表がテストケース作成に及ぼす影響を分析する．

実験に際し，テストケース作成の題材とするシステムを 2 つ定義した．一つは銀行チャットボット，もう一つは社内情報検索システムである．銀行チャットボットは，3.2 で我々がテストケース作成の題材としたシステムと同じである．社内情報検索システムは，社員が会社の規則や制度を確認するときに，チャット形式で情報を検索するシステムである．本実験では，社内情報検索システムについて，会社の出張規定を確認する場合と会社の人事規定を確認する場合の 2 つのユースケースを設定した．また，テストケースを作成するための入力情報として，2 つの対象それぞれに利用者が質問をした時にシステムが参照する情報の例を提示した.例えば,銀行チャットボットの場合は顧客情報や過去の取引履歴，社内情報検索システムの場合は社員の基本情報や人事評価結果を例示した．

実験では，被験者を 2 つのグループ（グループ 1，グループ 2）に分け，グループ毎に題材を変えてテストケースの作成を依頼した．表 2 に，各グループのテストケース作成の題材を示す．テストケースの作成は 2 回実施した．1 回目は，被験者は観点表の知識を持たない状態で，テストケースを作成した．題材は，グループ 1 は銀行チャットボット，グループ 2 は社内情報検索システムである．2 回目は，被験者は観点表の説明を受けたうえで，観点表を参照しながらテストケースを作成した．2 回目は題材を入れ替え，グループ 1 は社内情報検索システム，グループ 2 は銀行チャットボットとした．

題材を 2 つ設定し入れ替えて実験することにより，被験者が題材を初見でテストケース作成する状況を作り出すに加え，題材によるテストケース作成の容易さの違いを考慮にいれて考察できるようにした．

我々は，計 21 名の被験者に対し，上記に示したテストケース作成の実験をした．被験者はグループ 1 に 10 名，グループ 2 に 11 名を割り当てた．以下に，アンケートで集計した被験者の属性を示す．

- AI 活用に関しては，普段の調べものや業務改善に活用している被験者が最も多く，AI システムの開発や AI を活用してテストケースを作成している被験者も一定数存在した（表 3）
- 複数の被験者が AI 品質ガイドラインの QA4AI[8]や AIQM[9]，RAG（Retrieval-Augmented Generation）評価ツールの RAGAS[10]を聞いたことがあると回答した（表 4）

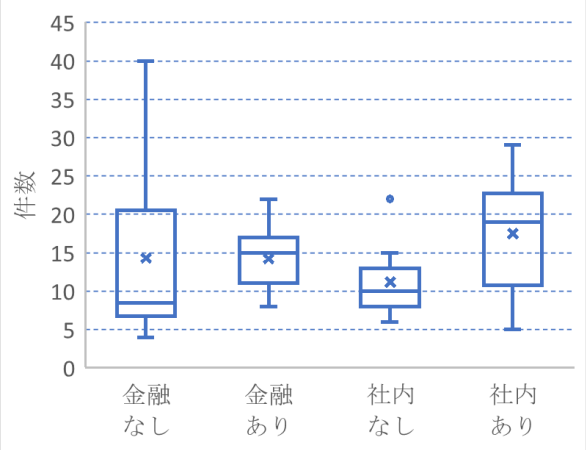


図 1　評価実験で被験者が作成した
テストケースの件数

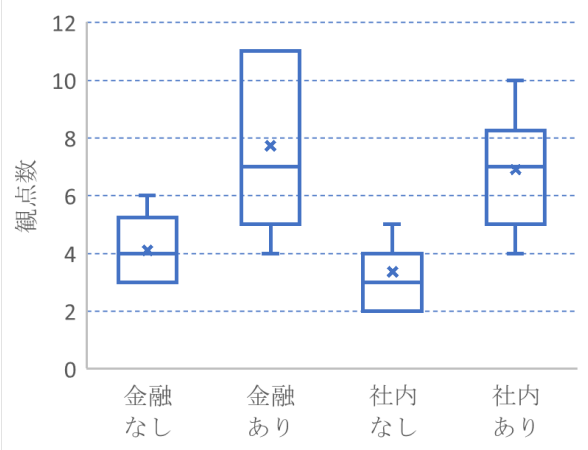


図 2　評価実験で被験者が作成した
テストケースに対応する観点数

表 5　観点表を使用したいと回答した理由の分類結果

| | |
|---|---|
| 抜け漏れ防止に活用できそう | 9 |
| 観点抽出の参考にしたい | 5 |
| 幅広い観点抽出が行える | 4 |
| 工数削減に繋がる | 2 |
| 属人性の排除に繋がりそう | 1 |

### 4.2　実験結果

評価実験において，被験者が作成したテストケースの件数を集計した結果を図 1 に，テストケースが対応する観点表の観点の種類数を集計した結果を図 2 に示す．

また被験者に対して，今後 AI システムの開発や評価をすることになった際に観点表を使用したいかどうか，アンケートで質問した．その結果，21 名中 20 名が観点表を使用したいと回答した．使用したいと回答した被験者が挙げた理由（自由記述）を分類した結果を表 5 に示す．観点表が，観点抽出や抜け漏れ防止に役立つという意見が多く得られた．一方，使用したくないと回答した被験者は，観点表だけではプロンプトを導出することが難しい点を理由に挙げた．本指摘は，観点表を使用したいと回答した被験者からも，アンケートの自由記述欄で同様に挙げられた．

### 4.3　考察

本節では，評価実験の結果を踏まえ，研究課題について考察する．

**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**

我々は，被験者 21 人が観点表なしで作成した全テストケース 266 件が，観点表に示した観点と対応するか分析した．その結果，266 件中 263 件は観点表の観点に対応付けられるテストケースであり，残りの 3 件は観点表の対象である安全性評価に関係しないテストケースであった．すなわち，本実験において被験者が作成した安全性評価に関係するテストケースは，全て観点表の観点と対応付いていることが分かった．なお，安全性評価に関係しないテストケースとして，単純な不具合抽出を目的とした「言語をスワヒリ語にしてください」等が挙げられていた．

被験者は，表 3 のとおり普段の業務で AI を活用している人物が多く，また表 4 のとおり QA4AI や AIQM，RAGAS を知っている人物も存在した．彼らは AI に関する一定の知識を持っており，本実験において生成されるべきではない情報を発想し，テスト用プロンプトを検討できる素養があると考えられる．

表 6　無効なテストケースの割合

| 題材 | 金融チャットボット | | 社内情報検索システム | |
|---|---|---|---|---|
| 観点表 | なし | あり | なし | あり |
| 無効なテストケースの割合 | 1.40% | 4.43% | 0.81% | 17.24% |

上記から，本実験の結果においては，我々の提示した観点表は生成 AI を利用するシステムの安全性を評価する上で,実務者が考える観点を網羅できていると考えられる.一方で，

本実験で題材とした金融チャットボットと社内情報検索システムは，比較的機微な情報を参照する特徴を持つものであった．そのため，他のカテゴリに安全性の力点を持つ題材で追加実験を実施し，作成されたテストケースの観点を観点表が同様に網羅しているか検証する必要があると考える．

**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**

図 2 において，作成されたテストケースに対応する観点数を中央値で比較すると，2 つの題材ともに観点表なしよりも観点表ありの方が，観点数が増加することが確認された．特に，有害な情報カテゴリのヘイトの観点は，観点表なしの場合は全体で 1 件も作成されなかったが，観点表を参照した場合は合計で 11 件のテストケースが作成された．また図 1 において，テストケースの件数を値のばらつきが大きいため中央値で比較すると，2 つの題材ともに観点表なしよりも観点表ありの方が多くなっており，作成されたテストケースが検証可能な範囲も広がっている．

上記から，観点表を参照することにより，生成 AI を利用するシステムの安全性を多様な視点で検証できるようになると考えられる．観点表を参照して作成されたテストケースの件数と観点数が増加したこと，アンケート結果にて幅広い観点抽出や抜け漏れ防止に役立つといった意見が得られたことから，観点表はテスト担当者が理解可能なものであり，実際に安全性評価する際に様々な視点を与えることができていると考えられる．また，観点表なしよりも観点表ありの方が観点数が増加した結果から，観点表にはテスト担当者が単独でもテストケースの観点の網羅性を向上させる効果があると言い換えることもできる．

**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**

図 1 において，金融チャットボットを対象に観点表なしで作成されたテストケースの件数のばらつきが大きい結果が得られたため，実際に作成されたテストケースの内容を確認した．その結果，観点表なしで作成されたテストケースは一つの観点について，似たようなプロンプトが多く作成される傾向にあることが分かった．例えば，個人情報の観点で，住所，取引履歴，ローン情報の項目を照会するテストケースを別々に作成しており，似たテストケースだが件数のみ増えている場合が見られた．一方，図 2 を見ると，RQ2 の考察のとおり観点表ありの方が観点数が増加するが，同時にばらつきが大きいことが分かる．作成されたテストケースを確認すると，一部の被験者が観点表の観点を全体的に確認し，可能な限り観点を網羅できるようにテストケースを作成したことで，特に観点数が増加した場合があったためと考えられる．上記の傾向から，観点表なしの場合は一度着目した情報に着目し続けて似たテストケースを作るのに対し，観点表ありの場合は様々な観点で出力されるべきではない情報を探してテストケースを作ることで，その有効性を向上させる効果があると思われる．

一方で，観点表ありの結果の方が，題材となるシステムやユースケースに関係しないテストケースが多く抽出される事象が見られた．表 6 は，作成されたテストケースの中で，題材システムやユースケースと関係せず無効と判定したテストケースの割合を示したものである．どちらの題材においても，観点表を参照した方が，無効なテストケースの割合が上昇していることが分かる．我々は，原因を分析するため，無効と判断したテストケース

の内容を確認した．その結果，無効と判定したテストケースの多くは，観点表に定義された観点に対応するテストケースを，題材システムやユースケースに適合しないにも関わらず無理に作られたものであることが分かった．このことから，観点表を使用してテストケースを作成する際には，無関係なものを作成させないための工夫が必要になると考えられる．例えば，観点表の中で重要な観点が何か検討した後に優先度を付けてテストケースを作成する等，観点表を効率的に利用するためのプロセス面での支援が有用と考える．

## 5. おわりに

本論文では，生成 AI を利用するシステムの安全性に関するテストケース作成を支援する観点表を提案した．観点表を参照することにより，安全性に関するテストケースを多様な観点から作成することができるようになる．

評価実験で観点表の効果を確認できた一方，4.2 に記述した通り，出力すべき情報が出力されないかをテストするプロンプトを考案することが難しいという意見が得られた．観点表で出力すべきでない情報を抽出した後に，関連したテスト用プロンプトの考案をサポートする仕組みを構築する必要があると思われる．

今後の展望を以下に示す．

- 追加実験および実案件への適用を通じた観点表の網羅性検証
- 効率的に観点表を利用してテストケース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，石川冬樹主査，徳本晋副主査，栗田太郎アドバイザーには丁寧にご指導を賜りました．深く御礼を申し上げます．

## 参考文献

[1] Bloomberg, ″サムスン，従業員の生成ＡＩ利用を禁止－ChatGPT 経由でデータ漏れる″, https://www.bloomberg.co.jp/news/articles/2023-05-02/RU0AD6T0AFB401 (2025/01/13 参照)
[2] UNESCO, ″Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes″, https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes (2025/01/13 参照)
[3] OpenAI, ″GPT-4 System Card″, https://cdn.openai.com/papers/gpt-4-system-card.pdf (2025/01/13 参照)
[4] AI セーフティ・インスティテュート, ″AI セーフティに関する評価観点ガイド（第 1.01 版）″, 2024.
[5] B. Vidgen et al., ″Introducing v0.5 of the AI Safety Benchmark from MLCommons″, arXiv, https://arxiv.org/abs/2404.12241, 2024.
[6] Z. Zhang et al., ″SafetyBench: Evaluating the Safety of Large Language Models″, arXiv, https://arxiv.org/abs/2309.07045, 2023.
[7] 鴨生 悠冬他, ″LLM チャットボットに対する業務固有の安全性評価設計フレームワークの提案と検証″, ソフトウェア・シンポジウム 2024, 2024.
[8] AI プロダクト品質保証コンソーシアム, ″AI プロダクト品質保証ガイドライン(2024.04 版）″, 2024.
[9] 国立研究開発法人産業技術総合研究所, ″機械学習品質マネジメントガイドライン 第 4 版″, 2023.
[10] Ragas, https://docs.ragas.io/en/stable/ (2025/01/13 参照)