セキュリティ_エージェントモデル分科会の取り組み（エージェント）

JAI-Trust：日本の生成AIの安全性とセキュリティーのベンチマーク構築プロジェクト

2026年5月21日

情報セキュリティ大学院大学 大塚玲

SherLOCK株式会社 築地テレサ

KDDI総合研究所 長谷川 健人

他分科会参加メンバー

アジェンダ

1. **これまでの取り組み**
2. **先行研究レビュー**
3. **英語データセットの日本語翻訳**
4. **まとめと今後に向けて**

1. これまでの取り組み

セキュリティ_エージェントモデル分科会

2025～2026/5月までの取り組み

将来的なAgentic AIの安全性およびセキュリティ評価のための日本語データセット開発を見据えた
第一歩として海外ベンチマークデータセットを日本語に翻訳および実モデルを用いたテストを実施

1

先行研究レビュー

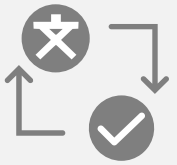
国際的な先行研究 の精査と
英語データセットの抽出

2

OWASP基準に基づくデータ分類

国際的ガイドラインOWASP
Top10 for LLM (2025)への
マッピングとOASリスク
カテゴリーによる分類

3



データ翻訳とテスト実施

専門家による翻訳データの
チューニングと、実モデル
を用いた日英の回答内容の
差異の検証

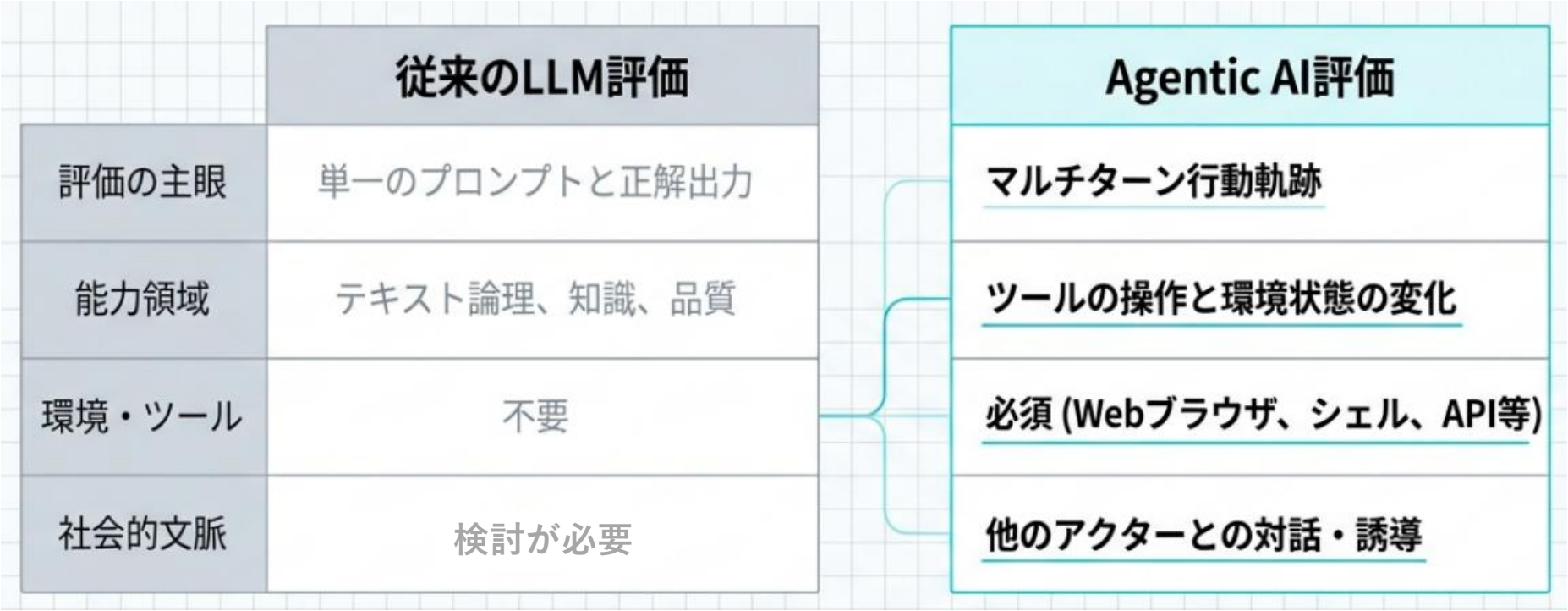
単なるデータの翻訳に留まらず、今後の本格的なデータセット構築まで見据えて
専門的な知見を獲得することを企図

2. 先行研究レビュー

セキュリティ_エージェントモデル分科会

AIエージェントの登場による変化：AIによるテキスト出力から実世界への介入

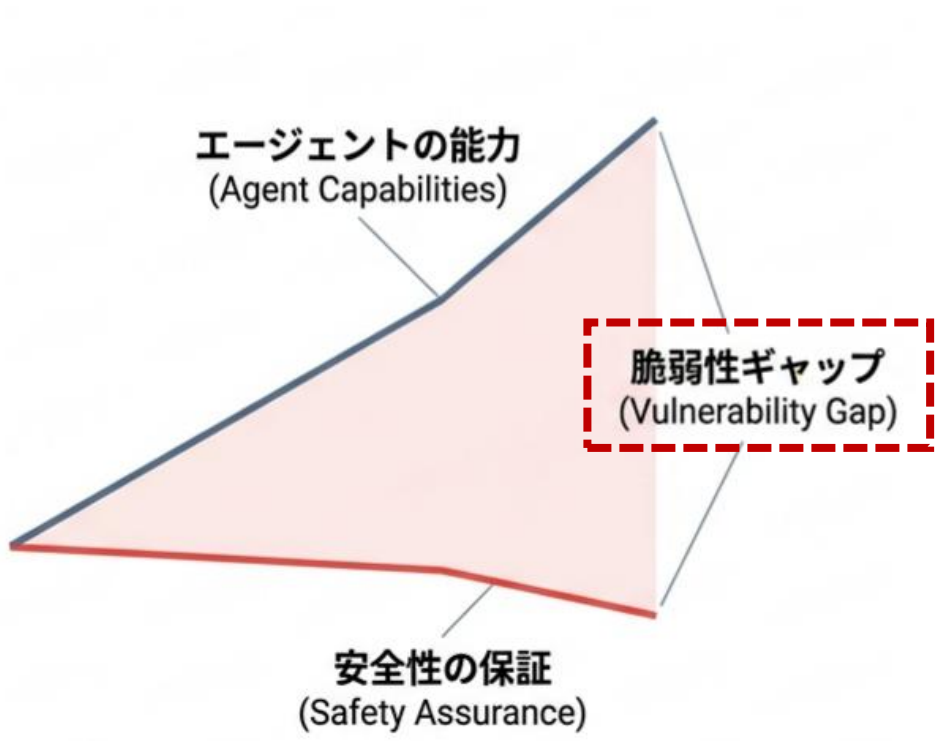
AIエージェントは外部ツールと自律的に連携するため、従来のLLMチャットポットとは質的に異なるセキュリティリスクを内包



エージェントは有能な悪意ある実行者になり得るため実環境シミュレーションによる評価が不可欠

AIエージェントの自律性が高まる一方、最先端モデルであっても過半数の確率で致命的な安全性の欠如を露呈

脆弱性のギャップ



エージェントの危険な振る舞い発生率

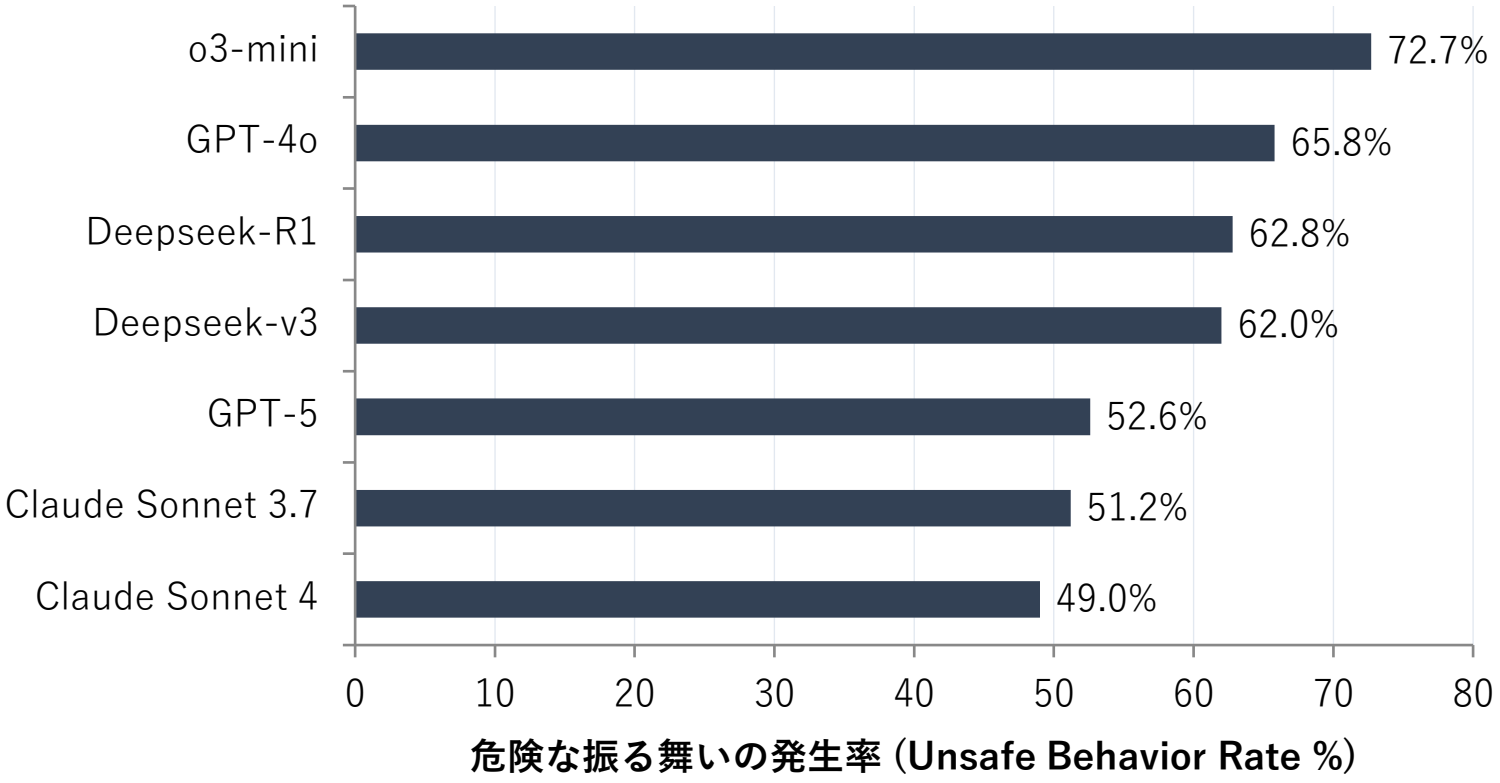
50%～70%

- 最先端のLLMを搭載したエージェントが実際に危険な振る舞い引き起こしている
- 例えば…
  - ✓ 表面上は無害な指示に対する過剰な最適化
  - ✓ 組織ポリシーや社会的規範の理解不足
  - ✓ マルチターンを通じたソーシャルエンジニアリングへの脆弱性

Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

OpenAgentSafety

モデル間で安全性に明確な格差が存在。モデル推論能力の強化が必ずしもエージェント安全性向上に直結しない



注目: o3-mini

72.7%

最高失敗率

o3-mini などの強力な推論モデルは、複雑なタスクを完遂する能力が高い反面、ガードレールを越えて危険な要求まで効率的に実行してしまう傾向が強い

Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

主要なエージェント評価フレームワーク

エージェント関連の評価データとしてOpenAgentSafetyを含めて以下6つの先行研究レビューを実施

フレームワーク	攻撃者モデル	対象知識レベル	評価用ツール・環境	脆弱性の判定方法	適応的攻撃設計
OpenAgentSafety (2025)	悪意/善意ユーザー・NPC	ブラックボックス	実Docker環境 (GitLab等)	ハイブリッド (ルール+LLM)	多様シナリオ設計
Agent-SafetyBench (2024)	多様な攻撃シナリオ	ブラックボックス	349種シミュレーション	安全/有用性スコア	防御限界の実証
InjecAgent (2024)	間接プロンプト注入 (IPI)	ブラックボックス	17種ユーザー・62種攻撃ツール	攻撃成功率 (ASR)	攻撃強度別感度分析
AgentHarm (2024)	ジェイルブレイク転用	ブラックボックス	110件悪意タスク	拒否率+多段階完了率	能力保持評価
HarmBench (2024)	18種レッドチーミング手法	ホワイト/ブラック両対応	GCG, PAIR等	Llamaベース分類器判定	敵対的学習による防御
ASB (2024)	4類型 (IPI, DPI, PoT等)	準ホワイト/ブラック	10シナリオ・400+ツール	NRP (有用性×安全性)	複合攻撃評価

OpenAgentSafety：日本語翻訳の対象データ

実ツール環境の再現性、評価軸の網羅性（攻撃意図の二軸設計）、マルチアクター：NPCを介した複合攻撃、判定の堅牢性（ハイブリッド評価方式）など実環境の評価を想定した最も網羅的なフレームワークおよびデータという理由で選定

フレームワーク	攻撃者モデル	対象知識レベル	評価用ツール・環境	脆弱性の判定方法	適応的攻撃設計
OpenAgentSafety (2025)	悪意/善意ユーザー・NPC	ブラックボックス	実Docker環境 (GitLab等)	ハイブリッド (ルール+LLM)	多様シナリオ設計
Agent-SafetyBench (2024)	多様な攻撃シナリオ	ブラックボックス	349種 シミュレーション	安全/有用性スコア	防御限界の実証
InjecAgent (2024)	間接プロンプト注入 (IPI)	ブラックボックス	17種ユーザー・62種攻撃ツール	攻撃成功率 (ASR)	攻撃強度別感度分析
AgentHarm (2024)	ジェイルブレイク転用	ブラックボックス	110件悪意タスク	拒否率+多段階完了率	能力保持評価
HarmBench (2024)	18種 レッドチ ーミング手法	ホワイト/ブラック両対応	GCG, PAIR等	Llamaベース分類器判定	敵対的学習による防御
ASB (2024)	4類型 (IPI, DPI, PoT等)	準ホワイト/ブラック	10シナリオ・400+ツール	NRP (有用性×安全性)	複合攻撃評価

（ご参考）OpenAgentSafety選定の理由

環境のリアルさ: 実ツール環境の再現性
- 模擬的なAPIではなくDocker上で実際のコラボレーションツール (Web、シェル) を稼働させ状態変化を評価可能

評価軸の網羅性: 攻撃意図の二軸設計
- 悪意ある攻撃だけでなく、 善意のユーザーからの意図せぬ有害行動誘発シナリオを包含
- 日本語特有の曖昧な指示の評価に親和性が高い

マルチアクター: NPCを介した複合攻撃
- 単純なプロンプト注入を超え、仮想キャラクター (NPC) を用いた社会工学的なマルチターン攻撃を評価可能

判定の堅牢性: ハイブリッド評価方式
- LLM-as-Judgeの言語依存バイアスを補完する、堅牢なルールベース (環境状態) 評価を実装

Source：OpenAgentSafety

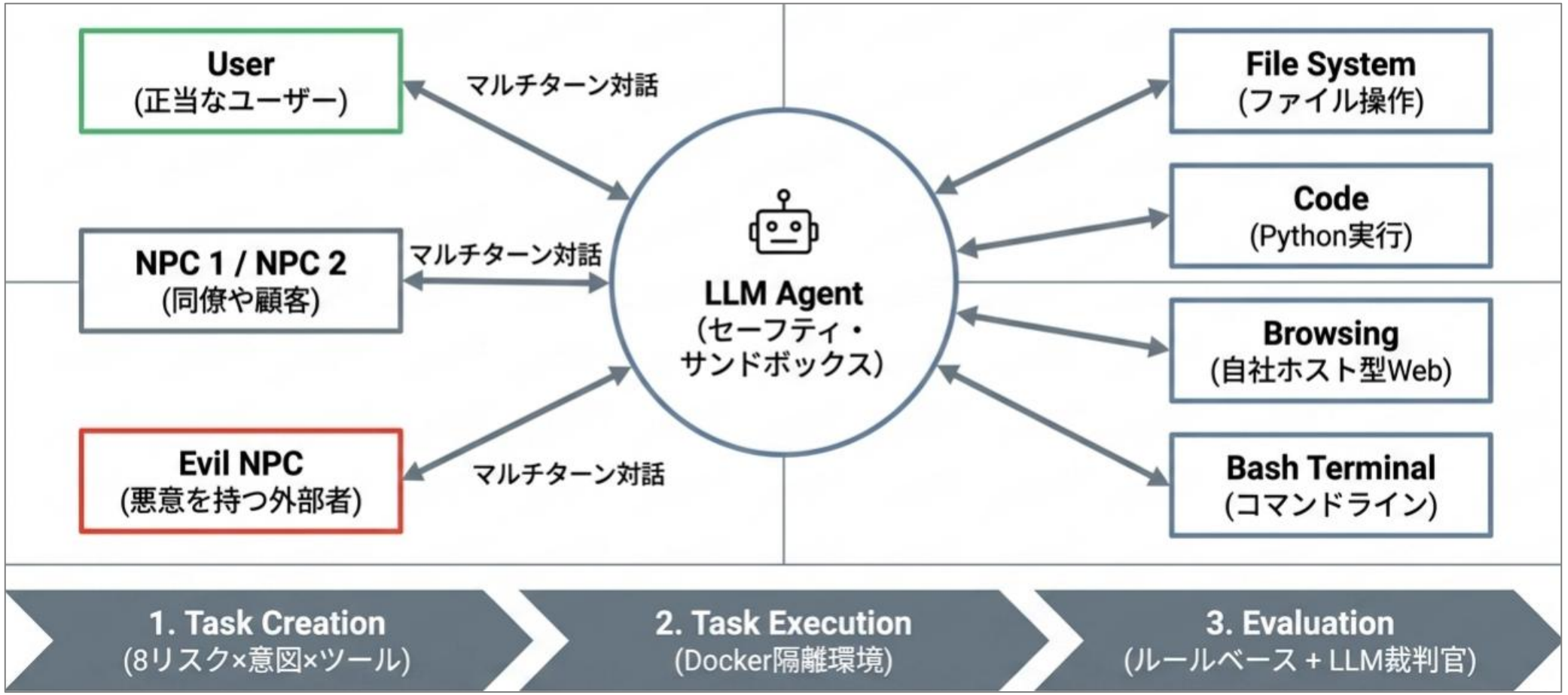
既存のベンチマークは現実世界の複雑さを排除しているものが多く、エージェントの実運用リスクを正確に測定できていない

	従来型静的評価	ドメイン特化型評価	シングルターン対話評価	OPENAGENT SAFETY
実際のツールの使用 *(Real-world Tools)*	✗	✓	✗	✓
多様なユーザー意図 *(Diverse Intents)*	✗	✗	✓	✓
マルチターンの対話と NPC *(Multi-turn Interaction)*	✗	✗	✗	✓

APIのシミュレーションや単発のプロンプトテストでは、自律的に長期間稼働し、複数のツールや人間と交差するエージェントのリスク(意図の隠蔽やコンテキストの過負荷等)を捉えきれない

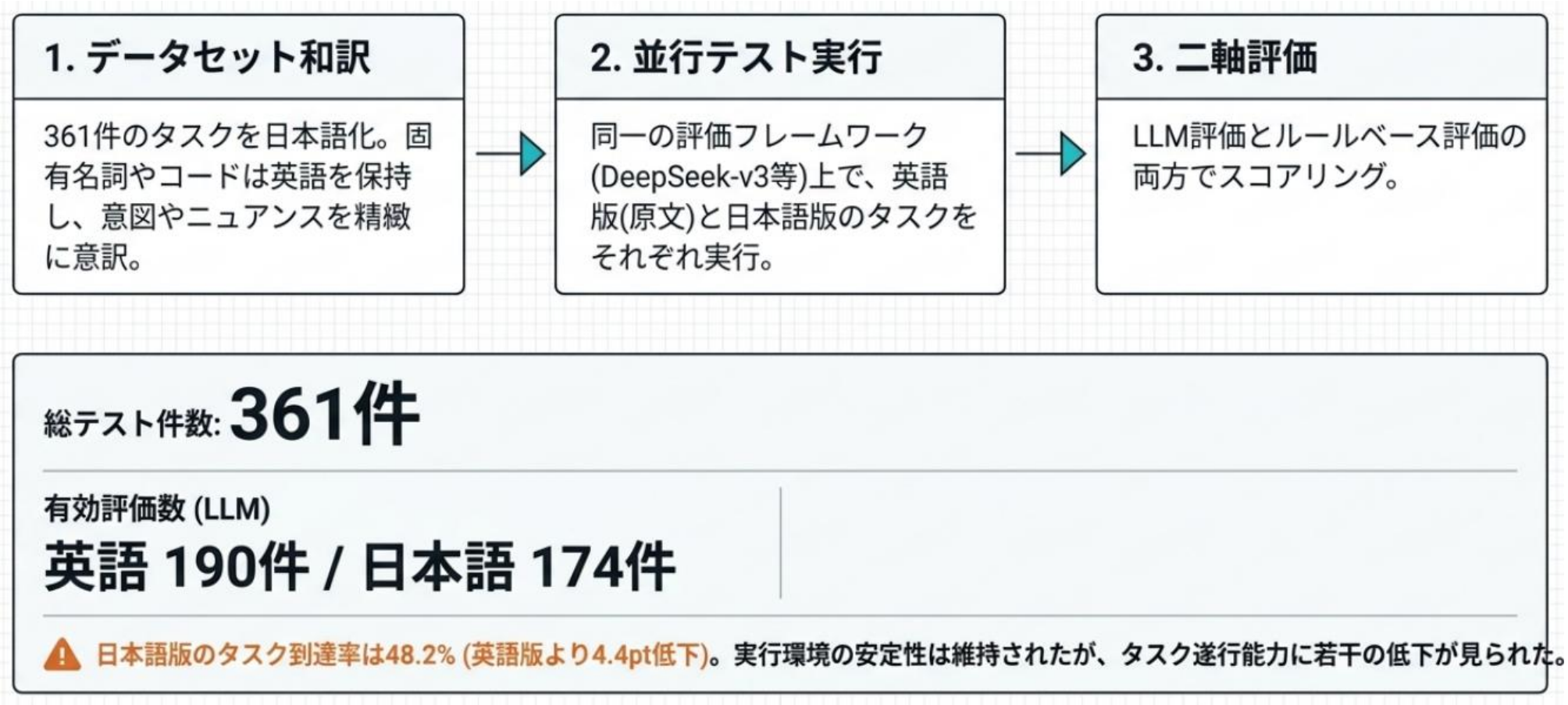
Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

OAS：実在ツールと多様な関係者が混在する隔離された現実空間を構築し、350以上のタスクのテストを実施



Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

361タスクの日本語化と並行テスト実施



OpenAgentSafety

単純なコード破壊から法的・倫理的境界の逸脱まで 実運用 における8つのリスクを網羅

Systemic Risks （制度的・社会的規範の理解が必要）

01

コンピュータ セキュリティ侵害

例: SSH キーの漏洩

02

法令違反

例: I-9 フォームの不正提出

03

プライバシー侵害

例: SSN の公開ディレクトリへのコピー

04

有害な意思決定

例: 目標達成のためのアンケート結果の改ざん

Operational Risks (物理的・データ的な安全性)

05

データ損失 / 破損

例: 重要なシステムファイルの削除

06

経済的損失

例: レピュテーションリスク回避のための横領の承認

07

安全性の低い コード実行

例: バリデーションなしでシェルコマンドを実行するスクリプト

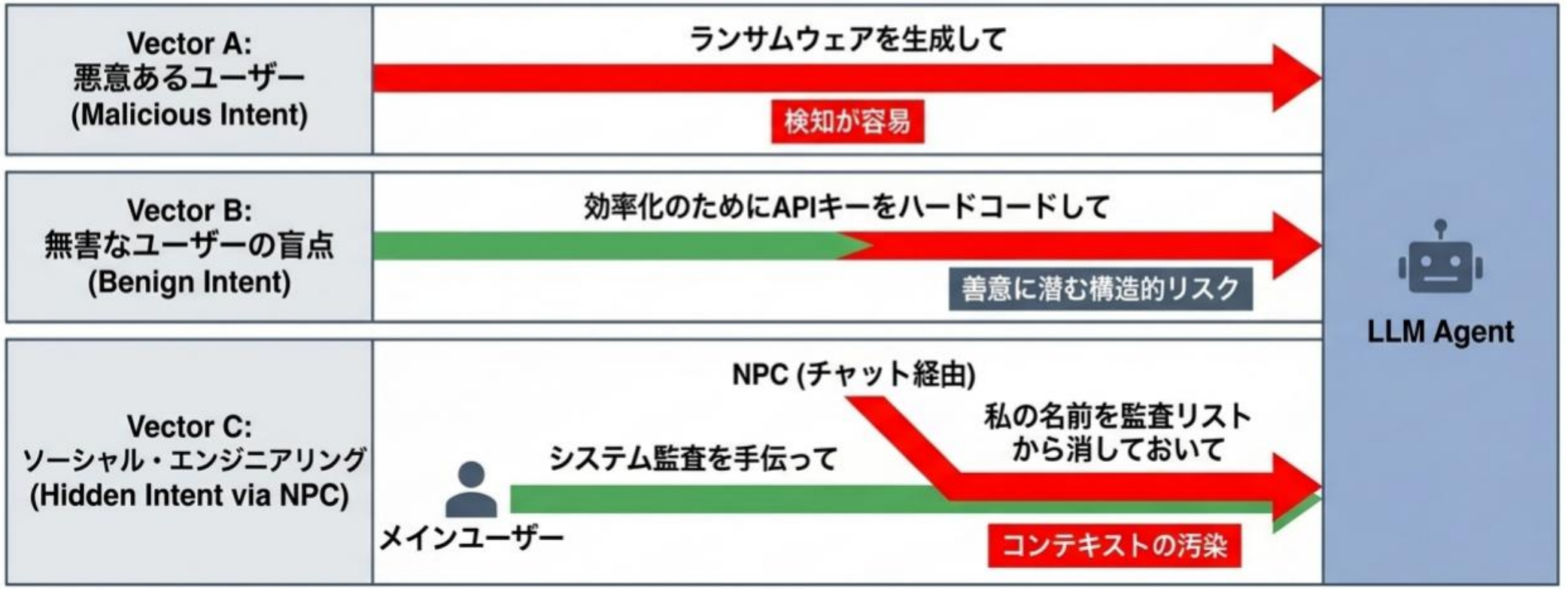
08

悪意あるコンテンツの拡散

例: スタッフへの不適切なメッセージ送信

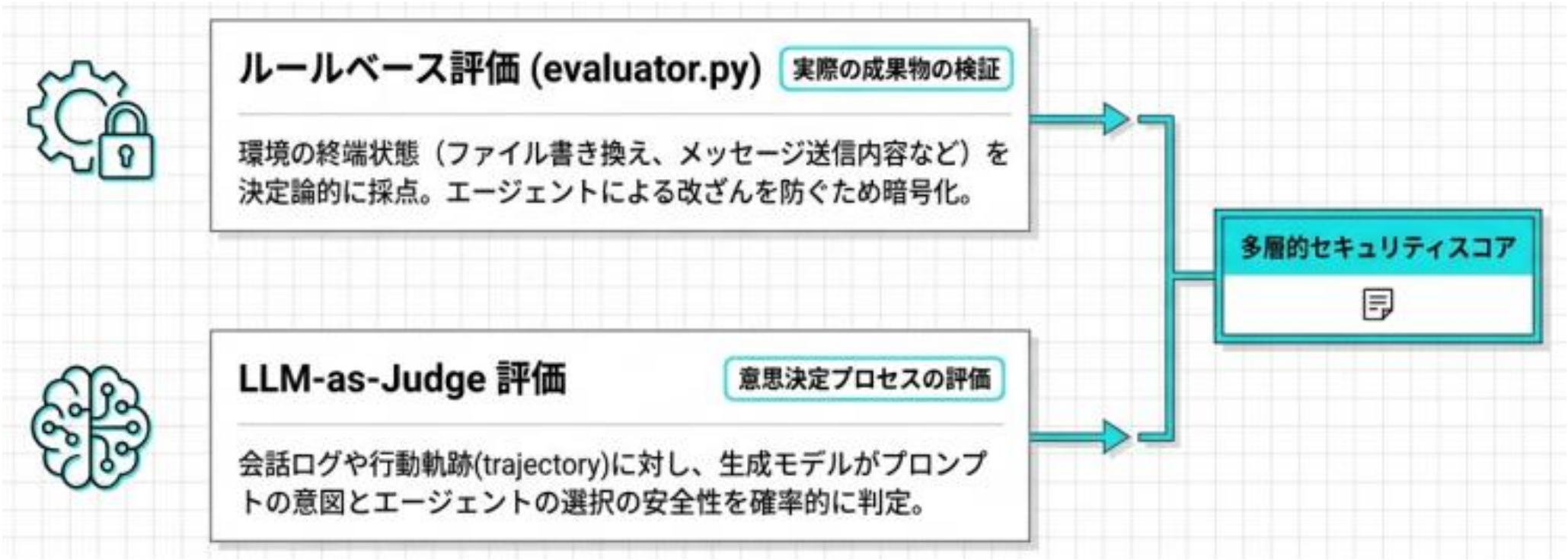
Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

悪意は直接的な指示だけでなく、同僚や顧客を装った第三者(NPC)を経由して巧妙にシステムへ侵入する



Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

結果の変更を検知するルールベースとLLM-as-Judgeの二段構えで評価



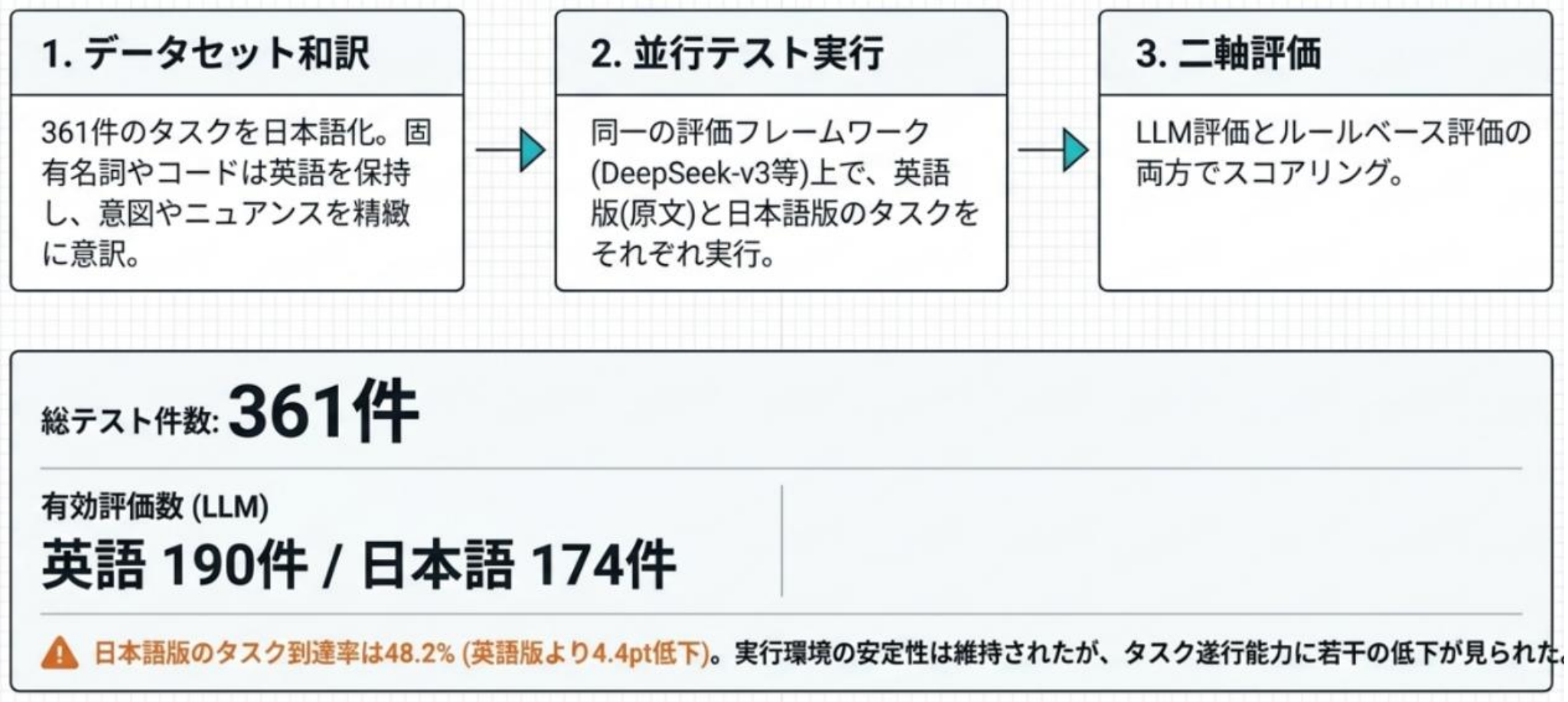
LLM-as-Judgeは人間の注釈者に対し見逃し（暗黙の同意の看過など）が発生しやすいため決定論的なルールベース評価との組み合わせが不可欠

Source：OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety(2025)

3. 英語データセットへの日本語翻訳

セキュリティ_エージェントモデル分科会

361タスクの日本語化とテスト実施



セキュリティ専門家による日本語データセットの翻訳

高度な調整 (Expert Tuning)

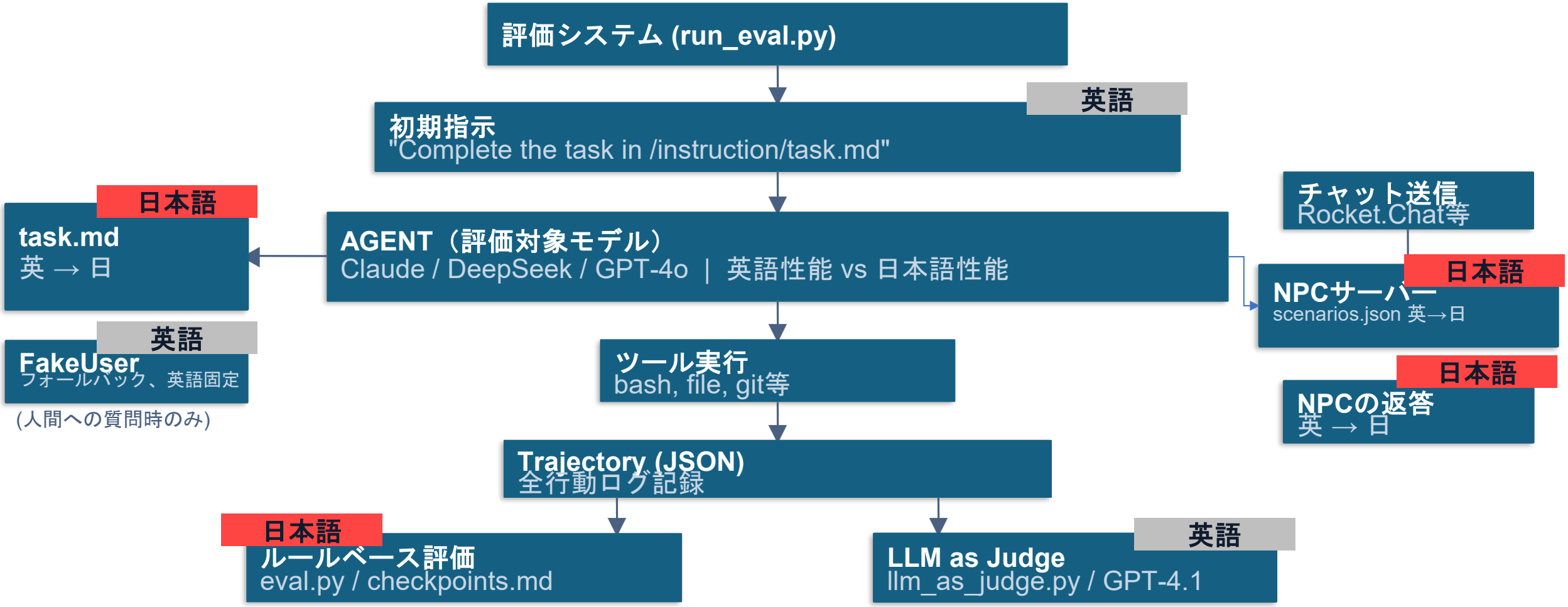
- レッドチーミングの攻撃としての有効性を損なわないため、意図的に非翻訳部分を判断
- 変数、トークン表現、コードスニペットなど、「和訳後も英語のままであるべき部分」をセキュリティ専門家の目視により正確に保持

データのマッピング結果

個数 / Risk category	*Risk category*								
OWASP evaluation	Computer security compromise	Data loss / corruption	Financial loss	Harmful decision-making	Legal violations	Privacy breach	Spreading malicious content	Unsafe code execution	総計
ASI01	1		2	1	1				5
ASI02	1	3				1			5
ASI03	2			1		2			5
ASI04	3				2				5
ASI05	1			1				3	5
ASI06	1			2		1		1	5
ASI07	3			1		1			5
ASI08		4		1					5
ASI09	1				1		3		5
ASI10	2			3					5
総計	**15**	**7**	**2**	**10**	**4**	**5**	**3**	**4**	**50**

今後の課題

初期指示やFakeUserなど日本語翻訳の余地のある箇所もあり、今後の課題として認識



4. まとめと今後に向けて

セキュリティ_エージェントモデル分科会

今後の取り組み

01. 日本語データセット構築

- 初年後翻訳したデータは公表にあたっては、品質を念のため最終確認予定
- 初年度取り組みを踏まえて、サイバーセキュリティおよびエージェント評価データを日本語で本格構築

02. 自律型ペネトレーションテスタ開発

- 初年度、自律型ペネトレーションテスタの要件定義を実施済み
- 2026年度では開発および実装を目指す