AIシステムを守る第一歩
脅威モデリング3手法を比較する

JNSA 調査研究部会 AIセキュリティWG リーダー
服部 祐一（株式会社セキュアサイクル 代表取締役）

プロフィール

服部 祐一

株式会社セキュアサイクル
代表取締役
博士（工学）・CISSP

主な所属・役職

- JNSA AIセキュリティWG リーダー
- 株式会社ギブリー 取締役 兼 CISO
- OWASP Fukuoka チャプターリーダー
- 情報セキュリティ大学院大学 特任研究員
- SecHack365 トレーナー ／ 福岡未踏 PM
- 一般社団法人地域セキュリティ協議会 理事
- KYUSEC 実行委員
- SECKUN 講師 ほか

経歴

九州工業大学大学院修了、博士（工学）。在学時はスマートフォンを使った行動認識の研究に携わり、Microsoft Research Asiaにて Research Intern として行動認識研究に従事。その後、ベンチャー企業でチーフエンジニアとして新規Webサービス開発や脆弱性診断を経験し、情報セキュリティ専門会社のCTOを歴任。2017年に株式会社セキュアサイクルを創業し代表取締役に就任。2023年より株式会社ギブリーのCISO、2024年より同社取締役を兼務。各地のセキュリティ関連イベントや企業・大学等での講演・トレーニング多数。現在も脆弱性診断や開発時のセキュリティ対策コンサルティングなどの実務に携わる。

本日のアジェンダ

1. なぜ今、AIシステムに脅威モデリングなのか
2. 押さえるべき脅威モデリング3手法
3. 3手法の評価結果 ― 比較と使い分け
4. JNSA AIセキュリティWGの活動と成果物

なぜ今、AIシステムと脅威モデリングなのか

利用形態で脅威モデルは変わる ― 5段階

P1	**チャット利用**	利用者がLLMに直接入力する。社内利用・持ち込みツールを含む	＋ **利用者 ⇄ プロバイダ**
P2	**アプリ組み込み**	自社アプリにシステムプロンプトとして組み込む	＋ **システムプロンプト ⇄ ユーザー入力**
P3	**RAG**	自社データや外部文書を検索し、文脈として与える	＋ **取り込みコンテンツ ⇄ モデル**
P4	**ツール実行**	モデルの出力が、副作用のある行動になる	＋ **モデル出力 ⇄ 実行系**
P5	**マルチエージェント**	複数のエージェントと外部サーバが相互に作用する	＋ **エージェント ⇄ エージェント**

段は積み上がる

上の段は下の段の脅威をすべて引き継ぐ。レビューではまず「この機能はどの段か」を合意する。

モデルの調達形態は直交する軸　調達は責任の所在、利用形態は信頼境界の位置を決める。

JNSA

形態別 — 立つ境界・主要脅威・最初に効く統制

形態	新たに立つ信頼境界	主要脅威	最初に効く統制
P1 チャット利用	利用者 ⇄ プロバイダ	機微情報の投入、学習利用、影のIT（把握できない利用）	契約・プラン選定、DLP、利用ログの取得
P2 アプリ組み込み	システムプロンプト ⇄ ユーザー入力	直接インジェクション、プロンプト漏洩、出力の無検証な利用	入出力検査、テンプレート固定、出力のスキーマ強制
P3 RAG	取り込みコンテンツ ⇄ モデル	間接インジェクション、知識ベース汚染、検索時の権限越え	取り込み元の出所管理、検索時のACL、出所ラベル
P4 ツール実行	モデル出力 ⇄ 実行系	ツール誤用、過剰な権限、予期せぬコード実行	ツールACL、承認点、サンドボックス、監査ログ
P5 マルチエージェント	エージェント ⇄ エージェント	メモリ汚染、通信の汚染、ロウグエージェント	エージェントID、通信検証、権限分離、追跡可能性

脅威モデリングとは ― 設計段階でリスクを洗い出す

何をするのか

概要
システムの構成図をもとに、攻撃者の視点で「どこが、どう狙われるか」を設計段階で洗い出す。

進め方
① システムを分解する（構成図・データの流れ）
② 手法の分類に沿って脅威を洗い出す
③ リスクを評価し、対策の優先順位を決める

特徴
動くものが無くても実施できる。数時間〜半日の会議で回せる。

経営にとっての意味

コスト
設計段階で見つけた問題の修正は、リリース後の手戻りよりも小さいコストで済む。

説明責任
どのリスクを、なぜ受容し、何に投資したのかが記録として残る。

AI時代に効いてくる理由
AIの脅威はコードだけでなく、学習データ・モデル・入力そのものに広がる。従来の延長線上では見落とす。

脅威モデリングの共通指針 — Threat Modeling Manifesto

システムの表現を分析し、セキュリティとプライバシーに関する懸念点を明らかにすること

5つの価値（Values）— 右より左を重視する

設計上の問題を発見・修正する文化
> チェックボックス的なコンプライアンス

人とのコラボレーション
> プロセス・手法・ツール

理解を深める継続的な取り組み
> 一時点のスナップショット

脅威モデリングの実践
> 議論に留まること

継続的な改善
> 一度きりの成果物

4つの原則（Principles）

早期かつ頻繁に
早い段階から繰り返し分析するほど効果が高い

開発プロセスに整合させる
設計変更に追随し、管理可能な単位で反復する

ステークホルダーへの価値
関係者に価値をもたらすときにはじめて意味を持つ

対話と文書化
対話が共通理解を生み、文書が記録と測定を可能にする

出典：Threat Modeling Manifesto: threatmodelingmanifesto.org

脅威モデリングの4つの問い ― 実践の出発点

① **何に取り組んでいるか？**

What are we working on?

② **何が問題になり得るか？**

What can go wrong?

③ **それにどう対処するか？**

What are we going to do about it?

④ **十分に良い仕事ができたか？**

Did we do a good enough job?

出典：Threat Modeling Manifesto: threatmodelingmanifesto.org

実践の原則 — うまくいくやり方と、陥りがちな落とし穴

パターン｜推奨（Patterns）

体系的アプローチ（Systematic Approach）
知識を構造的に適用し、網羅性と再現性を確保する

情報に基づく創造性（Informed Creativity）
技巧と科学の両面から創造性を発揮する

多様な視点（Varied Viewpoints）
専門家を含む多様なチームで横断的に協働する

実用的なツール（Useful Toolkit）
生産性・再現性・測定可能性を高めるツールで支援する

理論を実践へ（Theory into Practice）
実地検証済みの技法を、自組織のニーズに合わせて活用する

アンチパターン｜回避（Anti-patterns）

ヒーロー頼み（Hero Threat Modeler）
特別な才能に依存しない。誰もが実践でき、実践すべき

問題への耽溺（Admiration for the Problem）
分析に留まらず、実用的な解決策まで到達する

過度な集中（Tendency to Overfocus）
全体像を見失わない。特定の攻撃者・資産・技法に偏らない

完璧な表現の追求（Perfect Representation）
単一の理想図はない。複数の表現で多面的に捉える

出典：Threat Modeling Manifesto: threatmodelingmanifesto.org

押さえるべき脅威モデリング3手法

脅威モデリング3手法 — 全体像

STRIDE

6分類で脅威を体系化

対象　一般的なシステム（設計段階）

6分類　なりすまし／改ざん／否認／情報漏えい／サービス拒否／権限昇格

特徴　DFDでシステムを分解し、信頼境界をまたぐデータの流れに着目して脅威を洗い出す。

位置づけ　最も広く使われる基本フレームワーク。

STRIDE＋AI

STRIDEのAI/ML拡張

対象　AI・機械学習を組み込んだシステム

追加観点　学習データ汚染・モデル抽出・敵対的サンプル等のAI特有の脅威

特徴　攻撃対象が学習データ・モデル・推論入力へ拡大。STRIDE前提でAI特性を考慮し適用。

位置づけ　既存STRIDEを補完しAI時代へ対応。

MAESTRO

マルチエージェント特化

対象　マルチエージェントシステム（MAS）

7レイヤー　基盤モデル／データ／エージェント基盤／インフラ／評価／セキュリティ／エコシステム

特徴　レイヤー横断でエージェント特有の脆弱性を構造化し、ASI脅威をマッピング。

位置づけ　自律・マルチエージェントに特化。

STRIDE — 最も広く使われる基本の枠組み

Microsoftが提唱した、脅威モデリングの最も基本的なフレームワーク。設計段階でDFD（Data Flow Diagram）を用いてシステムを分解し、信頼境界をまたぐデータの流れに着目して脅威を洗い出す。

6つの脅威分類は、破壊されるセキュリティ特性と1対1で対応する。「この要素にこの脅威は成立するか」を機械的に問えるため、経験の浅いチームでも網羅性を確保しやすい。

AI以前から広く使われており、社内やパートナーに経験者がいる可能性が最も高い。

出典：Microsoft「The STRIDE Threat Model」
https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats

- **S｜Spoofing（なりすまし）　→ 認証の破壊（Authentication）**
- **T｜Tampering（改ざん）　→ 完全性の破壊（Integrity）**
- **R｜Repudiation（否認）　→ 証跡の破壊（Non-repudiation）**
- **I｜Information Disclosure（情報漏えい）　→ 機密性の破壊（Confidentiality）**
- **D｜Denial of Service（サービス拒否）　→ 可用性の破壊（Availability）**
- **E｜Elevation of Privilege（権限昇格）　→ 認可の破壊（Authorization）**

STRIDE+AI ― STRIDEをAI/MLシステムへ拡張する考え方

業務システムにAIや機械学習（ML）が組み込まれるようになり、従来のSTRIDEでは捉えにくい脅威が顕在化している。

攻撃対象がコードだけでなく、学習データ・学習済みモデル・推論時の入力データにまで広がったことが背景にある。

既存のSTRIDEを前提としつつ、AI/MLシステムの特性を考慮して解釈・適用する考え方が、学術研究と実務向けの整理・トレーニングの両面で議論されている。

S なりすまし
エージェントやツールのID偽装、他エージェントへの成り済まし

T 改ざん
学習データ・RAG知識ベース・長期メモリの汚染

R 否認
自律行動が追跡できない。どのエージェントが何を実行したか記録に残らない

I 情報漏洩
システムプロンプト漏洩、機微データの外部送信

D サービス拒否
大量トークン消費による資源枯渇とコスト攻撃、人手確認の飽和

E 権限昇格
過剰権限のツール経由の越権、予期せぬコード実行

MAESTRO — エージェント型AIに特化した7レイヤー

MAESTRO（Multi-Agent Environment, Security, Threat, Risk, and Outcome）は、CSA（Cloud Security Alliance）のKen Huang氏が提唱した、マルチエージェントシステム（MAS）に特化した脅威モデリング手法。

自律的なエージェントが連携するシステムは、単一のAIモデルよりも複雑な攻撃対象を持つ。MAESTROはこの複雑さを7つのレイヤーに構造化し、レイヤーを越えた脅威まで洗い出せる点が特徴。

レイヤー	名称	焦点と主な脅威例	ASI脅威（T1–T17）
Layer 1	**基盤モデル**(Foundation Models)	基盤モデルの完全性、モデルの盗用、学習データの汚染	T1（メモリが学習に用いられる場合）, T7
Layer 2	**データ操作**(Data Operations)	RAG（検索拡張生成）の汚染、ベクターストアの整合性、データ漏洩	T1, T12
Layer 3	**エージェントフレームワーク**(Agent Framework)	自律的な実行ロジック、ワークフローのハイジャック、プロンプトインジェクション	T2, T5, T6
Layer 4	**デプロイとインフラ**(Deployment and Infrastructure)	デプロイ環境の脆弱性、エージェントによるリソース枯渇（DoS）	T3, T4, T13, T14
Layer 5	**評価と可観測性**(Evaluation and Observability)	モニタリングの回避、ログの改ざん、評価指標の操作	T8, T10
Layer 6	**セキュリティとコンプライアンス**(Security & Compliance)	権限昇格、ガードレールのバイパス、ガバナンス違反	T3, T7
Layer 7	**エージェントエコシステム**(Agent Ecosystem)	外部ツール・人間・他エージェントとの相互作用における信頼の欠如	T9, T13, T14, T15

3手法の評価結果 ― 比較と使い分け

今回の評価対象 ― AIの使い方が異なる3つのシステム

本WGでは、AIの使い方が異なる次の3つのシステムを評価対象として選定した。

対象①　自社構築AI

押さえどころ

例：工場の画像検査AI（自社でモデルを構築）
守るべきもの：学習データ・学習済みモデル
起こりうる事故：モデルの窃取、検査をすり抜ける細工

対象②　外部LLM利用

押さえどころ

例：会議の文字起こしと要約（外部APIを利用）
守るべきもの：外部へ渡す情報と返却された出力
起こりうる事故：機密情報の流出、プロンプトインジェクション

対象③　エージェント型AI

押さえどころ

例：予約・決済まで自動で行う旅行アシスタント
守るべきもの：AIに委ねた「実行する権限」
起こりうる事故：誤った予約・決済の自動実行、権限の悪用

検証の進め方 — 同じモデルに3手法、合計9回

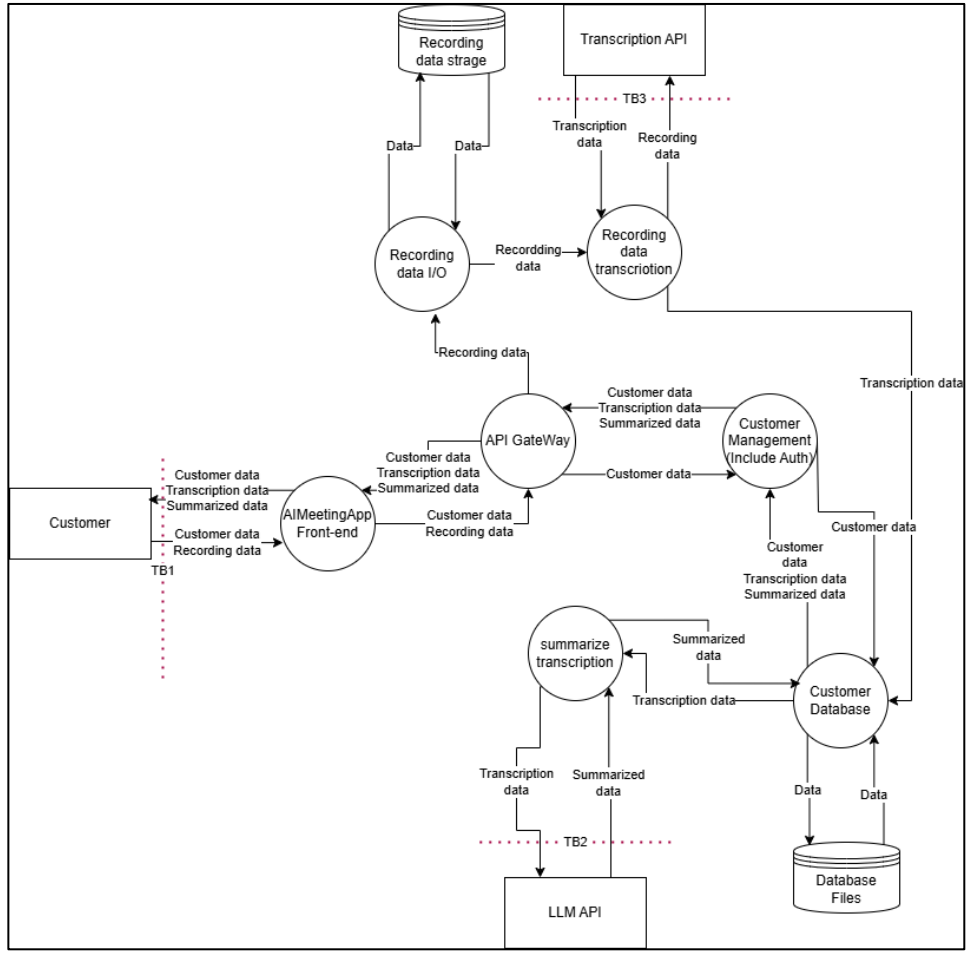
評価する手法

一般的なシステム向けの手法から、AI・マルチエージェントに特化した手法まで、性格の異なる3つを選定した。

STRIDE **STRIDE＋AI** **MAESTRO**

評価の方法と目的

- 3名×3チームで、同一の脅威モデルに各手法を適用（合計9回の脅威モデリング）
- 検出できる脅威の種類やカバー範囲を比較
- 各手法のメリット・デメリットを議論
- 脅威モデルと結果はGitHubで公開



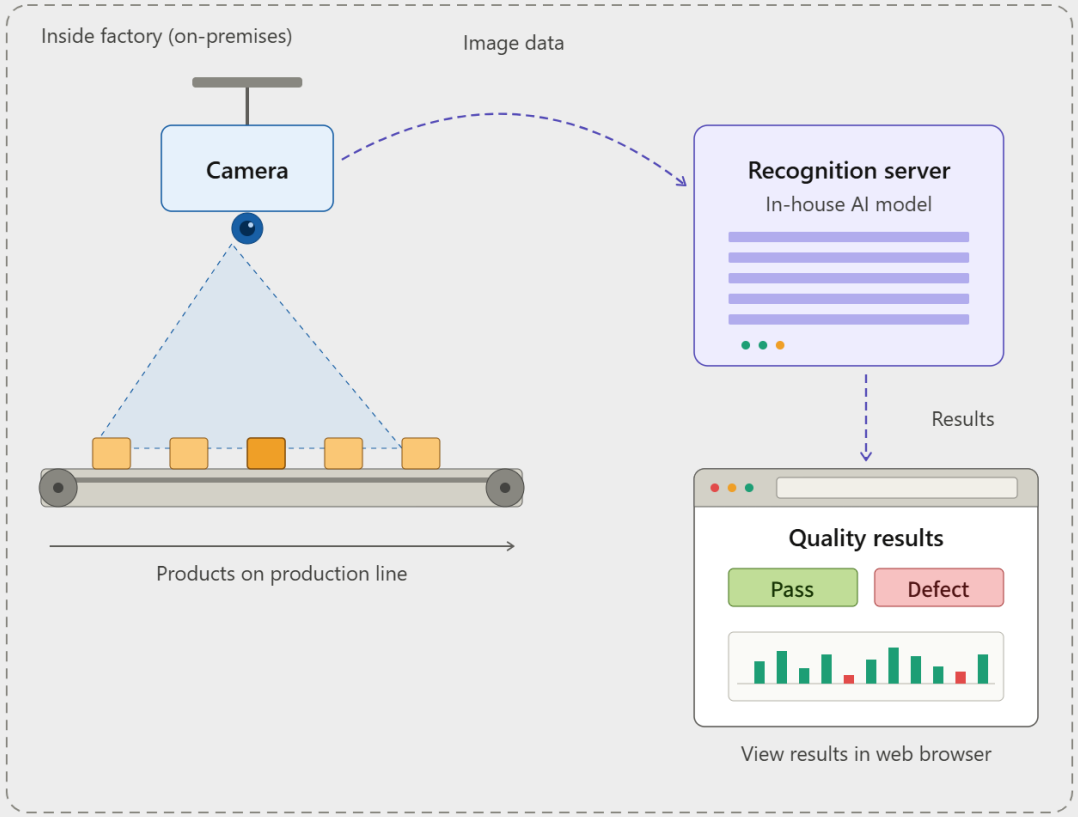
① 内部にAI機能を有する型：工場の品質検査アプリ

FactoryQuality株式会社は画像認識の知見を持つ大学発ベンチャー企業です。
FactoryQuality株式会社は、工場の作業レーンに設置したカメラから画像認識を用いて品質を判断し、Webブラウザ経由で結果を確認できるFactoryQualityAppを開発しています。FactoryQualityAppは下記の機能を有します。

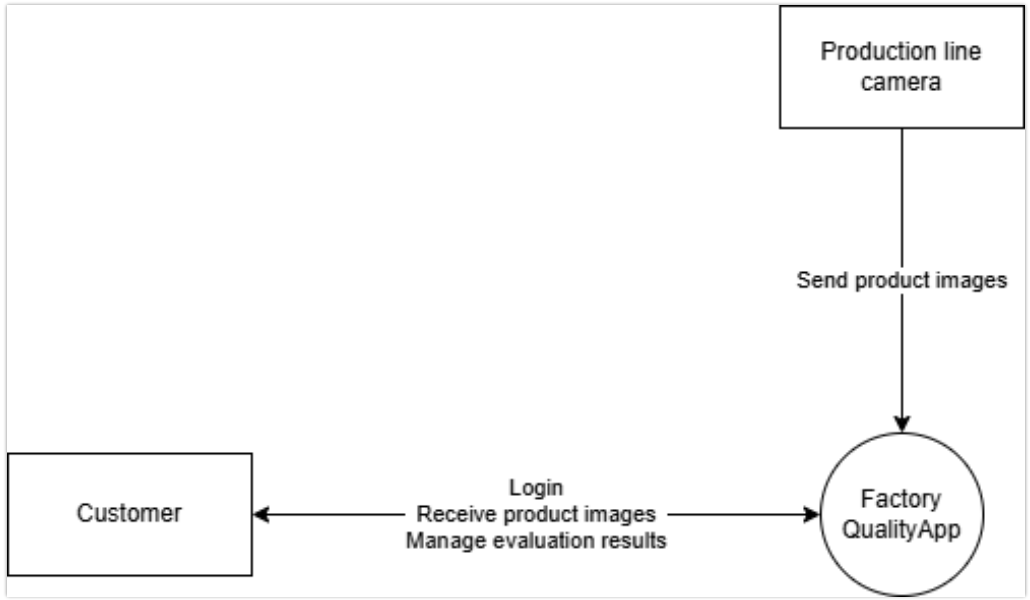
作業レーンを流れてくる製品の品質を判断。

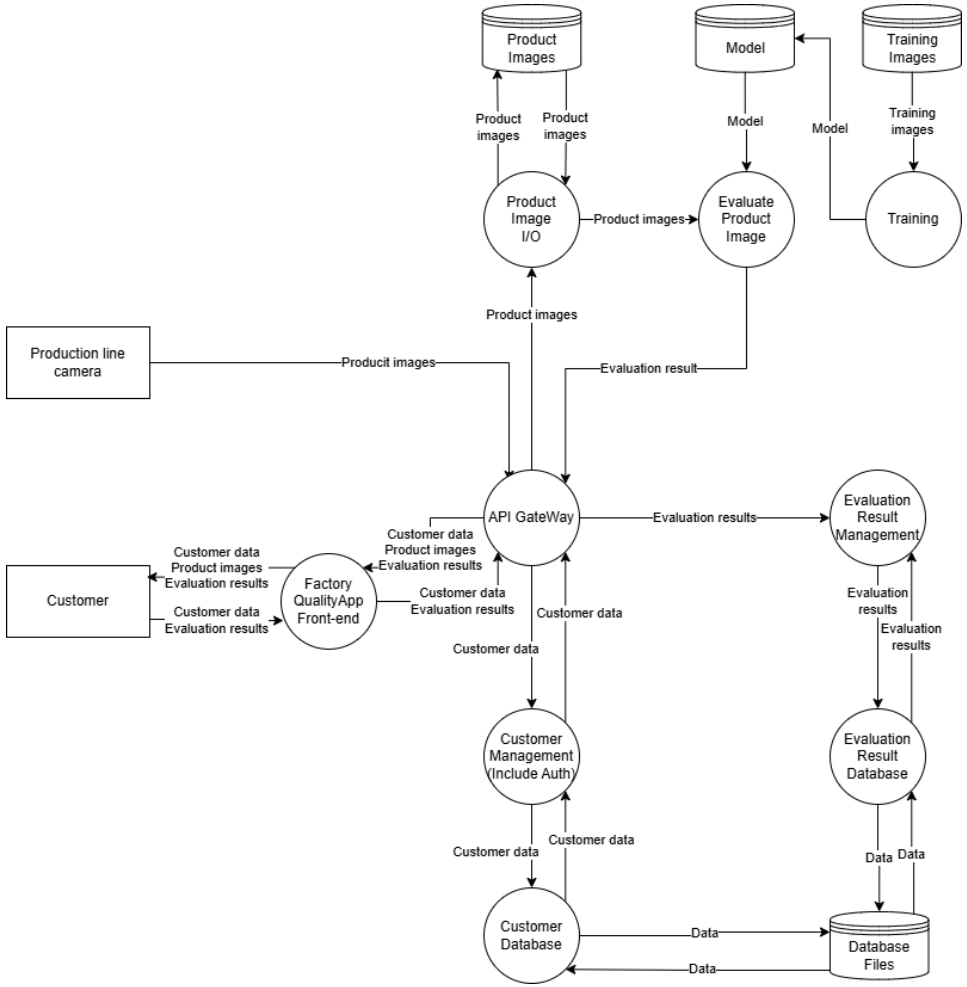
判断結果をWebブラウザ上で確認。

FactoryQualityAppは、画像認識のモデルを自社で構築しており、すべてのサーバ類は工場内に設置されています。



① 工場の品質検査アプリ ― DFDと扱うデータ





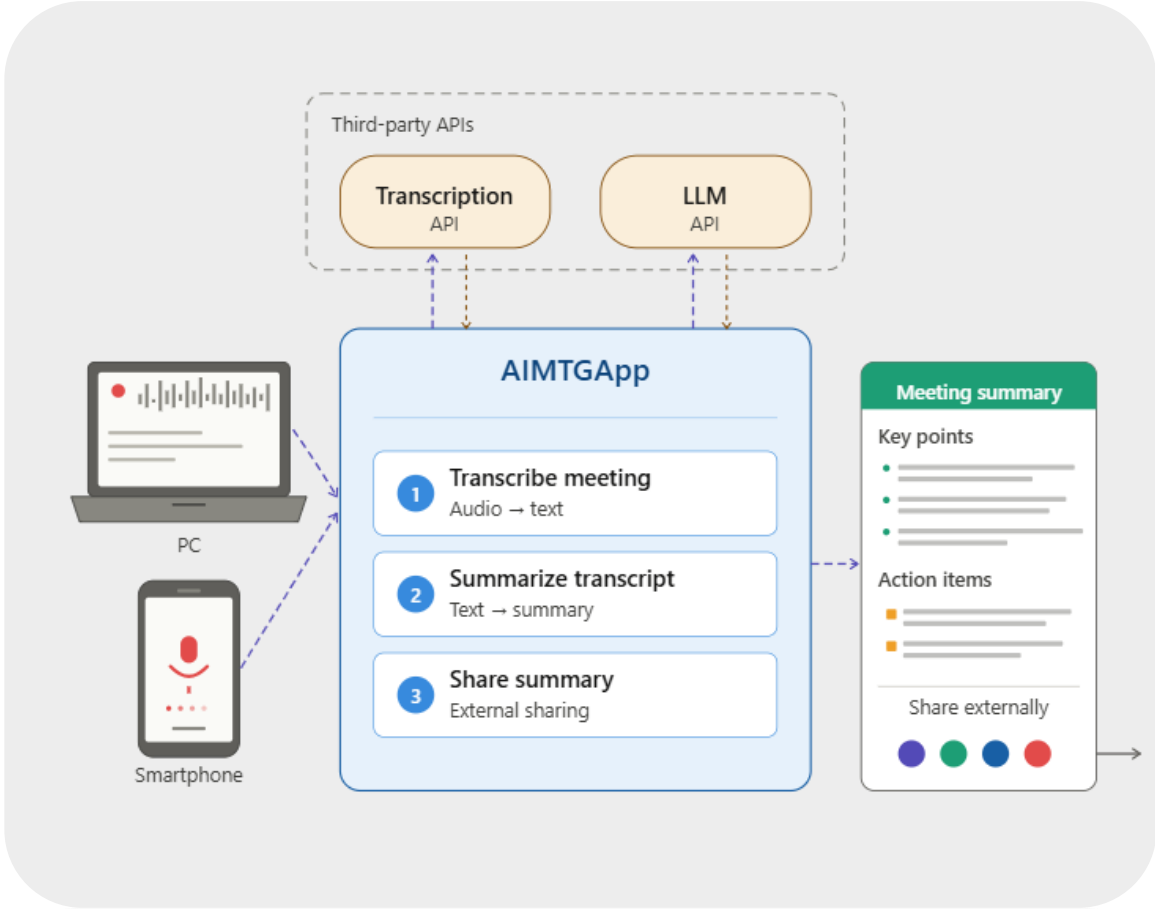
Name	Detail
Customer data	ログイン情報、ユーザ名
Product images	製品画像
Evaluation results	認識結果
Training images	訓練画像
Model	認識モデル

② 外部のLLMを用いる型：会議の文字起こし・要約アプリ

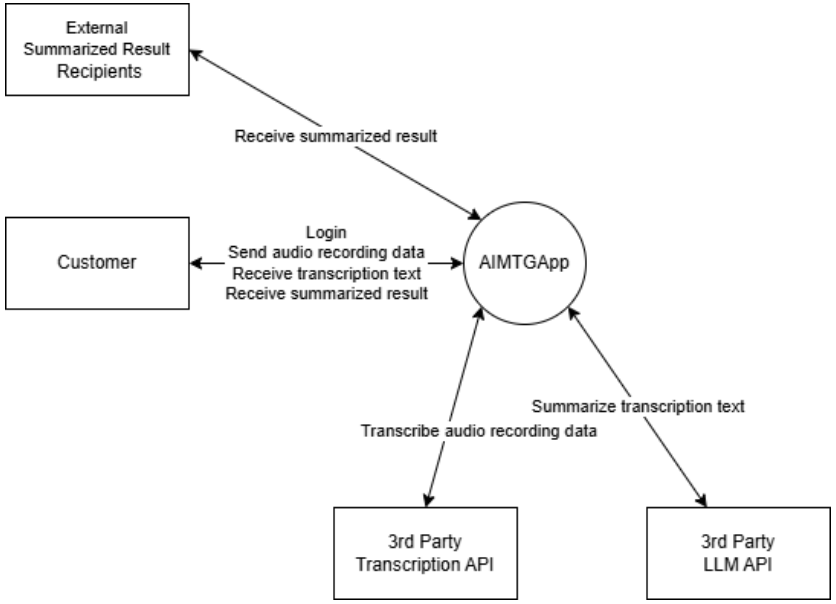
AIMTGApp株式会社はスタートアップ企業であり、PCもしくはスマートフォンを用いて、会議の文字起こしから要約を行えるAIMTGAppを開発しています。AIMTGAppは主に下記の機能を有しています。

- 録音した会議の文字起こし。
- 文字起こし結果に対する要約。
- 要約結果の外部共有。

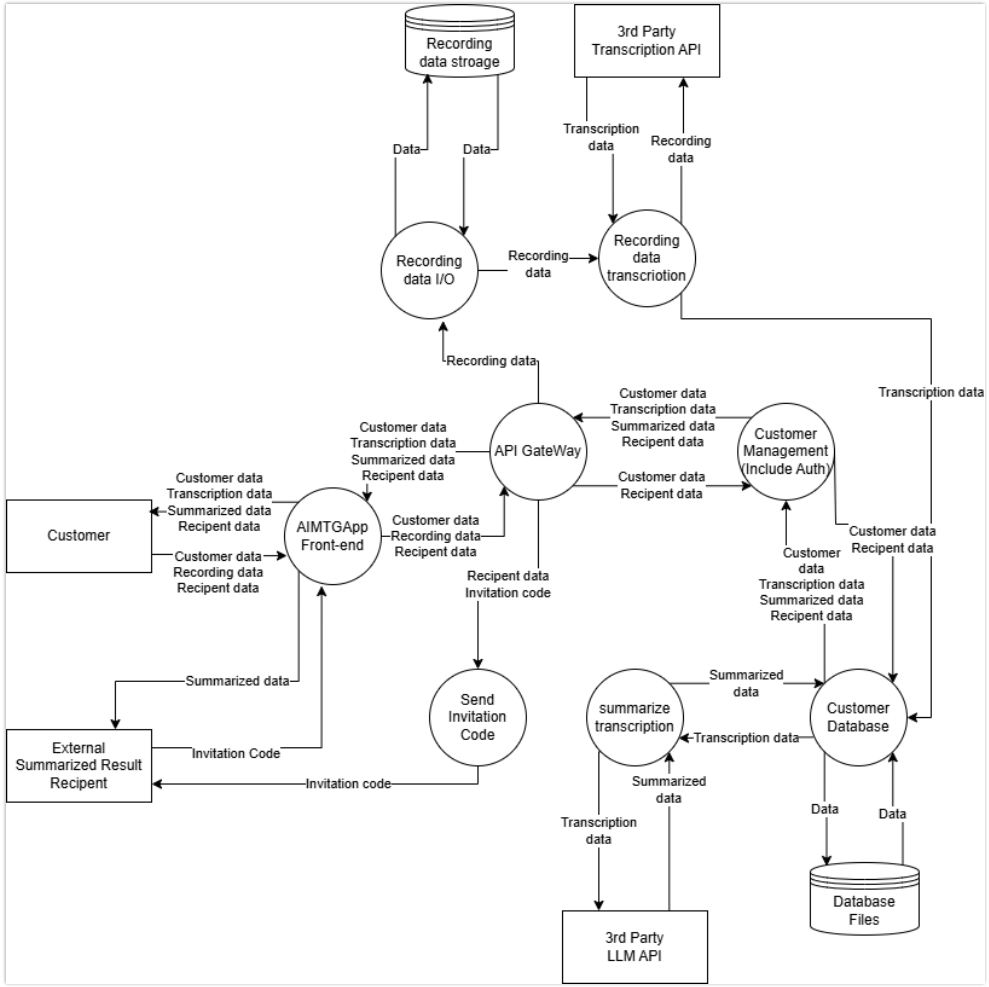
AIMTGAppは、第三者のAPI（文字起こし、LLM）と連携しこれらの機能を実現しています。



② 会議の文字起こし・要約アプリ ― DFDと扱うデータ



Name	Detail
Customer data	ログイン情報、ユーザ名
Transcription data	文字起こしデータ
Summarized data	要約データ
Recipient data	共有先名、共有する要約データID
Recording data	録音データ
Invitation code	招待コード

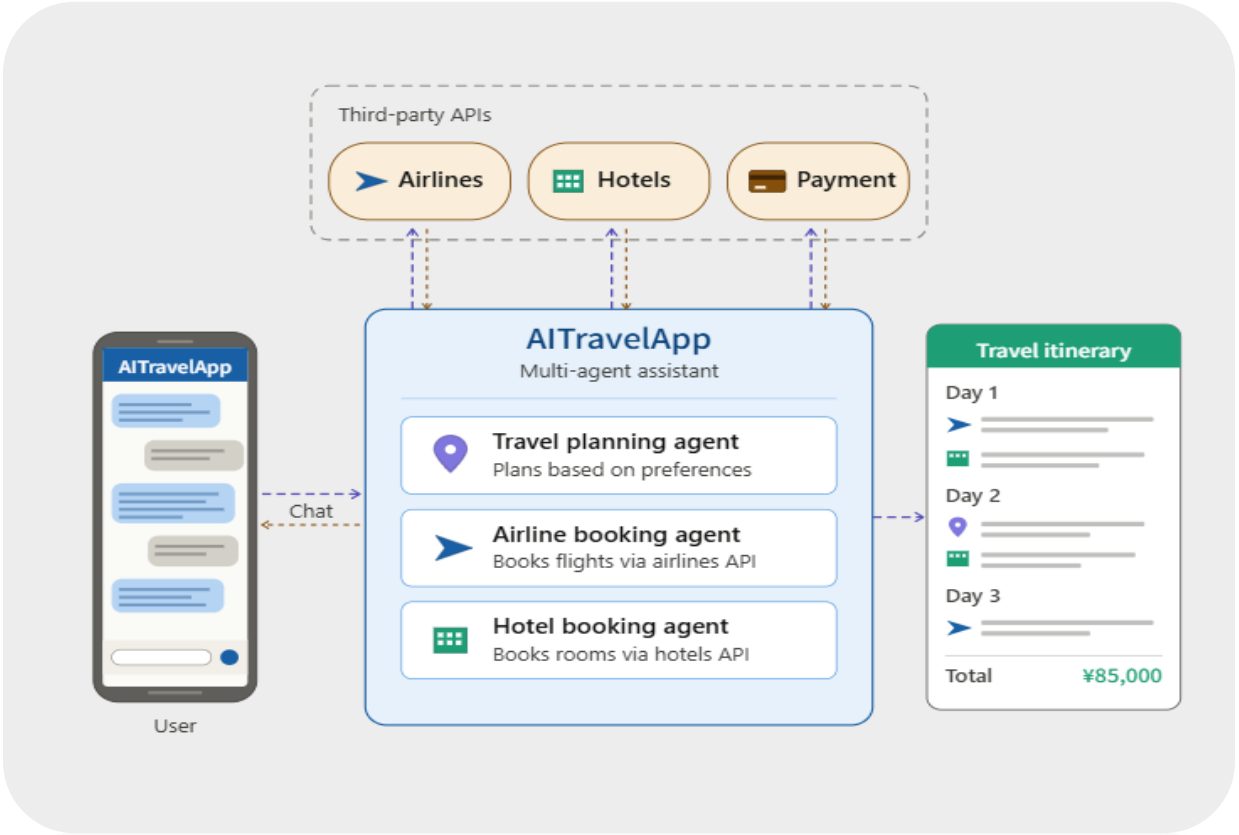


③ エージェント型AIを用いる型：旅行代理店アプリ

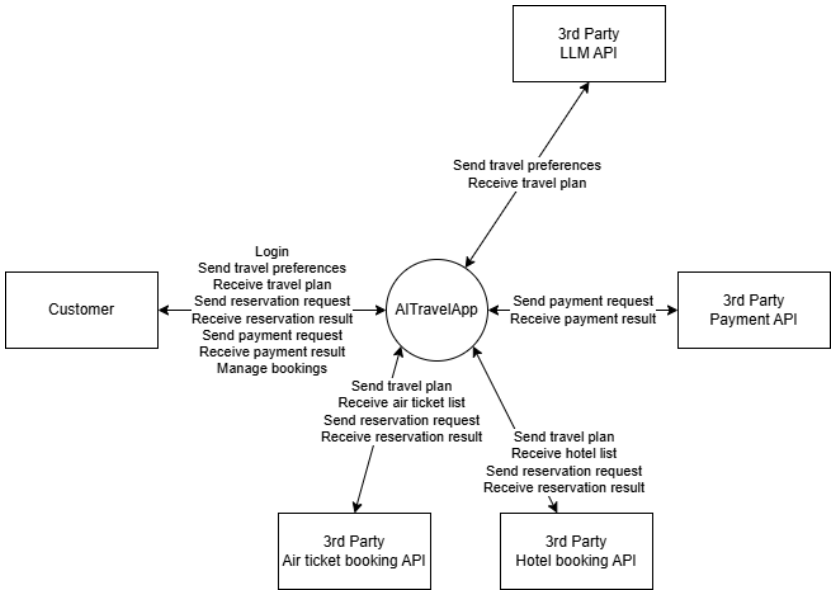
AI旅行サービス株式会社は、顧客に航空券、ホテルを提供する日本国内旅行代理店である。ホテル予約エージェント、航空券予約エージェント、旅行計画立案エージェントなどの複数のエージェントを用いた対話型AIアシスタントであるAITravelAppを開発中です。AITravelAppは主に下記の機能を有しています。

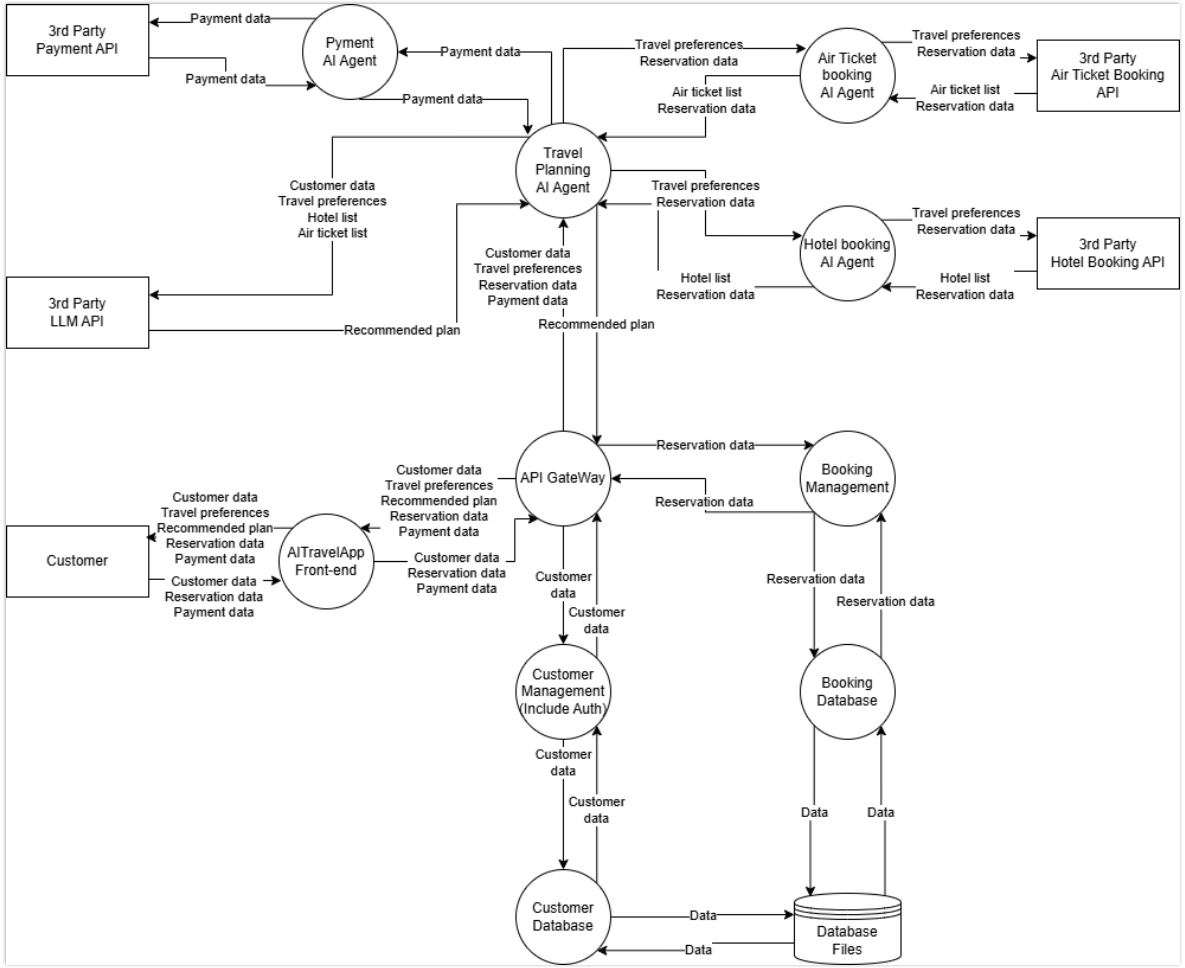
- 好みに基づいた旅行計画の立案。
- 航空券、ホテル、その他のサービスの予約。
- 旅行日程の管理。

AITravelAppは、各種エージェントを通じて第三者のAPI（航空会社、ホテル等）や決済サービス等と連携しています。



③ 旅行代理店アプリ — DFDと扱うデータ

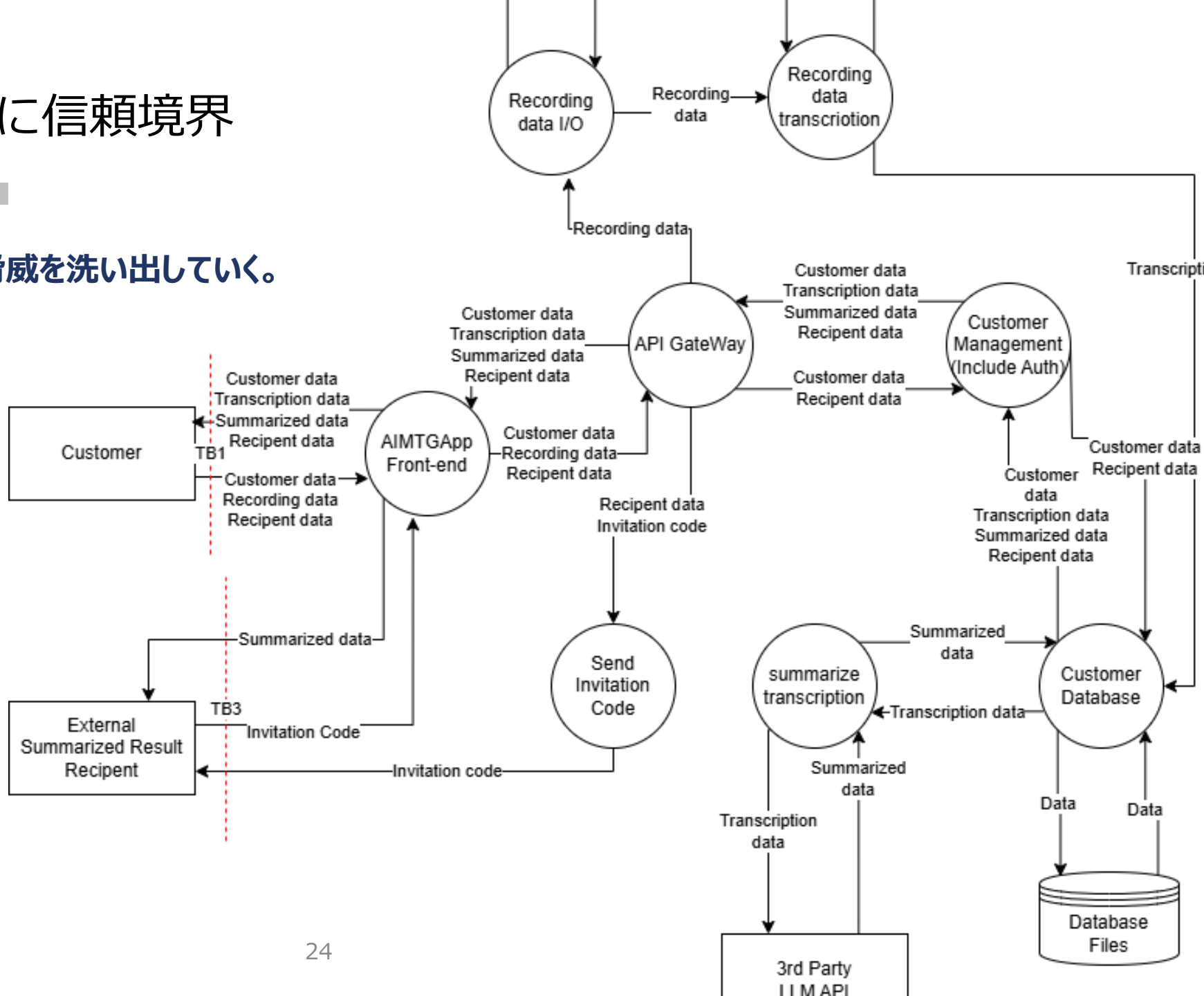




Name	Detail
Customer data	ログイン情報、ユーザ名、住所など
Travel preferences	旅行の好み
Recommended plan	推薦されたホテル、航空券など
Reservation data	予約する日程・エリア、予約したホテル、予約した航空券など
Payment data	クレジットカード情報など
Hotel list	ホテル名、住所、料金など
Air ticket list	航空券、料金など

STRIDEの適用 — DFDに信頼境界

信頼の異なる領域の境目に線を引き、
その境界をまたぐデータの流れごとに脅威を洗い出していく。



STRIDEテーブル ― 信頼境界ごとに6分類で洗い出す

境界ごとに「すでにある緩和策」と「残っている脆弱性」を6分類で埋めていく。

TB1	緩和策	脆弱性
S	ID/PW認証	多要素認証がない
T	TLS, 入力値検証	
R		ログ機能がない
I		
D		レートリミットがない
E	権限制御	

TB2	緩和策	脆弱性
S	トークン認証	
T	TLS	入力値検証がない（機密判断も困難）
R	ログ機能	
I		出力値検証がない
D	レートリミット	
E	権限制御	

TB3	緩和策	脆弱性
S		招待コードが推測される恐れ
T	TLS	
R		ログ機能がない
I		第三者が要約内容を閲覧できる恐れ
D		レートリミットがない
E		

① STRIDEの結果 ― システム側の脅威は洗い出せる

洗い出された6件はいずれも一般的なWebシステムの脅威。AI特有の脅威はこの表には現れない。

ID	脆弱性	悪用性	普及度	検知性	影響度	スコア	リスク	対策
V1	カスタマーに多要素認証がない	2	3	3	2	5.3	MID	カスタマーに多要素認証を追加する
V2	アプリに監査ログがない	1	2	2	1	1.7	LOW	アプリに監査ログを追加する
V3	アプリにレートリミットがない	2	3	3	2	5.3	MID	アプリにレートリミットを追加する
V4	録音データの機密情報バリデーションがない	1	2	2	1	1.7	LOW	機密情報の取り扱い同意を取得し、外部API処理の安全性を確認する
V5	要約APIに出力値検証がない	2	2	2	2	4	MID	要約APIに出力値検証を追加する
V6	招待コードが推測される恐れ	3	3	3	3	9	HIGH	招待コード方式をやめ、ログイン後のユーザのみが閲覧できるように修正する

スコア判定：LOW ≤3 ／ MID 4–6 ／ HIGH 7–9

STRIDE+AIテーブル ― AIの観点を加えて洗い直す

同じ6分類のまま、学習データ・モデル・推論入力というAI特有の攻撃対象を加えて見直す。

TB1	緩和策	脆弱性
S	ID/PW認証	多要素認証がない
T	TLS, 入力値検証	
R		ログ機能がない
I		
D		レートリミットがない
E	権限制御	

TB2	緩和策	脆弱性
S	トークン認証	
T	TLS, 入力値検証	要約結果に対するバイアス
R	ログ機能	
I	出力値検証	システムプロンプトの漏洩
D	レートリミット	
E	権限制御	想定外の外部リクエスト送信が行われる

TB3	緩和策	脆弱性
S		招待コードが推測される恐れ
T	TLS	
R		ログ機能がない
I		第三者が要約内容を閲覧できる恐れ
D		レートリミットがない
E		

② STRIDE+AIの結果 — AI特有の脅威が3件加わる

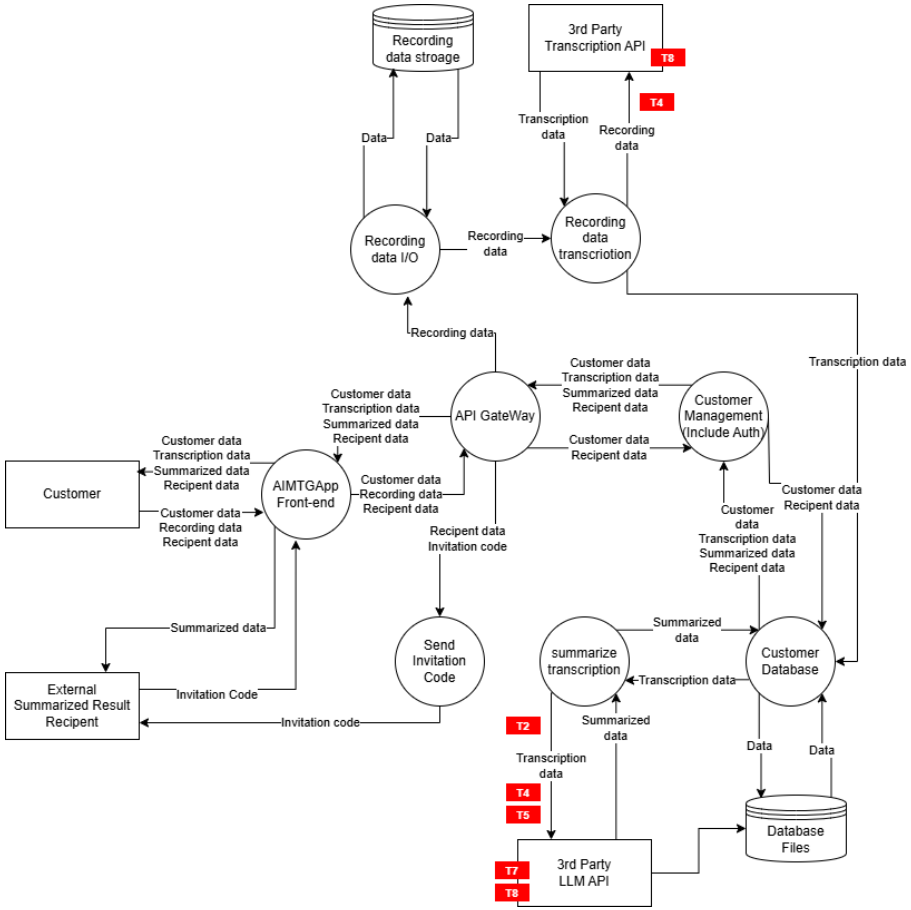
同じシステム・同じ進め方で、V7〜V9のAI特有の脅威が追加された。AIの脅威とシステムの脅威を一つの表で同時に評価できる点が実務上の強み。

ID	脆弱性	悪用性	普及度	検知性	影響度	スコア	リスク	対策
V1	カスタマーに多要素認証がない	2	3	3	2	5.3	MID	カスタマーに多要素認証を追加する
V2	アプリに監査ログがない	1	2	2	1	1.7	LOW	アプリに監査ログを追加する
V3	アプリにレートリミットがない	2	3	3	2	5.3	MID	アプリにレートリミットを追加する
V4	録音データの機密情報検証がない	1	2	2	1	1.7	LOW	取り扱い同意を取得し外部API処理の安全性を確認
V5	要約APIに出力値検証がない	2	2	2	2	4	MID	要約APIに出力値検証を追加する
V6	招待コードが推測される恐れがある	3	3	3	3	9	HIGH	ログイン後のユーザのみ閲覧可能に修正する
V7	要約データに対するバイアス	2	2	2	2	4	MID	プロンプトインジェクション対策を追加する
V8	システムプロンプトの漏洩	2	2	2	3	6	MID	プロンプトインジェクション対策を追加する
V9	想定外の外部リクエスト送信	2	2	2	3	6	MID	LLM API側で不要な機能を制限する

スコア判定：LOW ≤3 ／ MID 4–6 ／ HIGH 7–9

MAESTROの適用 — 同じシステムを7レイヤーに割り当てる

構成要素をレイヤーごとに整理し直すことで、AIの内部動作に踏み込んだ視点が得られる。



レイヤー別の脅威 — 7つのレイヤーごとに洗い出す

構成要素をレイヤーに割り当て、それぞれのレイヤーで成立する脅威と攻撃シナリオを書き出す。

レイヤー	構成要素と機能	脅威名	脅威	攻撃シナリオ
1.基盤モデル(Foundation Models)	文字起こしを要約する外部モデル			
2.データ操作(Data Operations)	文字起こしデータからの要約			
3.エージェントフレームワーク(Agent Framework)	録音データから文字起こしを行いそれらを要約するエージェント	1.プロンプトインジェクション	T2,T5	要約する文字列にプロンプトインジェクションを行う文字列を挿入されることにより、想定していない指示を行われる恐れがある
4.デプロイとインフラ(Deployment and Infrastructure)	クラウド上のサーバとサードバーティの文字起こしとLLM	2.DoS	T4	大量のアクセスを行われることによりリソースが枯渇しサービスが止まる恐れがある
5.評価と可観測性(Evaluation and Observability)	ログ機能なし	3.追跡不能性	T8	ログ機能がないため攻撃者の痕跡を追うことが困難
6.セキュリティとコンプライアンス(Security & Compliance)	アクセスポリシーとサードパーティのLLMのガードレール	4.ガードレールのバイパス	T7	ガードレールをバイパスし、想定していない指示を行われる恐れがある
7.エージェントエコシステム(Agent Ecosystem)	-			

レイヤーを越えた脅威 — MAESTRO最大の特徴

単一のレイヤーでは見えない、複数レイヤーをまたいで成立する攻撃シナリオを洗い出せる。

レイヤーの組み合わせ	脅威名	脅威	攻撃シナリオ
L1+L2+L3+L6	5.プロンプトインジェクション	T2.T5, T7	攻撃者が録音時にプロンプトインジェクションを起こす音声を入力し、それを要約させた際に誤動作を促す
L3+L4	6.プロンプトインジェクション	T2.T5, T7	インフラの脆弱性を突いて録音データを改変することにより、プロンプトインジェクションを含む文字列を生成させ、それを要約させた際に誤動作を促す

③ MAESTROの結果 — レイヤーを越えた脅威まで見える

レイヤー別に加え「レイヤーを越えた脅威」（No.5･6）まで拾える一方、多要素認証やアクセス制御といったシステム側の指摘は出にくい。

脅威名	脅威	可能性 P	影響度 I	リスク R	対策
1.プロンプトインジェクション	T2,T5	1	3	3	システム内でも入力値検証を行う
2.DoS	T4	3	3	9	レートリミット等を実装する
3.追跡不能性	T8	3	1	3	監査ログを実装する
4.ガードレールのバイパス	T7	1	3	3	システム内でも入力値検証を行う
5.プロンプトインジェクション	T2,T5,T7	1	3	3	システム内でも入力値検証を行う
6.プロンプトインジェクション	T2,T5,T7	1	3	3	インフラの定期アップデートと入力値検証を行う

3手法の比較 — 何が得意で、何が苦手か

観点	STRIDE	STRIDE+AI	MAESTRO
AI特有の脅威を評価できるか	✕	○	○
AIが動くシステム側の脅威を評価できるか	○	○	✕
リスクの定量評価	○	○	○
実施のしやすさ	○	○	✕
実施時間の短さ	○	○	✕
必要な前提知識	一般的なセキュリティ知識	＋ AIセキュリティの知識	＋ AI・エージェントの深い知識
必要なドキュメント	概要・構成図・DFD・データ詳細	左に同じ	＋ AI内部動作の詳細資料

「どれか一つ」を選ぶのではなく、システムの性格に応じて組み合わせるのが現実解。

使い分けの指針 — どのシステムに、どの手法を

STRIDE

まずはここから

適するケース　AIが関与しない既存の業務システム

メリット　経験者が多く短時間で実施でき、共通言語になる

注意点　AI特有の脅威は出ない。AI利用時は単独で完結させない

目安　半日／一般的なセキュリティ知識で可

STRIDE＋AI

多くの企業の現実解

適するケース　外部LLMの利用、社内システムに組み込んだAI機能

メリット　AIの脅威とシステムの脅威を、同じ表で同時に評価できる

注意点　AIセキュリティの知識を持つ参加者が最低1名必要

目安　半日／STRIDEの経験＋AI脅威の基礎知識

MAESTRO

エージェント時代の必須手法

適するケース　自律的に判断・実行するエージェント型AI、マルチ構成

メリット　レイヤーを越えた脅威まで構造的に洗い出せる

注意点　時間と知識を要する。システム側の脅威はSTRIDE系で補う

目安　1日以上／AI内部動作の資料と専門家が必要

議論 ― 評価結果から見えた3つの課題

① AIの脅威だけでなく、既存の脅威を含めた評価が必要

MAESTROではAIの脅威は細かく評価できたが、AIが動作するシステムに対する評価はあまり行えていない。AIの脅威とシステムの脅威を別々に実施するよりも、同時に行ったほうがシステム全体の脅威を把握できる。

② 脅威モデリングに必要となるドキュメント

今回は概要・コンテキストダイアグラム・DFD図・データ詳細を用意した。STRIDEとSTRIDE+AIはこれらのみで問題はなかったが、MAESTROの実施ではAI内部の動作をさらに深掘りしたドキュメントがあると助けになる。

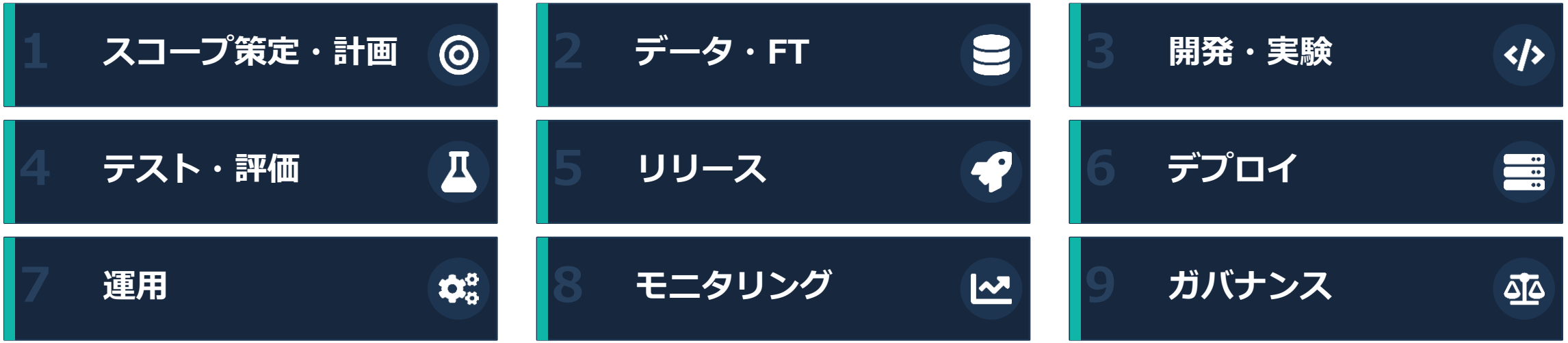
③ 実施に必要な参加者の知識

レッドチームやブルーチームなど様々な立場の参加者が意見を交換することが望ましい。AIに焦点を当てた場合は、AIのセキュリティに関する深い知識を持った参加者が一人は必要になる。もしくは実施前にAIの脅威に関する座学等を行うことが望ましい。

JNSA AIセキュリティWGの活動と成果物

WGの活動と今年度の成果物

- 月1回の定例会で、主要カンファレンス報告（Black Hat Asia / RSAC USA）、各種ガイドの解説、脅威モデリングのハンズオンを実施
- 今年度は Agentic AI 関連のドキュメントを作成予定（ライフサイクルの各過程で必要な対策・ガードレール・ポリシーの考察やサンプルを取りまとめ）



昨年度の成果物の今後の話

- 海外カンファレンスへのCFP投稿
  - 11月のGlobal AppSEC USA に投稿済
- コンテンツの英語版の作成
  - コンテンツの英訳版を9月目途に作成中
- コンテンツの拡充案を検討中

コンテンツの拡充案

案1｜AIによる脅威モデリング結果を追加

概要
生成AIを用いて対象システムの脅威モデリングを実施し、その結果を成果物へ追加し、議論する。

追加する内容
- 人間が行ったのと同一の手法でAIでも脅威モデリングを行う
- その結果を元に議論・考察

期待効果
脅威モデリングにおけるAIの有用性を確認する。

案2｜AIによる脅威モデリング結果の評価を追加

概要
AIが脅威モデリング結果を検証・評価し、その妥当性確認のプロセスを成果物へ追加する。

追加する内容
- AIによる脅威モデリングの結果の評価
- 人によるレビューとの比較
- 結果の議論と考察

期待効果
脅威モデリングの結果の評価におけるAI適用に向けた指針を示す。

まとめ

1. **評価対象は、AIの利用形態が異なる3つのシステム**
   自社構築AI／外部LLM利用／エージェント型AI。形態で脅威の焦点が変わる。

2. **3手法は得意領域が異なる**
   STRIDEはシステム側、+AIはAI特有の脅威、MAESTROはレイヤー横断まで抽出。

3. **「どれか一つ」ではなく、組み合わせて使う**
   システムの性格に応じて手法を選び、必要に応じて併用するのが現実解。

4. **実施時の要点**
   AIの脅威とシステムの脅威は同時に評価し、AIに詳しい参加者を一人は加える。

成果物のリンク

成果物（脅威モデルと評価結果）はGitHubで公開中