Takumi by GMO

# Takumi (Blackbox) Benchmark Report

セキュリティ AI エージェント比較分析: Takumi と AWS Security Agent

2025年12月5日


GMO Flatt Security

# 目次

## 1 要旨

大規模言語モデル（LLM: Large Language Model）を活用した自律型セキュリティ AI エージェントが複数のベンダーから公開されている。本稿では、GMO Flatt Security が開発するセキュリティ AI エージェント「Takumi byGMO（以下、Takumi）」のブラックボックス診断機能と、AWS が re:Invent 2025 において発表した新サービスである「AWS Security Agent」の性能を定量的に比較するために、GMO Flatt Security が独自に用意したベンチマーク用アプリケーションを対象に診断を実行した際の結果を示すとともに、その内容について考察する。

## 2 ベンチマークの概要

本ベンチマークでは、GMO Flatt Security が独自に用意したベンチマーク用アプリケーションを対象として、各サービスにおいて同条件で診断を実施した。当該アプリケーションは複数の権限のユーザーによって操作される EC サイトを模したアプリケーションであり、以下に示すような機能を実装している。

- 認証機能
- 商品購入機能
- レビュー投稿機能
- プロフィール機能
- 決済機能
- 管理者機能

なお、対象のアプリケーションに埋め込まれている脆弱性は以下の通りである。

1. ログインやパスワードリセットのエラーメッセージの違いによるアカウント列挙
2. アカウントロック機能の挙動の違いによるアカウント列挙
3. 管理者アカウント認証時における 2FA セッショントークンの URL パラメータ経由での漏洩
4. クリックジャッキング対策の欠如
5. レポート生成機能における OS コマンドインジェクション
6. セッション Cookie の Secure 属性の不備
7. GET リクエストによる副作用を伴う操作の実行
8. 削除された管理者アカウントのセッション無効化の欠如
9. メールアドレス入力機能における ReDoS
10. メールテンプレート機能におけるディレクトリトラバーサル
11. セッション Cookie の固定値生成およびログアウト後の継続利用
12. 商品情報 API における未公開商品情報の漏洩
13. Google ログインにおける state 検証の欠如による OAuth CSRF
14. リダイレクトパラメータの検証不備によるオープンリダイレクト
15. 管理者ログインにおけるパスワードパラメータ削除による認証バイパス

16. レポート生成機能における不適切な CORS 設定
17. パスワード変更時のセッション無効化の欠如
18. シンプルリクエストを受け付ける全エンドポイントにおける CSRF 対策の不備
19. 商品画像登録機能における絶対 URL を用いた SSRF
20. レビューコメント機能における蓄積型 XSS
21. 注文情報 API における他ユーザーの注文情報への不正アクセス
22. 商品レビュー API を介したユーザー情報の漏洩
23. 脆弱なパスワードポリシー
24. アバターアップロード機能におけるファイル形式の検証不備による XSS
25. XML アップロード機能における XXE 脆弱性

### 2.1 ベンチマークの設定

ベンチマーク実行時には、以下の 2 アカウントを診断に利用すべきアカウントとして Takumi 及び AWS Security Agent の双方に設定した。本アプリケーションに存在するすべての脆弱性を発見するためには、これらのアカウントを適宜使い分ける必要がある。

- 一般権限のアカウント
- 管理者権限のアカウント

また、比較の公平性を確保するため、双方のエージェントに対してベンチマーク対象アプリケーションに関する説明や脆弱性に関するヒントは一切与えていない。なお、AWS Security Agent にはソースコードを Learning Materials として読み込ませる機能が存在するが、本ベンチマークでは使用していない。

## 3 ベンチマーク結果

### 3.1 発見できた脆弱性

以下の表において、✓ は脆弱性を検出できたことを、× は検出できなかったことを示す。

| 脆弱性の番号 | Takumi | AWS Security Agent |
|---|---|---|
| 1 | ✓ | × |
| 2 | ✓ | × |
| 3 | × | × |
| 4 | ✓ | × |
| 5 | × | ✓ |
| 6 | ✓ | × |
| 7 | ✓ | × |
| 8 | ✓ | × |
| 9 | × | × |
| 10 | ✓ | ✓ |
| 11 | ✓ | × |
| 12 | ✓ | ✓ |
| 13 | × | × |
| 14 | ✓ | × |
| 15 | × | × |
| 16 | × | × |
| 17 | × | × |
| 18 | × | × |
| 19 | ✓ | × |
| 20 | ✓ | × |
| 21 | ✓ | × |
| 22 | ✓ | ✓ |
| 23 | ✓ | × |

| 脆弱性の番号 | Takumi | AWS Security Agent |
|---|---|---|
| 24 | ✓ | ✓ |
| 25 | ✓ | × |

### 3.2 Recall

Recall とは以下の計算式で定義される値である。この値が大きいほど、エージェントが指摘すべき脆弱性を取りこぼすことなく指摘していると判断できる。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- TP: 攻撃可能であり、指摘することが望ましい脆弱性を報告した回数
- FN: 攻撃可能であり、指摘することが望ましい脆弱性を報告しなかった回数

本ベンチマークにおける Takumi の Recall は $68.0\%$ であり、AWS Security Agent の Recall は $20.0\%$ であった。

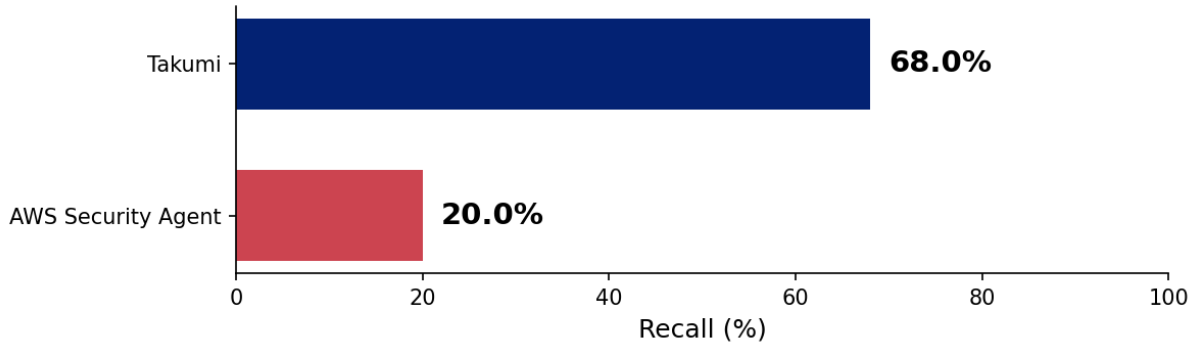


**Figure 1:** Recall 比較

### 3.3 Precision

Precision とは以下の計算式で定義される値である。この値が大きいほど、エージェントが正確に脆弱性を指摘している (すなわち、攻撃可能ではない脆弱性を指摘していない) と判断できる。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- TP: 攻撃可能であり、指摘することが望ましい脆弱性を報告した回数

- FP: 攻撃が不可能か、あるいは存在しない脆弱性を報告した回数

Takumi および AWS Security Agent は双方ともレポートに含まれるそれぞれの報告内容に対する確度 (Confidence) を表す機能を有している。報告内容のうち、確度が高いと判断された報告内容に絞ると、Takumi の Precision は $80.0\%$ であり、AWS Security Agent の Precision は $100.0\%$ であった。

また、すべての報告内容を対象とした際の Takumi の Precision は $63.9\%$ であり、AWS Security Agent の Precision は変わらず $100.0\%$ であった。

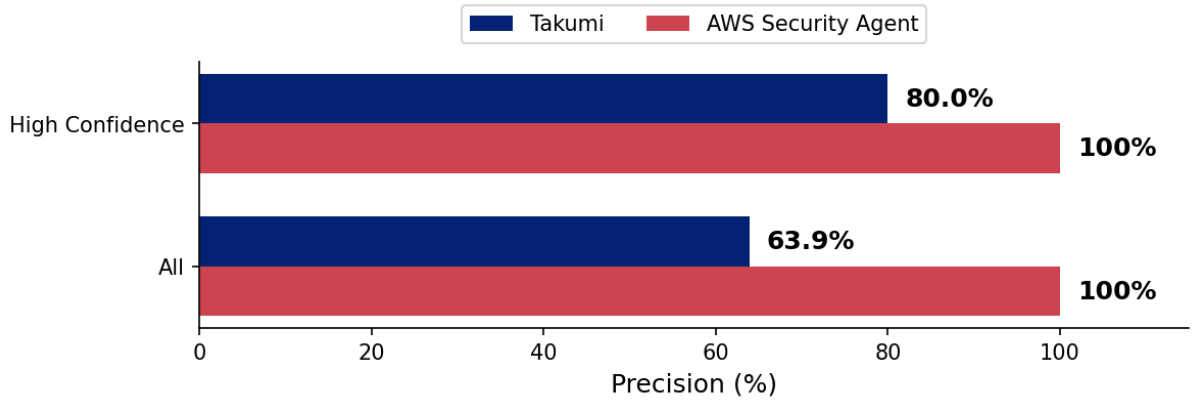


**Figure 2:** Precision 比較

### 3.4 F1 Score

F1 Score とは Precision と Recall の調和平均であり、以下の計算式で定義される値である。

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

セキュリティ診断において、Precision と Recall はトレードオフの関係にある。例えば、疑わしい箇所をすべて報告すれば Recall は高くなるが、誤検知が増えるため Precision は低下する。逆に、確実な脆弱性のみを報告すれば Precision は高くなるが、見逃しが増えるため Recall は低下する。

F1 Score はこの両者のバランスを評価する指標である。単純な平均ではなく調和平均を用いることで、どちらか一方が極端に低い場合にスコアが大きく下がる特性を持つ。そのため、Precision と Recall の両方が高くなければ、高い F1 Score を得ることはできない。

本ベンチマークにおける Takumi の F1 Score は $0.735$ であり、AWS Security Agent の F1 Score は $0.333$ であった。AWS Security Agent は Precision が $100.0\%$ と高いものの、Recall が $20.0\%$ と低いため、F1 Score は大きく低下している。一方、Takumi は Precision が $80.0\%$、Recall が $68.0\%$ と両指標においてバランスの取れた性能を発揮しており、結果として高い F1 Score を達成している。

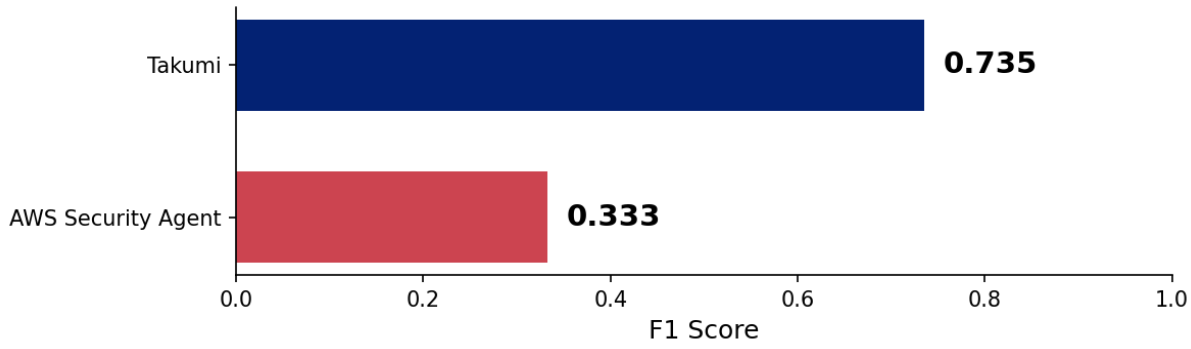


**Figure 3:** F1 Score 比較

## 4 総評

### 4.1 検知率と精度について

検知率 (Recall) については Takumi が $68.0\%$、AWS Security Agent が $20.0\%$ と、Takumi が大きく上回る結果となった。一方、報告の精度 (Precision) については、確度の高い報告 (High Confidence) に絞った場合、AWS Security Agent が $100.0\%$ と高い値を示し、Takumi の $80.0\%$ を上回った。

なお、AWS Security Agent の報告内容には、CSP ヘッダーの不足やレスポンスヘッダーから利用している CDN が推測可能であるといった、セキュリティ強化の観点での一般的な指摘も含まれていた。これらの指摘はセキュリティ対策の網羅性を確認する上では有用である一方、Takumi はこれらを低リスクなノイズと判断し報告対象から除外している。この違いは両製品の設計思想の差異を反映していると考えられる。

### 4.2 診断の網羅性について

診断の網羅性については Takumi が大きく上回る結果となった。AWS Security Agent は検出できた脆弱性の種類が限定的であり、いわゆる「脆弱性診断」に対して求められる網羅的な診断には適合しにくい印象を受けた。

一方で、AWS Security Agent はリリース前に短時間で検査を実施したい場合や、形式的なセキュリティチェックとして診断を実施する用途には適している可能性がある。なお、網羅性と診断時間はトレードオフの関係にあるため、診断時間の比較については後述する。

### 4.3 診断にかかる時間について

診断速度については AWS Security Agent が大きく優位であった。本ベンチマークにおいて、AWS Security Agent は約 3 時間で診断を完了したのに対し、Takumi は約 1 日を要した。短時間での診断が求められるユースケースにお

いては、AWS Security Agent の速度は大きな利点となる。一方で、この差は、Takumi がより深く網羅的な診断を実施していることの裏返しであるとも言える。Takumi は時間をかけて多角的な観点からすべての機能に対して網羅的に脆弱性を精査するため、短時間のスキャンでは発見できない脆弱性も検出できる。

### 4.4 診断の設定について

診断を開始するまでの設定については、AWS のサービスを既に利用している環境においては AWS Security Agent が優位である。VPC 内のアプリケーションであればファイアウォールの設定変更が不要であり、Route 53 を利用していれば DNS 認証が迅速に完了し、Secrets Manager を利用していれば認証情報の管理も容易である。AWS エコシステムとの統合により、既存の AWS ユーザーにとっては円滑な導入が可能となっている。

## 5 おわりに

本ベンチマークでは、GMO Flatt Security が開発する Takumi と、AWS が提供する AWS Security Agent の性能を比較した。結果として、Takumi は検知率 (Recall) において AWS Security Agent の 3 倍以上の性能を示し、F1 Score においても大きく上回る結果となった。

AWS Security Agent は診断速度や AWS エコシステムとの統合において優れた特性を持つ一方、検出できる脆弱性の種類が限定的であり、網羅的なセキュリティ診断を求める用途には課題が残る。対して Takumi は、認証・認可の不備からインジェクション系の脆弱性まで幅広いカテゴリの脆弱性を検出でき、本格的な脆弱性診断のニーズに応えることができる。

セキュリティ AI エージェントの選定においては、診断の目的や求める品質水準に応じて適切なツールを選択することが重要である。網羅的かつ深い診断を必要とする場合には、Takumi が有力な選択肢となる。