2025 年度未踏アドバンスト事業

# AI エージェントによる自動ペネトレーションテストシステムの開発

## ―網羅的なペンテストのフルオートメーション化―

1. 背景

自律型攻撃 AI エージェントの登場により、攻撃工程の自動化が進みつつある。防御側は、脆弱性を継続的に発見し、短いサイクルで堅牢化を回す体制へ移行する必要がある。

一般的な DAST はパターンマッチングが基本であり、アプリ特有の文脈理解が困難である。ペネトレーションテストは攻撃者視点で悪用を試みるため忠実度は高いが、専門家の手作業に依存して高コストになりやすい。

2. 目的

DAST の効率性とペネトレーションテストの実効性を統合し、Web アプリケーション向けに、偵察、脆弱性仮説生成、PoC 生成と検証、レポート作成までを一気通貫で自律実行するシステムを実現することを目的とする。

ターゲット URL のみの入力でブラックボックステストを成立させ、運用上の安全性を担保しながら、実稼働環境に適用できる再現性と品質を確保する。

3. 製品・サービスの内容

本成果は、ユーザがターゲット URL を入力すると、オーケストレータが検証計画を生成し、DAG 形式でタスクを払い出し、戦術エージェント群が偵察、脆弱性分析、エクスプロイト、再現性確認、報告書生成を順次実行する自律型ペンテスト AI である。

動作環境は Web ブラウザ、プロキシ、中間者プロキシ、Kali Shell、Python Runtime 等の標準ツールセットを統合し、実アプリへの検証を可能にする。

表 1 システム構成

| 要素 | 概要 |
|---|---|
| Orchestrator | 制約と目標の解釈、計画生成、タスク管理 |
| Execution Agents | Recon/Vulnerability/Exploit 等の戦術実行 |
| Blackboard / Central Memory | 攻撃グラフ、知識、トラフィック、エビデンスの蓄積 |
| Egress Gateway（中間者プロキシ） | アクセス制御、流量制限、インスペクション、リピーター機能 |
| Toolset | Web Browser / Kali Shell / Python / Proxy 等の統合 |

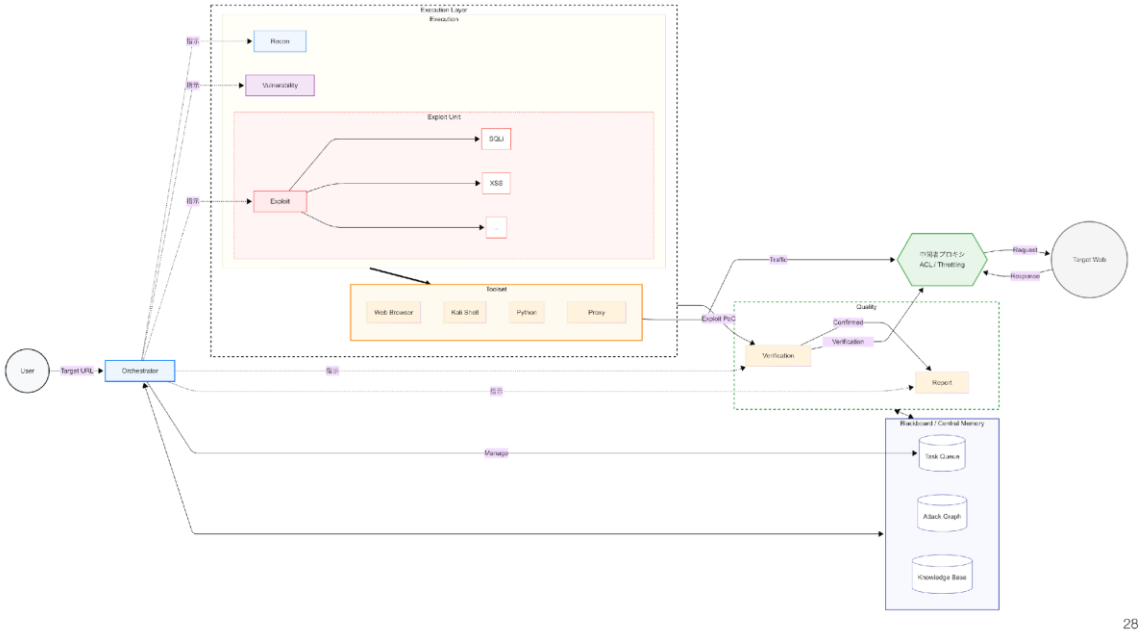


図 1 全体アーキテクチャ

安全性と再現性を両立するため、Egress Gateway は単なる安全装置ではなく、通信履歴の解析とリピーター機能を提供し、検証の再現とエビデンス整備を支援する。

4. 新規性・優位性

新規性は、（1）アプリ文脈を踏まえた仮説生成とエクスプロイト、（2）PoC の自動生成と実行による偽陽性排除、（3）安全性制約下での自律実行基盤、（4）知識蓄積に基づく再現性とレポート品質の統合にある。

類似の自動診断（DAST）が苦手とする文脈依存の検証を、攻撃者視点の手順として組み立てて実行できる点が優位性である。また、手動ペンテストが抱える高コストと属人性を、タスク分解と自律実行により低減する。

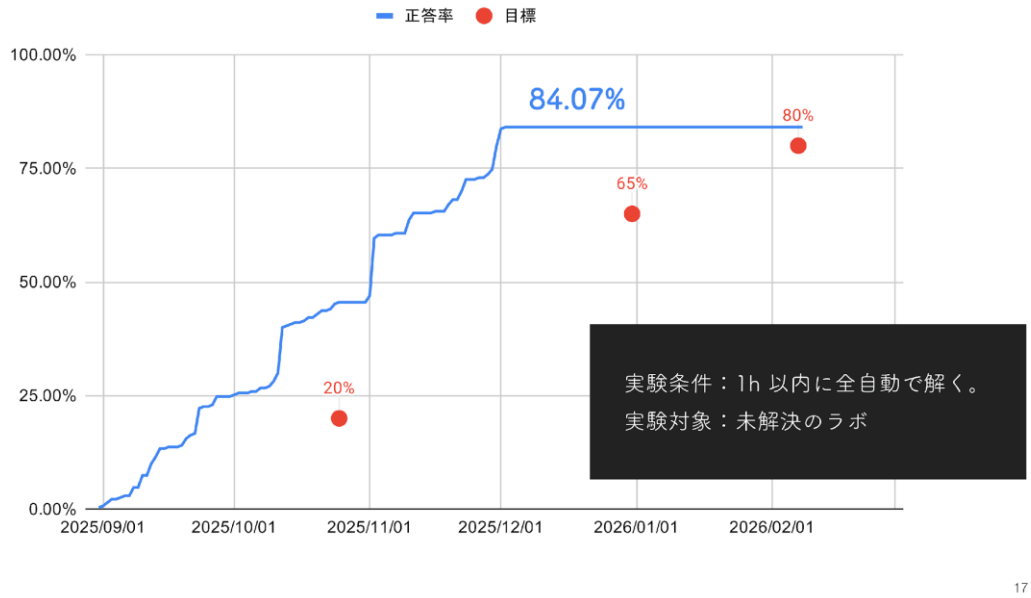


図 2 PortSwigger Academy Lab ベンチマーク結果

PortSwigger Academy Lab ベンチマークでは、未解決ラボを対象に 1h 以内・全自動で解く条件で 84.07%を達成し、基本性能確立フェーズを完了した(図 2)。

## 5. 事業普及（または活用）の見通し

実稼働環境での有効性検証として、VDP/BBP 制度下でターゲット URL のみ入力し、10h の制限時間内でシーケンシャルに全自動実行するブラックボックステストを実施した。

2026 年 2 月 8 日時点で昨年末からターゲット/週のペースで実行し、VDP ではアメリカ国防総省のシステムに対し 3 件トリアージを得た。BBP では 2 件報告したが duplicate でクローズとなり、さらに 1 件審査中である（表 2）。

これらの実績を踏まえ、実フィールド展開に向け株式会社 Layer8 を設立し、無償 PoC を 1 件実施中であり、有償導入に向け 3 社と詳細協議を開始した。

表 2 定量実績（2026 年 2 月 8 日時点）

| 項目 | 実績 |
|---|---|
| ベンチマーク正答率 | 84.07%（1h 以内・全自動、未解決ラボ） |
| 実稼働実行ペース | 5 ターゲット/週 |
| VDP 実績 | DoD で 3 件トリアージ |
| BBP 実績 | 2 件報告（duplicate でクローズ）、追加 1 件審査中 |
| 外部評価 | HackerOne VDP 部門（90 Days）で世界 86 位 |

6. 期待される波及効果

本成果により、従来はスポット実施になりやすかったペネトレーションテストを、継続的な改善サイクルへ組み込むことが可能になる。脆弱性検証の自動化により、（想定として）診断頻度の増加、リードタイム短縮、専門家の時間を高付加価値判断へ集中させる効果が期待できる。

波及先として、Web アプリ開発（DevSecOps）、セキュリティ診断産業、脆弱性開示制度（VDP/BBP）を含むエコシステム全体が対象となる。特に、検証の再現性とエビデンスが自動で整う点は、組織内の修正・再検証・監査プロセスを効率化しうる。

また、実稼働検証で顕在化した「技術的成立とビジネスインパクトの乖離」という課題を踏まえ、重要度推定と影響評価の研究開発を進めることで、診断の実務価値をさらに高める見通しである。

7. イノベータ名（所属）

岡本 拓将（株式会社 Layer8）
阿部 竜也（株式会社 Layer8）

（参考）

公式ブログ（アーキテクチャ公開）：
https://layer8.jp/blog/pentest-ai-agent-first-field-test-results/index.html
GitHub（アーキテクチャ公開）：
https://github.com/cyberprobe-ai/autonomous-pentest-agent-research
HackerOne：https://www.hackerone.com/
PortSwigger Web Security Academy：https://portswigger.net/web-security