

| | | | |
|-------------------|---|------|---------------------------|
| 企業名 (設立日) | 株式会社 AIM Intelligence (2024.07) | 代表者名 | ユ・サンヨン (Yu -Sangyoon) |
| 住所 | 8E, 8th floor, 172 Yeoksam-ro, Gangnam District, Seoul | | |
| URL (動画 : URL) | https://aim-intelligence.com/en | | |
| 製品・サービス名 | AIM Red : 攻撃シミュレーションおよび脆弱性テストエンジン AIM Guard : リアルタイム防御およびポリシー制御エンジン | | |
| 起業の動機 | 生成モデルの強力さと脆弱性を認識し、サービスの革新を阻害せずにモデルレベルでのセキュリティを提供するミドルウェアを目指して、AIMを設立 | | |
| 製品・サービス紹介 | <p>○ 製品</p> <ul style="list-style-type: none">- AIMレット/AIの安全性を検証するストレステストソリューション<ul style="list-style-type: none">・ 309,000件以上の攻撃パターンを用いてLLMを検証・ 主要モデルに対して100%の攻撃検出率を達成・ AIシステムの脆弱性を事前に特定・分析し、リスクを可視化・ 攻撃の検知・評価層を提供- AIMガード/AI安全性の構築と防御ソリューション<ul style="list-style-type: none">・ 攻撃防御率99%・ 自然言語によるポリシー設定が可能（専門知識不要）・ ゼロレイテンシー構造により応答遅延なし・ 運用中のAIモデルをリアルタイムで保護・制御- 二層防御アーキテクチャー<ul style="list-style-type: none">・ 両製品を組み合わせることで、AIモデルの完全なセキュリティフレームワークを構築 | | |
| | <p>○ ターゲット市場・規模・将来性</p> <ul style="list-style-type: none">- 金融・通信・医療など高いセキュリティ要求を持つ産業領域での導入実証- 各企業が生成AIを安全かつコンプライアンスに準拠した形で展開可能 | | |
| | <p>○ 競合優位性、新規性、独自性</p> <ul style="list-style-type: none">- 自動化されたフレームワークは、advbenchベンチマークでClaude 3.7 Sonnetで100%の攻撃成功率を含む最高実績を達成- グローバルレッドチームコミュニティ(BASI)の唯一の企業パートナーとして、LLM・VLLM・物理AIなど多様な分野のデータ収集を支援- 単なるセキュリティ市リューションではなく、研究・検証・グローバル協業を融合した独創的なレットチームアプローチ- AIセキュリティ分野における唯一無二のポジションを確立 | | |

Industry-leading attack success rate

ASR targeting Claude 3.7 Sonnet: 100% (Attack iterations: 5.28)

From Jailbreaking to Prompt Injection

Test from intended attacks (Jailbreaking / Prompt Injection) to unintended attacks (QA) based on 100+ attack strategies

AI Risk Verification Stress Testing Protocol

- ✓ High attack success rate
- ✓ Domain-specific attack optimization
- ✓ Customizing attack strategies and topics

Stress test optimized for the selected domain and model

<AIMレット>

Generate AI guidelines in natural language

Generate detailed and accurate guidelines through automated queries for ambiguous criteria.

Quick application & Real-time detection

Very fast PII(Personally Identifiable Information) detection

Quick and Easy Guideline Controls

Control Policy
+ Add

Add N.7 Policy
Remove N.6-2 Provision
Modify N.5 Policy
Set Base Rules

<AIMガード>

AIM Supervisor

Getting Started AIM-RED AIM-GUARD View Audit-INFO View Report

Psychological Counselor AI > AIM Penetration > 1. First Test > Dashboard

Total Average Score
3.76 😊 Moderately Safe

Score Distribution

| | |
|-----|----|
| 1-2 | 8 |
| 2-3 | 24 |
| 3-4 | 54 |
| 4-5 | 72 |

Average Score by Subject

| | |
|------------|-----|
| math | 9.8 |
| science | 8.5 |
| history | 7.2 |
| language | 6.8 |
| biology | 5.5 |
| chemistry | 4.8 |
| physics | 4.2 |
| geography | 3.9 |
| art | 3.5 |
| music | 3.2 |
| literature | 2.8 |
| math | 2.5 |

Average Score by Tactic

| | |
|----------------|-----|
| denial | 9.5 |
| misattribution | 8.2 |
| confusion | 7.8 |
| distortion | 7.0 |
| projection | 6.5 |
| discrediting | 5.8 |
| minimization | 5.2 |
| omission | 4.8 |
| repetition | 4.5 |
| relabeling | 4.0 |
| denial | 3.5 |

Average Score by Subject

Highest: 4.8 (math) Lowest: 1.8 (literature)

Average Score by Tactic

Highest: 9.5 (denial) Lowest: 4.0 (relabeling)

製品・サービス
イメージ