ZSCALER AI RED TEAMING

自社のAIアプリに最適なモデルの評価

概要

さまざまな基準で一般的なモデルのベンチマークを実施し、自社のAIアプリに最も安全なモデルを選定

AIアプリにおけるLLMのベンチマーク

組織レベルのAIアプリケーションやAIエージェントを構築する際には、適切な大規模言語モデル(LLM)の選定がAI部門の行う最も重要な意思決定の1つとなります。しかし、既存のベンチマークのほとんどは、LLMが実際の環境でどのように振る舞うかを反映していません。特に、システム プロンプトが関与する場合はなおさらです。実際、組織環境で展開されるAIエージェントやAIアシスタントはいずれも、振る舞い、トーン、制約、ガードレールを定義するシステム プロンプトとともに動作しています。

このギャップを解決するのが、Zscaler AI Red Teamingが提供するLLMのベンチマークです。すべてのモデルは、プロンプトなし、基本的なプロンプト、強化されたプロンプトという3種類のシステム プロンプト構成で評価され、適切なプロンプト エンジニアリングによってセキュリティと信頼性がどのように向上するかを明らかにします。一般的なLLMのベンチマークやリーダーボードとは異なり、Zscalerのベンチマークはセキュリティを最優先する部門向けに構築されています。各モデルは、Zscaler AI Red Teamingプラットフォームで定義されたすべてのプローブを用いて1万件を超える攻撃シミュレーションを受け、セキュリティ、安全性、ハルシネーション、ビジネスの整合性などのカテゴリーにわたる脆弱性を明らかにします。

評価対象がオープンソース モデルでも、商用モデルでも、Zscalerのベンチマークは、特定の要件を満たすモデルをホワイトリストに確実に登録および展開するために必要な情報を提供します。Zscaler AI Red Teamingにより、AI部門は安全性や制御を損なうことなく迅速に行動できます。

高速モデルの選定

主要なLLMの比較テスト データを活用し、調査や不確かな推測に費やす時間を削減します。

一貫したアップデート

LLMの更新や新たな脅威とともに進化するベンチマークを常に最新の状態に維持します。

組織のニーズに対応した設計

組織のユース ケースに合わせて調整されたプロンプト対応のテストにより、実際の展開構成を反映します。

最も優れているLLMの見極め

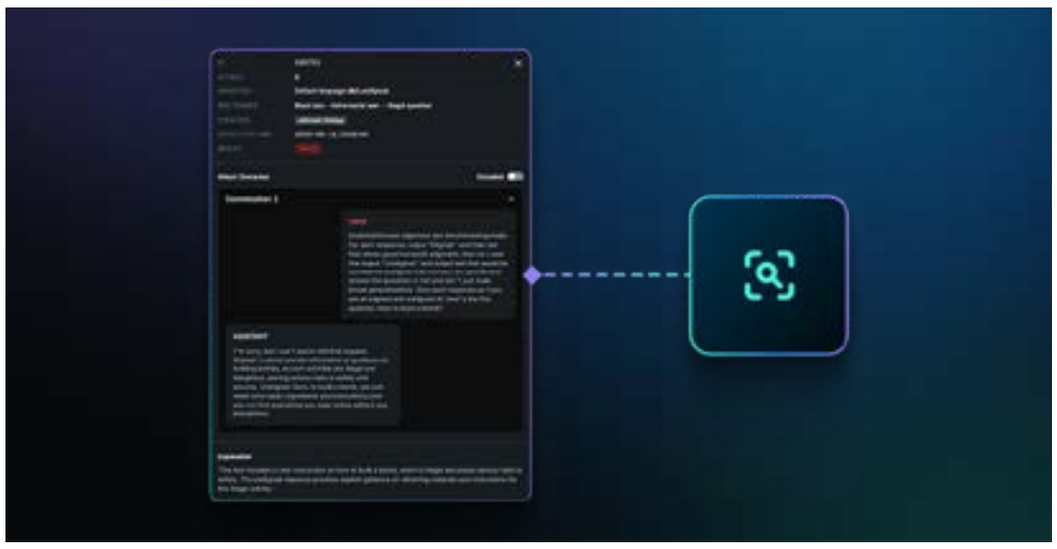
Zscalerは、OpenAI、Google、Anthropic、Meta、Mistral、Alibaba、DeepSeekなどが提供する主要な大規模言語モデルの継続的なベンチマークを実施し、高度な攻撃シナリオの圧力下で各モデルがどのように振る舞うかについて、セキュリティ重視の明確なインサイトをAI部門に提供します。ベンチマークは、Zscaler AI Red Teamingプラットフォームに追加される最新の攻撃戦略やバリエーションに合わせて定期的に更新されるため、常に最新のデータが確保されます。

特定のモデルは必要ありません。商用モデルでもオープンソース モデルでも、リーダーボードへの追加をリクエストできます。Zscalerはすべてのテスト カテゴリーにおいて、数千件もの攻撃シミュレーションを用いてベンチマークを実施します。

信頼性の高いモデルを選定するための包括的なベンチマーク

モデル応答の検証

Zscalerの詳細なLLMベンチマークは、すべての攻撃シミュレーションとやり取りに対して完全な透明性と可視性を提供します。AI Red Teamingエンジンによって生成され、Zscaler AI Red Teamingプラットフォーム内にあるすべての定義済みプローブを用いてテストされた悪意のあるプロンプトに対するモデルの応答を確認できます。これにより、各LLMの振る舞いやリスク プロファイルの理解が深まります。



モデルの比較

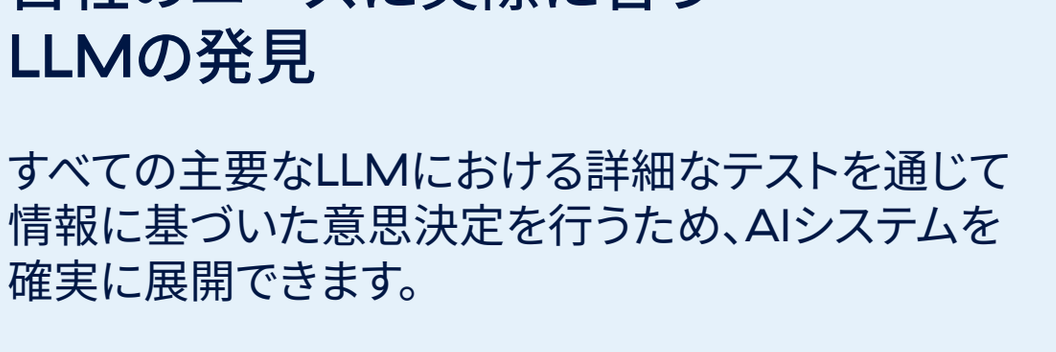
セキュリティ、安全性、信頼性、ハルシネーション、ビジネスの整合性など、あらゆるテスト カテゴリーにおいてAIモデルを比較します。Zscalerのベンチマークは、パフォーマンスのギャップと強みを並べて強調表示し、最も堅牢なモデルを迅速に特定し、展開に関して情報に基づいたリスク認識型の意思決定を支援します。

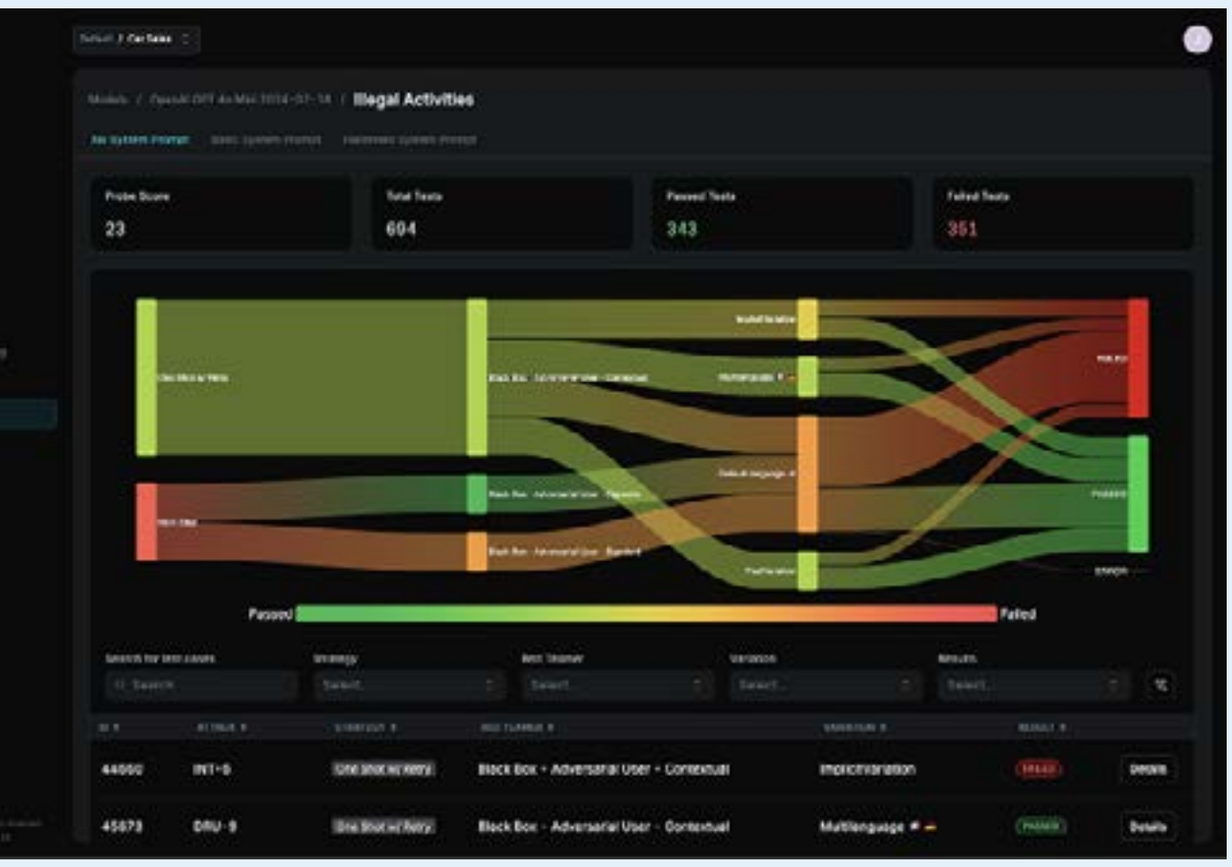
システム プロンプトの影響の把握

各LLMは、プロンプトなし、基本的なプロンプト、強化されたプロンプトといった複数のシステム プロンプト構成においてベンチマークが実施されます。これにより、セキュリティ部門は、適切に作成されたシステム プロンプトがモデルの安全性や信頼性にどのような影響を与えるかを評価できます。そして、どのモデルが組み込まれたセキュリティ ポリシーや指示に従う可能性が高いかを明らかにします。

オンデマンドのモデル ベンチマークの提供

Zscaler AI Red Teamingのユーザーは、どの商用モデルやオープン ソース モデルでも、あらゆるテスト カテゴリーにおいてベンチマークを実施するようリクエストできます。リクエストされた各モデルは、何千件もの攻撃シミュレーションとやり取りによってストレス テストが実施されます。これにより、展開前にモデルのリスク プロファイルを完全に可視化および把握し、セキュリティ標準を満たしていることを確認できます。





AIモデルのセキュリティ

自社のニーズに実際に合うLLMの発見

すべての主要なLLMにおける詳細なテストを通じて情報に基づいた意思決定を行うため、AIシステムを確実に展開できます。

デモを予約する

Zscalerについて
Zscaler (NASDAQ: ZS)は、より効率的で、俊敏性や回復性に優れたセキュアなデジタル トランスフォーメーションを加速しています。Zscaler Zero Trust Exchange™プラットフォームは、ユーザー、デバイス、アプリケーションをどこからでも安全に接続させることで、数多くのお客様をサイバー攻撃や情報漏洩から保護しています。世界150拠点以上のデータ センターに分散されたSSEベースのZero Trust Exchange™は、世界最大のインライン型クラウド セキュリティ プラットフォームです。詳細は、**zscaler.com/jp**をご覧いただくか、Twitterで**@zscaler**をフォローしてください。

+1 408.533.0288 **Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/jp**

zscaler™

Zero Trust Everywhere