ケーススタディー

ソフトウェア・サービス
AI 統合

Prediction Guard で LLM アプリケーションのリスクを回避

インテル® デベロッパー・クラウドは、AI スタートアップに回復機能を備えたコンピューティング・リソースを提供し、生成 AI アプリケーションをクラウドで運用する際のピーク・パフォーマンスと一貫性を保証します。

大規模言語モデルを使用すると AI を活用したイノベーションを推進できますが、多くの企業は、求める結果を得るためのリソースや専門知識が不足していることに気付いています。

大規模言語モデル (LLM) は、企業が業務の効率化を実現して強力な AI 駆動型ツールを構築するのに役立つ、大きな可能性を秘めています。AI 統合の最前線で活動しているスタートアップ企業の Prediction Guard は、組織が可能性をより早く発揮できるように支援したいと考えています。

LLM は幅広い生成 AI (GenAI) アプリケーションで有用性を証明していますが、信頼性、変動性、セキュリティーに関するリスクが発生する可能性があります。そのため、多くのユースケース、特に人々の安全や企業のコンプライアンスが最優先事項であるユースケースにデプロイすることは困難です。

Prediction Guard は、組織が LLM を AI 対応のビジネス・アプリケーションに統合できるように、インテル® デベロッパー・クラウドとインテル® Gaudi® 2 プロセッサーを活用して効率的かつ効果的な運用を行い、これらのリスクを軽減するために取り組んでいます。

概要

Prediction Guard の API プラットフォームは、幻覚、有害な出力、プロンプト・インジェクションなどのセキュリティーと信頼性の問題を軽減しながら、企業が大規模な言語モデルの可能性を最大限に活用できるようにします。

インテル® デベロッパー・クラウドは、企業がスケーラビリティー、可用性、セキュリティー、コスト削減を実現しながら、スループットを向上するのに役立ちます。

インテル® Liftoff プログラムにより、企業はインテル® デベロッパー・クラウドを使用して運用能力、価値、効率を最大化できます。

課題

2023年の初めに設立された Prediction Guard は、さまざまな業界で顧客ポートフォリオを急速に構築してきました。顧客の多くは、LLM がゲーム・チェンジャーになり得ることを認識しています。これらの組織は、情報抽出、カスタマーサービスと分析、マーケティング・キャンペーンの計画、財務報告、「Copilot」ソリューションなどの生成 AI アプリケーションで LLM を活用することを検討しています。

主要なユースケースの 1 つは、医師の口述から抽出された情報に基づいて医療フォームを自動的に作成することにより医療サービス提供者をサポートします (個人データと LLM の出力を安全かつ確実に処理する必要があるシナリオ)。ほかのユースケースは、個人顧客データを統合して、サプライチェーンの質問に対する回答を生成したり、購入行動、在庫、注文情報に基づいて小売の顧客に対応します。

「ユースケースはほぼ無限です。企業はどこにでも可能性を見つけます。彼らは、大規模言語モデルなどの生成 AI ツールを使用してデータの能力を最大限に活用し、ビジネスの成果を次のレベルに引き上げることができる新しい洞察とエクスペリエンスを生み出す方法を理解しています。しかし、彼らはまだ克服すべき大きな障害が存在することも認識しています。」
— Daniel Whitenack 氏、Prediction Guard 創設者

重要な課題の 1 つは、LLM の出力が異なり、LLM の信頼性が低くなることです。LLM はしばしば「幻覚」を引き起こし、結果が事実上不正確であったり、有害なものとなることがあります。LLM は、攻撃者が悪意のある入力を使用してモデルから意図しない応答やデータ侵害を引き出す、プロンプト・インジェクションと呼ばれる新しいタイプのセキュリティーの脅威に対しても脆弱です。

「これらの課題により、組織が LLM を活用することに消極的になる可能性があります。彼らは責任を回避したいと考えて、個人を特定できる情報などの機密データが失われるリスクを負いたくないのです。多くの組織では、これらの問題に対処するために必要なデータを処理できる人材やその他のリソースも不足しています。Prediction Guard のソリューションは導入に対するこれらの障壁を取り除くのに役立ちますが、インテル® デベロッパー・クラウドで提供される、信頼性の高いコンピューティング能力が必要になります。」— Daniel Whitenack 氏

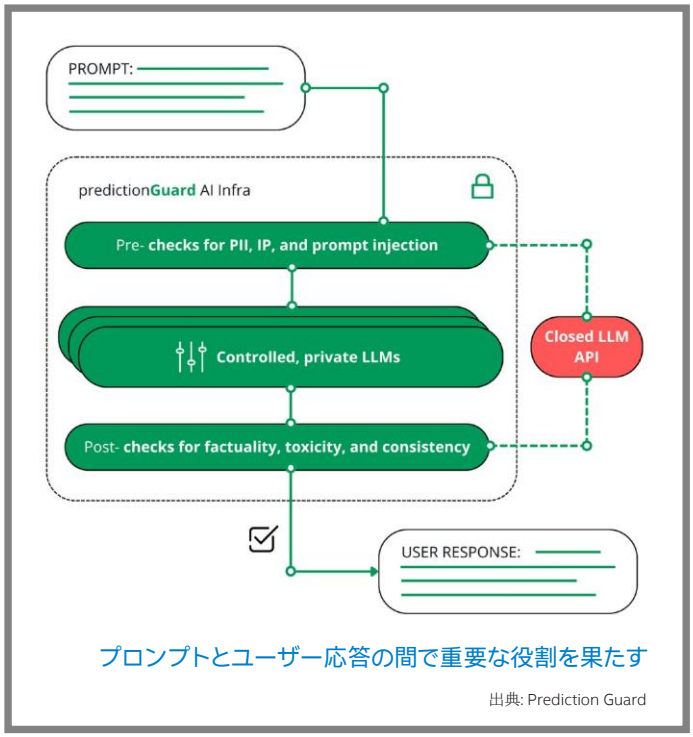
ソリューション

Prediction Guard のプラットフォームは、顧客に複数の LLM へのアクセスを提供しながら、有害な入力 (プロンプト・インジェクションなど) と有害な出力 (不正確性や有害性など) に対処する追加の機能とテクノロジーを統合します。プラットフォームはすべて、インテル® デベロッパー・クラウド上の、インテル® Gaudi® 2 AI アクセラレーターを利用したセキュアなプライベート環境でホストされます。Prediction Guard は、インテル® デベロッパー・クラウドで Hugging Face Optimum Habana ライブラリー (インテルと Hugging Face のコラボレーション) を使用してプロセッサー上で実行するモデルを最適化しています。

インテル® デベロッパー・クラウドを利用して、ユーザーは最新のインテルのハードウェアとソフトウェアのクラスター上でアプリケーションを学習、テスト、実行できます。発売前の開発とテストのために最新のインテルのテクノロジーにアクセスできるだけでなく、AI アプリケーションを大規模に構築およびデプロイするためのフルスタックのソリューションも提供されます。インテル® デベロッパー・クラウドは、ハードウェアの選択肢と独自のプログラミング・モデルからの自由を開発者に提供し、アクセラレーテッド・コンピューティング、コードの再利用と移植性をサポートします。

インテル® デベロッパー・クラウドで提供されるインテル® Gaudi® 2 プロセッサーは、Prediction Guard がビジネスの成長と、迅速で信頼性の高い生成 AI の成果に対する顧客の期待に伴う大きなニーズに対処するのに役立ちます。AI アクセラレーターとして設計されたインテル® Gaudi® 2 プロセッサーは、7nm プロセス・テクノロジー、ヘテロジニアス・コンピューティング、24 のテンソル・プロセッサー・コア、デュアル行列乗算エンジンを備えています。24 の 100 ギガビット・イーサネットをチップ上に統合し、96GB HBM2E メモリー、48MB SRAM、インテグレーテッド・メディア・コントロールも含まれています。

インテル® デベロッパー・クラウド上のインテル® Gaudi® 2 プロセッサーに移行する前に、Prediction Guard は別のプロバイダーのプロセッサーを試しました。新しいインテルの環境は、Prediction Guard とその顧客に複数の分野で大幅なスピードアップをもたらしました。

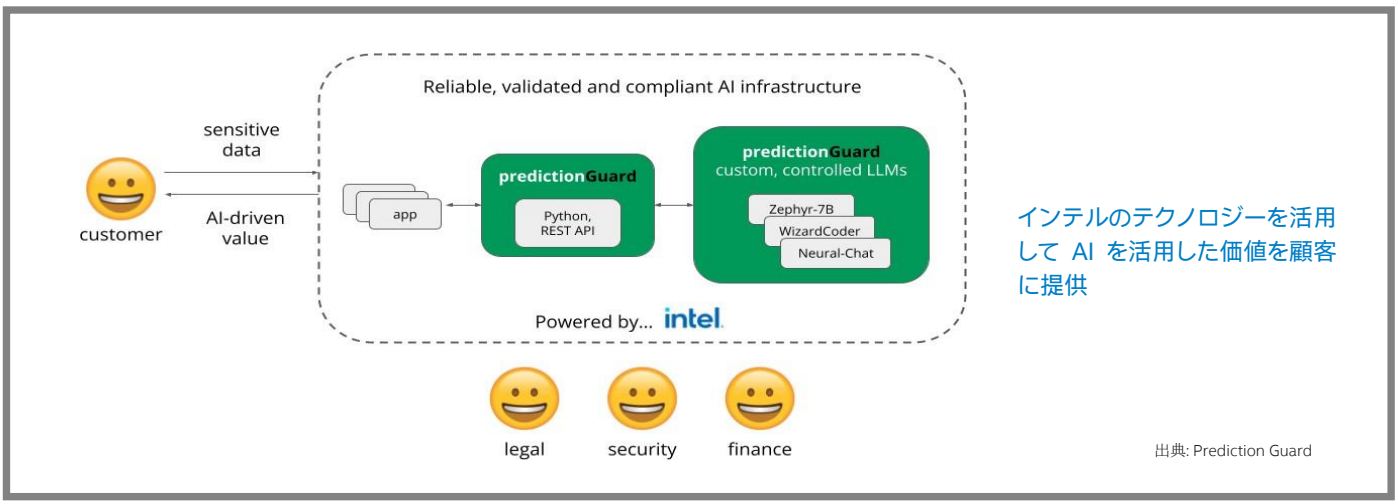


プロンプトとユーザー応答の間で重要な役割を果たす

出典: Prediction Guard

「特定のモデルでは、インテル® Gaudi® 2 への移行により、コストが削減され、スループットが 2 倍に向上しました。当社の顧客は、迅速かつコスト効率良く出力を生成できるようになりました。[1] また、以前のプロバイダーでは不足することが多かった、クラウドで必要な処理能力に簡単にアクセスできるようになりました。さらに、可用性の向上により、顧客や自社のビジネスの増大するニーズに合わせて、容易にスケールアップできるようになりました。特に機密データの処理に慎重な Prediction Guard の顧客の場合、インテル® デベロッパー・クラウドで使いやすい Prediction Guard API を使用して顧客データを保護するクライアント固有のデプロイメントを作成できます。」— Daniel Whitenack 氏

インテル® デベロッパー・クラウドのもう 1 つのメリットは、Prediction Guard がインテル® Liftoff プログラムに参加したことに由来します。このプログラムは、スタートアップ企業に必要な計算能力へのアクセスを提供し、スタートアップ企業間のコラボレーションを促進して、サービスの改良と強化を支援します。



インテルのテクノロジーを活用して AI を活用した価値を顧客に提供

出典: Prediction Guard

[1] インテルは、サードパーティーのデータについて管理や監査を行っていません。ほかの情報も参考にして、正確かどうかを評価してください。

「インテル® Liftoff への参加は、我々が現在の地位を確立するために非常に重要なことでした。このプログラムに参加することにより、インテル® Gaudi® 2 AI アクセラレーターで実現可能な技術を調査し、必要な速度と規模で、インテル® デベロッパー・クラウド上でシステムを効果的に実行できることを早い段階で実証することができました。」— Daniel Whitenack 氏

Prediction Guard はインテル® デベロッパー・クラウド上のインテル® Gaudi® 2 プロセッサーによる運用に完全に移行しており、インテル® Liftoff のメリットを引き続き享受しています。

「このプログラムでは、Prediction Guard がインテルのテクノロジーを使用して革新を続ける中で、質問したり問題に対処するのに役立つインテルのエキスパートへの迅速なアクセスが提供されます。この関係の最も素晴らしい点の 1 つです。コンピューティングとコミュニティーがここにあるのです。結果も同様です。インテル® デベロッパー・クラウドは、革新と成長を続ける当社のビジネスをサポートできる本格的な AI クラウド環境を提供します。」— Daniel Whitenack 氏

結果

Prediction Guard は、インテル® デベロッパー・クラウド上でインテル® Gaudi® 2 AI アクセラレーターを使用することにより、速度、セキュリティー、拡張性、信頼性の向上とコストの削減を実現しています。このテクノロジーにより、一部の大規模言語モデルのスループットが 2 倍に向上し、さまざまな革新的な生成 AI アプリケーションに LLM を使用する顧客をサポートできるようになりました。[2]

インテル® デベロッパー・クラウドは、Prediction Guard の収益に目に見える効果を与えています。Prediction Guard は重要なモデルのホスティングをインテル® デベロッパー・クラウド上のインテル® Gaudi® 2 マシンに移行したことで、固定費を正確に把握できるようになり、より正確なコスト計画を実施して、実行するモデルを柔軟に調整できるようになったのです。

ソリューションの構成要素

インテル® Tiber™ デベロッパー・クラウド

インテル® Gaudi® 2 プロセッサー

Optimum Habana

PyTorch* 向けインテル® エクステンション

Transformers* 向けインテル® エクステンション

関連情報

Prediction Guard (英語)

インテル® Liftoff

インテル® Tiber™ デベロッパー・クラウド

インテル® Gaudi® 2 AI アクセラレーター (英語)

oneAPI