Prediction Guard がインテル® Gaudi® 2 AI アクセラレーターで信頼できる AI を実現した方法

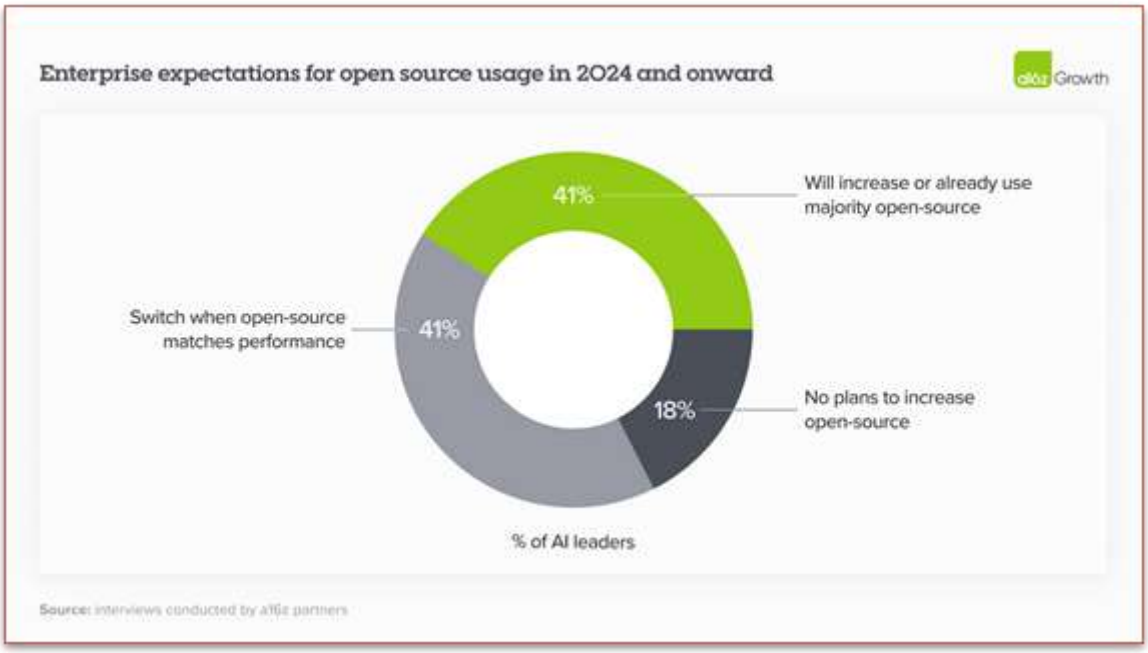
この記事は、インテルのブログで公開されている「How Prediction Guard Delivers Trustworthy AI on Intel® Gaudi® 2 AI Accelerators」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

エンタープライズ・レベルでオープンソースのツールとソフトウェアの利用、特に生成 AI と大規模言語モデル (LLM) の導入が増えるにつれて、エンタープライズ・アプリケーション向けに安全でスケーラブルかつ効率良い LLM の実装に必要な基本的な戦略とテクノロジーについて議論することが重要になります。Prediction Guard (英語) の創設者である Daniel Whitenack 博士とスタートアップ向けのインテル® Liftoff プログラムのエンジニアリング・リーダーである Rahul Unnikrishnan Nair が主導するこのインテルのウェビナーでは、オープンモデルを使用して LLM を導入し、データ・プライバシーを確保し、高精度を維持するという重要な側面について説明します。

ビデオ: Prediction Guard がインテル® Gaudi® 2 AI アクセラレーターで信頼できる AI を実現した方法 (英語)

エンタープライズ LLM 導入の主要要件

ウェビナーでは、エンタープライズ LLM 導入を成功させるには、オープンモデルの使用、データ・プライバシーの確保、高精度の維持という 3 つの主要な要件があることが示されています。Llama* 3 (英語) や Mistral* (英語) などのオープンモデルでは、モデルウェイトをダウンロードしたり、推論コードにアクセスできるため、より高度な制御とカスタマイズが可能です。これは、プロセスの透明性がないまま API 経由でアクセスされるクローズドモデルとは対照的です。特にエンタープライズでは、個人識別情報 (PII) や保護対象健康情報 (PHI) などの機密情報を扱うため、データ・プライバシーの確保は最も重要です。このような場合、HIPAA などの標準への準拠が必要になります。高精度も不可欠であり、ハルシネーション (文法的に正しく一貫性があるにもかかわらず、出力に誤った情報や誤解を招く情報が生成されること) などの問題を軽減するため、基準となるデータで LLM の出力を検証する堅牢なメカニズムが求められます。



クローズドモデルの課題

OpenAI (英語) や Cohere (英語) などが提供するクローズドモデルには、いくつかの課題があります。これらのモデルでは、入力と出力がどのように処理さているか確認できないため、バイアスやエラーが発生する可能性があります。透明性がないため、ユーザーは原因が分からないモデレーション・エラーやレイテンシーの変動に遭遇する可能性があります。さらに、プロンプト・インジェクション攻撃によりクローズドモデルが悪用され、機密データが漏洩し、重大なセキュリティー・リスクが生じる可能性があります。これらの問題は、エンタープライズ・アプリケーションにオープンモデルを使用することの重要性を強調しています。

Prediction Guard のアプローチ

Prediction Guard のプラットフォームは、安全なホスティング、堅牢な保護、パフォーマンスの最適化を組み合わせることで、これらの課題に対処します。安全なホスティングは、インテル® Tiber™ AI クラウド内のプライベート環境でモデルをホスティングすることで実現し、インテル® Gaudi® 2 AI アクセラレーターを活用することでパフォーマンスとコスト効率を向上します。入力フィルターを使用して、プロンプト・インジェクションをブロックし、PII が LLM に到達する前にマスクまたは置換します。出力バリデーターは、LLM 出力を基準データと比較することで、LLM 出力の事実上の一貫性を保証します。

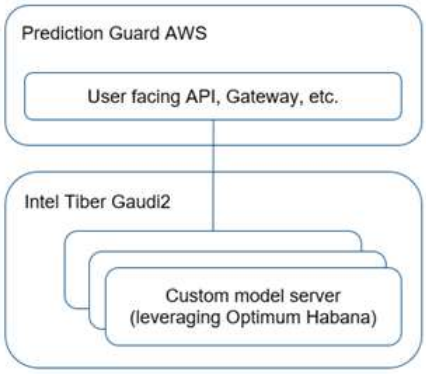
インテル® Gaudi® 2 AI アクセラレーターへの移行

Prediction Guard のインテル® Gaudi® 2 AI アクセラレーターへの移行は、特定の技術要件と最適化に対応するため、複数のフェーズで実行されました。最初の移行フェーズ (2023 年 7 月～9 月) では、カスタム・モデル・サーバーがベアメタルのインテル® Gaudi® 2 AI アクセラレーターに導入されました。Optimum Habana (英語) を使用して、Prediction Guard は標準の Hugging Face* クラスをインテル® Gaudi® 2 AI アクセラレーター向けに最適化されたバージョンに交換しました。使用量の急増に対応する動的バッチ処理が実装され、推論効率を最適化するため静的シェイプが管理されています。

Initial Migration

July 2023 to September 2023

- Bare Metal Gaudi2 server(s)
- Hybrid Cloud
- Earlier Synapse versions
- No clustering/ orchestration
- Highly custom model servers
- Critical:
  - Manage static shapes
  - Implement dynamic batching

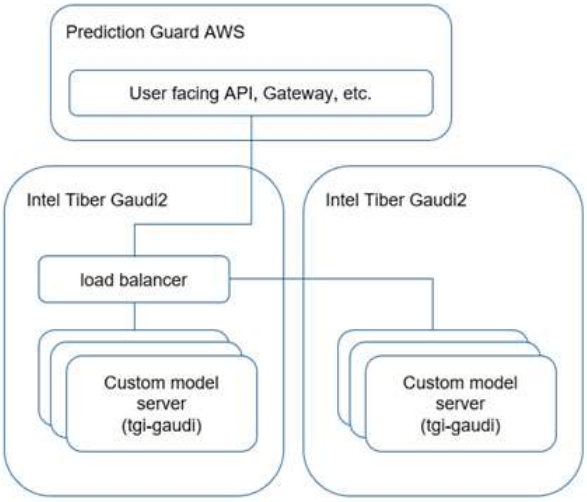


最適化フェーズ (2023 年 9 月～2024 年 4 月) では、複数のインテル® Gaudi® 2 プラットフォーム間で負荷を分散し、同様のサイズのプロンプトをバケット化してスループット向上のためパディングすることでプロンプト処理を最適化し、モデルサーバー管理を合理化するため TGI Gaudi (英語) フレームワークに移行しました。

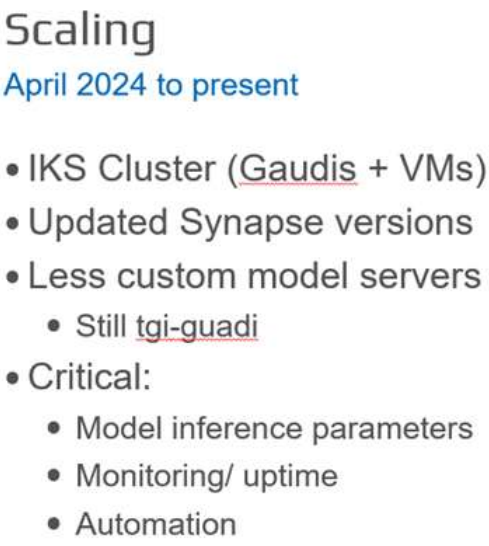
Optimization

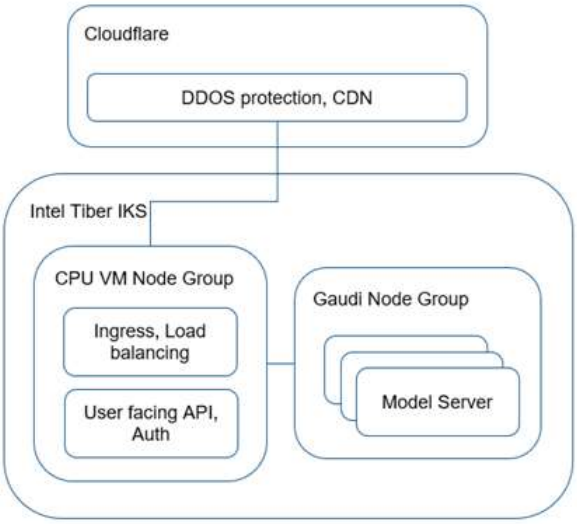
September 2023 to April 2024

- Bare Metal Gaudi2 server(s)
- Hybrid Cloud
- Updated Synapse versions
- No clustering/ orchestration
- Less custom model servers
  - Eventually tgi-guadi
- Critical:
  - Load balancing, model replicas
  - Bucketing and padding



現在のスケーリング・フェーズ (2024 年 4 月～現在) では、Prediction Guard はインテル® Tiber™ AI クラウド内の Kubernetes ベースのインフラストラクチャーに移行し、CPU とインテル® Gaudi® 2 AI アクセラレーター・ベースのノードグループを組み合わせています。デプロイの自動化、稼働時間とパフォーマンスの監視、DDoS 防御と CDN サービスのため Cloudflare の統合が実装されました。





パフォーマンスとコストのメリット

インテル® Gaudi® 2 AI アクセラレーターへの移行は、大幅な改善をもたらしました。Prediction Guard は、エンタープライズ・ワークロードのスループットを 2 倍に向上し、以前の GPU ソリューションと比較して計算コストを 10 分の 1 に削減しました。最初のトークンまでのレイテンシーが 200 ミリ秒未満に短縮されたことで、Prediction Guard は業界最先端のパフォーマンスを実現しました。これらのメリットはパフォーマンスを犠牲にすることなく実現され、インテル® Gaudi® 2 AI アクセラレーターのコスト効率とスケーラビリティーを実証しています。

技術的な洞察と推奨事項

ウェビナーでは、堅牢なエンタープライズ AI ソリューションには、LLM API 以上のものが必要であることを強調しています。出力の正確性と信頼性を確保するには、基準データでの厳格な検証が必要です。機密データを統合するには、強力なプライバシーとセキュリティー対策が必要であり、データの取り扱いは AI システム設計における重要な考慮事項となっています。Prediction Guard のインテル® Gaudi® 2 AI アクセラレーターの使用を最適化する段階的なアプローチは、他の開発者にとってモデルとなります。基本機能の検証から始めて、パフォーマンス・メトリックとユーザー・フィードバックに基づいて段階的に最適化およびスケーリングすることが、導入を成功させる鍵となります。

技術実装の詳細

移行の初期段階では、静的な形状を管理するため、可変長のプロンプトをあらかじめ決められたサイズにパディングして処理するようにモデルサーバーを構成し、メモリー使用量と計算量を最適化しました。動的バッチ処理により、システムはリクエストのウィンドウを蓄積し、それらを一括して処理することで、スループットの向上とレイテンシーの短縮を実現しました。最適化段階では、トラフィックを効率的に管理し、ボトルネックを回避するため、複数のインテル® Gaudi® 2 AI アクセラレーター間で負荷分散が実装されました。入力プロンプトをサイズに基づいてバケットに分類し、各バケット内でパディングを行うことで、入力プロンプトの処理を洗練させ、パフォーマンスをさらに向上しました。TGI Gaudi フレームワークへの移行により、モデルサーバーの管理が効率化されました。

スケーリング段階では、CPU とインテル® Gaudi® 2 AI アクセラレーター・ベースのノードグループを組み合わせたインテル® Kubernetes Service (IKS) クラスターをデプロイすることで、スケーラブルかつレジリエントなデプロイを実現しました。デプロイプロセスの自動化とモニタリング・ツールの導入により、高可用性とパフォーマンスを確保しました。推論パラメーターの設定とキー値キャッシュの管理により、モデルの提供効率を最適化しました。

実践的な実装のヒント

同様の AI ソリューションの実装を検討している開発者や企業には、制御とカスタマイズ機能を維持するためオープンモデルから始めることをお勧めします。機密データが安全に取り扱われ、関連する標準に準拠していることを確認することが重要です。基本機能から始めて、メトリックとフィードバックに基づいて徐々にパフォーマンスを改善する段階的な最適化アプローチを採用することも、導入を成功させる鍵となります。最後に、Optimum Habana や TGI Gaudi などのフレームワークを活用すると、統合と最適化の取り組みを効率化できます。

まとめ

Prediction Guard の包括的なアプローチは、インテルとの協業により、企業がいかに安全でスケーラブルかつ効率良い AI ソリューションを導入できるかを示しています。インテル® Gaudi® 2 AI アクセラレーターとインテル® Tiber™ AI クラウドを活用することで、Prediction Guard は、制御、カスタマイズ、データ・プライバシー、精度に関する重要な懸念に対処し、エンタープライズ AI を導入するための堅牢なプラットフォームを提供します。ウェビナーで共有された技術的な洞察と実用的な推奨事項は、LLM 導入の複雑な作業に取り組む開発者と企業にとって貴重なガイダンスを提供します。

また、インテルのほかの **AI ツール** (英語) と**フレームワーク**の最適化をチェックし、インテルの AI ソフトウェア・ポートフォリオの基盤を形成する、統一されたオープンな標準ベースの **oneAPI** プログラミング・モデルについて学んでください。

ウェビナーの講演者について

Daniel Whitenack 博士
Prediction Guard の創設者兼 CEO であり、経験豊富なデータ・サイエンティストです。大規模なマシンラーニング・モデルの開発とデプロイに 10 年以上の経験があり、2 つのスタートアップ企業と 4,000 人以上のスタッフを擁する国際 NGO でデータチームを構築しました。Practical AI ポッドキャストの共同ホストを務め、世界中のカンファレンスで講演し、パーデュー大学でデータサイエンス/アナリティクスを教えています。

Rahul Unnikrishnan Nair
スタートアップ向けインテル® Liftoff プログラムのエンジニアリング・リーダーです。応用 AI とエンジニアリングの豊富な経験を活かして、アーリーステージの AI スタートアップを指導しています。マシンラーニングと応用ディープラーニングの分野で 10 年以上の経験を持ち、生成 AI の分野でも重要な役割を担ってきた Rahul は、ユースケースを重視した実践的なエンジニアリングと最適化により、スタートアップ企業が革新的なアイデアを本格的な市場対応可能な製品に転換できるよう支援しています。