生成 AI のリスクを低減し、信頼性を向上させる Prediction Guard のソリューション

インテル® Liftoff 参加企業によるインテル® Tiber™ AI クラウド採用事例

エクセルソフト株式会社

2025/01/17

生成 AI アプリケーションへの期待

LLM (大規模言語モデル) は、企業が業務の効率化を実現するための
AI 駆動型ツールの構築に役立つ、大きな可能性を秘めている

オープンアクセスの LLM
(Mistral、Llama3、Deepseek など)

自然言語による多様な指示へ
柔軟に回答する能力を発揮

選択の自由、ベンダーロックインの回避

RAG (検索拡張生成)
※ 派生手法を含む

前提知識を与えて専門性を補強
役割に合った回答を生成

機微情報を明かさずに活用

Prediction Guard: 生成 AI 導入支援企業

- 米国、2023年初頭に設立 https://predictionguard.com/ (英語)
- 創設者、CEO: Daniel Whitenack
  - データ・サイエンティスト、AI 専門家、業界で 10年以上の経験
  - インストラクター https://datadan.io (英語)
  - ポッドキャスター https://changelog.com/practicalai (英語)

安全で私有できる AI 機能を使いやすい API を通して提供
お客様のデータを保持しない
HIPAA 準拠、要件に応じて BAA 締結可

HIPAA: Health Insurance Portability and Accountability Act (医療保険の相互運用性と説明責任に関する米国の法律)

課題: 有効な計算資源の確保

LLM 運用に対する要求

ハルシネーション (幻覚) 抑制
→ 回答は信頼できるか

情報セキュリティーの確保
→ リスクを把握し、対応できるか

処理能力と合理的なコスト
→ 有効な運用を継続できるか

インテル® Tiber™ AI クラウド

- インテルによる AI 向けブティック型クラウドサービス
  - 一次的な評価および継続的な運用ができる
    インテルの各種プロセッサーに基づくシステムのリモートアクセスを提供
    (ベアメタル、VM インスタンス)

CPU

GPU

AI アクセラレーター

インテル® Liftoff プログラム 🚀

AI や機械学習 (ML) を活用してビジネスを行う
スタートアップ企業向けのパートナー企画
無料 (あるいは低価格) で、インテルの
ハードウェアを用いる技術的な支援、情報提供、
ネットワーキングの機会などを提供

詳細、事例など: 英語 日本語

Prediction Guard は、インテルの クラウド・プラットフォームをビジネスに採用

インテル® Gaudi® 2 AI アクセラレーターで
LLM の運用コストを削減しつつ、
最大 2 倍のスループット向上を達成[1]

既存のシステム

NVIDIA A100
Tensor コア GPU ベース

Prediction Guard

移行

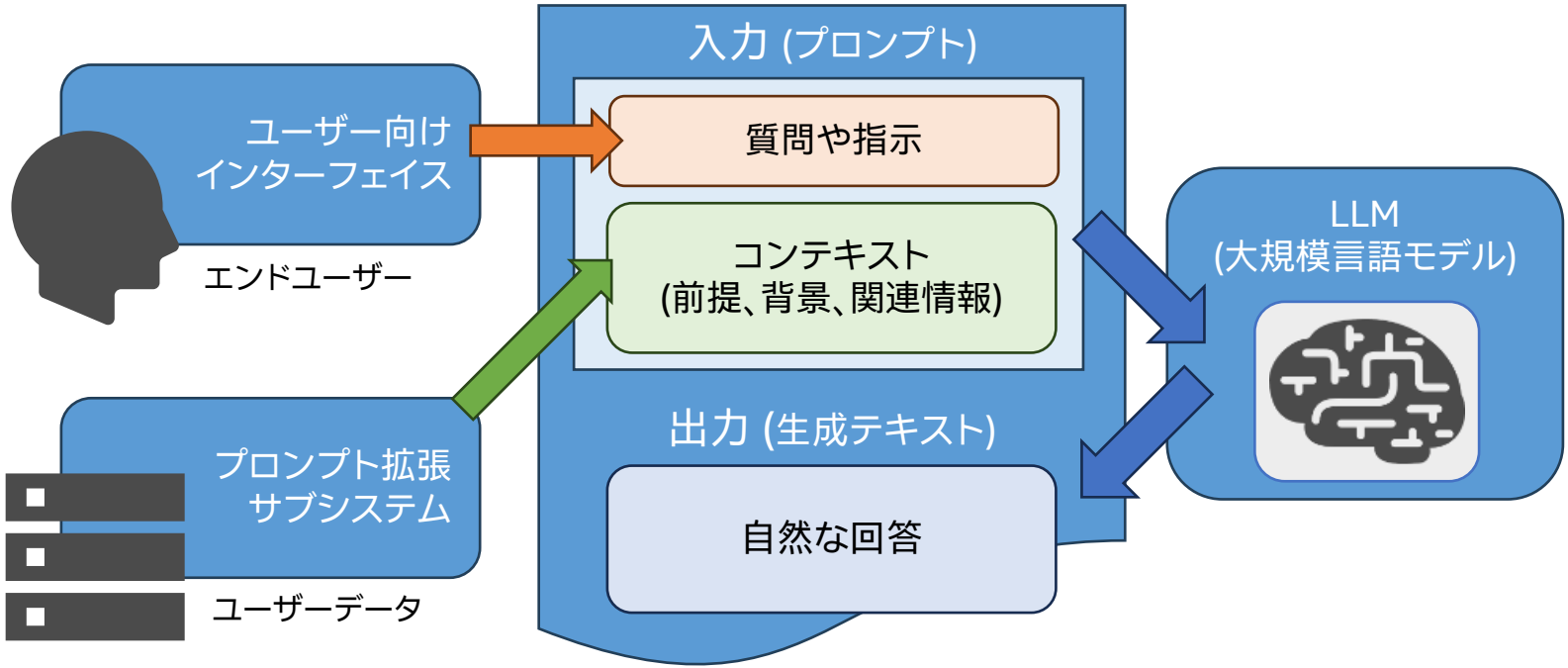
intel® tiber™ AI Cloud

旧称: インテル® デベロッパー・クラウド

参照記事 1　参照記事 2

[1] 2024年1月31日時点の Prediction Guard による報告値
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

LLM (大規模言語モデル) に基づく生成 AI システムの基本構成



Prediction Guard の取り組み

- 生成 AI の可能性
  - ✓ 2 つの異なるパートナー事例
- 生成 AI のリスク (懸念材料)
  - ✓ 想定されるリスクと対応
- 生成 AI のコスト (費用)
  - ✓ コストの構造、およびインテルとの協業による効率的な実装

Prediction Guard のソフトウェア・システム、採用モデル (LLM など) はドキュメントを参照
https://docs.predictionguard.com/guides-and-concepts/getting-started/quick-start (英語)

事例 1: 救急医療の臨床意思決定支援ツール

SimWerx

🚑 https://simwerx.com/ (英語)

- 救急医療に関する訓練と臨床意思決定の支援ツールを開発、提供

生成 AI アプリケーション
- 情報収集、質疑応答、タスクの優先順位付けの自動化
- 応急処置のガイドラインと一貫した自然な指示テキストの生成

Prediction Guard が提供した価値
- PII、PHI の安全な取り扱い
- ハルシネーションの検知と抑止
- 応答速度の向上
  (インテルのハードウェアによる)

PII: 個人識別情報、PHI: 保護対象保険情報

参照: Saving Lives With AI: Prediction Guard Teams Up With Simwerx (英語)

参照: SimWerx - Prediction Guard を使用して AI で信頼性の高い救命ツールを実現

事例 2: データに基づく経営判断支援ツール

Antique Candle Co.

https://antiquecandleco.com/pages/our-candles (英語)

- キャンドルの製造、全米とカナダの販売店 (400 以上) への卸売、小売 (EC)

生成 AI アプリケーション
- 財務、在庫、販売、マーケティング・データの統合分析
- 自然言語による分析指示と回答

Prediction Guard が提供した価値
- 私有の使いやすい AI 機能
- 複数のモデルを選択、連携可能

参照: Antique Candle Co. - Prediction Guard の LLM を活用した分析によりデータに基づく意思決定を推進

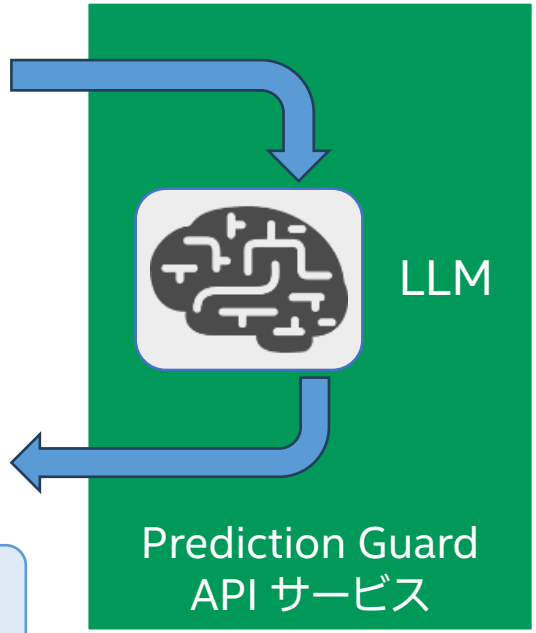
生成 AI のリスク評価は盛んに行われている

- OWASP TOP10 for LLM Applications (英語)
  - 非営利団体 OWASP による、生成 AI に関する重大脆弱性 10 件の紹介
    - 前年の研究、インシデント報告などに基づいた国際的な情報共有
  - 第一位はプロンプト・インジェクション　LLM01:2025 Prompt Injection (英語)
    - 自然言語による指示は LLM/生成 AI を混乱させることも可能
    - 機密情報の漏洩、誤った/偏った出力の誘発など、多くの問題の起点

絶対的なセキュリティーを提供できる製品またはコンポーネントはありません

Prediction Guard API の対応機能

- LLM 入力に対する事前処理
  - 🧬 /pii (個人特定情報の検出と前処理) (英語)
  - 😈 /injection (プロンプト・インジェクション検知) (英語)
- LLM 出力に対する検証
  - 🕵️ /factuality (テキストの事実性評価) (英語)
  - 🤬 /toxicity (テキストの有害性評価) (英語)

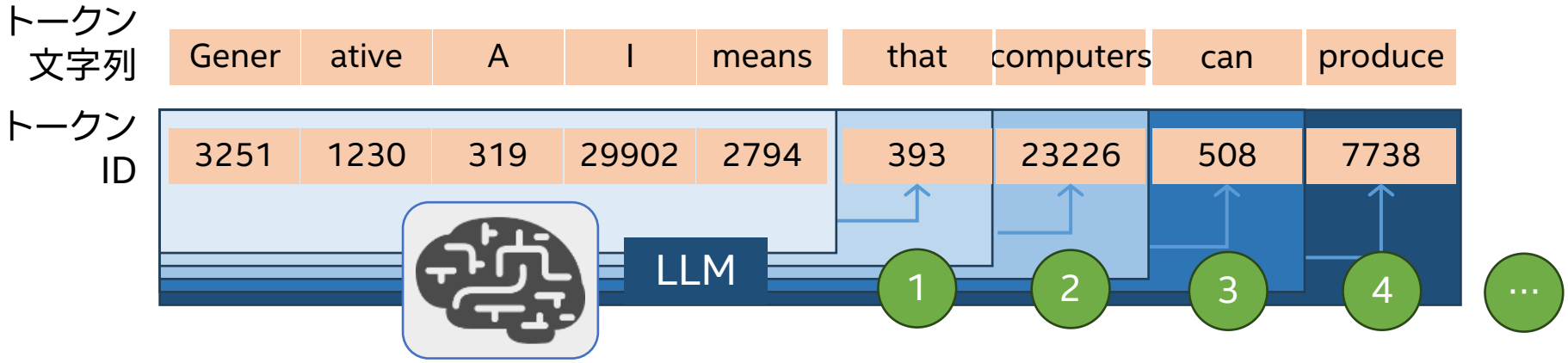


Prediction Guard によるリスク分析と対応のブログ記事
System level security for open source AI models (英語)

コスト計算の単位: LLM の入出力トークン数

- LLM はテキストを分解した「トークン」を取り扱う
  - 一度に生成するのは次のトークン 1 つのみ: 処理量はトークンの数に比例

入力テキスト例: Generative AI means



(補足) 日本語ではトークン数が多くなりやすい

生成テキスト全文
計 34 トークン

Generative AI means that computers can produce new content, such as images, text, or music, that resembles or mimics human creativity.

Gener	ative	_A	I	_means	_that	_comput ers	_can	_produc e	_new	_content	,	_such	_as	,	_images	,	_text	,	_or
3251	1230	319	29902	2794	393	23226	508	7738	716	2793	29892	1316	408	29892	4558	29892	1426	29892	470

_music	,	_that	_res	emb	les	_or	_m	im	ics	_human	_cre	ativity	.
4696	29892	393	620	1590	793	470	286	326	1199	5199	907	28157	29889

上記の参考訳
62 トークン

生成 AI とは、画像やテキスト、音楽など新しいコンテンツを、人間の創造性を真似て生成できるコンピューターのことです。

	生	成	_A	I		と	は	、	画	像	や	テ	キ	ス	ト	、	音	楽	な
29871	30486	30494	319	29902	29871	30364	30449	30330	31046	31551	31111	30572	30454	30255	30279	30330	30941	31739	30371

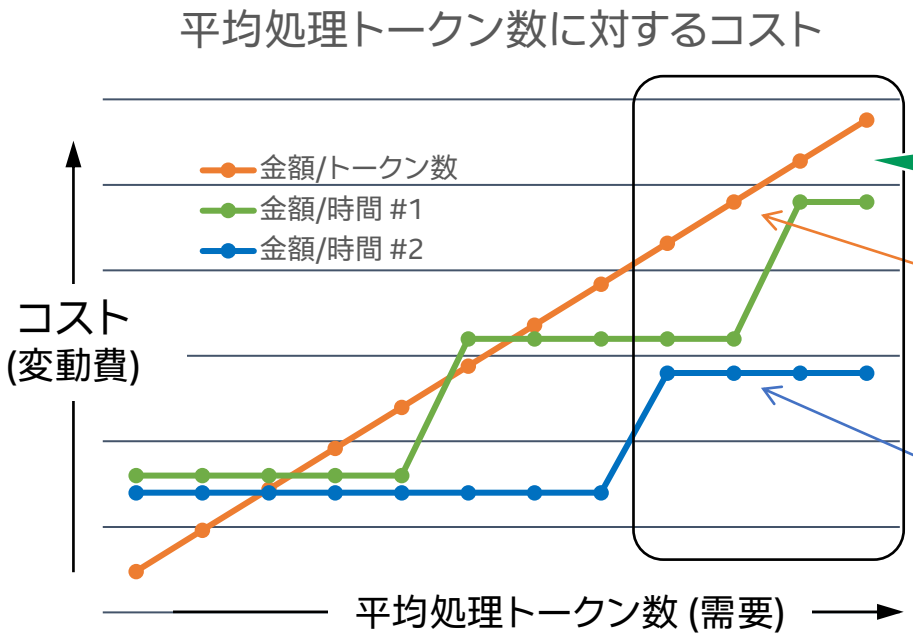
ど	新	し	い	コ	ン	テ	ン	ツ	、	人	間	の	創			造	性	を	真
30748	30374	30326	30298	30459	30203	30572	30203	30844	30330	30313	31069	30199	232	140	184	31420	30952	30396	30848

似		て	生	成	で	き	る	コ	ン	ピ	ュ	ー	タ	ー	の	こ	と	で
231	191	191	30466	30486	30494	30499	30538	30332	30459	30203	31172	30645	30185	30369	30185	30199	30589	30364	30499

す	。
30427	30267

使用モデル: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct (英語)

選択された私有のシステムは
トークン数に基づく有料 API に競争力を持つ



✓ 需要が高いときのコスト低減
✓ コストを予測しやすい

処理トークン数で
従量課金されるウェブ API

高スループットな私有のシステム
(クラウドまたはオンプレミス)

※ コストの関係は、各社の技術や
サービスによって変わる場合があります

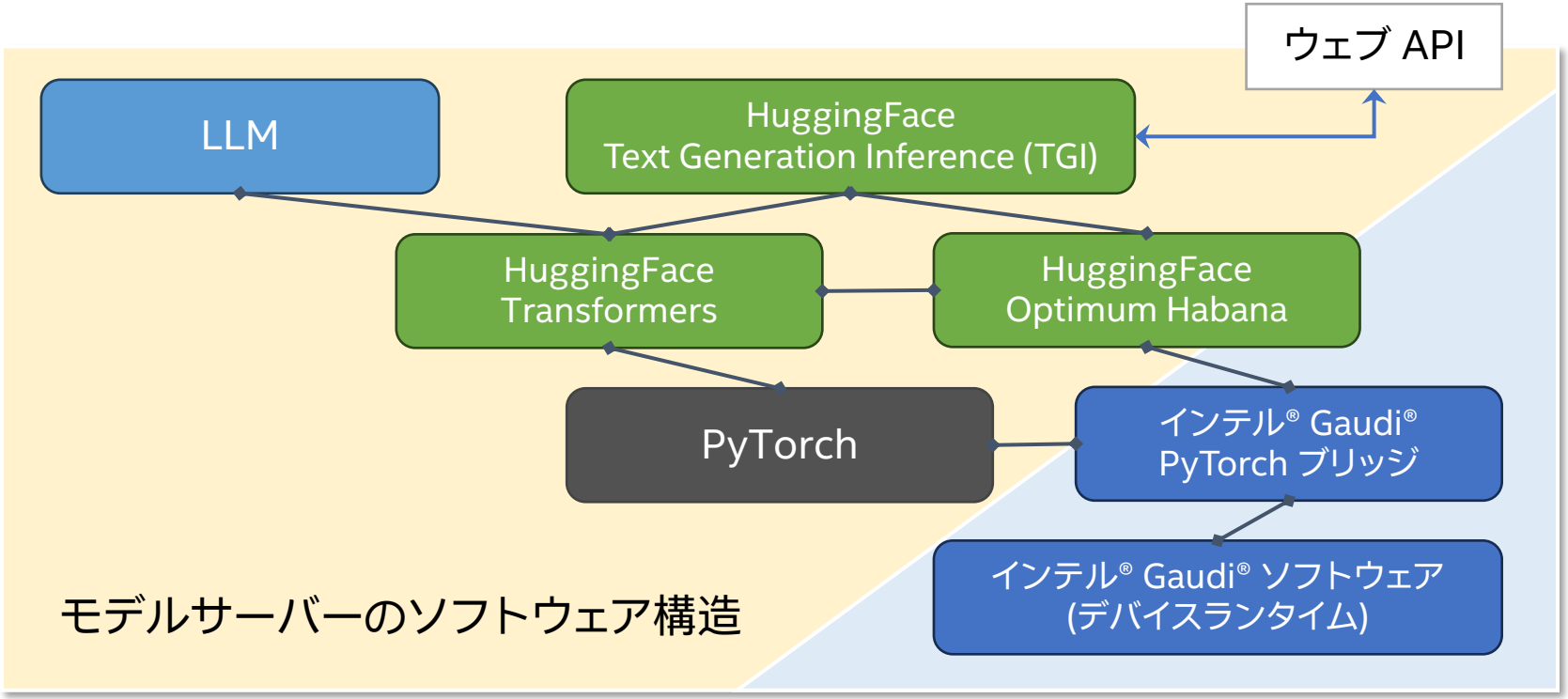
実行効率の良い実装でコスト削減を実現

- インテル® Gaudi® 2 AI アクセラレーター搭載インスタンスで最大限のパフォーマンスが得られるよう、モデルサーバーを調整
  - 静的形状: 入力トークン数を切り上げ、パディング
  - 動的バッチ処理: 一定の遅れの代わりにスループットを向上
  - 複数のシステムでの負荷分散 … など

参照: インテル® Gaudi® 2 AI アクセラレーター上での Prediction Guard のプライバシー保護 LLM プラットフォームのスケーリング

参照: 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

オープンソースによる共通化が
専用アクセラレーター採用の障壁を低減



モデルサーバーのソフトウェア構造

まとめ

- Prediction Guard は、インテル® Tiber™ AI クラウド上に
  信頼できる LLM/生成 AI 運用基盤システムを構築
  - 有益な生成 AI 利活用の提案を、予測可能で合理的なコストにより実施

- インテルのソフトウェアとサービスは、スムーズな移行を支援
  - オープンソース・フレームワーク PyTorch を基に、
    インテル固有のハードウェア機能を積極的に統合、有効化

お問い合わせはこちらまで
https://www.xlsoft.com/jp/qa