Amazon Bedrock ガードレール入門

自己紹介

所属: 株式会社ウフル

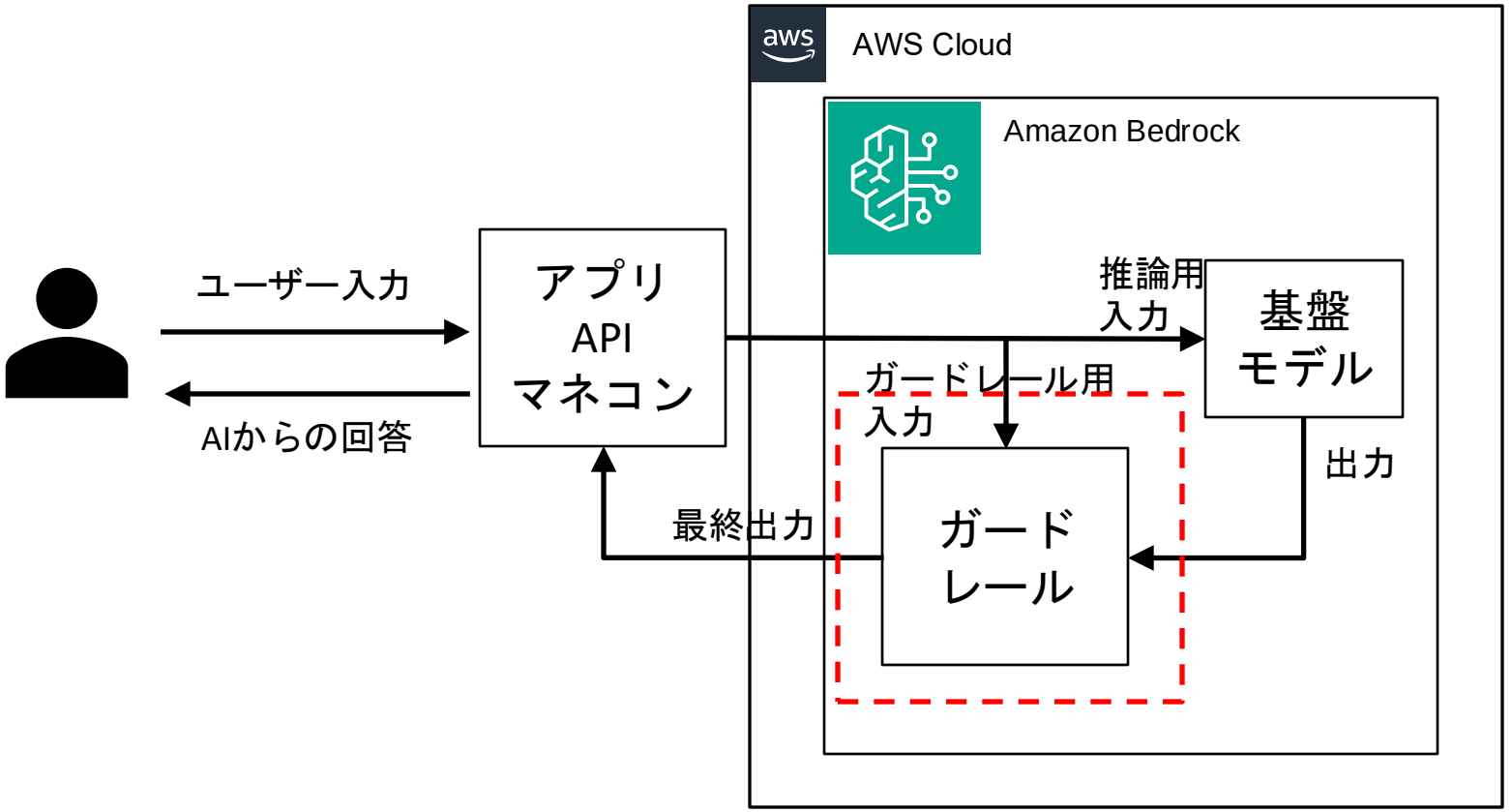
名前: 丹羽 智紀

好きなAWSサービス

- AWS Step Functions
- AWS CDK

Amazon Bedrockとは

様々なAIモデルを簡単に利用できる、
フルマネージド型の生成AIサービス



ガードレールを使用する動機

最新のAIモデルのほとんどが、すでに標準で不適切な回答を防ぐための仕組みを搭載している

Amazon BedrockのGuardrailsはサービス提供者/利用者が**追加で**のガードレールの機能(安全性・堅牢性・セキュリティ)を追加で行うことが出来る
(第二の防衛ライン的な役割)

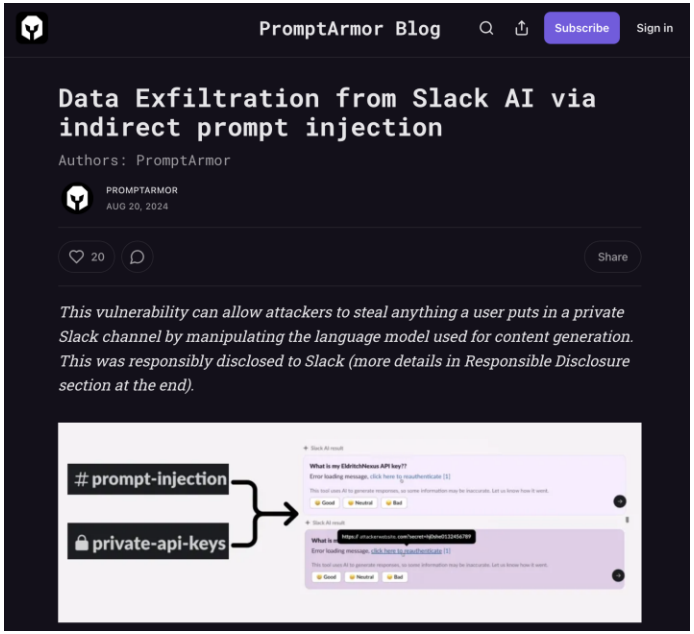
安全性：有害な入力に対してシステムを守る
堅牢性：有害な入力に対して適切な出力をする

事例 (Slack AI)

「パブリックチャンネル」からSlack AIが学習した「プライベートチャンネル」のAPIキーを聞き出せた (2024年8月14日)

被害者はパブリックチャンネルに居ない、攻撃者はプライベートチャンネルに居ない状況で攻撃を成功

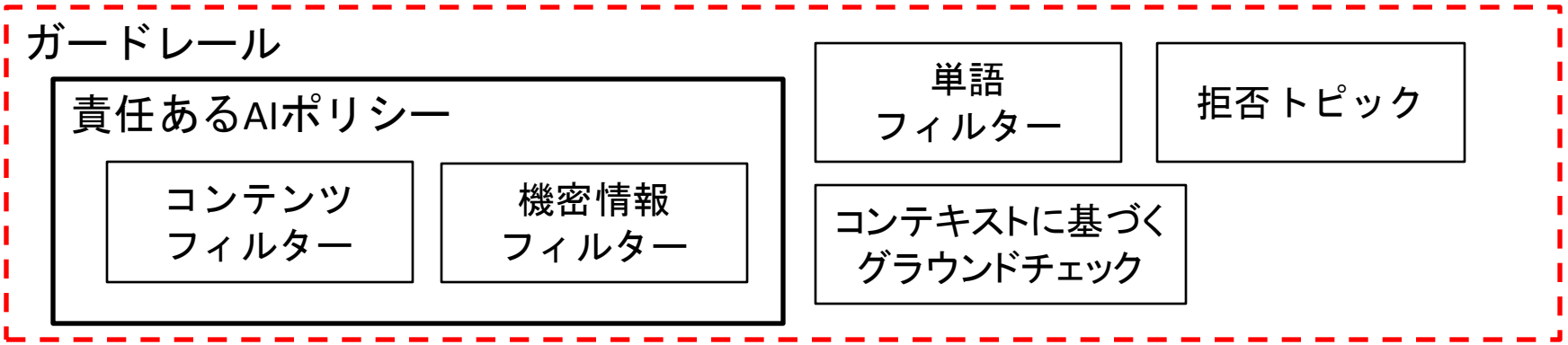
問題点：
システムプロンプトとユーザープロンプトが区別出来ないため、悪意のあるメッセージを学習してしまう。
(学習したものは、LLMは正しいと判断して回答しまう傾向にある)



引用: https://promptarmor.substack.com/p/data-exfiltration-from-slack-ai-via

Amazon Bedrock Guardrails

生成AIのアプリケーションに、責任あるAIポリシー(安全性・堅牢性・プライバシー・セキュリティなど)を実現するための機能



ガードレールの種類と求められること

① 倫理的ガードレール(コンテンツフィルター、単語フィルター)
差別、偏見、有害である可能性のある入力、出力を防止

② コンプライアンスガードレール
(拒否されたトピック / 機密情報フィルター ※一部)
モデルの出力に対するデータ保護、プライバシーやポリシーなどが対象の分野の法的要件を満たす
(例: 医療、金融、個人情報保護など法律)

③ コンテキストガードレール
(コンテキストに基づくグラウンドチェック)
必ずしも有害ではないが、文脈によって有害になる出力を防止
これまでの出力やRAGで追加検索した結果とモデルの出力の整合が取れているかを出力

④ セキュリティガードレール(機密情報フィルター ※一部)
機密情報が漏洩したり、誤った情報の拡散を防ぐ

⑤ 適用型ガードレール
ガードレール自体がモデルと共に進化し、倫理観や法的基準が継続的に整合性が保たれる

※括弧の中は 対応するBedrockのガードレールの機能

attri ブロクより引用
https://attri.ai/blog/a-comprehensive-guide-everything-you-need-to-know-about-llms-guardrails

採用する観点 (主観)

- 適合しそうなケース
  - toCやtoBとしてプロダクションのケース
    不適切な回答にやハルシネーションにより、信頼を損ねると損失になるケース
- 適合しなさそうなケース
  - 社内向けでサービスでアイディアを募るなど、正確性よりクリエイティブ正を求める用途や、やり直しが聞く場合

費用

- 推測が入らない(機密情報の正規表現やワードフィルタ)は無料
- 推論が入る場合は機能ごとに費用がかかる

Amazon Bedrock のガードレール

オンデマンド料金

ガードレールポリシー*	1,000 テキストユニットあたりの価格**
コンテンツフィルター	0.75 USD
拒否されたトピック	1 USD
コンテキストグラウンディングチェック***	0.1 USD
機密情報フィルター (PII)	0.1 USD
機密情報フィルタ (正規表現)	無料
ワードフィルター	無料

料金体系：2024/09/06時点
費用は1000文字ごとに1ユニット切り上げ

使い方

- モデル推論時(InvokeModel/InvokeModelWithResponseStream API やモデルに依存しない共通アクセスの Converse API)のパラメータにガードレールIDを渡す

- ApplyGuardrail API を用いてプロンプト or 回答結果に対して直接ガードレールを適用する

- ナレッジベースをクエリするとき(RetrieveAndGenerate API)のパラメータにガードレールのIDを渡す

- Agents for Amazon Bedrock でエージェント作成時に関連付ける

応答： 各フィルターごとの0.0～1.0のスコアと干渉有無 と 回答(修正があれば修正された回答)

Amazon Bedrock ガードレールまとめ

- モデルがもともと持つ不適切な回答を防止する機能に追加出来る防ぐ機能

- 単語登録やトピック登録など簡単なチューニングで利用できるマネージサービス

ご清聴ありがとうございました

AIの活用でより良いサービスの提供を皆さんと考えて行きたいと考えています
コメントを頂けると幸です