

生成AIで迷惑SMS対策やってみた

KDDI株式会社 コア技術統括本部 オペレーション本部
プラットフォームエンジニアリング部 鹿野良太

2025/10/24

▶ 自己紹介



■ 名前

・ 鹿野 良太

・ 経歴

・ 2024年度～

・ KDDI株式会社入社

・ SMS設備担当

→SMS設備の開発・保守運用

→生成AIを使用した迷惑SMS対策

■ 出身

・ 東京都

■ 趣味

・ 美味しいものを食べる

→高知で鰹のたたきを食いたい

今回の発表テーマ

・生成AIを活用するに至った背景

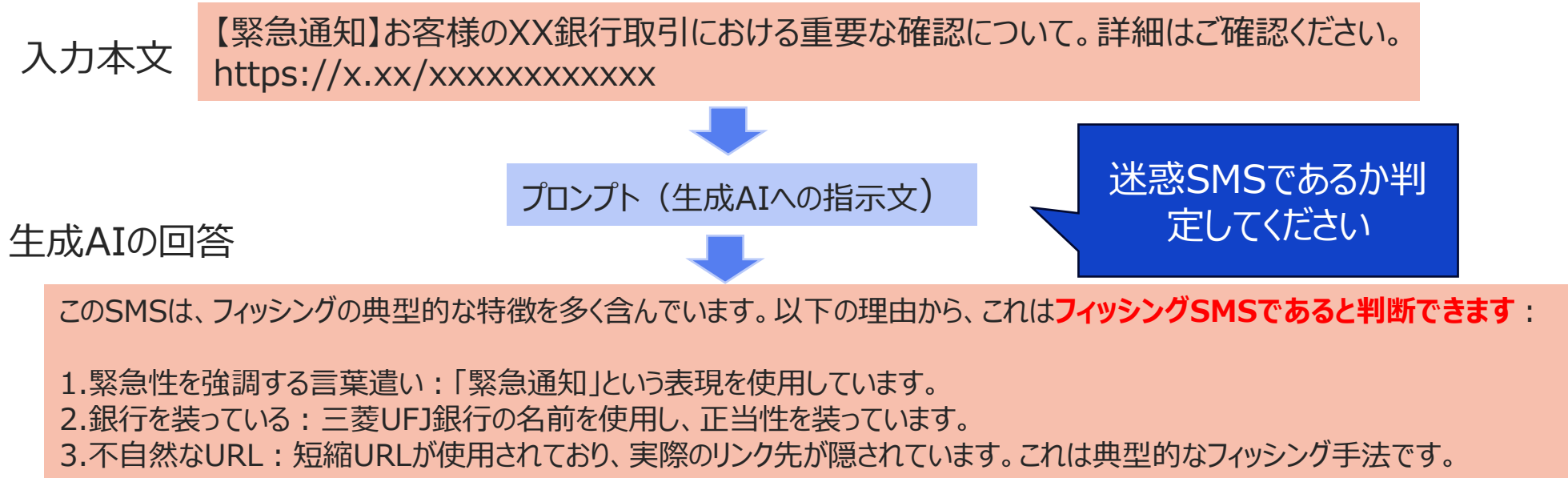
<従来の課題>

・迷惑SMSの疑いのあるお客様からの申告データをもとに、SMS設備に適用する迷惑SMSブロック用のファイルを人力で作成していた

稼働	目視で正規SMSと迷惑SMSに分類し ファイル作成 専任で約1.5営業日	鮮度	分析～設定までに一週間 迷惑SMSのトレンドを追えない	精度	一部の申告データのみで分析作業 を行うためブロック率が上がらない
----	---	----	--------------------------------	----	-------------------------------------

<仮説> 生成AIは申告SMSから正規のSMSと迷惑SMSを判別できるのではないかと。もし可能であれば、課題の解決になる。

<検証> 迷惑SMSの判定を指示するプロンプトを作成し、実際の迷惑SMSを入力



<結果> **生成AIは正規のSMSか迷惑SMSかおおむね判別できる。**

しかし、商用適用にあたって、正規SMSを迷惑SMSと誤判定する（FP）をゼロにすることが必要である。

▶ FP=0への取り組み ~SMS本文判定~

■ FPゼロ対策実施内容

- ・無理に白か黒かの2値分類をさせない

→わからないものはわからない（判定不可）と答えてもらう

■ 各社HPの注意喚起をプロンプトに反映

- ・業界ごとに判定プロンプトを作成し、各社HP上で「この本文（URL）に注意」と、**注意喚起されている内容をプロンプトへ反映**

→しかし、各社の公式HP内容を継続的に追従する必要がある

【迷惑メール・SMS】最近多い迷惑メール・詐欺メールの事例と特徴を知りたい

KDDIやauを装う迷惑SMS

Q 【迷惑メール・SMS】 KDDIやauを装う迷惑SMSの特徴を知りたい

A KDDIやau、通信会社を装う迷惑SMSの事例です。
迷惑SMSに記載されたURLを開いたことで、不正アプリのダウンロード、偽のMy auへのログイン、KDDIで受け付けていない電子マネー（ギフトカード）での料金支払いなどが求められる事例が報告されています。

KDDIやauが下記のようなSMSをお客さまに配信することは一切ありません。

身に覚えのないものや不審に感じた場合は、記載のURLなどは絶対に開かないようにご注意ください。

KDDI系に特化させたプロンプトの例

```
<task>
  <role>あなたは大手通信会社セキュリティ
  <instructions>
    SMSの文章を見て、<KDDI、au、UQモバイル、POVOなどのサービス>のサービス]</KDDI、au、UQモバイル、POVOなどのサービス>であるか判定してください。
    以下のフィッシングSMSの特徴を参考に
  </instructions>
```

各社HPのURLとHTML文をDBに保存し更新があるたびに**赤枠プロンプト**を**自動再作成**

```
<features>
  <feature># KDDI、au、UQモバイル、POVOなどのサービスにおける正規のSMSの特徴:
  - 公式アプリやあらかじめ登録したブックマークからアクセスしてログインする
  - 2段階認証時は内容をしっかり確認する
  - アプリのインストールは公式サイトから行う
  - 送信元が正規のものである(例: auto@connect.auone.jp)
  - 公式サイトURLが正しく記載されている
  - 正式な社名・サービス名が使用されている
  - 日本語として自然な文章で書かれている

  # KDDI、au、UQモバイル、POVOなどのサービスにおける迷惑SMSの特徴:
  - auやKDDIを装ったメッセージとURLが記載されている
  -
  -
  "訴訟最終通知"</feature>
</features>
```

```
<judgmentCriteria>
  <criteria>#判定結果:</criteria>
```

```
<options>
  <option>フィッシングの場合: black</option>
  <option>フィッシングでない場合: white</option>
  <option>判定不可: gray</option>
</options>
```

```
</judgmentCriteria>
```

判定結果は分析しやすい形式で指定

▶ FP=0への取り組み ~HTML判定~

■ 判定材料がSMS本文だけでは FP=0が実現せず

実例

お荷物について
<https://xxxxxxxx-xxxxxx.xx/xxxxxxxx/xxxxxx/xxxx/xxxx>

URL先は正規の
ホテルサイト

HTMLの中に不審なリンク
が含まれている

HPに連絡先や問い
合わせ情報が載って
いない

HTMLの構文が不自然

HTML文 判定プロンプトの例

```
<role>あなたは大手通信会社セキュリティ部門のプロフェッショナルです。</role>
<instructions>
  URLのHTMLの中身を見て、迷惑SMSメッセージであるか判定してください。
  以下の迷惑SMSメッセージのHTMLの特徴を参考にしてください。
</instructions>
<features>
  <feature>不審なリンクやURLが含まれている</feature>
  <feature>緊急性を強調する表現が使われている</feature>
  <feature>送信者の正当性が疑わしい（公式なものではない）</feature>
  <feature>個人情報の提供を要求する内容</feature>
  <feature>文法やスペルの誤りが多い</feature>
  <feature>一般的な挨拶や名前の欠如</feature>
  <feature>偽のロゴやデザインが使用されている</feature>
  <feature>連絡先情報や問い合わせ先の情報が明確に記載されていない</feature>
  <feature>html構文が不自然</feature>
</features>
<judgmentCriteria>
  <criteria>#判定結果：</criteria>
  <options>
    <option>迷惑SMSの場合: black</option>
    <option>迷惑SMSでない場合: white</option>
    <option>判定不可: gray</option>
  </options>
</judgmentCriteria>
<outputInstructions>
  <instruction>#出力方法：</instruction>
  <format><ans>判定結果</ans>(出力では<ans>判定結果</ans>のみ出力してください)
</format>
</outputInstructions>
<example>
  <description>例：</description>
  <sampleInput>入力: HTMLの中身</sampleInput>
  <sampleOutput>出力:<ans>black</ans></sampleOutput>
</example>
<prompt>
  以下のURLのHTMLの中身を判定してください：
```

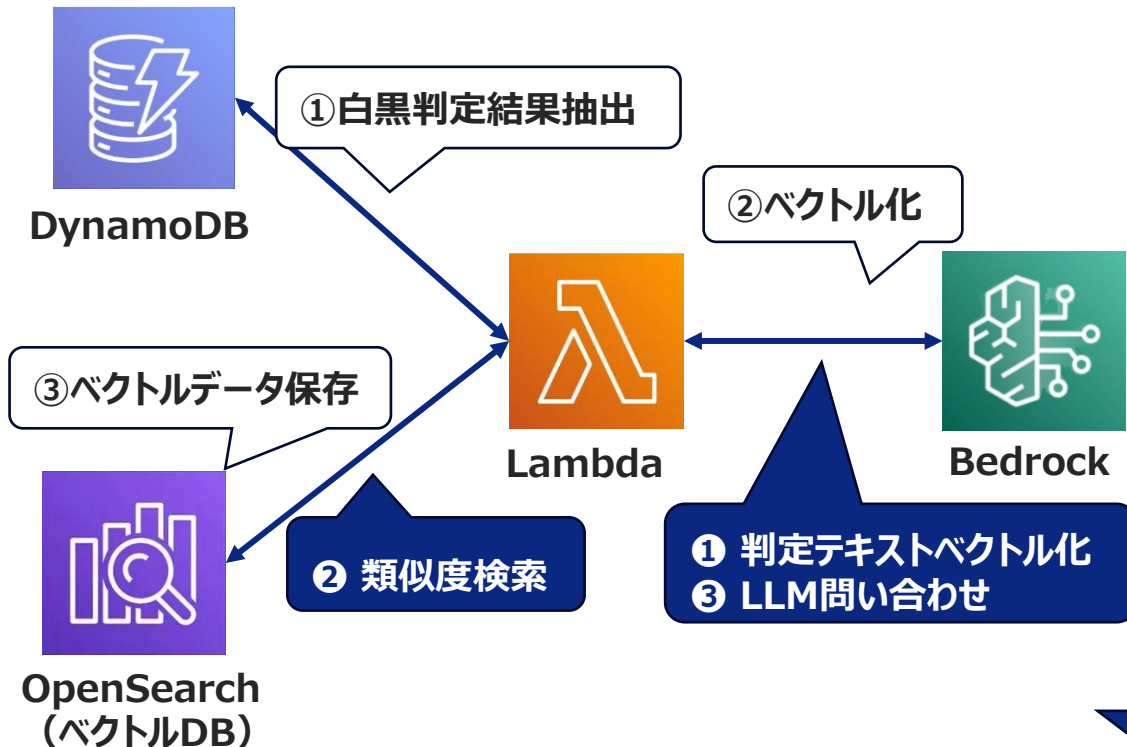
- URL先のHTML文も生成AIの判定材料、
特にSMS本文では読み取れない情報を判断材料
に白黒判定を行う
→本文とHTMLの2重で判定することで誤判定を防ぐ

▶ FP=0への取り組み ～RAGの導入検討～

■ 精度を更に高めるために

過去に白黒判定した本文をベクトル化（数値化）しDBに保存し、今後判定するSMSも同様にベクトル化して、過去のナレッジと類似度を計算することで、白黒判定に活用したい

• AWS構成図



• 類似度検索検証結果

入力 黒データ

“○更新有：期限内にアクセスの
ほど、お願いいたします。
http://xxx.xxx/xxxxxxxxxxx
xxxxx”

入力 白データ

“お荷物について
[https://xxxxxxxx-
xxxxxx.xx/xxxxxxxx/xxxxxx
xxxxx/xxxxx](https://xxxxxxxx-xxxxxx.xx/xxxxxxxx/xxxxxx
xxxxx/xxxxx)”

類似度データ抽出結果

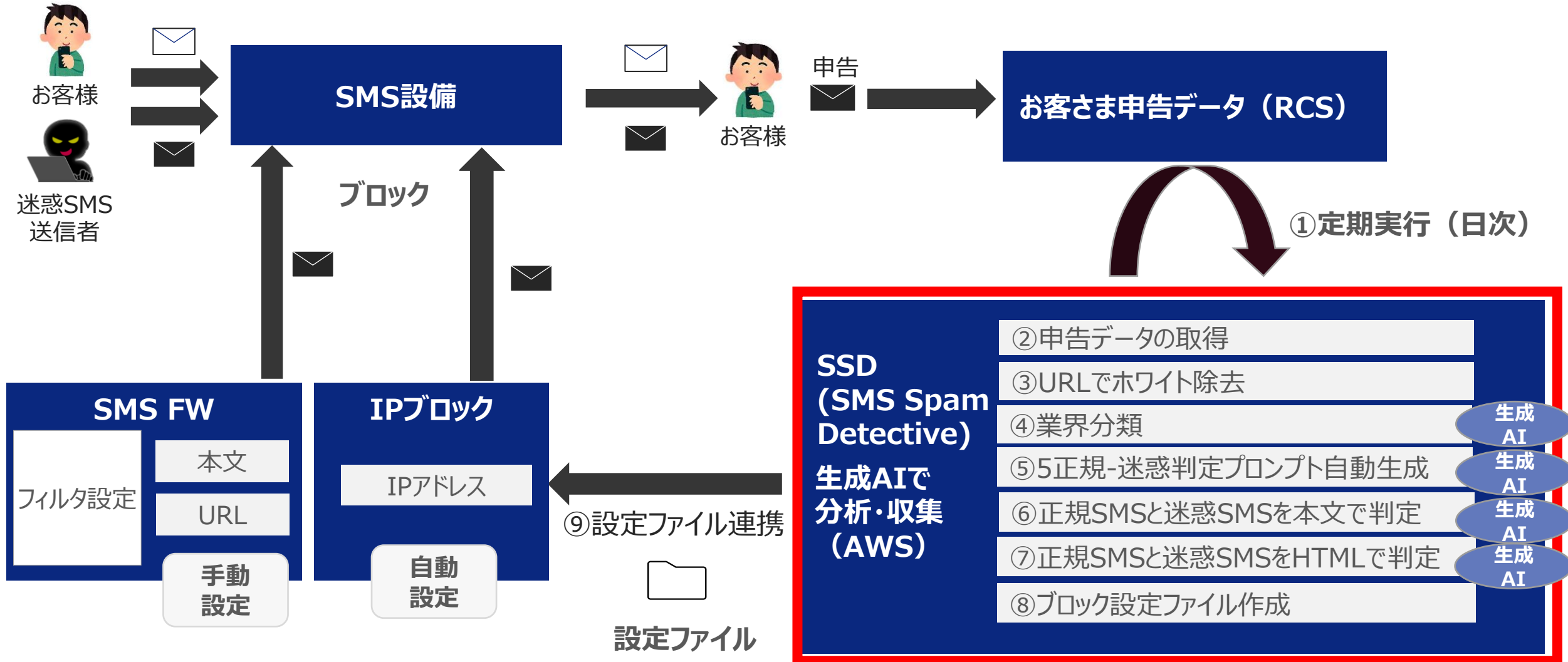
{'text': '期限を過ぎますと受取不可
となってしまいます。
[https://xxxxxxxx.xxx/xxxxxx/xxxxxx
xxxxxx](https://xxxxxxxx.xxx/xxxxxx/xxxxxx
xxxxxx),
'label': '迷惑SMS', 'src': 'finance',
'類似度': 0.889452}

{'text': '配達できず保管中。ご都合
の良い時間帯をお知らせください。
http://xxxxxxxxxxx.xxxxxx.xxxx/xx
xxxxx,
'label': '迷惑SMS', 'src': 'delivery',
'類似度': 0.8745303}

cos類似度計算のため1が完全一
致、-1が最大の外れ値

類似度判定の結果をどのように本システムに活用するかが肝に

構築したシステム(SSD)の概要



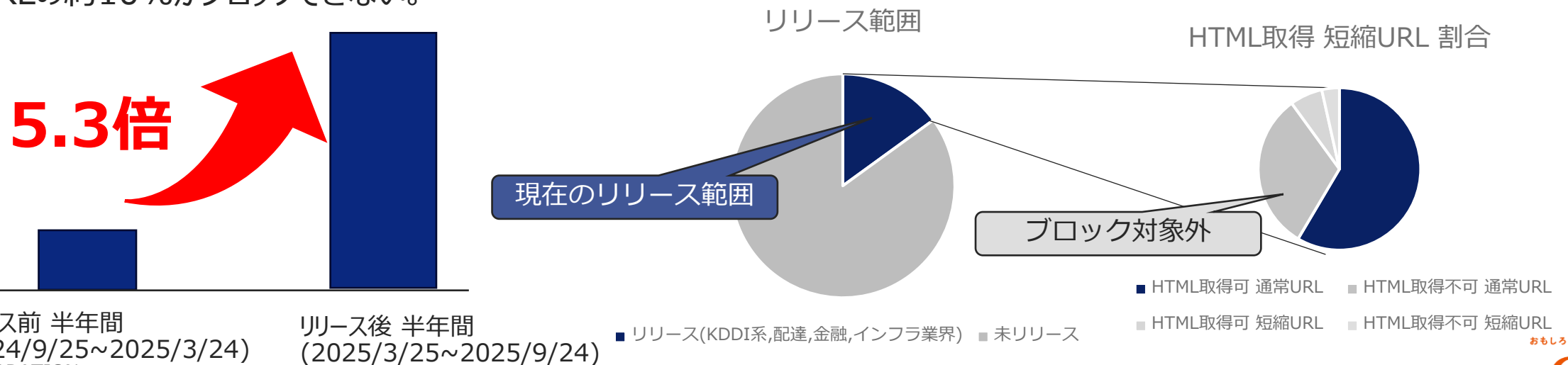
運用実績と課題

■ 運用実績(2025/3/25~2025/9/24)

- ・ リリース範囲：業界分類でKDDI、配達、金融、インフラ業界に分類された全体15%の出力IPアドレスだけSMS設備と連携
- ・ 半年間 FP数：**0件**
- ・ IPブロック数 リリース前後の半年間の比較で**5.3倍**に

■ 現状の課題

- ・ HTML判定をする際、リンク先サイトが削除されており、HTMLを取得出来ずに判定不可が35%。
- ・ 短縮URLから返されるIPアドレスをブロックすると、正規のSMSもブロックしてしまう可能性があるため、短縮URLの約10%がブロックできない。



・ 今後の方針と活動を通じての学び

■ 今後の方針

- ・ 本文判定にRAGを取り入れる
- ・ 業界分類の残り85%リリースする
- ・ IPブロックだけではなくURLブロックを取れ入れる
- ・ 迷惑メール対策にも適用する

■ 活動を通じて

- ・ 生成AIをシステム内で利用する際の難しさを感じた。
→生成AI出力結果の確認は最終的に人が行うことになり、リリース後もどの程度まで確認を続けるべきか、不明確になる
- ・ 生成AIはブラックボックスであり、出力結果はモデルの性能の良さに依存してしまうと考えていたが、
プロンプトやRAGなどの工夫により出力結果を変えられる

「つなぐチカラ」を進化させ、
誰もが思いを実現できる社会をつくる。

KDDI VISION 2030

