

問い合わせ履歴の匿名化と利用： プライバシー保護を確保したチャットボット開発

森山 有理名^{1,2,a)} 魚井 宏高³

受付日 2024年5月13日, 採録日 2024年8月27日

概要: 本稿では、問い合わせ履歴の匿名化処理を行い、そのデータを利用したチャットボットを開発した。提供しているサービスに関するドキュメントに加え問い合わせ履歴を活用することで、個人情報を保護しながらもより適切な回答を生成することを目指した。開発されたチャットボットは、社内業務における問い合わせ処理の効率化に貢献し、将来的に顧客対応の自動化を目指す。

キーワード: 生成 AI, RAG, 個人情報, チャットボット, プライバシー保護

Anonymizing and Utilizing Inquiry Histories: Developing a Privacy-Preserving Chatbot

ARINA MORIYAMA^{1,2,a)} HIROTAKA UOI³

Received: May 13, 2024, Accepted: August 27, 2024

Abstract: In this paper, we developed a chatbot utilizing anonymized inquiry histories. By leveraging documentation related to the services provided as well as the inquiry histories, we aimed to generate more appropriate responses while protecting personal information. The developed chatbot has contributed to the efficiency of internal operations by streamlining inquiry processing and aims to automate customer support in the future.

Keywords: LLM, RAG, PII, chatbot, privacy

1. はじめに

近年生成 AI 技術の発展はめざましく、文章、画像、音楽、動画といった様々なコンテンツを、人間が作ったものに近いレベルで生成することが可能になってきている。

その中でも、LLM (Large language Models) は Llama [1], Claude [2], GPT [3] など国内外の多くの事業者が開発しており、一般消費者が使えるサービスとして日本語での自然な文章を生成することが可能なモデルが提供され始めている。

文章生成は質問文に対し応答文を生成するチャットボットとして広く使われており、企業のサポートに利用されるケースもみられる。

経済産業省と総務省が取りまとめた AI 事業者ガイドライン [4] では、AI 事業者は“AI 開発者”、“AI 提供者”、“AI 利用者”の3つに大別される。LLM を使用してチャットボットを提供する事業者は AI 提供者にあたる。

AI 提供者はプライバシー保護に関して、プライバシー・バイ・デザインの観点から

“AI システムの実装の過程を通じて、採用する技術の特性に照らし適切に個人情報へのアクセスを管理・制限する仕組みの導入等のプライバシー保護のための対策を講ずる” [5] 必要があると記載されている。

生成 AI を利用したチャットボットの運営にはプライバシー保護に関して懸念が存在する。既存の問い合わせ事例を元に学習をするとき、既存の LLM に対しトレーニングデータと別に知識ベースを検索し、参照して応答を作成する検索拡張生成 (retrieval augmented generation

¹ 大阪電気通信大学 大学院
Osaka Electro-Communication University, Shijyouawate, Osaka
575-0063, Japan

² 株式会社 Urbs
Urbs Inc., Osaka 541-0045, Japan

³ 大阪電気通信大学 総合情報学部
Osaka Electro-Communication University, Shijyouawate, Osaka
575-0063, Japan

a) dt22a001@oecu.jp

(RAG)[6]を行うときに入力情報に個人情報が含まれてしまい、生成 AI が回答文を出力する際に個人情報を出力してしまう可能性がある。チャットボットを公開して顧客が利用する場合は、利用者が自身の個人情報を入力してしまう可能性がある。

機械学習を使用して入出力された情報に個人情報 (personally identifiable information (PII)) が含まれているかどうかを検出する AWS Comprehend といったサービス [7] も存在するが、Comprehend の個人情報検出は日本語に対応しておらず、ユーザがカスタマイズすることもできない。日本語での個人情報の検出を自動化することは難しいといえる。本稿では LLM を個人情報検出に使用する実証も目的であったため、Comprehend を使用しなかった。

また、問い合わせチャットボットである場合、提供しているサービスの実態に即した事実に基づいた返答をする必要がある。しかし、生成 AI にはハルシネーションを起こすリスクがあり、間違った返答を事実であるかのように返答する。RAG により、信頼できるドキュメントを提供してより正確な回答を生成する手法が存在する。

2. 目的

株式会社 Urbs では、学会・国際会議を主なターゲットとした決済、参加者管理オンラインサービス Payvent を提供している [8]。Payvent にはイベント主催者、イベント運営代行業者、イベント参加者、決済事業者など複数の立場の人物が関わり、利用方法が立場によって大きく異なる。そのため、問い合わせ者の立場に合わせた返答を提供する必要があり、問い合わせの返答に要するコストが大きい。また、間違った案内をしてしまう懸念も存在する。

株式会社 Urbs は少人数の組織であるため、問い合わせへの返答コストを下げるために生成 AI を利用した。

顧客への問い合わせに対し適切な回答をするチャットボットを作成することが最終的な目標である。しかし、RAG を使用して返答精度を高められたとしてもハルシネーションを完全に無くすことは不可能であると考えられるため、本稿では前段としてサポートスタッフが問い合わせに返答する際に返答案を生成することで返答にかかる時間コストを減らすことを目的とした。

既存の問い合わせ情報とヘルプとして提供しているドキュメントや利用規約を元に問い合わせに返答するチャットボットを想定し、以下の生成 AI を作成した。

- ・ 既存の問い合わせ情報から個人情報を除去して匿名化する生成 AI
- ・ 問い合わせ返答を評価するための架空の問い合わせをデータセットとして生成する生成 AI
- ・ ヘルプセンターおよび利用規約、プライバシーポリシーを RAG で検索し、返答を生成する生成 AI
- ・ 上記に加え、個人情報を除去した既存の問い合わせを

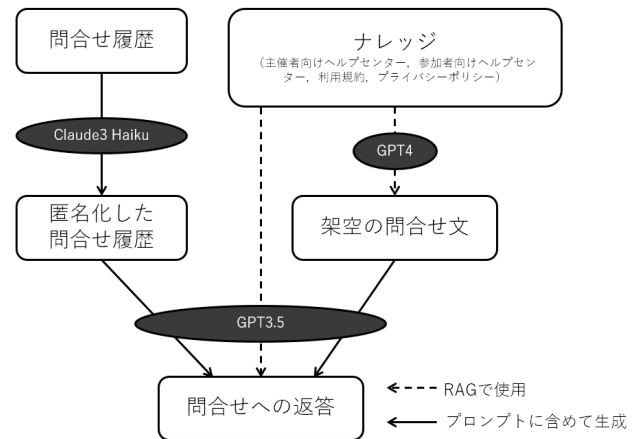


図 1 全体の生成の流れ

Fig. 1 Overall generation flow.

RAG で検索し、返答を生成する生成 AI

全体の生成処理の流れを図 1 に示す。

なお、現段階では社内でのみ使用しており顧客への公開は行っていない。Payvent は決済サービスであるため、金銭が絡む問い合わせは誤答による被害が大きいことが予想される。また、返金や不審請求対応など、人の意思決定が必要とされる問い合わせも多く、チャットボットですべての問い合わせに対応することは不可能であると考えられる。したがって、顧客への公開の際は操作方法や機能、仕様に関する問い合わせのような、返答が一意に定まるものについてはチャットボットが返答し、そうでないものはサポートスタッフにエスカレーションするようにチャットボットに判断させる運用を想定している。

3. 個人情報保護のための匿名化

生成 AI を活用するにあたり、既存の問い合わせ情報を利用することを想定した。問い合わせ情報は個別の事例であり、個人名やメールアドレス、電話番号等の個人情報が含まれることから、生成 AI チャットボットにそのまま入力してしまうと、チャットボットが回答に個人情報を使ってしまうリスクが存在する。

したがって、問い合わせ情報の履歴から個人情報を除去する試みを行った。なお、個人情報を研究開発に使用することはプライバシーポリシーに記載されている。AWS bedrock [9] を利用して Anthropic Claude3 Haiku を呼び出し、個人情報を削除する試みを行った。

プロンプトは以下のとおりである。発見した個人情報を { 項目名 } の形に置換させる指示を記述した。temperature は生成される回答のランダム性の量を指定するパラメーターであり、0.0 から 1.0 の値を取る。文章の創造の際はこの値を上げることになるが、今回は元文章を維持することが求められるため 0.0 を指定した。System 内の {} は入力された問い合わせ文章に置換される。名前の定義はプロンプトに記載していないため、人名の検出は Claude3

Haiku の判断による.

System
<p>以下はあなたがこのタスクで参照すべき文書です。 {}</p> <p>以下はこのタスクの基本ルールです。</p> <p><rule> 個人情報を除去してください。個人情報以外の部分は元データそのまま出力してください。途中でやめず、入力された情報の最後まで置き換えてください。 変換したデータのみを出力してください。 判断基準は以下のとおりです。 “-”や“(“や”)”が含まれる可能性のある 10 桁程度の数字は電話番号です。 半角文字で @ を挟んでいる文字列はメールアドレスです。 “県”や“市”といった文字が含まれる文字列は住所です。 個人情報とは以下の情報です。：以降の文字列に置き換えてください。 ・名前：{name} ・電話番号：{tel} ・メールアドレス {email} ・住所 {address} ・銀行口座番号 {bank}</p> <p></rule></p> <p>あなたは企業の事務担当で個人情報の文書からの除去を依頼されています。</p>
Human
<p>Human: ルールにしたがって <ticket> 内の文書から個人情報を除去し変換した XML を出力してください。</p> <p>Assistant:</p>

Claude では XML を使用することが推奨されている [10] ため、文章の入力には XML を使用し、プロンプト内で XML タグを参照した。

株式会社 Urbs では問い合わせ情報の管理に株式会社インゲージの Re:lation [11] を使用している。Re:relation が出力した CSV ファイルを XML ファイルに変換し、入力に使用した。XML ファイルの形式は以下のとおりである。

```

<root>
<ticket>
<ticketid> 問合せ ID</ticketid>
<receive> 問合せとして受け取った文章 </receive>
<send> 問合せに返答した文章 </send>
</ticket>
</root>
    
```

<ticket> は問合せの件数だけ存在し、<ticket> には <receive> タグと <send> タグがそれぞれ 0 件以上含まれる。チケットの数は 150, そのうち返答が含まれたチケットの数は 127 である。

Python を用いたプログラムを作り、<ticket> ごとに上記のプロンプトで個人情報を除去した XML を生成させた。その後、人間の目視によるチェックで残留する個人情報を確認し除去した。生成 AI の出力では、情報の除去精度に加えて元情報からの不必要な改変が含まれる懸念がある。本稿では、問い合わせ 1 件ごとに個人情報を除去した文章を生成し、都度生成結果に不必要な改変が含まれていないかチェックを行い、文章が改変されていた場合再生成を行った。生成 AI 出力時点での置換結果と、人間のチェック後の置換結果を以下に示す (図 2) (表 1)。

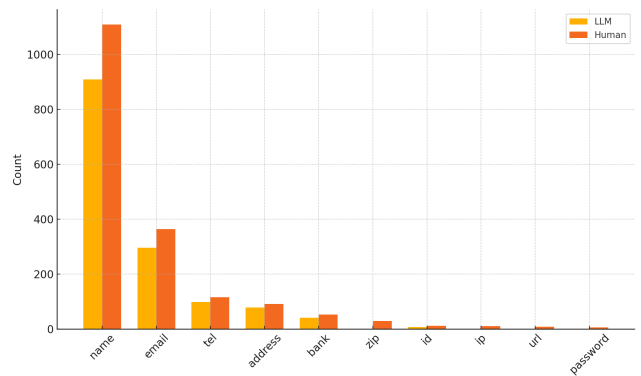


図 2 生成 AI と人間のチェック後の個人情報置換数

Fig. 2 PII replacements count of generated by LLM and checked by human.

表 1 生成 AI と人間のチェック後の個人情報置換数

Table 1 PII replacements count of generated by LLM and checked by human.

	生成 AI	人のチェック後	生成 AI で除去できた割合
Name(個人名)	909	1108	82%
Email(メールアドレス)	296	363	81.5%
Tel(電話番号)	98	116	84.5%
Address(住所)	78	91	85.7%
Bank(銀行口座番号)	41	53	77.4%
Id(ID)	7	11	63.6%
Ip(IP アドレス)	0	10	0%
Password(パスワード)	0	6	0%
Zip(郵便番号)	0	29	0%
url(URL)	0	8	0%
合計	1429	1795	79.6%

前述のプロンプトのとおり、指示したものは名前、電話番号、メールアドレス、住所、銀行口座番号であるが、アカウントのIDについては指示に含まれていないにも関わらず置換を行った。人間のチェック後に追加された項目はチェック者の主観による。

名前、メールアドレス、住所、電話番号は具体的な定義をプロンプトで与えたため精度が比較的高く、80%を超えて生成AIでの除去ができています。銀行口座番号やIDに関しては定義を与えていないにも関わらず60%を超えて除去ができています。

生成AIで置換できず、人間がチェックした際に置換したものを以下に例示する。掲載にあたり一部情報を■に置換した。

置換前：お世話になっております、Payvent 運営サポート担当、林です。
置換後：お世話になっております、Payvent 運営サポート担当、{name} です。
置換前：デバイスの場所：Japan, Ōsaka, Yao (■■■■■■)
置換後：デバイスの場所：Japan, Ōsaka, Yao ({ip})
置換前：〒■■■■■ {address}
置換後：〒{zip} {address}
置換前：イベント名「第■■■回■■■会ならびに学術講演会（管理アドレス：■■■■■■■■）」において
置換後：イベント名「第■■■回■■■会ならびに学術講演会（管理アドレス：{email}）」において

林はPayventサポートスタッフの名字である。このような、文字数が少なく一般名詞と同じ名称は固有名であると判断できない場合があり、置換漏れが起こったと考えられる。また、メールアドレスについてはプロンプトの定義と一致するにも関わらず漏れがあった。

置換するようにプロンプトに指示をしたにも関わらず置換できなかったものについては、文単位で判断させるように指示を追加する対応が考えられる。アドレスという単語が文内にあればメールアドレスが近くに存在する可能性が高く、単語が主語である（“が”、“は”が後に続く）場合や挨拶の直後にある場合は固有名である可能性が高いと考えられる。

個人情報の除去前と除去後の比較を以下に例示する。掲載にあたり一部情報を■に置換した。会社名や個人情報は特にメールのフッターに多く含まれている。

個人情報の除去は絶対的に行われなくてはならず、少しでも漏れが発生してしまうことで漏洩リスクが発生するため、いずれの場合でも、人間のチェックは必要となる。

しかし、合計で比較して80%近い個人情報を生成AIで除去することに成功しており、すべてを人間が置換するよ

個人情報の除去前	個人情報の除去後
Payvent 運営サポート担当 林様 平素より大変お世話になっております。 株式会社■■■の■■■です。 先日、3つのイベントの入金申請を同日に行ったのですが、エラーで入金できなかったと 貴社からご連絡がありました。 口座名義が違っているのですが、以前に別のイベントの入金申請を行い、無事に完了してから 内容は何も触っていない状態です。設定口座の変更や情報修正も行っておりません。 解決策をご教示いただけましたら幸いです。	Payvent 運営サポート担当 {name} 様 平素より大変お世話になっております。 株式会社■■■の{name}です。 先日、3つのイベントの入金申請を同日に行ったのですが、エラーで入金できなかったと 貴社からご連絡がありました。 口座名義が違っているのですが、以前に別のイベントの入金申請を行い、無事に完了してから 内容は何も触っていない状態です。設定口座の変更や情報修正も行っておりません。 解決策をご教示いただけましたら幸いです。
■■氏名 ■ローマ字氏名■■ 株式会社■■■ TEL：06-■■■-■■■ FAX：06-■■■-■■■ Email: ■■■@■■■.co.jp Web: https://www.■■■.co.jp/ 〒■■郵便番号■■ 大阪府大阪市■■区■■■■番■■号 ■■ビル 5F	{name} 株式会社■■■ TEL：{tel} FAX：{tel} Email: {email} Web: https://www.■■■.co.jp/ 〒{address}

りもはるかに人的コストを抑えることができるという知見が得られた。

4. 架空の問い合わせの生成

問い合わせチャットボットを評価するためのデータセットとして、架空の問い合わせを生成する生成AIを作成した。OpenAI API から gpt-4-turbo-preview を使用した。Payvent は前述のとおり様々な立場の人物が利用し、問い合わせ内容も立場によって異なるため、実際の問い合わせ

に近づけるために、問い合わせ者のペルソナを想定させた。プロンプトは以下を使用した。

この GPT はウェブサービス「Payvent」に来ると思われる問合せ文章を想像して XML 形式で出力します。
 まず、架空の Payvent に対する質問者を想像します。質問者は以下の情報を持ちます。

- ・法人または個人
- ・法人の場合、法人名
- ・質問者名
- ・業種
- ・職種
- ・主催者または参加者
- ・Payvent を既に利用しているか、これから利用しようとしているか
- ・主催者で既に利用している場合、主催しているイベントの名称

なお、Payvent の利用者は学会関係者、国際会議関係者、大学関係者、または学会運営代行会社が多いです。

次に、Payvent の web サイトに掲載している主催者向けヘルプセンター、参加者向けヘルプセンター、利用規約、プライバシーポリシーをそれぞれマークダウン形式のファイルにしてナレッジとして OpenAI API の Vector Stores にアップロードし、RAG 対象とした。以降これらのファイル群およびそれらを RAG のために保存した Vector データベースをナレッジと呼称する。プロンプトは以下のとおりである。

ナレッジは以下のファイルです。

- ・プライバシーポリシー. md
- ・利用規約. md
- ・主催者向けヘルプセンター. md
- ・参加者向けヘルプセンター. md

これらはナレッジとして提供されているマークダウン形式のファイルです。
 提供されたナレッジの中から日本語のキーワードを1つ任意に選択し、ナレッジからキーワードを検索して、その結果の文章を元にして想定した利用者が問合せすると思われる文章を作ります。
 問合せ文章は1つのトピックを含み、3-10 個の文の組み合わせから成り、回答するのに十分な情報量を持ちます。

ナレッジの中から日本語のキーワードを選択することで、生成される問い合わせ文が特定のトピックにフォーカスされる。また、ペルソナを先に作成し、問い合わせ者の立場を明確にすることで実際のユースケースに近い文章が生成できた。生成は基本的に生成する件数のみ指定して行ったが、ヘルプセンターの内容を鑑みて不足していると感じたトピックに関しては、キーワードを指定して生成を行った。

生成した架空問い合わせは 110 件で、具体的な生成例を以下に示す。うち生成結果が英語のものが 1 件あったため除外した。生成された問い合わせの例を以下に示す。

ペルソナを先に定義させ、ナレッジ内のキーワードを元

サステナブルエネルギーシンポジウムの参加者データを CSV ファイルでダウンロードしたいのですが、その手順を教えてください。また、特定のカラムだけを選択してダウンロードすることは可能ですか？

主催者向けヘルプセンターに CSV ファイルで参加者リストをダウンロードする機能の記述がある

グローバルビジネスサミットの参加費用収集のために Payvent を検討中です。クレジットカード以外に、銀行振込やコンビニ決済など、日本国内で一般的な方法での支払い受付が可能かどうか知りたいです。

主催者向けヘルプセンターにクレジットカード決済と銀行振込決済を受付可能である記述がある

今後の国際健康医療学会の参加費収集のため Payvent の導入を検討しています。参加費収集の際、特に留意すべき事項や記載する必要がある情報、また Payvent を利用して参加費を収集するメリットについて教えてください。参加者が適切に支払いを行えるよう、事前に準備すべき項目等もあれば知りたいです。

主催者向けヘルプセンターに Payvent を利用するメリットを記載している

初めて Payvent Hub を利用して国際教育カンファレンスに参加申し込みします。登録後のダッシュボードで参加ステータスや支払情報を確認する方法を教えてください。また、イベントの詳細やスケジュールの更新情報を確認する方法も知りたいです。

参加者向けヘルプセンターに Payvent Hub (参加者向けダッシュボードの名称) で申込情報や主催者からのお知らせを確認する方法を記載している

新技術研究所の橋本です。Payvent を使い革新技術国際シンポジウムを運営しています。しかし、利用規約違反に該当する可能性がある参加申込みを受けてしまいました。利用規約違反と判断される具体的な行為と、その場合の対処方法について教えてください。

利用規約に禁止行為に該当する申込の一覧が記載されている

に架空の問い合わせを生成することで、質問者の立場が明確で、提供サービスに対して適切な質問ができ、実際の問い合わせに近い文章を作成が可能であるという知見が得られた。

5. ナレッジを使用した返答の生成

ナレッジを使用して問い合わせに返答するチャットボットを作成した。OpenAI API から gpt-3.5-turbo-0125 を使用した。利用コストの観点から問い合わせ返答に gpt-4 を使用することは困難であった。

プロンプトは以下のとおりである。

あなたは学会向け決済・参加者管理サービス Payvent の運営スタッフです。あなたの役割は Payvent に来た問い合わせに対して、適切な回答を提示することです。回答は日本語で行ってください。

Payvent のドキュメントは”主催者向けヘルプセンター”, ”参加者向けヘルプセンター”, ”利用規約”, ”プライバシーポリシー” にまとめられています。必ず、これらのドキュメントから問い合わせに関連する情報を探し出し、それを元に回答を提示してください。

ドキュメントに記載のない項目や、ドキュメント内に項目を見つけることが出来なかった場合、回答できない旨を示してください。手順や機能を想像しないでください。

回答はサポートスタッフが顧客に向けるような丁寧な文体を心がけてください。

回答の後、ドキュメント内の回答生成のために使用した文章をそのまま引用してください。

プロンプトの後半部はハルシネーションを防ぐためにナレッジに記載されている情報のみを元に回答するよう記述した。ただし、“文章をそのまま引用してください。”と記述したにも関わらずナレッジに記載された文章がそのまま引用されず、文章の元となったファイルへのリンクを示すのみにとどまった。

前項で生成した架空問い合わせ 109 件に対し回答を生成した。生成された回答例を以下に示す。

6. 生成された返答の評価

前項のチャットボットで生成された返答を人力で定量評価した。評価軸を以下に示す。

・情報の正確性

正しい情報を見つけ出せるかを 5 段階で評価した (表 2)。

ナレッジはすでに Payvent の web サイトで提供している情報であり、追加された機能や問い合わせの多い項目が重点的に記述されている。

・網羅性

問い合わせで質問された内容に的確に答えているかどうか

質問文：グローバルビジネスサミットの参加費用収集のために Payvent を検討中です。クレジットカード以外に、銀行振込やコンビニ決済など、日本国内で一般的な方法での支払い受付が可能かどうか知りたいです。

回答文：“Payvent では、日本国内で一般的な支払い方法として、クレジットカード決済と銀行振込を受け付けています。クレジットカードとしては、VISA, Mastercard, American Express, JCB, ダイナース, ディスカバーの 6 種類に対応しており、安全な SSL 通信を介して情報が送信されるため、海外からの参加者でも利用可能です。また、銀行振込を利用する場合は、指定期間内に銀行で支払うように求められます。

返金の際には、クレジットカード決済の場合は全額返金されますが、Payvent 決済手数料として 6% が差し引かれます。一方、銀行振込の場合は、6% の Payvent 決済手数料と銀行振込手数料が差し引かれた金額が返金されます。返金手続きは参加者詳細ページで行われ、手続きが完了すると「返金済み」と表示されます。”

質問文：私どもは、近く国際教育交流のための研究会を開催予定で、Payvent を利用して参加者の登録および管理を行いたいと考えています。ただし、特定の国からの参加者に対しては参加費を無料または割引にする予定です。Payvent にはこのような割引設定を適用する機能がありますか？また、どのような形で参加者に案内すれば良いのでしょうか？

回答文：Payvent には特定の国からの参加者に対して参加費を無料または割引にする機能は明示されていません。代わりに、料金設定を自由に行うことができ、たとえば「早割」「学割」「子ども料金」などの価格設定が可能です。ただし、特定国からの参加者に割引を適用したい場合は、別途オプション料金を追加して、その割引を適用する可能性があります。オプション料金はイベント編集画面で自由に追加でき、申込画面やメールにも明記されます。割引の具体的な適用方法や条件については、主催者が必要な情報を含めて詳細を提示することが重要です。

かを 3 段階で評価した (表 3)。

・自然さ

返答文が日本語として自然かどうかを 3 段階で評価した (表 4)。文意は問わない。

・明確さ

内容の理解のしやすさを示す (表 5)。専門用語を説明せずに使用した場合、回答に無関係または不要な情報が含まれている場合は質問者を混乱させるため評価を下げた。

評価は株式会社 Urbs に所属する 3 人で行った。評価結

表 2 情報の正確性の評価

Table 2 Evaluating of information accuracy.

5	情報がナレッジに存在し、適切な返答を行った
4	情報がナレッジに存在し、トピックにあった情報を選択した
3	情報がナレッジに存在するが、誤った情報を選択した
2	情報がナレッジに存在せず、情報が無いと回答をした
1	情報がナレッジに存在せず、誤った回答をした

表 3 網羅性の評価

Table 3 Evaluation of comprehensiveness.

3	問合せに含まれるすべての質問に回答した
2	問合せに含まれる一部の質問に回答した
1	問合せに含まれる質問を正しく認識できていない

表 4 自然さの評価

Table 4 Evaluation of naturalness.

3	自然な言葉や表現で書かれており、会話の流れに違和感がない
2	一部分が自然でない言葉や表現があり、会話の流れに若干の違和感がある
1	不自然な言葉や表現で書かれており、会話の流れが不自然である

表 5 明確さの評価

Table 5 Evaluation of clarity.

3	わかりやすい形式で提示されており、理解しやすい
2	一部分かりにくい言葉や表現があり、理解に若干の困難さがある
1	回答が非常に分かりにくく、理解がほとんどできない

表 6 ヘルプセンター、利用規約、プライバシーポリシーを RAG して生成した回答の評価

Table 6 Evaluation of answers generated by RAG from Help Center, Terms of Use and Privacy Policy.

	情報の正確性	網羅性	自然さ	明確さ
1	28	12	2	7
2	3	28	39	29
3	18	69	68	73
4	30			
5	30			

果を表 6 に示す。

情報の正確性については、ナレッジに記述の無い項目について誤った情報から回答を生成してしまう（評価 1）や、記述があるにも関わらず間違った情報から回答を生成してしまう（評価 3）ものが目立った。

網羅性については質問文に含まれるすべての質問に回答していないものや、一部の質問の要旨を間違ったものを評価 2 とした。

自然さについては使用した LLM のモデルに依拠する。プロンプトで生成 AI の立場“運営スタッフ”を明確にし、問い合わせの返答文を生成するように指示したにもかかわらず、回答文が第三者からの視点で書かれているものや、文章のつながりが不自然なものは、評価 2 とした。

明確さについては概ね良好な結果を示したが、評価を 2 としたものは余分な情報を記載していることが多かった。特に、銀行振込についての質問の場合に質問に対する返答に加えて返金について答えてしまう傾向がみられた。これは、銀行振込での参加費受付に関して過去に振込先の間違い、振込金額の間違いといったトラブルが多く、ヘルプセンターの銀行振込に関する記述が返金について案内しているものが多いことによる。

以下に内容について評価が低かった回答の例をいくつか抜粋して示す。

この回答は情報の正確性を 1 とした。記載されている写

<p>質問：参加者から提出される発表資料をアップロードしてもらおう機能について知りたいです。Payvent ではどのようにファイルのアップロードを管理できますか？ファイル形式やサイズの制限はありますか？</p>
<p>回答：参加者が提出する発表資料をアップロードする際に、特定のファイル形式やサイズの制限はドキュメントから明確には示されていませんでした。ただし、主催者向けヘルプセンターには写真のアップロードに関する情報が記載されており、以下の要件が提示されています：</p> <ul style="list-style-type: none"> - 写真は鮮明でカラー写真であること - サイズが 5MB 以下 - .png または .jpeg のフォーマット - 全ての情報が閲覧可能な状態であること

真の条件は登録時の本人確認書類に関するものであり、“ファイルのアップロード”という状況でのみナレッジを検索しており、誤った回答である。

この回答は情報の正確性を 1 とした。支払い方法を後から変更することは不可能であり、支払方法の変更についてはナレッジに一切記載はない。

この回答は網羅性を 1 とした。質問文は参加者の立場であるのに対し回答で主催者の管理画面について記述されて

質問：既に申し込みを完了した後でも、支払い方法をクレジットカードから銀行振込みに変更することは可能でしょうか？また、その手続き方について詳細を知りたいです。

回答：Payvent では、支払い方法を変更することが可能です。既に申し込みを完了した後でも、クレジットカードから銀行振込みに変更できます。クレジットカードからの変更方法については、支払い手続きを進める際、指定の「銀行振込する」ボタンを押すことで可能です。この場合、返金の手続きが行われますが、支払ったクレジットカードに返金される額から Payvent 決済手数料や銀行振込手数料が差し引かれる点にご留意ください。支払い方法変更や返金において、は迅速かつ透明な手続きを提供しています。

質問：こんにちは、Payvent を利用してカンファレンスに申し込みたいと思っています。銀行振込の場合、入金確認までどれくらいの時間がかかるのか、そして入金後の手続きがあるのか教えてください。さらに、うっかり振込金額を入力ミスしてしまった場合の対処法についても知りたいです。

回答：銀行振込による入金確認には通常数日かかります。入金後、入金申請・履歴画面で入金の状況を確認できます。主催者の管理画面からは、入金履歴を確認することが可能です。入金手続きで金額を入力ミスした場合、60日経過後に余った金額について返金手続きが開始されます。振込先の銀行口座番号を間違えてしまった場合は、早急にご利用の銀行に組戻しの手続きを依頼する必要があります。詳細は銀行にお問い合わせください

いる。また、記載の方法は主催者の売り上げを銀行口座に入金する際の方法であり、銀行振込決済とは異なる。

これらの評価から、RAG 対象とするドキュメントが十分に網羅されていても情報の偏りにより検索結果に影響を及ぼしてしまう知見が得られた。ナレッジベースをより構造化し、トピックごとの記述量の偏りをなくすことで改善することができる。また、プロンプトについては、回答例を提示する Few-Shot プロンプティングが改善案として考えられる。

また、類似した用語が別の文脈で使われている場合に混同してしまう例がみられた。具体的には、イベント主催者が売上を主催者の銀行口座に振り込む「入金」と、イベント参加者が銀行振込決済をする場合に参加費を振込「入金」を混同した。上述の偏りと合わせて、売り上げの入金について質問した場合に混同したうえで参加費の返金につ

いての記述が追加されてしまうといった例があった。

7. 問い合わせ履歴を使用した返答の生成

ナレッジを使用して問い合わせに返答するチャットボットは多くの場合正しい情報を選択し、回答文として自然な日本語で返答を生成することができた。ナレッジに加え、前項で作成した個人情報除去した問い合わせ履歴を RAG 対象にすることで、既存の問い合わせ事案も参考にしより精度の高い回答を生成させることを目指した。使用モデルは前項と同じく OpenAI API から gpt-3.5-turbo-0125 を使用した。Claude は XML を扱うことに長けていたが、OpenAI API で利用できるファイル形式に XML が含まれていなかったため、json 形式に変換し inquiries.json としてアップロードした。

問い合わせ履歴も含めた返答の生成は、ナレッジと問い合わせ履歴の情報を分けて検索したうえでソースを元に条件分岐した手順が必要になる。以下の 3 パターンのプロンプトを用意して生成を試みた。

- ・ 問い合わせ履歴とナレッジを合わせて検索して回答を生成する
- ・ 問い合わせ履歴をまず探して見つからなかった場合ナレッジを検索して生成する
- ・ ナレッジをまず探して見つからなかった場合問い合わせ履歴を検索して生成する

各プロンプト共、役割の指定や文体の指示は前項の「ナレッジを使用した返答生成」と共通であるが、ナレッジおよび問い合わせ履歴を検索する指示が異なる。異なる部分を抜き出して以下に提示する。

回答の生成に利用した質問文は問い合わせ履歴のものとして架空問い合わせから以下の条件で選択した。

- ・ 問い合わせ履歴とナレッジ双方に記述がある質問 4 件
- ・ 問い合わせ履歴に記述が存在するがナレッジに記述がない質問 5 件
- ・ ナレッジに記述が存在するが問い合わせ履歴に記述が

問い合わせ履歴とナレッジを合わせて検索して回答を生成するパターン

Payvent のドキュメントは「主催者向けヘルプセンター」、「参加者向けヘルプセンター」、「利用規約」、「プライバシーポリシー」にまとめられています。問い合わせ履歴は inquiries.json にまとめられています。

「inquiries.json」の tickets は事例ごとの配列、ticket 中の receive は受信した問合せ文、send は問合せに対する返信文です。

これらのドキュメントおよび問合せ履歴から問い合わせに関連する情報を探し出し、それを元に回答を提示してください。

問合せ履歴をまず探して見つからなかった場合ナレッジを検索して生成する

問合せ履歴から問い合わせに関連する情報を探し出し、それを元に回答を提示してください。

問い合わせ履歴は inquiries.json にまとめられています。“inquiries.json”の tickets は事例ごとの配列、ticket の中の receive は受信した問合せ文、send は問合せに対する返信文です。

問合せ履歴に類似事例が見つけれなかった場合ドキュメントから問い合わせに関連する情報を探し出し、それを元に回答を提示してください。

Payvent のドキュメントは”主催者向けヘルプセンター”、“参加者向けヘルプセンター”、“利用規約”、“プライバシーポリシー”にまとめられています。

ナレッジをまず探して見つからなかった場合問合せ履歴を検索して生成する

Payvent のドキュメントは”主催者向けヘルプセンター”、“参加者向けヘルプセンター”、“利用規約”、“プライバシーポリシー”にまとめられています。

まず、これらのドキュメントから問い合わせに関連する情報を探し出し、それを元に回答を提示してください。

ドキュメントから関連する情報を探し出せなかった場合、問合せ履歴から問い合わせに関連する情報を探し出し、それを元に回答を提示してください。

問い合わせ履歴は inquiries.json にまとめられています。“inquiries.json”の tickets は事例ごとの配列、ticket の中の receive は受信した問合せ文、send は問合せに対する返信文です。

ない質問 3 件

- ・ 問い合わせ履歴とナレッジどちらにも記述がない質問 5 件

問い合わせ履歴に記述がある質問は実際に返答を行ったため実際の問い合わせ文を使用し、問い合わせ履歴に記述がない質問は前項で生成した架空の問い合わせ文を使用した。

なお、inquiries.json には Re:lation のファイル出力の制約上回答が全文記録されていないものが存在した。問い合わせ履歴に回答が全文含まれていない質問を行った場合、質問文を見つけることができても回答例が存在しないため、“過去に事例があった”と紹介するような回答を生成することがあった。そのため、返答を生成させる質問は回

答が全文記録されているものを選択した。

いずれのプロンプトでも、問い合わせ履歴とナレッジを使用して回答を生成することができ、ナレッジに加え問い合わせ履歴を RAG 対象にすることで回答を改善できるという知見が得られた。

問い合わせ履歴に存在するがナレッジに記述がない質問では、問い合わせ履歴から返答事例を見つけて正しい回答を提示し、返答が改善された。

ナレッジに存在するが問い合わせ履歴に記述がない質問では、ナレッジのみで生成したときと同等の回答を提示した。

しかし、問い合わせ履歴とナレッジどちらにも記述がない質問に関しては、いずれのプロンプトでも間違った情報を提示してしまう例がみられた。

問い合わせ履歴、ナレッジ双方に記載のある質問文では、検索する順番の指定に従わない返答がみられた。問い合わせ履歴を先に探すよう指示したプロンプトであっても、双方の文章を引用して回答を生成した。

以上のプラクティスから、RAG 対象のドキュメントを網羅的に検索することはできるが、検索する順番の指示は難しいと考えられる。

また、問い合わせ履歴に類似事例を発見できたかどうかにかかわらずナレッジを検索して、両方の情報を使用して回答を生成するようなプロンプトを記述した場合に、“ナレッジの内容を精査せず、問い合わせ履歴のみで回答してしまう”という現象もみられた。

さらに、問い合わせ履歴から生成した回答を A、ナレッジから生成した回答を B として回答 A、回答 B 両方を記載せよとプロンプトに記述した場合、手順が複雑になりすぎたためか、回答 A、回答 B を記載しわけることができなかった。

モデルを gpt-4-turbo-preview に切り替えた場合は回答 A、回答 B を記載しわけることはできたが、問い合わせ履歴からの生成とナレッジからの生成を混同してしまい、回答 B に問い合わせ履歴の情報を使用して答えてしまう現象がみられた。

8. 課題と活用

本稿では、個人情報が含まれる問い合わせ履歴とヘルプセンターを活用して適切な返答を生成するチャットボットの開発を試みた。

個人情報の除去の試みでは、プロンプトに記載した個人情報の定義が不十分であった。特に、住所の定義を「”県”や”市”といった文字が含まれる」としたが、“市場”、“県営”といった住所ではない文字列を誤検出する可能性がある。また、“都道府区町村”といった文字が含まれていれば住所である可能性が高いが今回は未対応であった。人名と地名が同一である場合もある。これらを正しく

検出するための対策として、プロンプトに前後の文章を考慮するように指示することが考えられる。たとえば、単語の前後の文章から判断して当該の単語が主語である場合人名の可能性が高く、郵便番号や番地といった数字が前後に存在する場合は住所である可能性が高い。文内に“都道府県市区町村”のいずれかの文字が連続して出現する場合住所である可能性が高い。これらの指示で、単語そのものでは個人情報であるかどうかの判断が難しい場合でも検出精度を高められると考えられる。

個人情報保護については、生成 AI では完全な匿名化を保証することはできず、不十分であった際の責任問題もあるため、どれだけ精度を高めたとしても人の手による確認、修正は必須である。また、どこまでの個人情報を保護すべきかの要件は時代に合わせて変化していくため、長期的に継続してプロンプトに記述する個人情報の定義を更新していく必要がある。

本稿の試行では生成 AI によって 80% 近くの個人情報を除去できた。現時点でもすべてを人の手で除去するよりも大きく労力を下げることができている。前述したプロンプトの改善で検出精度を高めることはできると考えるが、生成 AI の精度目標は高いほど大きなコストがかかる。本稿の見解では生成 AI による個人情報の除去はあくまで人による除去の補助と考え、生成 AI では電話番号やメールアドレス等の定義が明確なものをより確実に検出できるようにし、個人情報の定義の継続的な更新および、生成 AI による除去後のチェック、漏れの修正は人の手で行っていくのが望ましい。

生成された回答の中には、問い合わせ履歴にもナレッジにも基づかない誤った情報を含むものがあつた。記述の無い項目を無いと答えるのは、記述を見つけるよりも難しいと考えられる。回答の網羅性が不足しており、問い合わせのすべての質問に対して回答できていない例が多くみられた。

RAG 対象のドキュメントの記述が十分に網羅している項目であっても偏っていると検索結果もそれに伴い偏ってしまう現象がみられた。例として、ヘルプセンターに記載されている銀行振込に関する項目は返金に関する記述がほとんどであり、銀行振込について聞くだけで返金についても追加で答えてしまうことがあつた。

また、ヘルプセンターの記述で、違う文脈で同じ文言が使われている場合に RAG が混同してしまう現象がみられた。例として、ヘルプセンターでは“入金”という言葉が売り上げの入金と銀行振込の参加費の入金の 2 パターンの記述が存在した。同様に、参加者の立場での質問文に対し主催者向けヘルプセンターの記述から回答を生成してしまう現象や、質問文に含まれる単語の記述はあるが異なる内容の項目から回答を生成してしまう現象がみられた。人間のサポートスタッフは文脈から適切な返答が可能だが生成

AI が判別することは難しいと考えられる。

これらの課題は、RAG でのドキュメントの検索に原因があると考えられる。永江らは、章で区切られたドキュメントを検索する際に、単一の章を検索単位とするよりも、近隣の 3 つの章を検索単位とするほうが回答生成の性能があつたと報告した [12]。このようなドキュメントの検索範囲の設定や、ナレッジの偏りをなくし、より構造化するといったドキュメントの前処理段階での改善で生成する回答の質を向上できると考えられる。

質問者の立場によってナレッジ内のヘルプセンターのファイルに分けていたにもかかわらず、プロンプトの指示ではまとめて“ナレッジ”として扱ってしまったことも原因として考えられる。質問者の立場によって検索対象のファイルを指定するように指示することでも回答の質を向上できる可能性がある。さらに、異なる項目で意味の似た単語（銀行振込決済の代金の“入金”と売上の主催者口座への“入金”など）を使用しないようにヘルプセンターの記述を変更することでナレッジの検索結果に異なる項目が混同されることは回避できると考えられる。

今回の試行ではナレッジを使用した生成 AI を人間の手で定量評価したが、この方法は問い合わせ数が増えるほどに人的、時間的なコストが高くなるため、RAG の評価の自動化が求められる。評価の自動化により、同じ質問に対して複数回回答させた結果を評価して回答の揺れについても評価することが可能であると考えられる。RAG の評価を自動化する手法としては RAGAS [13] が存在する。

試行で判明した業務の改善案としては、ヘルプセンターに追記が必要な情報の洗い出しや記述ミスの発見ができたこと、架空の問い合わせを実際のサポートスタッフが答えるトレーニングに使用できることがあげられる。

Payvent のヘルプセンター内の項目には、文章での説明が不足しておりスクリーンショットを読み取る必要がある項目があり、画像を RAG 対象に含められないことにより当該項目に関する質問では回答の生成に支障をきたした。副次的に、音声でブラウジングしている閲覧者などの画像を読めない環境で問題があることが判明した。

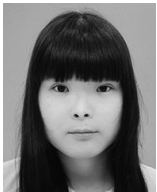
顧客による個人情報の入力リスクや誤回答リスクがあるため、現段階では社内でのみの活用にとどまっているが、今後チャットボットを顧客に向けて公開しサポート業務の自動化を目指す。

謝辞 株式会社 Urbs 関係者ならびに Payvent を利用していただいた皆様に、謹んで感謝の意を表す。

参考文献

- [1] Meta: Meta Llama (オンライン) 入手先 <<https://llama.meta.com/>> (参照 2024-05-11).
- [2] ANTHROPIC: Introducing the next generation of Claude (オンライン), 入手先 <<https://www.anthropic.com/news/claude-3-family>> (参照 2024-05-11).

- [3] OpenAI: GPT-4 | OpenAI (オンライン), 入手先 <https://openai.com/index/gpt-4/> (参照 2024-05-11).
- [4] 総務省 経済産業省: AI 事業者ガイドライン (第 1.0 版), p.5 (2024).
- [5] 総務省 経済産業省: AI 事業者ガイドライン (第 1.0 版), p.32 (2024).
- [6] Amazon Web Services: RAG とは何ですか? (オンライン), 入手先 <https://aws.amazon.com/jp/what-is/retrieval-augmented-generation/> (参照 2024-05-11).
- [7] Amazon Web Services: PII エンティティの検出 (オンライン), 入手先 https://docs.aws.amazon.com/ja_jp/comprehend/latest/dg/how-pii.html (参照 2024-05-11).
- [8] 株式会社 Urbs: Payvent (オンライン), 入手先 <https://payvent.net/> (参照 2024-05-11).
- [9] Amazon Web Services: Amazon Bedrock (オンライン), 入手先 <https://aws.amazon.com/jp/bedrock/> (参照 2024-05-11).
- [10] ANTHROPIC: XML タグを使用する (オンライン), 入手先 <https://docs.anthropic.com/claude/docs/use-xml-tags> (参照 2024-05-11).
- [11] 株式会社インゲージ: Relation (オンライン), 入手先 <https://ingage.jp/relation/> (参照 2024-05-11).
- [12] 永江尚義, 吉田尚水, 小林優佳, 久島務嗣, 岩田憲治: マニュアルの章構造にロバストなマニュアル検索技術, 人工知能学会研究会資料 言語・音声理解と対話処理研究会, 人工知能学会, p.190(2023)
- [13] Shahul Es, Jithin James, Luis Espinosa-Anke, Steven Schockaert.: RAGAS: Automated Evaluation of Retrieval Augmented Generation, arXiv: 2309.15217, pp.1-2



森山 有理名 (学生会員)

大阪電気通信大学大学院, 株式会社 Urbs.



魚井 宏高 (正会員)

大阪電気通信大学総合情報学部.