更新履歴
2024/05/13 初版公開
2024/05/17 p1の誤字修正
2024/05/17 p1に更新履歴を追加

デジタル庁R5年度　検証結果共有

行政における生成AIの
適切な利活用に向けた技術検証の環境整備

※本スライドはR5年度の検証について、委託事業者の株式会社FIXERの最終報告書に基づいて作成されています
そのため文体が、委託事業者からデジタル庁向けになっている箇所がございます。ご了承ください
また、文中の主張や解釈は政府を代表した意見ではなく事業者による解釈に基づいております。

エグゼクティブ・サマリー (1/2: 本プロジェクトの実績)

はじめに

急速なAIの進歩・普及と、生成AIの性能向上を受けて、行政において生成AIを活用する環境を構築するために、デジタル庁で安全な基盤上で生成AIを扱うことができる実証環境を整備し、2023年12月から2024年3月にかけて他府省庁・自治体のみなさまにもご参画いただきながら、技術検証プロジェクトを実施した

本報告書においては、技術検証の実施内容をご報告し、技術検証の論点についての考察を提示する

実績

実施内容	支援内容(当初案)	実績
検証環境の提供	• 要件定義書の要件を満たす検証環境を構築、検証に参加する府省庁・自治体職員に提供	• 技術検証基盤を767人 (13府省庁・26自治体)、共創PF Slack上で158人に提供 • ご要望を踏まえ、30回のアップデートを実施 – 最新のLLMであるClaude 3 Sonnetや国産LLM Swallowもご提供
検証環境の運用・保守	• 環境を安定稼働させ、運用レポートおよびダッシュボードを提供 • 利用状況・問い合わせ状況・作業実績等の運用レポートを月次で提出	• 運用レポート・ダッシュボードを提供 • 貴庁セキュリティ班とISMAPチェックリストの確認を行い、必要となるシステム改修を実施
技術検証支援	• 操作マニュアル・導入研修を提供、サポート体制を構築 • 課題管理を行い、プロンプト/学習データの改良等で課題解決を支援	• ヘルプデスクを56日間稼働、計64件の質問に対応 • 導入研修を計12回実施し、計128人が受講 • 9個のユースケースでの有用性を検証

エグゼクティブ・サマリー (2/2: 技術検証課題のご報告要旨)

技術検証課題	ご報告要旨
1 LLMを用いた国・自治体職員の業務効率化や成果物の品質向上等に係る有用性の検証	**・技術検証基盤・共創PF Slack上で生成AIを提供し、職員の業務効率化や成果物の品質向上に一定の効果** – ユーザと対話しつつ**LLMとその設定、プロンプト、RAGの対象となる学習データを整備**し、ご利用いただいた。 – 利用者調査では、**90%以上の利用者が業務効率化および成果物の品質に効果あり**と回答した。
2 本件有用性の検証に係る試験的なデータの整備、データモデル等の提案	**・ユースケースの検証を通じて、学習データの整備についての知見を蓄積** – ガイドブック・規則等を用いたユースケースで、**学習データとして適切なデータ形式の知見**が得られた。 – **用例による法令検索（検索単語の一致に加え、 LLMで用例一致法令を抽出）を構築**し、さまざまな文書(例: 過去の国会答弁 等)で有用な可能性を見出した。本件では、このアルゴリズムも設計書として整理した。
3 LLMを業務利用のために調達する場合の要件定義・品質管理等のあり方	**・必要な回答速度・精度を踏まえ、丸投げではなく調達するLLMを検討する必要** – ユースケースで**GPT-4 TurboとClaude 3の回答品質を比較したところ、課題によって異なる結果**となった。 – **調達仕様の策定に際しては、複数LLMを比較・検討が実施できるSaaS等で試行を行うことも検討**すべき。
4 LLMの性能評価、競争性、相互運用性、リスク管理、選定基準などの整理	**・複数LLM・パラメータを比較し、性能評価できる環境を提供。リスクある利用にアラートを上げる仕組みも構築** – 本検証では**PlayGround機能を通じて、複数のLLM・パラメータの性能を評価できる環境を構築**した。 – リスク管理について、**情報公開法の記述を参考に機密性2情報を本基盤で取り扱う上での目安を策定**し、**機密情報の入力にアラートを上げる仕組みを試験的に実装**した。
5 その他、政府・自治体がLLMを活用するにあたり留意すべき課題の整理	**・政府・自治体が組織内でLLMを活用していく上で、業務に応じたプロンプト等の整備拡充が必要** – LLMが活用環境な可能であっても、**組織内でのユーザの拡大および適用対象業務の拡大に向けては、研修等の実施に加え、業務に応じたLLM・プロンプト・学習データをセットで提供する必要**がある。

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

プロジェクトの背景と目的

	調達仕様書・要件定義書より	＋	自治体・企業の現状を踏まえた補足（当社理解）
背景	**急速なAIの進歩・普及 ＋ 生成AIの性能向上** • デジタル社会の実現に向けたAIの社会的な影響力増大 • AIの適切・効果的な活用で生産性向上・競争力強化 • 行政運営の効率化・行政サービス品質向上の必要性		**行政における生成AIの活用はまだ限定的で、非効率な状況** • 過半数の自治体首長が活用に賛成[1]しているが、試行・導入している自治体は限定的 • 但し、自治体ごとに業務効率化に向けたプロンプトを開発しているなど、車輪の再発明が起こっている可能性
目的	**行政において生成AIを活用する環境を構築するためにデジタル庁で安全な基盤上で生成AIを扱うことができる実証環境を整備** • 安全な環境で行政が生成AIを利活用するための検証 – 機密情報を取り扱うことができる環境整備（機密性2情報[2]） – 業務における適用範囲拡大のためのポイント • 生成AIにおける将来的なベンダーロックインの回避 – 性能や技術的特性の理解、相互運用性の評価 – 競争の促進、将来の調達における公平性確保		**最終的には今回の成果物をデジタルマーケットプレイスを通じ、他府省や自治体にご提供し、ご利用いただくことも視野に入れてプロジェクトを運営** • 機密情報を取り扱える安全な環境を自治体に広くご提供することで、活用を加速化 • 行政・IT双方に知見あるデジタル庁の皆様と実験を通じて磨き込み、業務効率化に寄与 • ベンダーロックインを防ぐことにより、適正な価格で、最適な生成AIの導入を支援

1. 毎日新聞「生成AI、知事や市長の67％が活用に「賛成」 毎日新聞アンケート」（2023年6月17日）
2. 総務省「政府機関の情報セキュリティ対策のための 統一管理基準」における定義は「行政事務で取り扱う情報のうち、秘密文書に相当する機 密性は要しないが、漏えいにより、国民の権利が侵害され又は行政事務の遂行に支障を及ぼすおそれがある情報」

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

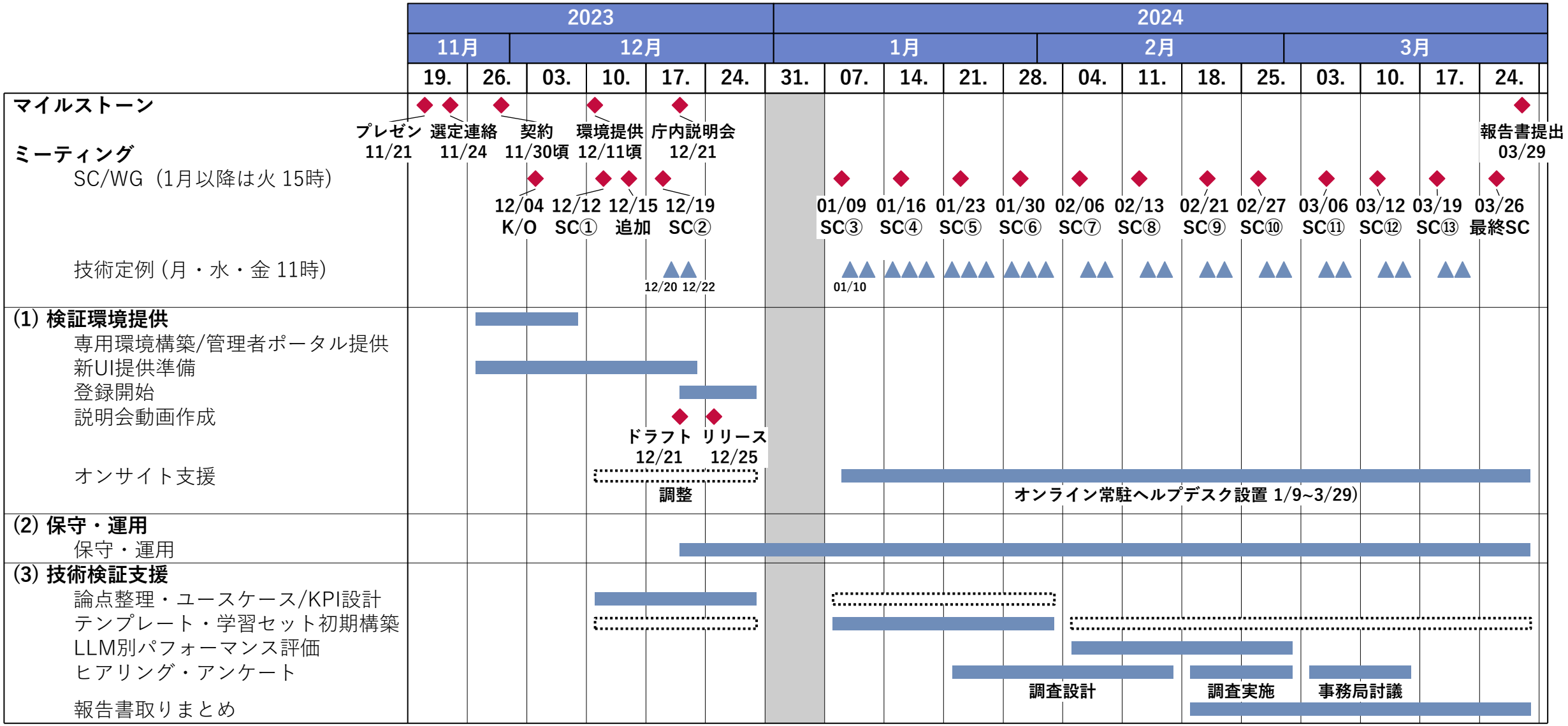
Appendix

Appendix A. 検証基盤の利用状況

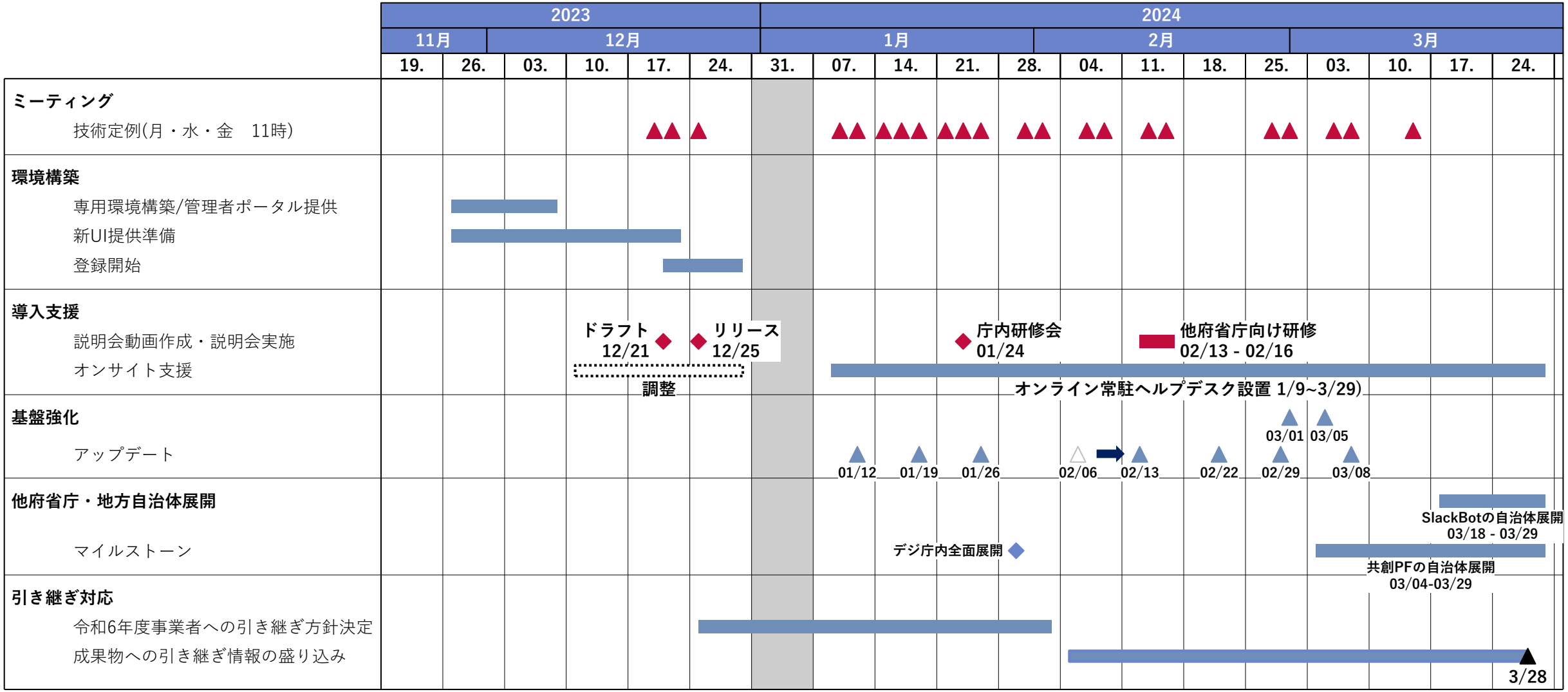
Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

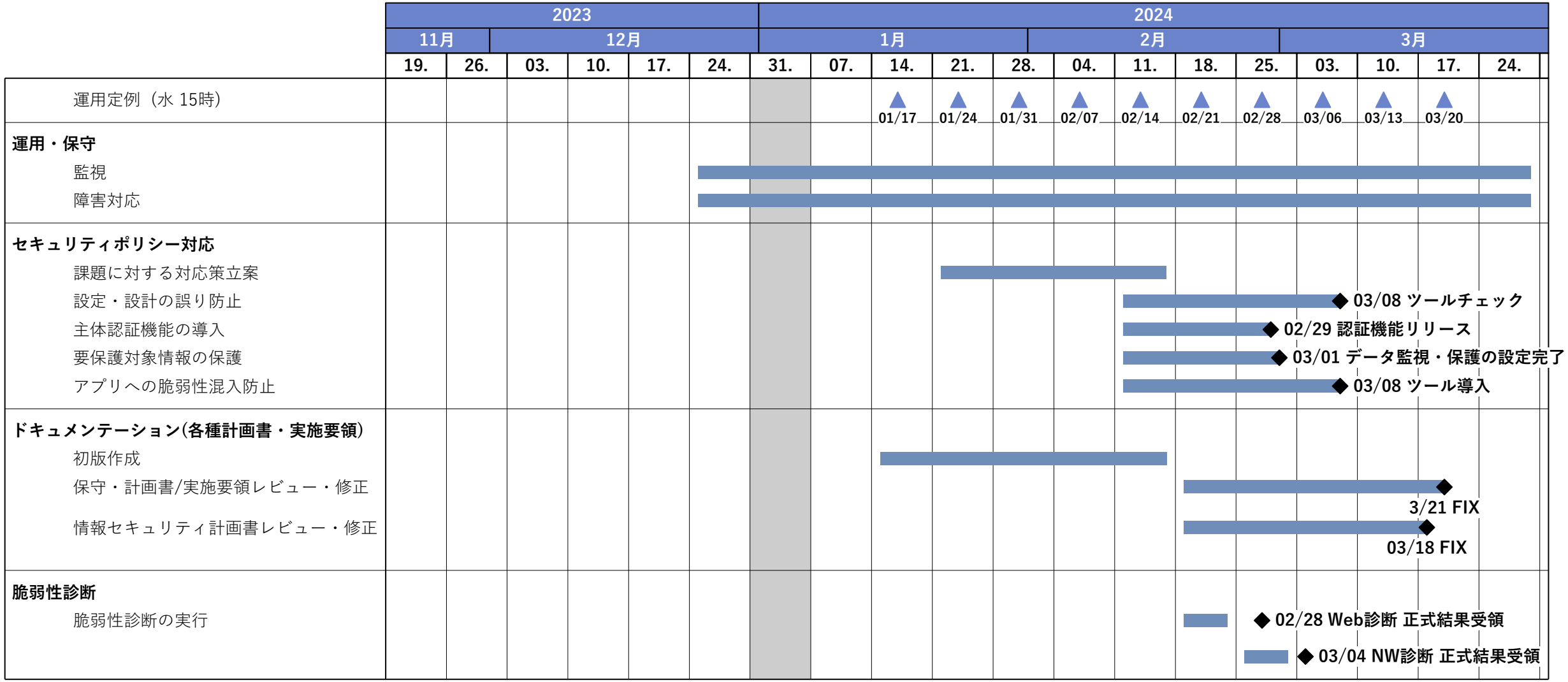
マスタースケジュール


2023
2024
11月
12月
1月
2月
3月
19. 26. 03. 10. 17. 24. 31. 07. 14. 21. 28. 04. 11. 18. 25. 03. 10. 17. 24.
マイルストーン
プレゼン 11/21
選定連絡 11/24
契約 11/30頃
環境提供 12/11頃
庁内説明会 12/21
報告書提出 03/29
ミーティング
SC/WG（1月以降は火 15時）
12/04 K/O
12/12 SC①
12/15 追加
12/19 SC②
01/09 SC③
01/16 SC④
01/23 SC⑤
01/30 SC⑥
02/06 SC⑦
02/13 SC⑧
02/21 SC⑨
02/27 SC⑩
03/06 SC⑪
03/12 SC⑫
03/19 SC⑬
03/26 最終SC
技術定例（月・水・金 11時）
12/20 12/22
01/10
(1) 検証環境提供
専用環境構築/管理者ポータル提供
新UI提供準備
登録開始
説明会動画作成
ドラフト 12/21
リリース 12/25
オンサイト支援
調整
オンライン常駐ヘルプデスク設置 1/9~3/29)
(2) 保守・運用
保守・運用
(3) 技術検証支援
論点整理・ユースケース/KPI設計
テンプレート・学習セット初期構築
LLM別パフォーマンス評価
ヒアリング・アンケート
調査設計
調査実施
事務局討議
報告書取りまとめ

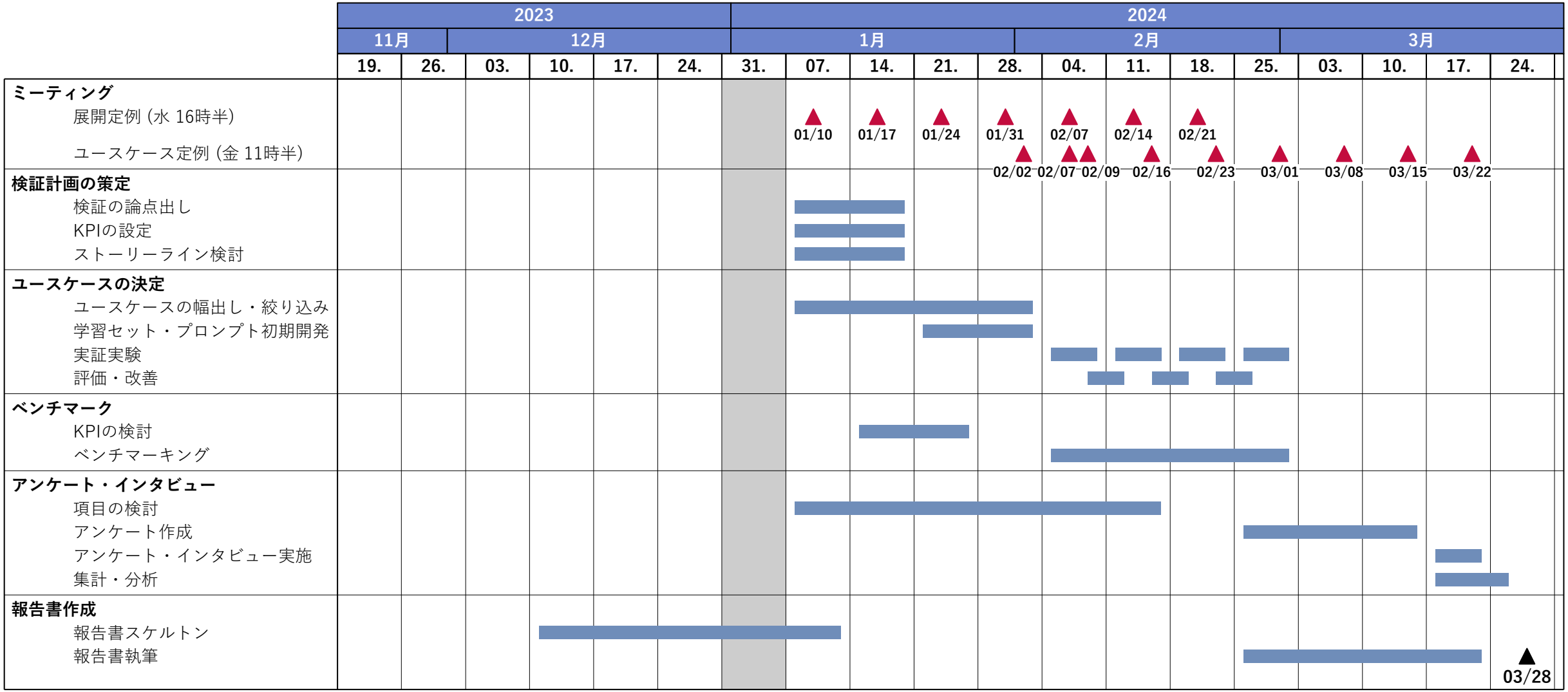
詳細スケジュール: (1) 検証環境整備



詳細スケジュール: (2) LLM運用・保守

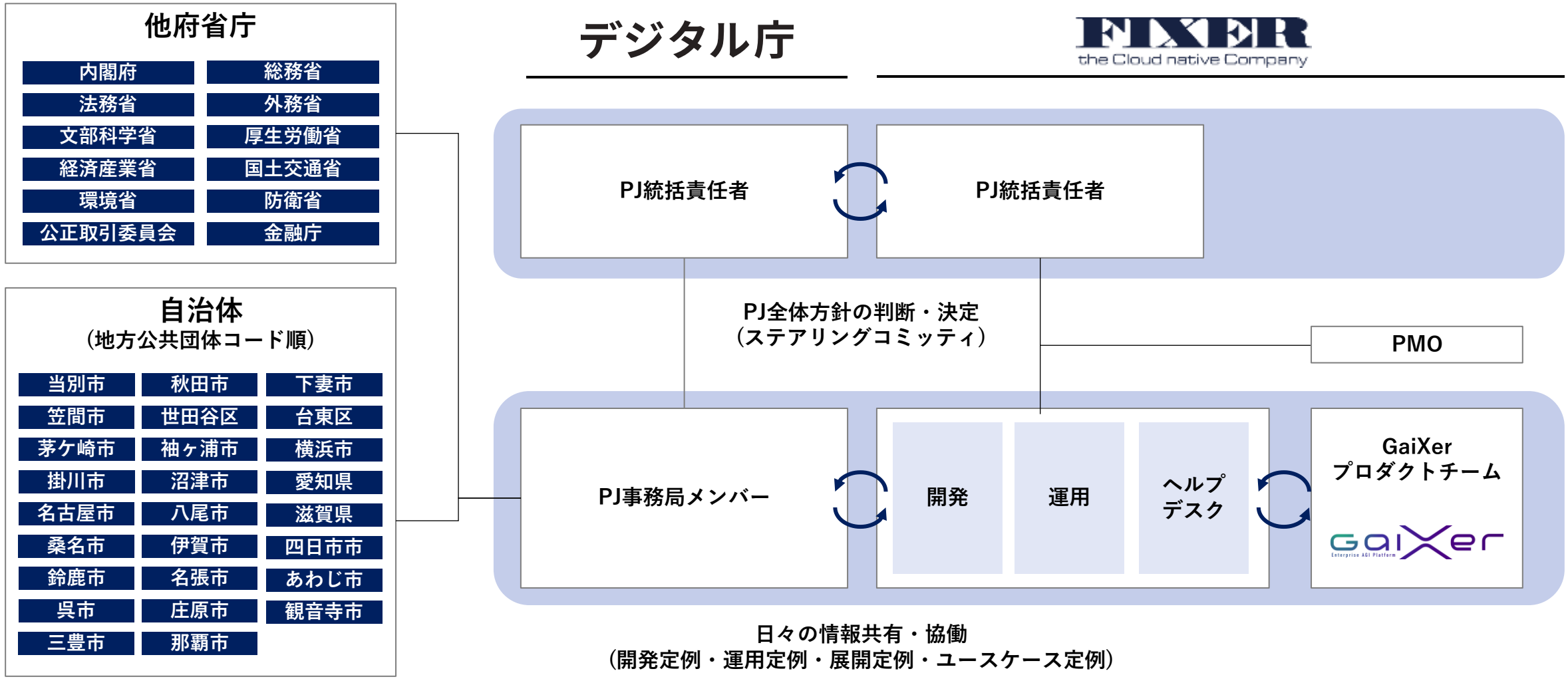


詳細スケジュール: (3) 技術検証支援



プロジェクト体制

▶ デジタル庁と委託先の株式会社FIXERが中心となってプロジェクトを推進した。検証には13府省庁・26自治体が参加



目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

業務内容（企画提案書より再掲）

▶ 調達仕様書「5. 調達の範囲」に記載されている内容は、下記の通りご提供いたします

調達仕様書の内容	提供内容
(1) 検証に参加する国・地方公共団体の職員に対する検証環境の提供	**要件定義書の各要件を満たす検証環境を構築し、検証に参加する関係府省庁、自治体の職員に提供いたします** • 要件定義の充足はP30-33で示します • 利用者はデジタル庁職員の皆様の約2倍、2,000人を想定しております
(2) LLM 検証環境の運用・保守	**環境を安定稼働させ、運用レポートおよびダッシュボードをご提供します** • 利用状況・問い合わせ状況・作業実績等の運用レポートを月次で提出します • ダッシュボードの内容は、調達仕様書5(2)の内容を元に協議の上で決定します
(3) 技術検証支援	**生成AIコンサルタント/プロンプトエンジニアが、調達仕様書5(3)記載の通り技術検証をご支援します** • 技術検証の計画案を作成し、プロジェクト終了までに報告書を作成します • 操作マニュアル・導入研修を提供、サポート体制を構築します • 課題管理を行い、プロンプト/学習データの改良等で課題解決を支援します

業務の実施状況

▶ 調達仕様書に記載されている各業務について、下記の通り実施

実施内容	支援内容(当初案)	実績
(1) 検証環境の提供	• 要件定義書の要件を満たす検証環境を構築、検証に参加する府省庁・自治体職員に提供	① 技術検証基盤を767人 (13府省庁・26自治体)、共創PF Slack上で158人に提供 ② ご要望を踏まえ、30回のアップデートを実施 – 最新のLLM Claude 3 Sonnetや国産LLM Swallowを含む、10のLLMをご提供
(2) 検証環境の運用・保守	• 環境を安定稼働させ、運用レポートおよびダッシュボードを提供 • 利用状況・問い合わせ状況・作業実績等の運用レポートを月次で提出	③ 運用レポート・ダッシュボードを提供 ④ 貴庁セキュリティ班とISMAPチェックリスト、セキュリティポリシーとの整合性を確認し、脆弱性診断および必要となるシステム改修を実施
(3) 技術検証支援	• 操作マニュアル・導入研修を提供、サポート体制を構築 • 課題管理を行い、プロンプト/学習データの改良等で課題解決を支援	⑤ ヘルプデスクを78日間稼働、64件の質問に対応 ⑥ 導入研修を12回実施し、計128人が受講 ⑦ 6分野・9個のユースケースにおける有用性を検証 ⑧ 検証過程において、68個のプロンプトを作成

注: 各数値は2024年3月28日10時時点

業務の実施状況(1)検証環境の提供

① 技術検証基盤・共創PF上でのLLM提供

▶ 当社生成AIプラットフォームGaiXerをベースとした技術検証基盤を767ユーザ、共創PF Slack botを158ユーザに提供

技術検証基盤の提供 (Webブラウザでアクセス)

当社の生成AIプラットフォームGaiXerをベースに
Webブラウザからアクセスできる技術検証基盤を提供し、
767名(13府省庁・26自治体)がユーザ登録し、利用した。

デジタル改革共創PF Slack上での提供

デジタル改革共創プラットフォームSlackの参加者に呼びかけ
158名(3府省庁・100自治体)が専用チャンネルに参加。
チャンネル上でメンションすると回答するSlack botを提供した。

業務の実施状況(1)検証環境の提供

組織別参加人数一覧

府省庁

組織名	検証基盤	共創PF
内閣官房	25	
内閣府	16	
デジタル庁	451	○
総務省	19	
法務省	7	
外務省	15	
文部科学省	4	
厚生労働省	18	
経済産業省	19	
国土交通省	26	○
環境省	8	
防衛省	11	
公正取引委員会	4	
金融庁	4	
農林水産省	ー	○
合計	627	

自治体

組織名	検証基盤
当別町	2
秋田市	11
下妻市	6
笠間市	10
世田谷区	14
台東区	3
茅ケ崎市	1
袖ケ浦市	2
横浜市	10
掛川市	4
沼津市	19
愛知県	1
名古屋市	3
八尾市	1
滋賀県	2
桑名市	5
伊賀市	7
四日市市	1
鈴鹿市	2
名張市	5
南あわじ市	15
呉市	1
庄原市	1
観音寺市	2
三豊市	10
那覇市	1
合計	140

デジタル改革共創PF 生成AIチャンネル参加自治体				
下呂市	美唄市	別府市	石狩市	奈良市
福井市	小美玉市	西和賀町	上越市	村山市
大田原市	伊豆の国市	鳥取	武富氏	志木市
東久留米市	野洲市	南関町	岩泉氏	八戸市
千葉県	熊谷市	八王子市	横手市	武蔵野市
当別町	桜井市	柳津町	駒ヶ根市	浦添市
久留米市	敦賀市	宮城県	佐賀県	大泉町
台東区	岐阜県	黒石市	稚内市	印西市
田原市	久留米町	萩市	青森県	
沖縄市	小田原市	海津市	浜松市	
横浜市	千葉県	周防大島町	善通寺市	
豊島区	名古屋市	黒石市	中標津町	
高知市	湯沢市	広島県	川俣町	
京都府	太子町	牧方市	江戸川区	
京田部市	江津市	群馬県	美幌町	
大阪市	朝来市	室蘭市	遠野市	
兵庫県	常滑市	浦添市	大泉市	
奈良市	横瀬	薩摩川内市	埼玉市	
西脇市	知立市	宝塚市	あわじ市	
平内町	三田市	伊丹市	東伊豆町	
滋賀県	東温市	古城市	岡山県	
名張市	福岡市	北谷町	東海市	
大田市	上島町	萩市	豊川市	

業務の実施状況(1)検証環境の提供

② 技術検証基盤の機能追加

▶ GaiXerが具備するWeb検索RAG・プロンプトテンプレート・学習セット機能に加え、ユースケース検証に必要な機能を随時追加
技術検証基盤のアップデートを30回実施し、調達仕様書で3つ以上と定められたLLM(大規模言語モデル)は、最終的に10個提供

技術検証基盤のアップデート

当初機能

- 専用環境
- LLM選択
- Web検索によるRAG
- プロンプトテンプレート
- 追加データ学習

期間中の追加機能

- テンプレート作成時のLLM選択
- サイト/ドメイン指定Web検索RAG
- システムプロンプト
- 用例による法令検索
- 入力可能文字数の増加
- 機密情報チェッカー
- PlayGround
- 学習セットへのメモ追加
- Web検索RAG設定
- Slack連携
- プロンプトテンプレートにおけるWeb検索RAG設定

※ 太枠はご要望に基づき実装した機能。設計書で仕様をご提供

大規模言語モデルの追加

当初から仕様で定められた3種のLLMを提供、
その後も最新のLLMを追加、最終的に10個提供

#	追加日時	追加LLM
1	2023/12/4	・GPT-3.5 ・GPT-3.5-16K
2	2023/12/8	・GPT-4 ・GPT-4-32K ・Bedrock Llama 2 Chat 13B ・Bedrock Claude2
3	2023/1/12	・GPT-4 Turbo ・Bedrock Claude 2.1
4	2023/1/19	・Swallow 70B instruct
5	2023/3/8	・Claude 3 Sonnet

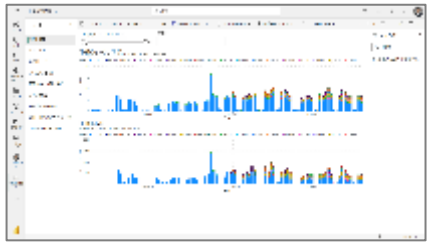
業務の実施状況(2) LLM検証環境の運用・保守

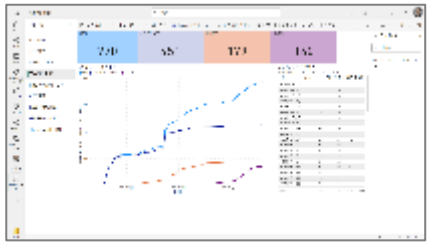
③ ダッシュボード・運用レポート提供

▶ 運用レポート・ダッシュボードを提供

ダッシュボード

PowerBIベースのダッシュボードを提供



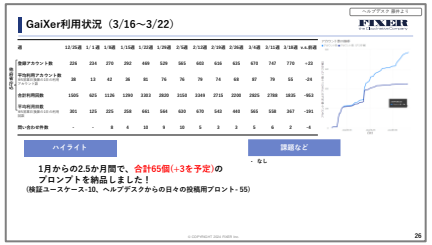


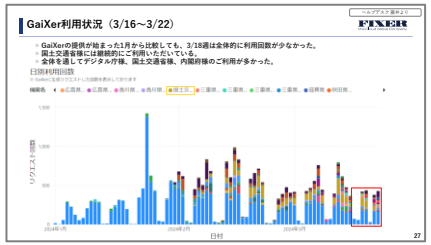
提供情報

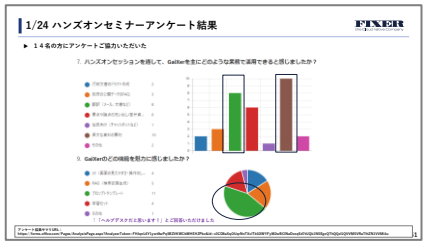
- アカウント数
- 利用継続アカウント数
- 利用ログ
- 機密性情報検出
- 利用回数（組織別・LLM別・週/月別）
- 学習セット利用回数
- Web検索利用回数
- テンプレ利用回数
- 消費トークン数
- Playground利用数 等

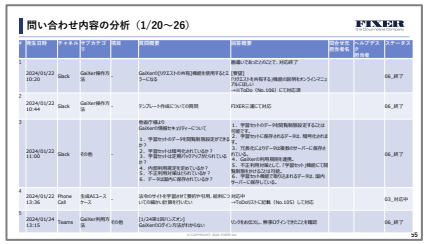
運用レポート

毎月最終のステコミにて運用報告を実施









提供情報

- アカウント数
- 利用回数
- 問い合わせ件数・内容
- 研修実施状況および研修アンケート結果 等

業務の実施状況(2) LLM検証環境の運用・保守

④ ISMAPチェックリスト・セキュリティポリシーへの対応

▶ セキュリティ班とISMAPチェックリスト・セキュリティポリシーへの適合性を確認、脆弱性診断および必要となるシステム改修を実施

ISMAPチェックリスト・セキュリティポリシーを踏まえた課題および対応

No.	基準	項番	内容	課題	対応
1	ISMAP チェック リスト	8.1.2.7.PB 10.1.2.20.PB	クラウドサービス事業者は、クラウドサービス利用者に対し、当該利用者の資産を管理するため、次のいずれか提供している。 (a)当該利用者の管理する資産を、記録媒体に記録する（バックアップを含む）前に暗号化し、当該利用者が暗号鍵を管理し消去する機能 (b) 当該利用者が、当該利用者の管理する資産を記録媒体に記録する(バックアップを含む）前に暗号化し、暗号鍵を管理し消去する機能を実装するために必要 となる情報	利用者暗号鍵管理機能(HYOK)はなし	• 長期保持する情報は有していないため、HYOKまでは不要 • 情報は透過的に暗号化している • 特権的アクセスは制限されており、必要時のみ承認を得て利用している • 利用終了時は、AzureやAWSのサブスクリプションごと削除し、AzureやAWSによってデータが消去される
2	セキュリティ ポリシー	4.2.1.(6).(a).(エ)	設計・設定時の誤りの防止	IasCでの構築や複数名体制でのチェックを行っているが、ツール等での客観的チェックはなし	• Azure/AWSが提供するセキュリティツールでのチェック • 脆弱性診断ツールの導入 • セキュリティツールでのチェック、および脆弱性診断ツールを導入
3	セキュリティ ポリシー	6.1.1.(1).(c) 6.1.1.(1).(g)	• 主体認証機能の導入(多要素認証) • 主体認証情報の不正利用対策(利用停止、主体認証情報の再設定)	多要素認証なし	• ID/Password認証に、メールOTP認証を組み合わせる方針。 • 認証機能をリリース
4	セキュリティ ポリシー	6.1.5.(1).(a).(ア) 6.1.5.(1).(a).(イ) 7.1.2.(2).(d).(ウ)	• 要機密情報を取扱う情報システムでの暗号化の付与・検証の検討と対応 • 要保全情報を取扱う情報システムでの電子署名の付与・検証の検討と対応 • サーバー監視としてのファイル完全性チェックツール(サーバーの利用環境等から必要な場合)	左記の３機能なし	• 学習セットを管理しているシステムの設定完全性チェック(VM、NSG) • 学習セットを扱うVMへの不正アクセス監視 • 上記以外の、アプリケーションログや利用者情報など保護が必要な情報への不正アクセスの監視
5	セキュリティ ポリシー	6.3.1.(2).(a).(イ)	• 提供するアプリケーション・コンテンツが脆弱性を含まないこと	都度のアプリケーションリリースのタイミングでの脆弱性診断は未実施	• 脆弱性診断ツールの導入 • ツールを導入

業務の実施状況(3) 技術検証支援

⑤ ヘルプデスク運用

- ▶ ヘルプデスクを56日間稼働、64件の質問に対応。ユーザとの対話で発見された不具合等を開発チームに迅速に連携
- ▶ 地方公共団体に対する技術検証基盤参画の呼びかけ、申請受付からユーザ追加まで対応し、基盤利用者の拡大を支援

ヘルプデスク運用

ヘルプデスクを1月9日～3月29日の計56日間運用(10-17時)、
利用者からの計64件の質問に生成AIコンサルタントと連携して対応。
日次で朝会を実施し、タイムリーに対応状況を確認

朝会ログの例 (3月5日実施分)

GaiXer ヘルプデスク 朝会 (3/5 10:00)

- 昨晩から今朝にかけての時間外チャットなし
- 本日のFAQ投稿内容の確認
  - [新機能] 3/1リリースの新機能：PDFの画像データを学習対象から外す機能
  - [新機能] 3/1リリースの新機能：Playgroundのカスタムモデルから作成したテンプレートを区別できるように
  - [不具合] 3/1リリースの新機能：同一チャット内でシステムプロンプトが途中で消える不具合を修正
- GaiXerの利用状況について
  - 登録アカウント数：636（前日から+1）
  - 利用アカウント数：44（前日から+4）
  - 利用した回数：248（前日から+5）
- 昨日から他府省庁向けにヘルプデスクを開放したがウェブ会議から抜けるとチャットが途切れてしまい以降のやり取りができない件
  - デジタル庁職員も他の府省庁とは繋がっておらず連絡手段はメールしかない
  - Teamsチャットで回答ができない状態が発生したらヘルプデスクからSlackの朝会スレッドに共有してもらう

利用者の声からのシステムアップデート

問い合わせ内容はヘルプデスクチームで精査し、
システム改修を必要とする場合、速やかに開発チームに
フィードバックし、バックログに掲載。順次改修を実施

ヘルプデスクに寄せられた声を踏まえたアップデート

#	日時	変更内容
3	2023/12/11	ユーザ情報取得APIを正しくコールするように修正
4	2023/12/19	Admin-PortalのContainerApps hostedservice-blueのスケールアップ作業
6	2023/12/21	新UIのトピック生成用 ContainerApps、AzureFunctionsを構築
8	2023/12/26	LLM選択箇所の増加・各回答に利用したLLMの表示対応、軽微なレイアウト改修
9	2023/12/26	一部のテンプレートで編集ができない不具合の修正対応
10	2023/12/28	Admin Portal上でユーザーをCSVで取得した際に、管理者情報が正しく取得されない不具合を修正
11	2023/12/28	通常の生成とテンプレート作成時の「試しに生成」時にプロンプトの送信方法が違うことで、 出力結果に著しい差が出てしまう不具合の修正。プロンプト設定の追加
13	2023/12/28	登録時に確認コードを自動送信する際の電子メールが空メッセージになっている不具合を解消
14	2024/01/09	ログイン直後にフリーズする問題の解決 (要約元のテキスト量が多く、履歴表示の途中でエラーが発生)

業務の実施状況(3) 技術検証支援

⑥ 研修コンテンツ制作・研修実施

▶ 技術検証基盤の研修コンテンツを制作。研修を12回実施し、計128人が受講(研修資料は本報告書の別添1としてご納入)

研修名	実施日	参加人数	目的・内容
庁内説明会 (All Hands)	12月21日	–	**デジタル庁メンバーに対する、技術検証プロジェクトの周知および基盤利用開始に当たってのアナウンス** • アカウントの登録方法 • 基本的な利用方法
勉強会	1月13~16日 (8回開催)	79人 (8回合計)	**生成AIの導入と基本的な使い方の理解** • 利用登録方法 • 基本的な使い方 • プロンプティング入門 • テンプレートの使い方 • 学習セットの使い方
ハンズオンセッション	1月24日 (4回開催) ※ オフライン	49人 (4回合計)	**オフラインでのハンズオン研修を通じたより高度なプロンプト作成手法の定着** • 最新のLLM紹介 • 高度なプロンプティング

ハンズオンセッションの様子
(於 デジタル庁オフィス)

業務の実施状況(3) 技術検証支援

⑦ ユースケース検証

▶ 調達仕様書記載のユースケースA～Eに加え、独自提案の「F. 機密度判定」を加え、技術検証参画メンバーとの協働を通じて計9個のユースケースでLLMの有用性を検証し、今後のデータ整備・データモデルに関する示唆を検討

ユースケース	対象業務	2024 1月	2月	3月
A 背景情報および特定の論点に基づく文書作成（国際会議や部内会議資料）	A-1. パブリックコメント対応品質の向上	概要/内容 検証手法 → 実験 → ユーザー評価 → ブラッシュアップ → 考察/結論 → レポート		
	A-2. 調達仕様書のラベリング	概要/内容 検証手法 → 実験 → ユーザー評価 → ブラッシュアップ → 考察/結論 → レポート		
B 特定分野における問合せ対応案の自動作成	B-1. デジタル・ガバメント推進標準ガイドライン実践ガイドブック ※自治体向けにBot展開	概要/内容 検証手法 → 実験 → 考察/結論 → レポート		
		概要/内容 → 仕様検討 → 実装 → 実験 → 考察/結論 → レポート		
	B-2. 人事院規則	概要/内容 検証手法 → 実験 → ユーザー評価 → 考察/結論 → レポート		
	B-3. デジタル庁問い合わせ窓口	A-1と合流して検証・レポート作成		
C 特定の文書ルールに沿った文書校正（仕様書・挨拶文等）	C-1. 用例による法令検索	概要/内容 → 業務要件定義 → 実装 / 検証手法 → 実験 → 考察/結論 → レポート		
	C-2. 用語集作成業務の効率化	概要/内容 → 業務要件定義 → 実験 → 考察/結論 → レポート		
D 既存の質問主意書に対する答弁案の作成	D-1. 質問主意書に対する答弁ドラフト	ユースケースの一部を検証・レポート作成		
E 既存コードの解析による技術文書の作成、コード変換など	E-1. ソースコードの言語変換	ユースケースの一部を検証・レポート作成		
F 機密性情報の分類・検知	F-1. ドキュメントの機密度判定	概要/内容 → 検証手法 → 実験 → 考察/結論 → レポート		

業務の実施状況(3) 技術検証支援

⑧ プロンプト開発

▶ 検証過程において、ユースケースから12個、ヘルプデスクからの啓蒙発信として62個、合計74個のプロンプトを作成
（作成プロンプト集は本報告書の別添2としてご納入）

背景

各ユースケースの検証過程の他、
利用者への啓蒙として、
毎日ヘルプデスクから発信した
プロンプトを知見として集約

各プロンプトに対し、利用LLM・Web検索RAGのON/OFFといった設定、リクエスト、システムプロンプトを提供

作成プロンプト一覧

No.	プロンプト名
1	A-1_パブコメ分類
2	A-1_パブコメ意見変換
3	A-1_パブコメ意見元抽出
4	A-1_パブコメ回答案作成
5	A-2_調達仕様書ラベル付け
6	B-1_デジタル社会推進実践ガイドブック
7	B-2_人事院規則
8	C-2_用語集生成
9	D-1_質問主意書に対する答弁ドラフト生成
10	E-1.Java言語をPython言語に変換するプロンプト
11	E-1.COBOL言語をJava言語に変換するプロンプト
12	F-1_機密性情報分類
13	テクノロジーニュースの検索
14	【1・31投稿】フォーマルな返答文の生成
15	【1・31投稿】学習セットを参考に回答
16	【2・1投稿】自己PR作成
17	【2・2投稿】誤字脱字チェック＆修正
18	【2・2投稿】〇〇に向けた文章変換
19	【2・5投稿】サービス説明文作成
20	【2・6投稿】文章の品質向上

No.	プロンプト名
21	【2・6投稿】専門用語解説プロンプト
22	【2・6投稿】観光案内チャットボット
23	【2・7投稿】感情分析
24	【2・7投稿】文章に改行を挿入
25	【2・8投稿】ネガティブな文章を中立的な文章に変換
26	【2・16投稿】課題解決アドバイザー
27	【2・16投稿】日報分析まとめ
28	【2・16投稿】プロンプト改善案提案
29	【2・20投稿】SNSに投稿する文章作成
30	【2・20投稿】間違った主張を見抜く
31	【2・20投稿】業務に用いるツールの選定
32	【2・21投稿】適切なアイスブレイクを考える
33	【2・21投稿】適切な表現を用いて文章を補完する
34	【2・22投稿】ピッチを考えてもらう
35	【2・22投稿】イベント広報案作成
36	【2・22投稿】アンケート作成
37	【2・26投稿】議事録から議事内容の評価をしてもらう
38	【2・26投稿】政策に対する意見生成
39	【2・26投稿】会議の事前準備のチェックリスト生成
40	【2・28投稿】条文改正によって想定される影響生成

No.	プロンプト名
41	【2・28投稿】発言からニーズを見抜く
42	【2・29投稿】会議の文字起こしの内容をわかりやすく要約する
43	【2・29投稿】アンケート結果の傾向や意見を抽出
44	【2・29投稿】評価要件を洗い出し段階評価基準を作成
45	【3・1投稿】ディベートのシミュレーション
46	【3・1投稿】目標達成事項を考えてもらう
47	【3・8投稿】業務を複数のタスクに分解する
48	【3・8投稿】英文をネイティブが用いる表現に変える
49	【3・12投稿】英会話の練習をする
50	【3・12投稿】データの分類分け
51	【3・13投稿】リスクの推定と危機管理シナリオの考案
52	【3・13投稿】課題解決をロジックツリーを用いて手助け
53	【3・13投稿】難しい単語を分かりやすく説明する
54	【3・14投稿】ダミーデータを作成する
55	【3・14投稿】外国の方から見た印象を教えてもらう
56	【3・14投稿】面接の質問を作成する(Claude3)
57	【3・15投稿】プロダクトの命名(Claude3)

No.	プロンプト名
58	【3・18投稿】顧客満足度を上昇させる手助けを行う
59	【3・18投稿】慣用句の意味を教える-Claude3-
60	【3・19投稿】多言語でキャッチコピー作成
61	【3・21投稿】excel関数の提案(Claude3)
62	【3・21投稿】長所と短所を分析する
63	【3・21投稿】研修のアジェンダを作成
64	【3・22投稿】プロジェクト計画書の作成
65	【3・22投稿】最適な学習プランを提案
66	【3・22投稿】DX案を提示する
67	【3・25投稿】プログラミング言語勉強先生
68	【3・25投稿】文章から用語集を作成
69	【3・25投稿】送信シーンに適した周知文を作成する
70	【3・26投稿】業務での生成AI活用法を提案する
71	【3・26投稿】csvコンバータ
72	【3・26投稿】絵文字が多い文章に変換する
73	【3・27投稿】単語の発音方法とコツを教えてくれる
74	【3・27投稿】複数のタスクに分解(Claude3)

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

対象ユースケースの全体像

業務品質向上が主軸	業務改善、時間短縮が主軸
他ユースケースに合流	

▶ 行政の業務において、生成AIが活用可能である(=生成AIに任せられる領域がある)業務に特化して検証を実施した

	ユースケース	対象業務	業務概要
A	背景情報および特定の論点に基づく文書作成（国際会議や部内会議資料）	A-1. パブリックコメント対応品質の向上	提出を受けたパブリックコメントに対する分析および回答案の作成
		A-2. 調達仕様書のラベリング	調達仕様を特徴付けるラベリング/タグ付け
B	特定分野における問合せ対応案の自動作成	B-1. デジタル・ガバメント推進標準ガイドライン実践ガイドブック ※自治体向けにBot展開	デジタル・ガバメント推進標準ガイドライン実践ガイドブックに関する問い合わせの回答案を作成
			同上の業務概要を自治体向けbotにて展開
		B-2. 人事院規則	人事院規則に関する問い合わせの回答案を作成
		B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合
C	特定の文書ルールに沿った文書校正（仕様書・挨拶文等）	C-1. 用例による法令検索	類似法令の検索と要約
		C-2. 用語集作成業務の効率化	対象ドキュメントの用語集の作成
D	既存の質問主意書に対する答弁案の作成支援	D-1. 質問主意書に対する答弁ドラフト生成	質問主意書に対し、過去の答弁を参考にして答弁ドラフトを作成
E	既存コードの解析による技術文書の作成、コード変換など	E-1. ソースコードの言語変換	Java⇒Python・COBOL⇒Javaの言語変換が可能なことを確認
F	機密性情報の分類・検知	F-1. ドキュメントの機密度判定	政府統一基準に則り、情報/文書の機密性を判定

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

ユースケース検証手法の全体像

(1) 検証プロセス

(2) 検証体制

(3) LLMの選定

(4) プロンプトの開発

(5) パフォーマンスの測定

(6) テストケースの作成

ユースケース検証手法(1) 検証プロセス

検証手順

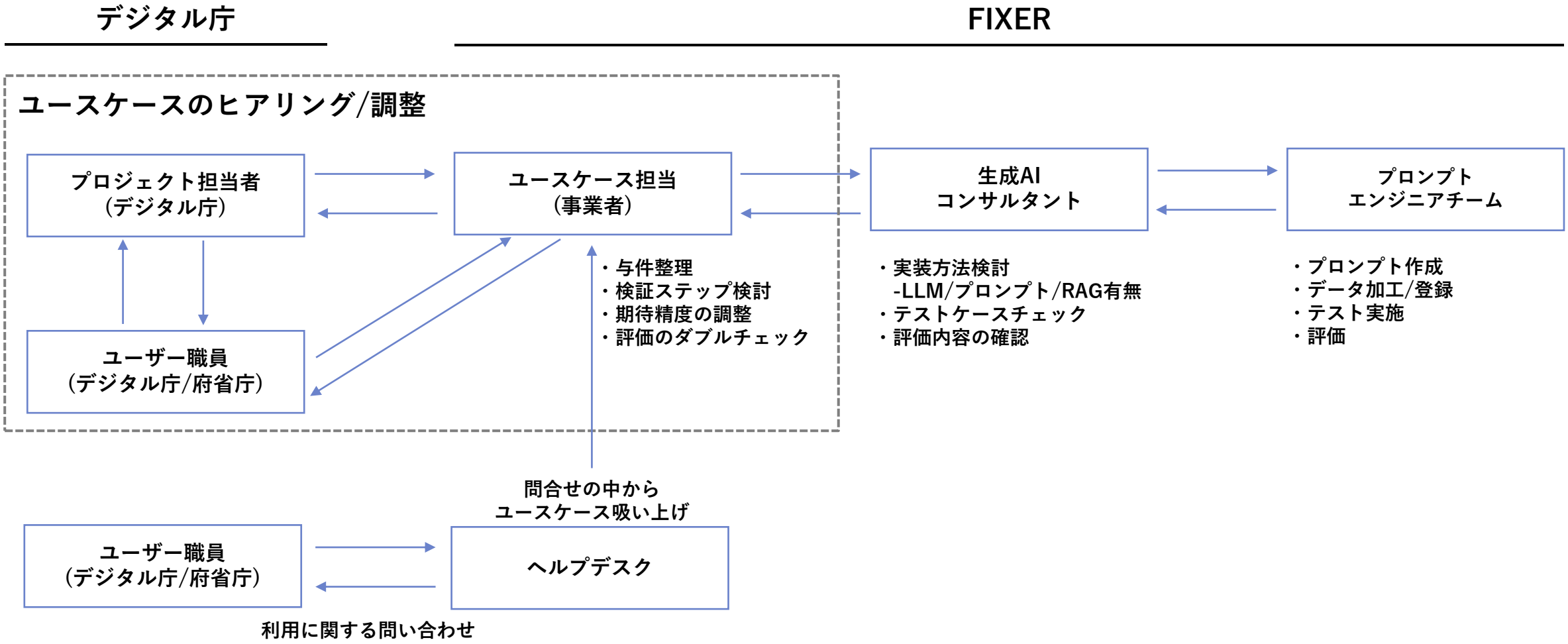
- ▶ 用途のヒアリングを行った時点でクイック検証を実施、生成AIによる課題解決の可能性を評価し、ユースケース化を判断
- ▶ 本検証の前にスモールテストを実施して方針を調整し、全体工数を低減

大分類	中分類	担当	インプット/アウトプット/テストケース	作業
ヒアリング	用途のヒアリング	ユースケース担当	インプット: 架空データ＋シンプルなプロンプト テストケース: 2~3問 アウトプット: 一部のみ出力	・生成AIへの向き/不向き ・実際の業務との整合性（机上の空論は×） ・想定される回答精度 → ユースケースとして取り上げるかを判断
	アウトプット/インプットの定義	ユースケース担当		
	クイック検証の実施	プロンプト エンジニア		
	担当者/ユーザーによる確認	ユースケース担当 ・ユーザー職員		
スモールテスト 方針調整	検証対象データの入手	ユースケース担当	インプット: 実データの一部（必要に応じて加工） テストケース: 各ステップで約10問 アウトプット: 各ステップで出力	・LLM/プロンプト調整による回答精度差異 ・ステップ毎での精度の違い → ユースケースの実施範囲を調整 （e.g. 低精度 or 効率的でない手順は対象外）
	検証ステップの定義	ユースケース担当		
	評価項目の定義	生成AI コンサルタント		
	実装(プロンプト・学習データ)	プロンプト エンジニア		
	各ステップ毎での小規模検証	プロンプト エンジニア		
本検証 ユーザー評価 （納品）	対象データ全体での検証	プロンプト エンジニア	インプット: 全ての実データ テストケース: 網羅性のあるテストケース - データのほぼ全ての要点を - 複数箇所にまたがる/選択問題等、様々な角度から アウトプット: 全ての出力	・データ増加に伴うハルシネーション ・多角的なテストケースへの回答傾向 ・上記に基づく最終回答精度 ・納品に向けた単純化の余地 → ユーザー業務での実践検証を調整
	ユーザーによる評価	ユーザー職員		
	実装を単純化してユーザーに納品 （他のデータでも実装しやすい工夫）	ユースケース担当		

ユースケース検証手法(2) 体制

機能開発を伴わないケース

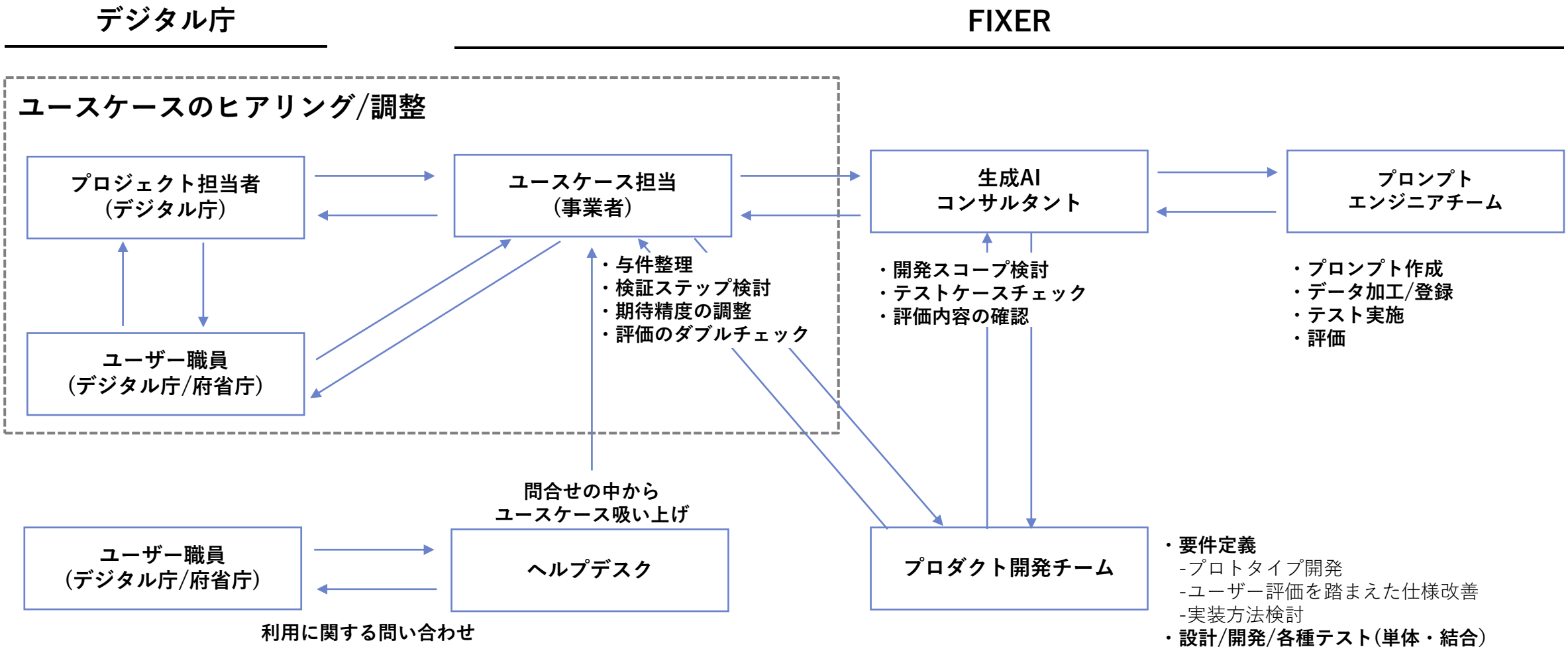
▶ ユーザー職員とFIXERのユースケース担当者が密に協働してユースケースを推進。推進に際しては、FIXERの生成AIコンサルタントとプロンプトエンジニアが最新の技術トレンドや事例に基づくナレッジを提供。ユースケースの発掘においては、ヘルプデスクとも連携



ユースケース検証手法(2) 体制

機能開発を伴うケース

▶ ユースケースの検証にあたり、大量データの前処理・後処理や他システムとの連携が必要な場合[1)]、FIXERのユースケース担当者が与件を整理し、プロダクト開発チームがバックログに掲載した上で、アジャイルに実装



1. 具体的には「B-1. デジタル・ガバメント推進標準ガイドライン実践ガイドブック」「C-1. 法令による用例検索」が該当

ユースケース検証手法(2) 体制

各ユースケースのテストケース数/実施工数

▶ 各ユースケースにおけるテストケース数・作業工数は下記の通り

検証ユースケース	体制	テストケース数	実施工数
A-1. パブリックコメント対応品質の向上	3人	10(コメント数)	12人日
A-2. 調達仕様書のラベリング	3人	10(調達仕様書分)	15人日
B-1. デジタル・ガバメント推進標準ガイドライン実践ガイドブック	7人	588	56人日
B-2. 人事院規則	4人	62	25人日
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合		
C-1. 法令による用例検索	B-3の事例を参考にプロダクト開発・機能提供を実施し、テストは実施せず		
C-2. 用語集作成業務の効率化	2人	1(1ドキュメントのみで検証)	5人日
D-1. 質問主意書に対する答弁ドラフト生成	2人	10	6人日
E-1. ソースコードの言語変換	2人	20	15人日
F-1. ドキュメントの機密度判定	2人	50	12人日

主要LLMの特徴

▶ ユースケースの検証において採用した言語モデルおよびその特徴は下記の通り。GPT-4 Turboは高品質が求められるユースケース、Claude 2.1/3 Sonnetはプロンプトに入力するトークン数が多いユースケースに適していると想定

	GPT-4 Turbo	Claude 2.1	Claude 3 Sonnet
生成品質	100※1	59※1	83※1
コンテキストウィンドウ(入力可能トークン数)	128,000※1	200,000※1	200,000※1
プロンプト試行コスト	入力：0.01ドル/kトークン※2 出力：0.03ドル/kトークン※2	入力：0.008ドル/kトークン※3 出力：0.024ドル/kトークン※3	入力：0.003ドル/kトークン※3 出力：0.015ドル/kトークン※3
回答生成速度(生成トークン数/1s)	15※1	34※1	54※1
利用に適していると想定されるユースケース	高品質が求められるユースケース	プロンプトに入力するトークン数が多いユースケース	

※1 「Independent analysis of AI models and hosting providers」より引用
※2 「Azure OpenAI Serviceの価格」よりGPT-4-Turboの単価をベースに引用
※3 「Claude 3 Pricing」よりClaude 2.1、Claude 3 Sonnetの料金プランを引用

ユースケース検証手法(3) LLMの選定

各ユースケースの性格分析/LLM選定

▶ 各ユースケースの性格を分析し、前述の特徴を踏まえて利用するLLMを選定

検証ユースケース	最適な言語モデル	ユースケースの性格
A-1. パブリックコメントの対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1.デジタル・ガバメント推進標準ガイドライン実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

意図理解に優れ、柔軟にプロンプトへ対処できるとという観点から、 GPT-4 Turboを採用。但し、長大なコンテキストフレームを求められるケース(B-2)はClaude(2.1/3 Sonnet)を採用。また、正確かつ簡潔な抽出を求められるB-3のようなユースケースは、検証を進める内にClaudeの方が正確な引用スコアを出す傾向が見られたため同じく採用。

プロンプト改善のポイント

▶ LLMの特徴・原理を考慮に入れたうえで、各課題において以下のような手法でプロンプトを改善し、アウトプット品質を向上

	課題	改善方針
①	コンテキストフレームに収まらない文書やテキスト情報を扱う場合	• コンテキストを落とさないような情報圧縮を生成AIで補助
②	RAGの精度に課題がある場合	• ファイルの拡張子やJSON形式、およびSystemプロンプトの利用などをユースケースごとに情報設計を最適化
③	一般的な質問では解決できないユースケースを扱う場合	• Chain-of-ThoughtやFew-shot、JSONのような論文を根拠とした汎用性のある手法や、LLMごとの多く機械学習されているであろうデータ形式(マークダウンやXMLタグ)などをLLMごとにディレクション
④	分類の正答率が上がらない場合	• LLMのエンティティ分析能力の柔軟性を理解し、できる前提でプロンプトのディレクションを選択肢にして解決 • PlaygroundによるLLMごとのパラメータの調整も選択肢に入れて、プロンプト×モデル選定×パラメータの3変数でのディレクション
⑤	そもそも課題を定義できない場合（LLM特有の性質等により）	• 原理である確率的推論に基づいて、そもそもの想定生成を点ではなく範囲で、再現性を評価 • プロンプトの量が多い場合には、全てを要件定義せず、範囲の両端のみを定義し、グラデーションをコンテキストドリブンで任せ、LLMが得意とする点と点を線で結ぶ能力を発揮

ユースケース検証手法(5) パフォーマンスの測定

定量評価・定性評価の枠組

▶ 各ユースケースにおけるLLM・プロンプトのパフォーマンス測定として、あらかじめ定めた評価基準に基づいた定量評価とユーザーによる定性評価を実施

定量評価

- ユーザーヒアリングをもとに、各ユースケースにおける評価基準を設定
- 評価基準ごとに点数をつけるための評価要件を設定
- 評価要件に基づいて、各テストケースにおける生成結果を採点
- 採点結果を集計し、ユースケースにおける生成AIのパフォーマンスを測定

評価基準の決定方法

以下①②をベースとして、
各ユースケースにあてはまる評価基準を設定

① 回答の正しさ
- 引用箇所が正しいか、引用元と内容に齟齬がないかといった正確性に関する基準

② わかりやすさ
- 理解しやすい文章になっているか、質問内容に答えられているかといったわかりやすさに関する基準

評価要件/点数の決定方法

ユースケースの性質に応じて、
以下①②の評価要件/点数のパターンを設定

① 部分点なし(0点 or 1点)
- 機密性情報の判定のような正誤が明確な評価基準においては、部分点を設定せず

② 部分点あり
- 上記以外の設問においては、部分点を与えるように評価要件/点数を設定

定性評価

- 各ユースケースの実業務担当者にヒアリングを実施
- 生成結果に対する評価、および改善方針に関するコメントを受領

担当者目線での評価要件

実業務で利用可能かいう観点で評価

① 生成結果に対して、担当者がどの程度の手直しを行う必要があるか

② 現状の実装方式に基づく生成結果が、担当者にとって使い勝手がいいか
(自動化/システム化が必要な場合は、将来像をシーケンス図として表現)

各ユースケースにおける評価基準

▶ 各ユースケースにおいて、前頁に記載した定量評価・定性評価を下記の通り実施

※一部ユースケースにおいては、検証期間の関係で担当者の時間が取れなかったため未実施

検証ユースケース	定量評価				定性評価
	評価基準		評価要件/点数		
	回答の正しさ	わかりやすさ	部分点あり	部分点なし	
A-1. パブリックコメント対応品質の向上	○	○	○	ー	ー
A-2. 調達仕様書のラベリング	○	○	○	ー	○
B-1. デジタル・ガバメント推進標準ガイドライン実践ガイドブック	○	○	○	ー	○
B-2. 人事院規則	○	○	○	ー	○
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合				
C-1. 用例による法令検索	ー	ー	ー	ー	○
C-2. 用語集作成業務の効率化	○	○	○	ー	ー
D-1. 質問主意書に対する答弁ドラフト生成	○	○	○	ー	ー
E-1. ソースコードの言語変換	ー	ー	ー	ー	○
F-1. ドキュメントの機密度判定	○	ー	ー	○	○

テストケースの観点と作成方法

▶ テストケースの作成では、下記の観点を考慮

	観点	適用対象	作成方法	具体例
①	評価要件による場合分け	• 評価基準が機械的に判断可能なケース	• 評価要件に基づく点数が異なるパターンを網羅 • 評価要件が複数ある場合は複数の基準を組み合わせ	機密度判定では、評価要件の個人情報の有無 × 行政業務への影響有無でテストケースを作成 ① 個人情報あり・行政業務影響あり ② 個人情報あり・行政業務影響なし ③ 個人情報なし・行政業務影響あり ④ 個人情報なし・行政業務影響なし
②	設問形式による場合分け	• 問い合わせ対応等、多様な質問が想定されるケース	• オープンクエスチョンと選択式のクローズドクエスチョンの両者を作成	実践ガイドブックでは、様々な問い合わせを想定し、方法を質問するようなオープンクエスチョンに加え、選択式の正誤問題をケースとして追加
③	ハルシネーション防止を確認するケースの追加	• 指定学習セット以外の情報(例: Web検索)を参照してはいけないケース	• 指定学習セットには含まれない情報を元に作成	質問主意書への答弁案生成では、学習データに回答が含まれる場合、回答が含まれない(答えようがない)場合の2ケースを作成 • 後者の場合にハルシネーションを回避し、わからないと答えられるかを確認

各ユースケースにおけるテストケース

▶ 各ユースケースにおいて、前頁に記載した観点を踏まえ、下記の通りテストケースを作成

検証ユースケース	評価要件による場合分け	設問形式による場合分け	ハルシネーション防止を確認するケースの追加
A-1. パブリックコメント対応品質の向上	○	–	–
A-2. 調達仕様書のラベリング	○	–	–
B-1. デジタル・ガバメント推進標準ガイドライン実践ガイドブック	○	○	○
B-2. 人事院規則	○	○	–
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合		
C-1. 法令による用例検索	B-3の事例を参考にプロダクト開発・機能提供を実施し、テストは実施せず		
C-2. 用語集作成業務の効率化	○	–	○
D-1. 質問主意書に対する答弁ドラフト生成	○	–	–
E-1. ソースコードの言語変換	○	–	–
F-1. ドキュメントの機密度判定	○	–	○

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ハンドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ハンドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

A-1. パブリックコメント対応品質の向上 (1/4)

1. ユースケース概要

対象業務	• 政策・法案等に対して提出されたパブリックコメントの分類や分析
前提	• 政府各担当課は通常、関連政策や法令に関して国民や民間企業から広くパブリックコメントを収集する。 • パブリックコメントはそのすべてに対して国が真摯に対応し、考慮すべきものであり、 今回の生成AI検証では、AI任せではなく行政官の対応品質向上に主眼を置いて検証した。
課題	• 行政官は、時に膨大な件数(1万件)にもなるコメントの集計/分類に基づく意見作成に忙殺されている。 その際、難易度が高くない業務工程の効率化→高難度業務への対応時間確保が課題となっている。
実現したいこと	• パブリックコメントに基づく意見作成のための難易度が高くない作業工程の効率化を図り、 行政官の対応品質向上を図る

A-1. パブリックコメント対応品質の向上 (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	1. パブリックコメントの主旨の分類(賛同、今後の施策の参考意見、関係ない意見)が正確か 2. 不適切なコメントの変換(表現改善、論点分割)が正確か 3. コメントへの回答案の品質が担保されているか 4. コメントと資料のマッピング精度が担保されているか
	仮説	1. 定性タスクでもある事から、比較的高精度を期待される 2. 同上 3. 文脈を踏まえた高度な回答が求められることもあり、職員レベルの回答品質は疑問視される 4. 資料のフォーマットがまばらであり、媒体への依存が想定される
検証手法	実験設計	• **実験デザイン** – Excelにて手動で分類/回答済みのものを、プロンプト生成内容と比較する – 運用利便性を重視し、参考資料は学習データやプロンプトに組み込まない • **実験手順** – 1.)パブリックコメントのExcel表をテキスト化 – 2.)プロンプトを組み、1.)を組み込む – 3.)主にコメント変換タスクの結果をもとに元データの提供者と相談し、業務内容を精緻にヒアリング
	取得データ	• 総務省から受領したパブリックコメント(10件)
	KPI	• 現職職員の作業結果との合致度

A-1. パブリックコメント対応品質の向上 (3/4)

4. 実験結果 / 5. 考察

評価内容	評価手法	• **評価項目** ○ 意見分類: 正しく分類できているか / 分類の判断理由を説明しているか /分類の判断元箇所を引用しているか ○ 意見変換: 形式的な修正の指摘 / 複数回答が必要な意見の分割 / 不適切な文言の言い換え /意見の要約 ○ 回答案作成：質問に対して回答をしているか / 正解データと同じ内容を回答しているか ○ 意見元抽出：意見元の場所が提示されているか / 意見元の場所が正しいか
	評価結果	• **定量評価(検証項目：得点率)** ○ 意見分類：90% ○ 意見変換：90% ○ 回答案作成：90%(回答部分の得点率は60%) ○ 意見元抽出：60%(参照箇所の得点率は40%)
考察	実験結果の意味合い	• **分類 / 変換は概ね実施できた。回答案作成 / 意見元抽出は、文脈理解が必要になる難易度の高い作業であり、かつ回答に必要な情報のデータ形式が煩雑であることから得点率が上がらなかった**
	当初期待との差分	• **検証当初の仮説の通り**
	実験の限界点	• **回答案作成 / 意見元抽出は、文脈理解が必要となるため、参考材料として採用する運用が現実的である**

A-1. パブリックコメント対応品質の向上 (4/4)

6. 結論

実験の成果/影響	• 生成AIを用いたパブリックコメントの分類や論点整理は担当者の業務負担を改善しながらコメントへの対応品質向上に貢献するものであり、有効と考える
今後の展望	• 年間約2000案件に対応しているものと仮定した場合、分類/論点整理などのタスクについては、補助的にAIを活用する事で対応品質の向上が期待できる • コメント件数が多いもの(1万件)等に対しても十分に力を発揮する事ができるため、システム化する事で生成AIの対応範囲を標準化、スケール適用の効果が期待できる
推奨や提案	• システム実装時に府省庁ごとでのネットワークやデータベースの違いなどを考慮して標準導入することを推奨する

A-2. 調達仕様書のラベリング (1/4)

1. ユースケース概要

対象業務	• 新たに調達仕様書を作成する際、参考資料として類似する過去調達仕様書を参照する作業
前提	• 府省庁は年間15,000件(R3年度実績)の入札を実施しており、全て調達仕様書に基づき調達がなされる • 政府にとって調達品質の改善と効率化は重要命題であり、そのデータベースの充実や整備ニーズは高い
課題	• 担当者が本業務を行う際、各仕様書の内容を確認しないと案件の属性が抽出できず、業務難易度に対して一定の作業時間を要することが課題
実現したいこと	• 調達仕様書の作成業務効率化と品質向上を図ること – 調達仕様書の内容から案件名と案件概要の作成を行う – 調達仕様書のラベルの各項目を案件概要のテキスト情報から抽出もしくは解釈して割り当てる

A-2. 調達仕様書のラベリング (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	1. 生成された案件名と案件概要の精度が担保されているか 2. 生成されたラベルが現場担当者の作業レベルで担保されているか
	仮説	1. 定性タスクであるため、比較的高精度を期待される 2. 同上
検証手法	実験設計	• **実験デザイン** – Excelの人力でラベリング済みのものとGaiXerで生成したものを比較する • **実験手順** – 1.)プロンプトを作成 – 2.)調達仕様書を読み込み案件名、案件概要を作成。ラベリングを実行 – 3.)出力された結果を人力で作成したものと比較する
	取得データ	• デジタル庁から提供された調達仕様書10件(未ラベリング)
	KPI	• 担当者によるラベリングとの合致度 / 修正所要時間

A-2. 調達仕様書のラベリング (3/4)

4. 実験結果 / 5. 考察

評価内容	評価手法	• **評価項目** ○ 回答の正しさ1:調達仕様書の案件名が完全一致 ○ 表現の適切さ1:案件概要の適切な記述 ○ 回答の正しさ2:ラベル漏れの項目はないか ○ 回答の正しさ3:選択したラベルが適切か
	評価結果	• **定量評価(検証項目：得点率)** ○ 調達仕様書の案件名が完全一致:100% ○ 案件概要の適切な記述:83% ○ ラベル漏れの項目はないか:100% ○ 選択したラベルが適切か:50%(正解ラベルを含むが、付与過多があるため部分点となり50%)
考察	実験結果の意味合い	• **選択したラベルが適切かという評価基準における得点率は50%にとどまったものの、全てのケースで正解データのラベルを選択したうえで付与したラベルの付与が多い事象であり、ユーザーによる評価は実業務利用において許容範囲内であった**
	当初期待との差分	• **検証当初の仮説の通り**
	実験の限界点	• **正解データのラベル以上に余分にラベルを付けてしまう事例が目立った**

A-2. 調達仕様書のラベリング (4/4)

6. 結論

実験の成果/影響	• 調達仕様書のラベリングは生成AIの活用用途として適切であり、導入対効果が見込まれる
今後の展望	• 担当者からの定性評価として、本ユースケースは有用とのコメントも受領しており、今後の本格実装が期待できる
推奨や提案	• 将来的に本ユースケースをシステム化する際には、ラベリング結果や内容に対してユーザーがフィードバック/指摘できるインターフェースを実装する事 • システム実装時に府省庁ごとでのネットワークやデータベースの違いなどを考慮して標準導入することを推奨する

B-1. 標準ガイドライン実践ガイドブック (1/4)

1. ユースケース概要

対象業務	• デジタル・ガバメント推進標準ガイドライン実践ガイドブックに対するFAQ対応
前提	• 複数の政府ガイドラインが記された文書と人の知見を用いて問合せ対応を行っている職員の業務(本務ではない)において、生成AIによる代行対応の可能性を検証することで、行政におけるコミュニケーションコストの最適化を検証するもの。
課題	• 今回はデジタル庁で調達関連の問い合わせを受けるチームのタスクに着目。本チームは、各職員からの調達に関する問い合わせを人の手で対応しており、回答までに工数と時間を要することが、質問者・回答担当者ともに業務負担になっている点を課題とした
実現したいこと	• デジタル・ガバメント推進標準ガイドライン実践ガイドブックに関するFAQ対応の業務効率化を図ること

B-1. 標準ガイドライン実践ガイドブック (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	1. 引用箇所が正しいかどうか 2. 回答文が引用箇所を正しく反映しているか 3. 表現の仕方が適切であるか
	仮説	1.2.関連する質問であれば本文より引用して回答を生成出来る 3.関係のない質問には「わかりません」と回答することができる
検証手法	実験設計	• **実験デザイン** メソッド1.プロンプトに要約された本ドキュメント全文(PDF446頁)を格納することで参照(長大なプロンプト故、claude2.1のみで生成) メソッド2.RAGでドキュメントのPDFを登録することで参照 • **実験手順** – 1.)本ドキュメントの塊ごとの要約 or PDFそのものの学習セット登録 – 2.)claude2.1を前提に、XMLタグを用いてプロンプト開発 – 3.)各章や章を跨いだ網羅的なテストケースを作成して検証
	取得データ	• デジタル・ガバメント推進標準ガイドライン実践ガイドブック
	KPI	• 正解データとの合致度

B-1. 標準ガイドライン実践ガイドブック (3/4)

4. 実験結果 / 5. 考察

評価内容	評価手法	• **評価項目** ○ 回答の正しさ1:引用箇所が正しいかどうか ○ 回答の正しさ2:回答文が引用箇所を正しく反映しているか ○ わかりやすさ:表現の仕方が適切であるか
	評価結果	• **定量評価 (検証項目：得点率)** ○ 引用箇所が正しいかどうか:80% ○ 回答文が引用箇所を正しく反映しているか:91% ○ 表現の仕方が適切であるか:85% • **定性評価** ○ 長大なコンテキストを踏まえた回答を確認できる一方で、あいまいな質問への回答精度にはやや難ありな印象
考察	実験結果の意味合い	• **総合回答得点率は85%と高スコアとなった。得点に至らなかったケースでも、引用箇所の表記のみの誤りが見られた場合が多く、回答は十分に生成できていた**
	当初期待との差分	• **検証当初の仮説の通り**
	実験の限界点	• **複数章に跨る質問は得点率が低くなりやすい**

B-1. 標準ガイドライン実践ガイドブック (4/4)

6. 結論

実験の成果/影響	• RAG有無にかかわらず、LLMを用いて行政職員レベルの正確性を以て回答代行させることは容易でない結果となった一方で、職員が行う一部の業務(ドキュメントからの要約、個別質問の一般化)等への活用は有効といえる
今後の展望	• FAQのレビューや回答内容の確認ステップなど、全自動運用は難しいため、将来のLLM性能の向上に応じたLLM対応領域の拡大が求められる
推奨や提案	• 職員の問合せや情報収集には正確性が求められるため、FAQの整備やチャットボットとの組み合わせなどを通じて、シナリオ作成の効率化を実現しながら回答正確性を担保することが推奨される

B-2. 人事院規則 (1/4)

1. ユースケース概要

対象業務	• 人事院規則に対する問い合わせの回答案を自動生成する (職員の確認が前提)
前提	• 政府では人事関連の規則類が人事院勧告により年度ごとに更新される。その際、府省庁から自治体に至るまで給与や人事システムなどの担当者が法令に基づいた確認と更新作業を受けており、法令の検索、読み込みに膨大な労力を割いている。
課題	• 難解かつ膨大な文章量で、毎年変更がなされる人事院規則を読み込んで理解し給与計算等を行うことは困難であり、業務に多大な時間がかかることが課題
実現したいこと	• 法令の該当箇所の検索、内容のわかりやすい要約、個別相談内容への対策の提示を行うことで、人事院規則に対する問い合わせ対応の業務効率化を図ること

B-2. 人事院規則 (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	1. 引用箇所が正しいかどうか 2. 回答文が引用箇所を正しく反映しているか
	仮説	似通った文言が複数箇所に分散のため、正解箇所をピンポイントで抽出するのが困難。関連ユーザーからヒアリングの結果、関連すると思われる箇所を複数抽出しその要約を提示するのが現実的であり有効と想定。
検証手法	実験設計	• 実験デザイン – 法令を学習データに組み込む形で、Claude2.1を活用して引用 • 実験手順 – 1.)回答用プロンプトと法令をテキストfileにした学習セットを用意 – 2.)質問を投げ、評価基準から出力された結果を評価
	取得データ	人事院規則九―六(俸給の調整額)(昭和三十二年人事院規則九―六) / 一般職の職員の給与に関する法律 / 一般職の任期付職員の採用及び給与の特例に関する法律 / 一般職の任期付研究員の採用、給与及び勤務時間の特例に関する法律 / 別紙第４ 職員の給与の改定に関する勧告
	KPI	• 担当者による誤謬出現率評価

B-2. 人事院規則 (3/4)

4. 実験結果 / 5. 考察

評価内容	評価手法	• **評価項目** ○ 回答の正しさ1:引用箇所が正しいかどうか ○ 回答の正しさ2:回答文が引用箇所を正しく反映しているか ○ わかりやすさ:表現の仕方が適切であるか
	評価結果	• **定量評価** ○ 引用箇所が正しいかどうか：72% ○ 回答文が引用箇所を正しく反映しているか：84% ○ 表現の仕方が適切であるか：81% • **定性評価** ○ あいまい検索のような質問でも引用を回答しており、実践に近いユースケースにも一定度合い耐えうると評価
考察	実験結果の意味合い	**80%前後の得点率であり、得点に至らなかったケースにおいてもユーザーへの回答のヒントとなる内容を抽出・提示できているため、実用できると考えられる**
	当初期待との差分	• **検証当初の仮説の通り**
	実験の限界点	**学習情報が長大且つ難解な内容であるため、引用箇所の抽出得点率で高スコアを出すことは難しい**

B-2. 人事院規則 (4/4)

6. 結論

実験の成果/影響	• GPT-4やClaude 2.1での正確な引用や抽出は一定精度までとみられるが、LLMの性能次第でより実践的な利用が可能になると推量されるため、今後とも法令などの検索/引用においてLLMは利用検討すべき
今後の展望	• 検証タイミングの都合上Claude2.1での検証が主体であったが、Claude3 Sonnetで試用した結果、極めて高精度での引用や生成が可能になることが分かった。LLMの性能向上が、ユースケースを実用的なレベルまで拡張できる可能性があるものと考える
推奨や提案	• 難解な計算ロジックである点以上に、対象職種や俸給パターンが多種難解である点から、自動化を目指すには入念な異常値テストが必要の為、人の手でのサポートや試験を継続しながら運用することを推奨する

C-1. 用例による法令検索 (1/2)

1. ユースケース概要

対象業務	• 用例に関連する法令の検索
前提	• 政府では府省庁や地方自治体に至るまで、様々な組織単位で法令や条例整備、更新を進めている。その際、関連法令を検索する業務がその起点となる。関連法令を探すときには、その用例(その法令の用いられ方)をベースにした検索がカギとなる。
課題	• 現在e-lawsのウェブサイトに政府による全ての法令がデータベース化されている。しかしながら、その検索は完全一致ベースとなるため、適切な法令検索には一定の知見が求められる。特に、用例をベースにした検索性までは実装されていないのが現状。
提供した機能	• 課題をベースに、本検証において「用例に基づく法令検索」を開発実装し、仮検証できるサービスとして提供するに至った。本サービスは検証環境(GaiXer)のUI上の一機能として実装されつつ、個別インターフェイスとアーキテクチャで提供する事で、ユーザー利便性が一定確保しやすいものとして提供。(次頁参照)

C-1. 用例による法令検索 (2/2)

2. 結論

期待される効果	・用例に基づく法令検索を行うことにより、作業時間の単純削減が期待できるに加え、属人の検索ノウハウが一定不要となる事から、特に人手不足となりうる地方自治体等において、行政の運用持続性への貢献が期待できる。
今後の展望	・ユーザーヒアリングに基づいて作成した、本機能の本格実用するにあたっての実装案(次頁参照)を実現する。

C-2. 用語集作成業務の効率化 (1/4)

1. ユースケース概要

対象業務	• 公開前の文書に対する用語集の作成
前提	• 既存の用語集を正義として、その拡張性に可能性があるかに主眼を置いて検証した
課題	• 公開前の文書に対する用語集を作成する担当者の業務を効率化する為に、新たに当該文書の用語集に追加するべきだと思われる用語を指摘し、その定義を生成することで、用語集の完成までのリードタイムと工数削減を狙うユースケースとして検証した
実現したいこと	• 既存の用語集と同等もしくはそれ以上の高品質な用語集を短時間で作成し、業務効率化を図ること

C-2. 用語集作成業務の効率化 (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	1. 既存の用語集から削除した用語の生成を正確に行えるか 2. 既存の用語集にある用語と重複せず用語の追加が行えるか
	仮説	1. 抽出と定義説明はLLMの得意分野であるため、高精度でドキュメントからIT関連の用語抽出可能な見込み 2. 生成AIによる用語の追加は可能
検証手法	実験設計	• **実験デザイン** – 既存の用語集とGaiXerで生成された用語集との差分を検証 • **実験手順** – 1.)既存の用語集から、特定の用語を削除する – 2.) 1で用意した削除済み用語集を、プロンプト内に入れる – 3.) 「新規の用語群」と「その用語群の定義説明」を生成 – 4.) GPT-4 TurboとClaude3 Sonnetの2つで、かつそれぞれに特化したプロンプトで比較
	取得データ	重点計画用語集
	KPI	重点計画用語集との合致度

C-2. 用語集作成業務の効率化 (3/4)

4. 実験結果 / 5. 考察

評価内容	評価手法	• **評価項目** ○ 回答の正しさ1:重複のない抽出 ○ 表現の適切さ1:適切な定義生成
	評価結果	• **定量評価(検証項目：得点率)** ○ 重複のない抽出： 削除した用語(17%)、追加した用語(81%) ○ 適切な抽出：100% • **定性評価** ○ 元々の用語集のコンセプトや用途などでディレクションをすれば生成確率を上げることができることが判明
考察	実験結果の意味合い	• **既存の用語集にない新規用語を複数生成できたことと、用語の説明は適切に生成することができたため、高品質な用語集生成が可能であると思われる。**
	当初期待との差分	• **削除した項目を的確に抽出することは十分に出来なかった。一方で、新規用語の提案は十分に有効活用できる。**
	実験の限界点	• **削除した用語をすべて的中して生成することはできなかった**

C-2. 用語集作成業務の効率化 (4/4)

6. 結論

実験の成果/影響	• 削除した用語を的中して生成することは難易度が高かったが、「既存用語集と重複のない抽出」や「網羅的かつトンマナに沿った説明」は高品質のものが生成できた
今後の展望	• 用語集の作成時間と労力を大幅に削減し、職員がより重要な業務に集中できるようになる • 大幅な業務削減効果のみならず、生成AIによる新規用語提案等のサポートによって、業務改革に新しいアプローチを提供できることまで期待される
推奨や提案	• 事前に要件を明確に設定すると用語選定の精度が高まるため、 用語集の目的やコンセプトなどの補足情報をプロンプトに組み込むことを推奨する

D-1. 質問主意書に対する答弁ドラフト (1/4)

1. ユースケース概要

対象業務	府省庁および自治体の職員が、議員からの質問主意書に対して答弁の案を作成する。
課題	答弁案の作成に行政職員の多大な時間が割かれており、効率化の余地が大きい。また、新人等が過去の膨大な資料・答弁を踏まえて、新たに答弁案を作成する負担が大きく、成果物の品質向上余地も存在すると考える。
実現したいこと	生成AIのドラフトを元に、府省庁および自治体の職員が答弁案を最終化することにより、業務の効率化および成果物の品質向上を実現する

D-1. 質問主意書に対する答弁ドラフト (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	**政府や自治体職員が、議会質問に対して答弁ドラフトを行う際、生成AIはどこまで有用か？** • 答弁ドラフトに適したLLMおよび設定、プロンプトはどのようなものか？ • RAGを行う際の学習データを、どのように構成すべきか？
	仮説	**質問が事実確認の場合には対応できるが、それ以上を求める場合は学習セットの拡充が必要** • 事実についての質問には一定正答率を示すが、将来についての質問への回答 • 過去の国会答弁以外に、府省庁Webサイトや政府見解を学習セットに追加することが必要
検証手法	実験設計	**実験デザイン** • 第148回(2000年)-211回(2023年)国会の「質問情報」「答弁情報」を学習セット化 • 上記に含まれない、212回以降の国会での質問を用いて、回答の妥当性を検証 **実験手順** • 1. 衆議院Webサイトの第213回国会 質問の一覧から、番号順質問を10個抽出する • 2. **実験1** 質問の「質問情報」を所定のプロンプトテンプレートに入力する • 3. 生成された結果と実際の「答弁情報」を比較し、類似度を評価する • 4. **実験2** 質問に加え、ヒントとして答弁情報で使われてる「キーワード」を所定のプロンプトテンプレートに入力する • 5. 生成された結果と実際の「答弁情報」を比較し、類似度を評価、3.から改善が見られたかを確認する ※ 今回は初期的な検証・実験デザインの妥当性検証が目的のため、質問は10問に留めている
	取得データKPI	**実験1 KPI: 質問情報のみを与えた際の、答弁情報(正解データ)との合致度** **実験2 KPI: 答弁で使われているキーワードをヒントとして与えた際の、答弁情報(正解データ)との合致度**

D-1. 質問主意書に対する答弁ドラフト (3/4)

4. 実験結果 / 5. 考察

検証内容

評価手法

KPI: 正解データとの合致度

- 生成結果を実際の「答弁情報」と比較し、以下の3段階でランク付けする
  - ○: 答弁と大枠で一致
  - △: 答弁ドラフトとは異なるが、内容は答弁として妥当
  - ×: 内容が答弁として不適切

評価結果

右表の通り

- 縦軸＝実験1
- 横軸＝実験2

※ 実験1→実験2で回答精度向上が見られたのが赤セル

	○	△	×	総計
○	1			1
△	2	2		4
×	4		1	5
総計	7	2	1	10

考察

過去の答弁データからの生成精度については限定的な結果に留まった

- 事実・データを問う質問に対して、正しく回答を生成することは難しい
- 政府として把握していないものを、把握していないと答えるのは難しい

追加データとして、回答作成の担当者が考えているキーワードを与えることで、回答精度の向上が望める

- 事実となるファクト
- 理由のキーワード
- 根拠となる法令

D-1. 質問主意書に対する答弁ドラフト (4/4)

6. 結論

実験の成果/影響

今回は実験デザインのための小規模な検証であり、データそのものの意味は限定的であるが、どのような質問には回答精度が高くなるのか、答弁の核となるキーワードにはどのような種類があるのかといった点について、一定の理解を得ることができた。

- LLMや過去答弁に含まれない最新の事実・データを問われていなければ、形式的な生成は可能
- キーワードとして、具体的な事実・データ、回答根拠となる法令を与えると精度が高まる可能性

今後の展望

今後、より回答精度を高めるための方向性として、下記の2つが考えられる。

- 担当者の文章作成支援用ツールとして、あくまで答弁キーワード(回答の核となる概念)は職員が与える前提とする
- 答弁キーワードを抽出できるよう、学習データの範囲を拡大する(例: 各省庁のWebサイト、首相・大臣や政府関係者の会見 等)
  - **上記を効率よく行えるよう、政府関連データの様式統一、データベース化を検討**する

推奨や提案

令和6年度、実験にご協力いただける方にご参画いただき、検討を継続することをご提案したい。

- 担当者を決め、学習セットにドメイン知識を充実させた上で、答弁生成のドラフトを実践する
- 担当者が着想した「キーワード」を与えて生成させ、それに業務支援効果があるかを確認する
- 担当者からのフィードバックを受けて、**答弁の類型化をプロンプトに盛り込む**
  - **どのようなタイプの質問かを判別し、盛り込むべきキーワードをAI自ら探索**できるようにする

D-1. 質問主意書に対する答弁ドラフト 実験結果のまとめ

集計結果

質問ID	実験1評価	実験2評価	評価
1	×	○	向上
2	×	○	向上
3	△	△	変化なし
4	△	○	向上
5	△	○	向上
6	×	○	向上
7	×	×	変化なし
8	○	○	変化なし
9	△	△	変化なし
10	×	○	向上

	○	△	×	総計
○	1			1
△	2	2		4
×	4		1	5
総計	7	2	1	10

考察

過去の答弁データからの生成精度については限定的な結果に留まった

- 事実・データを問う質問に対して、正しく回答を生成することは難しい
  - 例: No.1 中規模・小規模の飲食店の倒産件数等の経営状況
  - 例: No.7 出入力管理施設の常勤医師の在籍状況
- 政府として把握していないものを、把握していないと答えるのは難しい
  - 例: No.6 上陸特別許可の件数

追加データとして、回答作成の担当者が考えているキーワードを与えることで、回答精度の向上が望める

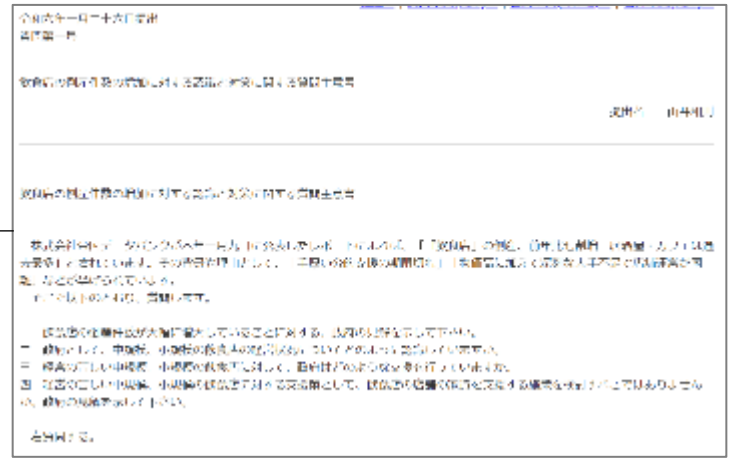
- 事実となるファクト
  - 例: No.2 システム改修の完了見込み時期
- 理由のキーワード
  - 例: No.4 買春に罰則が設けられていない理由
- 根拠となる法令
  - 例: No.10 過去の農地対策の根拠となる法令

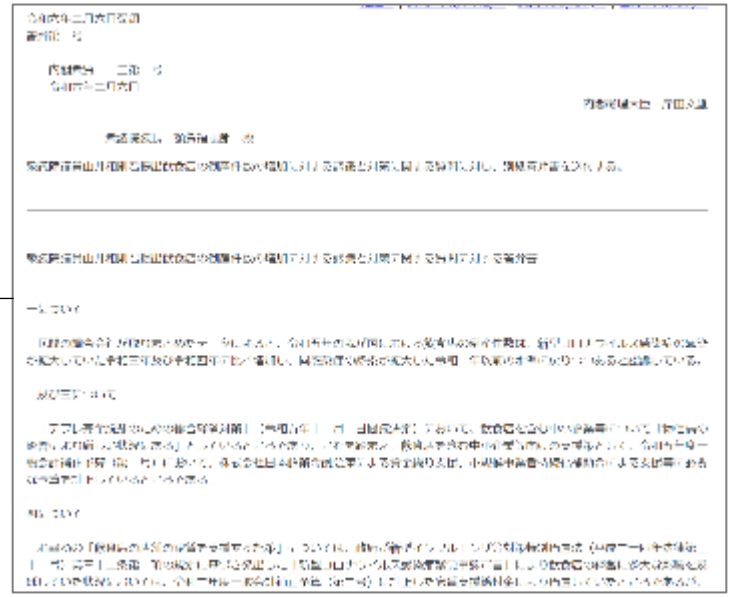
今後の方向性としては、2つが考えられる

- 担当者の文章作成支援用ツールとして、あくまで答弁キーワード(回答の核となる概念)は職員が与えるものとする
- 答弁キーワードを抽出できるよう、学習データの範囲を拡大する
  - 例: 各省庁のWebサイト、首相・大臣や政府関係者の会見 等

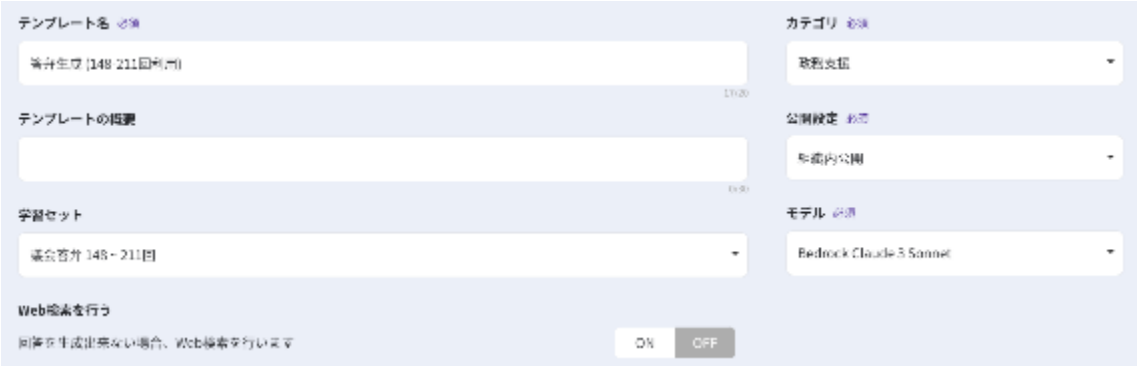
(参考) 学習データ

衆議院
The House of Representatives, Japan
本会議・委員会等
立法情報
議員情報
国会関係資料
各種手続
English
第213回国会　質問の一覧





(参考) プロンプトテンプレート



※ 上記のプロンプトテンプレートを実施すると、
GaiXerが、学習セットのベクトル検索を行い、
RAGが行われる。具体的には

<context>
Use the following information in your answer
Context [検索結果の抽出部分]
</context>

という形でContextが指定され、生成される。

出所: 福井県「生成AI業務活用タスクフォース実証結果報告」を参考にプロンプトを改変

E-1. ソースコードの言語変換 (1/4)

1. ユースケース概要

対象業務	• 対象プログラムのソースコードを異なる言語に変換する
前提	• ソースコードの言語変換に生成AIを活用できる可能性を探るため、 簡易的なソースコードの変化が可能かを確認することを主眼を置いて検証した
課題	• 言語の違いによる、構文やライブラリ、コードスタイルの違いに生成AIがどれだけ対応できるかを検証する必要がある
実現したいこと	• 対象プログラムのソースコードを異なる言語に変換すること

E-1. ソースコードの言語変換 (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	1. 生成AIによってソースコードから処理概要・設計書を作成することは可能か 2. 生成AIによってソースコードの可読性を向上させることは可能か 3. 生成AIによってソースコードの言語変換を行うことは可能か
	仮説	1. 生成AIによってソースコードから処理概要・設計書を作成することは可能 2. 生成AIによってソースコードの可読性を向上させることは可能 3. 生成AIによってソースコードの言語変換を行うことは可能
検証手法	実験設計	• **実験デザイン** – 500行程度の短いプログラムを生成AIを用いて、言語変換可能かを確認する • **実験手順** – 1.)対象プログラムをGaiXerで生成 – 2.)対象プログラムの処理概要とプログラム設計書を生成 – 3.)コード変換を行う
	取得データ	なし
	KPI	プログラムを別の言語に変換できること

E-1. ソースコードの言語変換 (3/4)

4. 実験結果 / 5. 考察

検証内容	評価手法	• **評価項目** ○ ソースコードから処理概要・設計書を作成することができているか ○ ソースコードの可読性を向上させることができているか ○ ソースコードの言語変換を行うことができているか
	評価結果	• **定性評価** ○ ソースコードから処理概要・設計書を作成することは可能 ○ ソースコードの可読性を向上させることは可能 ○ ソースコードの言語変換を行うことは可能
考察	実験結果の意味合い	• **ソースコードからの処理概要・設計書の作成、ソースコードの可読性向上、ソースコード変換に対して、生成AIの活用が見込める**
	当初期待との差分	• **検証当初の仮説の通り**
	実験の限界点	• **本検証では500行という短いプログラムでの言語変換を実施したため、長いプログラムにおいても変換が可能か更なる検証が必要**

E-1. ソースコードの言語変換 (4/4)

6. 結論

実験の成果/影響	• 本検証では500行という短いプログラムでの検証ではあったものの、Claude 3 Sonnet にてソースコードの言語変換を行うことは可能であることが判明した
今後の展望	• 言語変換を行った際に元のプログラムと同じ処理をコーディングできているかを確認するステップが必要であり、そのステップにも生成AIの活用余地があると考えられる
推奨や提案	• ソースコードの言語変換への生成AIは活用の見込みがあり、本格的な検証の実施を推奨する ※ ソースコードの言語変換においては、NVIDIAが600以上の言語を学習させ、コード生成に特化したオープンなLLM「StarCoder2」等が昨今発表されたが、今回については一般的に使用される言語モデルの可能性を検討するため、Anthropicの「Claude 3 Sonnet」を使用して実験を行った。

F-1. ドキュメントの機密度判定 (1/4)

1. 概要

対象業務	• 行政業務の効率化を図るため、生成AIに機密性を含む情報を入力する業務全般
前提	• 行政で取り扱う情報は機密性に基づき分類がなされており、2以上にあたる情報は、その取扱いに当たり注意が必要。 • 誤った運用や漏えいは、国民の権利侵害や行政業務への支障となり、国民への信頼を大きく損なうリスクとなりかねない
課題	• 行政業務で生成AIを活用するうえで、機密性の高い情報を取り扱う場合のリスク管理体制を整える必要がある
実現したいこと	• GaiXerを利用して、「政府機関の情報セキュリティ対策のための統一基準(第 4 版)」に記載されている機密性の格付けの区分に基づき、文書の機密度判定を自動化すること

F-1. ドキュメントの機密度判定 (2/4)

2. 検証方法 / 3. 検証手法

検証内容	論点	• 機密度判定の根拠となる「個人情報の有無(特定の個人を特定できるものとできないもの)」「行政業務への影響」を検知できるか
	仮説	• 個人情報の定義が複雑であるため、100%の正答率を出すことは困難であると想定
検証手法	実験設計	• **実験デザイン** – 人が判定した機密度と生成AIで生成した機密度を比較する • **実験手順** – 1.)命令、出力形式、前提知識の3つの命令ブロックから構成されるプロンプトを作成 – 2) 22通りの網羅的なテストケースを作成して検証 – 3.)機密性2情報と個人情報の定義をプロンプトに追加 – 4.)50通りの網羅的なテストケースを作成して検証
	取得データ	• 行政機関の保有する情報の公開に関する法律 / 個人情報の保護に関する法律
	KPI	• 個人情報有無と行政の業務への影響に基づいてあらかじめ定めた機密性分類との合致度

F-1. ドキュメントの機密度判定 (3/4)

4. 実験結果 / 5. 考察

評価内容	評価手法	• **評価項目** ○ 回答の正しさ1:個人情報判定、業務に支障があるかの判定が適切か ○ 回答の正しさ2:機密性情報分類が適切か ○ 回答の正しさ3:機密性情報分類した根拠が適切か
	評価結果	• **定量評価(検証項目：得点率)** ○ 個人情報判定、業務に支障があるかの判定が適切か:90% ○ 機密性情報分類が適切か:98% ○ 機密性情報分類した根拠が適切か:80%
考察	実験結果の意味合い	• **機密性情報入力の未然防止での利用は、100%の精度を求められるため実用は難しい。**
	当初期待との差分	• **情報分類の最終正答率が98%と想定以上の精度を出すことができた**
	実験の限界点	• **機密性分類２判定は100%の精度での回答は困難**

F-1. ドキュメントの機密度判定 (4/4)

6. 結論

実験の成果/影響	• 機密性分類２判定は100%精度での回答は困難。よって防止策的な利用は非推奨。
今後の展望	• 事後的にログなどを監査する目的でのバッチ処理は有効と考えられる
推奨や提案	• FAQのレビューや回答内容の確認ステップなど、全自動運用は難しいため、将来のLLM性能の向上に応じたLLM対応領域の拡大を推奨する

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

利用者サーベイ概要

目的

- 技術検証基盤の利用者に対し、利用頻度・利用目的、
  および業務効率化・成果物の品質向上の寄与可能性について調査する

対象

- 調査1: 技術検証基盤のユーザー登録者 (約750人)
- 調査2: デジタル改革共創プラットフォームの生成AIチャンネル参加者 (約130人)

期間

- 3月22日～3月28日 12時

回収状況

- 調査1: 技術検証基盤のユーザー登録者
  – 回答者: 256人 (回収率: 約34%)
- 調査2: デジタル改革共創プラットフォームの生成AIチャンネル参加者
  – 回答者: 18人 (回収率: 14%)

※ サーベイ設計に際しては、利用者へのインタビューを実施し、設問の妥当性について検証を実施

利用者サーベイ集計結果のアウトライン

調査1「技術検証基盤のユーザー登録者向け調査」

(1) 基本属性

(2) 全体利用状況と評価

(3) 機能・LLM別利用状況と評価

(3) 業務効率化の可能性

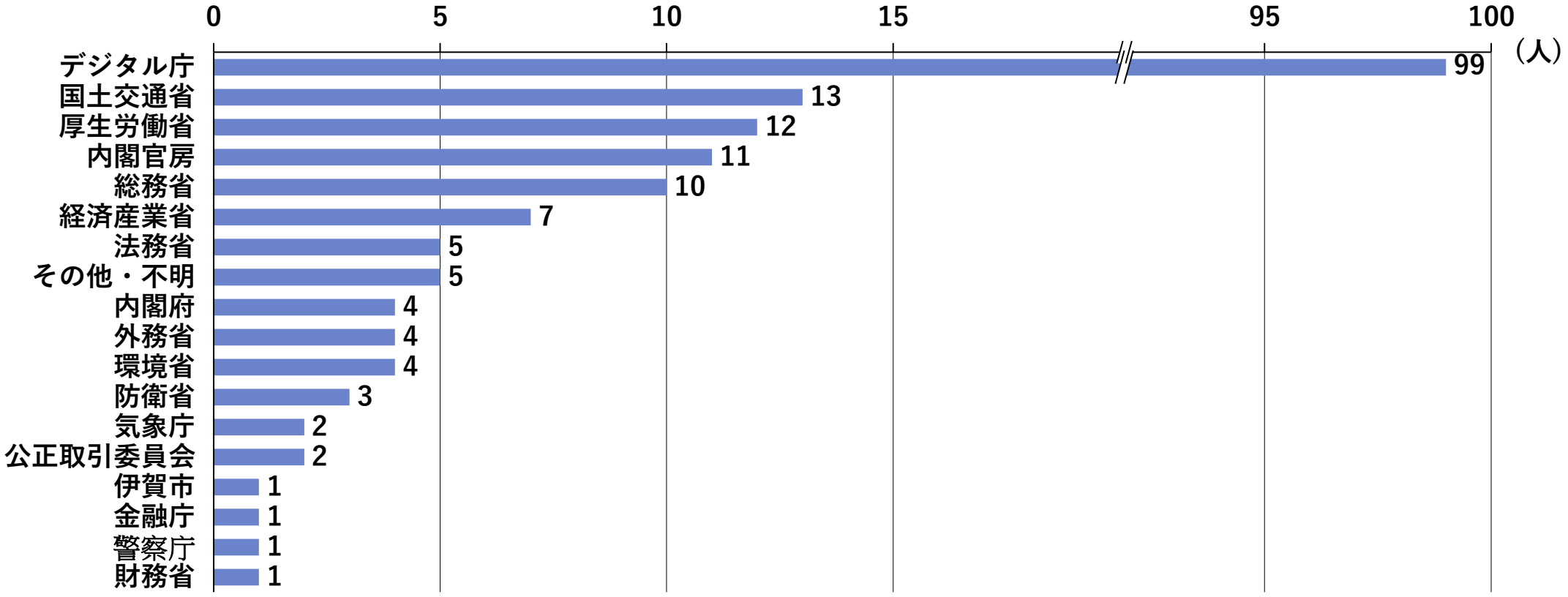
(4) 成果物品質向上の可能性

(5) その他

※ 調査2「デジタル改革共創プラットフォームの生成AIチャンネル参加者」については
回答数が限定的であり統計処理が難しいため、今後の施策に定性的に反映する取り扱いとする。

利用者の所属組織 [中央省庁 (n=185)]

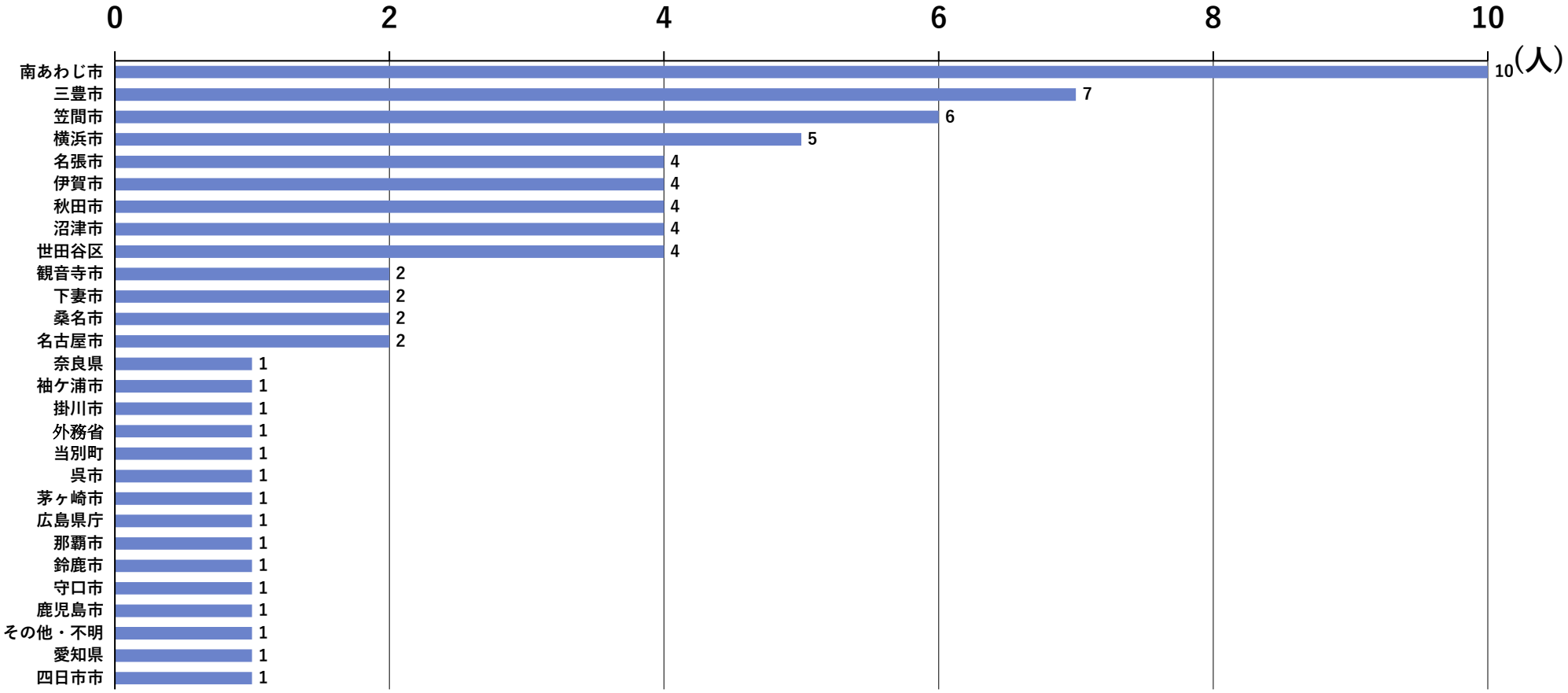
Q.あなたの所属組織を教えて下さい。 所属がデジタル庁以外の場合は、府省庁名または自治体名を教えて下さい



出所: ユーザー調査1 術検証基盤のユーザー登録者 (n=256)

利用者の所属組織 [地方公共団体 (n=71)]

Q.あなたの所属組織を教えて下さい。 所属がデジタル庁以外の場合は、府省庁名または自治体名を教えて下さい



出所: ユーザー調査1 術検証基盤のユーザー登録者 (n=256)

利用者の役職/勤続年数

Q. あなたの所属組織における役職について、次の中から最も近いものを選んで下さい (※ その他FAあり)

	中央省庁	地方公共団体	総計
部長級	5	3	8
室長級	1		1
参与	1		1
課長級	15	10	25
課長補佐	2	1	3
課長補佐級	6		6
補佐・専門官級	1		1
補佐級	2		2
主任・係長級	61	29	90
一般職員	37	27	64
任期付き職員	1		1
期間業務職員	1		1
非常勤職員	51		51
CIO補佐監		1	1
デジタル統括アドバイザー	1		1
総計	185	71	256

以降の質問の集計では“課長補佐級以上”

以降の質問の集計では“その他”

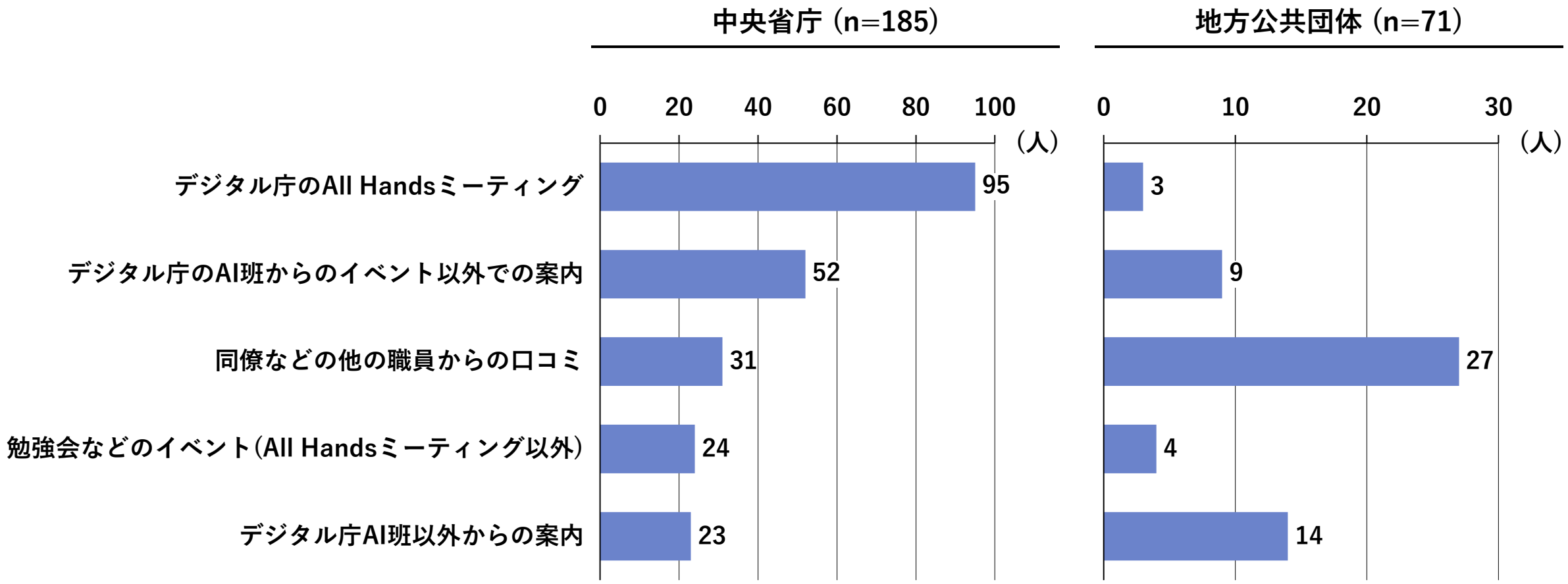
Q. あなたの公的機関でのキャリアの長さを選んで下さい (公的機関からの出向等は、その期間も含みます)

	中央省庁	地方公共団体	総計
半年未満	13	1	14
半年以上1年未満	36	2	38
1年以上3年未満	38	6	44
3年以上5年未満	12	5	17
5年以上10年未満	32	15	47
10年以上	54	42	96
総計	185	71	256

出所: ユーザー調査1 術検証基盤のユーザー登録者 (n=256)

技術検証基盤を知ったきっかけ

Q. あなたがGaiXerを知ったきっかけを、次の中から当てはまるものをすべて選んで下さい。

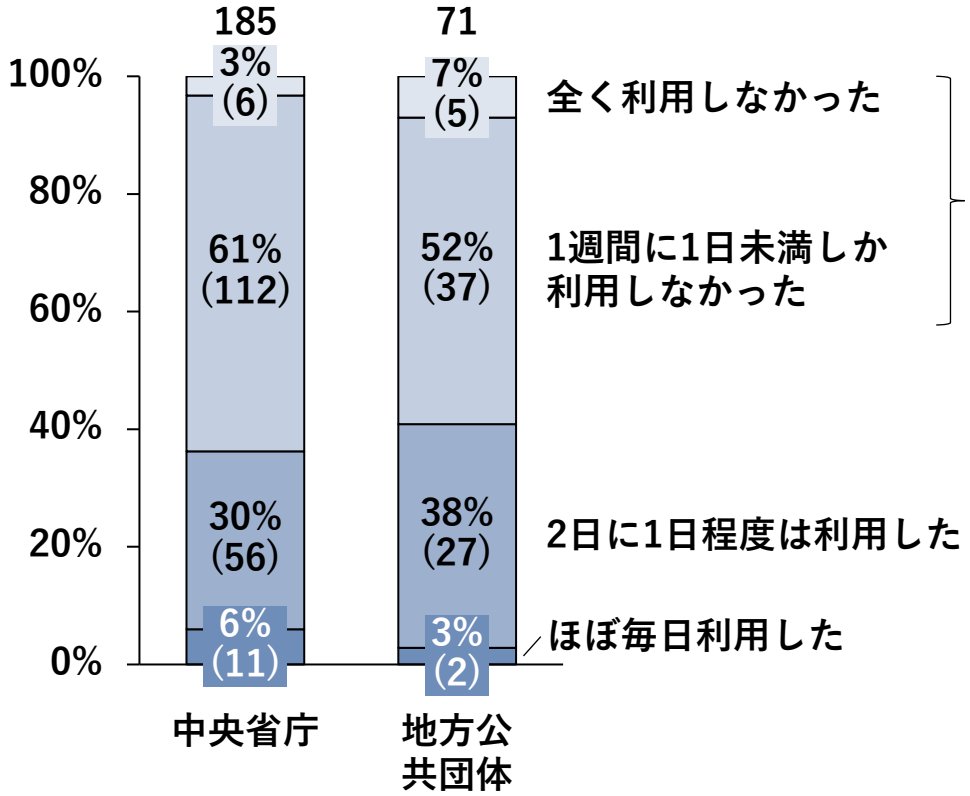


出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

技術検証基盤の利用頻度/低頻度の理由 [組織別]

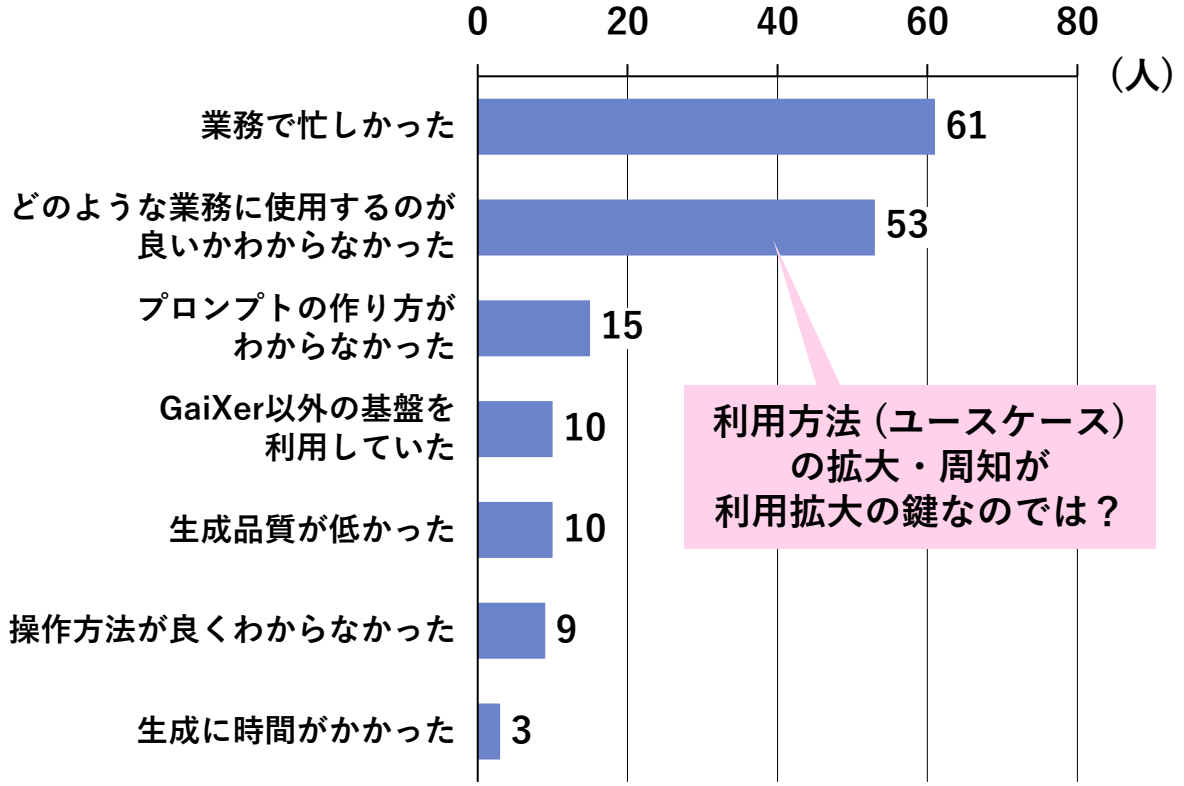
技術検証基盤の利用頻度

Q. あなたが業務で生成AIの技術検証基盤(GaiXer)を利用した日は営業日のうちどのくらいの割合ですか、最も近いものを選んで下さい。



低頻度の理由

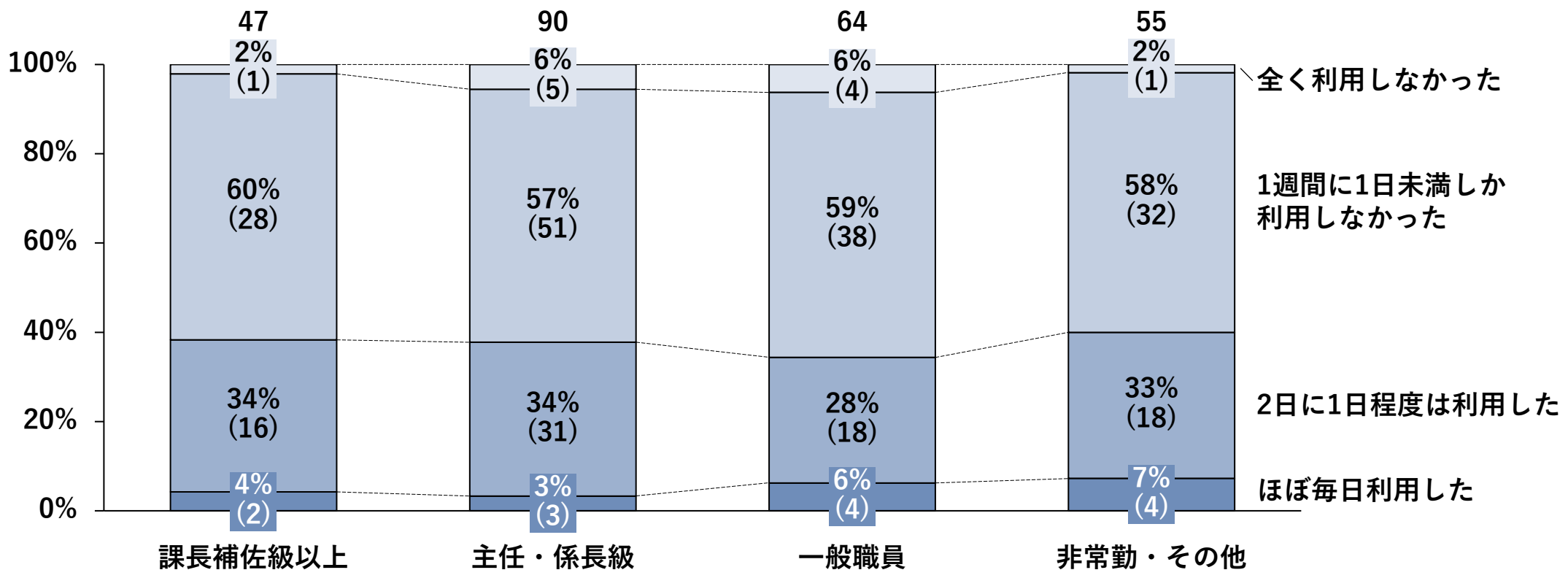
Q. 前問で「1週間に1日未満しか利用しなかった」「全く利用しなかった」と回答した方にお伺いします。その理由として当てはまるものを次のうちから全て選んで下さい。



出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

技術検証基盤の利用頻度 [役職別]

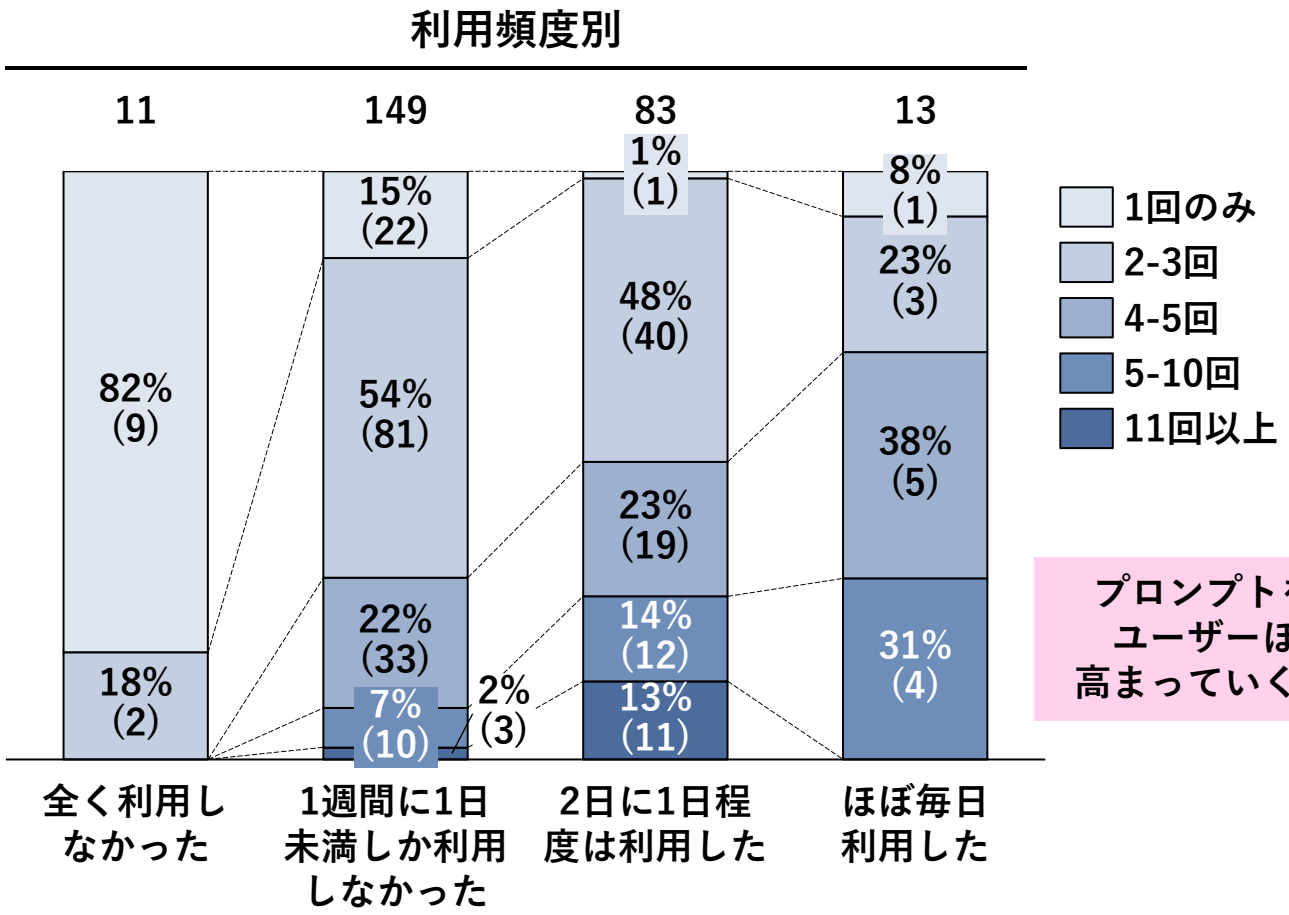
Q. あなたが業務で生成AIの技術検証基盤(GaiXer)を利用した日は営業日のうちどのくらいの割合ですか、最も近いものを選んで下さい。



出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

1日あたりの平均リクエスト回数

Q. あなたが生成AIの技術検証基盤(GaiXer)を利用したとき、1日あたり平均何回のリクエストを送信しましたか。
意図した回答を得られず質問を複数回繰り返した場合は、その全てをカウントして下さい。
(例: 1つの回答を得るまでに修正や対話のために、3回リクエストを送信した場合は「3回」とカウント)

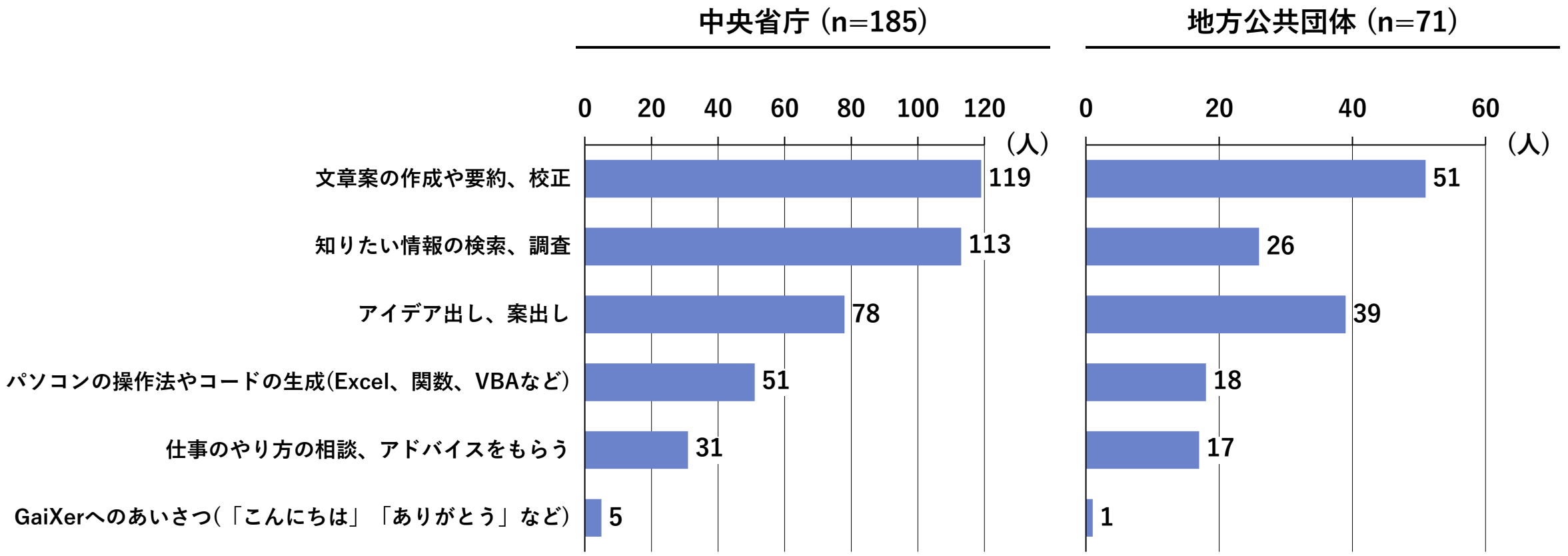


プロンプトを試行錯誤するユーザーほど利用頻度が高まっていく傾向がありそう

出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

技術検証基盤を利用した業務 [組織別]

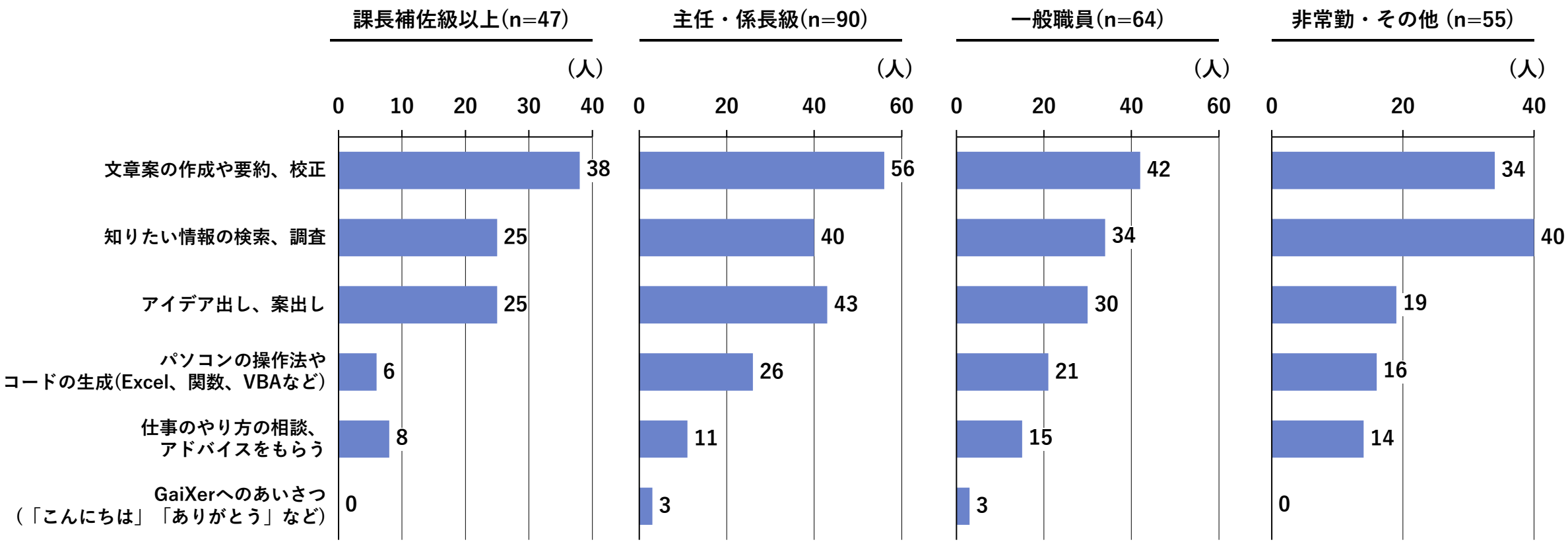
Q. あなたが生成AIの技術検証基盤(GaiXer)を利用した業務はどのようなものですか。当てはまるものをすべて選んで下さい。



出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

技術検証基盤を利用した業務 [役職別]

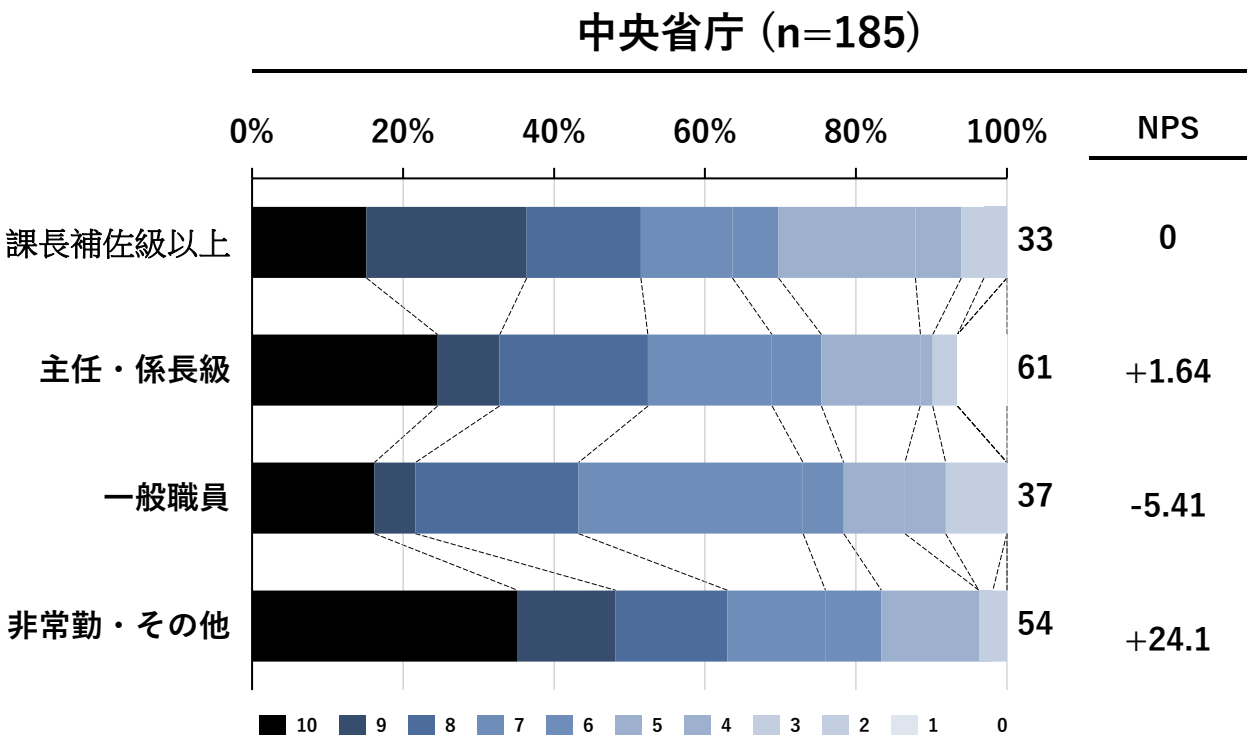
Q. あなたが生成AIの技術検証基盤(GaiXer)を利用した業務はどのようなものですか。当てはまるものをすべて選んで下さい。

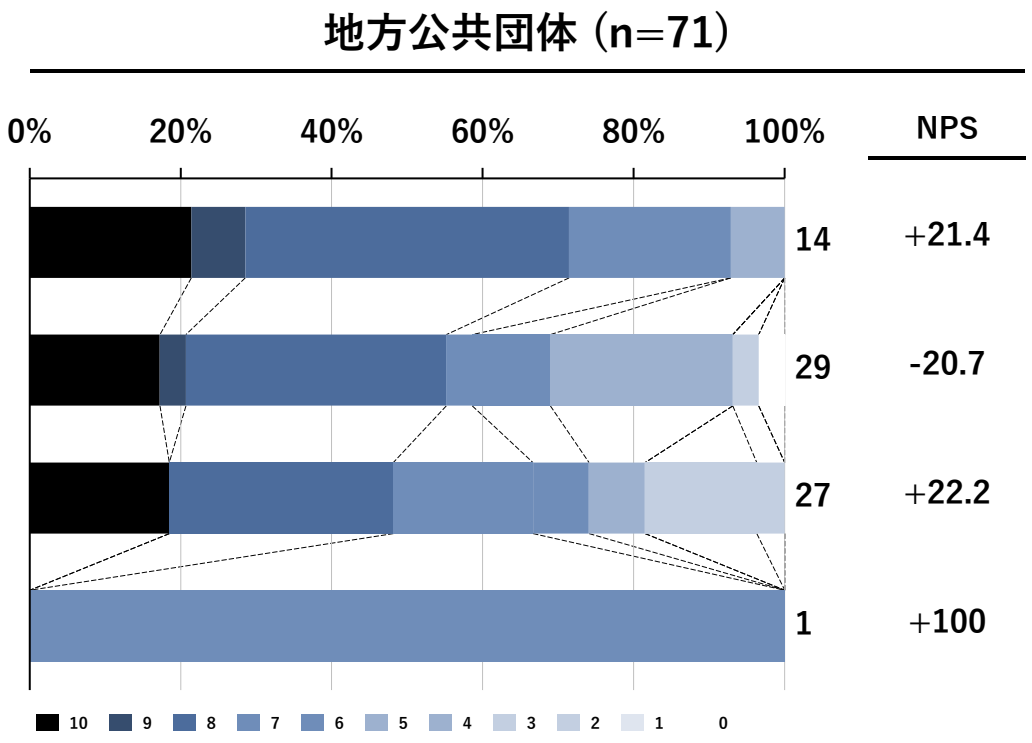


出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

技術検証基盤全体NPS [組織別/役職別]

Q. あなたは、生成AIの技術検証基盤(GaiXer)と同等のものを業務において利用してみることを同僚・友人などに勧めてみたいと思いますか、最も近いものを選んで下さい。

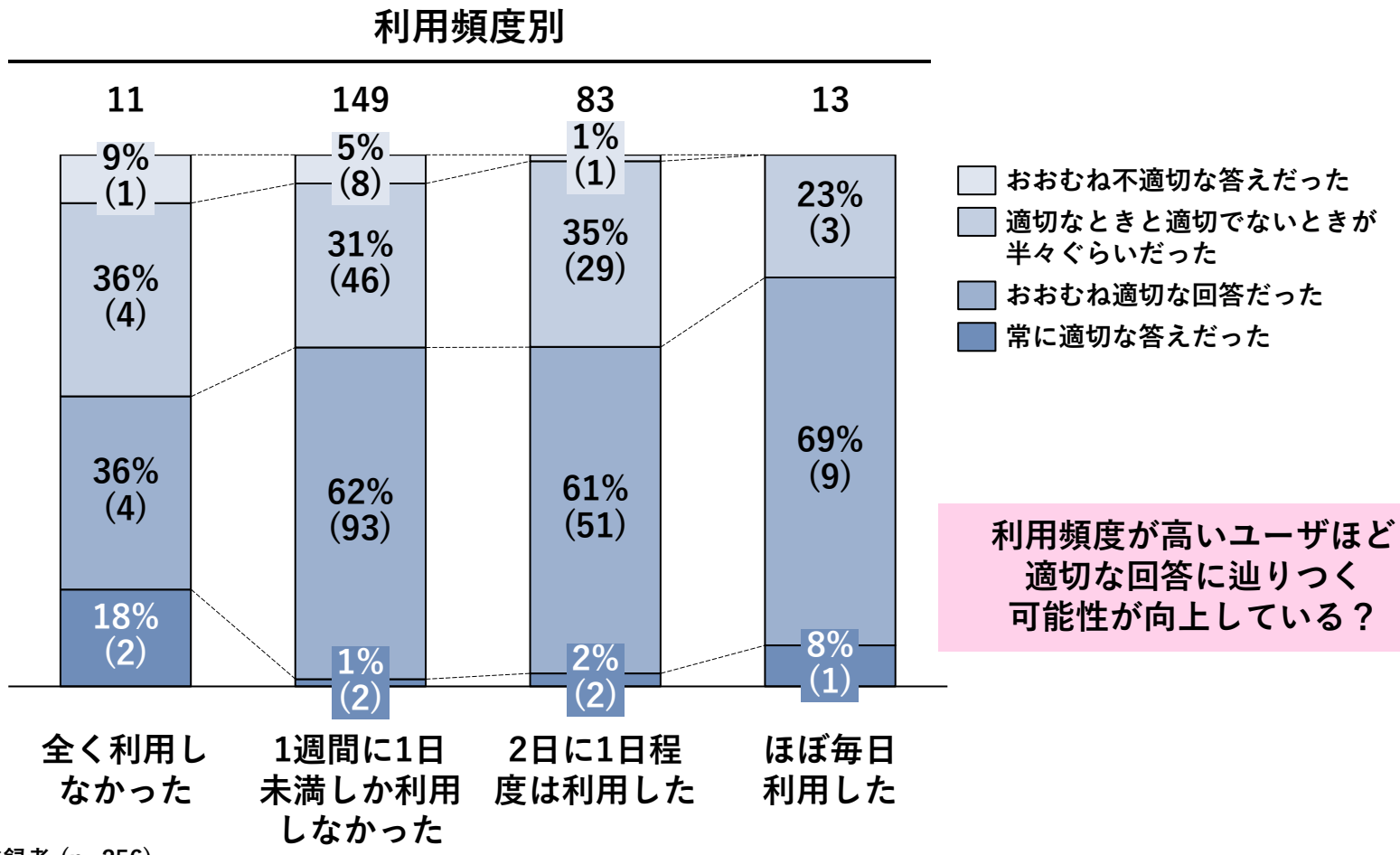




出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

回答の適切さ

Q. あなたが生成AIの技術検証基盤(GaiXer)を利用したとき、回答はどの程度適切だったと感じましたか。



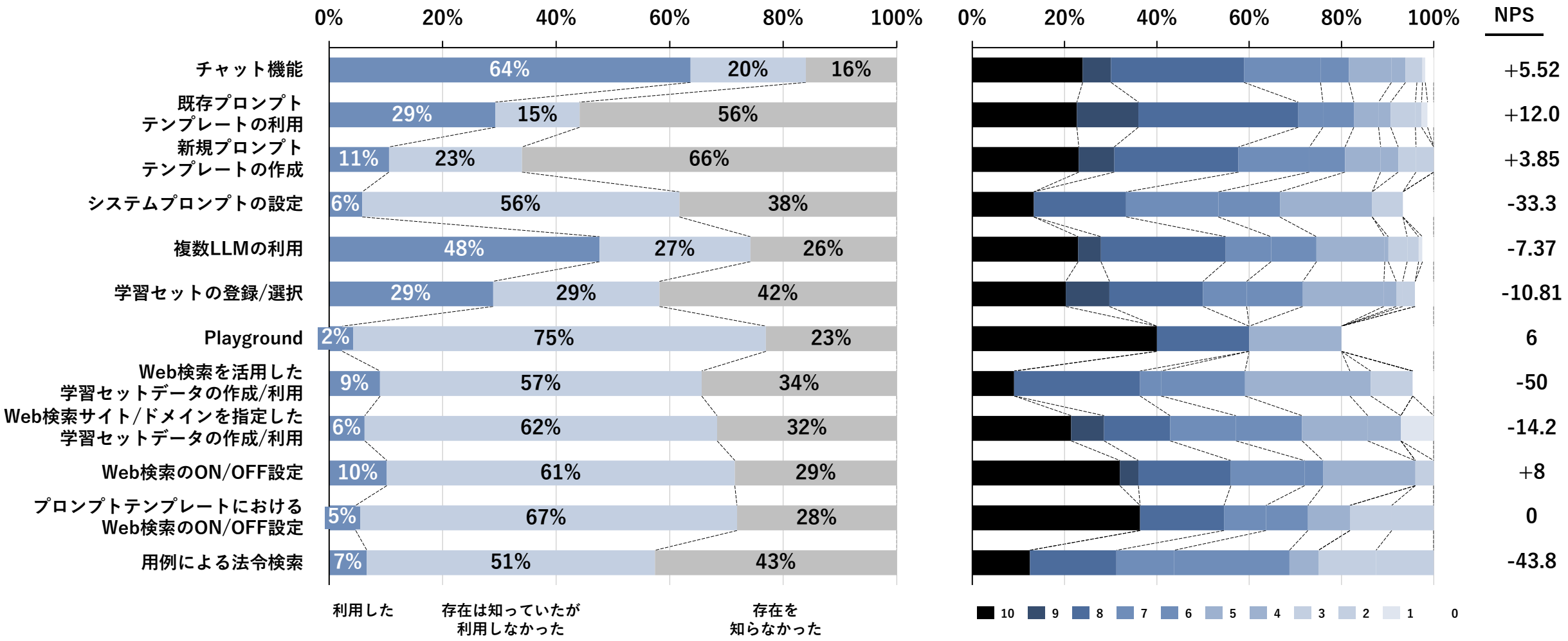
利用頻度が高いユーザほど
適切な回答に辿りつく
可能性が向上している？

出所: ユーザー調査1: 技術検証基盤のユーザー登録者（n=256）

機能別利用状況・NPS(Net Promoter Score)

Q. 生成AIの技術検証基盤(GaiXer)が有する「○○機能」の
あなたの利用状況について最も近いものを下記から選んで下さい。

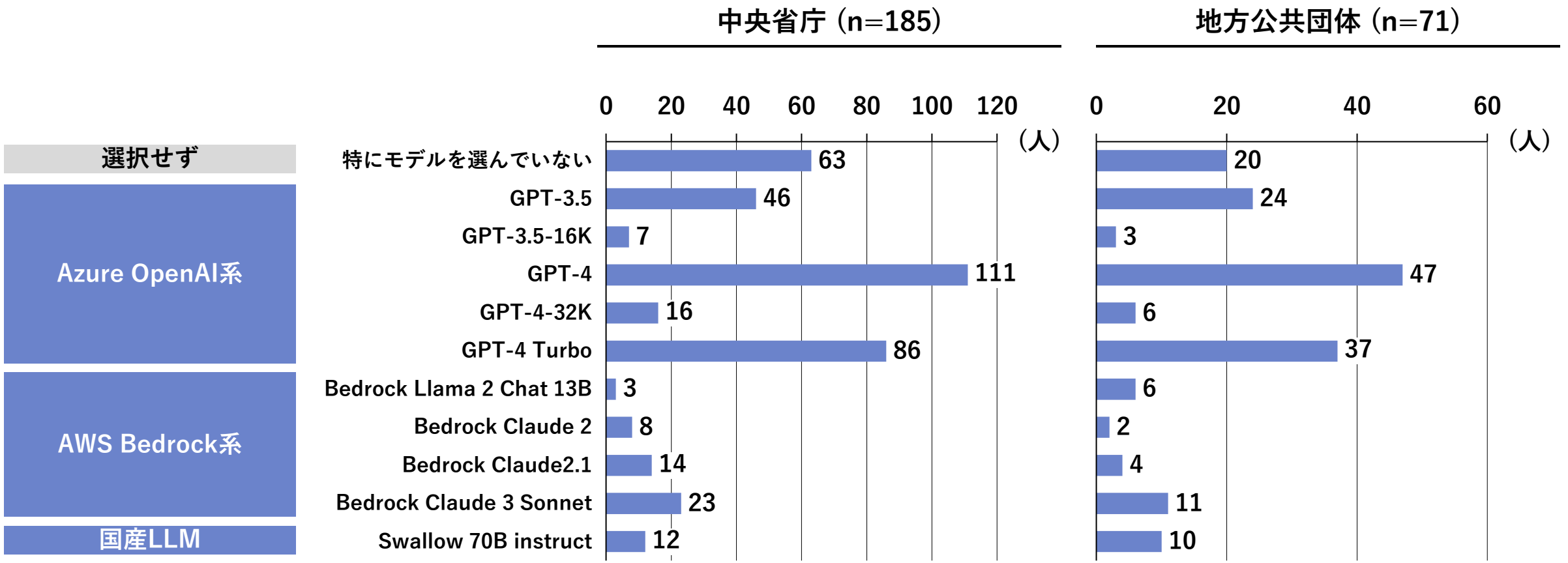
Q. あなたが利用した「○○機能」を同僚・友人などに
勧めてみたいと思いますか、最も近いものを選んで下さい。



出所: ユーザー調査1: 技術検証基盤のユーザー登録者（n=256）

LLM別利用状況

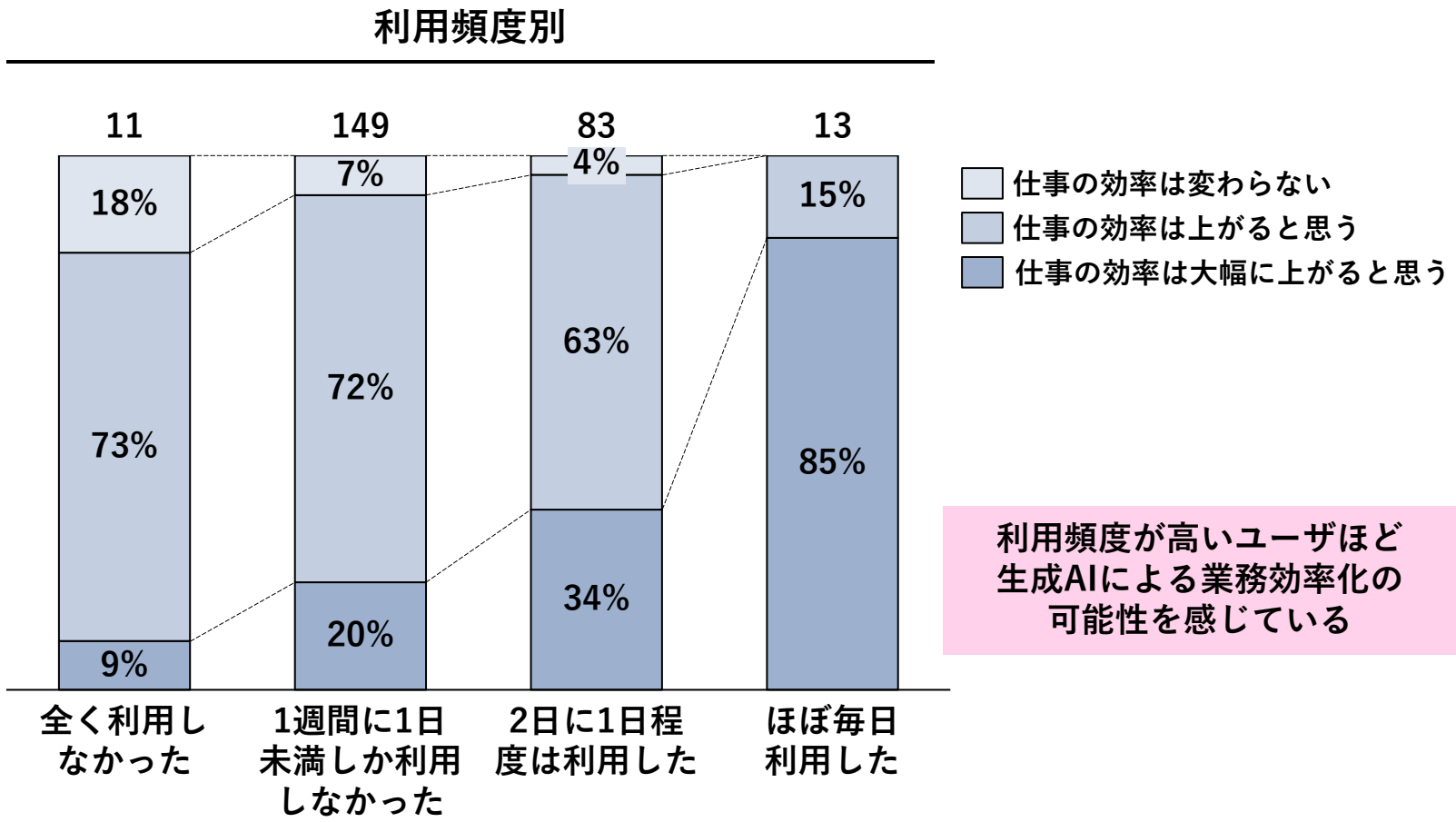
あなたは生成AIの技術検証基盤(GaiXer)においてどの大規模言語モデル(LLM)を利用しましたか。当てはまるものをすべて選んで下さい



出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

業務効率化に対する期待

Q. 生成AIの活用により、あなたが日常的に行っている業務の効率は向上すると思いますか。次の中から最も近いものを選んで下さい。

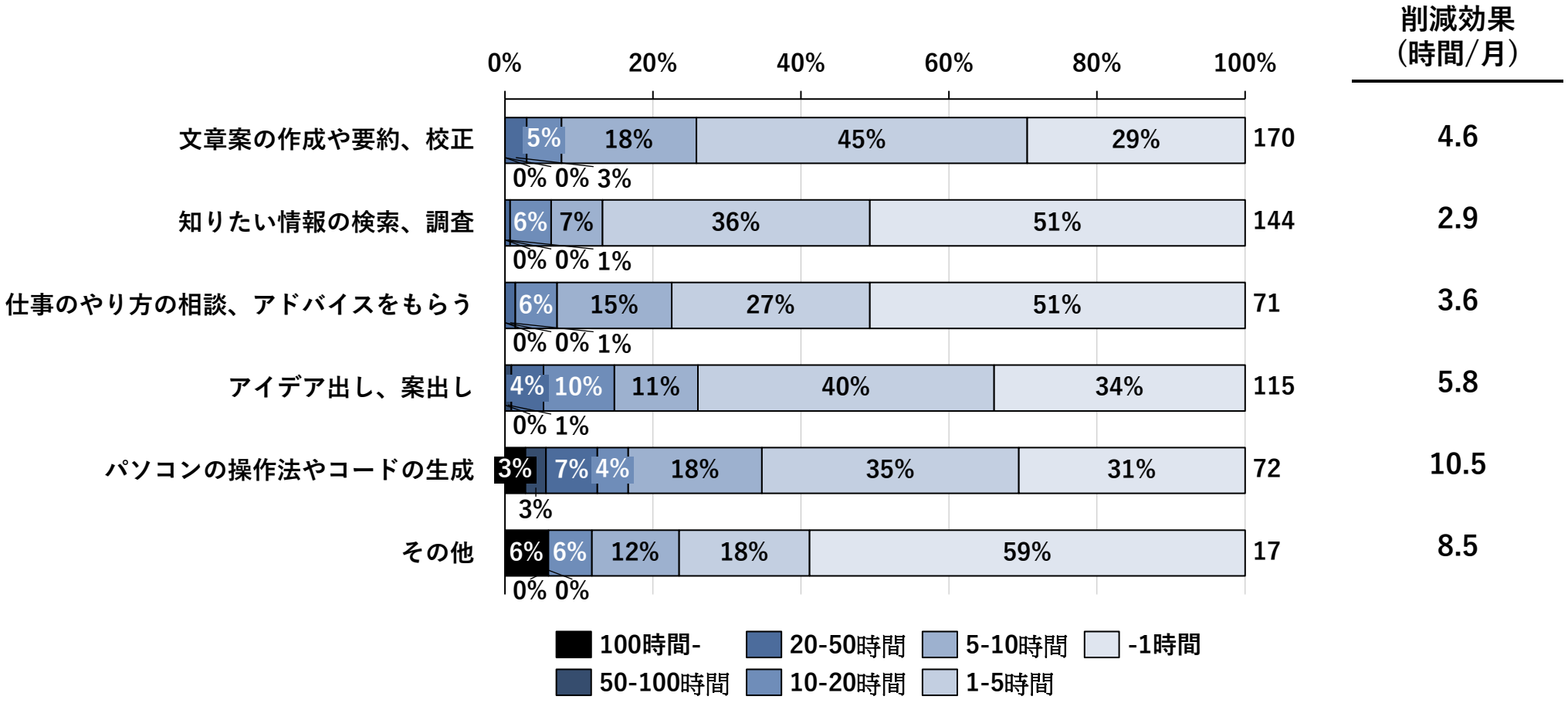


利用頻度が高いユーザほど生成AIによる業務効率化の可能性を感じている

出所: ユーザー調査1: 技術検証基盤のユーザー登録者（n=256）

業務効率化効果の一次試算

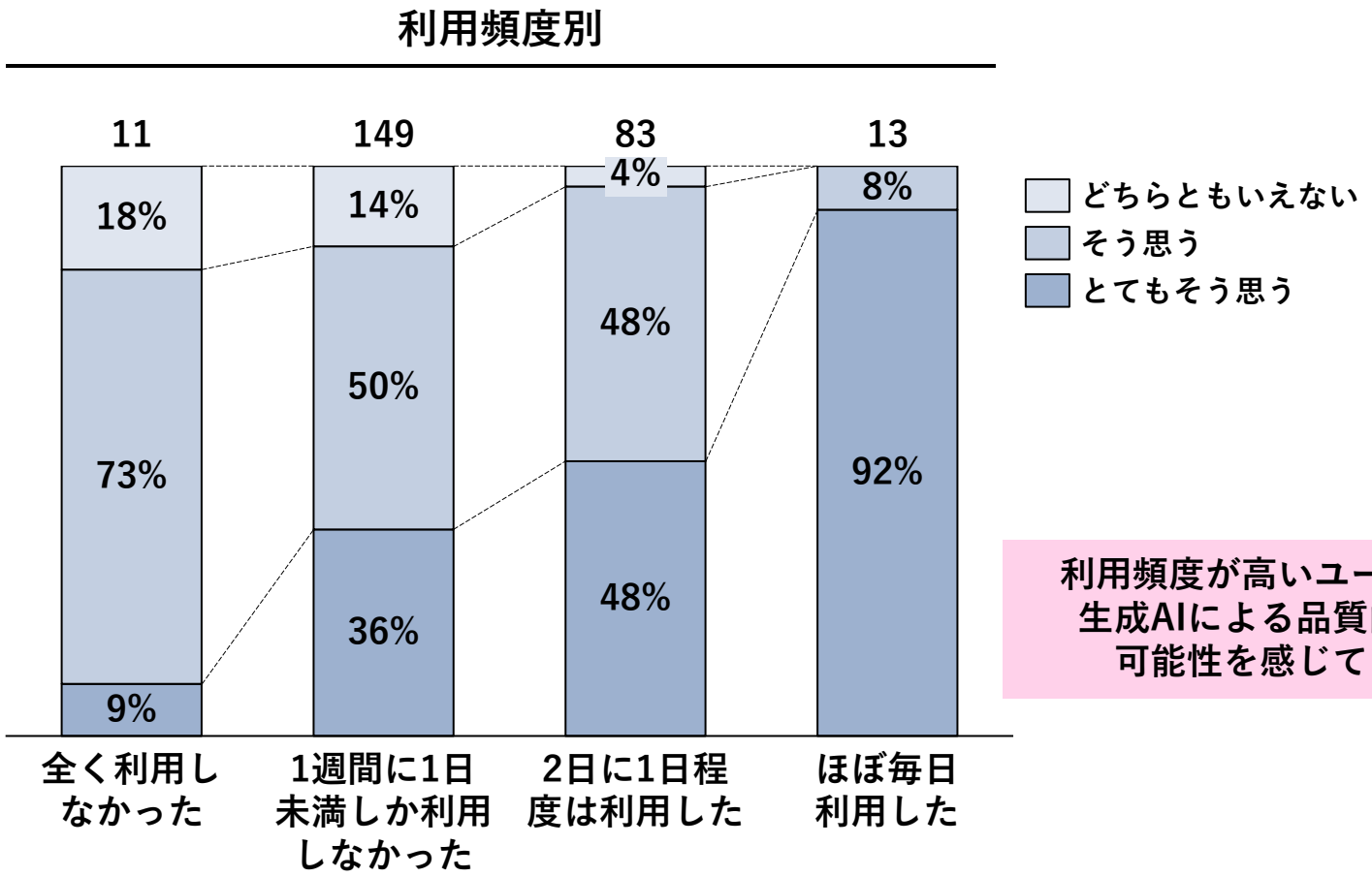
Q. 生成AIの活用により、回答した業務は月平均でどの程度効率化されると思いますか。
見込まれる削減効果について最も近いものを選んで下さい。（感覚値で構いません）



出所: ユーザー調査1: 技術検証基盤のユーザー登録者（n=256）

成果物の品質向上に対する期待

Q.あなたは生成AIを適用することにより、業務の品質向上が図れると思いますか、最も近いものを選んで下さい。



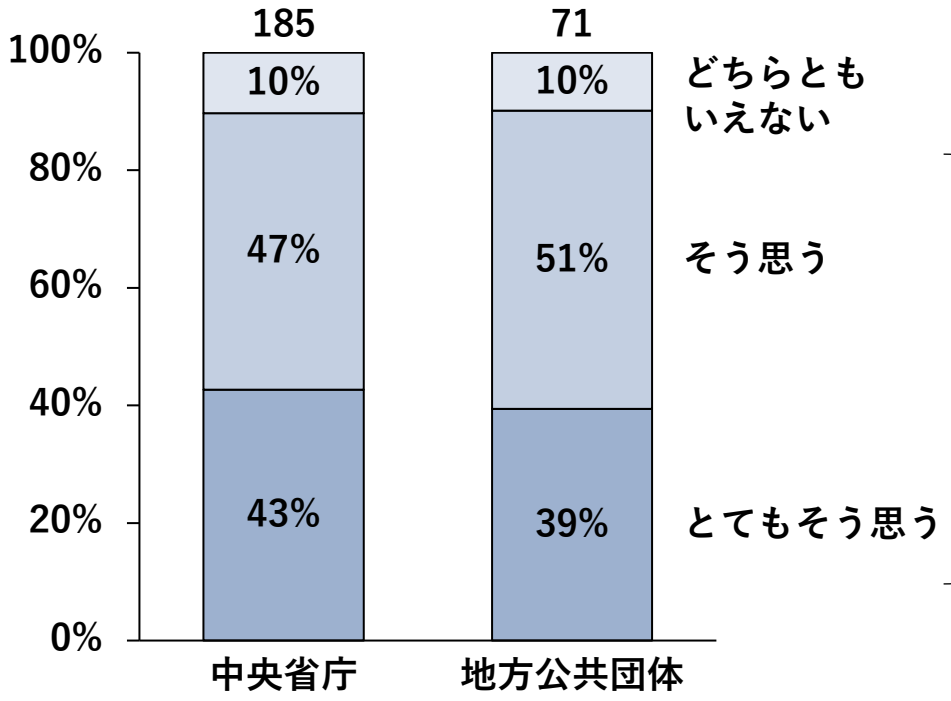
利用頻度が高いユーザほど生成AIによる品質向上の可能性を感じている

出所: ユーザー調査1: 技術検証基盤のユーザー登録者（n=256）

具体的な品質向上のポイント

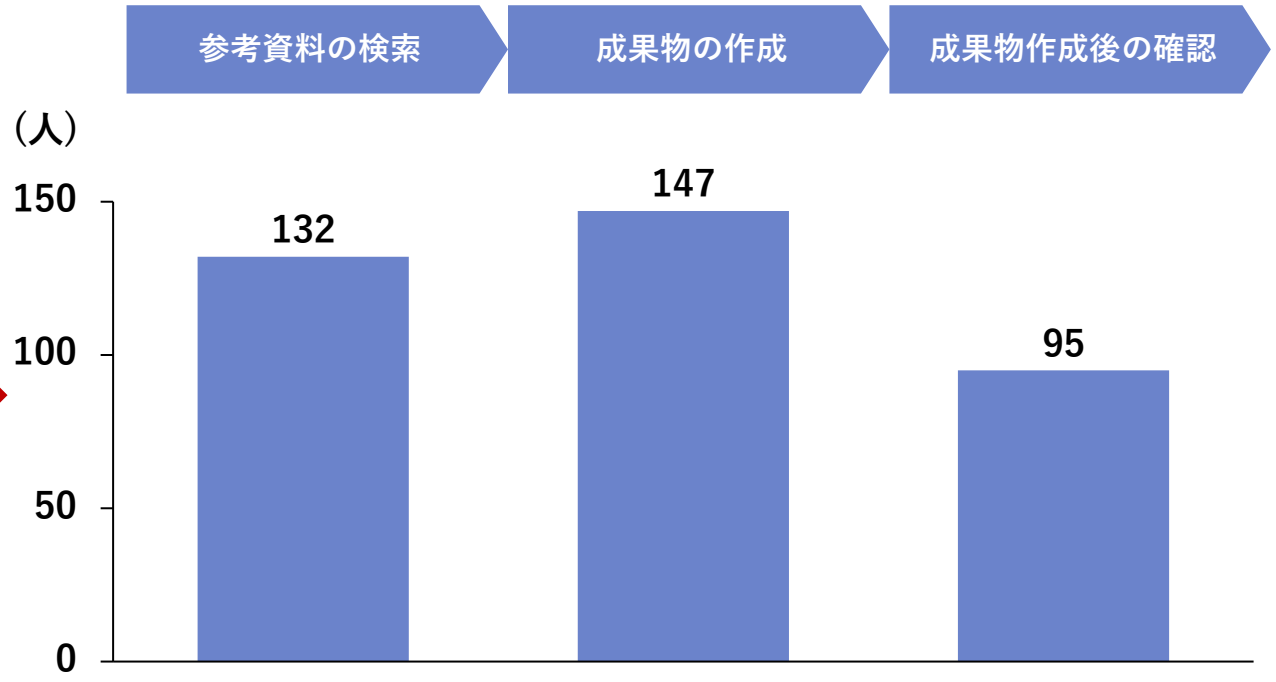
成果物の品質向上に対する期待

Q.あなたは生成AIを適用することにより、
業務の品質向上が図れると思いますか、
最も近いものを選んで下さい。



具体的な品質向上のポイント

Q. 前問で「とてもそう思う」「そう思う」と回答された方に
お伺いします。具体的にどのような点で品質向上が
図れると思いますか。当てはまるものをすべて選んで下さい。

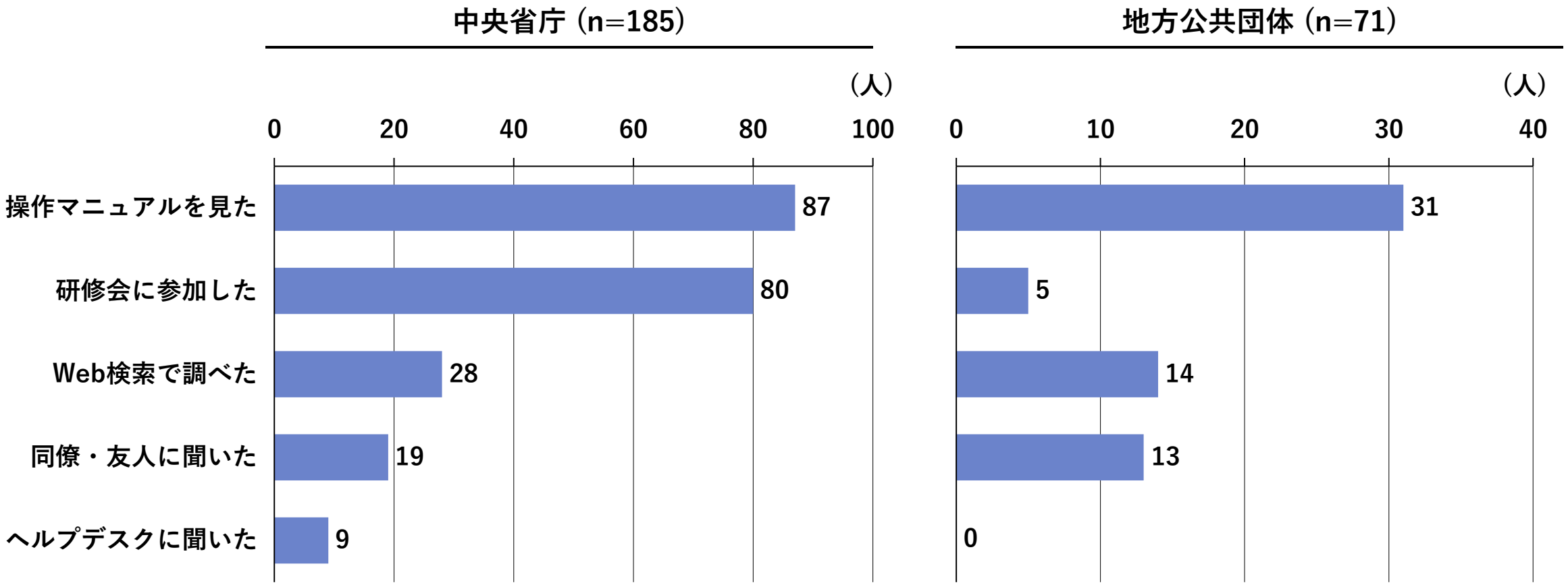


出所: ユーザー調査1: 技術検証基盤のユーザー登録者（n=256）

調査1 (5) その他

学習方法

あなたは生成AIの技術検証基盤(GaiXer)の操作方法について、どのように学習しましたか。当てはまるものをすべて選んで下さい



出所: ユーザー調査1: 技術検証基盤のユーザー登録者 (n=256)

調査1「技術検証基盤のユーザー登録者向け調査」からの示唆

全体利用状況と評価

- 技術検証基盤を**週2回以上利用したのは登録者の約4割**であった。中央政府/地方自治体や役職等による大きな差異は見られなかった。
- 利用が少ない理由は「忙しかった」「どのような業務に使用するのがよいのか分からなかった」「プロンプトの作り方が分からなかった」が多く、**明快なUI/UXや初期研修の重要性**が見えてきた。
- 一度に4回以上リクエストする登録者は、利用頻度と共に増えていき、ほぼ毎日利用している人の7割が該当する。**1回で答えを出すのではなく、繰り返し使うものという認識共有**が重要である。

機能・LLM別利用状況と評価

- 回答がおおむね適切と考えている利用者は、ほとんど利用していない層では約5割に留まるが、ほぼ毎日利用している層では約8割まで上昇。**利用頻度を高めることで、適切な使い方を学習**できる。
- 機能別に見ると、**プロンプトテンプレートの新規作成・利用機能のNPSが高位**であった。**サポートから毎日新たなテンプレートを発信するなど、啓蒙に一定の成果**があったと思慮
- LLM別に見るとGPT-4およびGPT-4 Turboの利用者が多数。**他のLLMの強みも周知する余地**がある。

業務効率化の可能性/成果物品質向上の可能性

- 業務効率化の可能性認識も、利用頻度とともに上昇する。ほとんど利用していない層は約8割だが、毎日利用してい層で全員が「上がる」「大幅に上がる」と回答しており、**効果を体感することが重要**である。
- 業務別に見ると、**パソコンの操作法やコード生成等のIT作業での効率化余地が大きい**と見られている。
- 成果物の品質向上認識も、利用頻度と相関する。**品質向上に対し「とてもそう思う」と回答した割合はほとんど利用していない層で約1割、毎日利用している層では約9割と顕著な差**が出た。

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

技術検証課題の報告要旨（エグゼクティブ・サマリーより再掲）

技術検証課題	報告要旨
1 LLMを用いた国・自治体職員の業務効率化や成果物の品質向上等に係る有用性の検証	**・技術検証基盤・共創PF Slack上で生成AIを提供し、職員の業務効率化や成果物の品質向上に一定の効果** – ユーザと対話しつつ**LLMとその設定、プロンプト、RAGの対象となる学習データを整備**し、ご利用いただいた。 – 利用者調査では、**90%以上の利用者が業務効率化・成果物の品質に効果あり**と回答した。
2 本件有用性の検証に係る試験的なデータの整備、データモデル等の提案	**・ユースケースの検証を通じて、学習データの整備についての知見を蓄積** – ガイドブック・規則等を用いたユースケースで、**学習データとして適切なデータ形式の知見**が得られた。 – **用例による法令検索（検索単語の一致に加え、 LLMで用例一致法令を抽出）を構築**し、さまざまな文書(例: 過去の国会答弁 等)で有用な可能性を見出した。本件では、このアルゴリズムも設計書として整理した。
3 LLMを業務利用のために調達する場合の要件定義・品質管理等のあり方	**・必要な回答速度・精度を踏まえ、丸投げではなく調達するLLMを検討する必要** – ユースケースで**GPT-4 TurboとClaude 3の回答品質を比較したところ、課題によって異なる結果**となった。 – **調達仕様の策定に際しては、複数LLMを比較・検討が実施できるSaaS等で試行を行うことも検討**すべき。
4 LLMの性能評価、競争性、相互運用性、リスク管理、選定基準などの整理	**・複数LLM・パラメータを比較し、性能評価できる環境を提供。リスクある利用にアラートを上げる仕組みも構築** – 本検証では**PlayGround機能を通じて、複数のLLM・パラメータの性能を評価できる環境を構築**した。 – リスク管理について、**情報公開法の記述を参考に機密性2情報を本基盤で取り扱う上での目安を策定**し、**機密情報の入力にアラートを上げる仕組みを試験的に実装**した。
5 その他、政府・自治体がLLMを活用するにあたり留意すべき課題の整理	**・政府・自治体が組織内でLLMを活用していく上で、業務に応じたプロンプト等の整備拡充が必要** – LLMが活用環境な可能であっても、**組織内でのユーザの拡大および適用対象業務の拡大に向けては、研修等の実施に加え、業務に応じたLLM・プロンプト・学習データをセットで提供する必要**がある。

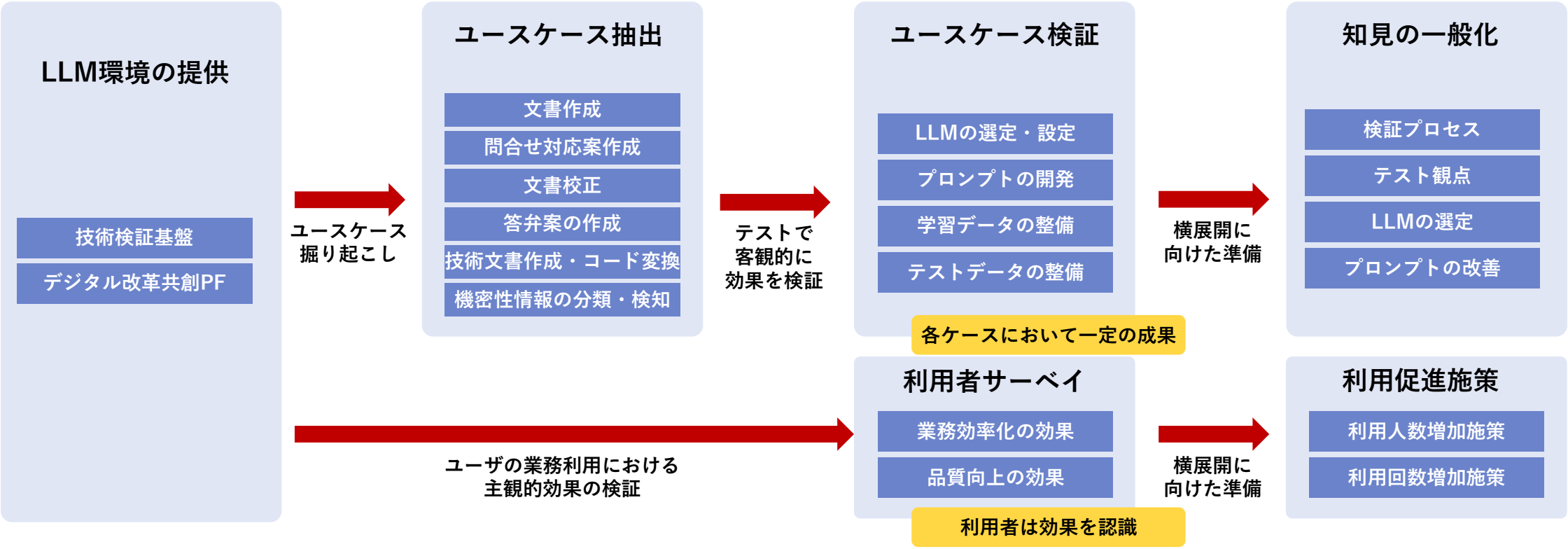
技術検証課題①: 論点・アプローチ・実施結果

検証テーマ	LLMを用いた国・自治体職員の業務効率化や成果物の品質向上等に係る有用性の検証

論点	アプローチ	実施結果
• **適用対象** – どんなユースケースに適用できるのか？ – ビジネスプロセスはどう変わるのか？	• **ユーザ職員へのヒアリング** • **ユースケース候補の抽出** • **クイック検証による有望性の検証**	• **文書作成、質問回答、文書校正、答弁案作成、コード変換の各ユースケースを検証し、一定以上の有用性があることが確認できた。**
• **適用方法** – 各ユースケースにおけるLLM選定、プロンプト開発、学習データ構築等をどのように行うべきか	• **各ユースケースからの手法一般化** –LLM選定 –プロンプト開発 –学習データ構築	• **ユースケース検証、プロンプト集の開発を通じてプロンプト改善プロセスの知見を得た。** • **主要LLMの特徴を踏まえて比較実験を行い、業務への向き/不向きについて示唆を得た。**
• **有用性** – 各ユースケースへのLLM適用は、業務効率化や成果物の品質向上にどの程度寄与するのか？	• **利用者アンケート**	• **利用者アンケートでは、利用者の90%以上が業務効率化・成果物の品質向上にLLMが寄与すると回答した。**
• **利用普及** – LLMの利用回数が多い人/少ない人に見られる特徴は何か？ – 利用者数増加・エンゲージメント向上にはどのような施策が有効か？	• **利用データ分析** • **利用者インタビュー** • **利用者アンケート**	• **利用頻度が高いユーザほど、1回ではなく繰り返しリクエストすることを認識し、業務効率化・品質向上を強く認識していた。一定程度まではハンズオンで導入し、LLMの価値を共有していくことが重要である。**

技術検証課題①: 結論

検証テーマ	**LLMを用いた国・自治体職員の業務効率化や成果物の品質向上等に係る有用性の検証**
サマリ	**・技術検証基盤・共創PF Slack上で生成AIを提供し、職員の業務効率化や成果物の品質向上に一定の効果** – ユーザと対話しつつ**LLMとその設定、プロンプト、RAGの対象となる学習データを整備**し、ご利用いただいた。 – 利用者調査では、**90%以上の利用者が業務効率化・成果物の品質に効果あり**と回答した。



技術検証課題②: 論点・アプローチ・実施結果

検証テーマ	本件有用性の検証に係る試験的なデータの整備、データモデル等の提案

論点	アプローチ	実施結果
・データ整備 – 法令APIなど、政府が公開するオープンデータをどう活用すべきか？ – 今後、さらなるLLM活用のためにどのようなデータ整備が必要か？ – 府省庁間・自治体間などで、RAG用の学習データなどをどう共有すべきか？	**・ユースケースを通じた課題の洗い出し**	**・ユースケースC-1(用例による法令検索)で、e-gov法令検索の全法令データを利用** –法令APIのキーワード検索機能はプロトタイプのため、今回は利用せず –こうしたキーワード検索APIが他データ(社会統計 等)においても、広く公開されることが、LLM活用促進のために望まれる **・ユースケースD-1(質問主意書への回答生成)では、国会の質問・答弁データを利用** –過去の国会質問・答弁データの学習だけでは、生成精度は限定的 (正解データのキーワードを与えることで精度は向上) –各種会見・議事録等のデータが広く整備され、一元的に提供されることが望ましい
・データモデル – 追加データ学習で精度を高めるには検索結果を用いたRAGと、プロンプトへの詰め込みをどう使い分けるべきか？	**・ユースケースを通じた有効な手法検討**	**・ユースケースB-1(デジタル・ガバメント推進標準ガイドライン実践ガイドブック)では、長文データを学習させる方式を検証** –長文に対応したLLMにより要約行い、システムプロンプトに盛り込む方式がRAGよりも高い回答精度を達成

技術検証課題②: 結論

検証テーマ	本件有用性の検証に係る試験的なデータの整備、データモデル等の提案
サマリ	・ユースケースの検証を通じて、学習データの整備についての知見を蓄積 – ガイドブック・規則等を用いたユースケースで、**学習データとして適切なデータ形式の知見**が得られた。 – **用例による法令検索（検索単語の一致に加え、LLMで用例一致法令を抽出）を構築**し、さまざまな文書(例: 過去の国会答弁 等)で有用な可能性を見出した。本件では、このアルゴリズムも設計書として整理した。

データ整備にかかわる提案

政府公開データの拡充

- 法令データ
- 国会データ
- 統計データ
- 府省庁Webデータ

公開方式の最適化

- キーワード検索API
- データ様式の統一化
- 用例一致検索機能の提供

具体的なデータの洗い出し・フォーマットの策定に向けた深堀り検討の余地

LLM活用におけるさらなる検証課題

学習データ化ノウハウ

- RAG or プロンプト詰め込み
- ノイズ情報の除去
- ファイル形式の変換

学習データ共有化

- 政府内での共有
- 自治体間での共有
- 政府ー自治体間の共有

整備されたデータがLLM上で活用されるよう追加学習ノウハウやデータ共有の仕組みを要検討

技術検証課題③: 論点・アプローチ・実施結果

検証テーマ	LLMを業務利用のために調達する場合の要件定義・品質管理等のあり方

論点	アプローチ	実施結果
・要件定義上の課題 – LLMの要件をどのように定めるべきか？（マルチLLM、最新LLMの追加可能性 等） – 追加開発を伴う場合、開発方式等をどうすべきか？（Waterfall vs Agile 等）	**・プロジェクト中に発生した追加要望とリリース実績の振り返り**	**・生成AI領域は特に進化が早く、プロジェクトの期間中も試行の価値があると思われるLLMが出現した（例: AnthropicのClaude 3 Sonnet）** **⇒ 取組内容次第ではあるが、要件に最新のLLMを追加できる等の条項を設けることで検討結果の陳腐化を回避しうる可能性** **・ユースケースの検討を通じ、主要LLMであっても、課題によって向き/不向きが存在** **⇒ どのLLMを調達するか、ベンダーに任せず、要件で方針を示すことが望ましい** –デジタル庁が、情報提供を担ったりガイドラインを策定する可能性も？
・品質管理上の課題 – LLMの”品質”をどう捉えるべきか？ – どの程度の品質を追い求めるべきか？	**・ユースケースにおけるテストケースの設計およびパフォーマンス評価**	**・LLMの品質管理に関しては、本技術検討で品質管理プロセスの“型”を提示** **⇒ ユースケース単位でテストケースを設計し、試行錯誤を通じて品質を評価** **・人間との協働が前提である以上、LLMの回答精度だけを評価する意味は限定的** **⇒ ユースケース上、必要な精度を設定**

技術検証課題③: 結論

検証テーマ	LLMを業務利用のために調達する場合の要件定義・品質管理等のあり方
サマリ	・必要な回答速度・精度を踏まえ、丸投げではなく調達するLLMを検討する必要 – ユースケースでGPT-4 TurboとClaude 3の回答品質を比較したところ、課題によって異なる結果となった – 調達仕様の策定に際しては、複数LLMを比較・検討が実施できるSaaS等で試行を行うことも検討すべき

	要件定義	プロジェクト進行	品質評価
LLM調達の難しさ	生成AIの進化が非常に早く 検討が陳腐化するリスク	ユースケースによっては LLMの活用を支援する 追加機能開発の必要が発生	絶対的な品質評価が困難 カタログ性能は実態に合わず 100%の精度というものもない
対応策（素案）	要件定義で 複数LLMの併用・比較や 最新LLM対応を盛り込む	ユースケースに必要な 機能追加が行えるよう アジャイルを基本にする	ユースケース単位で品質評価 自らテストケースを作り 必要品質を定義・評価する

技術検証課題④: 論点・アプローチ・実施結果

検証テーマ	LLMの性能評価、競争性、相互運用性、リスク管理、選定基準などの整理

論点	アプローチ	実施結果
• 性能評価・選定基準 （③の品質管理において述べた通り）	ー	–
• 競争性・相互運用性 – LLMベンダーのロックインを避け、どう競争性・相互運用性を担保するか？	• マルチクラウド、マルチLLMが選択可能な環境を提供し、競争性・相互運用性を確認 • Playground機能により、複数のLLMでの実行結果を比較・検討できる仕組みを構築	• ユースケースの課題に応じてLLMを選定できる基盤を確保することで、LLM間の競争性を担保することが可能 • LLM選定における、Playground機能の有用性を確認。調達仕様策定の支援が可能
• リスク管理 – 機密情報がLLMに入力されるリスクをどのように低減するか？ – 機密情報を取り扱うために、政府や自治体が確認すべきことは何か？	• セキュリティ班とISMAPチェックリスト及びセキュリティポリシーとの整合性を確認、課題を洗い出し、打ち手を実施 • 上記を踏まえた運用を運用計画書および実施要領としてドキュメント化 • 入力時に機密情報の有無を確認するチェッカーを実装し、入力状況を確認	• ISMAPチェックリスト及びセキュリティポリシーに基づき、メールOTPを実装し、ソフトウェア・ネットワークの脆弱性を診断 • 今後のLLM調達に活用可能な形で運用計画書及び実施要領をドキュメント化 • ダッシュボードにおいて制限情報の入力率や具体的内容を集計。今後の活用ガイドライン制定等に活用

技術検証課題④: 結論

検証テーマ	LLMの性能評価、競争性、相互運用性、リスク管理、選定基準などの整理
サマリ	・複数LLM・パラメータを比較し、性能評価できる環境を提供。リスクある利用にアラートを上げる仕組みも構築 – 本検証では<u>PlayGround機能を通じて、複数のLLM・パラメータの性能を評価できる環境を構築</u>した – リスク管理について、<u>情報公開法の記述を参考に機密性2情報を本基盤で取り扱う上での目安を策定</u>し、<u>機密情報の入力にアラートを上げる仕組みを試験的に実装</u>した

性能評価・選定基準のポイント

- 意味のあるユースケース選定
- ユースケースとLLMの適合性
- 適切なテスト観点の設定
- 適切な目標精度の設定

競争性・相互運用性のポイント

- マルチクラウド環境
- LLMの追加対応・相互比較
- データの移行可能性
- ブラックボックスの除外

リスク管理のポイント

- システム自体の脆弱性診断
- 安全な運用・保守体制
- 適切なガイドライン設定
- 機密情報入力の未然防止

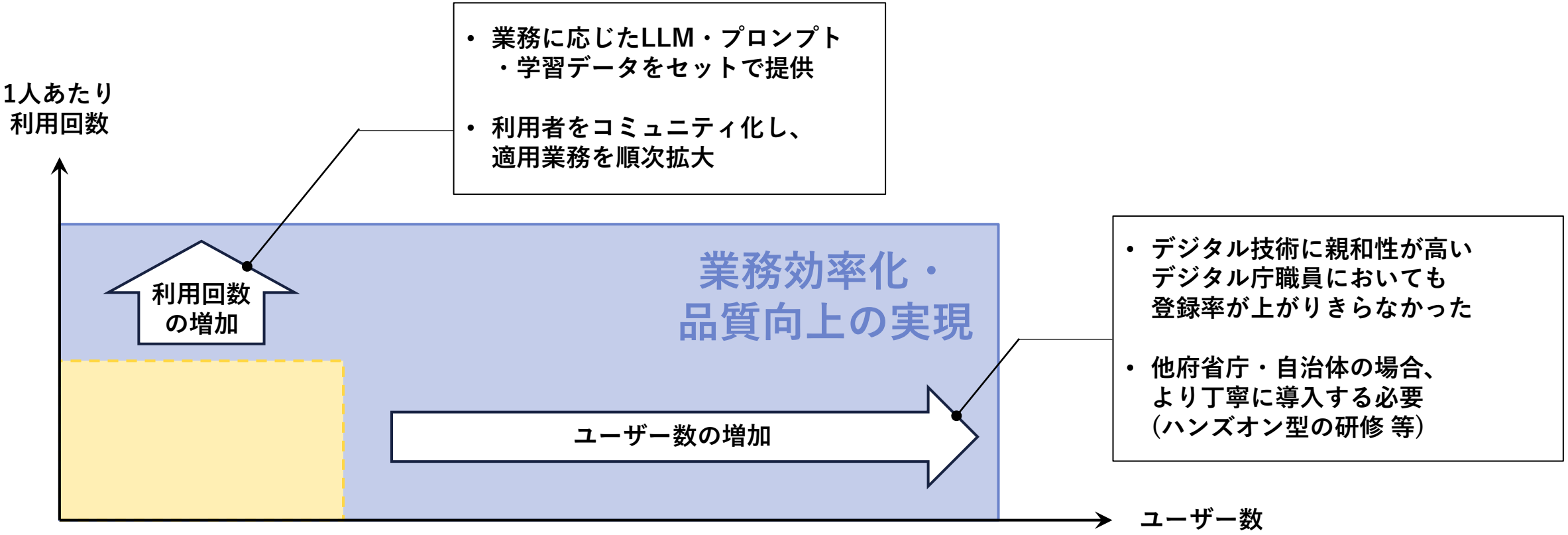
参考: 技術検証環境(GaiXer)を利用可能な機密性2情報の目安

機密性2情報のうち、情報公開法第5条第5号・第6条に該当し、修正・編集の上で将来公開が予定されている情報は技術検証環境を利用可能とし、具体的には担当部局の管理職が判断する。参考のため、以下に目安の一例を示す。

項	号	技術検証環境利用可能 (一例)	技術検証環境利用不可 (一例)
五 内部又は相互間における審議、検討又は協議に関する情報		・審議会等の議事メモ(公開前) ・パブリックコメント(公開前)	
六 事務又は事業に関する情報	イ　監査、検査、取締り、試験又は租税の賦課若しくは徴収に係る事務(後略)	・固有の監査等に関する情報を含まない一般論	・金融機関に対する検査に関する情報 ・交通違反の取り締まりに関する情報 ・未実施の試験問題に関する情報 ・税務調査の実施予定に関する情報
	ロ　契約、交渉又は争訟に係る事務(後略)	・固有の契約等に関する情報を含まない一般論	・入札に係る予定価格等に関する情報
	ハ 調査研究に係る事務(後略)	・調査/研究にかかわる情報全般	
	ニ　人事管理に係る事務(後略)	・人事管理に関連する内規等、組織内で公開され、個別の決定にかかわらない情報	・人事異動、昇給・昇格、勤務評定、懲戒処分の決定経過等が明らかになる情報
	ホ　独立行政法人等、地方公共団体が経営する企業又は地方独立行政法人に係る事業(後略)	・地方公共団体が経営する企業からの報告等、組織内で公開されている情報	

技術検証課題⑤: 結論（成果を挙げるための利用拡大の方策）

検証テーマ	その他、政府・自治体がLLMを活用するにあたり留意すべき課題の整理
サマリ	・政府・自治体が組織内でLLMを活用していく上で、業務に応じたプロンプト等の整備拡充が必要 － LLMが活用環境な可能であっても、組織内でのユーザの拡大および適用対象業務の拡大に向けては、研修等の実施に加え、業務に応じたLLM・プロンプト・学習データをセットで提供する必要がある。



目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

Appendix A. 検証基盤の利用状況

1. **登録アカウント数(日次)**
   - 全体
   - デジタル庁
   - 他府省庁
   - 自治体

2. **利用アカウント数・リクエスト数(週次)**
   - 全体
   - デジタル庁
   - 他府省庁
   - 自治体

3. **継続利用者数(週次)**
   - 全体
   - デジタル庁
   - 他府省庁
   - 自治体

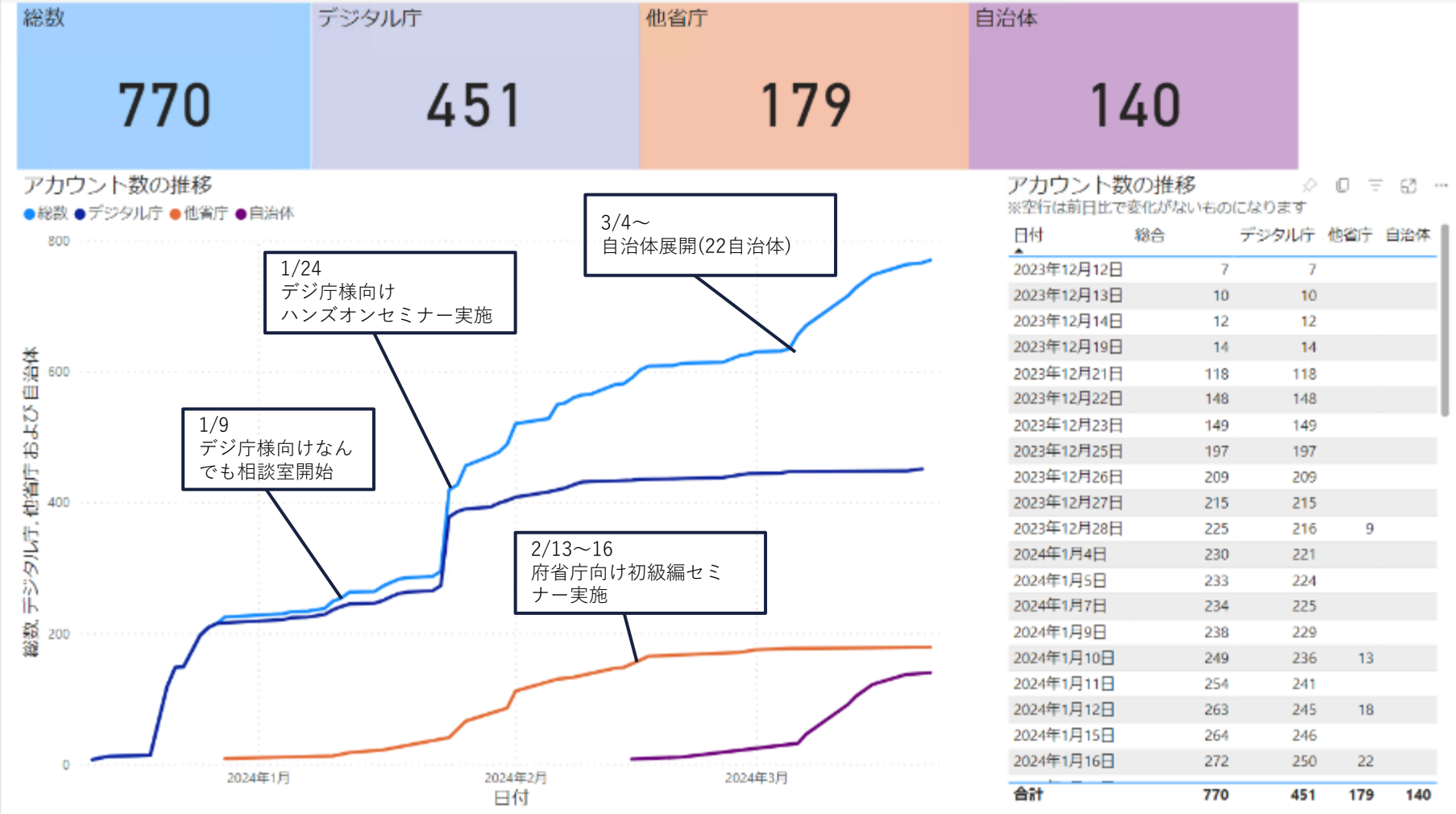
4. **学習セット利用アカウント数・利用回数(週次)**
   - 全体
   - デジタル庁
   - 他府省庁
   - 自治体

5. **モデル別利用比率(週次)**
   - 全体
   - デジタル庁
   - 他府省庁
   - 自治体

6. **問い合わせ内容**

1. 登録アカウント数(日次)

- ▶ 1/9「なんでも相談室」開始後、緩やかにアカウント数が増加
- ▶ 1/24 デジタル庁様向けハンズオンセミナーを実施した際に、急激なアカウント数の増加がみられた
- ▶ 3/4以降は自治体への展開を行ったため、自治体のアカウント数増加が大きくなった一方で、デジタル庁や府省庁の大きな増加はみられなかった



日付	総合	デジタル庁	他省庁	自治体
2023年12月12日	7	7		
2023年12月13日	10	10		
2023年12月14日	12	12		
2023年12月19日	14	14		
2023年12月21日	118	118		
2023年12月22日	148	148		
2023年12月23日	149	149		
2023年12月25日	197	197		
2023年12月26日	209	209		
2023年12月27日	215	215		
2023年12月28日	225	216	9	
2024年1月4日	230	221		
2024年1月5日	233	224		
2024年1月7日	234	225		
2024年1月9日	238	229		
2024年1月10日	249	236	13	
2024年1月11日	254	241		
2024年1月12日	263	245	18	
2024年1月15日	264	246		
2024年1月16日	272	250	22	
合計	**770**	**451**	**179**	**140**

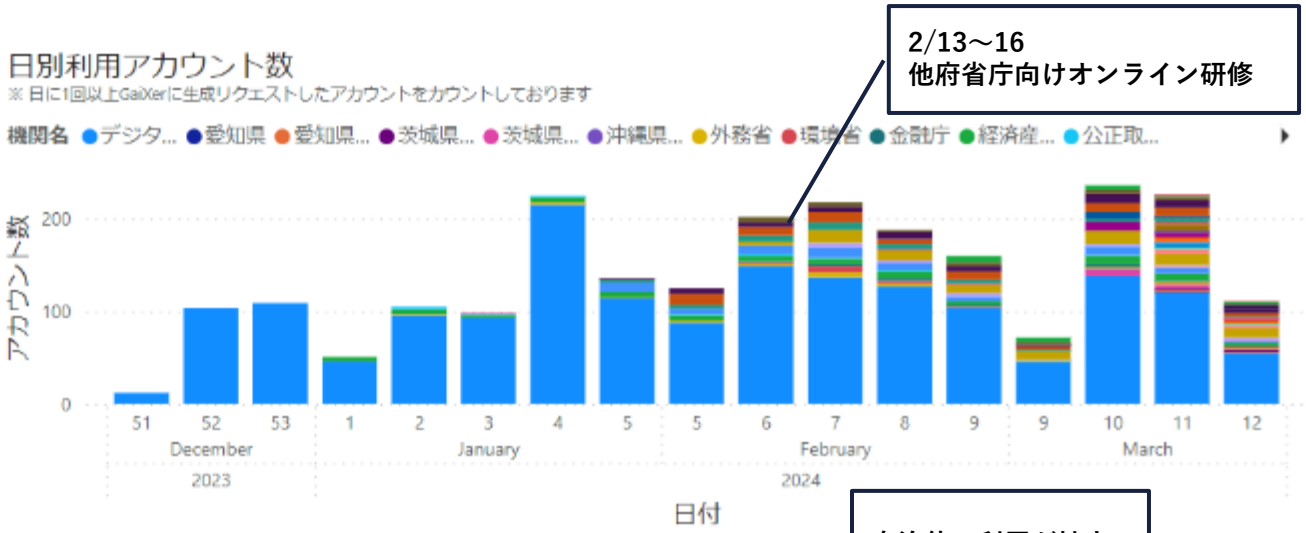
2. 利用アカウント数・リクエスト数(週次)全体

- ▶ 利用回数においては、2/5～11週で最大値となった(3480回/週) 。2/13～16に他府省庁向けのセミナーを開催したので、その影響と考える。
- ▶ 利用アカウントにおいては、3/4～10週で最大値となった(236アカウント/週)。3/4～自治体への展開が始まったためと想定する。

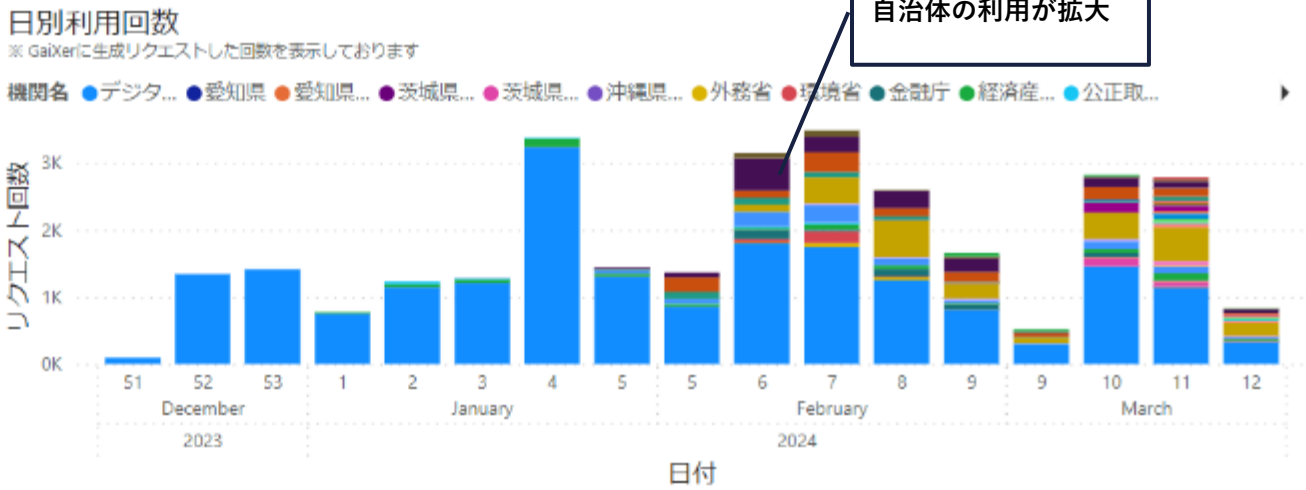
週番号	StartOfWeek	利用可数	利用アカウント数
51	2023/12/11 0:00:00	96	12
52	2023/12/18 0:00:00	1341	104
53	2023/12/25 0:00:00	1411	109
1	2024/01/01 0:00:00	774	51
2	2024/01/08 0:00:00	1229	105
3	2024/01/15 0:00:00	1281	98
4	2024/01/22 0:00:00	3375	225
5	2024/01/29 0:00:00	2807	195
6	2024/02/05 0:00:00	3142	202
7	2024/02/12 0:00:00	3480	218
8	2024/02/19 0:00:00	2593	188
9	2024/02/26 0:00:00	2179	185
10	2024/03/04 0:00:00	2822	236
11	2024/03/11 0:00:00	2780	226
12	2024/03/18 0:00:00	831	111

利用回数最大(218)

利用アカウント数最大(236)



2/13～16
他府省庁向けオンライン研修



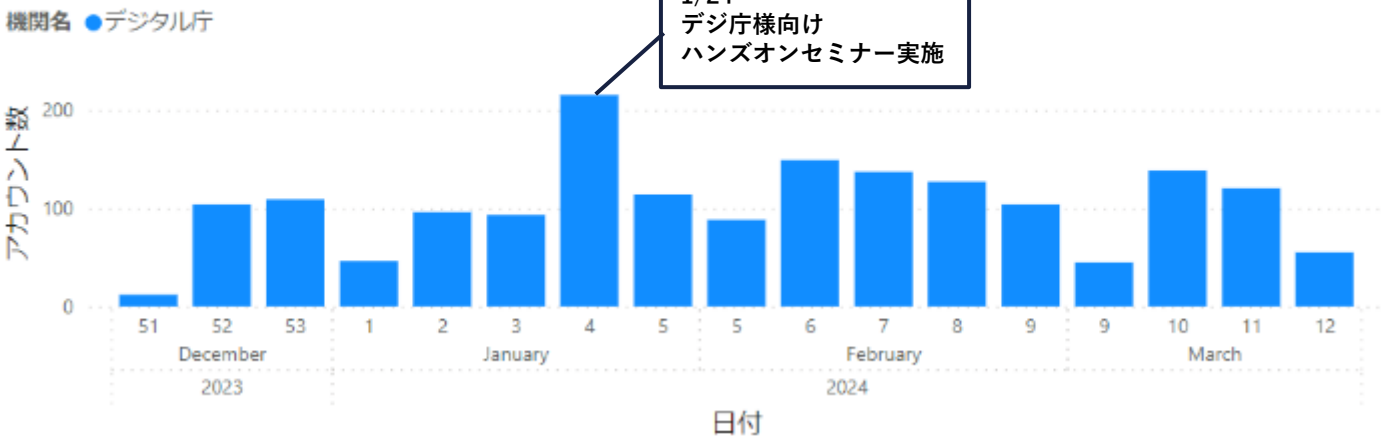
自治体の利用が拡大

2. 利用アカウント数・リクエスト数(週次)デジタル庁

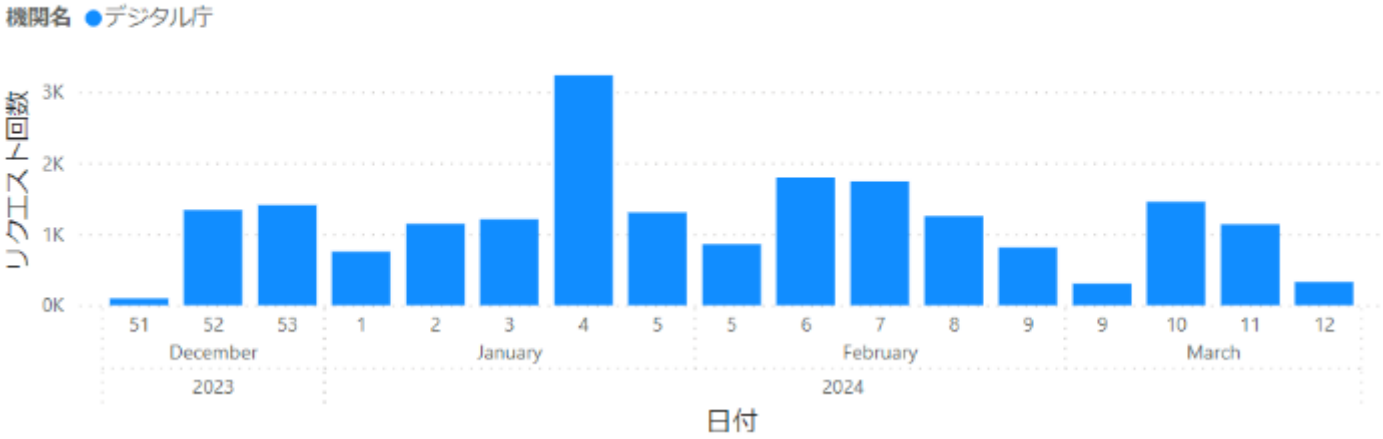
週番号	StartOfWeek	利用可数	利用アカウント数
51	2023/12/11 0:00:00	96	12
52	2023/12/18 0:00:00	1341	104
53	2023/12/25 0:00:00	1411	109
1	2024/01/01 0:00:00	753	46
2	2024/01/08 0:00:00	1147	96
3	2024/01/15 0:00:00	1211	93
4	2024/01/22 0:00:00	3235	215
5	2024/01/29 0:00:00	2160	143
6	2024/02/05 0:00:00	1797	149
7	2024/02/12 0:00:00	1740	137
8	2024/02/19 0:00:00	1251	127
9	2024/02/26 0:00:00	1112	122
10	2024/03/04 0:00:00	1455	138
11	2024/03/11 0:00:00	1138	120
12	2024/03/18 0:00:00	330	55

利用回数最大(3,235)
利用アカウント最大(215)

日別利用アカウント数
※ 日に1回以上GaiXerに生成リクエストしたアカウントをカウントしております



日別利用回数
※ GaiXerに生成リクエストした回数を表示しております

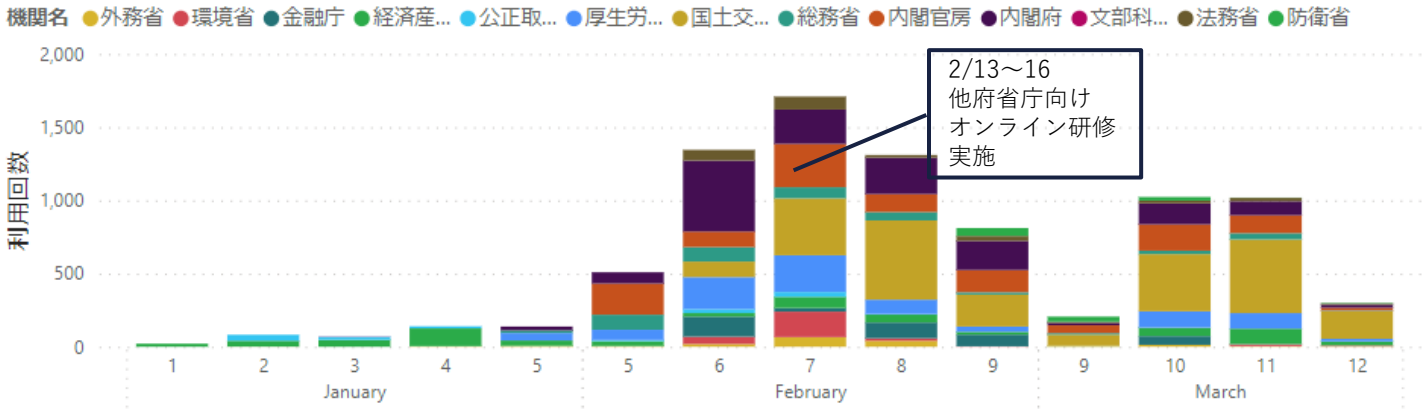


2. 利用アカウント数・リクエスト数(週次)他府省庁

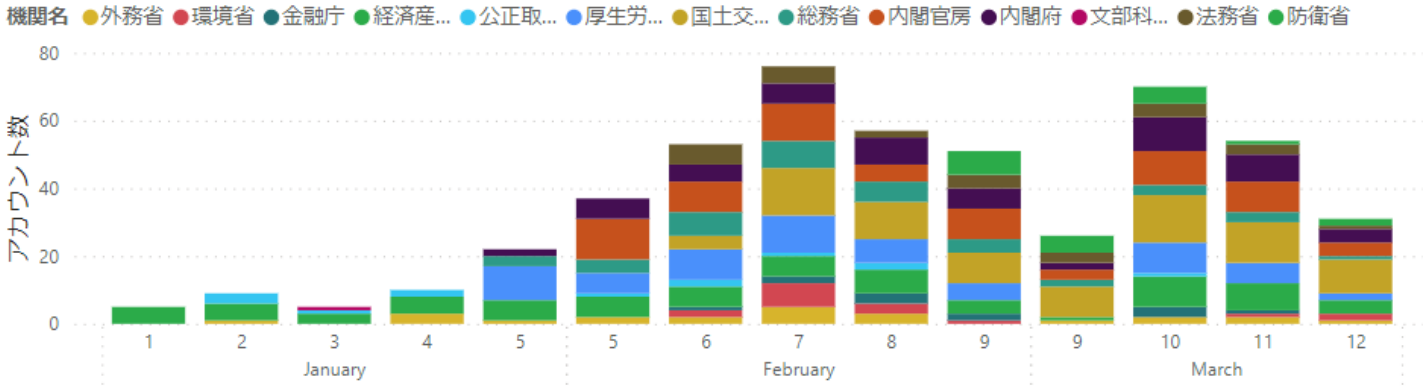
利用回数最大(1,709)
利用アカウント数最大(76)

週番号	開始日	利用回数	利用アカウント数
1	2024/01/01 0:00:00	21	5
2	2024/01/08 0:00:00	82	9
3	2024/01/15 0:00:00	70	5
4	2024/01/22 0:00:00	140	10
5	2024/01/29 0:00:00	647	52
6	2024/02/05 0:00:00	1345	53
7	2024/02/12 0:00:00	1709	76
8	2024/02/19 0:00:00	1310	57
9	2024/02/26 0:00:00	1017	58
10	2024/03/04 0:00:00	1023	70
11	2024/03/11 0:00:00	1018	54
12	2024/03/18 0:00:00	300	31
合計		**8682**	**137**

日別利用回数
※ GaiXerに生成リクエストした回数を表示しております



日別利用アカウント数
※ 日に1回以上GaiXerに生成リクエストしたアカウントをカウントしております



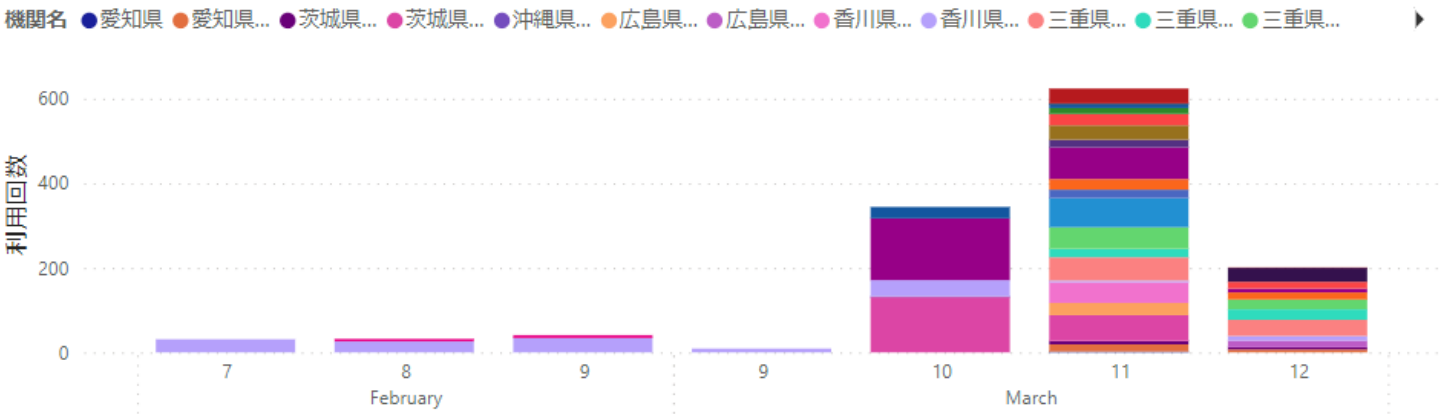
2. 利用アカウント数・リクエスト数(週次)自治体

利用回数最大(624)
利用アカウント数最大(52)

週番号	開始日	利用回数	利用アカウント数
7	2024/02/12 0:00:00	31	5
8	2024/02/19 0:00:00	32	4
9	2024/02/26 0:00:00	50	5
10	2024/03/04 0:00:00	344	28
11	2024/03/11 0:00:00	624	52
12	2024/03/18 0:00:00	201	25
合計		**1282**	**82**

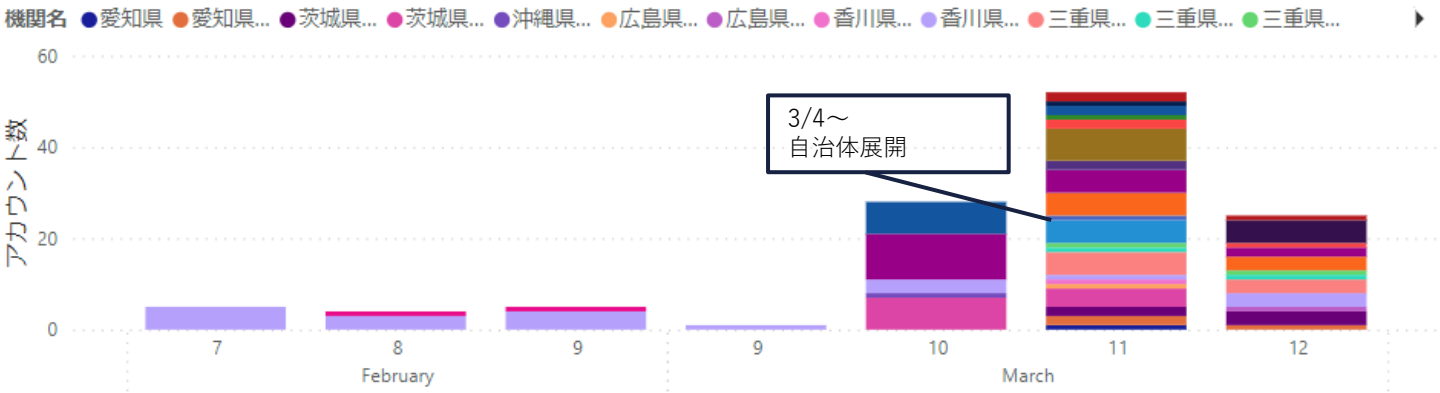
日別利用回数

※ GaiXerに生成リクエストした回数を表示しております



日別利用アカウント数

※ 日に1回以上GaiXerに生成リクエストしたアカウントをカウントしております



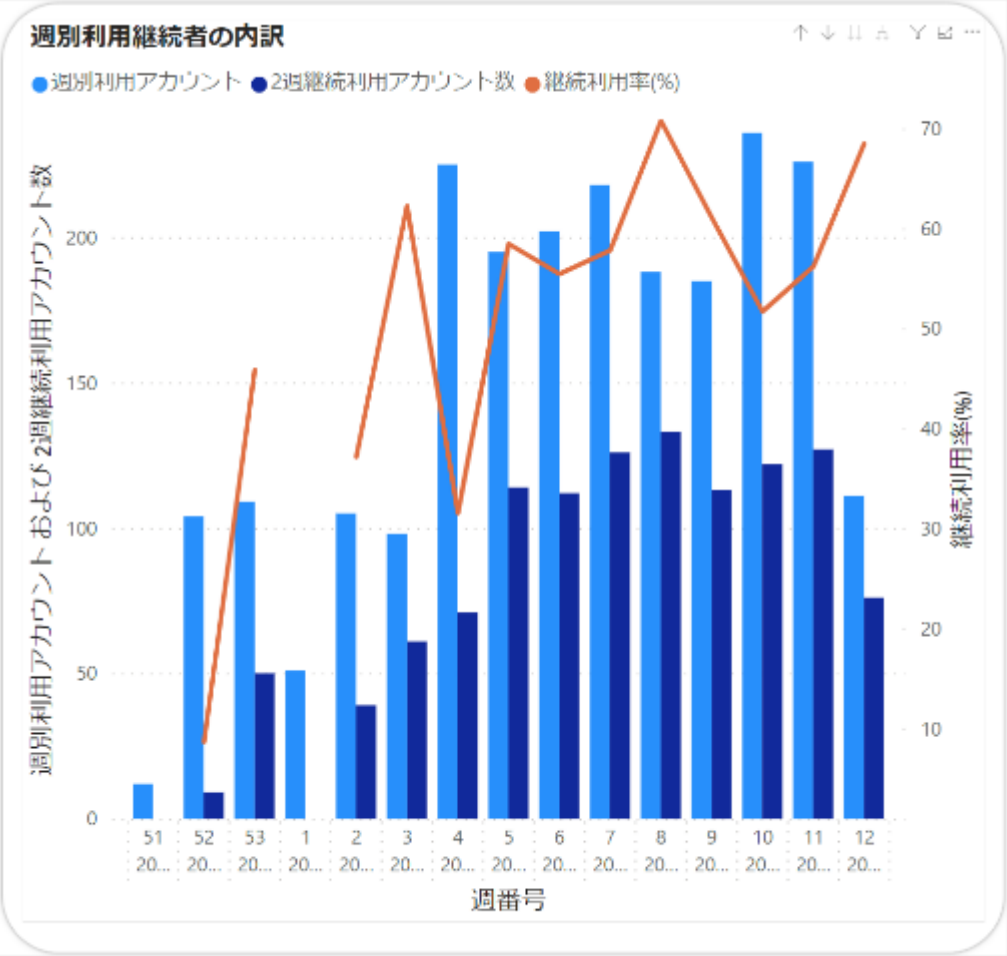
3. 継続利用者数(週次)全体

- ▶ 毎週平均170名の方にご利用いただき、継続利用率30%以上を維持できた。
- ▶ 2月以降は、継続利用率50%以上を維持し、一定のユーザーに対し利用を定着できたといえる。

週別利用継続者の内訳

※週に1度でもリクエストをしたアカウントを対象に集計をしています。

週番号	開始日	週別利用アカウント	2週継続利用アカウント数	継続利用率(%)
51	2023/12/11 0:00:00	12		
52	2023/12/18 0:00:00	104	9	8.65
53	2023/12/25 0:00:00	109	50	45.87
1	2024/01/01 0:00:00	51		
2	2024/01/08 0:00:00	105	39	37.14
3	2024/01/15 0:00:00	98	61	62.24
4	2024/01/22 0:00:00	225	71	31.56
5	2024/01/29 0:00:00	195	114	58.46
6	2024/02/05 0:00:00	202	112	55.45
7	2024/02/12 0:00:00	218	126	57.80
8	2024/02/19 0:00:00	188	133	70.74
9	2024/02/26 0:00:00	185	113	61.08
10	2024/03/04 0:00:00	236	122	51.69
11	2024/03/11 0:00:00	226	127	56.19
12	2024/03/18 0:00:00	111	76	68.47

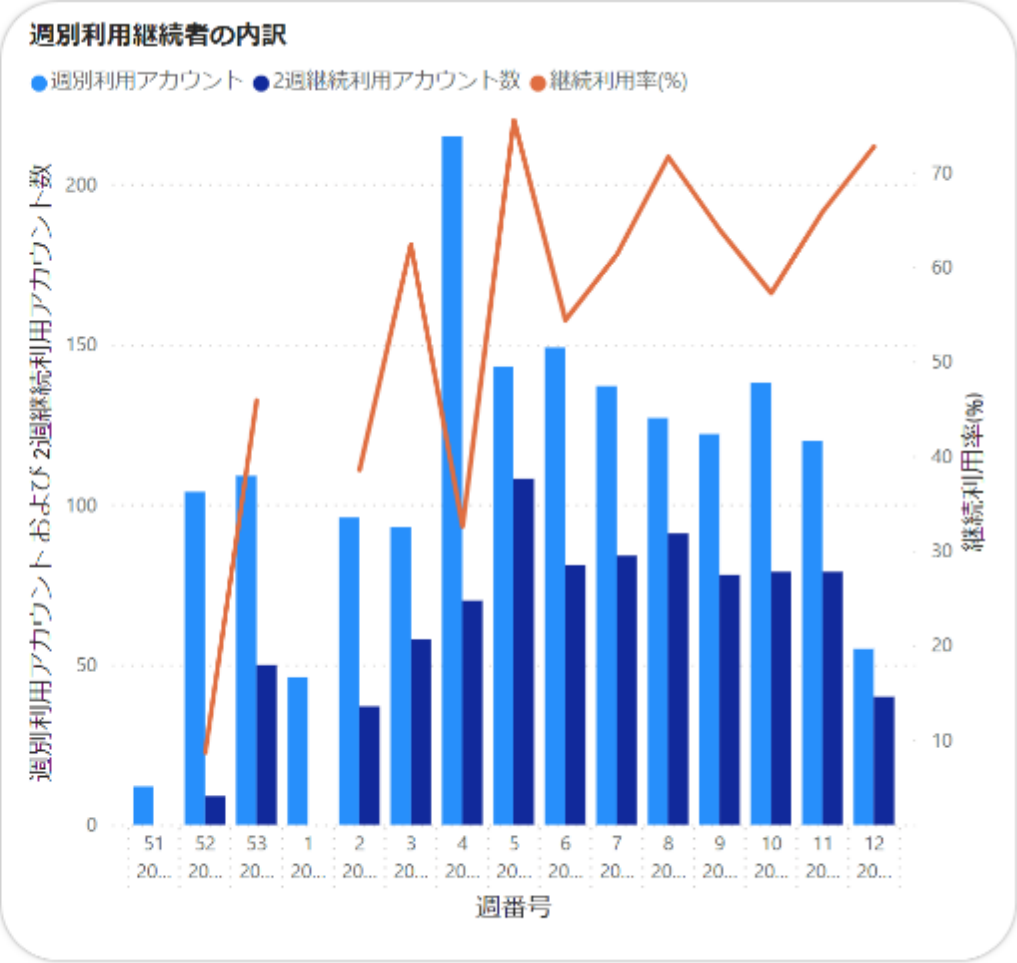


3. 継続利用者数(週次)デジタル庁

週別利用継続者の内訳

※週に1度でもリクエストをしたアカウントを対象に集計をしています。

週番号	開始日	週別利用アカウント	2週継続利用アカウント数	継続利用率(%)
51	2023/12/11 0:00:00	12		
52	2023/12/18 0:00:00	104	9	8.65
53	2023/12/25 0:00:00	109	50	45.87
1	2024/01/01 0:00:00	46		
2	2024/01/08 0:00:00	96	37	38.54
3	2024/01/15 0:00:00	93	58	62.37
4	2024/01/22 0:00:00	215	70	32.56
5	2024/01/29 0:00:00	143	108	75.52
6	2024/02/05 0:00:00	149	81	54.36
7	2024/02/12 0:00:00	137	84	61.31
8	2024/02/19 0:00:00	127	91	71.65
9	2024/02/26 0:00:00	122	78	63.93
10	2024/03/04 0:00:00	138	79	57.25
11	2024/03/11 0:00:00	120	79	65.83
12	2024/03/18 0:00:00	55	40	72.73



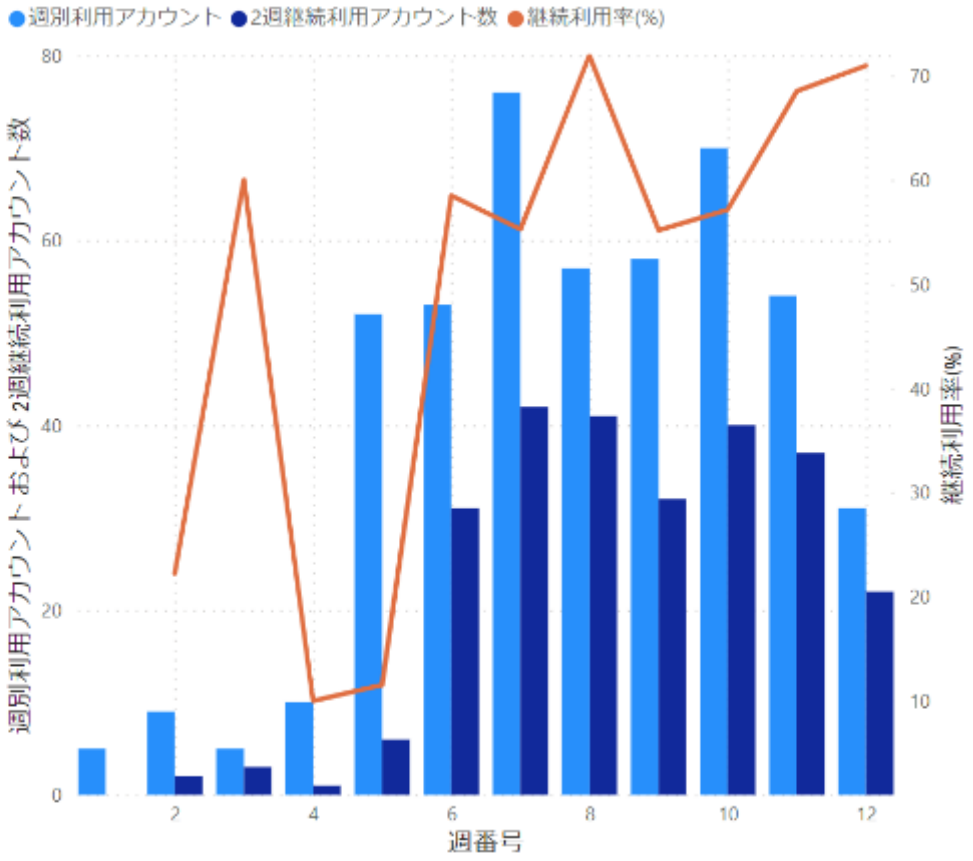
3. 継続利用者数(週次)他府省庁

週別利用継続者の内訳

※週に1度でもリクエストをしたアカウントを対象に集計をしています。

週番号	開始日	週別利用アカウント	2週継続利用アカウント数	継続利用率(%)
1	2024/01/01 0:00:00	5		
2	2024/01/08 0:00:00	9	2	22.22
3	2024/01/15 0:00:00	5	3	60.00
4	2024/01/22 0:00:00	10	1	10.00
5	2024/01/29 0:00:00	52	6	11.54
6	2024/02/05 0:00:00	53	31	58.49
7	2024/02/12 0:00:00	76	42	55.26
8	2024/02/19 0:00:00	57	41	71.93
9	2024/02/26 0:00:00	58	32	55.17
10	2024/03/04 0:00:00	70	40	57.14
11	2024/03/11 0:00:00	54	37	68.52
12	2024/03/18 0:00:00	31	22	70.97

週別利用継続者の内訳



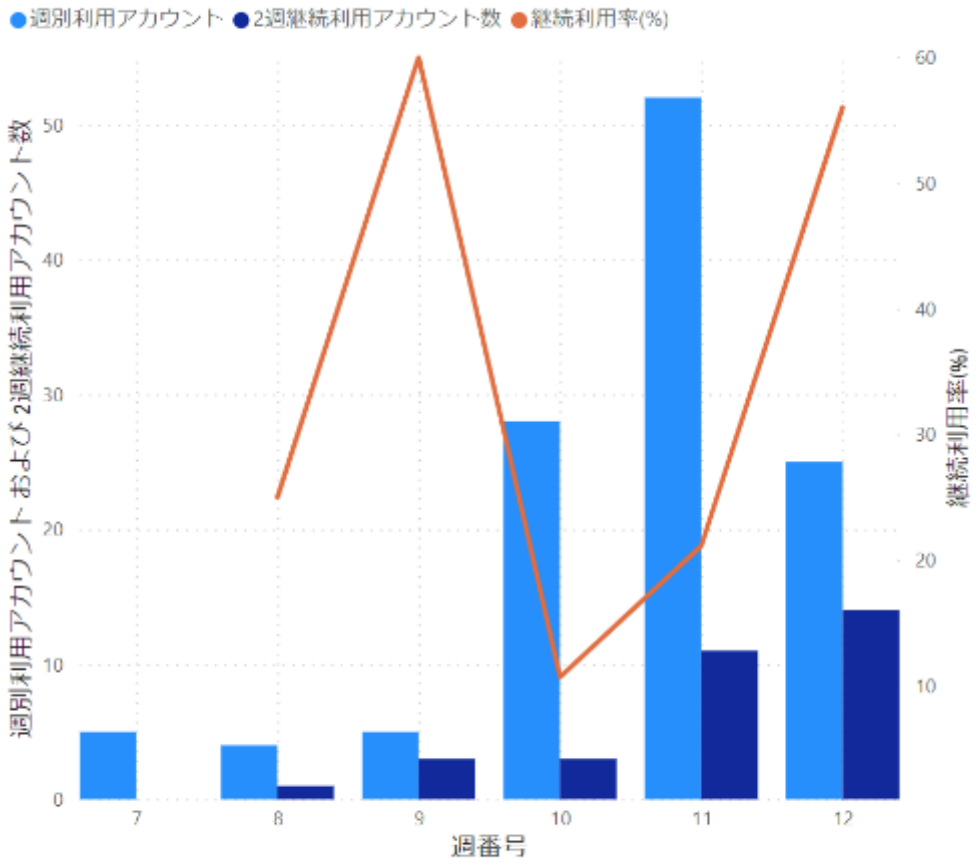
3. 継続利用者数(週次)自治体

週別利用継続者の内訳

※週に1度でもリクエストをしたアカウントを対象に集計をしています。

週番号	開始日	週別利用アカウント	2週継続利用アカウント数	継続利用率(%)
7	2024/02/12 0:00:00	5		
8	2024/02/19 0:00:00	4	1	25.00
9	2024/02/26 0:00:00	5	3	60.00
10	2024/03/04 0:00:00	28	3	10.71
11	2024/03/11 0:00:00	52	11	21.15
12	2024/03/18 0:00:00	25	14	56.00

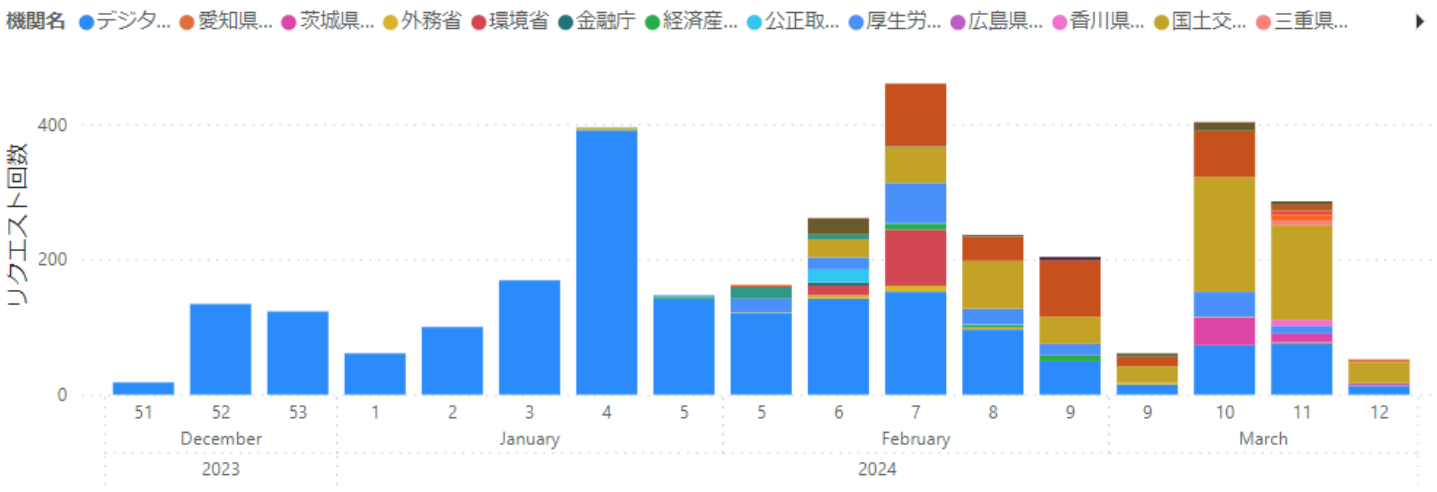
週別利用継続者の内訳



4. 学習セット利用回数(週次)全体

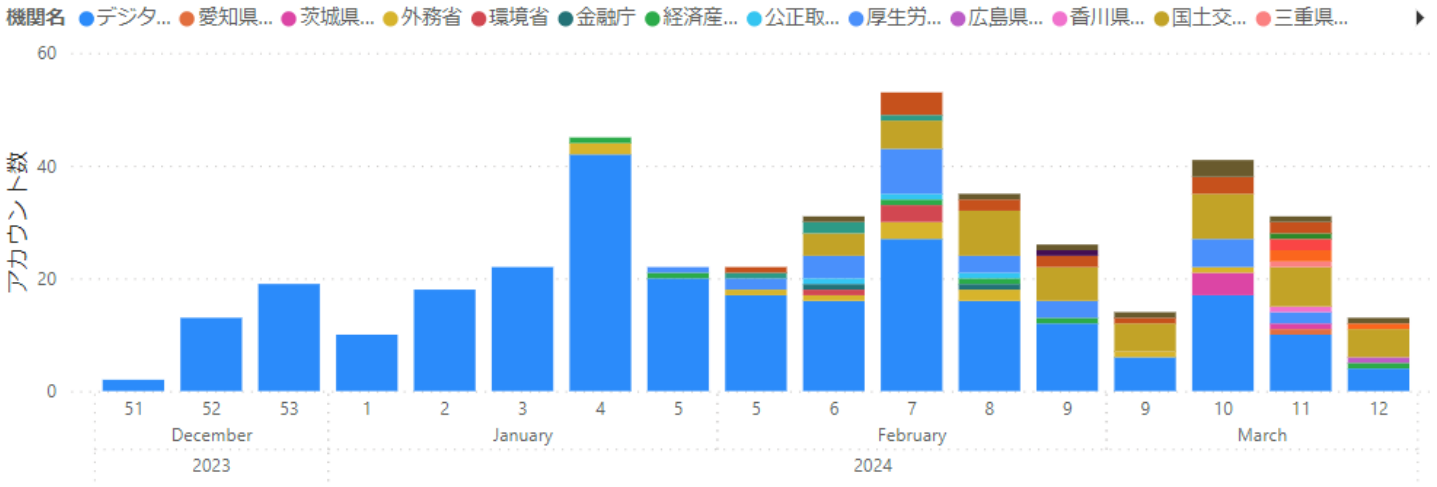
週番号	StartOfWeek	利用回数
1	2024/01/01 0:00:00	61
2	2024/01/08 0:00:00	100
3	2024/01/15 0:00:00	169
4	2024/01/22 0:00:00	395
5	2024/01/29 0:00:00	309
6	2024/02/05 0:00:00	261
7	2024/02/12 0:00:00	460
8	2024/02/19 0:00:00	236
9	2024/02/26 0:00:00	265
10	2024/03/04 0:00:00	403
11	2024/03/11 0:00:00	286
12	2024/03/18 0:00:00	52
51	2023/12/11 0:00:00	18
52	2023/12/18 0:00:00	134
53	2023/12/25 0:00:00	123
合計		**3272**

日別学習セット利用回数



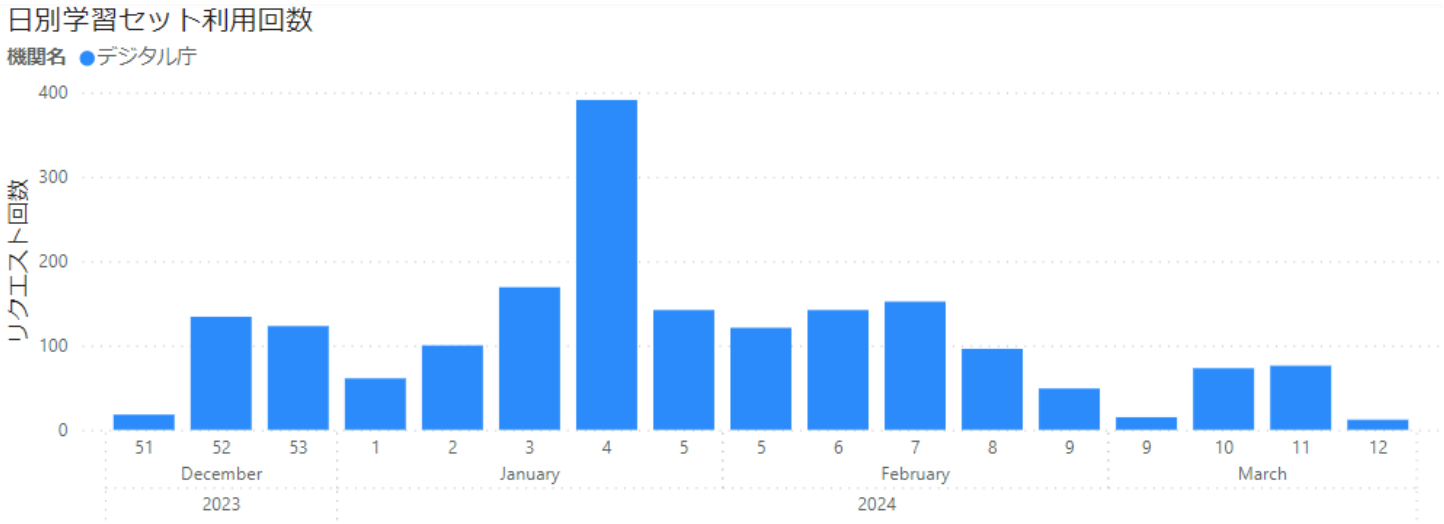
週番号	StartOfWeek	利用アカウント
1	2024/01/01 0:00:00	10
2	2024/01/08 0:00:00	18
3	2024/01/15 0:00:00	22
4	2024/01/22 0:00:00	45
5	2024/01/29 0:00:00	34
6	2024/02/05 0:00:00	31
7	2024/02/12 0:00:00	53
8	2024/02/19 0:00:00	35
9	2024/02/26 0:00:00	36
10	2024/03/04 0:00:00	41
11	2024/03/11 0:00:00	31
12	2024/03/18 0:00:00	13
51	2023/12/11 0:00:00	2
52	2023/12/18 0:00:00	13
53	2023/12/25 0:00:00	19
合計		**157**

日別学習セット利用アカウント数

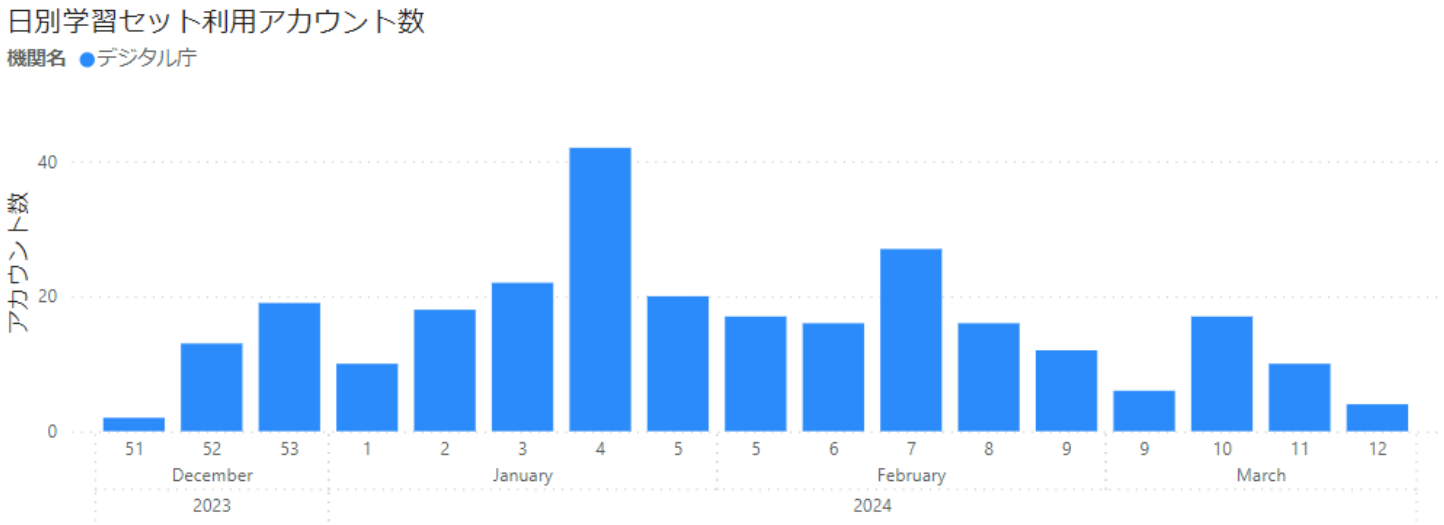


4. 学習セット利用回数(週次)デジタル庁

週番号	StartOfWeek	利用回数
1	2024/01/01 0:00:00	61
2	2024/01/08 0:00:00	100
3	2024/01/15 0:00:00	169
4	2024/01/22 0:00:00	395
5	2024/01/29 0:00:00	309
6	2024/02/05 0:00:00	261
7	2024/02/12 0:00:00	460
8	2024/02/19 0:00:00	236
9	2024/02/26 0:00:00	265
10	2024/03/04 0:00:00	403
11	2024/03/11 0:00:00	286
12	2024/03/18 0:00:00	52
51	2023/12/11 0:00:00	18
52	2023/12/18 0:00:00	134
53	2023/12/25 0:00:00	123
合計		**3272**



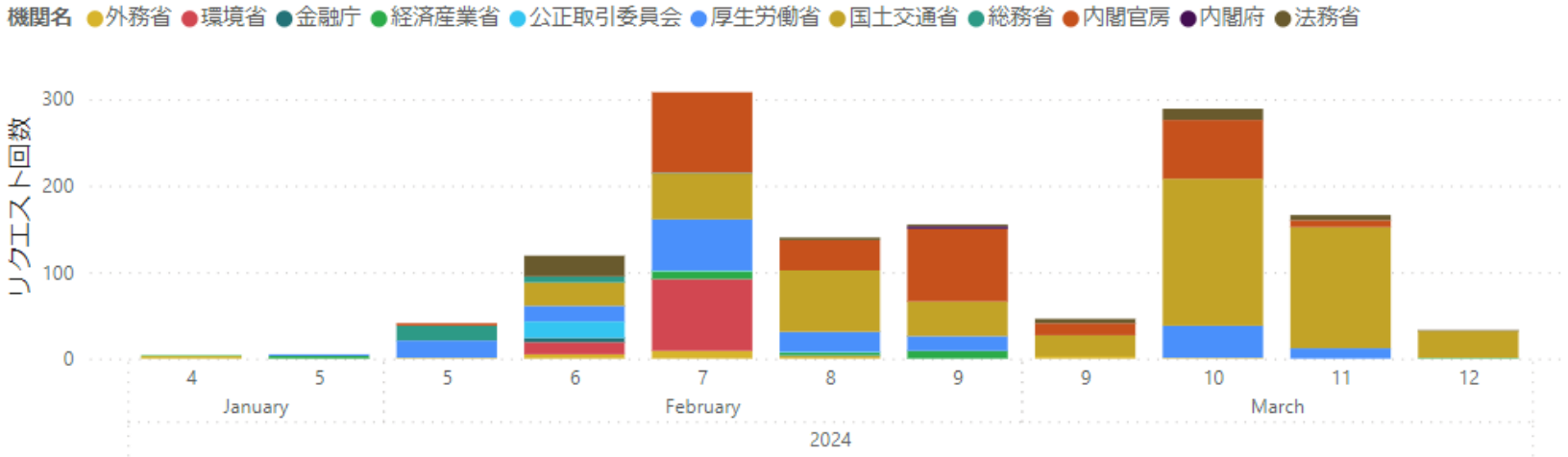
週番号	StartOfWeek	利用アカウント
1	2024/01/01 0:00:00	10
2	2024/01/08 0:00:00	18
3	2024/01/15 0:00:00	22
4	2024/01/22 0:00:00	45
5	2024/01/29 0:00:00	34
6	2024/02/05 0:00:00	31
7	2024/02/12 0:00:00	53
8	2024/02/19 0:00:00	35
9	2024/02/26 0:00:00	36
10	2024/03/04 0:00:00	41
11	2024/03/11 0:00:00	31
12	2024/03/18 0:00:00	13
51	2023/12/11 0:00:00	2
52	2023/12/18 0:00:00	13
53	2023/12/25 0:00:00	19
合計		**157**



4. 学習セット利用回数(週次)他府省庁

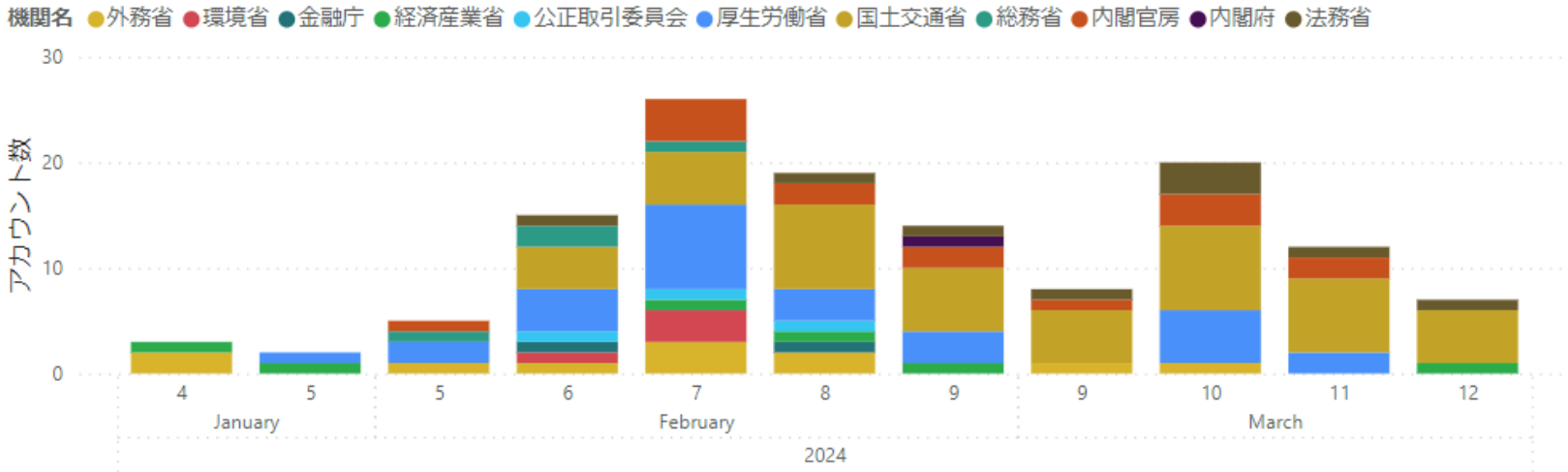
週番号	StartOfWeek	利用回数
4	2024/01/22 0:00:00	4
5	2024/01/29 0:00:00	46
6	2024/02/05 0:00:00	119
7	2024/02/12 0:00:00	308
8	2024/02/19 0:00:00	140
9	2024/02/26 0:00:00	201
10	2024/03/04 0:00:00	289
11	2024/03/11 0:00:00	166
12	2024/03/18 0:00:00	33
合計		**1306**

日別学習セット利用回数



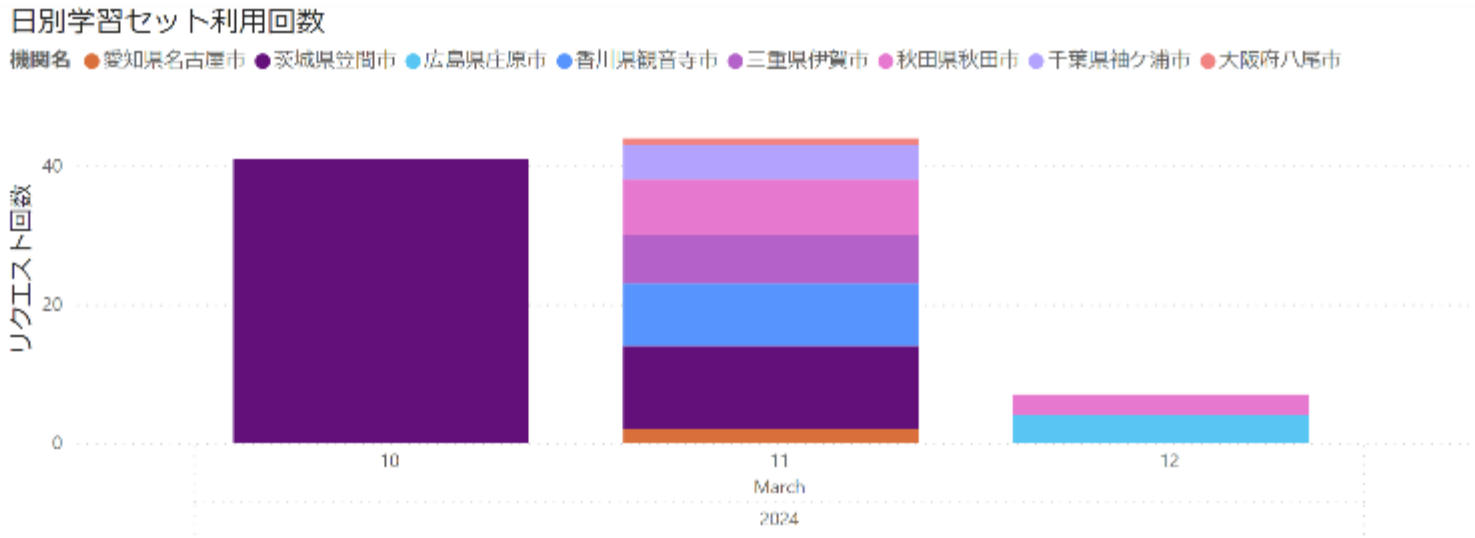
週番号	StartOfWeek	利用アカウント
4	2024/01/22 0:00:00	3
5	2024/01/29 0:00:00	6
6	2024/02/05 0:00:00	15
7	2024/02/12 0:00:00	26
8	2024/02/19 0:00:00	19
9	2024/02/26 0:00:00	20
10	2024/03/04 0:00:00	20
11	2024/03/11 0:00:00	12
12	2024/03/18 0:00:00	7
合計		**50**

日別学習セット利用アカウント数

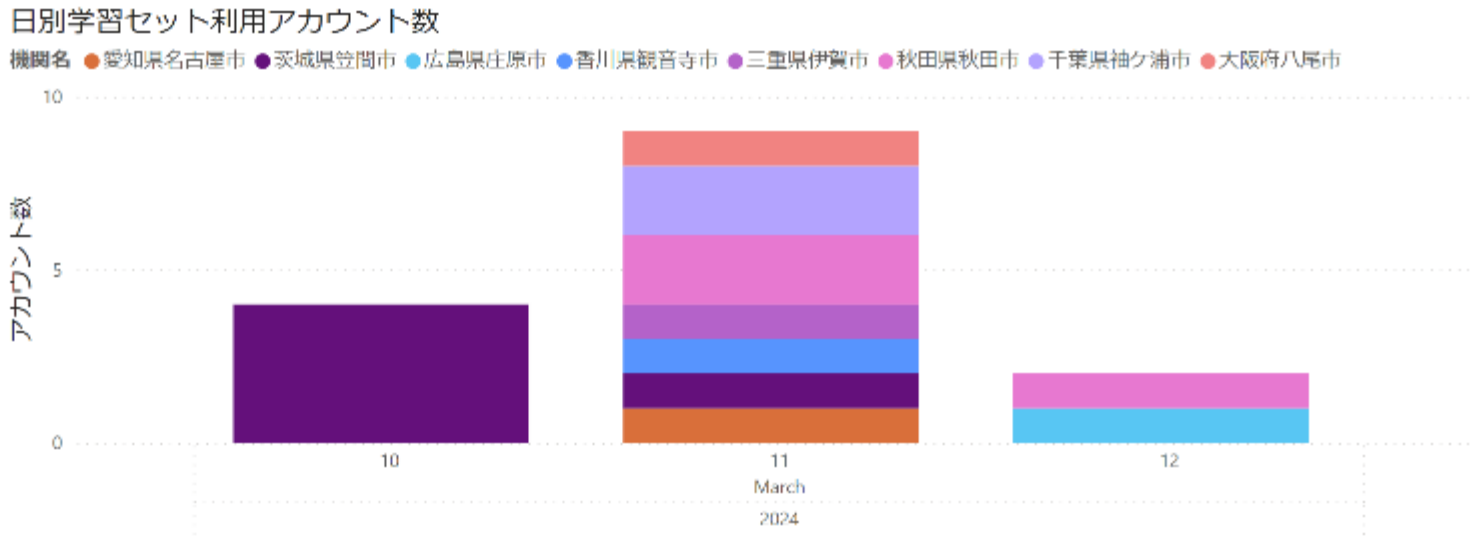


4. 学習セット利用回数(週次)自治体

週番号	StartOfWeek	利用回数
10	2024/03/04 0:00:00	41
11	2024/03/11 0:00:00	44
12	2024/03/18 0:00:00	7
合計		**92**



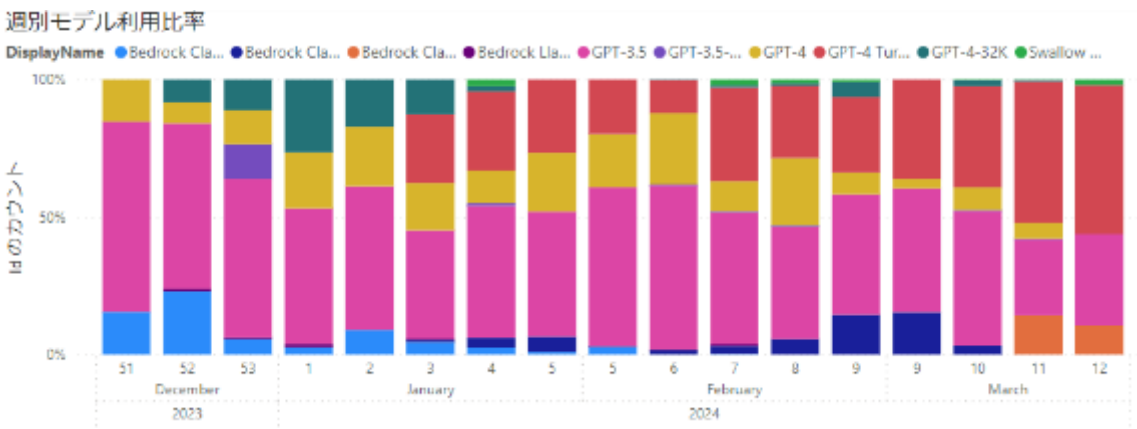
週番号	StartOfWeek	利用アカウント
10	2024/03/04 0:00:00	4
11	2024/03/11 0:00:00	9
12	2024/03/18 0:00:00	2
合計		**13**



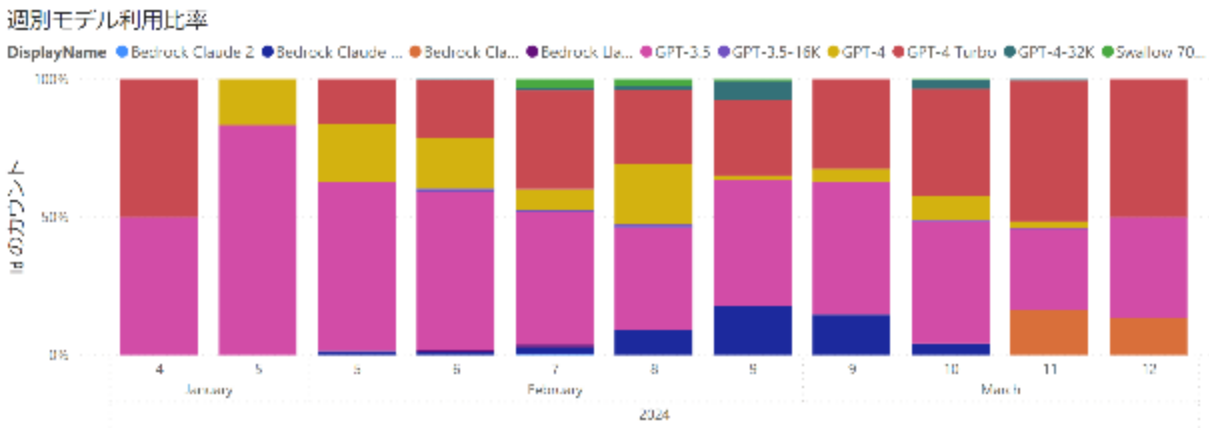
5. モデル別利用比率(週次)

- ▶ 全体として、GPT系の利用が活発であった。その中でもGPT3.5が多く利用された。3月に入るとGPT4-Turboの利用が増加
- ▶ 中央省庁でのCloude2.1の利用多数

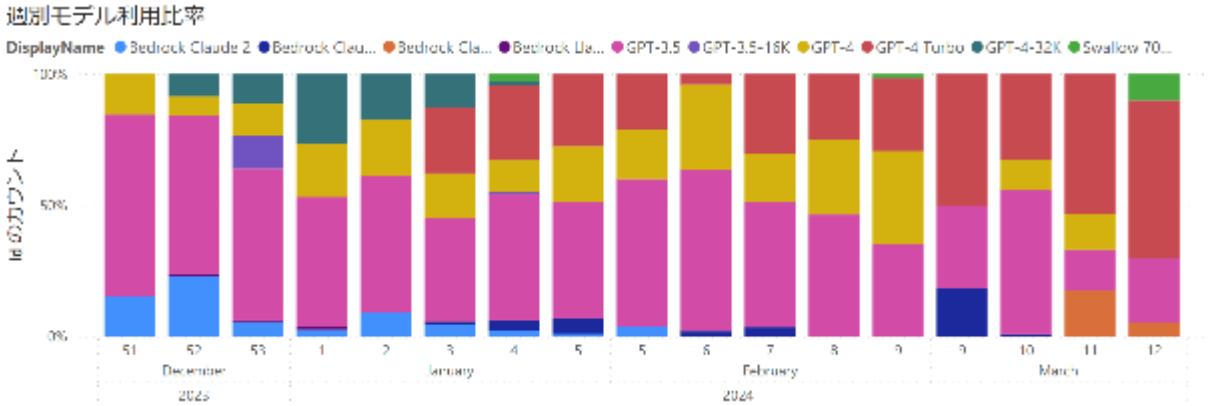
全体



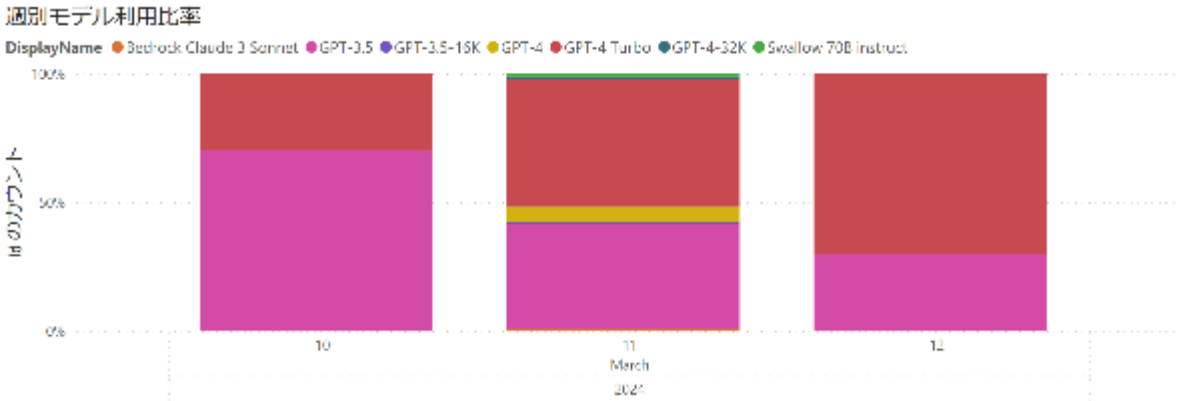
他府省庁



デジタル庁



自治体

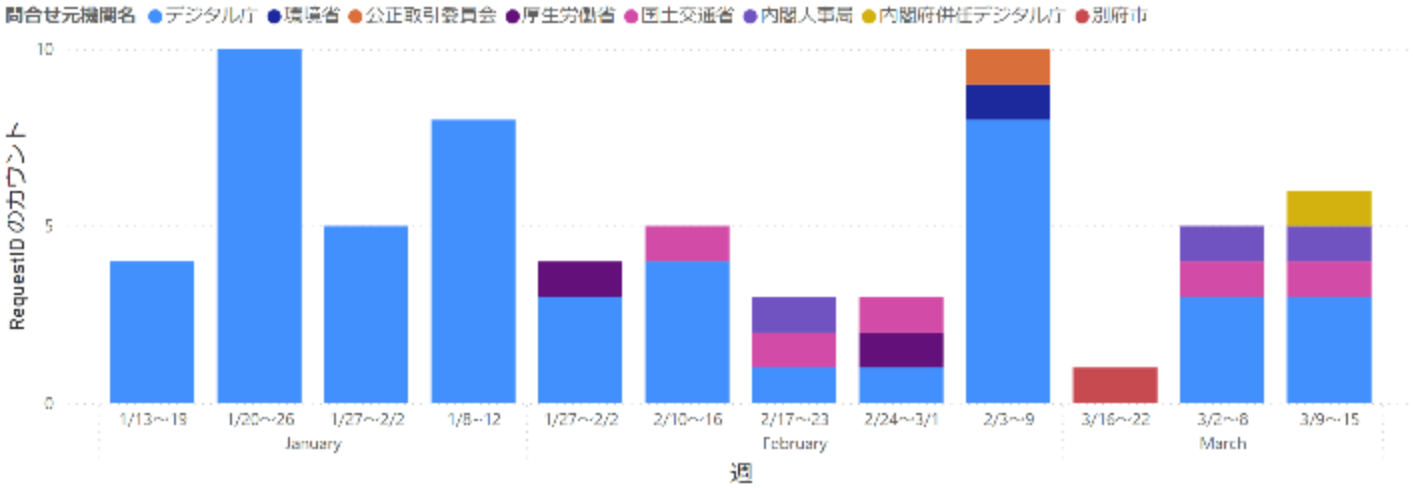


6. 問い合わせ内容(1/2)

- ▶ 問合せ件数はデジタル庁様(50件)、 次いで国土交通省様(5件)
- ▶ 問合せ種別はGaiXer利用方法が多く(27件)、「テンプレートの質問項目の使い方」から「プロンプトの記載方法」まで幅広い質問に対応

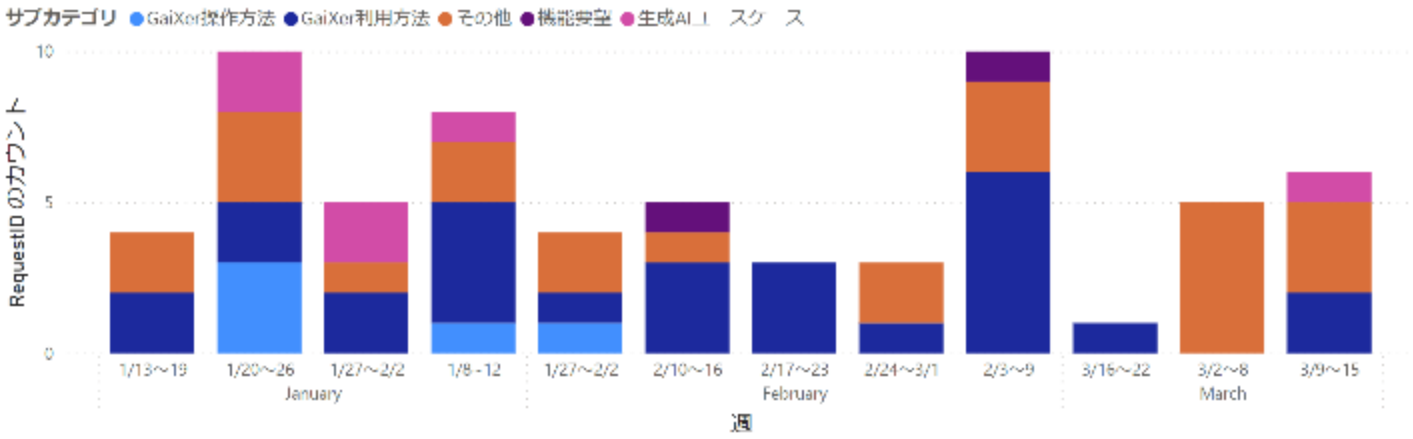
問合せ元機関名	問合せ件数
デジタル庁	50
国土交通省	5
内閣人事局	3
厚生労働省	2
環境省	1
公正取引委員会	1
内閣府併任デジタル庁	1
別府市	1
合計	64

週次_機関別お問合せ件数



サブカテゴリ	問合せ件数
⊟ GaiXer利用方法	27
その他	16
プロンプトの書き方	7
学習データの使い方	4
⊞ その他	24
⊞ 生成AIユースケース	6
⊞ GaiXer操作方法	5
⊞ 機能要望	2
合計	64

週次_種類別お問合せ件数



6. 問い合わせ内容(2/2)

▶ カテゴリ別の問い合わせ内容例

#	サブカテゴリ	項目	問い合わせ内容例
1	GaiXer 利用方法	学習データの使い方	海外の論文を学習データとして作成したのですが、参照されません。参照できるプロンプトの書き方について教えてください。
		プロンプトの書き方	テンプレートのリクエストとシステムプロンプトはどのように使い分ければいいでしょうか
2	GaiXer 操作方法	ー	全リクエスト一括削除は可能でしょうか
3	機能要望	ー	GPTからの戻り値でコードブロックが含まれるとき、そこだけをコピーする方法はありますか
4	生成AI ユースケース	ー	GaiXerを用いて、Excelシート上にあるテキストの分類を行いたい
5	その他	ー	GPT-3.5よりGPT-4Turboの方が回答速度が遅い気がしますが、スピードは3.5の方が早いのでしょうか

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

リリースノート(1/5)

▶ プロジェクト開始後、下記のリリースを実施した。

#	日時	ユーザー影響	変更対象	変更内容
1	2023/12/4	あり	アプリ＆インフラ	お客様提供用環境の構築
2	2023/12/8	あり	アプリ&インフラ	提供するLLMとしてGPT-4, GPT-4-32k, Bedrock Llama 2 Chat 13B, Bedrock Claude 2を追加
3	2023/12/11	なし	アプリ	ユーザ情報取得APIを正しくコールするように修正
4	2023/12/19	なし	インフラ	Admin-PortalのContainerApps hostedservice-blueのスケールアップ作業
5	2023/12/20	あり	アプリ	新UIの取り込み
6	2023/12/21	なし	インフラ	新UIのトピック生成用 ContainerAppsを構築
7	2023/12/21	なし	インフラ	新UIのトピック生成用 AzureFunctionsを構築
8	2023/12/26	あり	アプリ	LLM選択箇所の増加・各回答に利用したLLMの表示対応、軽微なレイアウト改修
9	2023/12/26	あり	アプリ	一部のテンプレートで編集ができない不具合の修正対応
10	2023/12/28	あり	アプリ	Admin Portal上でユーザーをCSVで取得した際に、管理者情報が正しく取得されない不具合を修正
11	2023/12/28	あり	アプリ	通常の生成とテンプレート作成時の「試しに生成」時にプロンプトの送信方法が違うことで、出力結果に著しい差が出てしまう不具合の修正。プロンプト設定の追加
12	2023/12/28	あり	アプリ	パスワードリセット画面にあるパスワードの必須形式について記載追加
13	2023/12/28	あり	アプリ	登録時に確認コードを自動送信する際の電子メールが空メッセージになっている不具合を解消
14	2024/01/09	あり	アプリ	ログイン直後にフリーズする問題の解決 (要約元のテキスト量が多く、履歴表示の途中でエラーが発生)

リリースノート(2/5)

▶ **プロジェクト開始後、下記のリリースを実施した。**

#	日時	ユーザー影響	変更対象	変更内容
15	2024/01/12	あり	アプリ&インフラ	• 選択LLM毎の文字数制限の追加 • 「申し訳ありません」のテキストが繰り返し表示される事象の修正 • 提供するLLMとしてGPT-4 Turbo, Bedrock Claude 2.1 を追加 • テンプレート、トピックの左右にボタンを追加
16	2024/01/19	あり	アプリ&インフラ	• 国産LLMのSwallowの利用追加 • テンプレートに生成に使用するモデルを設定可能 • テンプレートにシステムプロンプトを設定可能
17	2024/01/25	なし	インフラ	• Swallowの基盤設定の見直し
18	2024/01/26	あり	アプリ&インフラ	•Playground機能の追加 •Enter：送信、Ctrl+Enter/Shift+Enter：改行 •入力文字数制限の緩和 ■不具合対応事項 •テンプレートの「試しに生成」ボタンのアクティブ条件改善 •テンプレートの「試しに生成」エリアの入力項目アクティブ制御の改修 •テンプレートにXMLタグを入力すると、保存後の再表示時にXMLタグが消える事象の改修

19	2024/02/13	あり	アプリ	• Playground (step2)
				  • アシスタントテンプレート書き出し
				  • アシスタント作成・編集・削除・エクスポート
				• 入力文字数制限の緩和 (step3)
				  • 生成結果のトークン数がオーバーする場合のメッセージ表示
				• ニュースのカテゴリ表示

リリースノート(3/5)

▶ **プロジェクト開始後、下記のリリースを実施した。**

#	日時	ユーザー影響	変更対象	変更内容
20	2024/02/22	あり	アプリ	• Playground (step3) 　　• ワークスペース
21	2024/02/29	あり	インフラ	• メールOTP
22	2024/03/01	あり	アプリ	■機能追加 • テンプレート：学習セットの紐付け • テンプレート：プロンプトの非表示オプション ■改善 • PDFファイルの学習データ登録時、画像を学習対象にしないオプションを追加 • テンプレート機能：システムプロンプトの永続適用 • 「申し訳ありません」をフロントにて繰り返し表示している事象の改善 • Playgroundから作成したテンプレートと判断できるUI項目の追加 ■不具合対応 • Swallowの生成結果が意図しないフォーマットで生成される • GPT4-Turbo利用時、特定のプロンプトでトークン超過が発生する (「National Security by Platformという2021年のpaperについて日本語で教えて」)
23	2024/03/01	なし	インフラ	■セキュリティ強化(要保護対象情報の保護) • リソース設定の改ざん、DB・ストレージ・サーバー等への不正アクセスを検知・通知する、プラットフォーム事業者提供機能の有効化・設定 ■脆弱性試験向け事前準備 • 脆弱性試験でイエラエ社が使用するIPアドレスのWAF通信許可設定

リリースノート(4/5)

▶ **プロジェクト開始後、下記のリリースを実施した。**

#	日時	ユーザー影響	変更対象	変更内容
24	2024/3/5	あり	アプリ、インフラ	■改善 • ユーザーが明示的にWeb検索をしないオプションを追加 • Web検索する際にドメインを絞った検索に対応 • 法例検索機能を実装 • 共通設定項目の追加 • リクエストフォームの追加 ■セキュリティ強化 • 脆弱性診断指摘事項への対応
25	2024/3/8	あり	アプリ	• 制限チェッカー機能の追加 • Claude 3 の追加 • サイト指定時、会話履歴が効かない不具合の修正 • 文言の改修(「用例による法令検索」、「検索サイト指定」、「Web検索をRAGに利用しない」)
26	2024/3/8	あり	インフラ	• 学習セットのリソースのバージョンアップ等、安定化およびセキュリティ向上のためのインフラ作業
27	2024/03/15		アプリ	• ダッシュボードで週単位での利用継続率可視化
28	2024/03/15		アプリ	• Slackとの連携機能の追加(Slackにて実際に利用可能となるのはデロイト社での作業後)
29	2024/03/18		アプリ	• Azure OpenAI APIバージョンの変更 • 制限チェッカーのプロンプト修正 • B2Cの新規登録がうまくいかない問題の修正

リリースノート(5/5)

▶ プロジェクト開始後、下記のリリースを実施した。

#	日時	ユーザー影響	変更対象	変更内容
30	2024/03/22		アプリ	• Slack App版GaiXerの更新 ・任氏入力項目に任意を追加表示 ・利用したモデルを追加表示 ・Web検索ON/OFFの制御改善 ・テンプレート使用時、ユーザの入力内容が他のユーザから見えるように改善

改修事項の対応状況

▶ 改修対応の全量(要望リスト＋α、不具合、脆弱性診断の指摘内容)を以下に示す。

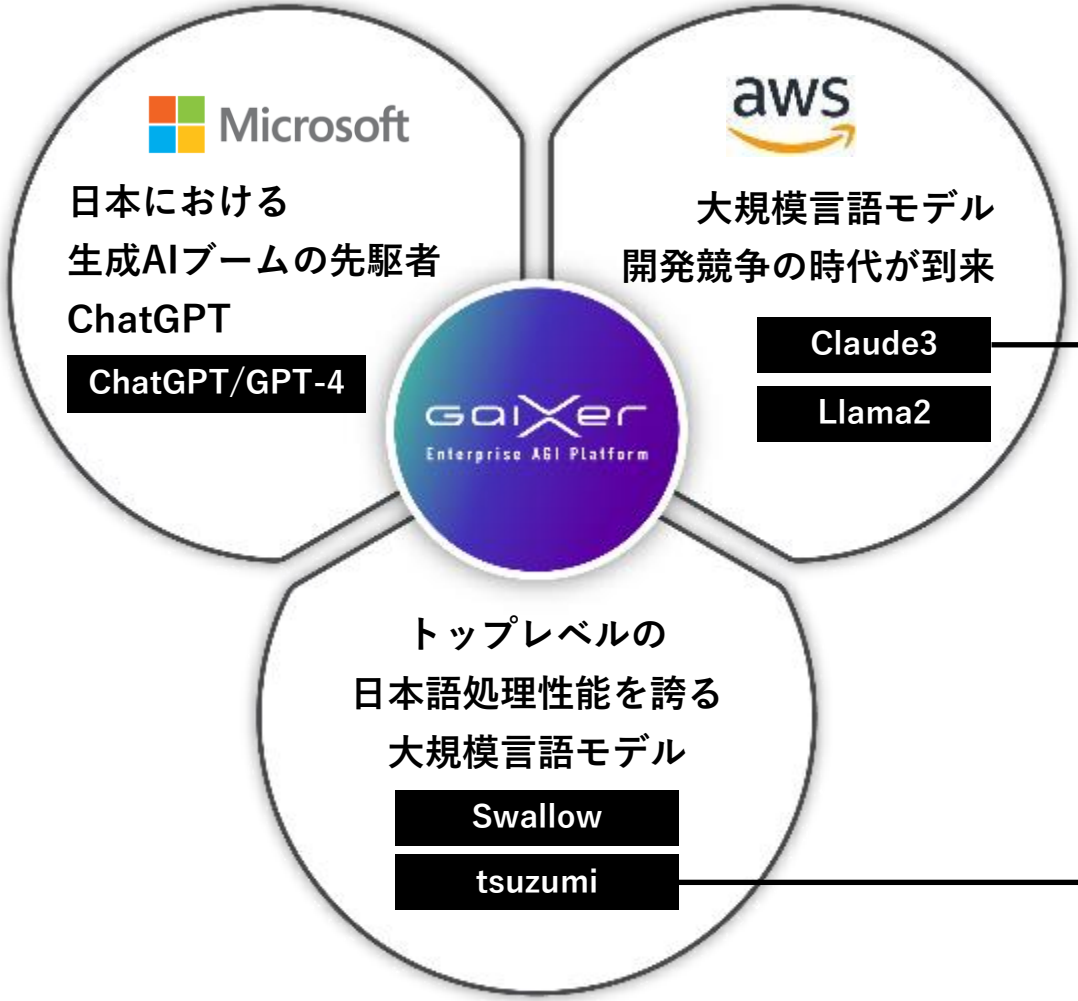
対応完了：

#	要望リスト＋α
d-215	認証に以下の機能を設ける。 1.SSOとID/PASS認証の併用 2.SSOできる省庁はSSO強制(ID/PASS不可) 3.ID/PASS認証を守る対策を設ける(メールOTPや、利用者セッション確認機能) ※メールOTPの発火タイミングは以下の２つ。 ・初回ログイン(管理者経由で初期PWDが流出するリスクを軽減) ・別デバイス、別ブラウザでのログイン時(意図しない端末による不正ログイン抑止)
d-217	ユーザーが明示的にWeb検索をしないオプションを作ってください
d-041	Web検索するさいにドメインを絞った検索に対応してほしい
d-031	用例による法令検索
α	共創プラットフォームとSlackの連携
α	制限チェッカー
α	ダッシュボードで週単位での利用継続率可視化
α	制限チェッカー機能に関して、プロンプトの調整前後で機密性情報の抽出率変化を閲覧できる機能をダッシュボードに追加する→3/18に対応済み
α	OpenAIのAPIバージョン更新→3/18に対応済み

#	不具合
1	デジタル庁の一部のユーザーにおいて学習セット登録ができない
2	Admin Portal 関連のリソースで再起動が発生し、学習セットの登録が行えない

#	脆弱性診断(Web診断)での指摘
1	[低] X-Frame-Options ヘッダの推奨事項
2	[低] 長いセッション保持期間
3	[低] キャッシュ制御の不備
4	[低] プロンプトインジェクション→対象外
5	[低] 不用意な情報公開
6	[低] 任意ファイルのアップロードが可能
7	[情報] X-Content-Type-Options ヘッダの推奨事項
8	[低] X-Frame-Options ヘッダの推奨事項
9	[低] 長いセッション保持期間
10	[低] ソースコード内でのパスワードの保持→対象外
11	[低] キャッシュ制御の不備
12	[低] アカウントロックアウトの欠如→対象外
13	[低] エラーメッセージによる情報推測
14	[低] オープンリダイレクト
15	[低] メール本文の改竄が可能
16	[情報] X-Content-Type-Options ヘッダの推奨事項
17	[情報] パスワードオートコンプリートが有効→対象外
18	[情報] ログアウト機能の不備
19	[情報]適切でないCross Origin Resource Sharing ポリシー設定

生成AIプラットフォームGaiXer コンセプト



GaiXer Enterprise AGI Platform は業務に最適なLLMを順次追加搭載

複数LLMのいいとこどり　が可能

2021年にOpenAIの元共同設立者によって設立されたAnthropic社が開発。
ChatGPTと比較して、より安価だが約1.5倍の長さの文章を扱える高性能LLM

NTTの長年にわたる自然言語処理技術の研究成果を基に開発されたLLMであり、
高い日本語の自然言語処理能力を持ちつつ、低消費電力で動作することが特徴

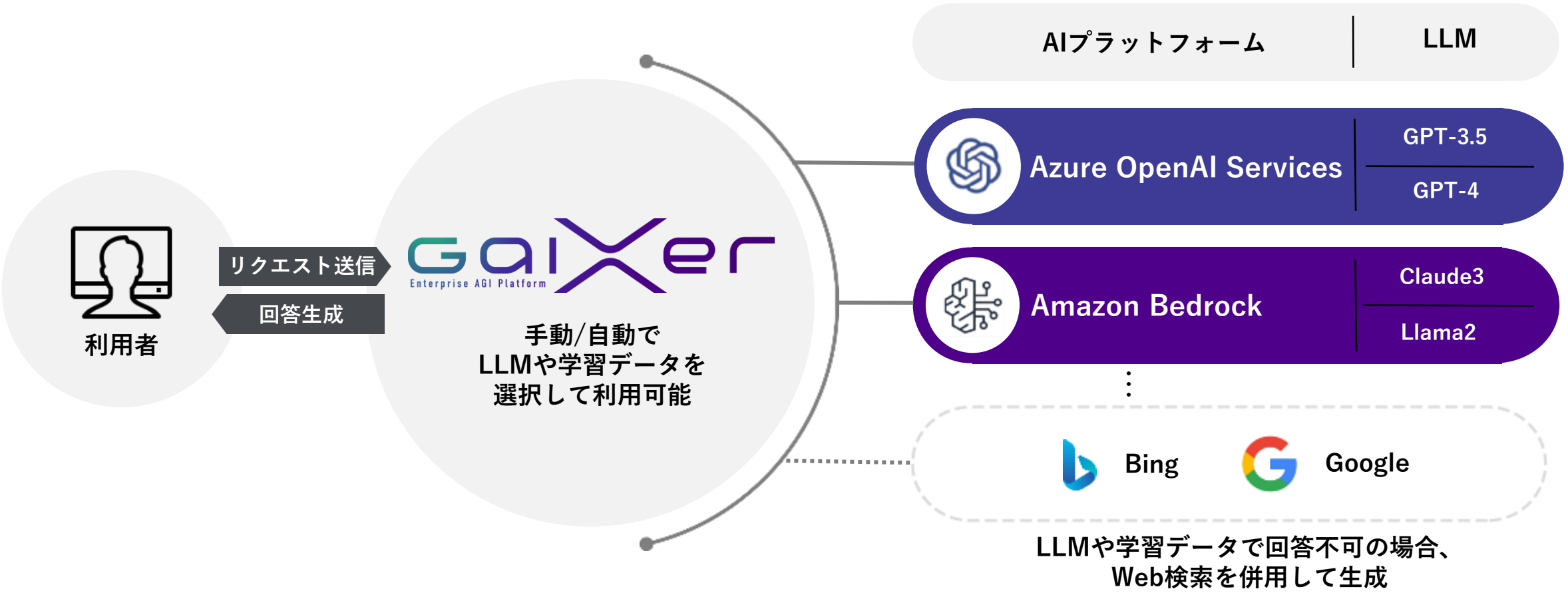
※Microsoft、Azureは、米国Microsoft Corporationの米国およびその他の国における登録商標または商標です。
※Amazon Web Services 、AWS は、Amazon.com, Inc. またはその関連会社の商標です。
※その他記載の会社名、製品名は、各社の商標または登録商標です。

当初からの提供機能: 専用環境

▶ GaiXerの全てのリソースをクラウド上の専用環境でご提供するため高いセキュリティレベルを確保

当初からの提供機能:マルチクラウド

▶ マルチクラウドにより、利用目的に合わせて適切なLLMを選択肢、高い生成品質を実現



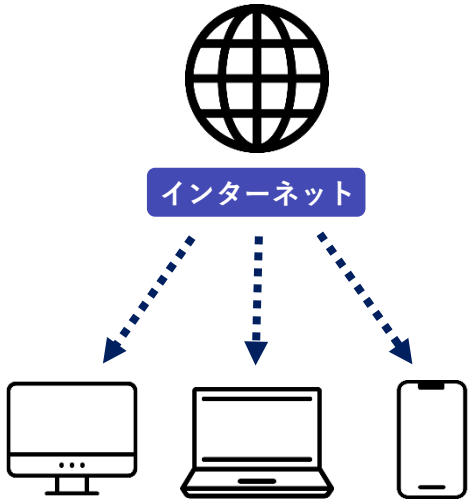
※ Microsoft 、Azure 、Azure OpenAI Services 、Bing は、米国Microsoft Corporationの米国およびその他の国における登録商標または商標です。
※ Amazon Web Services 、AWS 、Amazon Bedrock は、Amazon.com, Inc. またはその関連会社の商標です。
※ Googleは、 Google LLC の商標です。

当初からの提供機能: RAG・プロンプトテンプレート・データ学習

▶ 生成品質を高めるため、RAG・プロンプトテンプレート・データ学習機能を提供

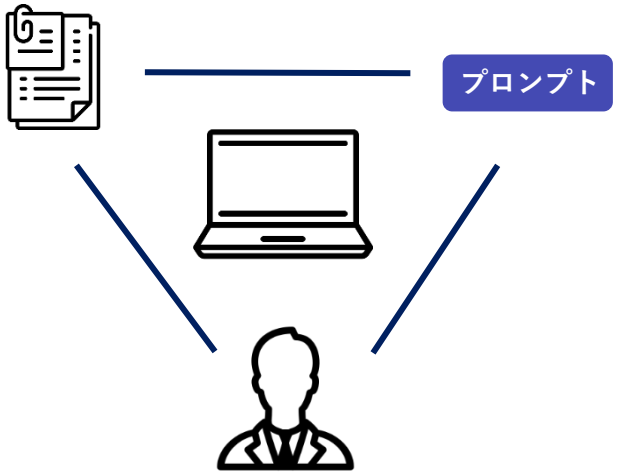
RAG

LLMだけで回答生成できない場合、GoogleやBingなどの検索エンジンと自動連携、最新情報に基づき回答



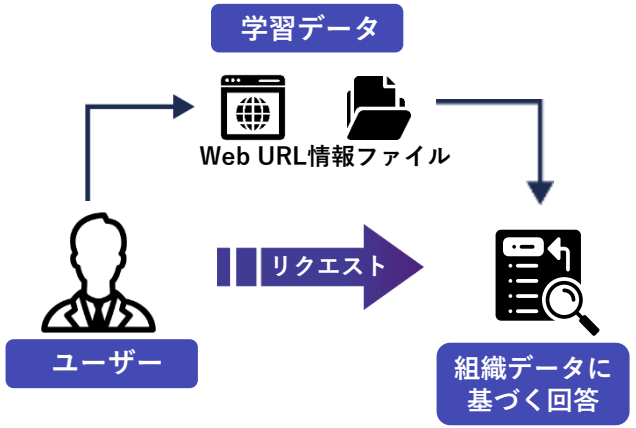
プロンプトテンプレート

様々な業務タスクを実行できるプロンプトテンプレートが利用でき、お客様独自のテンプレート作成・共有も可能



データ学習機能

Webサイトやマニュアル・FAQなどを追加学習可能。テキスト・Excel・PowerPoint・PDFなど、様々なファイル形式に対応



注: RAG = Retrieval-Augmented Generation; 外部の知識ベースを検索し、最新の正確な情報に基づいて大規模言語モデル(LLM)に回答を生成させること

追加提供機能: 国産LLM「Swallow」

概要	国産のLLMである「Swallow」を使用することが可能になった。英語の言語理解に長けるLLM「Llama 2」の日本語能力を拡張したもので、日本語の文章の読み込みや生成を得意としている。（※この後紹介するテンプレート作成時のモデル選択でも選択することが可能である。）

モデル変更の画面で選択できる。

追加提供機能: テンプレート作成時のモデル選択機能

概要	テンプレート作成時にモデルを選択することが可能となった。これにより、テンプレートの内容を生成する際にそれぞれのテンプレートに適しているモデルで生成することができる。

こちらのボタンからモデルを変更できる。

追加提供機能: システムプロンプト機能

概要	テンプレート作成時にシステムプロンプトの編集が可能となった。 これにより、生成するときにシステムプロンプトを使用できる。

こちらからシステム
プロンプトを設定できる。

追加提供機能: 入力できる文字数の増加

概要	「リクエスト」欄や「システムプロンプト」欄に入力できる文字数が増加。 これにより以前よりも長い文章を入力可能。

こちらから文字数を
確認できる。

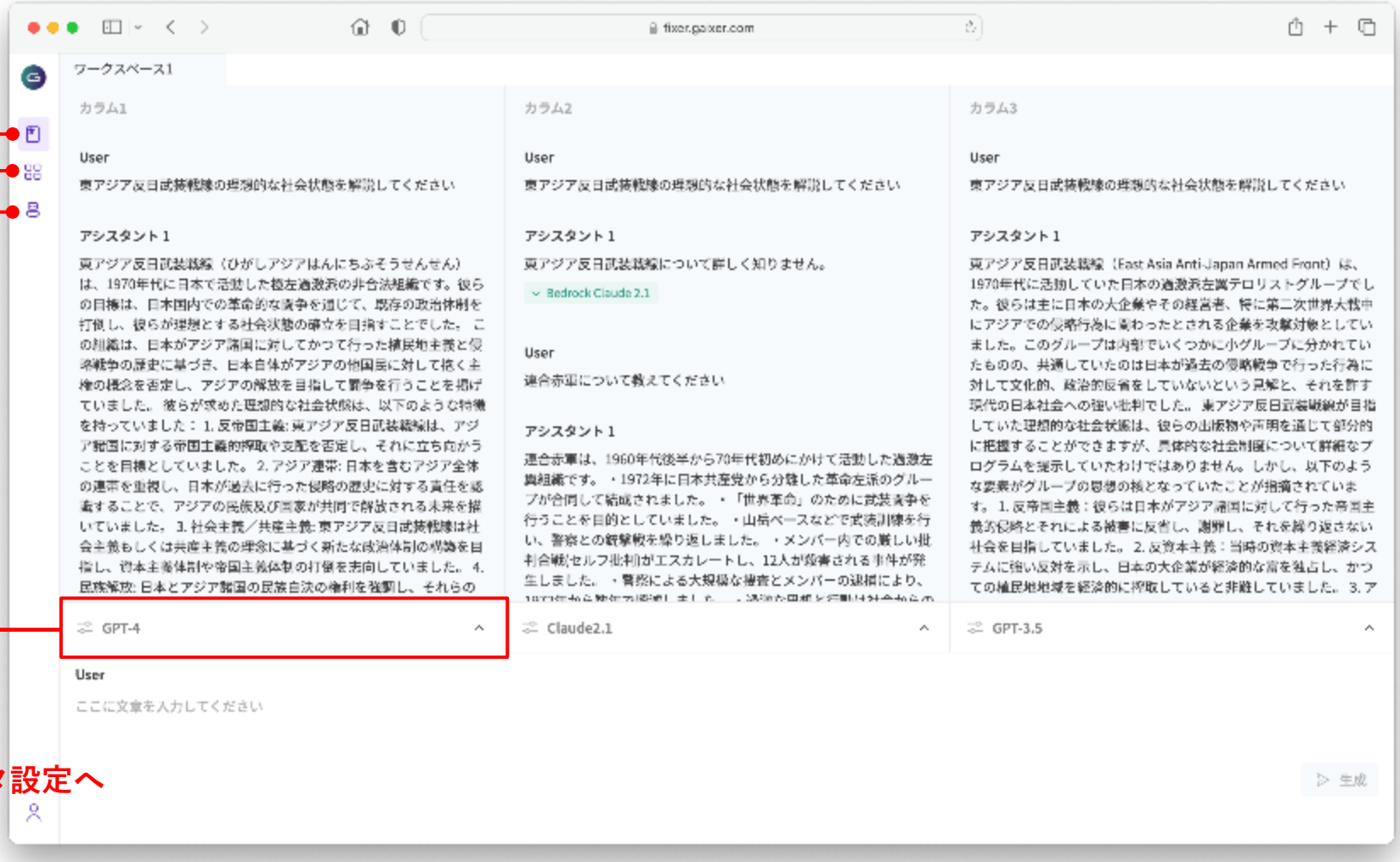
追加提供機能: PlayGround

概要	言語モデル（LLM）を複数用いて同時生成し、比較検証できる機能

Playground
言語モデルを比較画面を
表示します

ワークスペース
生成結果を作業場として
分ける機能

アシスタントリスト
各カラムのパラメータ設定は
「アシスタント」として保存可能



押下でパラメータ設定へ

追加提供機能: アシスタントリスト

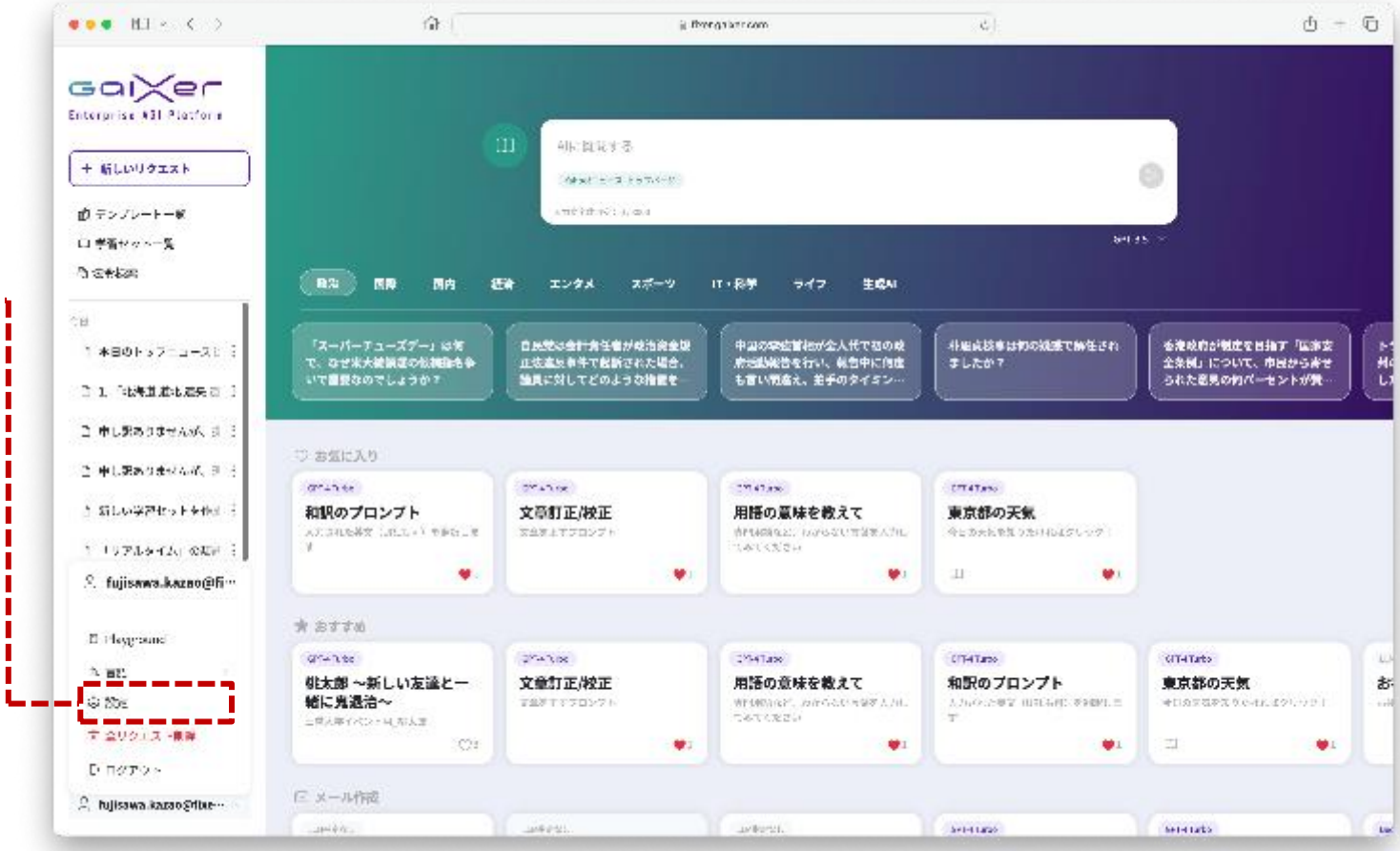
テンプレートとして書き出し
パラメータ設定を
「テンプレート新規作成」
へダイレクトに反映

アシスタントを削除

編集
パラメータへアクセスし調節が可能

追加提供機能: Web検索設定

▶ ユーザーメニューの「設定」からRAG機能の内、Web検索をしてデータ取得する動作を無効化することが可能
ここでの設定は、プロンプトテンプレート利用には影響ない



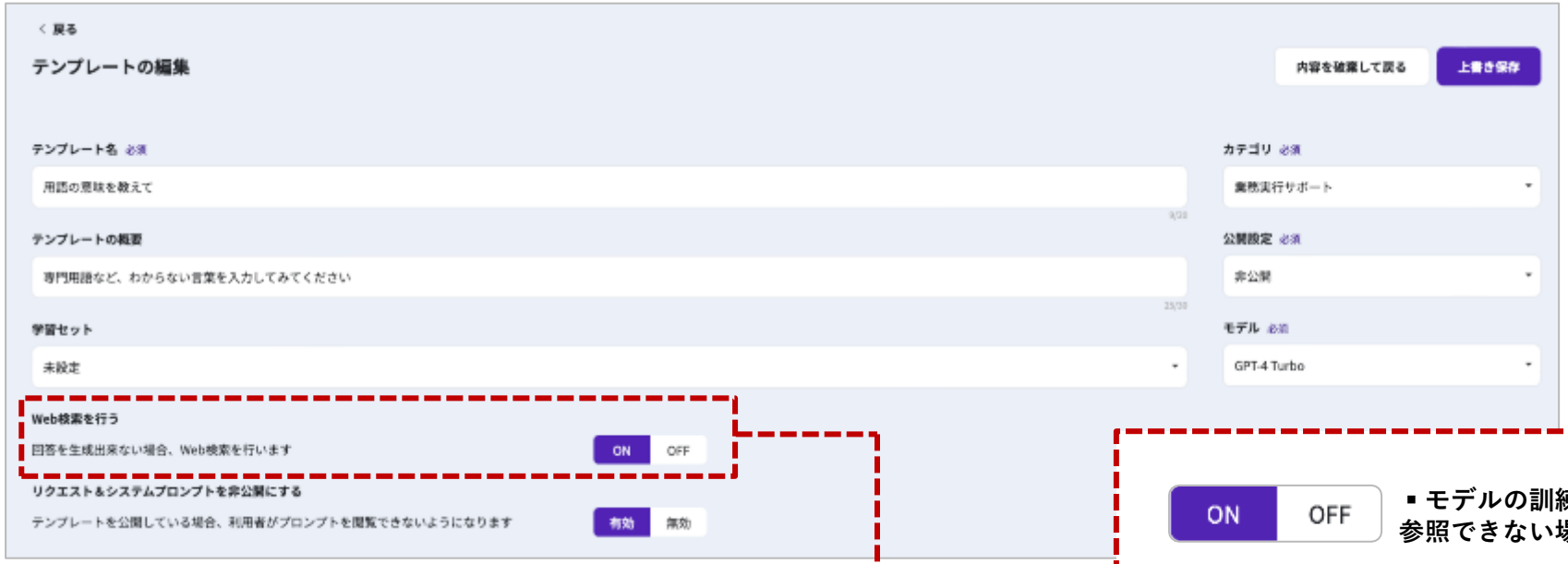
- **スイッチOFF**
  リクエストに対して、回答をより良くするために、必要に応じてWeb検索を自動で実行する。

- **スイッチON**
  リクエストに対して、Web検索を実行しない。

追加提供機能: プロンプトテンプレートにおけるWeb検索指定

▶ 「プロンプトテンプレート」では、新規テンプレートの作成や既存テンプレートの編集時に、回答を生成できない場合Web検索機能をオンまたはオフが可能

▪ テンプレートの新規作成/編集画面



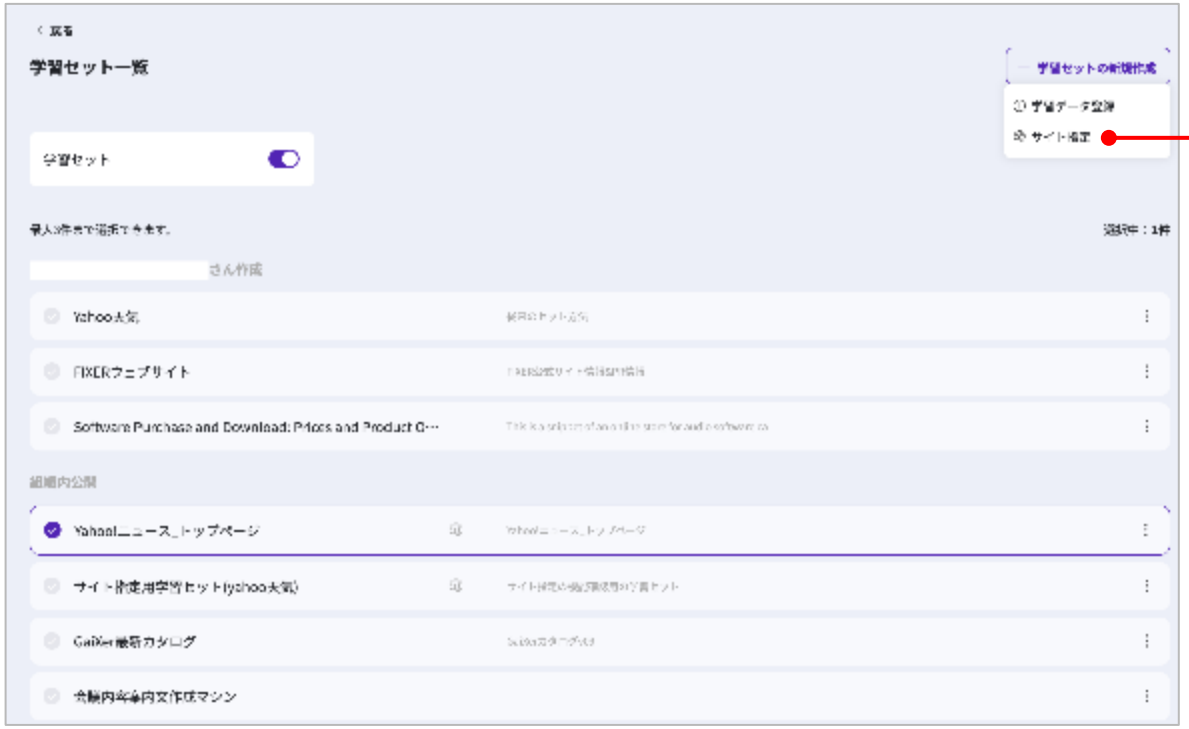
▪ モデルの訓練データもしくは学習セットから情報を参照できない場合にWeb検索を実行する。

▪ モデルの訓練データもしくは学習セットから情報のみを参照し、Web検索を実行しない。

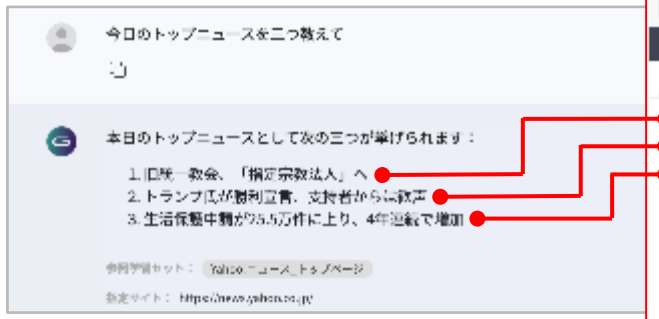
追加提供機能: サイト/ドメイン指定検索

▶ 新しい学習セットを作成する際に、「サイト指定」オプションを選択することにより、URLに記載された最新情報に基づいた回答生成が可能

▪ 学習セット画面



▪ 回答生成画面

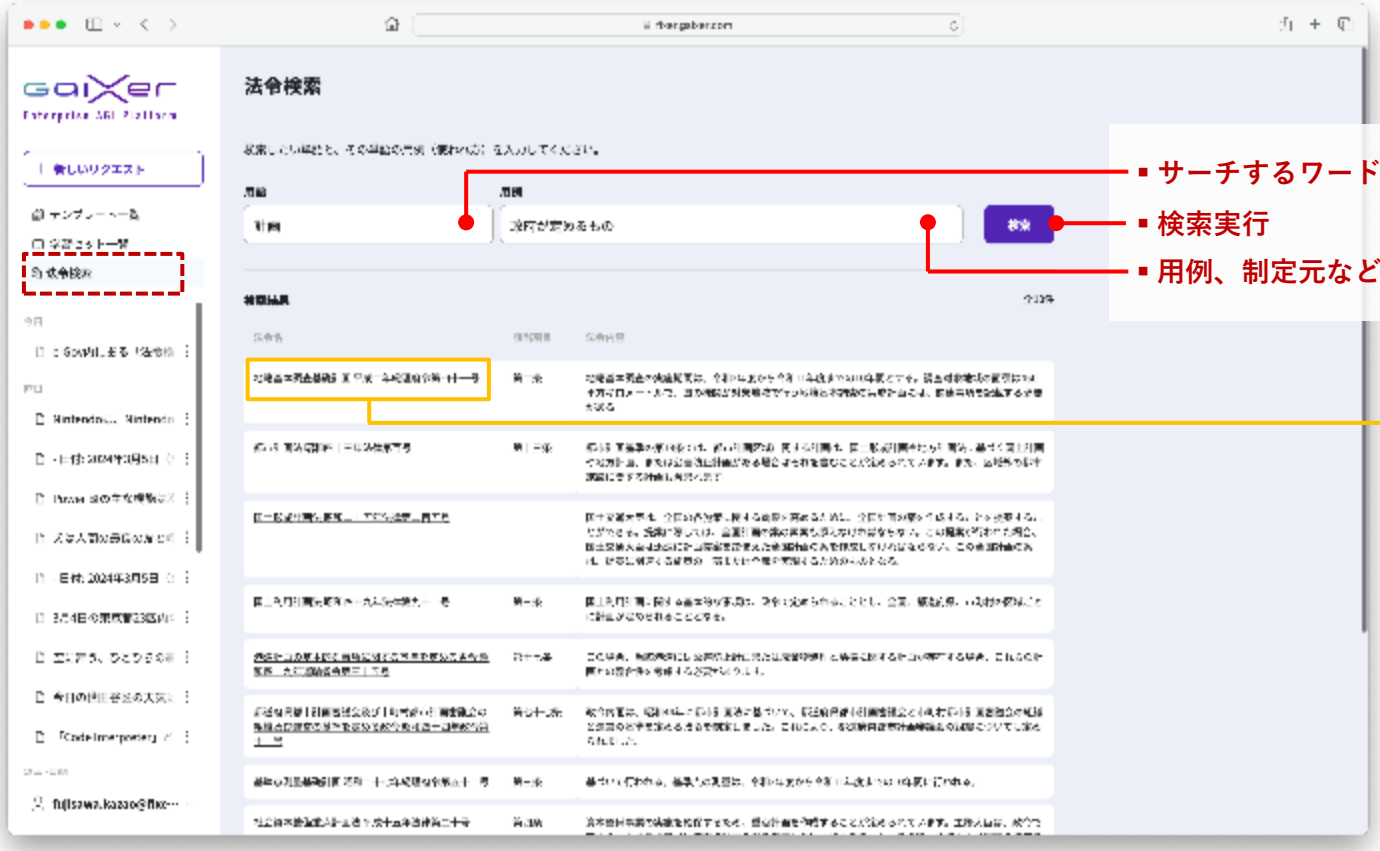


参照：Yahoo!ニュース3月6日14:00時点

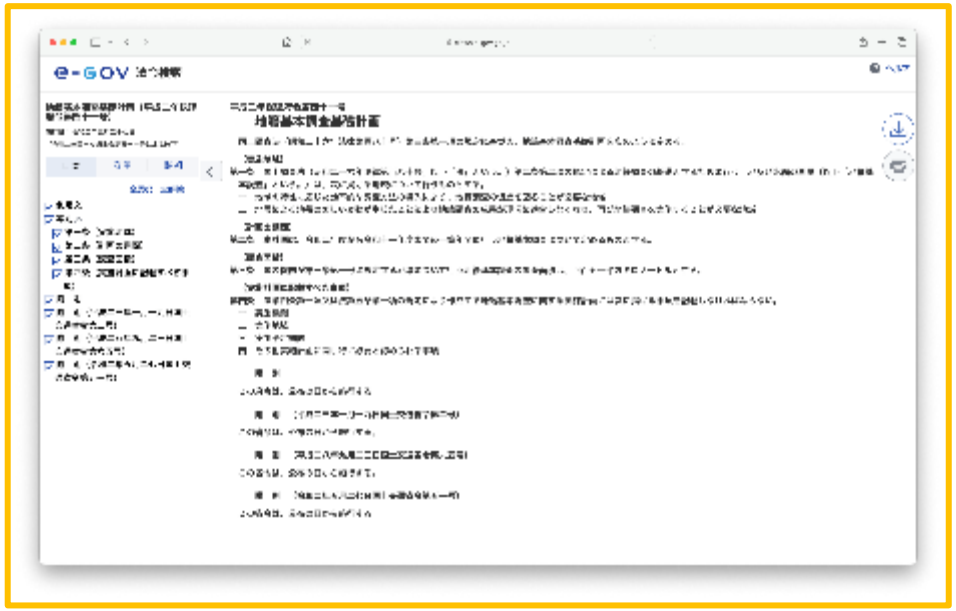
追加提供機能: 用例による法令検索

▶ 「用例による法令検索」は、日本の憲法、法律、政令、府省令、規則などの法令情報を検索し、確認可能
この機能を利用することで、最新の法令内容を確認したり、特定の法律に関する詳細情報を参照可能

▪ 用例による法令検索画面



▪ サーチするワード
▪ 検索実行
▪ 用例、制定元などの情報



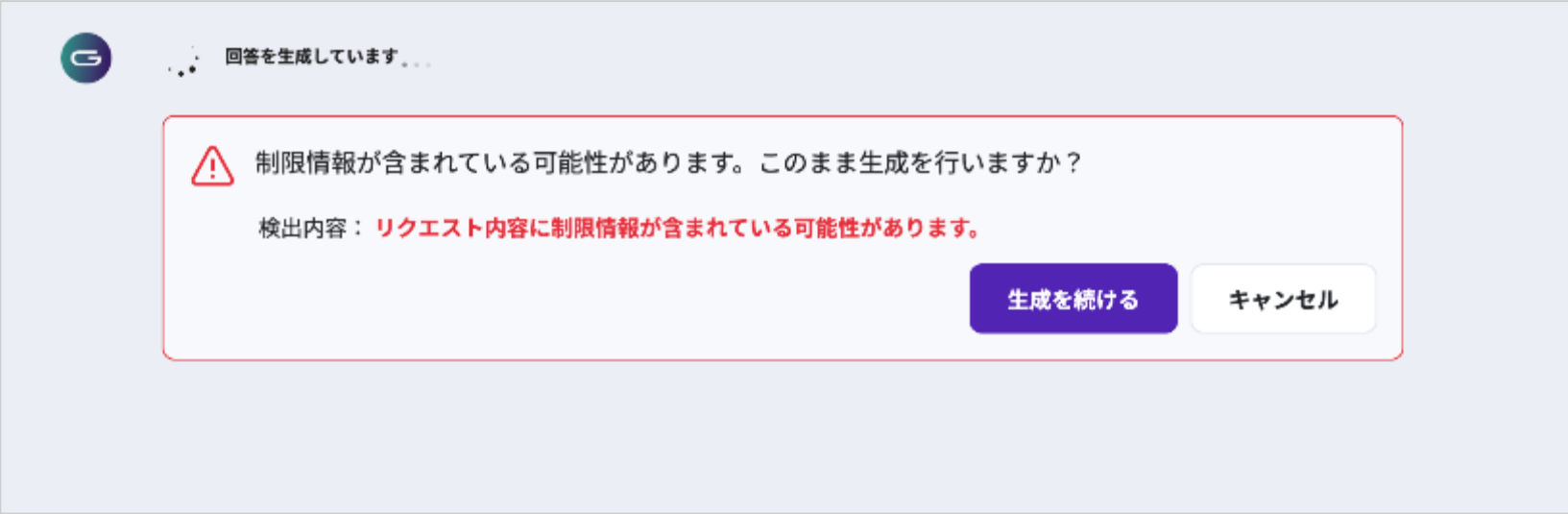
▪ 参照ページ（e-Govのウェブサイト）を開く。

追加提供機能: 機密情報制限チェッカー

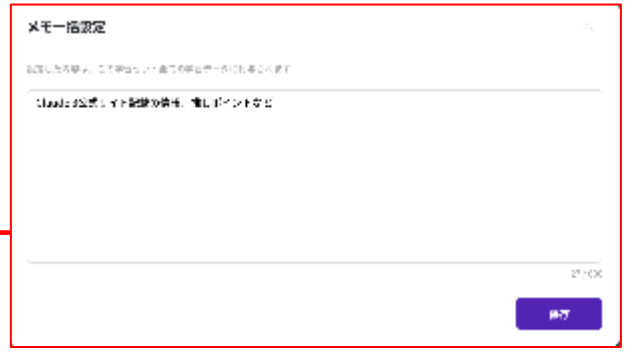
▶ 機密性２、３情報に関わる情報を生成させようとした際にアラートを出す機能

機密性についての格付けの定義

格付けの区分	分類の基準
機密性３情報	行政事務で取り扱う情報のうち、秘密文書に相当する機密性を要する情報
機密性２情報	行政事務で取り扱う情報のうち、秘密文書に相当する機密性は要しないが、漏えいにより、国民の権利が侵害され又は行政事務の遂行に支障を及ぼすおそれがある情報
機密性１情報	機密性２情報又は機密性３情報以外の情報



追加提供機能: 学習セットのメモ機能



メモ機能
データに関するメモ書きや共有相手に
セットの意図などを伝える際に利用する機能

追加提供機能: Slack連携

・一部チャンネルで利用可能
・DM利用は未定

目次

エグゼクティブ・サマリー

1. プロジェクトの背景と目的

2. スケジュールおよび体制

3. 業務の実施状況

4. ユースケース検証結果

4.1 対象ユースケースの全体像

4.2 ユースケース検証手法

4.3 各ユースケースの検証結果（サマリ）

5. 利用者サーベイ結果

6. 技術検証課題に対する考察および総括

Appendix

Appendix A. 検証基盤の利用状況

Appendix B. リリースノートおよび提供機能イメージ

Appendix C. ユースケースにおける検証結果（詳細）

検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

A-1. パブリックコメント対応品質の向上

ユースケース概要

前提	・政府各担当課は通常、関連政策や法令に関して国民や民間企業から広くパブリックコメントを収集する ・パブリックコメントはそのすべてに対して国が真摯に対応し、考慮すべきものであり、今回の生成AI検証では、AI任せではなく行政官の対応品質向上に主眼を置いて検証した
課題	行政官は、時に膨大な件数(1万件)にもなるコメントの集計/分類に基づく意見作成を実施し、忙殺されている。その際、難易度が高くない業務工程の効率化→高難度業務への対応時間確保が課題

今回の検証範囲

1.分類

パブリックコメントの内容で性質を分類し回答の生成が必要かを判断

・賛同意見
・今後の施策の参考意見
・関係がないもの

→

2.意見変換/分割/要約

一般に公開する際に変換すべき言語の検知と変換を行い、冗長と思われる部分は要約。また、複数意見の混在したパブリックコメントに関しては意見を分割

→

3.意見元抽出

回答の生成が必要と分類されたパブリックコメントに対し、パブリックコメント内容の背景を政府発布の資料から検知し抽出

→

4.回答担当課担当者の割り振り

パブリックコメントの内容と回答案によって、パブリックコメント募集担当課以外の担当課からの回答が必要なケースもあるため、適切な担当課へ割り振り

5.回答案作成

回答の生成が必要と分類されたパブリックコメントに対し、パブリックコメント内容の背景を政府発布の資料から検知。その内容に沿って適切な回答案を作成

→

6.集計/集計資料の作成

パブリックコメントの集計結果と概要資料に掲載するための集計資料を作成

→

7.資料の作成

一般公開を目的としたパブリックコメントの実施結果概要資料の作成(結果概要、提出意見、意見の考慮、結果・理由等を記載)

A-1. パブリックコメント対応品質の向上

ユーザーの作業分解とプロンプトの関連性

目的	パブリックコメントの公表資料作成時の難易度の高くない作業を効率化することで、行政官の高難度業務への対応品質向上を目的とする

	作業工程	プロンプト	検証アプローチ	検証結果
1	賛同意見/今後の検討の参考/形式的な修正事項/何ら関係がないもの/回答の作成が必要なもの	1.意見の分類 担当者コメント	・パブリックコメントを1つずつ入れ込み、その内容を分類するプロンプトを作成 ・担当者コメントに合わせたコメントも記入	✓ ・意見の分類の最終得点率：95%
2	コメント内の表現の変換	2.分割、表現変換 要約	・コメントの分割/意見変換/要約するプロンプトを作成 ・変換後の表現を記載	✓ ・コメントの分割の得点率：95% ・意見変換の得点率：90% ・要約の最終得点率：100 %
3	意見の分割			
4	冗長な意見の要約			
5	意見が指しているパブリックコメント対象資料内の記載の探知	3.回答案作成	・パブリックコメントの大まかな背景をプロンプトに入れ込み作成	▲ ・質問に対して回答はしているが、回答内容の最終得点率は60%にとどまった
6	意見の回答作成担当課・担当者の割振り	4.参照箇所抽出	・パブリックコメントの参考箇所となる部分を指摘 ・学習データでPDF資料をuploadして抽出	✕ ・参照先および参照先形式が多岐にわたるため、参照先の内容を正確に抽出する作業にLLM活用は不向きであることが判明した
7	意見に対する考え方(回答案)の作成			
8	パブリックコメントの集計及び集計資料の作成	開発性が伴う要件でもあり、一旦対象外 - 集計作業/集計結果のテーブル出力は可能(1に点線) - 概要資料の要点や素案作成は可能(学習データが必要)		―
9	概要資料の作成			

A-1. パブリックコメント対応品質の向上

現行の作業ステップ

▶ 10,000件のコメントが寄せられたケースの対応を行う場合、現行では意見の分類などの単純タスクだけで約670時間分の作業を要している

■生成AI導入前

生成AI導入前後で差分のあるステップ

No	作業ステップ	インプット	アウトプット	工数	工数合計（単純タスク）
1	意見の分類	パブリックコメント	分類のラベル付け	10,000分 (1分×10,000件)	670時間 (40,000分)
2	コメント内の表現の変換	パブリックコメント	パブリックコメント(変換後)	30,000分 (3分×10,000件)	
3	意見の分割				
4	冗長な意見の要約				
5	意見が指しているパブリックコメント対象資料内の記載の探知	パブリックコメント(変換後) パブリックコメント対象資料	パブリックコメント対象資料中の該当箇所	—	
6	意見の回答作成担当課・担当者の割振り	パブリックコメント(変換後)	回答作成の担当課・担当者	—	
7	意見に対する考え方(回答案)の作成	パブリックコメント(変換後) パブリックコメント対象資料	パブリックコメントに対する回答案	—	
8	パブリックコメントの集計及び集計資料の作成	パブリックコメント(変換後)	パブリックコメントの集計結果	—	
9	概要資料の作成	パブリックコメントに対する回答案 パブリックコメントの集計結果	概要資料	—	

A-1. パブリックコメント対応品質の向上

生成AI導入前後の作業ステップの変化

▶ **10,000件のコメントが寄せられた事例に対応する場合、単純タスクの作業工程で約670時間を要するが、生成AIを業務フローに導入する事で、約500時間の作業時間を短縮することが理論上可能と推測(670→170時間)**

■生成AI導入前

No	作業ステップ	インプット	アウトプット	工数	工数算出根拠	工数合計
1	意見の分類	パブリックコメント	分類のラベル付け	10,000分 (1分×10,000件)	コメントの文章が1件当たり約200文字あり、1分程度要すると考えられるため	**670時間(40,000分)**
2	不適切な文言の言い換え	パブリックコメント	パブリックコメント(変換後)	30,000分 (3分×10,000件)	手作業で文章を書き換える作業として、3分程度を妥当と考える	
3	意見の分割					
4	冗長な意見の要約					

■生成AI導入後(バッチ処理ができるシステムとして構築した場合)

No	作業ステップ	インプット	アウトプット	工数	工数算出根拠	工数合計
1	プロンプト試行 (意見の分類/不適切な文言の言い換え/意見の分割/冗長な意見の要約)	パブリックコメント(元データ)	分類のラベル付け パブリックコメント(変換後)	10,000分 (1分/件) →バッチ処理で夜間対応を行う為、本タスクは自動で実行される	システム化を行い、パブリックコメントを一括で生成AIに流し込むとした場合の推定	**ー**
2	人間によるチェック	分類のラベル付け パブリックコメント(変換後)	分類のラベル付け パブリックコメント(変換後)※チェック後	10,000分 (1分×10,000通)	コメントの文章が1件当たり約200文字あり、1分程度要すると考えられるため	**170時間(10,000分) 導入前比 -500時間**

A-1. パブリックコメント対応品質の向上

業務削減効果とコスト

▶ 政府のパブリックコメント対応品質の向上に対して、生成AIを導入すると、**年間約610万円で50,000時間の業務削減効果の見込み**（1,000円当たり8.2時間の削減）

■業務削減効果

業務削減効果（年間）　＝1事例当たりのパブリックコメント対応の業務削減時間 × 政府のパブリックコメント対応の総件数（年間）
＝500時間 × 100件※1
＝50,000時間

■必要コスト

分類/意見変換1回あたりのプロンプト試行コスト
＝入力トークン数（k） × 入力料金（ドル/kトークン）＋出力トークン数（k） × 出力料金（ドル/kトークン）
＝1.35※2 × 0.01※3 ＋ 0.911※2 × 0.03※3
＝0.0408ドル

必要コスト（年間）　＝分類/意見変換1回あたりのプロンプト試行コスト × パブリックコメントの件数（10,000通と仮定） × 政府のパブリックコメント対応の総件数（年間）
＝0.0408ドル × 10,000通 × 150（円/ドル） × 100件※1
＝6,120,000円

※1 「e-gov パブリック・コメント」よりパブリックコメントを収集する事例の年間平均は約2,000件であり、内パブリックコメントが1万通単位で寄せられる事例が100件と仮定
※2 本検証におけるプロンプト（最終版）で試行した際の平均値
※3 「Azure OpenAI Serviceの価格」よりGPT-4-Turboの単価をベースに引用

A-1. パブリックコメント対応品質の向上

生成AI導入による効果

▶ パブリックコメントのボリューム業務を生成AIに対応させることで、前頁の工数削減とコメント整理タスクの標準化を実現
これらにより、パブリックコメント業務そのものの品質向上、並びに次回政策策定時の遡及的な振り返りが可能

効果	効果内容
パブリックコメント対応そのものの質的な改善	パブリックコメントの単純業務部分を効率化する事で、担当職員の本件にかける時間の多くを振り分け先の指定や、意見作成や集計につぎ込むことで、パブリックコメントそのものの対応品質を向上させることが可能
	こうした対応スキームを横展開する事で標準化が期待できる。標準化を通じて、省庁ごとでの対応のばらつきをなくし、政府全体としての品質改善が可能
次回政策に向けた効率的な振り返り	集計フォーマットも含めたスキームの標準化を図ることで、パブリックコメントのデータベースを整備する事が可能になる(政策資料、パブリックコメント、それらの集計結果と対応結果等が格納)
	データベース化を通じ、政策策定時に関連政策とそれに付随するパブリックコメントの内容などを遡及的に振り返る事で、政策策定の品質改善に貢献が期待できる

パブリックコメント分類　評価基準と評価要件

ポイント1)

意見の中に複数に分類できるものがあり、完全一致ではないが正解データを包含している場合には部分点を与えることとした。

ポイント2)

分類の判断理由を明確にすることにより、生成結果の確認の効率化につながるため評価基準に加えることとした。

ポイント3)

分類の判断理由を明確にすることにより、生成結果の確認の効率化につながるため評価基準に加えることとした。

評価基準	要件定義	評価(点数)
回答の正しさ1:正しく分類できているか	妥当：正しく分類できている	2
	部分点：一部、分類できていない	1
	不適切：全く分類できていない	0
回答の正しさ2:分類の判断理由について説明しているか	妥当：適切に判断理由を説明できている	2
	部分点：一部、判断理由が誤っている	1
	不適切：適切な判断理由を説明できていない	0
回答の正しさ3:分類の判断元となる箇所を引用しているか	妥当：適切な判断元が引用できている	2
	部分点：一部、違った判断元から引用している	1
	不適切：適切な判断元が引用できていない	0

パブリックコメント意見変換　評価基準と評価要件

ポイント1)

形式的な修正事項の指摘がすべてできていなくても、一部できていれば作業の効率化につながるため、部分点を与えることとした。

ポイント2)

不適切な文言の言い換えがすべてできていなくても、一部できていれば作業の効率化につながるため、部分点を与えることとした。

ポイント3)

要約のわかりやすさの基準として、元文章の内容を改変することなく要約できているかどうかを判断し、抜け漏れがあるが致命的でないものに関しては部分点を与えることとした。

評価基準	要件定義	評価(点数)
回答の正しさ1:形式的な修正事項の指摘	妥当：すべての形式的な修正事項が指摘出来ている	2
	部分点：一部、修正事項が指摘出来ている	1
	不適切：修正できていない	0
回答の正しさ2:複数回答が必要なコメント内の意見の分割	妥当：適切な意見分割ができている	2
	部分点：一部、適切な意見分割ができている	1
	不適切：適切な意見修正ができていない	0
回答の正しさ3:不適切な文言の言いかえ	妥当：すべての不適切な文言の言いかえが出来ている	2
	部分点：一部、言い換えが出来ている	1
	不適切：言いかえが出来ていない	0
わかりやすさ:意見の要約	妥当：内容が適切に要約されている	2
	部分点：一部抜け漏れがあるが、適切に要約されている	1
	不適切：適切に要約されていない	0

パブリックコメント回答案作成　評価基準と評価要件

ポイント1)

質問に対して回答をしているが一部ずれた回答をしている場合でも、回答案としては致命的でないため、部分点を与えることとした。

ポイント2)

質問に対しての回答が正解データと一部ずれた回答をしている場合でも、回答案としては致命的でないため、部分点を与えることとした。

評価基準	要件定義	評価(点数)
回答の正しさ1:質問に対して回答をしているか	妥当：適切に回答している	2
	部分点：一部ずれた回答をしている箇所がある	1
	不適切：適切に回答できていない	0
回答の正しさ2:正解データと同じ内容を回答しているか	妥当：正解データと同じ趣旨の回答をしている	2
	部分点：一部違う回答をしている	1
	不適切：違う回答をしている	0

パブリックコメント意見抽出　評価基準と評価要件

ポイント1)

意見元が提示されないと業務として使えないため「明確に提示されているか/提示されていないか」の2項目での評価基準とした。

ポイント2)

意見元が候補として複数生成される場合があり、その中でも正解データが包含されている場合には致命的でないため、部分点を与えることとした。

評価基準	要件定義	評価(点数)
回答の正しさ1:意見元の場所が提示されているか	妥当：意見元の場所が提示されている	1
	不適切：意見元の場所が提示されていない	0
回答の正しさ2:意見元の場所が正しいか	妥当：適切な意見元が抽出できている	2
	部分点：一部誤った意見元を抽出している	1
	不適切：適切な意見が抽出できていない	0

A-1. パブリックコメント対応品質の向上

パブリックコメント分類　作業/評価ステップ

No.	検証段階	アクション	アクションの理由(簡潔に)
1	要件整理	パブリックコメントの検証資料の入手(総務省担当者より入手)	実際のパブリックコメントを入手する事で、具体的なコメントの中身に正しく対処する検証を行う為
2	要件整理	パブリックコメントの対応業務の中で人の手でのチェックを入れながらAIに任せてもよいもの/任せてはいけないものに分解	1)本業務においては対応品質向上を前提とした業務効率化を主軸に置くべき 2)特定政策に特化せず、手順の標準適用ができる領域を選定
3	クイック検証	LLMを選定(GPT-4-Turbo)	Claude2.1と並行で検討したが、コンテキスト長としてGPTでも問題なく収まる事が確認できたため
4	クイック検証	アクションごと(コメントの分類/表現変換/参照元の抽出/回答素案作成)でのプロンプトをプロンプトエンジニアにて作成	シングルプロンプトで実行した場合、命令が複雑になり精度が毀損されるため
5	クイック検証	アクションごとでの評価要件/評価基準を策定	-
6	クイック検証	政策から10個コメントを抽出して検証を実施	聞くたびに回答が変わることは無いか、正確に答えられているかなどをプロンプトを調整しながら確認するため
7	本検証	各プロンプトの構造の見直し(COT(step-by-step)化)	回答精度が想定より悪かったため (COTで意図理解を出力に含めることで精度改善を図った)
8	本検証	結果をベースに担当者との面談実施	評価基準や結果を担当者に説明し、実業務で役立つかどうか、to beのユースケース(システム化した将来像)のヒアリングのため
9	本検証	面談をベースに考察結果をとりまとめ	-

パブリックコメント分類　テストケースに関する詳細

テストケース作成方針

・分類：分類の根拠となる意見を正確に抽出できるか評価したかったため、参考とすべき意見を多く含むパブリックコメントを主に検証した
・意見変換：変換箇所の抽出および修正案の正確性を評価するため、変換の必要があるものを多く検証した
・回答生成：報告書の内容を踏まえた観点から回答を作成できるか評価したかったため、報告書を踏まえた回答が必要なものを多く検証した
・意見元抽出：意見元を報告書から探して抽出できるか評価したかったため、意見内部に意見元が明記されていないものを多く検証した

検証項目	賛同意見	参考とすべき意見	混在意見	合計
分類	1	7	2	10

検証項目	修正のないもの	修正のあるもの	合計
意見修正	2	8	10

検証項目	報告書の内容を元にそのまま回答できるもの	報告書を踏まえた回答が必要なもの	合計
回答生成	4	6	10

検証項目	意見に意見元が明記されているもの	意見に意見元が明記されていないもの	合計
意見元抽出	2	8	10

パブリックコメント分類　プロンプトの発展過程試行１回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト

命令
入力された{パブリックコメント}が「（1）賛成の意を示している意見」、「（2）今後検討が必要な意見」、「（3）本件とは関係のない意見」「（1）（2）（3）に該当しない、回答が必要な意見」のどれに該当するか、{出力形式}に則って分類してください。
また、分類した際に判断材料とした箇所を{パブリックコメント}から引用してください。

背景
＜背景や概要＞

内容（目次など）
＜報告書の目次など＞

出力形式
分類の結果：
判断材料となる箇所の引用：

作成者の評価

10問生成を行い正しく意見分類できたのは7問

課題#1

- 検討が必要な意見の多くが賛同意見に分類されてしまうなど、賛同意見に分類されるケースがとても多かった

課題#2

- 複数に該当するケースに対して、1つのみにしか分類されず、このプロンプトでは対応できていない

A-1. パブリックコメント対応品質の向上

パブリックコメント分類　プロンプトの発展過程試行２回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト

命令
政府が提出した特定の資料に対して入力された{パブリックコメント}の性質を分類してください。{出力形式}に従い、{パブリックコメント}を「（1）政府方針に対する賛同意見」、「（2）今後の施策の参考とすべき意見」、「（3）本件とは関係のない意見」の３つの分類にわけてください。ただし、一つの意見に複数の項目が含まれる場合があるため、その場合は複数の項目を指定する形で問題ありません。
分類した際に判断材料とした箇所を{パブリックコメント}から引用し、判断した理由についても答えてください。
どの項目にも分類できない場合には、「（4）分類不可」に分類してください。

背景や概要（任意）
<背景や概要など>

内容（任意）
<内容（目次など）>

出力形式
分類の結果：
判断材料となる箇所の引用：
判断した理由：

作成者の評価

10問生成を行い正しく意見分類できたのは9問

改善方針#1

- 単に賛成の意を示しているだけでなく、政府の意見に対して明確に賛同するものとして分類するようプロンプトを修正

改善方針#2

- 複数に該当するケースに対応できるよう、複数に分類できるケースがあることをプロンプトに明記

A-1. パブリックコメント対応品質の向上

パブリックコメント意見変換　プロンプトの発展過程試行１回目

LLM	GPT-4 Turbo

学習セット
無し

プロンプト

＃命令
[[条件]]に従って、[[パブリックコメント]]の内容をID別に編集してください

＃条件
- 以下の手順に従って回答を生成してください。
1.全ての[[パブリックコメント]]について、文脈にそぐわない表現、変換ミスと思われるような漢字、暴力的・反社会的な表現、倫理観に欠けた表現、その他の誤字・脱字などがないかチェックする。とくに変換ミスと思われる漢字が多く見つかると考えられるので、重点的に確認すること。
2.[[パブリックコメント]]に、1.で示した表現があった場合は、「修正すべきと思われる点があります。」と出力した後に、対象の表現に該当する部分を、
それぞれ10文字程度で、「文脈にそぐわない表現」「変換ミスと思われる漢字」「その他の誤字・脱字」の三項目に分けて列挙してください。
3. 次に、2.で列挙した表現について、修正案を提示してください。最適な修正案が分からない場合は、「確認をお願いします。」と出力してください。
4. 全ての[[パブリックコメント]]に修正点がなかった場合は「修正の必要はありませんでした。」と出力してください。
5.[[パブリックコメント]]について、一連の内容を通して、いくつかの異なる主張が含まれている場合、「＜主張の数＞個の主張があります。」とその数を出力してください。
6. 含まれている主張の数が二つ以上であった場合は、それぞれの主張を簡潔にまとめて出力してください。
7. 一つの主張のみが含まれているなら、それを簡潔にまとめて出力してください。
8. [[パブリックコメント]]について、これまでの[[条件]]の修正を踏まえて、一連の内容に対する意見は残した上で500文字程度で要約した文章を作成してください。

＃パブリックコメント
ID1:”””xxxxxx“”“
ID2:”””xxxxxx“”“
…

作成者の評価	複数のパブリックコメントを一度にプロンプトに入れていたため、修正箇所の漏れが多くほとんどうまく生成されなかった

課題#1

- 条件に従って生成するよう指示しているものの、量が多いためか作業の抜け漏れが激しい

課題#2

- 一度に入れるパブコメの数が多ければ多いほど品質が下がっていた

A-1. パブリックコメント対応品質の向上

パブリックコメント意見変換　プロンプトの発展過程試行２回目

LLM	GPT-4 Turbo

作成者の評価	10問の生成を行い正しく意見修正できたのは8問

学習セット

無し

プロンプト

＃命令
[[条件]]に従って、Step by Stepで[[パブリックコメント]]の内容を編集してください

＃条件
- 以下の手順に従って回答を生成してください。
1.全ての[[パブリックコメント]]について、文脈にそぐわない表現、変換ミスと思われるような漢字、暴力的・反社会的な表現、倫理観に欠けた表現、その他の誤字・脱字などがないかチェックする。とくに変換ミスと思われる漢字が多く見つかると考えられるので、重点的に確認すること。
2.[[パブリックコメント]]に、1.で示した表現があった場合は、「修正すべきと思われる点があります。」と出力した後に、対象の表現に該当する部分を、
それぞれ10文字程度で、「文脈にそぐわない表現」「変換ミスと思われる漢字」「その他の誤字・脱字」の三項目に分けて列挙してください。
3. 次に、2.で列挙した表現について、修正案を提示してください。最適な修正案が分からない場合は、「確認をお願いします。」と出力してください。
4. 全ての[[パブリックコメント]]に修正点がなかった場合は「修正の必要はありませんでした。」と出力してください。
5.[[パブリックコメント]]について、一連の内容を通して、いくつかの異なる主張が含まれている場合、「＜主張の数＞個の主張があります。」とその数を出力してください。
6. 含まれている主張の数が二つ以上であった場合は、それぞれの主張を簡潔にまとめて出力してください。
7. 一つの主張のみが含まれているなら、それを簡潔にまとめて出力してください。
8. [[パブリックコメント]]について、これまでの[[条件]]の修正を踏まえて、一連の内容に対する意見は残した上で500文字程度で要約した文章を作成してください。

＃パブリックコメント
”””xxxxxx“”“

改善方針#1

- 各作業を明確に実施するようstep by stepをプロンプトに追記

改善方針#2

- プロンプトに記載するパブコメを1つに絞ることで、各作業工程の精度の向上を図った

A-1. パブリックコメント対応品質の向上

パブリックコメント回答案作成　プロンプトの発展過程試行１回目

LLM	GPT-4 Turbo

学習セット

パブリックコメントの意見元である報告資料

プロンプト

＃命令
あなたは、パブリックコメントの回答担当者の助手です。送られてくるパブリックコメントに対して、どんな考え方で回答を作るべきかの案を出してください。
また、コメントは複数件送られる場合がありますが、別々に回答してください。資料をよく読み、400字以内で簡潔に答えてください。
資料にのっとって、誤解を招かない表現をしてください。

作成者の評価

10問生成を行い、ほとんどが回答できないという主旨の結果が生成された

課題#1

- 学習セットから必要な情報が取ってこれていないためか、質問に対して妥当と評価できる回答案が生成されない

課題#2

- 正解データ（受領したパブコメの回答データ）と比べて感謝の表現が多く、質問に対して明確に回答が作成できていないケースが多かった

パブリックコメント回答案作成　プロンプトの発展過程試行２回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト

＃命令
あなたは、パブリックコメントの回答担当者の助手です。送られてくるパブリックコメントに対して、どんな考え方で回答を作るべきかの案を出してください。
また、コメントは複数件送られる場合がありますが、別々に回答してください。資料をよく読み、400字以内で簡潔に答えてください。
資料にのっとって、誤解を招かない表現をしてください。感謝の意を示す文章は省いてください。

#パブリックコメント募集の背景
＜パブリックコメントの概要や背景＞

作成者の評価	10問生成を行い最終的に正解データと似た回答をしたのは5問だった

改善方針#1

- パブリックコメント募集時の報告書の内容に関して、トークン数上限の都合上すべて含めることができないため、概要や背景となる箇所をプロンプトに記載

改善方針# 2

- 感謝の表現を省かせ、与えられた資料の概要や背景に則ることを指示に追記することで、質問に対して明確に回答案を作成

A-1. パブリックコメント対応品質の向上

パブリックコメント意見元抽出　プロンプトの発展過程試行１回目

LLM	GPT-4 Turbo

学習セット

パブリックコメントの意見元である報告資料

プロンプト

＃命令
政府が提出した特定の資料に対して入力された{パブリックコメント}がどの章についての意見なのか、{背景}や{内容}を参考に{Context}から探して答えてください。
見つからない場合は、その趣旨を答えてください。

#背景
＜背景や概要＞

作成者の評価	10問生成を行い、正しい意見元が抽出されたのは0件だった

課題#1

- 学習セットから必要な情報が抽出できていない

課題#2

- Web検索した情報を元に回答する場合があり、動作に不安定さが生じる

パブリックコメント意見元抽出　プロンプトの発展過程試行２回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト

＃命令
政府が提出した特定の資料に対して入力された{パブリックコメント}がどの章についての意見なのか、{背景}や{内容}を参考に{Context}から探して答えてください。
見つからない場合は、その趣旨を答えてください。

制約
Web検索を実施しない。回答ができない場合にも「申し訳ありません」と生成しない。

#背景
＜背景や概要＞

#内容(もくじ)
＜資料の目次など＞

作成者の評価

10問生成を行い最終的に正しい意見元を抽出できたのは3問だった

改善方針#1

- 学習セットではなく、報告書の目次をプロンプトに追記することで、意見を述べている対象となっている章を生成させる

改善方針#2

- Web検索を実施させないことを明確にすることで、一般的な回答をすることを防ぐ

※ 検証当初はなかったが、2024/3/21時点でのGaiXerでは、機能としてWeb検索をOFFにすることが可能

A-1. パブリックコメント対応品質の向上

試行2回目の評価結果

ポイント1)

意見分類の得点率はほぼ100%であり、部分点となった1問に関しても本質的には誤りではない内容であったため、実業務での利用が大きく期待できる。

ポイント2)

意見変換においても不適当回答はなく、要約においては得点率100%であるので、実業務での利用が十分期待できる。

ポイント3)

回答案生成と意見元抽出は、生成に必要となる情報量の多さと情報のドキュメント形式が多岐にわたることから、最終回答点数が低くなったため、今作業工程におけるLLM活用はやはり不向きであることが判明した。

【意見分類】評価基準	妥当	部分点	不適当	質問数	得点率
回答の正しさ1:正しく分類できているか	9	1	0	10	95%
回答の正しさ2:分類の判断理由について説明しているか	10	0	0	10	100%
回答の正しさ3:分類の判断元となる箇所を引用しているか	10	0	0	10	100%

【意見変換】評価基準	妥当	部分点	不適当	質問数	得点率
回答の正しさ1:形式的な修正事項の指摘	9	1	0	10	95%
回答の正しさ2:複数回答が必要なコメント内の意見の分割	9	1	0	10	95%
回答の正しさ3:不適切な文言の言い換え	8	2	0	10	90%
わかりやすさ：意見の要約	10	0	0	10	100%

【回答案作成】評価基準	妥当	部分点	不適当	質問数	得点率
回答の正しさ1:質問に対して回答をしているか	9	0	1	10	90%
回答の正しさ2:正解データと同じ内容を回答しているか	5	3	2	10	65%

【意見元抽出】評価基準	妥当	部分点	不適当	質問数	得点率
回答の正しさ1:意見元の場所が提示されているか	6	0	4	10	60%
回答の正しさ2:意見元の場所が正しいか	2	4	4	10	40%

A-1. パブリックコメント対応品質の向上

総括

結論　：　生成AIを用いたパブリックコメントの分類や論点整理は担当者の業務負担を改善しながらコメントへの対応品質向上に貢献するものであり、有効と考える

考察　：

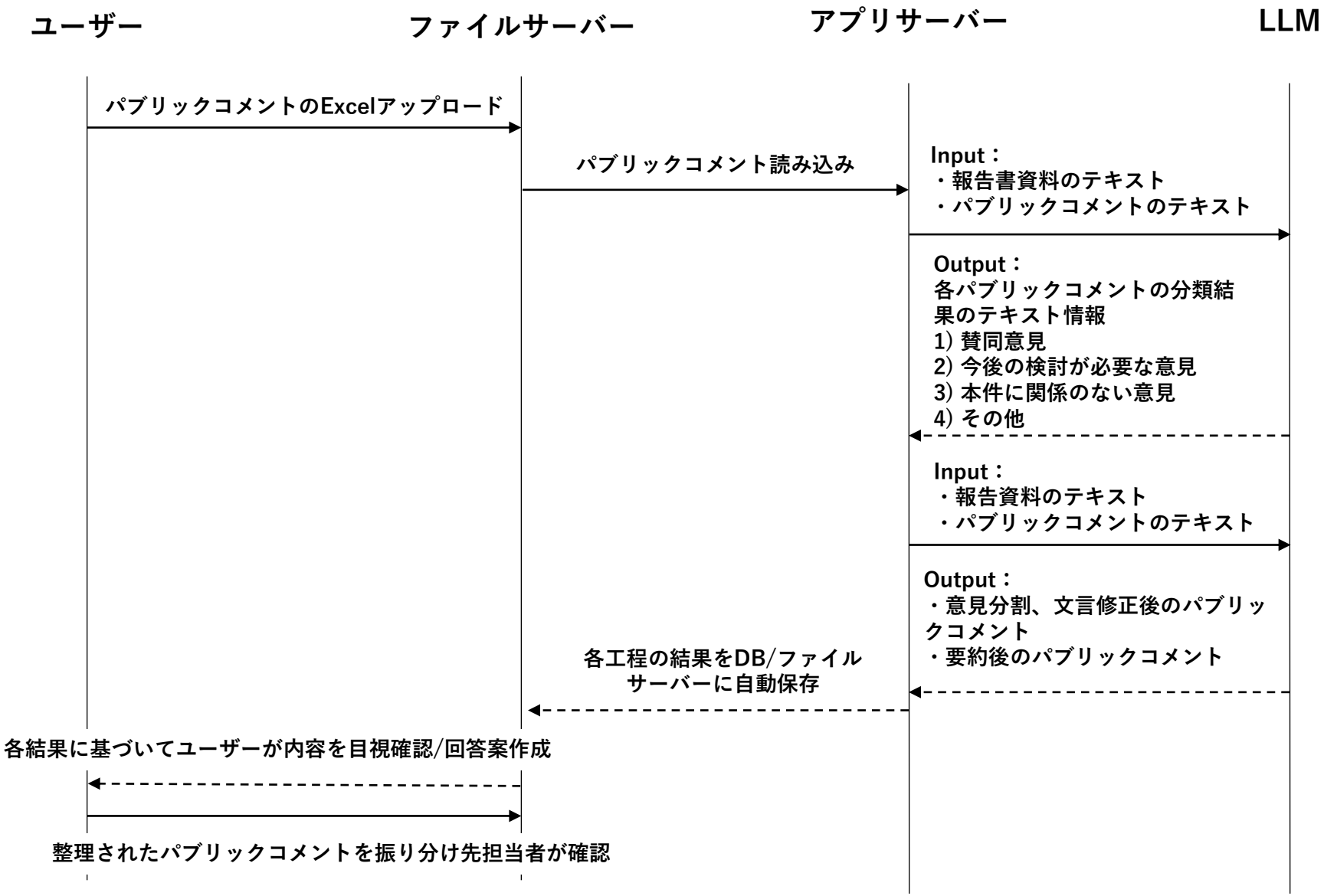
・年間約2000案件に対応しているものと仮定した場合、分類/論点整理などのタスクについては、補助的にAIを活用する事で対応品質の向上が期待できる

・特に重視したい点として、コメント件数が多いもの(1万件)等に対しても十分に力を発揮する事ができる点であり、システム化する事で生成AIの対応範囲を標準化、スケール適用の効果が期待できる

実装形式：

実装レベル	実装形式 (シーケンス図ご参照)	コスト構造	導入/実施にあたり考慮すべき課題
検証	生成AI型SaaSサービス上でプロンプトテンプレートを作成し、ユーザーがサービスにアクセスしてプロンプトテンプレートにコメント本文を挿入する事で結果を生成	• SaaS利用料(トークン料込)	• 更に複数種類のパブリックコメントでの網羅テストを行うことで、プロンプトや利用するLLMの普遍性を検証する • バルク対応、自動化が求められる場合は組織の業務フローに組み込む必要があるため、実装形式を要検討
将来像	LLMに実行させた資料の分類/意見整理結果を府省庁のファイルサーバーと自動連携し、府省庁のパブリックコメントDBを構築	• システム実装コスト • インフラ運用コスト • トークン利用料	・RAG機能の活用検証を通じ、パブリックコメントの回答作成等をフロー化できるかを検討 ・システム実装時に府省庁ごとでのネットワークやデータベースの違いなどを考慮した標準導入の検討

シーケンス図


ユーザー
ファイルサーバー
アプリサーバー
LLM
パブリックコメントのExcelアップロード
パブリックコメント読み込み
Input：
・報告書資料のテキスト
・パブリックコメントのテキスト
Output：
各パブリックコメントの分類結果のテキスト情報
1) 賛同意見
2) 今後の検討が必要な意見
3) 本件に関係のない意見
4) その他
Input：
・報告資料のテキスト
・パブリックコメントのテキスト
Output：
・意見分割、文言修正後のパブリックコメント
・要約後のパブリックコメント
各工程の結果をDB/ファイルサーバーに自動保存
各結果に基づいてユーザーが内容を目視確認/回答案作成
整理されたパブリックコメントを振り分け先担当者が確認

検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

A-2. 調達仕様書のラベリング

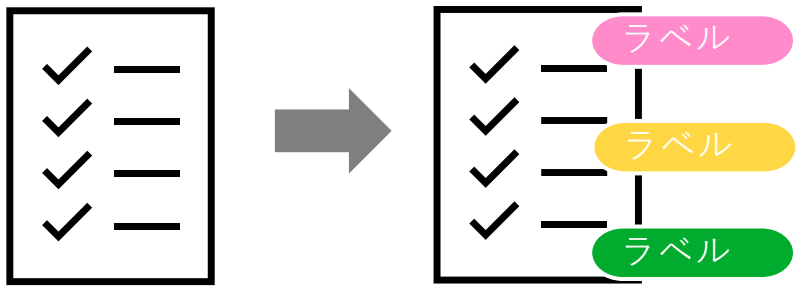
ユースケース概要

前提	・府省庁は年間15,000件(R3年度実績)の入札を実施しており、全て調達仕様書に基づき調達がなされる ・政府にとって調達品質の改善と効率化は重要命題であり、そのデータベースの充実や整備ニーズは高い
課題	最適ラベルを案件概要のテキスト情報から抽出もしくは解釈して割り当てるという単一作業に業務工数がかかることや、ラベリング済みの調達仕様書のデータベース整備ができていないことが課題

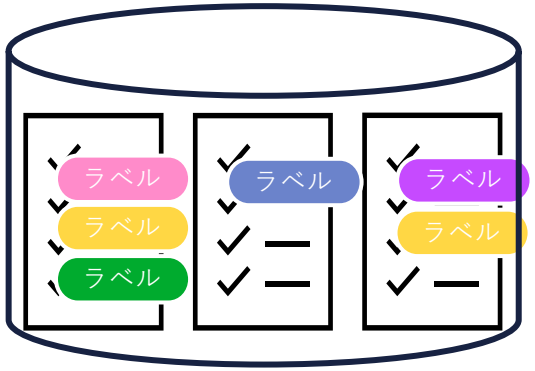
今回の検証範囲

1.調達仕様書をGaiXerに読み込ませる

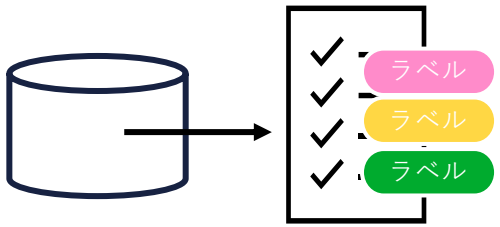
2.調達案件名と案件概要を作成調達項目にラベリング



3.過去調達仕様書のラベリング結果をデータベースへ登録



4.新規作成の調達仕様書と類似する過去の調達仕様書をラベルを基に検索
検索した調達仕様書をもとに新規作成



ラベルをもとに
類似の調達仕様書を検索

5.新規作成された調達仕様書を1.以降の手順で繰り返し行うことでデータベースの情報を拡充。継続的な品質向上を行う

A-2. 調達仕様書のラベリング

作業分解とプロンプトの関連性

目的	・ 調達仕様書作成時、過去の類似した調達仕様書を参照しやすいよう、調達仕様を特徴付けるラベリング/タグ付けを行う

作業工程	プロンプト	検証アプローチ	検証結果
1 調達仕様書の読み込み	案件詳細作成とラベリング	・調達仕様書の資料のテキストを入力し、案件名と案件概要を生成する	✓ 案件名と案件概要の高精度な生成に成功 ・生成案件名の最終得点率：100% ・生成案件概要の最終得点率：83%
2 案件名と案件概要の作成		・調達仕様書の各項目に対して決まったラベルの中からラベルを付与していく	✓ 正解データよりも多くラベルをつけてしまう場合が多いが、漏れはなく正解データを確実に含んでいる。最終得点は完全一致としては50点となっているが、付与過多は誤りではないため、実用として十分応えられる結果であるといえる
3 ラベリング			
4 各仕様書にリンクする表形式アプリ(MS List)への反映	生成AIタスクとしては対象外	-	

A-2. 調達仕様書のラベリング

生成AI導入前後の作業ステップの変化

▶ 政府調達案件の調達仕様書のラベリングに対して、生成AIを導入すると、**1件当たり14分間の業務時間が削減**

■生成AI導入前

No	作業ステップ	インプット	アウトプット	工数 (熟練者による作業)	工数算出根拠	工数合計
1	調達仕様書の読み込み	調達仕様書本文	概要理解	5分	担当者へヒアリング	15分
2	案件名抽出、案件概要の記述	調達仕様書本文	案件名などの文章	5分	担当者へヒアリング	
3	ラベル項目を表ツール(MS List等)に記載しラベル付け	調達仕様書本文	ラベルテキスト	5分	担当者へヒアリング	

■生成AI導入後

No	作業ステップ	インプット	アウトプット	工数	工数算出根拠	工数合計
1	ラベリングのプロンプト試行	調達仕様書本文	ラベルテキスト	0分	バッチ処理で自動化する想定	1分 **導入前比 -14分**
2	担当者による最終チェック	ラベルテキスト	チェック後のラベルテキスト	1分	明らかに異なるラベルが含まれていないかを確認する程度の想定	

業務削減効果とコスト

▶ 政府調達案件の調達仕様書のラベリングに対して、生成AIを導入すると、年間約70万円で3,500時間の業務削減効果の見込み（1,000円当たり5時間の業務削減）

■業務削減効果

業務削減効果(年間)　＝ 調達仕様書1件当たりの業務削減時間 × 政府の調達仕様書の総件数(年間)
＝ 14分 × 15,000件※1
＝ 270,000分
＝ 3,500時間

■必要コスト

ラベリング1回あたりのプロンプト試行コスト
＝ 入力トークン数(k) × 入力料金(ドル/kトークン) ＋ 出力トークン数(k) × 出力料金(ドル/kトークン)
＝ 25.0※2 × 0.01※3 ＋ 2.11※2 × 0.03※3
＝ 0.313ドル

必要コスト(年間)　＝ラベリング1回あたりのプロンプト試行コスト × 政府の調達仕様書の総件数(年間)
＝ 0.313ドル × 150(円/ドル) × 15,000件※1
＝ 704,250円

※1 「令和３年(暦年)における政府調達実績」より年間平均15,000件と仮定
※2 本検証におけるプロンプト(最終版)で試行した際の平均値
※3 「Azure OpenAI Serviceの価格」よりGPT-4-Turboの価格を引用

調達仕様書ラベリング業務への生成AI導入に関する定性効果

▶ 属人的に各担当者の裁量に任されている本業務を、苦労なく展開できるユースケースとして明示する事で、府省庁への展開を可能にするとともに、調達仕様書作成時の参照先を統一化が可能

効果	効果内容
仕様書ラベリング業務そのものの普及	ラベリング手順そのものが簡素化されたことで、府省庁/自治体で同様の業務に携わる(りたい)職員の参入障壁が下がり、作業担当者の母数が増える
	これにより、仕様書のラベリング業務そのものが政府全体で普及し、標準化へ向けた活動が加速化する
ラベリングDBの利用ユーザー増	各府省庁での普及を通じて、DBを参照して仕様書を作成する職員数が増える
	ユーザーが増える事で仕様書作成の結果や、ラベリング内容に対するフィードバックサイクルのPDCA化が可能になり、より高品質な調達仕様書の作成が期待できる

調達仕様書のラベリングに伴うデータベースを構築する意義

▶ 行政の調達仕様書をデータベース化する作業を府省庁、ひいては自治体単位で効率的に実施する事が可能
これにより、行政の調達仕様書としての品質改善に加えて、異なる組織間で、行政としての調達ナレッジを共有する仕組み構築に貢献

効果	効果内容
仕様書の品質改善	過去の類似した調達仕様書を多く検索できることで、将来の調達仕様書の品質向上が見込まれる
	本用例を拡張し、案件の遂行結果などもラベル化してDB化する事で、より充実したインプットを仕様書作成担当者に効率的に提供(事業者のパフォーマンスや、導入効果等、定量/定性評価履歴を蓄積)
組織間での調達ナレッジの共有	ラベリングスキームが標準展開でき、府省庁や自治体をも巻き込んだ仕組みとして展開する事で、多組織間での調達ナレッジの相互共有に貢献
	調達のベストプラクティスの共有に留まらず、案件遂行予算の考え方など、様々な切り口から行政調達を政府全体として振り返るツールとして拡張するロードマップを組むことができる

評価基準と評価要件

ポイント1)

案件名や案件概要は軽微の誤りがあっても致命的なものではないため部分点を与えることとし、ラベリング項目の漏れの有り無しで点差が出るよう設定した。

ポイント2)

ラベリングの漏れの有無が最重要評価基準であるため、「付与されている/漏れがある」の2択で評価点数も0点か2点の2択とした。

ポイント3)

現職の方へのヒアリングにより、正解データを含む形であれば許容範囲との意見があったため、余分に貼られている分には部分点を与えることとした。

評価基準	要件定義	評価(点数)
回答の正しさ1:調達仕様書の案件名が完全一致	完全一致	2
	不完全一致だが軽微な違い	1
	不完全一致	0
表現の適切さ1:案件概要の適切な記述	案件の調達内容、概要、目的が漏れなく正確に記載されている	2
	調達内容/概要/目的のいずれかが不明瞭	1
	案件の説明が誤っている	0
回答の正しさ2:ラベル漏れの項目はないか	ラベルを付与すべき全ての項目にラベルが付与されている	2
	ラベル漏れがある	0
回答の正しさ3:選択したラベルが適切か	択一、あるいは複数選択の箇所で、全てのラベルが完全一致	2
	複数選択箇所で、一部のラベルは一致するが余分に貼られている	1
	択一で不一致、あるいは複数選択で完全不一致	0

A-2. 調達仕様書のラベリング

作業/評価ステップ

No.	検証段階	アクション	アクションの理由(簡潔に)
1	要件整理	調達仕様書のラベリング業務について、デジ庁担当者からのヒアリング	担当者がどのように調達仕様書のデータベースを作ろうとしているのかを確認する為
2	要件整理	ラベリング項目の洗い出し	仕様書のラベル情報が定型化されているため、それらをプロンプトに内包すべく確認
3	要件整理	ラベリング業務において、担当者の目線での評価基準や品質を確認	LLMにラベリングさせた際、担当者がチェックしてほぼ修正するようでは使い物にならない為、具体的なラベリング作業の精度をすり合わせ
4	要件整理	ラベリング作業未実施の調達仕様書を10個受領	実際の仕様書でラベリングしていないものを、担当者目線で評価し、後日担当者がどの程度手入れしないといけないかを検証する為
5	本検証	各アクション(仕様書概要の作成とラベリングタスク)を1つにまとめたプロンプトを作成	本検証においては仕様書本文→LLMが抽出した概要→ラベル項目をシングルセットで出力する形式が最も精度確保できると考えたため
6	本検証	4&5を併せ、10のテストケースと定義して検証実行	-
7	本検証	各結果をベースに担当者との面談実施	評価基準や結果を担当者に説明し、実業務で役立つかどうか、to beのユースケース(システム化した将来像)のヒアリングのため
8	本検証	面談をベースに考察結果をとりまとめ	-

テストケースに関する詳細

デジタル庁より提供された調達仕様書10件にてテストを実施

【提供された調達仕様書名】

1. 340679_デジタル技術を活用340679_デジタル技術を活用した災害時の位置情報共有に関する調査研究
2. 340945_令和4年度ガバメントソリューションサービスにおけるセキュリティ脅威の検知・分析・遮断サービスの提供
3. 340964_デジタル庁ガバメントソリューションサービスの府省LAN統合に係る作業支援
4. 341056_防災情報のデジタル化推進に向けた防災アーキテクチャに関する調査研究
5. 341060_地方公共団体における防災業務の実態に関する調査研究
6. 342229_令和4年度デジタル庁ガバメントソリューションサービスに係るドキュメント整備支援
7. 342453_法人向けＧビズＩＤ発行申請の電子化に向けた検証事業
8. 349411_e-Govのクラウド移行及び機能拡充に関する設計・開発等
9. 359459_令和5年度 補助金申請システムの運用及び保守等業務
10. 359506_令和5年度 補助金申請システムの開発支援業務

A-2. 調達仕様書のラベリング

プロンプトの発展過程　試行１回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト1/2

Step-by-Stepの内容を出力しながら処理を行ってください。
Step1.{{調達案件名}と{案件概要}からそれぞれ適切なものを以下のタグから選択し{案件のラベリング}を選択してください。また、選択したタグごとに理由を記述してください。

{プロジェクトの特徴}を選択してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。
＜プロジェクトの特徴のラベル一覧＞

{調達工程}を選択してください。一つだけ選択可能で、当てはまらない場合は「該当なし」と出力してください。
＜調達工程の特徴のラベル一覧＞

{役務内容}を選択してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。
＜役務内容の特徴のラベル一覧＞

{物品・サービスの購入・賃貸借}を選択してください。複数選択可能で、当てはまらない場合又は{役務内容}が"物品・サービスの購入、賃貸借・保守"ではない場合は「該当なし」と出力してください。
＜物品・サービスの購入・賃貸借のラベル一覧＞

{開発方式}を選択してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。
＜開発方式のラベル一覧＞

{開発手法}を選択してください。一つだけ可能で、当てはまらない場合は「該当なし」と出力してください。
＜開発手法のラベル一覧＞

作成者の評価

正解データとラベルが完全一致することはほぼなかったものの、調達案件名と案件概要に対して正解のラベルが生成されたのは10問中6問であった

課題#1

- ラベリングのために調達案件名と案件概要の情報が必要であるものの、生成するStepが存在しない

A-2. 調達仕様書のラベリング

プロンプトの発展過程　試行１回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト2/2

```
## Step2.{案件のラベリング}から以下の出力形式に従い、結果を出力してください。
**プロジェクトの特徴**
{プロジェクトの特徴}
**調達工程**
{調達工程}
**役務内容**
{役務内容}
**物品・サービスの購入・賃貸借**
{物品・サービスの購入・賃貸借}
**新規・改修等**
{新規・改修等}
**システム基盤**
{システム基盤}
**開発方式**
{開発方式}
**開発手法**
{開発手法}
```

評価項目

回答の正しさ1:漏れがないか	○ / △ / ×
回答の正しさ2:内容に対して正しくラベリングされてるか	○ / △ / ×
回答の正しさ3:元データと同じ値がラベリングされてるか	○ / △ / ×

作成者の評価

正解データとラベルが完全一致することはほぼなかったものの、調達案件名と案件概要に対して正解のラベルが生成されたのは10問中6問であった

課題#2

- 各ラベルの定義がプロンプトに記載されていないため、ラベル付けの正解率が低くなっている可能性がある

課題#3

- 案件名と案件概要の生成も作業工程に追加したため、評価項目の見直しが必要

課題#4

- 正解データよりも多くラベルをつけてしまう場合が多いが、正解データを含む形であれば許容範囲との現職の方の意見があったため評価項目の見直しが必要

A-2. 調達仕様書のラベリング

試行１回目の生成結果

質問

案件名
情報提供ネットワークシステムの基盤対応に関する調達仕様書等作成支援（請負）

案件概要
第二期情報提供ネットワークシステムにおいて利用しているOS,M/Wに係るEoLに向けて、基盤改修対応を予定している。それらに係る調達仕様書の作成や要件定義書案の加筆修正に係る業務支援を調達するもの。

生成結果

Step1の指示に従い、提供された調達案件名と案件概要から適切なタグを選択し、理由を記載します。

プロジェクトの特徴
- ネットワーク関係: 案件概要に「情報提供ネットワークシステム」とあり、ネットワークに関連するシステムであることが分かるため。
- 基盤構築: 「基盤改修対応」という文言があり、システムの基盤に関する作業が含まれていることが分かるため。

調達工程
- 要件定義フェーズ: 「要件定義書案の加筆修正」という文言があるため、要件定義フェーズに関連する作業が含まれていることが分かる。

役務内容
- 調査研究（国内）: OS/MWのEoLに向けた基盤改修対応のために、調査研究が必要と考えられる。
- 要件定義: 「要件定義書案の加筆修正」という文言があるため、要件定義に関する作業が含まれている。

物品・サービスの購入・賃貸借
該当なし: 物品やサービスの購入・賃貸借に関する具体的な記載が案件概要にないため。

新規・改修等
- 機能追加・改修: 「基盤改修対応」という文言があるため、既存システムの改修に関連する作業が含まれていることが分かる。

システム基盤
該当なし: 特定のシステム基盤の種類（オンプレミス、クラウド等）に関する記述がないため。

開発方式
該当なし: 特定の開発方式（PaaS、SaaS、PKG、スクラッチ開発）に関する記述がないため。

開発手法
該当なし: 特定の開発手法（アジャイル、ウォーターフォール等）に関する記述がないため。

Step2の指示に従い、結果を出力します。
プロジェクトの特徴
ネットワーク関係、基盤構築

調達工程
要件定義フェーズ

役務内容
調査研究（国内）、要件定義

物品・サービスの購入・賃貸借
該当なし

新規・改修等
機能追加・改修

システム基盤
該当なし

開発方式
該当なし

開発手法
該当なし

A-2. 調達仕様書のラベリング

試行１回目の生成結果

生成されたラベル

プロジェクトの特徴：ネットワーク関係、基盤構築
調達工程：要件定義フェーズ
役務内容：調査研究（国内）、要件定義
物品・サービスの購入・賃貸借：-
新規・改修等：機能追加・改修
システム基盤：-
開発方式：-
開発手法：-

評価項目

回答の正しさ1:漏れがないか	各項目に対して漏れがなくラベリングされている	○ / △ / ×（評価：○）
回答の正しさ2:内容に対して正しくラベリングされてるか	案件名と案件概要に対して正しくラベリングされている	○ / △ / ×（評価：○）
回答の正しさ3:元データと同じ値がラベリングされてるか	元データのラベルを8割以上包含した結果になっている	○ / △ / ×（評価：△）

A-2. 調達仕様書のラベリング

プロンプトの発展過程　試行2回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト1/4

```
# 調達仕様書
＜調達仕様書の内容＞

# 条件
{{# Context}}の内容を一切参照しないでください。
{{# 調達仕様書}}の内容のみを参照してください。

#Step-by-Stepで実行
Step-by-Stepの内容を出力しながら処理を行ってください。

## Step1.{{# 調達仕様書}}からそれぞれ条件に従い生成してください。
{調達案件名}を抜き出してください。
{案件概要}を{調達仕様書}の内容を要約し作成してください。要約する際、常態(である調)にしてください。

## Step2.{{# 調達仕様書}}からそれぞれ適切なものを以下のタグから選択してください。
可能性で選択することは禁止とします。
{プロジェクトの特徴}を以下の項目から**関連値の高いもの**を選択し、どうして関連値が高いのかそれぞれ選択した物毎に詳細に説明してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。
- UI/UX重視:調達仕様書における記載から,UI/UXを重視している調達仕様書
- Webシステム(一般公開):構築・運用等の対象システムが一般公開されているWebシステムである調達仕様書
- Webシステム(官公庁内):構築・運用等の対象システムが府省共通システムである調達仕様書
- マイナンバーカード対応:構築・運用等の対象システムがマイナンバーカード対応システムである調達仕様書
```

作成者の評価

- 案件名は正解データとの一致率が100%であり、案件概要も8割以上が適切に生成された
- ラベリングに関しては、全10問において元データのラベルを包含する結果が得られた

改善方針#1

- ラベリングのために調達案件名と案件概要を生成する必要があったため、調達案件名と案件概要を生成するStepを追加

改善方針#2

- 各ラベルの定義をプロンプトに記載(次頁以降記載箇所も同様)

プロンプトの発展過程　試行2回目

プロンプト2/4

- スマートフォン対応:構築・運用等の対象システムがスマートフォン対応システムである調達仕様書
- 複数システムと連携:構築・運用等の対象システムが複数システムとの連携するシステムである調達仕様書
- クラウドサービス関係:パブリック,プライベート等のクラウドサービス関連の調達仕様書
- ネットワーク関係:ネットワーク関連の調達仕様書

{調達工程}を以下の項目から選択してください。一つだけ選択可能で、当てはまらない場合は「該当なし」と出力してください。また、選択した、または該当なしとなった理由を詳細に説明してください。可能性で選択することは禁止とします。
- 要件定義フェーズ:システムの要件定義が主であると捉えられる調達仕様書
- 調査研究フェーズ：システムに関わらず調査研究（国内・国外）や実証実験（プロトタイプ等の構築含む）が主であると捉えられる調達仕様書
- 設計・開発フェーズ:アプリケーション,システム基盤,ネットワーク基盤などの設計,開発,テスト工程が主であると捉えられる調達仕様書
- システム運用フェーズ:構築したアプリケーション,システム基盤,ネットワーク基盤などの運用,保守工程が主であると捉えられる調達仕様書
- システム以外:システム調達以外の調達仕様書

{役務内容}を以下の項目から**関連値の高いもの**を選択し、どうして関連値が高いのかそれぞれ選択した物毎に詳細に説明してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。また、選択した、または該当なしとなった理由を詳細に説明してください。可能性で選択することは禁止とします。
- 設計：**設計・開発フェーズにおける**各工程のうち、それぞれの役務の内容が含まれると捉えられる調達仕様書
- アプリケーション開発：**設計・開発フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- 基盤構築：**設計・開発フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- テスト：**設計・開発フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- 移行・導入：**設計・開発フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- 工程管理支援：**設計・開発フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- システム運用：*システム運用フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- アプリケーション保守：**システム運用フェーズにおける**各工程のうち、役務の内容が含まれると捉えられる調達仕様書
- 物品・サービスの購入,賃貸借・保守:クラウドサービスやSaaS,サーバーやパソコンのような物品の賃貸借と捉えられる調達仕様書 ※対象物品が主に何であるかもメタデータとして管理
- 調査研究(国内):国内外の事例等を調査・収集し,報告書等にまとめる調査研究が含まれると捉えられる調達仕様書
- 調査研究(海外):国内外の事例等を調査・収集し,報告書等にまとめる調査研究が含まれると捉えられる調達仕様書
- 実証実験(アプリ開発あり):机上,又はプロトタイプ等のアプリケーションを構築し,実証として効果検証等を実施,報告書等にまとめる実証実験が含まれると捉えられる調達仕様書
- 実証実験(アプリ開発なし):机上,又はプロトタイプ等のアプリケーションを構築し,実証として効果検証等を実施,報告書等にまとめる実証実験が含まれると捉えられる調達仕様書
- コンサルティング:調査の実施ではなく,会議体の事務局の運営支援業務や作業の支援全般のようなコンサルティング業務が含まれると捉えられる調達仕様書
- 運営・事務局等:調査の実施ではなく,会議体の事務局の運営支援業務や作業の支援全般のようなコンサルティング業務が含まれると捉えられる調達仕様書
- ヘルプデスク:コールセンターやヘルプデスク業務が含まれると捉えられる調達仕様書
- 労働者派遣:上記のうち,労働者派遣と明記されている調達仕様書
- システム関連以外:消耗品の購入や簡易な公共工事などのシステム関連以外の調達仕様書

プロンプトの発展過程　試行2回目

プロンプト3/4

{物品・サービスの購入・賃貸借}を以下の項目から**関連値の高いもの**を選択し、どうして関連値が高いのかそれぞれ選択した物毎に詳細に説明してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。可能性で選択することは禁止とします。
- サーバ等機器:サーバ等の機器を購入・賃貸借する役務の内容が含まれると捉えられる調達仕様書
- PC端末・周辺機器等:職員が直接利用するPCや周辺機器を購入・賃貸借する役務の内容が含まれると捉えられる調達仕様書
- ネットワーク回線・周辺機器等：ネットワーク回線（敷設含む）及び周辺機器等を購入・賃貸借する役務の内容が含まれると捉えられる調達仕様書
- ソフトウェア提供：ソフトウェアサービス（SaaSは除くパッケージソフトウェアを想定）を購入・賃貸借する役務の内容が含まれると捉えられる調達仕様書
- 施設・設備提供:施設・設備等の場所を提供する役務の内容が含まれると捉えられる調達仕様書
- クラウド提供：クラウドサービス（IaaS,PaaS）を提供する役務の内容が含まれると捉えられる調達仕様書
- SaaS提供:SaaSサービスを提供する役務の内容が含まれると捉えられる調達仕様書
- その他機器,ハードウェア:上記に紐づかないが,その他機器・ハードウェアを購入・賃貸借する役務の内容が含まれると捉えられる調達仕様書

{新規・改修等}を以下の項目から**関連値の高いもの**を選択し、どうして関連値が高いのかそれぞれ選択した物毎に詳細に説明してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。可能性で選択することは禁止とします。
- 既存システム:新規開発のシステムに係る要件定義,設計開発,システム運用等が対象と捉えられる調達仕様書
- 機能追加・改修:既存システムに対する機能追加,機能改修等が対象と捉えられる調達仕様書
- 再構築(更改):既存システムを基に新たに再構築,基盤更改,クラウドリフト・シフト等が対象と捉えられる調達仕様書
- 新規構築:既存システムに関連する業務であるが,アプリケーション開発が主ではないと捉えられる調達仕様書
{システム基盤}を以下の項目から選択してください。一つだけ選択可能で、当てはまらない場合は「該当なし」と出力してください。また、選択した、または該当なしとなった理由を詳細に説明してください。可能性で選択することは禁止とします。
- ガバメントクラウド:原則ガバメントクラウド上での構築を求めている調達仕様書
- パブリッククラウド:原則パブリッククラウド上での構築を求めている調達仕様書(ガバメントクラウドを除く)
- オンプレミス：クラウドサービスではない基盤上（オンプレミス、一部プライベートクラウド等）での構築を求めている調達仕様書
- 政府共通PF:第一期又は第二期政府共通プラットフォーム上での構築を求めている調達仕様書
{開発方式}を以下の項目から**関連値の高いもの**を選択し、どうして関連値が高いのかそれぞれ選択した物毎に詳細に説明してください。複数選択可能で、当てはまらない場合は「該当なし」と出力してください。可能性で選択することは禁止とします。
- OSS活用:オープンソースソフトウェアの積極的な活用が記載されている調達仕様書
- PKG活用:オープンソースソフトウェアではないパッケージソフトウェアの活用が明記されている調達仕様書
- PaaS活用:クラウドサービスが提供するPaaSの活用が明記されている調達仕様書
- SaaS活用:ノー/ローコードツールやプロジェクト管理ツールなどのSaaSの活用が明記されている調達仕様書

{開発手法}を以下の項目から選択してください。一つだけ可能で、当てはまらない場合は「該当なし」と出力してください。また、選択した、または該当なしとなった理由を詳細に説明してください。可能性で選択することは禁止とします。
- ウォーターフォール:システム開発手法としてウォーターフォール開発手法を求めていると捉えられる調達仕様書
- アジャイル：システム開発手法としてアジャイル開発手法（スプリント等）を求めていると捉えられる調達仕様書
- プロトタイプ:システム開発手法としてプロトタイプ開発手法を求めていると捉えられる調達仕様書

プロンプトの発展過程　試行2回目

プロンプト4/4

Step3.出力形式に従い、結果を出力してください。
調達案件名
{調達案件名}

プロジェクトの概要
{プロジェクトの概要}

プロジェクトの特徴
{プロジェクトの特徴}

調達工程
{調達工程}

役務内容
{役務内容}

物品・サービスの購入・賃貸借
{物品・サービスの購入・賃貸借}

新規・改修等
{新規・改修等}

システム基盤
{システム基盤}

開発方式
{開発方式}

開発手法
{開発手法}"

試行2回目の生成結果

質問

341056_防災情報のデジタル化推進に向けた防災アーキテクチャに関する調査研究/04_【別紙１】調達仕様書.pdf

生成結果1/2

Step1調達案件名}を抜き出してください。
防災情報のデジタル化推進に向けた防災アーキテクチャに関する調査研究

{案件概要}を{調達仕様書}の内容を要約し作成してください。要約する際、常態(である調)にしてください。
デジタル技術を活用して確実かつ迅速に災害対応を実施できる仕組みや、被災者支援の活動を効率化できるシステムの構築を目的とする。主にデジタル化を推進するための基礎調査として、既存の施策の概要整理、データ活用による防災業務の検討、人的被害軽減に向けた予防対策と応急対策の基礎調査、それらに関する調査整理、報告書の作成を行う。

Step2
{プロジェクトの特徴}を以下のリストの中から実際に係る内容と**非常に関連値の高いもの**を選択し、どうして関連値が高いのか、どこの情報を引用したのかを必ず示し、詳細に説明してください。
該当なし
理由：調達仕様書中にUI/UX重視、Webシステムの公開、マイナンバーカード対応等に特化した記述は見受けられず、また特定のシステム連携やクラウドサービス、ネットワーク関係に言及している内容も確認できないため。

{調達工程}を以下のリストの中から該当項目を参考に選択し、どこの情報を引用したのかを必ず示し、なぜ選択したのかを詳細に説明してください。
調査研究フェーズ
理由：調達仕様書ページ6の「２ 作業内容」において、プロジェクト計画書の作成、データ活用による防災業務の検討、予防対策と応急対策に関する基礎調査など、調査研究に関連する活動が主な内容として記述されているため。

{役務内容}を以下のリストの中から該当項目を参考に**非常に関連値の高いもの**を選択し、どうして関連値が高いのか、どこの情報を引用したのかを必ず示し、詳細に説明してください。
該当なし
理由：調達仕様書には具体的なシステム設計、開発、運用等に関する記述がなく、主に調査研究に関する内容が記載されているため、上記の役務内容には該当しない。

{物品・サービスの購入・賃貸借}を以下のリストの中から該当項目を参考に**非常に関連値の高いもの**を選択し、どうして関連値が高いのか、どこの情報を引用したのかを必ず示し、詳細に説明してください。
該当なし
理由：調達仕様書内で物品やサービスの購入、賃貸借に関する具体的な記述は確認できず、調査研究が主な目的であるため。

{新規・改修等}を以下のリストの中から該当項目を参考に**非常に関連値の高いもの**を選択し、どうして関連値が高いのか、どこの情報を引用したのかを必ず示し、詳細に説明してください。
該当なし
理由：調達仕様書内で既存システムの改修や新規システムの構築に関する記述はなく、調査研究に関連する活動が主要な役務内容であるため。

{システム基盤}を以下のリストの中から該当項目を参考に選択し、どこの情報を引用したのかを必ず示し、なぜ選択したのかを詳細に説明してください。
該当なし
理由：調達仕様書内で特定のシステム基盤上での構築に関する記述は見受けられないため。

試行2回目の生成結果

生成結果2/2

{開発方式}を以下のリストの中から該当項目を参考に**非常に関連値の高いもの**を選択し、どうして関連値が高いのか、どこの情報を引用したのかを必ず示し、詳細に説明してください。
該当なし
理由：調達仕様書内でオープンソースソフトウェア、パッケージソフトウェア、PaaS、SaaSなどの開発方式や活用に関する具体的な記述はなく、調査研究を主目的としているため。

{開発手法}を以下のリストの中から該当項目を参考に選択し、どこの情報を引用したのかを必ず示し、なぜ選択したのかを詳細に説明してください。
該当なし
理由：調達仕様書内でウォーターフォール、アジャイル、プロトタイプといった具体的な開発手法に言及している部分は確認できず、調査研究に関連する内容が中心であるため。

Step3
調達案件名
防災情報のデジタル化推進に向けた防災アーキテクチャに関する調査研究

プロジェクトの概要
デジタル技術を活用して確実かつ迅速に災害対応を実施できる仕組みや、被災者支援の活動を効率化できるシステムの構築を目的とする。主にデジタル化を推進するための基礎調査として、既存の施策の概要整理、データ活用による防災業務の検討、人的被害軽減に向けた予防対策と応急対策の基礎調査、それらに関する調査整理、報告書の作成を行う。

プロジェクトの特徴
該当なし

調達工程
調査研究フェーズ

役務内容
該当なし

物品・サービスの購入・賃貸借
該当なし

新規・改修等
該当なし

システム基盤
該当なし

開発方式
該当なし

開発手法
該当なし

試行2回目の生成結果

生成されたラベル

調達案件名
防災情報のデジタル化推進に向けた防災アーキテクチャに関する調査研究

プロジェクトの概要
デジタル技術を活用して確実かつ迅速に災害対応を実施できる仕組みや、被災者支援の活動を効率化できるシステムの構築を目的とする。主にデジタル化を推進するための基礎調査として、既存の施策の概要整理、データ活用による防災業務の検討、人的被害軽減に向けた予防対策と応急対策の基礎調査、それらに関する調査整理、報告書の作成を行う。

プロジェクトの特徴
該当なし

調達工程
調査研究フェーズ

役務内容
該当なし

物品・サービスの購入・賃貸借
該当なし

新規・改修等
該当なし

システム基盤
該当なし

開発方式
該当なし

開発手法
該当なし

試行2回目の評価項目

評価項目

評価基準	要件定義	評価(点数)
回答の正しさ1:調達仕様書の案件名が完全一致	完全一致	2
	不完全一致だが軽微な違い	1
	不完全一致	0
表現の適切さ1:案件概要の適切な記述	案件の調達内容、概要、目的が漏れなく正確に記載されている	2
	調達内容/概要/目的のいずれかが不明瞭	1
	案件の説明が誤っている	0
回答の正しさ2:ラベル漏れの項目はないか	ラベルを付与すべき全ての項目にラベルが付与されている	2
	ラベル漏れがある	0
回答の正しさ3:選択したラベルが適切か	択一、あるいは複数選択の箇所で、全てのラベルが完全一致	2
	複数選択箇所で、一部のラベルは一致するが余分に貼られている	1
	択一で不一致、あるいは複数選択で完全不一致	0

改善方針#3

- 案件名と案件概要の生成も作業工程に追加したため、評価項目を見直し

改善方針#4

- 正解データよりも多くラベルをつけてしまう場合が多いが、正解データを含む形であれば許容範囲との現職の方の意見があったため評価項目を見直し

A-2. 調達仕様書のラベリング

試行２回目の評価結果

ポイント1)

最重要評価基準である「ラベル漏れの項目はないか」において、得点率が100%であり、本ケースにおいて生成AIは実用可能と考えられる。

ポイント2)

「選択したラベルが適切か」においては、部分点を獲得できていれば実用上許容可能とユーザーから伺っており、すべてのケースで部分点を獲得しているため実用可能と考えられる。(詳細は次頁参照)

ポイント3)

今回の検証ではすでに案件名と案件概要が作成済みのテストケースで検証を行ったが、案件概要を生成させた結果高スコアを出せていることから、実業務において案件名と案件概要が未作成の場合でも高品質の生成結果を出せると期待できる。

評価基準	妥当	部分点	不適当	質問数	得点率
回答の正しさ1:調達仕様書の案件名が完全一致	9	0	0	9	100%
表現の適切さ1:案件概要の適切な記述	6	3	0	9	83%
回答の正しさ2:ラベル漏れの項目はないか	9	0	0	9	100%
回答の正しさ3:選択したラベルが適切か	0	9	0	9	50%

課題とみられた生成結果の例

生成結果のうち主に課題とみられるケースとして下記2種類が存在した
※1: 正解ラベルを包含しているものの、余分にラベルが付与されているケース
※2: 正解ラベルとは違うラベルが付与されているケース

この2種類のうち、※1についてはユーザーより許容可能との意見をいただいた
よって、※1に関する評価では部分点を付与し、※2に関しては配点しないような評価基準を設定した

案件名と概要	ラベリングの項目	【GaiXer生成ラベル】	【正解ラベル】
【案件名】 情報提供ネットワークシステムの基盤対応に関する調達仕様書等作成支援(請負) 【案件概要】 第二期情報提供ネットワークシステムにおいて利用しているOS,M/Wに係るEoLに向けて、基盤改修対応を予定している。それらに係る調達仕様書の作成や要件定義書案の加筆修正に係る業務支援を調達するもの。	プロジェクトの特徴	ネットワーク関係 基盤構築【※1】	ネットワーク関係
	調達工程	要件定義フェーズ	要件定義フェーズ
	役務内容	要件定義 調査研究(国内)　【※1】	要件定義
	物品・サービスの購入・賃貸借	該当なし	該当なし
	新規・改修等	機能追加・改修	機能追加・改修
	システム基盤	該当なし【※2】	オンプレミス
	開発方式	該当なし	該当なし
	開発手法	該当なし	該当なし

総括

結論　：　調達仕様書のラベリングは生成AIの活用用途として適切であり、導入対効果が見込まれる

考察　：

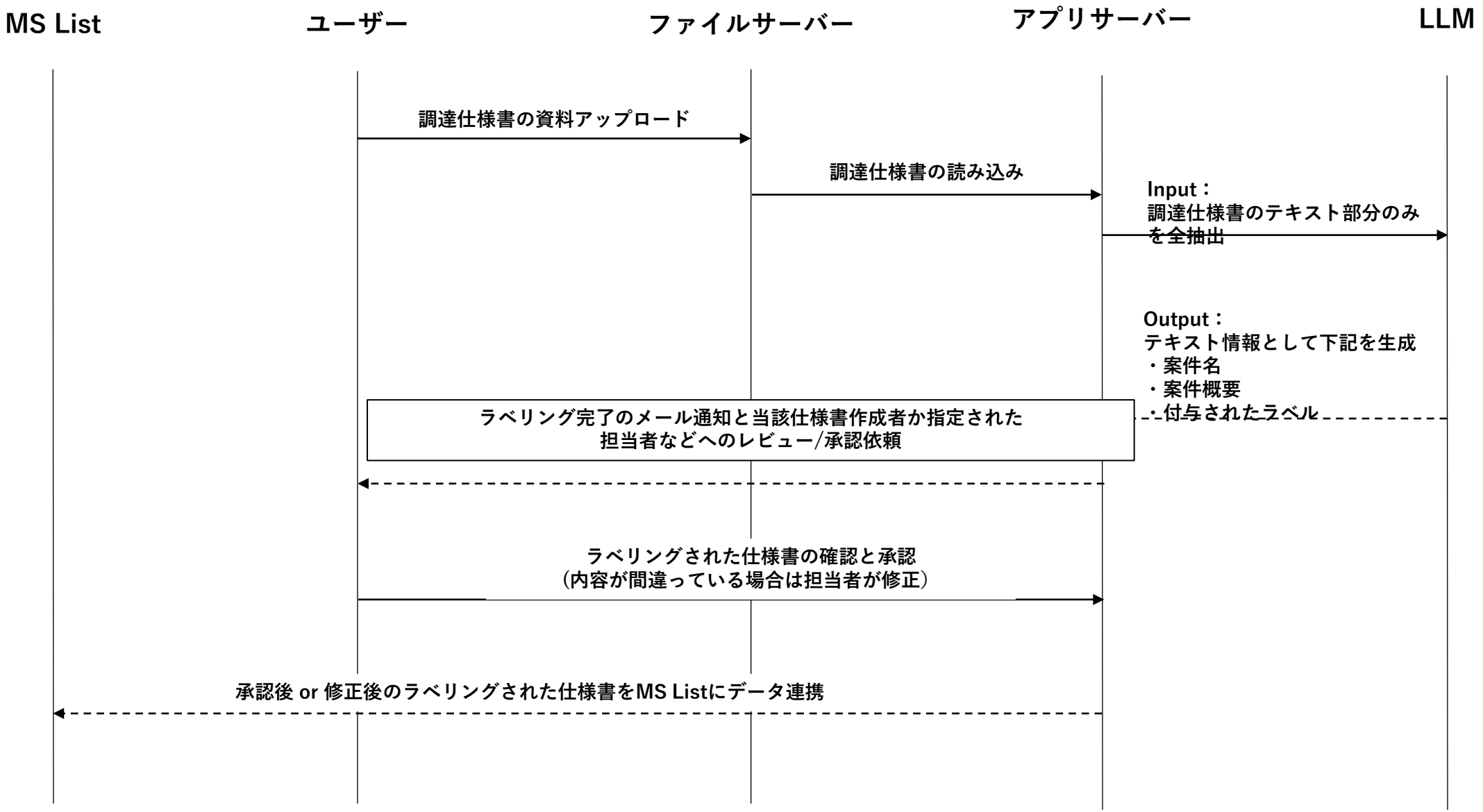
- 検証試行の過程で、複数選択可の項目で明らかに誤ったラベリングが見られたが、プロンプト修正において大きく改善
- 具体的には生成AIのtop_pを特にコントロールするようなプロンプティングが功を奏したように考えられる
  (e.g.「**非常に関連値の高いもの**を選択」と指示した箇所)
- 担当者からの定性評価として、本ユースケースは有用とのコメントも受領しており、今後の本格実装が期待できる
- なお、将来的に本ユースケースをシステム化した際には、ラベリング結果や内容に対してユーザーがフィードバック/指摘できる
  インターフェースを実装する事で、ラベリング精度の改善PDCAを構築でき、更なる利用率と調達品質改善への貢献が期待できると思慮する

実装形式

実装レベル	実装形式	コスト構造	利点	課題
検証	プロンプトテンプレート機能や、出力形式を指定できる生成AI SaaSサービスを導入	SaaS利用料(トークン量込)	本利用に向けた本格検証が可能。組織管理/共有が可能なため、部分的な業務利用も可能。	SaaSインターフェースに準拠した利用となるため、ワークフローとしての組み込みが困難。
将来像	後述のシーケンス図に従ったファイルサーバーからリスト化ツールへの自動反映を含めた手順をシステム化して提供。同実装に先立ち、既存の生成AI SaaSサービスでLLMやプロンプトの調整を実施。	• システム実装コスト • インフラ運用コスト • トークン利用料	府省庁内システムとの連携が確保されるため、業務フローの自動/半自動化が可能。	システム実装時に府省庁ごとでのネットワークやデータベースの違いなどを考慮した標準導入の検討が必要。

A-2. 調達仕様書のラベリング

シーケンス図



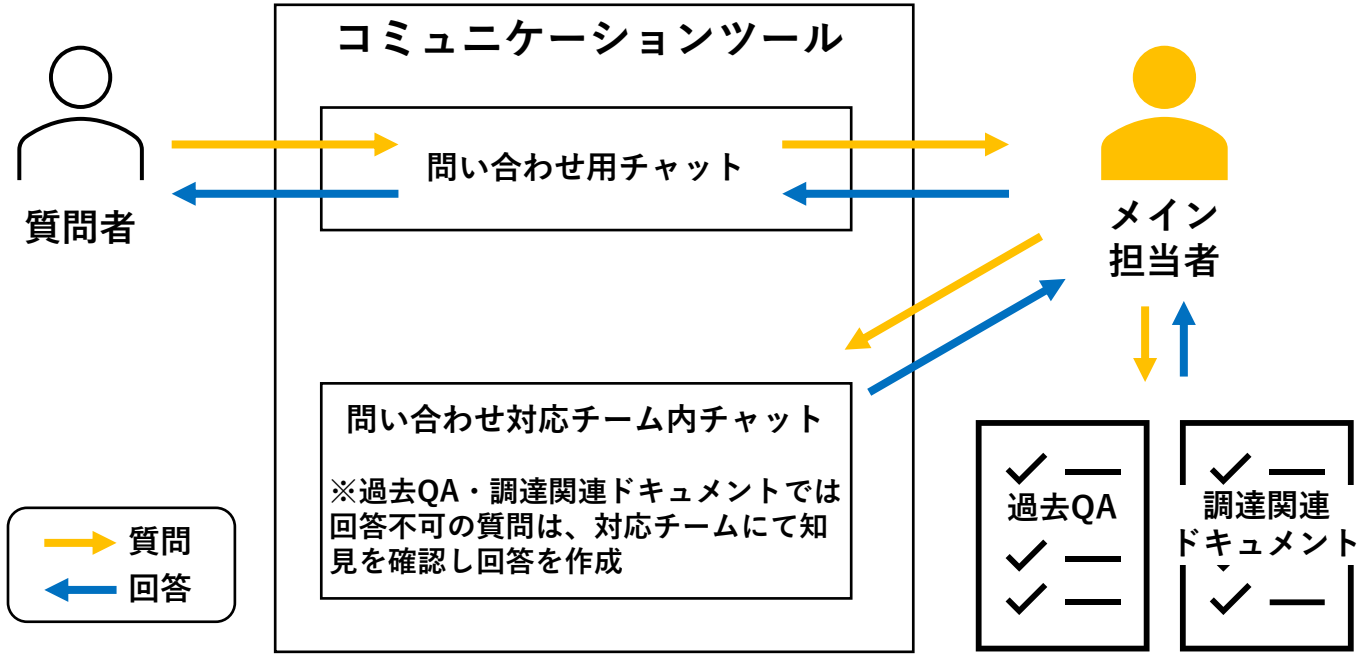
検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

ユースケース概要

前提	複数の政府ガイドラインが記された文書と人の知見を用いて問合せ対応を行っている職員の業務(本務ではない)において、生成AIによる代行対応の可能性を検証することで、行政におけるコミュニケーションコストの最適化を検証するもの。
課題	今回はデジタル庁で調達関連の問い合わせを受けるチームのタスクに着目。本チームは、各職員からの調達に関する問い合わせを人の手で対応しており、回答までに工数と時間を要することが、質問者・回答担当者ともに業務負担になっている点を課題とした。

デジタル庁における調達関連の問い合わせの回答フロー



デジタル・ガバメント推進標準ガイドライン実践ガイドブック(以降、実践ガイドブックと呼称)において、生成AIによるFAQ対応を検証

長大な文書をいかにして生成AIに学習させるかがポイント

ユーザーの作業分解とプロンプトの関連性

目的

- 長大なガイドライン等、様々なPDF文書等を参照しながら、職員からの問い合わせに対して担当職員が正確なアドバイスをしながらFAQを整備していく業務を生成AIで効率化する

作業工程

1. ユーザー(職員)から問合せ受領(Slackなど)
2. 担当職員が文書等から回答に参照する情報を検索
3. 必要に応じて他の職員とも相談
4. 記載内容を整理する
5. 回答を送信する
6. 受けた質問と回答(個別的な内容)を基に、QAを一般化してFAQ化
7. FAQを公開する事で、同様のQAを周知する

プロンプト

StepA
実践ガイドブックをサンプルに問い合わせ回答プロンプトを作成

StepB
質問を一般化するプロンプトを作成

StepC
SlackBotとして仮検証

検証アプローチ

・生成AIが一般論から回答することなく、ドキュメントの内容に基づいた正確な回答ができるようなプロンプティング

・データ参照のあり方として、RAG or プロンプトへの情報入力の双方を検証

職員が個別的な質問を一般化し、FAQとして昇華するタスクを実行

追加開発の伴う要件として、本検証期間中にSlackbotとしてデジタル庁/農水省のslackに実践ガイドプロンプトのbotを実装した

検証結果

✓ 実践ガイドブックの内容に基づいた回答の生成を行い、全588問で85.3%の得点率であった。行政職員レベルの正確性をもって回答代行させることは容易でない結果となった。

✓ 長大な文書を学習させるユースケースであったため、RAGによる検証を実施したものの、得点率が実用的でない数値(38.3%)であったため、生成品質が高いシステムプロンプトによる実装が適している結果となった。

✓ 職員の方が汎用的で一般化した質問とGaiXerで一般化した質問の生成を行い比較を行った結果、17問中14問は似たような質問を生成することができた。

・3/19(火) 登録アカウント数：93名 / リクエスト数：約60回
・3/21(木) 登録アカウント数：123名 / リクエスト数：約13回
・3/22(金) 登録アカウント数：146名 / リクエスト数：約60回
・3/25(月) 登録アカウント数：149名 / リクエスト数：約8回
・3/26(火) 登録アカウント数：157名 / リクエスト数：約45回

Step A 生成AI導入前後の作業ステップの変化

▶ 生成AIの導入により、実践ガイドブックを参照する**質問1件当たりの対応工数が85%削減**できる見込み

■生成AI導入前

No	作業ステップ	インプット	アウトプット	工数 （熟練者による作業）	工数算出根拠	工数合計
1	実践ガイドブックを参照し、疑問点を解消する箇所を参照する（必要に応じて複数章を参照）	質問	質問に回答するための参照箇所	5分	担当者へのヒアリング	**20分**
2	記載内容を整理する	質問に回答するための参照箇所	記載内容の理解	10分	担当者へのヒアリング	
3	回答を作成する	記載内容の理解	回答	5分	担当者へのヒアリング	

■生成AI導入後

No	作業ステップ	インプット	アウトプット	工数 （熟練者による作業）	工数算出根拠	工数合計
1	実践ガイドブックを参照して、回答を作成するプロンプトを実行する	質問	回答案	1分	プロンプトに質問を入力して試行する操作の時間として、妥当と考えられる	**3分** **導入前比 -17分**
2	人間によるチェック	回答案	回答	2分	100字程度の質問および回答を確認する時間として、妥当と考えられる	

Step A 業務削減効果とコスト

▶ 全省庁職員からの問い合わせ対応業務に対して、生成AIを導入すると、月間約7.3万円で283時間の業務削減効果の見込み（1,000円で約4時間の業務削減）

■業務削減効果（全省庁職員からの問い合わせ対応に生成AIを導入した場合）

業務削減効果（月間）
= 質問1件当たりの業務削減時間 × 全省庁職員からの質問件数
= 質問1件当たりの業務削減時間 × デジタル庁職員からの質問件数（月間） × 全省庁の職員数/デジタル庁の職員数
= 17分 × 10件※1 × 300,000人/3,000人
= 17,000分
= 283時間

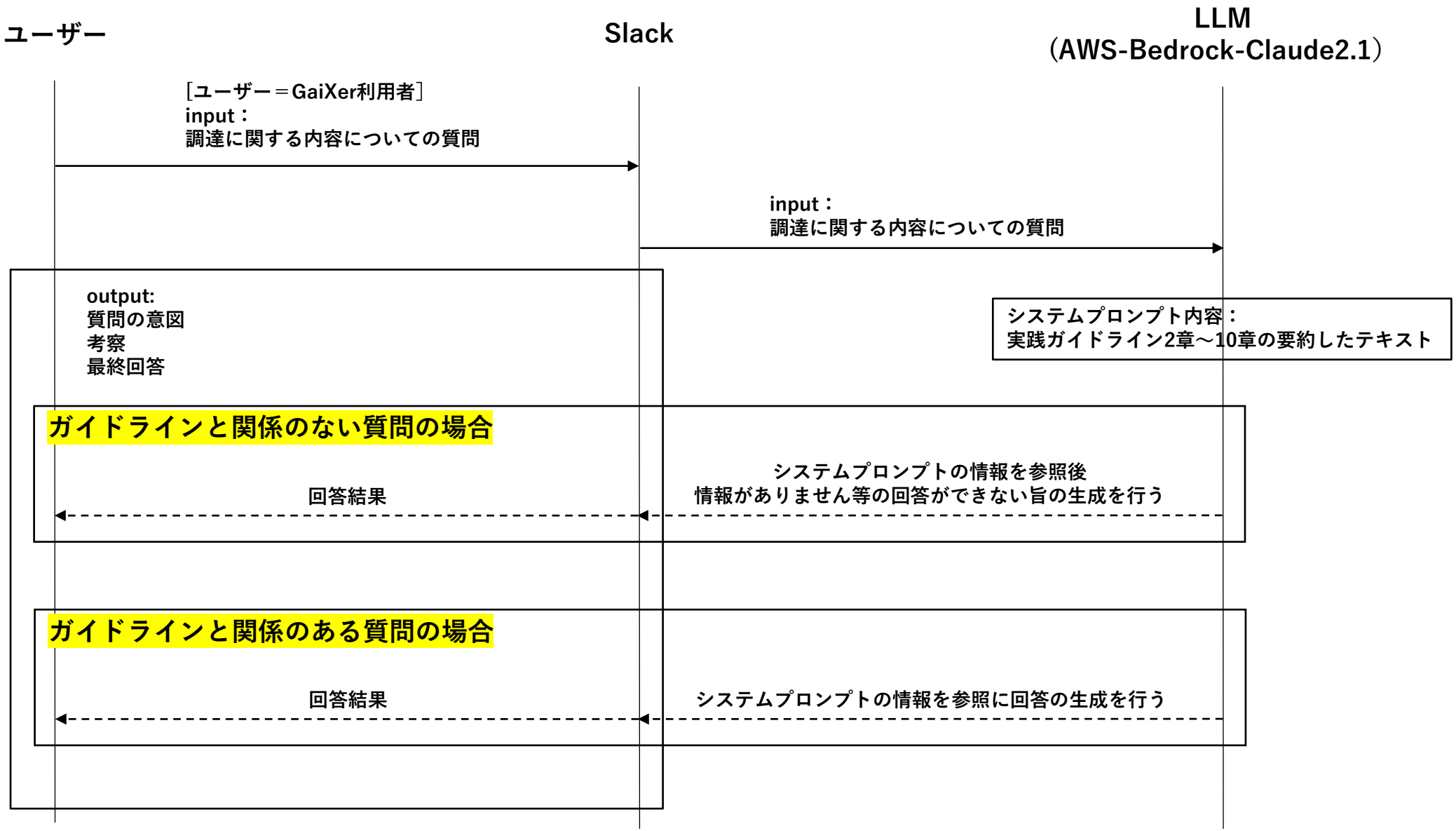
■必要コスト

問い合わせ回答1回あたりのプロンプト試行コスト
= 入力トークン数（k） × 入力料金（ドル/kトークン）＋ 出力トークン数（k） × 出力料金（ドル/kトークン）
= 161※2 × 0.003※3 ＋ 0.540※2 × 0.015※3
= 0.491ドル

必要コスト（年間）
= 問い合わせ回答1回あたりのプロンプト試行コスト × 全省庁職員からの質問件数（月間）
= 0.491ドル × 150（円/ドル） × 1,000件
= 73,650円

※1 デジタル庁職員からの問い合わせは月10~15件程度（担当者へのヒアリングより）、内10件を過去QA・関連ドキュメントより回答生成が可能と仮定
※2 本検証におけるプロンプト（最終版）で試行した際の平均値※実践ガイドブック（446頁）の要約をシステムプロンプトに入れた場合
※3 「Claude 3 Pricing」よりSonnetの料金プランを引用

Step A シーケンス図(問い合わせへの対応)



Step A 評価基準と評価要件

ポイント1)

実践ガイドブックの長大なテキスト情報から正確に参照して回答する必要があるため、「引用精度」を重要な評価基準とした

ポイント2)

ユーザー(職員)が見た際に、最終的に整理された回答が得られているか、を最終アウトプットと定義し、「表現の適切性」を３つ目の評価基準と設定

ポイント3)

本ユースケースにおいて最も重要なポイントはユーザーに対する回答が「ためになるかどうか」と判断し、情報量が多い分には問題ないとしつつ、部分点なども設定

評価基準	要件定義	評価(点数)
回答の正しさ1:引用箇所が正しいかどうか	元々テストケースを作成した時の本文の参照箇所と合致している	2
	テストケースの参照箇所と異なるが、引用箇所が質問に対する回答になりそう	2
	2か所以上引用しており、１つは合致しているが１つは関連がなさそう	2
	2か所以上引用しており、１つは合致しているが１つは明確に間違い	1
	2か所引用が必要だが、1か所しかとってきてない(章またぎの質問など)	1
	引用箇所は正解だが、ナンバリングや見出しが誤り	1
	明らかに引用箇所が誤っている/引用していない	0
回答の正しさ2:回答文が引用箇所を正しく反映しているか	引用箇所に関する表記が正しく記載されている	2
	引用箇所に関する表記がおおむね合っているが、一部誤り(見出しなど)	1
	引用箇所に関する表記が全く異なる(他の章の内容を抽出しているなど)	0
わかりやすさ:表現の仕方が適切であるか	回答が引用や考察を正しく踏まえており、ユーザーの質問に答えている	2
	おおむね答えられているが、不要な表現や小さな誤りも一部見られる	1
	ユーザーの質問に回答できていない	0

Step A 作業/評価ステップ

No.	作業手順	作業内容	評価件数
0	FAQ回答用プロンプトを作成	-	-
1	クイック検証を実施するべく、第２章と第６章の一部を検証	デジ庁職員から、職員が聞くと思われる質問を入手(xls)	33問
2	マス展開に向け、質問を生成するプロンプト	1相当の質問を作成できるレベル感を調整	-
3	プロンプト本文に本文を入れ込む判断により、文書を要約するプロンプトを作成	-	-
4	マス展開開始に備え、2~10章全文をプロンプトに挿入	-	
5	回答コンテキストに全章が参照されるかを確認すべく各章1問づつテスト	-	10問
6	マス展開	単一章への質問/複数章またぎ/選択問題、等複数バリエーションを作成	588問
7	ダブルチェック	担当者を、6と切り替えて全質問のダブルチェック	588問

Step A テストケースに関する詳細

テストケース作成方針

- 参照先が単一章の設問に加え、複数章にまたがる内容に関する設問と実践ガイドブックに無関係な内容の設問を用意した
- 質問のバリエーションが多いユースケースであり、オープンクエスチョン・クローズドクエスチョンのどちらにも対応できるかを確認するため、記述回答問題に加え、選択問題を用意した

テストケースの内訳 参照先(章)/設問タイプ	記述回答問題	選択問題(正を選択)	選択問題(誤を選択)	実践ガイドブックに無関係	総計
2	51	1			52
3	34	1			35
4	56	1			57
5	64	1			65
6	57	1			58
7	44	1			45
8	49		1		50
9	47				47
10	49		1		50
無し				10	10
3と4	34				34
5と6	26				26
7と8	30				30
9と10	29				29
総計	570	6	2	10	588

Step A プロンプトの発展過程　試行１回目 (1/2)

LLM	GPT-4 Turbo

学習セット(RAG)

第2章_プロジェクトの管理(.pdf又は.docx)

プロンプト

#命令
-あなたは優秀なデジタル庁職員です
-あなたは[#Context]内のデータ情報しか有しておりません。
-[#Context]を参照し、Userからの質問に回答してください
-Userの質問から何が知りたいのかの意図を<intention></intention>に記載してください
-回答を行う前にまず[#Context]内のみを参照し、Userの質問に関連する正確な情報を見つけて、<thinking></thinking>の中に一言一句書き出し、参照したファイルと見出しを書き留めてください
-正確な引用を見つけ、十分な情報に基づいて推測できる場合にのみ、Userの質問に回答して下さい。そうでなければ、分かりませんと出力してください
-以下の出力形式に従い出力してください
-学習データ以外からの参照(https~等)を行った場合、<thinking></thinking>の中に記載せず分かりませんと記載してください

＃出力形式
<intention>
</intention>
<thinking>
注意：
Userの質問が複数の回答伴う場合、以下の項目を複数記載してください
参照した学習データ:参照した学習セットの大見出しを記載
見出し：引用した文章の見出しを記載
ページ数：引用した個所のページ数を記載
引用した文章：引用した文章に手を加えずそのまま記載
</thinking>
<最終回答>
thinkingを元に最終回答を生成

作成者の評価

テストケース30問生成を行い、
最終的に引用箇所通りに正しく回答できたのは、
学習セットがPDFファイルの場合15問(Wordファイルの場合5問)
※テストケースの内訳
- ２章に回答がある質問(10問)
- 関係ない質問(10問)
- ２章に関連がある内容だが回答が存在しない質問(10問)

課題#1

- PDFファイル / Wordファイルからは、引用箇所の見出しとページ数の正しい出力を行うことができなかった

Step A プロンプトの発展過程　試行１回目（2/2）

LLM　**GPT-4 Turbo**

学習セット（RAG）

第2章_プロジェクトの管理(.pdf又は.docx)

プロンプト

#命令
-あなたは優秀なデジタル庁職員です
-あなたは[#Context]内のデータ情報しか有しておりません。
-[#Context]を参照し、Userからの質問に回答してください
-Userの質問から何が知りたいのかの意図を<intention></intention>に記載してください
-回答を行う前にまず[#Context]内のみを参照し、Userの質問に関連する正確な情報を見つけて、<thinking></thinking>の中に一言一句書き出し、参照したファイルと見出しを書き留めてください
-正確な引用を見つけ、十分な情報に基づいて推測できる場合にのみ、Userの質問に回答して下さい。そうでなければ、分かりませんと出力してください
-以下の出力形式に従い出力してください
-学習データ以外からの参照(https~等)を行った場合、<thinking></thinking>の中に記載せず分かりませんと記載してください

＃出力形式
<intention>
</intention>
<thinking>
注意：
Userの質問が複数の回答伴う場合、以下の項目を複数記載してください
参照した学習データ:参照した学習セットの大見出しを記載
見出し：引用した文章の見出しを記載
ページ数：引用した個所のページ数を記載
引用した文章：引用した文章に手を加えずそのまま記載
</thinking>
<最終回答>
thinkingを元に最終回答を生成

課題#2

- 図表のテキスト情報をGPTが正確に読み取れず、回答精度に問題が見られた

課題#3

- 与えた役割ではうまく回答することができなかった

Step A 試行１回目の生成結果

質問

2章に回答がある質問
プロジェクトの目標設定時に利用者が困っていることをどのように把握すべきですか？

解答参照個所

利用者が困っている点を把握するためには、現場に直接赴いて調査を行い、
利用者からのフィードバックを収集し、それに基づいて目標を設定することが重要です。
見出し：Step2：1 A 現場で発生している事実をつかんだ上で今後の目標を定める　ページ数：7

生成結果

<intention>
利用者が困っていることを把握する正確な方法を知りたい。
</intention>
<thinking>
参照した学習データ:プロジェクトの管理
見出し：誰が何に困っているのか
ページ数：- 7 -
引用した文章：プロジェクトの立ち上げ時には、利用者が困っていることを把握した上で、その困りごとを解消するための目標を立てます。
その後に詳細な現状把握を行って問題の発生原因を突き止めていくので、場合によっては目標の修正が必要になることもあるでしょう。
</thinking>
<最終回答>
プロジェクトの立ち上げ時には、まず利用者が困っていることを把握し、その困りごとを解消するための目標を立てることが重要です。
その後、詳細な現状把握を行い、問題の発生原因を突き止める過程で、目標の修正が必要になる可能性があります。

Step A プロンプトの発展過程　試行2回目

LLM	Claude2.1

学習セット(RAG)

なし

プロンプト

#命令
- あなたは#ドキュメントからの引用に基づく正確な相談回答ができるコンサルタントです。
-必ず#出力形式に従って出力してください。
-#ドキュメントを参照し、Userからの質問に回答してください
-Userの質問から何が知りたいのかの意図を<intention></intention>に記載してください
-<thinking></thinking>には<intention></intention>に関連する箇所を#ドキュメントから参照し、見出し、ナンバリング、引用を記載してください
- 参照すべき箇所が複数見られる場合は、すべてを<thinking></thinking>に記載してください
-<advice></advice>は<thinking></thinking>をベースにマークダウン形式や箇条書きを用いながら,informativeに回答を生成すること。
- #ドキュメントに関連しないトピックや、考察が不十分な場合はアドバイスを生成しないでください。

#ユーザーからの質問
{ここに質問を入力する質問項目を追加}

出力形式
<intention></intention>
<thinking>
[見出し][ナンバリング][引用文]
</thinking>
<advice></advice>

作成者の評価

全章を要約した文章をシステムプロンプトに組み込むことで、試行1回目よりも引用箇所の精度の向上が見られた。

改善方針#1

- 学習データとしてではなく、システムプロンプトにtext形式で挿入したことで、引用箇所の見出しとページ数を正しく生成できるケースが増えた

システムプロンプト(一部抜粋)

<document>

デジタル・ガバメント推進標準ガイドライン

実践ガイドブック

（第３編第２章 プロジェクトの管理）

| 目次 | ページ |

| Step.1 プロジェクト管理活動全体の流れ | 5 |

| Step.2 プロジェクトの立上げ、初動 | 6 |

| １ 目標とする成果を見定める | 6 |

| A. 現場で発生している事実をつかんだ上で今後の目標を定める | 6 |

| B. 上位計画の目標をブレークダウンし、プロジェクト目標と紐づける | 10 |

| ２ 手段の妥当性を確認する | 17 |

| ３ プロジェクトの投資対効果を算出する | 19 |

| ４ プロジェクトへの投資判断を行う | 20 |

| ５ 機能する体制を作る | 21 |

| A. 制度所管部門、業務実施部門等を含めたＰＪＭＯ体制とする | 21 |

Step A プロンプトの発展過程　試行2回目

LLM　Claude2.1

学習セット(RAG)

なし

プロンプト

#命令
- あなたは#ドキュメントからの引用に基づく正確な相談回答ができるコンサルタントです。
-必ず#出力形式に従って出力してください。
-#ドキュメントを参照し、Userからの質問に回答してください
-Userの質問から何が知りたいのかの意図を<intention></intention>に記載してください
-<thinking></thinking>には<intention></intention>に関連する箇所を#ドキュメントから参照し、見出し、ナンバリング、引用を記載してください
- 参照すべき箇所が複数見られる場合は、すべてを<thinking></thinking>に記載してください
-<advice></advice>は<thinking></thinking>をベースにマークダウン形式や箇条書きを用いながら,informativeに回答を生成すること。
- #ドキュメントに関連しないトピックや、考察が不十分な場合はアドバイスを生成しないでください。

#ユーザーからの質問
{ここに質問を入力する質問項目を追加}

出力形式
<intention></intention>
<thinking>
[見出し][ナンバリング][引用文]
</thinking>
<advice></advice>

改善方針#2

- 全章を要約した文章をシステムプロンプトに入れるため、Contextウィンドウが現時点で一番大きいClaude2.1に変更した

改善方針#3

- 「ドキュメントからの引用~」と役割を与えることで<document></document>内の個所のみを踏まえ回答を行うことができた

課題#4

- 1回目よりも精度は向上したが、プロンプトにマークダウン式とXMLタグが混合しているためノイズになっている可能性が高い

Step A 試行2回目の生成結果

質問

システム監査の全体の流れにおいて、リスクへの対応を可能にするためには、どのような手法が用いられていますか？

解答参照個所

システム監査は、プロジェクトの取り組みが目標に正しく向かっているか、実施プロセスが適切か、機能が仕様どおりか、仕様が目的に適しているかを客観的に評価する活動であり、問題を抽出しリスクへの対応を可能にし、プロジェクト成功へ近づける。見出し:第10章 Step2 1 B

生成結果

<intention>
ユーザーは、システム監査がリスクへの対応を可能にするために、どのような手法が用いられているか知りたいとしています。
</intention>

<thinking>
第3編第10章 システム監査のStep.1に以下のように記載があります。

Step.1: システム監査の全体の流れ
プロジェクトの目標達成に向けて、課題やリスクの管理をする中で、管理方法が適切かどうかの改善点を見落とすことがある。システム監査は、プロジェクトの取り組みが目標に正しく向かっているか、
実施プロセスが適切か、機能が仕様どおりか、仕様が目的に適しているかを客観的に評価する活動であり、問題を抽出しリスクへの対応を可能にし、プロジェクト成功へ近づける。
以下に、システム監査に向けた準備や実施に必要な作業を説明する。
</thinking>

<advice>
システム監査では、プロジェクトの取り組みやプロセス、機能、仕様などを客観的に評価することで、プロジェクトの課題やリスクを抽出し、それらへの対応を可能にしています。

具体的には以下のような手法が用いられています。

- プロジェクトの目標達成に向けた取り組みが適切かを評価
- プロジェクトの実施プロセスが適切かを評価
- 情報システムの機能が仕様どおりかを評価
- 仕様がプロジェクトの目的に適しているかを評価

これらの評価を通じて、プロジェクトの課題やリスクを抽出し、それらへの対応を可能にしています。
</advice>

Step A プロンプトの発展過程　試行3回目

LLM	Claude 2.1

プロンプト

あなたは<document></document>からの引用に基づく正確な相談回答ができるコンサルタントとして、以下のルールに基づいて<回答></回答>を出力してください。
1.userの質問:<question></question>に対して<document></document>を参照し、<回答></回答>を出力する
2.<question></question>の意図を<ユーザーの質問意図></ユーザーの質問意図>に記載してください
3.-<thinking></thinking>には<ユーザーの質問意図></ユーザーの質問意図>に関連する箇所を<document></document>から参照してください
4.answerの作成に複数箇所への参照が必要な場合、すべてを<thinking></thinking>に記載してください
5.<アドバイス></アドバイス>は<thinking></thinking>をベースにマークダウン形式や箇条書きを用いながら,informativeに回答を生成すること
6. <document></document>に関連しないトピックや、考察が不十分な場合はアドバイスを生成しないでください。

<question>
{質問をここに入力する質問項目を追加}
</question>

<回答>
<ユーザーの質問意図></ユーザーの質問意図>
<考察>
<quote>{{chapter}}{{step}}{{number}}{{letter}}{{引用文のテキスト}}</quote>
<!-- `number` と `letter` が存在しない場合は省略 -->
</考察>
<アドバイス></アドバイス>
</回答>

作成者の評価

第2章から第10章までのそれぞれの質問を回答させたところ、10問中8問正しい箇所を引用することができた。

改善方針#4

- Claude2.1に最適なXMLタグ形式で記載することにより認識精度を高めた

課題#5

- 小見出しを引用する場合、見出しではない個所を引用したり、そもそも小見出しが引用されないケースが多い

システムプロンプト(一部抜粋)

改善提案は報告の場で具体例を混ぜながら行う

監査報告書の完成後は、ＰＪＭＯに指摘内容を共有し、表面的解決ではなく効果的な改善へと導くための助言をする。ルール違反だけでなく、その原因を探ることが重要で、他のプロジェクトの好例を参考に具体的な改善提案を行う。

ある省でのシステム監査では、ＰＪＭＯの監査への前向きな改善を促すため、効率的で効果的な管理への提案、改善案の具体例の提示、報告ルートの明確化などの取り組みが行われた。

改善提案は報告の場で具体例を混ぜながら行う

システム監査報告書の様式を把握する

さらに、システム監査報告書の様式を理解し、時間を節約することも推奨されている。

Step.5: 指摘事項を踏まえた改善

PJM0はシステム監査報告書で指摘された内容を基に改善計画を立て、改善を進める。改善計画は監査報告書の改善提案に限らず、運用・保守業務で発生している問題を含めた優先順位を考慮して策定する。内容によっては改善が定着するまで時間がかかる場合があり、継続的なモニタリングが必要である。改善計画や結果、状況は監査責任者及びPMOへ報告し、PMOの確認・評価を受けることで的確な助言を得る。

-

19 -

</document>

Step A 試行3回目の生成結果

質問

総合評価落札方式における加点配分の工夫とはどのようなものがあるのでしょうか？

解答参照個所

参照＜Step.4: 加点の配分を工夫する＞
見出し：Step.4 2 B

生成結果

<intention>
ユーザーは総合評価落札方式における加点配分の工夫について知りたいとしています。
</intention>

<thinking>
[Step.4: 調達仕様書以外のドキュメント作成][2 提案依頼書の内容を工夫する][B. 加点の配分を工夫する]

総合評価落札方式では、評価事項ごとの配点を事前に決める必要があるが、均等に配点するわけではない。「情報システムの調達に係る総合評価落札方式の標準ガイドライン」では、総合評価落札方式を適用できる調達の要
件が示されている。

大規模なプロジェクトや多くの情報システムや組織が絡む案件では、技術的な評価を価格より重視する傾向にある。具体的には、情報システム開発では技術点と価格点の比率を3:1、プロジェクト管理支援では2:1、物品購入
では1:1に設定することが一般的だ。重要な項目には高い配点を設け、基礎点よりも加点に重みを置くことで、価格競争を避ける評価が可能になる。評価項目はプロジェクトの目的に応じて設定し、調達仕様書や要件定義書
と整合性を保つことが重要だ。
</thinking>

<advice>
総合評価落札方式における加点配分の工夫として、以下が挙げられます。

- 技術的な評価を価格よりも重視する傾向にある。情報システム開発では技術点と価格点の比率を3:1に設定することが一般的。

- 重要な評価項目には高い配点を設け、加点に重みを置くことで、価格競争を避ける。

- 評価項目はプロジェクトの目的に応じて設定し、調達仕様書や要件定義書との整合性を保つ。

- 大規模なプロジェクトや多くの組織が関わる案件では、技術評価を価格より重視する傾向にある。

</advice>

Step A プロンプトの発展過程　試行4回目

LLM	Claude 2.1

プロンプト

あなたは<document></document>からの引用に基づく正確な相談回答ができるコンサルタントとして、以下のルールに基づいて<回答></回答>を出力してください。
1.userの質問:<question></question>に対して<document></document>を参照し、<回答></回答>を出力する
2.<question></question>の意図を<ユーザーの質問意図></ユーザーの質問意図>に記載してください
3.-<thinking></thinking>には<ユーザーの質問意図></ユーザーの質問意図>に関連する箇所を<document></document>から参照してください
4.answerの作成に複数箇所への参照が必要な場合、すべてを<thinking></thinking>に記載してください
5.<アドバイス></アドバイス>は<thinking></thinking>をベースにマークダウン形式や箇条書きを用いながら,informativeに回答を生成すること
6. <document></document>に関連しないトピックや、考察が不十分な場合はアドバイスを生成しないでください。

<question>
{質問をここに入力する質問項目を追加}
</question>

<回答>
<ユーザーの質問意図></ユーザーの質問意図>
<考察>
<quote>{{chapter}}{{step}}{{number}}{{letter}}{{引用文のテキスト}}</quote>
<!-- `number` と `letter` が存在しない場合は省略 -->
</考察>
<アドバイス></アドバイス>
</回答>

作成者の評価	588問生成を行い、正解数：493問(得点率85.3%)

システムプロンプト(一部抜粋)

Step.4: 2 B 改善提案は報告の場で具体例を混ぜながら行う

監査報告書の完成後は、ＰＪＭＯに指摘内容を共有し、表面的解決ではなく効果的な改善へと導くための助言をする。ルール違反だけでなく、その原因を探ることが重要で、他のプロジェクトの好例を参考に具体的な改善提案を行う。

ある省でのシステム監査では、ＰＪＭＯの監査への前向きな改善を促すため、効率的で効果的な管理への提案、改善案の具体例の提示、報告ルートの明確化などの取り組みが行われた。

改善提案は報告の場で具体例を混ぜながら行う

Step.4: 2 C システム監査報告書の様式を把握する

さらに、システム監査報告書の様式を理解し、時間を節約することも推奨されている。

改善方針＃5

- ページ数を引用するのではなく、見出しを正確に引用することを重視し、引用する箇所の見出しの階層の小見出しにも目次の番号を付け加えた（Step4:2 B等）

Step A 試行4回目の生成結果

質問

プロジェクトの目標設定時に利用者が困っていることをどのように把握すべきですか？

解答参照個所

利用者が困っている点を把握するためには、現場に直接赴いて調査を行い、
利用者からのフィードバックを収集し、それに基づいて目標を設定することが重要です。
見出し：Step2：1 A 現場で発生している事実をつかんだ上で今後の目標を定める

生成結果

<回答>
<ユーザーの質問意図>
プロジェクトの目標設定時に、利用者がどのようなことに困っているかを把握する方法を質問している。
</ユーザーの質問意図>

<考察>
<quote>第3編第2章 Step.2 1 A 誰が何に困っているのか</quote>
利用者が困っていることを把握するには、現場を調査し、利用者の状況を詳細に把握する必要があると記載されている。
</考察>

<アドバイス>
プロジェクトの目標設定時に利用者が困っていることを把握するには、以下の方法が効果的です。

- 現場を訪問し、利用者に直接ヒアリングを行う
- 利用者を分類し、それぞれのニーズや課題を把握する
- 申請から結果通知までの一連の流れを可視化し、利用者の行動や感情を洗い出す
- 問い合わせや要望の内容を詳細に分析する
- 実データや統計情報から利用者の特徴や課題を把握する

これらの方法を組み合わせて利用者が困っていることを多角的に把握し、プロジェクトの目標設定に反映させることが重要です。
</アドバイス>
</回答>

Step A 試行4回目の評価結果

▶ 評価基準3項目において、平均得点率は85.3%であった

ポイント1）

「わかりやすさ」の指標における得点率は「回答の正しさ１」の得点率と相関が高い。引用がない場合に、一般論から回答する傾向を防止できている点が効いているように思慮される。

ポイント2）

「回答の正しさ2」において、複数章にまたぐ質問の得点率が低かった。
引用箇所は正確だがナンバリングや見出しが誤りとなる質問が多くみられた。
（複数章にまたぐ質問の内、引用得点が満点でなかった61問中26問が該当する。具体例は次ページに記載。）

	回答の正しさ1: 引用箇所が正しいかどうか ※複数テストケースを実施した平均点(2点満点中)	得点率	回答の正しさ2: 回答文が引用箇所を正しく反映しているか ※複数テストケースを実施した平均点 (2点満点中)	得点率	わかりやすさ: 表現の仕方が適切であるか ※複数テストケースを実施した平均点(2点満点中)	得点率
2章	1.88	94%	1.98	99%	1.90	95%
3章	1.60	80%	1.91	96%	1.77	89%
4章	1.65	82%	1.77	89%	1.73	87%
5章	1.37	68%	1.80	90%	1.83	91%
6章	1.31	66%	1.64	82%	1.64	82%
7章	1.73	87%	2.00	100%	1.98	99%
8章	1.40	70%	1.92	96%	1.76	88%
9章	1.48	74%	1.85	92%	1.76	88%
10章	1.66	83%	1.74	87%	1.58	79%
なし	2.00	100%	1.90	95%	2.00	100%
3＆4章	1.18	59%	1.65	82%	1.15	57%
5＆6章	1.46	73%	1.85	92%	1.54	77%
7＆8章	1.60	80%	1.60	80%	1.48	74%
9＆10章	1.34	67%	1.69	84%	1.41	71%
平均	1.88	77%	1.8	90%	1.7	84%

Step A システムプロンプトとRAGの評価結果の比較

- ▶ 単一章のテストケースの内、各章2問ずつシステムプロンプトによる生成の評価結果が低かったテストケースにおいて、RAGによる生成の評価を実施した
- ▶ システムプロンプトと比較して、RAGは得点率が26%低く、実用にあたってはシステムプロンプトを活用すべきという結果となった

章	システムプロンプト使用時			RAG使用時		
	回答の正しさ1	回答の正しさ2	分かりやすさ	回答の正しさ1	回答の正しさ2	分かりやすさ
2章	0	4	4	0	1	2
3章	0	4	4	0	1	1
4章	2	2	0	1	1	1
5章	0	3	0	0	2	1
6章	0	0	0	0	0	0
7章	0	4	4	0	1	0
8 章	2	4	2	1	4	2
9 章	2	4	4	2	3	2
１０章	4	4	4	2	2	2
合計得点	**10**	**29**	**22**	**6**	**15**	**11**
得点率	**27%**	**81%**	**61%**	**17%**	**42%**	**31%**
合計得点率	**56%**			**30%**		

Step A 複数章にまたぐ質問の詳細結果

質問

業務フローに紐づける効果の積算時に注意すべき点と、
情報セキュリティ対策が不十分で後日問題が発生した場合の対応について教えてください。

解答参照箇所

３章：原課の予算担当、会計課、ＰＭＯにおいても監査結果による新たな情報セキュリティ対策が必要なこと、
　　そして予算要求が必要だったことを誰も知りませんでした。
４章：事業者と役割分担して作業を進める

生成結果

<回答>
<ユーザーの質問意図>
ユーザーは、業務フローに効果を紐づける際の注意点と、情報セキュリティ対策が不十分な場合の対応について質問しています。
</ユーザーの質問意図>
<考察>
<quote>第3編第4章 Step.5 2 G</quote>
業務フローの効果と効果積算の方法

業務フローを作る際は、変更点とその効果をわかりやすく記載する。吹き出しの形式で効果を示すことで、関係者に目指す姿を共有できる。効果積算の基本は「1件あたりの効果」×「件数」である。

<quote>第3編第4章 Step.5 2 G</quote>
効果積算を行う際は、サンプリング調査と効果想定の誤差に注意が必要である。サービスや業務の処理単位、実施担当部門、処理場所等で適切に対象を分類し、詳細な効果積算を行うべきである。条件が同じと考えられるグループに分けて各グループで調査と積算を行うことがポイントである。

<quote>第3編第9章 Step.2 1 J</quote>
セキュリティ対策が不十分な場合の対応

情報セキュリティ侵害またはその恐れがある場合は、関係府省に速やかに報告する。
</考察>

<アドバイス>
業務フローに効果を紐づける際の注意点は以下の通りです。

- サービスや業務の処理単位、実施担当部門、処理場所などで適切に対象を分類し、詳細な効果積算を行う
- サンプリング調査と効果想定の誤差に注意する

情報セキュリティ対策が不十分な場合は、関係府省に速やかに報告する必要があります。
</アドバイス>
</回答>

本来の参照箇所は3章と4章であり、生成結果は正確に引用できているが、第3章の引用箇所なのに、第9章と記載されている(引用箇所は３章のものを参照している)。このようにナンバリングや見出しが誤っている生成結果が多数見られた

Step B 質問の変換(個別的内容の一般化)に関する精度検証結果

▶ 職員がSlackで受けた調達に関する個別具体的な質問を一般化するタスクを生成AIで実行したところ、13問/17問(76%)で職員の作成する粒度での質問を生成できた

評価結果	質問数	総評
合計	17	全体で76%の精度で一般化することが出来た。
職員が作成した質問＆回答と非常に似通っている	13	文章量が生成AIの方が多いが、職員が作成するレベル感/粒度で一般化することは出来た。
職員が作成した質問＆回答とやや似通っている	4	元の質問＆回答には記載していない記述が生成され、手直しが必要に思えた回答内容が見られた。
職員が作成した質問＆回答と異なる	0	元々の個別質問から逸脱したような変換は認められなかった。

▶ そのままチャットボットのFAQとして掲載

▶ 職員が微修正してFAQとして掲載

Step B 質問の変換(個別的内容の一般化)に関する検証 (質問)

元の質問	職員の方が一般化した質問	GaiXerが一般化した質問
プロポーザル型企画競争でのWTO案件について、限定入札にあたるので、資料提供招請や意見招請の手続きは不要と考えていますが、公告は、官報の掲載と50日以上の期間が必須となるでしょうか。	プロポーザル型企画競争でも、意見招請や官報掲載は必要でしょうか。	プロポーザル型企画競争における入札の場合、資料提供招請や意見招請の手続き、および官報掲載は必要でしょうか。
ベンダーロックイン防止チェックリストについてご相談です。 少額随契の場合、調達手法の選択肢にないのですがどうしたらよいでしょうか。	ベンダーロックイン防止チェックリストの作成対象となる調達手法をお教えください。	少額随意契約の場合、ベンダーロックイン防止チェックリストの対象外となっておりますが、どのように対応すればよいでしょうか。
プロポーザル型企画競争の審査員について質問させてください。 マニュアルには3名以上となっておりますが、監督職員、検査職員含めて3名以上となりますでしょうか(監督職員、検査職員とは別で3名以上となるのか)	審査者の選定について、何かルールはあるでしょうか。	プロポーザル型企画競争における審査員の構成について教えてください。マニュアルには審査員が3名以上と記載されていますが、監督職員や検査職員を含めて3名以上とするのでしょうか。それとも監督職員や検査職員とは別に3名以上の審査員を置く必要があるのでしょうか。

Step C 質問の変換(個別的内容の一般化)に関する検証 (質問)

元の回答	職員の方が一般化した回答	GaiXerが一般化した回答
ご認識の通り、企画競争の場合は、資料提供招請や意見招請の手続きは任意となります。また、官報の掲載も不要となります。詳細については、契約班のご担当者にご確認下さい。 ただ、意見招請を行わない場合、技術的対話による企画競争とは異なり、事業者側の意見を聞く機会が少なくなります。任意となりますが、意見招請を実施することや、RFI（情報提供依頼）をご活用頂ければと思います。	プロポーザル型企画競争は、資料提供招請・官報の掲載も不要となります。 ただし、意見招請を行わないため、事業者側の意見を聞く機会が少なくなります。必要に応じて、任意となりますが意見招請の実施や、RFI（情報提供依頼）をご活用頂ければと思います。	プロポーザル型企画競争の場合、資料提供招請や意見招請の手続きは任意です。また、官報掲載も不要です。ただし、意見招請を行わないと、事業者側の意見を聞く機会が少なくなります。必要に応じて意見招請の実施やRFI(情報提供依頼)の活用をお勧めします。詳細は契約班に確認してください。
ベンダーロックイン防止チェックリストは、少額随契は対象外となります。 もし契約班の方から、提出を求められる等、ございましたら、再度ご相談下さい。 チェックリストに、公募随契等は想定し、リストを作っておりましたが、ご指摘頂いた少額随契は考慮出来ておりませんでした。 次回のチェックリスト公開時に、反映したいと思います。	適切な競争環境を構築すること目的としており、チェックリストでは、下記の通り対象・対象外を示しています。 【作成対象】 ・一般競争入札（最低価格落札方式／総合評価落札方式） ・技術的対話による一般競争入札（総合評価落札方式） ・技術的対話による企画競争 ・企画競争入札（プロポーザル型企画競争含む） 【対象外：記載は不要で、ファイルのみ提出下さい】 ・特命随意契約となる案件 ・公募随契する案件 ・少額随契となる案件 ・変更契約	少額随意契約につきましては、ベンダーロックイン防止チェックリストの対象外となっております。契約担当者の方から提出を求められる場合は別途ご相談くださいますよう、お願い申し上げます。チェックリストは公募型随意契約等を想定して作成されており、少額随意契約の場合は考慮されていないことをご指摘いただき、誠にありがとうございます。今後、チェックリストの改訂の際には、少額随意契約の場合の対応についても検討させていただきます。
監督職員、検査職員含めて3名以上、がマニュアル上の推奨となります。（一般競争入札（総合評価）も同様です）この際、必ずしも監督職員、検査職員を審査者にする必要はありません。 ３名以上としているのは、１人の審査者の評価・意見に必要以上に引きずられないようにするためとなります。	一般競争入札（総合評価）やプロポーザル型企画競争、技術的対話等の調達における審査者については、３者以上で審査を行うことを原則としています。 また審査者選定の際、下記の知見について、審査者チームとして少なくとも１者以上有することを原則としています。 ①政策・サービス分野に関する知見 ②プロジェクトマネジメントに関する知見 ③システムの技術に関する知見	プロポーザル型企画競争においても、一般競争入札(総合評価)と同様に、審査員は監督職員や検査職員を含めて3名以上で構成することが推奨されております。ただし、必ずしも監督職員や検査職員を審査員に含める必要はございません。審査員を3名以上とする理由は、1人の審査員の評価や意見に過度に影響されないようにするためであります。

実践ガイドブック総括

総括　：　職員の問合せや情報収集には正確性が求められるため、FAQの整備やチャットボットとの組み合わせなどを通じて、シナリオ作成の効率化を実現しながら回答正確性を担保することが推奨される。
また、将来的なLLMの性能向上に応じたLLMの対応領域の拡大を検証したい

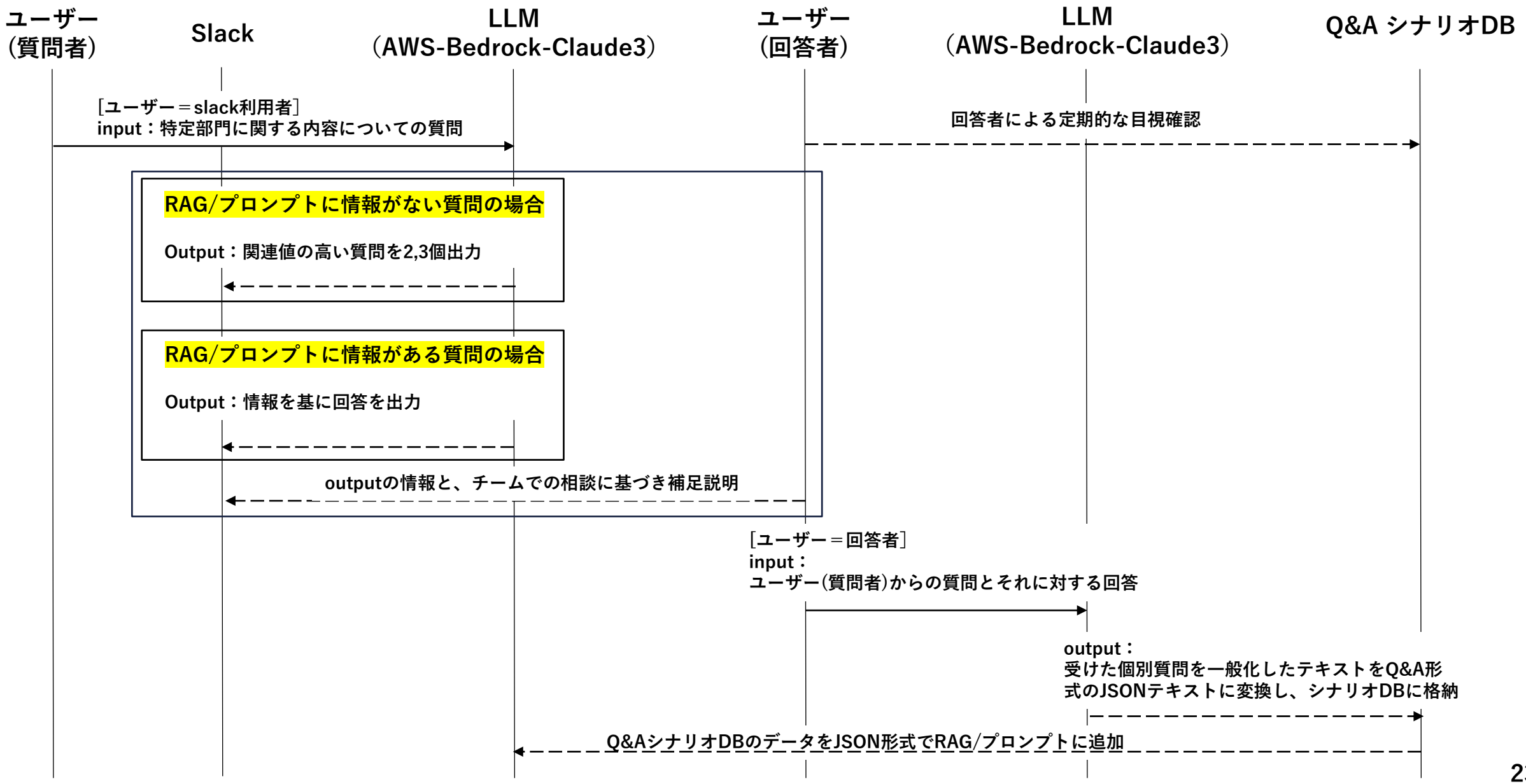
考察　：

- RAG有無にかかわらず、LLMを用いて行政職員レベルの正確性を以て回答代行させることは容易でない結果となった
- 一方で、職員が行う一部の業務(ドキュメントからの要約、個別質問の一般化)等への活用は有効と判定できる
- 故に、システム化の際には、従来型のチャットボット等との組み合わせを通じて、正確な回答と、効率的にFAQを増やす仕組みが有効であると思慮

実装形式

実装レベル	実装形式	コスト構造	利点	課題
検証	JSON形式化されたFAQをRAGで実装し、個別質問で受けた問い合わせをFAQとして一般化し、新たなQAを追加。個別質問を受けた際には、生成AIでFAQから回答を参照して正確に抽出	SaaS利用料(トークン量込)	本利用に向けた本格検証が可能。プロンプトの組織管理/共有が可能なため、業務利用も可能。	FAQから忠実に回答する正確性に対応できるLLMの選定(GPT-4-TurboやClaude 2.1等では一定精度までが限界だが、今後のLLM性能に応じてその限りでない)
将来像	後述のシーケンス図に従い、Slack環境にLLM、Chatbot、FAQ DBを実装し、個別質問への回答と共に、FAQナレッジベースの拡充ができるサービスとして実装	• システム実装コスト • チャットボット運用コスト • トークン利用料 • インフラ運用コスト	チャットボットとして搭載の為決められた回答のみを出力。生成AIは質問の受け手として機能の為、ハルシネーションリスクがヘッジできる。	FAQのレビューや回答内容の確認ステップなど、全自動運用は難しいため、将来のLLM性能の向上に応じたLLM対応領域の拡大が求められる

StepA-C 特定部門への問合せ対応に関する将来像 シーケンス図



検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

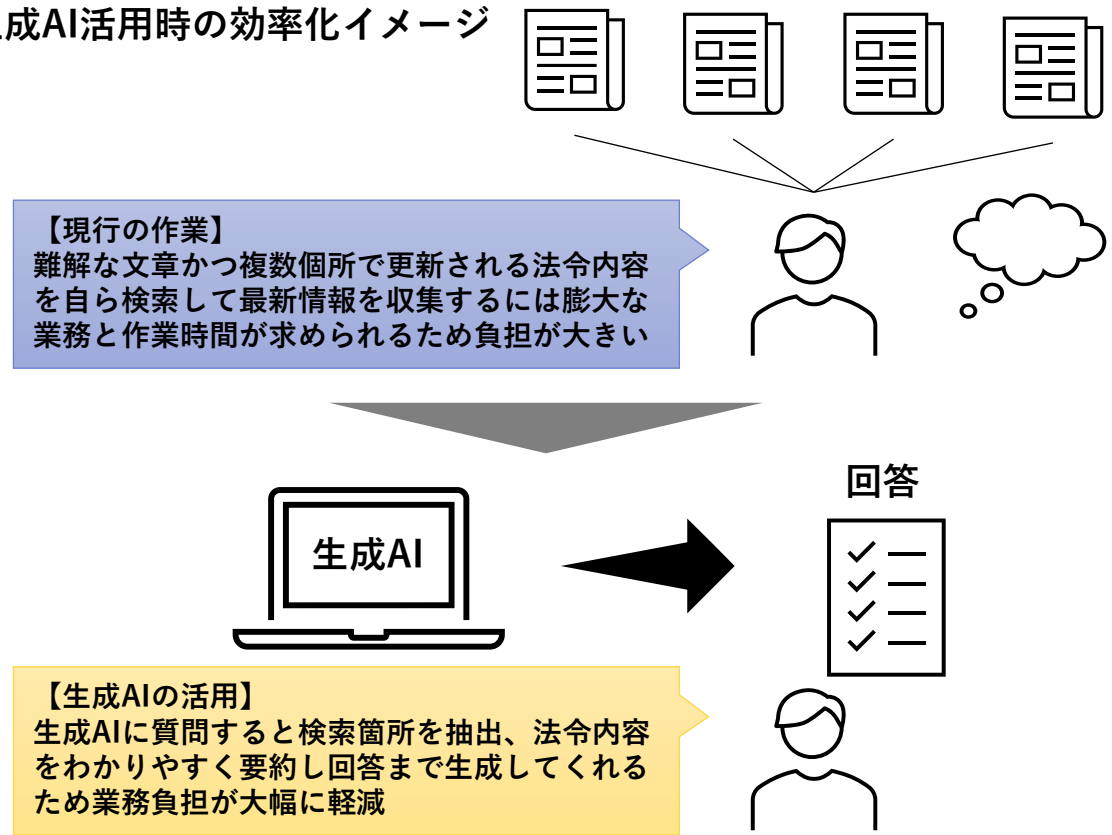
B-2. 人事院規則

ユースケース概要

前提	政府では人事関連の規則類が人事院勧告により年度ごとに更新される。その際、府省庁から自治体に至るまで給与や人事システムなどの担当者が法令に基づいた確認と更新作業を受けており、法令の検索、読み込みに膨大な労力を割いている
課題	難解かつ膨大な文章量で、毎年変更がなされる人事院規則を読み込んで理解し給与計算等を行うことは困難であり、業務に多大な時間がかかることが課題

人事院規則九―六(俸給の調整額)（昭和三十二年人事院規則九―六）

生成AI活用時の効率化イメージ



B-2. 人事院規則

作業分解とプロンプトの関係性

目的	・ 法令の該当箇所の検索、内容のわかりやすい要約、質問内容への対策の提示を行うことで行政職員の業務を効率化する

作業工程

1. 記載箇所を検索する
2. 検索箇所の内容を確認して回答を得る

プロンプト

1.参考箇所を出力し、それに基づいたアドバイスをする

検証アプローチ

・人事院規則のPDFファイルをテキストファイルにして学習させる。

・プロンプトをXMLタグを使用したものに変更し、引用箇所を細かく指定する

・ハルシネーション防止のためWeb検索をさせず、学習セット内の情報に限定して回答させる

検証結果

✓ 引用箇所の最終得点率は72%であったが、部分的には適正に抽出ができているものもあるため実用に耐えうる結果である

✓ 引用箇所の内容反映の最終得点率は84%であった。一部誤りの箇所が見受けられる場合は部分点としているが、ユーザーへの回答として一定評価できるだけの精度を出せている

B-2. 人事院規則

現行の作業ステップ

▶ 人事院規則に基づいて、俸給を算出するためのテーブルの最新化の対応において、現行では16人日の作業を要している
「俸給に基づいた改正内容確認」・「改正内容と改正方法の検討」において、生成AIを活用できる見込み

生成AI導入前後で差分のあるステップ

No	作業ステップ	インプット	アウトプット	工数(人日)	工数合計
1	部内でのupdate、人事院からの説明会(例年8月)	法令改正の連絡	最新の法令の詳細	-	16人日
2	年一回の俸給に基づいた改正内容を確認する	最新の法令の詳細	最新の法令の理解	1	
3	改正内容と改正方法を検討する	最新の法令の理解	システム更新に必要な情報	3	
4	改正に必要な各種資料をファイルサーバーに保管する	システム更新に必要な情報	システム更新に必要な資料	1	
5	各種俸給テーブルなどの手動での書き換え	システム更新に必要な資料	更新した各種俸給テーブル	3	
6	俸給の計算ロジックを理解し、エクセルファイル上に全て書き換える	更新した各種俸給テーブル	各種俸給テーブルチェックのためのエクセルファイル	2	
7	チェック	各種俸給テーブルチェックのためのエクセルファイル	正確な各種俸給テーブル	2	
8	事業者にエクセルファイルを送付	正確な各種俸給テーブル	SQL変換するエクセルファイル	1	
9	事業者がSQL変換しシステムに反映	SQL変換するエクセルファイル	SQL/システムへの反映	3	

B-2. 人事院規則

生成AI導入前後の作業ステップの変化

▶ 俸給を算出するためのテーブルの最新化の対応において、生成AIを活用すると1.3人日の業務削減の見込み

■生成AI導入前

No	作業ステップ	インプット	アウトプット	工数(人日)	工数算出根拠	工数合計
2	年一回の俸給に基づいた改正内容を確認する	最新の法令の詳細	最新の法令の理解	1	担当者へのヒアリング	4人日
3	改正内容と改正方法を検討する	最新の法令の理解	システム更新に必要な情報	3	担当者へのヒアリング	

■生成AI導入後

No	作業ステップ	インプット	アウトプット	工数(人日)	工数算出根拠	工数合計
2	年一回の俸給に基づいた改正内容を確認する	最新の法令の詳細(テキスト形式)	最新の法令の理解(テキスト形式)	0.7	生成AIによる参照箇所の検索・要約の提示により、改正内容の理解の改正方法の検討時間が3分の1程度削減できると考えられる	2.7人日 導入前比 -1.3人日
3	改正内容と改正方法を検討する	最新の法令の理解(テキスト形式)	システム更新に必要な情報(テキスト形式)	2		

業務削減効果とコスト

▶ 俸給を算出するためのテーブルの最新化の対応業務において、府省庁および全国の自治体に生成AIを導入すると、
約6,840万円で2,250人日の業務削減効果の見込み(1,000円で約47分の業務削減)

■業務削減効果(府省庁＋全国の自治体の俸給算出対応に生成AIを導入した場合)

業務削減効果(月間)
＝1組織当たりの業務削減時間 × 俸給算出業務の実施組織数(全国)
＝1組織当たりの業務削減時間 × (府省庁数 + 全国の自治体数)
＝1.3人日 × (13 + 1719※1)
＝2,250人日

■必要コスト

1回あたりのプロンプト試行コスト
＝入力トークン数(k) × 入力料金(ドル/kトークン) + 出力トークン数(k) × 出力料金(ドル/kトークン)
＝58.2※2 × 0.003※3 + 573※2 × 0.015※3
＝8.77ドル

必要コスト(年間)
＝1回あたりのプロンプト試行コスト× 本業務におけるプロンプト試行回数(30回と仮定) ×
俸給算出業務の実施組織数(全国)
＝8.77ドル × 150(円/ドル) × 30回 × 1,732件
＝6,840万円

※1 「総務省トップ > 政策 > 地方行財政 > 地方自治制度 > 広域行政・市町村合併」より引用
※2 本検証におけるプロンプト(最終版)で試行した際の平均値
※3 「Claude 3 Pricing」よりSonnetの料金プランを引用

生成AI導入による効果

▶ 法令文そのもののや計算ロジック意味理解の短縮化に加え、業務ナレッジへの依存度が下がる期待から、
属人的にならず業務持続性が改善される見込み

効果	効果内容
業務持続性	難解な法令を読み解き、計算ロジックを正確に計算するベテラン職員が必要だが、本ユースケースの展開により属人性から脱却できる可能性あり
	本サービスを府省庁のみならず自治体にまで展開の場合、特に人手不足で課題を抱える自治体の事業継続性に貢献する事が期待できる

B-2. 人事院規則

評価基準と評価要件

ポイント1)

回答のすべてが正確でなくとも、一部が正確な場合はユーザーにとって有用な情報であると判断したため、一部の内容が正確である場合には部分点を設定した。

ポイント2)

長大で複数あるテキストや表の情報から正確に参照して回答する必要があるため、「引用精度」を重要な評価基準とした。

ポイント3)

引用箇所が正しくとも、抽出した表の内容の一部でも誤っている場合は誤回答へ直結するため、評価を0点とした。

ポイント4)

ユーザーの質問に対して回答が得られているかを観点とし、「表現の適切性」を設定した。

評価基準	要件定義	評価(点数)
回答の正しさ1:引用箇所が正しいかどうか	元々テストケースを作成した時の本文の参照箇所と合致している	2
	テストケースの参照箇所と異なるが、引用箇所が質問に対する回答になりそう	2
	2か所以上引用しており、１つは合致しているが１つは関連がなさそう	2
	2か所以上引用しており、１つは合致しているが１つは明確に間違い	1
	2か所引用が必要だが、1か所しかとってきてない	1
	引用箇所は正解だが、ナンバリングや見出しが誤り	1
	明らかに引用箇所が誤っている/引用していない	0
回答の正しさ2:回答文が引用箇所を正しく反映しているか	引用箇所に関する表記が正しく記載されている(テーブル)	2
	引用箇所に関する表記が、全く異なる、もしくは一部誤り(テーブル)	0
	引用箇所に関する表記が正しく記載されている	2
	引用箇所に関する表記がおおむね合っているが、一部誤り(見出しなど)	1
	引用箇所に関する表記が全く異なる(他の文章の内容を抽出しているなど)	0
わかりやすさ:表現の仕方が適切であるか	回答が引用や考察を正しく踏まえており、ユーザーの質問に答えている	2
	おおむね答えられているが、不要な表現や小さな誤りも一部見られる	1
	ユーザーの質問に回答できていない	0

B-2. 人事院規則

作業/評価ステップ

No.	検証段階	アクション	アクションの理由(簡潔に)
1	クイック検証	プロンプト開発	引用に特化したプロンプトを開発した
2	クイック検証	「質問/相談/抽出」の3つの分類のテストケースを20件出力させる	様々な質問形式に適切な回答ができるかを確認
3	クイック検証	HTML/XML/PDF/Wordなど様々なファイルでの学習セット登録を検証	RAG利用時に、精度が高いファイル形式を確認する為
4	本検証	RAG(PDFテキスト)をベースに検証開始(テストケース数：20件)	その他形式では表などが崩れて反映され、精度を毀損してしまうため
5	本検証	ユーザーミーティング	現職ユーザーに使用感を確認してもらうため
6	本検証	職員が実利用する全てのデータを学習セットに登録し、最終検証として実施	現職の方からヒアリングした、実業務でよく利用する情報を学習セットへ追加することで精度の違いが表れるか確認するため
7	追加検証	使用するLLMを「Claude 3」に切り替えて、6で誤りだったケースを再度検証	言語モデルのアップグレードによる精度向上を確認するため

テストケースに関する詳細

テストケース作成方針

ユーザーの需要がニーズが高かった階級別の給与計算方法を回答するケースの検証のため、
階級表や俸給表等の表を正確に読み取れるか、正確な計算ロジックを提示できるかを検証すべく、
表の抽出に重きを置いたテストケースを作成した

令和５年 人事院勧告
勧告日　８月７日(月)

人事院総裁　川　本　裕　子

人事院は、国家公務員法、一般職の職員の給与に関する法律、一般職の職員の勤務時間、休暇等に関する法律等に基づき、公務員人事管理について別紙第１のとおり報告し、一般職の職員の勤務時間について別紙第２のとおり勧告するとともに、一般職の職員の給与について別紙第３のとおり報告し、別紙第４のとおり勧告する。

教育職俸給表
イ　教育職俸給表(一)

職員の区分	職務の級 / 号俸	1級 俸給月額	2級 俸給月額	3級 俸給月額	4級 俸給月額	5級 俸給月額
		円	円	円	円	円
	1	233,100	290,700	335,[illegible]00	410,200	535,900
	2	235,400	293,300	338,500	412,500	538,900
	3	237,600	295,700	341,500	414,600	542,000
	4	239,600	298,000	344,500	416,700	545,100
	5	241,700	300,300	347,400	418,600	548,100
	6	243,400	302,600	349,800	421,000	550,500
	7	245,100	304,700	352,300	423,200	553,000
	8	246,900	306,900	354,700	425,500	555,400

例）
令和五年人事院勧告(左記上段）にて、一般職の職員の給与について勧告された別紙４(左記下段）の俸給表の中から、質問の回答箇所にあたる「教育職職員の3級の号俸3の俸給月額」を抽出し回答している

質問：大学に準ずる教育職で働く定年前再任用短時間勤務職員以外の職員の3級の号俸3の俸給月額はいくらですか？

生成回答

<output>
<intention>ユーザーは大学に準ずる教育施設で働く定年前再任用短時間勤務職員以外の職員の3級の号俸3の俸給月額を知りたいとしています。</intention>
<thoughts>ユーザーの質問は具体的で明確です。法令の該当箇所を見つけ、要求された情報を正確に回答する必要があります。</thoughts>
<回答>
{{clause}}教育職俸給表(一){{number}}341,500円{{quote}}3級 俸給月額{{meaning}}定年前再任用短時間勤務職員以外の職員の3級の号俸3の俸給月額は341,500円である。
</回答>
</output>

B-2. 人事院規則

プロンプトの発展過程　試行1回目

LLM	GPT-4 Turbo

学習セット

人事院規則九―六（俸給の調整額）（昭和三十二年人事院規則九―六）(.PDF)

プロンプト

#命令
- あなたは#ドキュメントからの正確な引用に基づく回答ができる法律コンサルタントです。
-必ず#出力形式に従って出力してください。
-ユーザーの質問意図には、ユーザーが質問を通じて理解したいことを記載してください。
-引用には、#ドキュメントからユーザーの質問に関連するドキュメント名と条項を記載し、原文をそのまま引用してください。
-アドバイスは引用をベースにマークダウン形式や箇条書きを用いながら、informativeな回答を生成すること。
- #ドキュメントに関連しないトピックや、引用の情報量が不十分な場合はアドバイスを生成しないでください。

＃出力形式
1.ユーザーの質問意図：
2.引用：
[引用元のドキュメント名]の[条項番号]には下記記載があります;[原文]
3.アドバイス:[引用に基づいた最終回答]
4.引用箇所:[引用元のドキュメント名],[項目名],[見出し]

#ユーザーからの質問
端数計算の規定について教えてください。

作成者の評価	10問中6問で評価基準3項目の最高得点を獲得した。それなりに高い精度であるものの改善が必要である

課題#1

- 学習させたいURLからテキストデータを抽出することの難易度が高いため、PDF形式で学習セットに登録し検証したが、引用箇所の反映、特に表の読み取り精度が悪い

課題#2

- Web検索を行ってしまい、学習セットに登録した情報のみから引用を行わないケースが見られる

B-2. 人事院規則

試行1回目の評価項目

課題#3

・検証初期は明確な論点が定まっていない評価基準を用いていた。
・検証を繰り返しす中で得られた結果から、より明確な評価ポイントを検討し新たな評価基準を策定する必要がある

評価項目

評価基準	要件定義	評価(点数)
回答の正しさ1:引用箇所が正しいかどうか	元々テストケースを作成した時の本文の参照箇所と合致している	3
	引用箇所は正解だが、ナンバリングや見出しが誤り	2
	引用箇所の付近を引用している	1
	明らかに引用箇所が誤っている/引用していない	1
回答の正しさ2:回答文が引用箇所を正しく反映しているか	引用箇所に関する表記が正しく記載されている	3
	引用箇所に関する表記がおおむね合っているが、一部誤り	2
	引用箇所に関する表記が全く異なる(他の章の内容を抽出しているなど)	1
わかりやすさ:表現の仕方が適切であるか	回答が引用や考察を正しく踏まえており、ユーザーの質問に答えている	3
	おおむね答えられているが、不要な表現や小さな誤りも一部見られる	2
	ユーザーの質問に回答できていない	1
最終評価	質問に対して正しい回答をしている	1
	質問に対しての回答に一部誤りはあるがおおよそが合っている	0.5
	質問に対しての回答が全く異なっている	0

B-2. 人事院規則

試行1回目の生成結果

質問

「特殊警備救難艇」に乗り組む職員の調整数はいくつですか？

解答参照個所

（３）　特殊警備救難艇で全長二十メートル未満のもの（特殊警備救難艇にあっては、人事院の定めるものに限る。）に乗り組む職員

生成結果

1.ユーザーの質問意図：
特殊警備救難艇に乗り組む職員に関して適用される調整数について知りたい。

2.引用：
「海上保安庁の職員」の「（５）　巡視艇又は特殊警備救難艇で全長二十メートル未満のもの（特殊警備救難艇にあっては、人事院の定めるものに限る。）に乗り組む職員」

3.アドバイス:
特殊警備救難艇に乗り組む職員の調整数の具体的な数値については、引用された情報だけでは判断できません。詳細な調整数は、人事院の定めるところによりますので、関連する人事院の規定を確認する必要があります。

4.引用箇所:
海上保安庁の職員,（５）,巡視艇又は特殊警備救難艇に乗り組む職員

B-2. 人事院規則

プロンプトの発展過程　試行2回目

LLM	Claude 2.1

学習セット

人事院規則九―六（俸給の調整額）（昭和三十二年人事院規則九―六）（.text）

プロンプト

#命令 -あなたは優秀なチャットボットです。
-あなたは[#Context]内のデータ情報しか有しておりません。
-[#Context]を参照し、Userからの質問に回答してください
-Userの質問から何が知りたいのかの意図を<thinking></thinking>に記載してください
-回答を行う前にまず[#Context]内のみを参照し、Userの質問に関連する正確な情報を見つけて、<thinking></thinking>の中に一言一句書き出し、参照したファイルと見出しを書き留めてください
-正確な引用を見つけ、十分な情報に基づいて推測できる場合にのみ、Userの質問に回答して下さい。
そうでなければ、分かりませんと出力してください
-必ず以下の出力形式に従い出力してください
-学習データ以外からの参照(https~等)を行った場合、<thinking></thinking>の中に記載せず分かりませんと記載してください
出力形式
<intention>
</intention>
<thinking>
注意：
Userの質問が複数の回答伴う場合、以下の項目を複数記載してください
参照した学習データ:参照した学習セットの大見出しを記載
見出し：引用した文章の見出しを記載
引用した文章：引用した文章に手を加えずそのまま記載 </thinking>

<summary>
<引用した文章を法律に詳しくない人向けに要約>
</summary>

<最終回答>
thinkingを元に最終回答を生成

改善方針#1

- 学習データの形式をPDFやHTMLファイル、XMLファイルから.txt形式に変更

改善方針#2

- 学習セットに登録した情報のみから引用を行う類似の他ユースケースにおいて、Claude2.1に適した、[#Context]内の情報から回答するためのプロンプト（Web検索をする回答やハルシネーションを起こさないプロンプト）が作成できたため、同様にClaude2.1を採用

B-2. 人事院規則

プロンプトの発展過程　試行2回目

LLM	Claude 2.1

学習セット

人事院規則九―六（俸給の調整額）（昭和三十二年人事院規則九―六）(.text)

プロンプト

#命令 -あなたは優秀なチャットボットです。
-あなたは[#Context]内のデータ情報しか有しておりません。
-[#Context]を参照し、Userからの質問に回答してください
-Userの質問から何が知りたいのかの意図を<thinking></thinking>に記載してください
-回答を行う前にまず[#Context]内のみを参照し、Userの質問に関連する正確な情報を見つけて、<thinking></thinking>の中に一言一句書き出し、参照したファイルと見出しを書き留めてください
-正確な引用を見つけ、十分な情報に基づいて推測できる場合にのみ、Userの質問に回答して下さい。
そうでなければ、分かりませんと出力してください
-必ず以下の出力形式に従い出力してください
-学習データ以外からの参照(https~等)を行った場合、<thinking></thinking>の中に記載せず分かりませんと記載してください
＃出力形式
<intention>
</intention>
<thinking>
注意：
Userの質問が複数の回答伴う場合、以下の項目を複数記載してください
参照した学習データ:参照した学習セットの大見出しを記載
見出し：引用した文章の見出しを記載
引用した文章：引用した文章に手を加えずそのまま記載 </thinking>

<summary>
<引用した文章を法律に詳しくない人向けに要約>
</summary>

<最終回答>
thinkingを元に最終回答を生成

作成者の評価

引用箇所から内容を正しく反映できないことが課題であったが、20問中14問で適切に反映することができた

課題#4

- 現職の方による生成結果の評価、および現職の方へのヒアリングを行ったところ、給与に関して調査を行う際に実業務で利用している参考情報が複数存在することを伺ったため、学習セットに参考情報を追加して検証を行うと生成品質の向上が見込まれる

課題#5

- Claude2.1はXMLタグを読み込みやすいものの、XMLタグ以外も混在している

課題#6

- 参照箇所の記載が「参照した学習データ」と「見出し」だけでは「ページ数」や「章番号」が記載されておらず、実際にどこを参照したかわかりにくいため、細かく出力させる必要がある

B-2. 人事院規則

試行2回目の評価項目

評価項目

改善方針#3

- 引用箇所が完全一致でなくとも一定評価できるパターンも見受けられたことから、部分点評価項目を増やした
- 一部が正解であれば部分点を与えていたがテーブルの反映の誤りは回答内容への影響度が大きいため一部の間違いであっても評価0とする基準に変更

評価基準	要件定義	評価(点数)
回答の正しさ1:引用箇所が正しいかどうか	元々テストケースを作成した時の本文の参照箇所と合致している	2
	テストケースの参照箇所と異なるが、引用箇所が質問に対する回答になりそう	2
	2か所以上引用しており、1つは合致しているが1つは関連がなさそう	2
	2か所以上引用しており、1つは合致しているが1つは明確に間違い	1
	2か所引用が必要だが、1か所しかとってきてない	1
	引用箇所は正解だが、ナンバリングや見出しが誤り	1
	明らかに引用箇所が誤っている/引用していない	0
回答の正しさ2:回答文が引用箇所を正しく反映しているか	引用箇所に関する表記が正しく記載されている(テーブル)	2
	引用箇所に関する表記が、全く異なる、もしくは一部誤り（テーブル）	0
	引用箇所に関する表記が正しく記載されている	2
	引用箇所に関する表記がおおむね合っているが、一部誤り（見出しなど）	1
	引用箇所に関する表記が全く異なる(他の文章の内容を抽出しているなど)	0
わかりやすさ:表現の仕方が適切であるか	回答が引用や考察を正しく踏まえており、ユーザーの質問に答えている	2
	おおむね答えられているが、不要な表現や小さな誤りも一部見られる	1
	ユーザーの質問に回答できていない	0

B-2. 人事院規則

試行2回目の生成結果

質問

新しい給与法の下で勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算は、どのように行われますか？

解答参照個所

参照　前条の規定による俸給を支給される職員に関する給与法第十条の五第二項、第十九条の四第五項並びに附則第八項第二号から第四号まで、第六号及び第七号の規定の適用については、給与法第十条の五第二項中「俸給月額」とあるのは「俸給月額と一般職の職員の給与に関する法律等の一部を改正する法律（平成二十六年法律第百五号）附則第七条の規定による俸給の額との合計額」とする。
見出し：附　則　（平成二六年一一月一九日法律第一〇五号）　抄 （俸給の切替えに伴う経過措置） 第八条

生成結果

<intention>
ユーザーは、新しい給与法の下で勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算がどのように行われるか知りたいと考えています。
</intention>

<thinking>
参照した学習データ:一般職の職員の給与に関する法律等の一部を改正する法律
見出し:給与の内払
引用した文章:第一条の規定による改正後の一般職の職員の給与に関する法律(附則第三条において「改正後の給与法」という。)の規定、第三条の規定による改正後の一般職の任期付研究員の採用、
給与及び勤務時間の特例に関する法律(附則第三条において「改正後の任期付研究員法」という。)の規定及び第五条の規定による改正後の一般職の任期付職員の採用及び給与の特例に関する法律
(次条及び附則第三条において「改正後の任期付職員法」という。)の規定は、平成三十年四月一日から適用する。
</thinking>

<summary>
新しい給与法等の規定は平成30年4月1日から適用されます。したがって、その日以降、勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算も新しい給与法の規定に基づいて
行われることになります。
</summary>

<最終回答>
新しい給与法等の規定は平成30年4月1日から適用されています。したがって、その日以降、勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算も新しい給与法の規定に基づいて
行われています。具体的な計算方法については法律の規定や人事院規則に委ねられている部分もあると考えられます。

B-2. 人事院規則

試行2回目の生成結果

生成結果

<intention>
ユーザーは、新しい給与法の下で勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算がどのように行われるか知りたいと考えています。
</intention>

<thinking>
参照した学習データ:一般職の職員の給与に関する法律等の一部を改正する法律
見出し:給与の内払
引用した文章:第一条の規定による改正後の一般職の職員の給与に関する法律(附則第三条において「改正後の給与法」という。)の規定、第三条の規定による改正後の一般職の
任期付研究員の採用、給与及び勤務時間の特例に関する法律(附則第三条において「改正後の任期付研究員法」という。)の規定及び第五条の規定による改正後の一般職の任期付職員
の採用及び給与の特例に関する法律(次条及び附則第三条において「改正後の任期付職員法」という。)の規定は、平成三十年四月一日から適用する。
</thinking>

<summary>
新しい給与法等の規定は平成30年4月1日から適用されます。したがって、その日以降、勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算も新しい給与法の規定に
基づいて行われることになります。
</summary>

<最終回答>
新しい給与法等の規定は平成30年4月1日から適用されています。したがって、その日以降、勤務時間法の規定による勤務時間の割振りを行わない職員の給与計算も新しい給与法の
規定に基づいて行われています。具体的な計算方法については法律の規定や人事院規則に委ねられている部分もあると考えられます。

評価項目

回答の正しさ1:引用箇所が正しいかどうか	元々テストケースを作成した時の本文の参照箇所と合致している	0 / 1 / 2 点（0）
回答の正しさ2:回答文が引用箇所を正しく反映しているか	引用箇所に関する表記が正しく記載されている	0 / 1 / 2 点（2）
分かりやすさ:表現の仕方が適切であるか	回答が引用や考察を正しく踏まえており、ユーザーの質問に答えている	0 / 1 / 2 点（2）

B-2. 人事院規則

プロンプトの発展過程　試行3回目

LLM	Claude 2.1

作成者の評価	例題として16問検証したところ、10問正しい箇所から引用することができた

学習セット

人事院規則九―六（俸給の調整額）（昭和三十二年人事院規則九―六）（.text）
一般職の職員の給与に関する法律（.text）
一般職の任期付職員の採用及び給与の特例に関する法律（.text）
一般職の任期付研究員の採用、給与及び勤務時間の特例に関する法律 （.text）
別紙第４ 職員の給与の改定に関する勧告（.text）

改善方針#4

- 学習セットに情報を追加

改善方針#5

- <output></output>や<thoughts></thoughts>などXMLタグに統一する

改善方針#6

- 参照箇所を{{clause}}{{number}}{{quote}}{{meaning}}のように章番号やページ数を細かく出力させる

プロンプト

以下のルールに基づき、ユーザーの質問への回答を日本の法令から引用してください。
1.<output></output>に従い出力する
2.考察を<thoughts></thoughts>に記載する
3.<intention></intention>にUserの質問意図を記載する
4.回答に関連する情報が複数ある場合は、すべて<回答></回答>に記載する
5.関連情報が見当たらない場合は回答しない
6.{{meaning}}には{{quote}}で引用してきた文章の要約を記載する
<question>
質問をここに入力してください。
</question>
<output>
<intention></intention><thoughts></thoughts>
<回答>{{clause}}{{number}}{{quote}}{{meaning}}</回答>
</output>

B-2. 人事院規則

試行3回目の生成結果

質問

平成三十年四月一日から適用される改正後の給与法において、特定職員の俸給調整はどのように行われますか？

解答参照個所

切替日の前日から引き続き同一の俸給表の適用を受ける職員で、その者の受ける俸給月額が同日において受けていた俸給月額に達しないこととなるもの（人事院規則で定める職員を除く。）には、平成三十年三月三十一日までの間、俸給月額のほか、その差額に相当する額を俸給として支給する。

生成結果

<output>
<intention>
ユーザーは、切替日の前日から引き続き同一の俸給表の適用を受ける職員で、俸給月額が切替日において受けていた額に達しない職員に対する俸給の取扱いについて質問している。
</intention>
<thoughts>
ユーザーの質問は具体的で関連する情報が特定できそうだが、質問文の内容が複雑なので要点を絞って関連部分を抽出する必要がある。
</thoughts>

<回答>
附則第二項
切替日の前日から引き続き同一の俸給表の適用を受ける職員で、その者の受ける俸給月額が同日において受けていた俸給月額に達しないこととなるもの(人事院規則で定める職員を除く。)には、平成三十年三月三十一日までの間、俸給月額のほか、その差額に相当する額を俸給として支給する。
{{meaning}}
切替日前後で同じ俸給表適用の職員で、切替後の俸給月額が切替前の額に達しない職員(人事院規則で定める職員を除く)には、平成30年3月31日までの間、俸給月額に加えて差額を俸給として支給する。
</回答>

</output>

B-2. 人事院規則

試行3回目の生成結果

生成結果

<output>
<intention>
ユーザーは、切替日の前日から引き続き同一の俸給表の適用を受ける職員で、俸給月額が切替日において受けていた額に達しない職員に対する俸給の取扱いについて質問している。
</intention>
<thoughts>
ユーザーの質問は具体的で関連する情報が特定できそうだが、質問文の内容が複雑なので要点を絞って関連部分を抽出する必要がある。
</thoughts>

<回答>
附則第二項
切替日の前日から引き続き同一の俸給表の適用を受ける職員で、その者の受ける俸給月額が同日において受けていた俸給月額に達しないこととなるもの(人事院規則で定める職員を除く。)には、平成三十年三月三十一日までの間、俸給月額のほか、その差額に相当する額を俸給として支給する。
{{meaning}}
切替日前後で同じ俸給表適用の職員で、切替後の俸給月額が切替前の額に達しない職員(人事院規則で定める職員を除く)には、平成30年3月31日までの間、俸給月額に加えて差額を俸給として支給する。
</回答>

</output>

評価項目

回答の正しさ1:引用箇所が正しいかどうか	元々テストケースを作成した時の本文の参照箇所と合致している	0～2点（2）
回答の正しさ2:回答文が引用箇所を正しく反映しているか	引用箇所に関する表記が正しく記載されている	0～2点（2）
分かりやすさ:表現の仕方が適切であるか	回答が引用や考察を正しく踏まえており、ユーザーの質問に答えている	0～2点（2）

B-2. 人事院規則

試行3回目の評価結果

▶ 全体の最終得点率は79%と高スコアであり、実業務への利活用が十分に期待できる結果

ポイント1)

PDF形式からの情報の抽出が大きな課題であったが、学習セットへの取り込み方の工夫により、
回答の正しさ1(引用箇所の正確性）は53%から72%
・回答の正しさ2(回答文が引用箇所を正しく反映しているかどうか）は78%から84%
に向上した。

ポイント2)

検証途中で学習セットへの情報量を増やしたが回答精度の低下は見られなかったため、質の良いプロンプト作成ができたといえる

評価基準	妥当	部分点	不適当	質問数	得点率
回答の正しさ1: 引用箇所が正しいかどうか	10	3	3	16	72%
回答の正しさ2:回答文が引用箇所を正しく反映しているかどうか	13	1	2	16	84%
回答の正しさ3:表現の仕方が適切であるか	12	2	2	16	81%

B-2. 人事院規則

追加試行の評価結果

- ▶ プロンプトを変更せず、使用するLLMを「Claude 3」に切り替え追加検証を実施
- ▶ 対象となった質問は、最終試行において、回答の正しさ1の評価が「1」であったものを2つ、評価が「0」だったものを3つの合計5つの質問で、67%の得点率を確認

ポイント1）

Claude2.1を使用した場合の対象の質問のみの得点率が33%だったのに対し、claude3を使用すると67%になった

ポイント2）

回答の正しさ1の得点率が20%→80%になった。ユーザーの質問から適した箇所を参照する精度が向上したと考えられる

	再検証した質問数	Claude 2.1	Claude 3
回答の正しさ1:引用箇所が正しいかどうか	5	20%	80%
回答の正しさ2:回答分が引用箇所を正しく反映しているかどうか	5	40%	60%
回答の正しさ3:表現の仕方が適切であるか	5	40%	60%
得点率		33%	67%

B-2. 人事院規則

追加試行の回答比較

ポイント1)

Claude3使用時の方が、ユーザーの質問の意図理解がより具体的である。

生成結果(Claude2.1)	生成結果(Claude3)
<intention> ユーザーは平成30年4月1日以降の任期付職員の給与月額の計算方法の変更点について質問している。 </intention>	<intention> この質問は、平成30年4月1日以降の任期付職員の給与月額の計算方法が変更されたかどうか、そしてどのように変更されたかを尋ねています。 </intention>

ポイント2)

Claude2.1使用時は「変更点はない」と誤った回答をしているが、Claude3使用時はユーザーの質問に関連する箇所を複数引用し、適した回答を生成している。

生成結果(Claude2.1)	生成結果(Claude3)
<回答> 附 則（平成三〇年一一月三〇日法律第八二号）抄 （施行期日等） 第一条 この法律は、公布の日から施行する。ただし、第二条、第四条及び第六条の規定は、平成三十一年四月一日から施行する。 </回答> 平成30年改正の任期付職員法は平成31年4月1日から施行されるため、平成30年4月1日時点で任期付職員の給与月額の計算方法の変更点はない。	<回答> {{clause}}1{{number}}第1条{{quote}}この法律は、公布の日から施行する。ただし、第二条、第四条及び第六条の規定は、平成三十一年四月一日から施行する。{{meaning}}この法律の一部の規定は平成31年4月1日から施行されるが、任期付職員の給与月額の計算に関する規定は平成30年4月1日から適用される。}} {{clause}}2{{number}}第2条{{quote}}第一条の規定による改正後の一般職の職員の給与に関する法律（附則第三条において「改正後の給与法」という。）の規定、第三条の規定による改正後の一般職の任期付研究員の採用、給与及び勤務時間の特例に関する法律（附則第三条において「改正後の任期付研究員法」という。）の規定及び第五条の規定による改正後の一般職の任期付職員の採用及び給与の特例に関する法律（次条及び附則第三条において「改正後の任期付職員法」という。）の規定は、平成三十年四月一日から適用する。{{meaning}}任期付職員の給与月額の計算に関する規定が改正され、平成30年4月1日から適用される。}} {{clause}}3{{number}}第11条{{quote}}改正後の法の規定を適用する場合においては、改正前の法の規定に基づいて支給された給与は、改正後の法の規定による給与の内払とみなす。{{meaning}}平成30年4月1日以降は、改正後の任期付職員法の規定に基づいて給与月額が計算される。}} </回答>

回答比較

・出力形式に正確に従って出力していることから、Claude3の方がプロンプトの読み取りが優れていることがわかる。
・ユーザーの質問に対する意図理解や正確な引用と言う面で、Claude2.1使用時は引用箇所が適しておらず誤った回答をしているのに対し、Claude3使用時はユーザーの質問に適した箇所及び、参照箇所と同じ内容を引用しており、改正後の給与計算について述べている。

総括

総括　：　GPT-4やClaude 2.1での正確な引用や抽出は一定精度までとみられるが、LLMの性能次第でより実践的な利用が可能になると推量されるため、今後とも法令などの検索/引用においてLLMは利用検討すべき

考察　：

・熟練職員でなければ読解が困難な法令の計算ロジックやルール等を、LLMで簡易に読み解くことができる点は非常に有用と考える。
・また、検証タイミングの都合上Claude2.1での検証が主体であったが、Claude3 Sonnetで試用した結果、極めて高精度での引用や生成が可能になることが分かった。LLMの性能向上が、ユースケースを実用的なレベルまで拡張できる可能性があるものと考える。

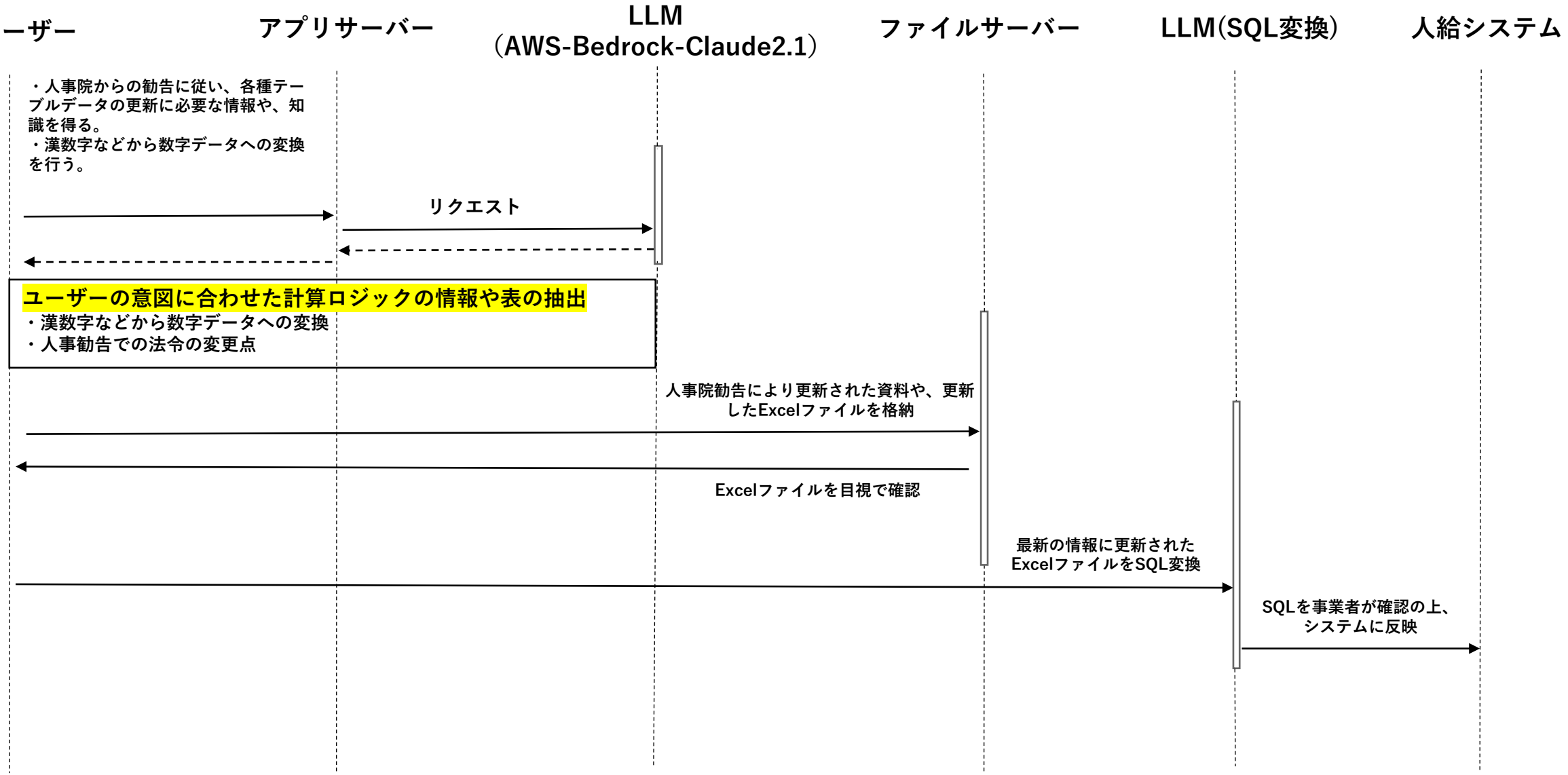
実装形式：

実装レベル	実装形式	コスト構造	利点	課題
検証	生成AI型SaaSサービスで、RAGで各種法令をテキストで登録したものを実装	SaaS利用料(トークン量込)	本利用に向けた本格検証が可能。プロンプトの組織管理/共有が可能なため、業務利用も可能。	法令を忠実に引用抽出できるLLMの選定(GPT-4-TurboやClaude 2.1等では一定精度までが限界だが、今後のLLM性能に応じてその限りでない)
将来像	後述のシーケンス図に従い、給与計算ロジックに基づいた俸給テーブルを半自動的に生成。表を出力してSQL化する手順までを実装範囲とする。	・システム実装コスト ・俸給テーブル作成ロジックDB(xlsへの出力) ・トークン利用料 ・インフラ運用コスト	俸給テーブルの法令文を計算ロジックへ返還し、システム反映可能なテーブルやSQLを一元出力できる点。	難解な計算ロジックである点以上に、対象職種や俸給パターンが多種難解である点から、自動化を目指すには入念な異常値テストが必要の為、人の手でのサポートや試験を継続しながら運用する必要がある。

B-2. 人事院規則

シーケンス図

人事院勧告の内容が更新される



検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

ユースケース概要

前提	政府では府省庁や地方自治体に至るまで、様々な組織単位で法令や条例整備、更新を進めている。その際、関連法令を検索する業務がその起点となる。関連法令を探すときには、その用例(その法令の用いられ方)をベースにした検索がカギとなる
課題	現在e-lawsのウェブサイトに政府による全ての法令がデータベース化されている。しかしながら、その検索は完全一致ベースとなるため、適切な法令検索には一定の知見が求められる。特に、用例をベースにした検索性までは実装されていないのが現状
ユースケース概要	こうした課題をベースに、本検証において「用例に基づく法令検索」を開発実装し、仮検証できるサービスとして提供するに至った。本サービスは検証環境(GaiXer)のUI上の一機能として実装されつつ、個別インターフェイスとアーキテクチャで提供する事で、ユーザー利便性が一定確保しやすいものとして提供。

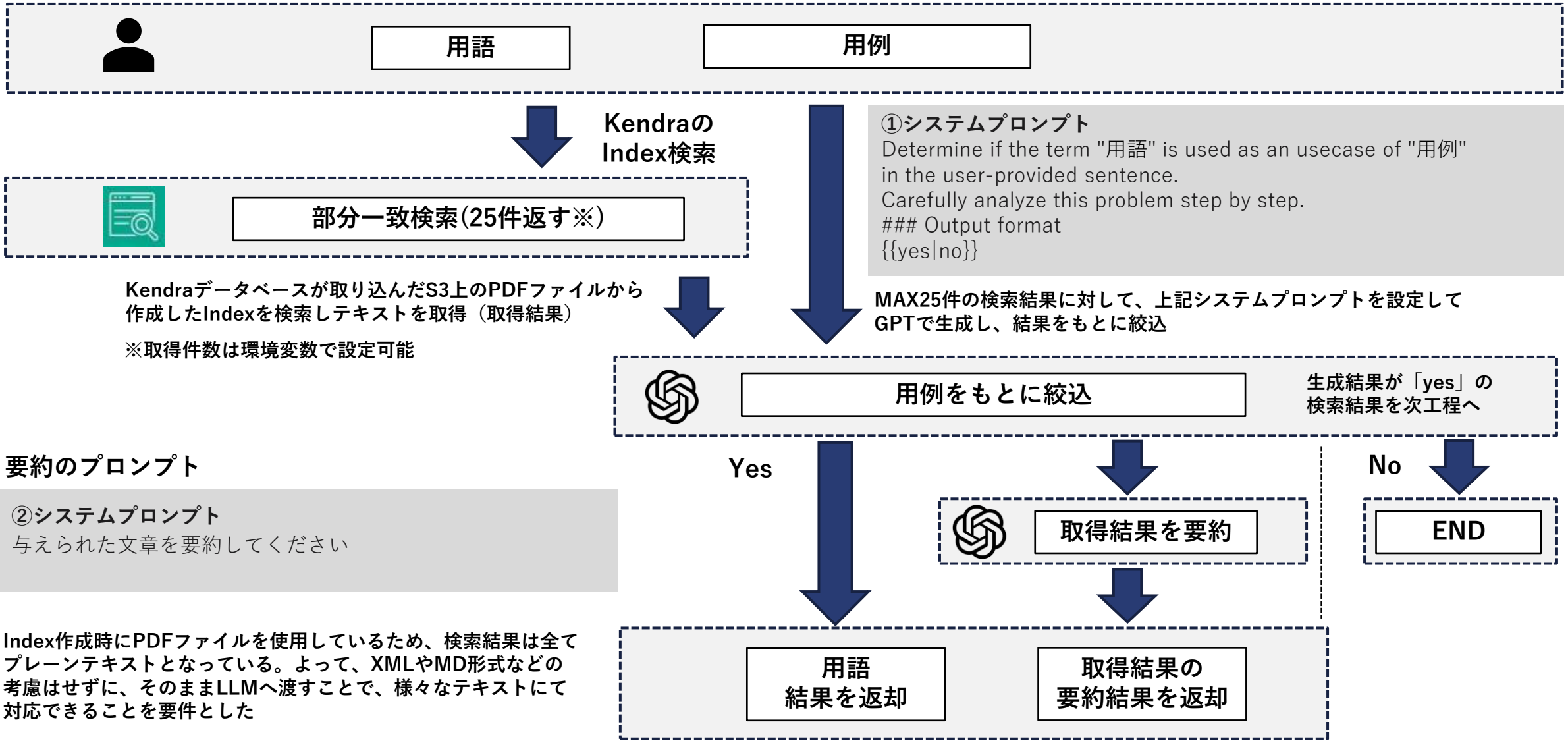
C-1. 用例による法令検索

作業分解とプロンプトの関連性

目的	用例に基づく法令検索ができるサービスを実装する事で、通常の検索プロセスを効率化する

	作業工程	プロンプト	検証アプローチ	検証結果
1	法令/条例を改定するニーズを把握	用例に基づく法令検索サービス	・2~4の作業工程をシステム化できるアーキテクチャを検証 ・LLMのみでの検索や実装はハルシネーションリスクが高く困難故、下記要件を検証； - あいまい検索が可能 - 用語は完全一致/用例はあいまい検索 - 法令原文の正確な出力 - その要約内容の解説の付与	✓ KendraとGPT3.5を組み合わせた機能としての実装により、検証アプローチ通りのサービスが実装できた
2	e-lawsで関連法令を検索			
3	検索した法令内で、用例が当てはまる条文を検索			
4	具体的な条文を読み込み意味を理解する			
5	法令の改定案等のドラフトを作成	-		

用例による法令検索機能



要約のプロンプト

②システムプロンプト
与えられた文章を要約してください

Index作成時にPDFファイルを使用しているため、検索結果は全てプレーンテキストとなっている。よって、XMLやMD形式などの考慮はせずに、そのままLLMへ渡すことで、様々なテキストにて対応できることを要件とした

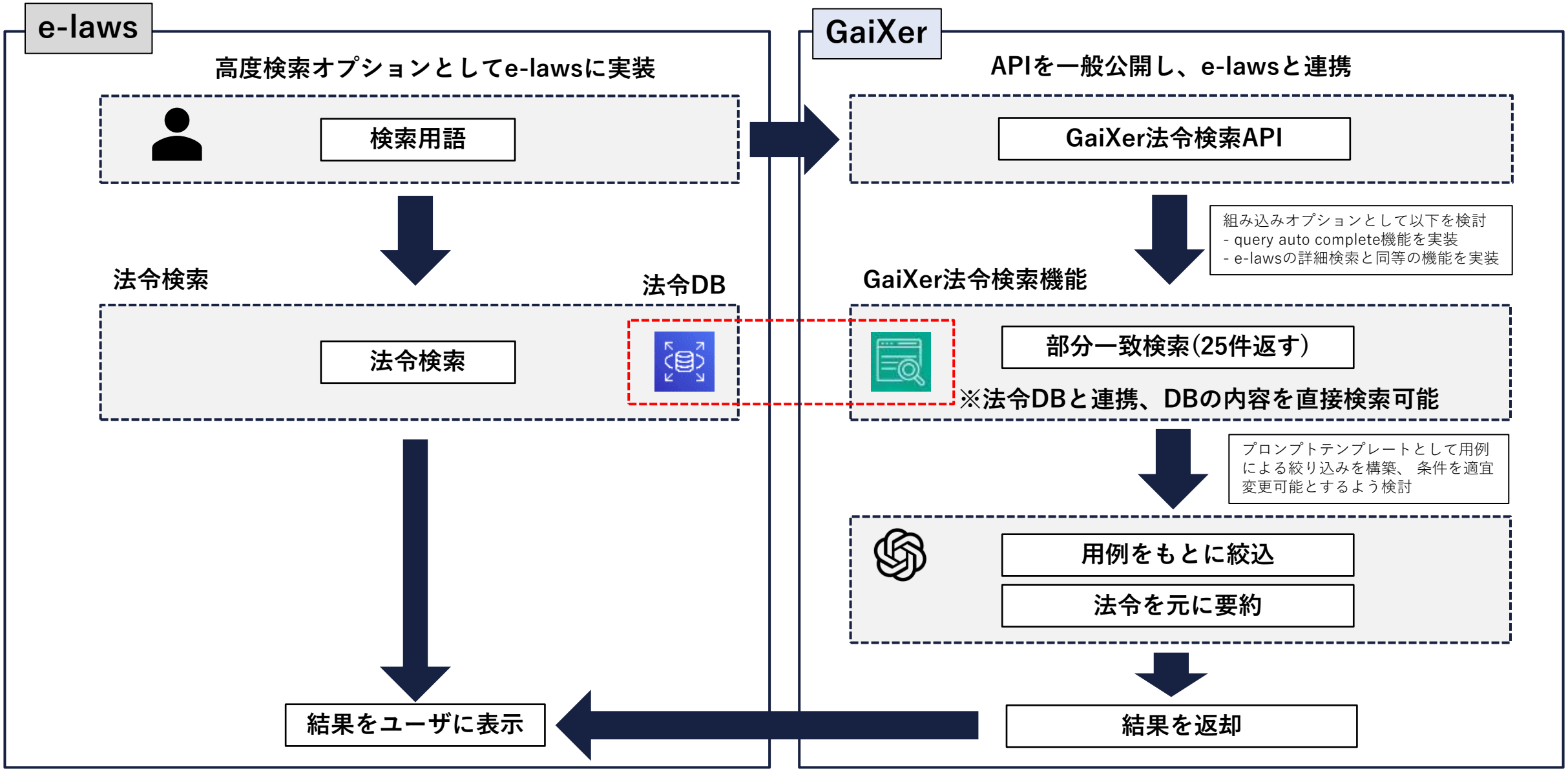
生成AI導入による効果

▶ 用例に基づく法令検索を行うことにより、作業時間の単純削減が期待できるに加え、属人の検索ノウハウが一定不要となる事から、特に人手不足となりうる地方自治体等において、行政の運用持続性への貢献が期待できる

効果	効果内容
作業時間の削減	現状ではe-lawsのサイト上で法令を特定用語(キーワード)や年次等で検索をかけることが通常である。その際、特定用語を用いた検索では、職員に対して一定の試行回数が求められる。本機能のような用例検索を通じた広範かつあいまいな検索による絞りこみを可能にすることで、そうした検索が一度できることによる効果が期待される。
属人性の排除/運用持続性	現行の検索では一定の熟練職員である事が効率的な検索のカギとなっている。本サービスを継続改善しながら本格実装した場合、そうした熟練者でなくとも、有効かつ効率的に法令検索が可能になることが期待できる。これにより、行政の実務において、人の特殊技能によらない持続的な運用継続性を確保する事が可能になると考えられる。

なお、現段階での本機能の本格実用するにあたっての必要な改善点をユーザーにヒアリングした。その実装案を次頁に示す。

あるべき将来像（e-laws連携）



e-laws：https://elaws.e-gov.go.jp/

検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

C-2.用語集作成業務の効率化
ユースケース概要

前提	既存の用語集を正として、その拡張性に可能性があるかに主眼を置いて検証した
課題	公開前の文書に対する用語集を作成する担当者の業務を効率化する為に、新たに当該文書の用語集に追加するべきだと思われる用語を指摘し、その定義を生成することで、用語集の完成までのリードタイムと工数削減を狙うユースケースとして検証した

既存用語集と対象ドキュメントを参照

新規用語の定義を生成

新規用語の候補を抽出

- デジタル社会形成基本法
- 情報通信技術
- 官民データ活用推進基本法
- マイナンバーカード
- eID
- テクノロジーマップ
- デジタル法制審査
- 情報連携基盤
- 公共サービスメッシュ
- 標準準拠システム
- 準公共サービス
- AI
- 包括的データ戦略
- トークン
- サイバーセキュリティ
- 防災デジタルプラットフォーム
- モビリティ・ロードマップ
- ４次元時空間 ID
- デジタル田園都市国家構想
- G ビズ ID
- J グランツ
- Society 5.0
- デジタル共生社会
- 情報リテラシー

その候補の用語集を作成

C-2.用語集作成業務の効率化

作業分解とプロンプトの関連性

目的	・ 新規用語の定義を生成する事で、公開前の文書に対する用語集を作成する担当者の業務を効率化する

	作業工程	ツール	検証アプローチ	検証結果
1	対象ドキュメントに既存用語が含まれているかをチェックする	Python	・ 本検証においてはスコープ外	
2	対象ドキュメントから、追加するべき新規の用語を抽出する	生成AI	・ 既存の用語集と重複せずに、新規用語を対象ドキュメントから抽出できたか？ ・ 既存の用語集のコンテキストに即した、新規用語の用語集を生成できたか？	▲ 24個の新規用語を対象ドキュメントから正しく抽出し、かつ既存用語集との重複は0だった。ただし、特定の用語を狙い撃ちで抽出することはできなかった。
3	新規用語の定義を作成			✓ 24個の用語の説明を正しく生成できた。
4	1,2を合わせて用語集を作成			

現行の作業ステップ

▶ 現行の作業ステップでは51時間の作業工数を要する

No	作業ステップ	インプット	アウトプット	工数 ※対象ドキュメントがpdf118ページ 新用語が300語あると想定 算出根拠は次頁に記載	工数合計
1	用語抽出の準備	対象ドキュメント	対象ドキュメントから作成すべき用語集の性質の理解	60分	3,060分 （51時間）
2	用語の抽出	対象ドキュメント	用語集に記載すべき用語	900分	
3	用語の定義と説明の準備	用語集に記載すべき用語	用語の定義作成のために必要なソース	120分	
4	用語の定義と説明の作成	用語の定義作成のために必要なソース	新たな用語の定義	1,800分	
5	校正と最終チェック	対象ドキュメントに対応した用語集	対象ドキュメントに対応した用語集	150分	
6	用語集のフォーマットと仕上げ	対象ドキュメントに対応した用語集	対象ドキュメントに対応した用語集	30分	

C-2.用語集作成業務の効率化

生成AI導入前後の作業ステップの変化

▶ 生成AIの導入により、用語集作成1件当たりの対応工数が約90%削減できる見込み

■生成AI導入前

No	作業ステップ	インプット	アウトプット	工数 ※対象ドキュメントがpdf118ページ、300用語あると想定	工数算出根拠	工数合計
1	用語抽出の準備	対象ドキュメント	対象ドキュメントから作成すべき用語集の性質の理解	60分	118ページのドキュメントを読みコンセプトを理解するまでに妥当な時間の推定	3,060分 (約51時間)
2	用語の抽出	対象ドキュメント	用語集に記載すべき用語	900分	1ページ約1,600文字文字に対して約8分と仮定(読む、見つける、リストアップする時間)	
3	用語の定義と説明の準備	用語集に記載すべき用語	用語の定義作成のために必要なソース	120分	適切なソースを手動で探すため、また、用語ごとの基準を設定するために必要な時間推定	
4	用語の定義と説明の作成	用語の定義作成のために必要なソース	新たな用語の定義	1,800分	1用語当たりの作成時間を6分と推定	
5	校正と最終チェック	対象ドキュメントに対応した用語集	対象ドキュメントに対応した用語集	150分	1用語あたり30秒でチェックできる推定	
6	用語集のフォーマットと仕上げ	対象ドキュメントに対応した用語集	対象ドキュメントに対応した用語集	30分	最終フォーマットの調整にかける妥当な時間推定	

■生成AI導入後

No	作業ステップ	インプット	アウトプット	工数	工数算出根拠	工数合計
1	既存の用語集を参照にして、プロンプト試行 (用語の抽出 / 新たな用語の定義を生成 / 用語集のフォーマット化)	対象ドキュメント	対象ドキュメントに対応した用語集	5 分	実作業を基に算出	305分 (約5時間) 導入前比 -46時間
2	担当者による最終チェック	対象ドキュメントに対応した用語集	対象ドキュメントに対応した用語集	300分	用語の整合性チェック 1分×300語	

C-2.用語集作成業務の効率化

業務削減効果とコスト

▶ 用語集作成業務に対して、生成AIを導入すると年間約540円で658.4時間の業務削減効果の見込み（1,000円当たり1,218時間の業務削減）

■業務削減効果

業務削減効果(年間)　＝ 用語集作成1件当たりの業務削減時間 × 用語集作成の総件数(年間)※1
＝ 3,292分 × 12
＝ 39,504分
＝ 658.4時間

■必要コスト

用語集作成1回あたりのプロンプト試行コスト
＝ 入力トークン数(k) × 入力料金(ドル/kトークン) ＋ 出力トークン数(k) × 出力料金(ドル/kトークン)
＝ 28※2 × 0.01※3 ＋ 1.2※2 × 0.03※3
＝ 0.3ドル

必要コスト(年間)　＝用語集作成1回あたりのプロンプト試行コスト×用語集作成の総件数(年間)
＝ 0.3ドル × 150(円/ドル) × 12件※1
＝ 540円

※1 月に1件×12か月と想定
※2 本検証におけるプロンプト(最終版)で試行した際の平均値
※3 「Azure OpenAI Serviceの価格」よりGPT-4-Turboの価格を引用

C-2.用語集作成業務の効率化

用語集作成業務への生成AI導入に関する定性効果

▶ 広範なデータベースから関連用語を抽出するAIの能力によって誰もが高品質な用語集を作成することが可能となることで、新たな知識(用語）の普及と作業の質の向上を実現できると期待できる

効果	効果内容
用語集業務の促進と品質の向上	AIプロンプトによって生成された用語は一貫性があり、庁内外での用語の共通理解を促進させることでコミュニケーションがスムーズになり、誤解のリスクを低減させることが可能となる
	専門知識がない職員でも高品質な用語集を作成可能になるため、用語集作成業務そのものが政府全体で普及し、標準化へ向けた活動が加速化する
用語集の品質向上から期待できる相乗効果	AIが従来の枠を超えた用語や関連性を発見することで、政策や業務改革に新しいアプローチを提供できる
	用語集のリアルタイム更新と最適化で常に最新の情報を保持しており、迅速な意思決定のサポートができることから組織の柔軟性と迅速性を強化

評価基準と評価要件

ポイント1)

**コンテキストを理解した上での抽出がどこまでできるか。
またモデル間に差はあるか**

ポイント2)

**コンテキストを理解した上での定義説明がどこまでできるか。
またモデル間に差はあるか**

評価基準	要件定義	評価(点数)
回答の正しさ1:重複のない抽出	用語集に含まれていた一部の用語を、抜き出した状態で検証した結果、用語集に加えるべきという指摘をした	1
	用語集に含まれていたが、抜き出した状態で検証した結果、用語集に加えるべきという指摘が漏れていた	0
	用語集に含まれていないが、新たに追加すべき用語であると提案をした	1
	用語集に含まれている用語と重複する用語を提案した	0
	用語集に含まれておらず、且つ新たに追加すべき用語であるとは認められない	0
	本文に含まれていない用語が登場している	0
表現の適切さ1:適切な抽出	用語集に含まれていた一部の用語に対して、元の説明と一致 or 類似した説明を生成した	1
	用語集に含まれており、用語の説明が元の説明と一致しておらず、説明が不適切	0
	用語集に含まれていないが、新たに追加すべき用語の適切な説明を生成した	1
	用語集に含まれておらず、且つ用語の説明が不適切	0

C-2.用語集作成業務の効率化

作業/評価ステップ

No.	検証段階	アクション	アクションの理由(簡潔に)
1	1回目の検証	既存用語のリストから、一部の用語を削除した状態の用語集を用意する	既存の用語集の完成度が高いことから、既存の用語を抽出のターゲットとして検証したいため
2		プロンプトを、GPT-4-Turbo/Claude3 Sonnetのそれぞれ個別で開発	コンテキスト理解の質で期待できるGPT-4 Turboと、コンテキストウィンドウの量で期待できるClaude3 Sonnetのどちらが有効かを検証するため
3		開発したプロンプトテンプレートの入力項目に、対象ドキュメントをコピー&ペーストで入力する	組織内での伝搬を期待できる、テンプレート機能を利用するため
4	1回目の修正	プロンプト内の「用語」を「既存のGeneric technical termリスト」に変更した	削除した特定の用語を生成することができなかったゆえ、その用語を拾えるよう、プロンプトのディレクションが必要と考えたため

テストケースに関する詳細

テストケース作成方針

既存の用語集から、一部の用語を削除した用語集を用意し、用語候補を含む文章から、追加すべき用語および説明をリストアップできるかを検証した

既存の用語集

用語	説明
5G	第5世代移動通信システム
API	システムやソフトウェアが公開している機能を外部から利用するためのプログラム上の規約
BCP	事業継続計画
Beyond5G	第5世代移動通信システムの次の世代の移動通信システム（いわゆる6G）
BPR	業務改革
・	・
・	・

一部の用語を削除した用語集

用語	説明
5G	第5世代移動通信システム
API	システムやソフトウェアが公開している機能を外部から利用するためのプログラム上の規約
BCP	事業継続計画
Beyond5G	第5世代移動通信システムの次の世代の移動通信システム（いわゆる6G）
BPR	業務改革
DMAT	災害派遣医療チーム
DX	デジタルトランスフォーメーションのこと

ポイント1）

削除した用語をリストアップできるか

ポイント2）

新たに追加すべき用語を追加できるか

C-2.用語集作成業務の効率化

プロンプトの発展過程　試行1回目

LLM	GPT-4 Turbo

学習セット

なし

プロンプト

#step-by-stepで実行
Step1. {既存の用語リスト}に追加するべき新しい仲間を、
{用語候補を含む文章}から見つけ出してください。
例えば、{既存の用語リスト}には、"DX"、"Web3.0"、
"インターフェース"などの用語が含まれています。
彼らと重複していない、新しい仲間を探してください。
注意：**つまり、{既存の用語リスト}にすでに含まれている用語の出力を禁じる。**
Step2. Step1で推定した用語リストに対して、「用語/わかりやすい表現または説明/備考」をテーブル形式で生成してください。

#出力形式
Step1. {既存の用語リスト}に追加するべき新しい仲間のリスト
Step2. テーブル形式の「用語/わかりやすい表現または説明/備考」
注意：**テーブル形式で出力するべき**

作成者の評価

抜き出した用語の生成以外は達成

課題#1

- その用語集自体のコンセプトや目的などの補足情報をプロンプトに落とし込むことで、課題特定の用語を狙い撃ちに期待できる

評価項目

回答の正しさ1: 重複のない抽出	• 用語集に含まれていたが、抜き出した状態で検証した結果、用語集に加えるべきという指摘が漏れていた • 用語集に含まれていないが、新たに追加すべき用語であると提案をした	0 ~ 2 点(結果：1点)
表現の適切さ1: 適切な抽出	• 用語集に含まれていないが、新たに追加すべき用語の適切な説明を生成した	0 ~ 2 点(結果：1点)

C-2.用語集作成業務の効率化

プロンプトの発展過程　試行2回目

LLM	GPT-4 Turbo

学習セット

なし

プロンプト

#step-by-stepで実行
Step1. {既存のGeneric technical termリスト}に追加するべき新しい仲間を、
{Generic technical term候補を含む文章}から見つけ出してください。
例えば、{既存のGeneric technical termリスト}には、"DX"、"Web3.0"、
"インターフェース"などのGeneric technical termが含まれています。
彼らと重複していない、新しい仲間を探してください。
注意：**つまり、{既存のGeneric technical termリスト}にすでに含まれている
Generic technical termの出力を禁じる。**
Step2. Step1で推定したGeneric technical termリストに対して、
「Generic technical term/わかりやすい表現または説明/備考」を
テーブル形式で生成してください。

#出力形式
Step1. {既存のGeneric technical termリスト}に追加するべき新しい仲間のリスト
Step2. テーブル形式の「Generic technical term/わかりやすい表現または説明/備考」
注意：**テーブル形式で出力するべき**

作成者の評価

試行1回目では抜き出した用語の生成は「0/3個」だったが、試行2回目では10回中5回で「1/3個」へ改善した

改善方針#1

- 試行1回目では「用語」だったところを、用語集に加えるべき用語の方向性を明示して、スコープを狭めた

評価項目

回答の正しさ1: 重複のない抽出	• 用語集に含まれていた一部の用語を、抜き出した状態で検証した結果、用語集に加えるべきという指摘をした • 用語集に含まれていたが、抜き出した状態で検証した結果、用語集に加えるべきという指摘が漏れていた	0 ~ 2 点(平均結果：1.5点)
表現の適切さ1: 適切な抽出	• 用語集に含まれていた一部の用語に対して、元の説明と一致 or 類似した説明を生成した • 用語集に含まれていないが、新たに追加すべき用語の適切な説明を生成した	0 ~ 2 点(平均結果：2点)

C-2.用語集作成業務の効率化
試行2回目の生成結果

生成結果

Step1. {既存のGeneric technical termリスト}に追加するべき新しい仲間のリスト

デジタル田園都市国家構想
データ連携基盤
クラウド・バイ・デフォルト原則
AI 活用
データ戦略
ガバメントクラウド
4次元時空間 ID
シビックテック
モビリティ・ロードマップ
サイバーセキュリティ
イノベーション
ステークホルダー
デジタル・スタートアップ
モジュール

Step2. テーブル形式の「Generic technical term/わかりやすい表現または説明/備考」

※生成された表は右図

Generic technical term	わかりやすい表現または説明	備考
家構想	地域の活性化を目的としたデジタル技術を活用した社会構造改革のビジョン	
データ連携基盤	異なるシステム間でデータを共有・活用するための基本的な技術インフラ	
クラウド・バイ・デフォルト原則	新しい情報システムは原則としてクラウドサービスを利用するという原則	
AI 活用	人工知能技術を利用して社会課題の解決や業務効率化などを図る取り組み	
データ戦略	データを活用して価値を生み出すための総合的な方針や計画	
ガバメントクラウド	国や地方公共団体の情報システムをクラウド化する取り組み	
4次元時空間 ID	空間情報を統一的な基準で一意に特定するためのID体系	
シビックテック	市民がテクノロジーを活用して地域の課題解決を図る取り組み	
モビリティ・ロードマップ	交通・移動に関する将来の技術開発やサービス展開を計画する指針	
サイバーセキュリティ	情報システムやネットワークをサイバー攻撃から保護するための取り組み	
イノベーション	新しいアイデアや技術が生み出され、実用化されることによって社会や経済に変化をもたらすプロセス	既存リストに同義の用語が存在するため、ここでは具体的な文脈での使用を想定
ステークホルダー	利害関係者、もしくはあるプロジェクトや事業に関わる関係者全般	既存リストに同義の用語が存在するため、ここでは具体的な文脈での使用を想定
デジタル・スタートアップ	デジタル技術を核として新しいビジネスや事業を興す新興企業	
モジュール	システムや製品を構成する個々の機能単位や部品	既存リストに同義の用語が存在するため、ここでは具体的な文脈での使用を想定

試行2回目の評価結果

ポイント1)

既存の用語集から削除した用語を狙いうちで生成する点以外においては、高得点である

ポイント2)

削除した用語を狙いうちで生成する場合でも、元々の用語集のコンセプトや用途などでディレクションをすることで、生成確率を上げることができることが判明した

評価基準		妥当	不適当	妥当な割合
回答の正しさ1:重複のない抽出	既存の用語集から削除した用語	5	25	17%
	新たに追加すべき用語	131	30	81%
表現の適切さ1:適切な抽出	既存の用語集から削除した用語	5	5	100%
	新たに追加すべき用語	131	131	100%

※同一プロンプト・同一単語を削除した場合で10回試行

総括

結論　：

単純な業務削減効果のみならず、生成AIによる新たな用語提案等のサポートによって全体的な用語集の高品質改善が見込まれるため、今後の活用が期待される

考察　：

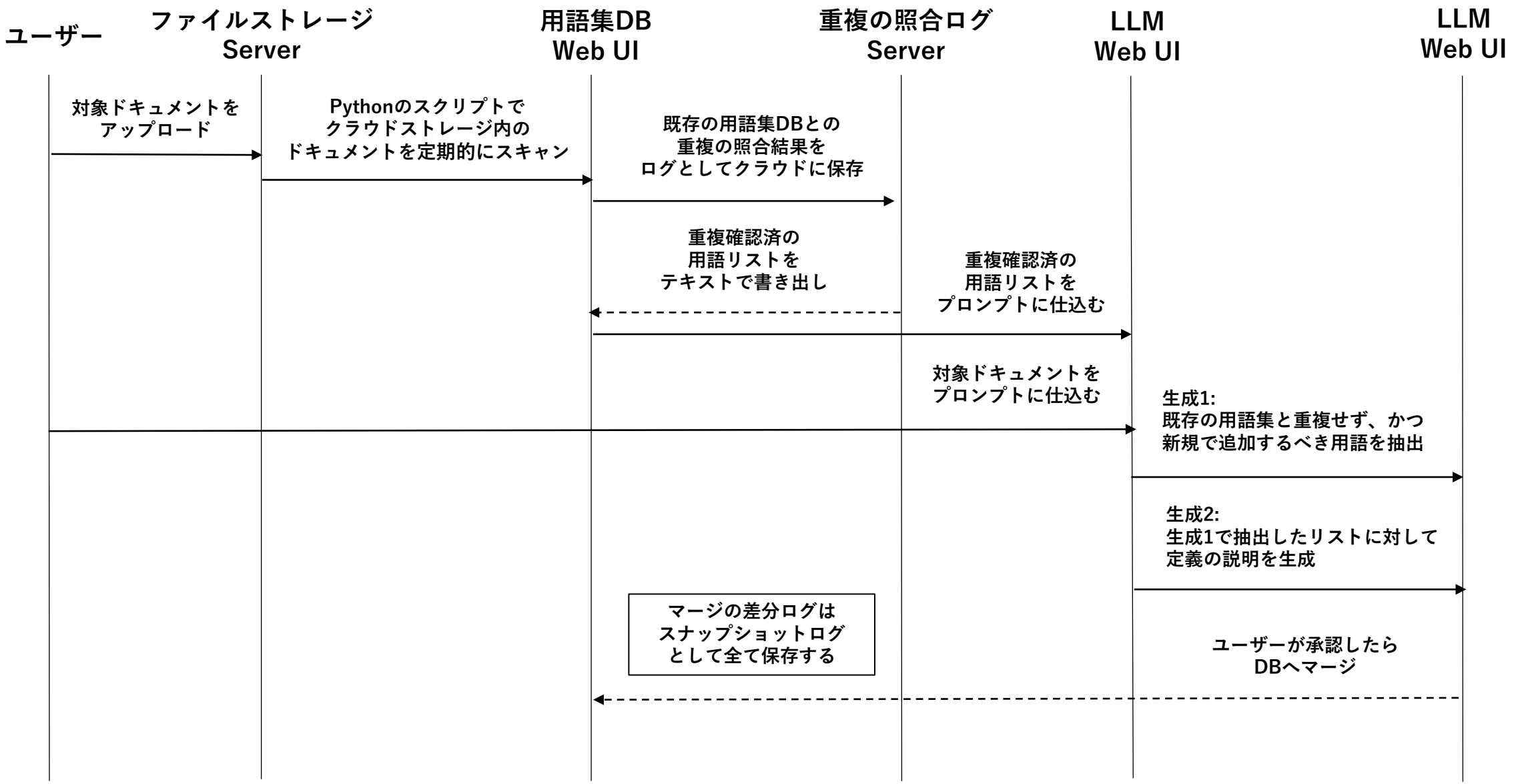
生成AIを活用した用語集の自動化プロセスにおいては、事前に用語集の目的やコンセプトなどの補足情報をプロンプトに組み込むことで用語選定の精度が高まることが判明した

実装形式

実装レベル	実装形式	コスト構造	利点	課題
検証	既存用語集を用いて、対象ドキュメントから用語集に記載すべき用語の抽出および定義の生成を行えるプロンプトを開発 ※生成AI型SaaSサービス上へは未導入	SaaS利用料(トークン量込)	用語集の作成時間と労力を大幅に削減し、職員がより重要な業務に集中できるようになる	AIへの適切なプロンプトの設定と維持には、専門知識が必要であり、適切な監督が求められる
将来像	新たな用語集をDBに蓄積して品質精度を上げるとともに、AIが自動的に目的やコンセプトを理解して用語を選定するレベルまで発展	• システム実装コスト • インフラ運用コスト • トークン利用料	AIの適応性と精度の向上により、用語集の質の向上が期待できる	AIツールの限界を理解し、最終的な出力に対する人間によるチェックを怠らないことが必要である

C-2.用語集作成業務の効率化

シーケンス図



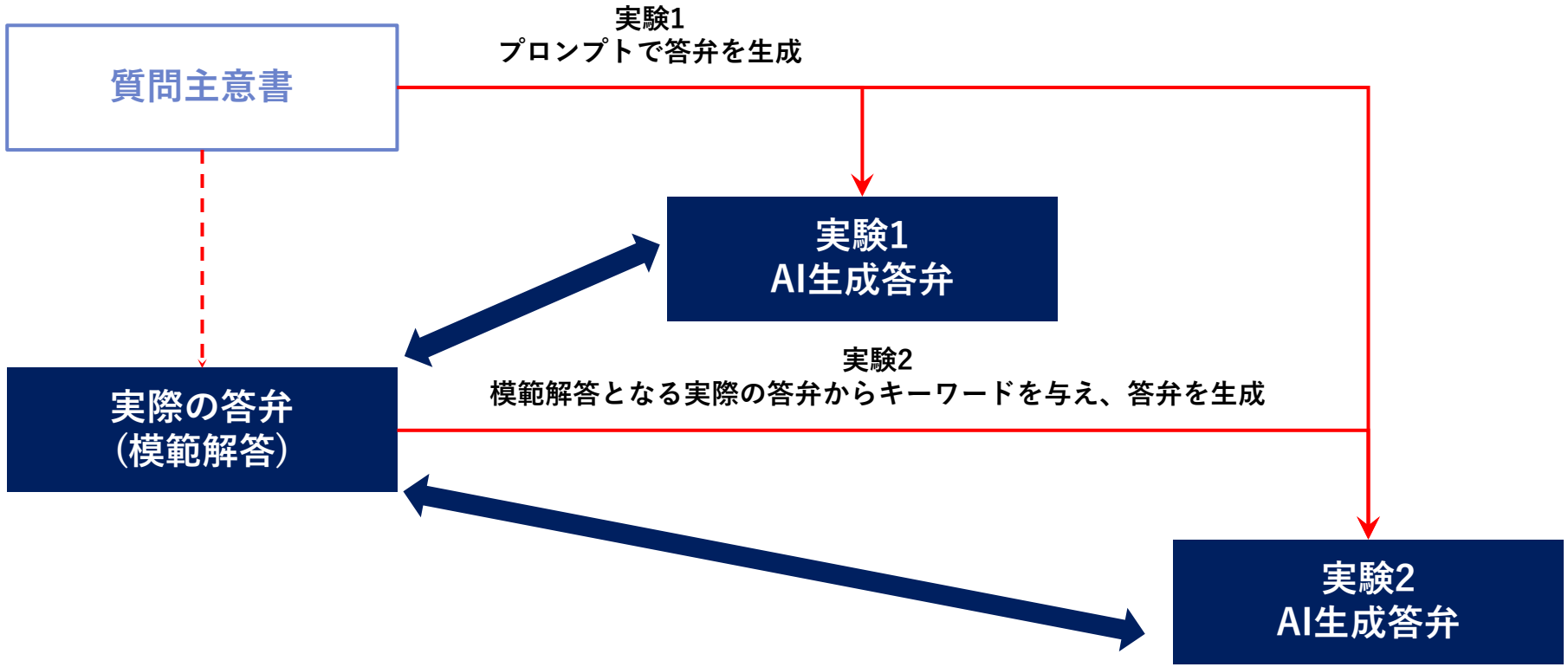
検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

D-1.質問主意書に対する答弁ドラフト生成

ユースケース概要

前提	初期的な検証・実験デザインの妥当性検証として、学習データは過去の国会の質問主意書＋答弁とし、他のデータは含めない。また、テストケースについても10問に留める。
課題	答弁案の作成に行政職員の多大な時間が割かれており、効率化の余地が大きい。また、新人等が過去の膨大な資料・答弁を踏まえて、新たに答弁案を作成する負担が大きく、成果物の品質向上余地も存在すると考える。



D-1.質問主意書に対する答弁ドラフト生成

作業ステップ

検証段階	#	アクション	アクションの理由(簡潔に)
準備	1	衆議院・参議院における過去の答弁(第148回=2000年～第211回＝2023年)の質問主意書・答弁をテキストで学習セットに登録	
	2	質問主意書に対して答弁ドラフトを生成するプロンプトを作成しプロンプトテンプレートに登録	
実験1	3	最新の国会答弁である第213回の質問主意書データを所定のプロンプトの「質問」に入力し、生成を実行 ※ 1つの質問に複数の質問が含まれる場合、判定を容易にするため、1つ目の枝問までを質問とみなして実験を行った	
	4	プロンプトの出力結果と、実際の答弁データを比較して評価	
実験2	5	実際の答弁データに含まれるキーワード (法令名・統計情報 等)を抽出し、答弁の「キーワード」としてプロンプトに追加入力し、再度生成を実行	
	6	プロンプトの出力結果と、実際の答弁データを比較して評価	

D-1.質問主意書に対する答弁ドラフト生成
試行過程 (1/2)

LLM	Claude 3 Sonnet

RAG	Web検索オフ

プロンプト

#命令
あなたは[国会議員からの質問に対して、
答弁を作成する担当者]である。
国会議員から[#質問]という質問を受け取った。
以下の[#前提]と[#制約条件]に従い、
[議会答弁]ドラフトを作成せよ。

#質問
質問

#前提
・答弁は政府の公な見解や意思である。
・答弁は質問議員に向けてだけでなく、
国民全体に向けての回答である。
・学習セットの過去答弁を参考に、回答を作成すること。
・答弁のキーワードはキーワードなので、
必ず答弁内で使うこと。
・回答字数は400字以内とすること。

制約条件
・答弁は簡潔かつ明瞭な表現を用いて作成
・答弁の語尾は常体「だ・である調」で生成。
敬体「です・ます調」は禁止
・個人的な感情や偏った思想に基づいた答弁作成は禁止

質問 例: 第213回 衆議院 No.10	我が国の農地面積がピーク時の昭和三十六年から三割減少していることに関連して、政府は、荒廃農地が発生する主な要因として、「高齢化、病気」や「労働力不足」により適切な農業生産活動を行うことが困難となっていることが挙げられると説明している。また、非農業用途等への転用が発生する原因として、例えば、民間企業が住宅や工場を建設する場合に土地の価格等の観点から農地が選好されやすいことが考えられると説明している。このような状況等を踏まえ、政府としては農地の維持・確保のためにどのような対策を講じようとしているのか明らかにされたい。
実際の答弁 (模範解答)	お尋ねの「農地の維持・確保」のための対策については、荒廃農地（現に耕作に供されておらず、耕作の放棄により荒廃し、通常の農作業では作物の栽培が客観的に不可能となっている農地をいう。以下同じ。）の発生に関しては、農業経営基盤強化促進法（昭和五十五年法律第六十五号）第六条第一項に規定する基本構想を定めている市町村で令和七年三月までに定めるものとされている同法第十九条第一項に規定する地域計画（以下「地域計画」という。）において、その区域において農業を担う者ごとに利用する農用地等（同法第四条第一項に規定する「農用地等」をいう。以下同じ。）を地図に表示し、これを特定した上で、農地中間管理事業の推進に関する法律（平成二十五年法律第百一号）第十七条第二項の規定に基づき、当該区域において農地中間管理事業（同法第二条第三項に規定する農地中間管理事業をいう。以下同じ。）を重点的に行い、農業の担い手への農地の集積・集約化を図ることにより、荒廃農地の発生を抑制していく考えである。また、農地転用に関しては、令和五年十二月二十七日に食料安定供給・農林水産業基盤強化本部で取りまとめた「「食料・農業・農村政策の新たな展開方向」に基づく具体的な施策の内容」において、「将来にわたっての農地の総量確保と適正・有効利用のための措置を強化する」として「令和六年の通常国会への改正法提出も視野」に、「地域計画内農地の転用規制強化の観点から、地域計画内農地の農用地区域への編入を促進する」等の施策の具体化の方向性について定めたところである。

試行過程（2/2）

実際の答弁（模範解答）

お尋ねの「農地の維持・確保」のための対策については、荒廃農地（現に耕作に供されておらず、耕作の放棄により荒廃し、通常の農作業では作物の栽培が客観的に不可能となっている農地をいう。以下同じ。）の発生に関しては、農業経営基盤強化促進法（昭和五十五年法律第六十五号）第六条第一項に規定する基本構想を定めている市町村で令和七年三月までに定めるものとされている同法第十九条第一項に規定する地域計画（以下「地域計画」という。）において、その区域において農業を担う者ごとに利用する農用地等（同法第四条第一項に規定する「農用地等」をいう。以下同じ。）を地図に表示し、これを特定した上で、農地中間管理事業の推進に関する法律（平成二十五年法律第百一号）第十七条第二項の規定に基づき、当該区域において農地中間管理事業（同法第二条第三項に規定する農地中間管理事業をいう。以下同じ。）を重点的に行い、農業の担い手への農地の集積・集約化を図ることにより、荒廃農地の発生を抑制していく考えである。また、農地転用に関しては、令和五年十二月二十七日に食料安定供給・農林水産業基盤強化本部で取りまとめた「「食料・農業・農村政策の新たな展開方向」に基づく具体的な施策の内容」において、「将来にわたっての農地の総量確保と適正・有効利用のための措置を強化する」として「令和六年の通常国会への改正法提出も視野」に、「地域計画内農地の転用規制強化の観点から、地域計画内農地の農用地区域への編入を促進する」等の施策の具体化の方向性について定めたところである。

実験1 生成結果

政府は、荒廃農地の発生要因として指摘のとおり、高齢化や労働力不足による生産活動の困難さ、非農業利用のための農地転用など、様々な課題を認識している。これらを踏まえ、政府としては以下の対策を講じている。
まず、人材確保の面では、令和2年度補正予算において、就農相談体制の整備や、農業研修支援、農業法人への雇用支援などに必要な経費を計上。農業従事者の高齢化対策と担い手の確保に努めている。
次に、農地の利用集積を進めるため、「人・農地プラン」の実質化に向け、集落営農組織の育成や農地中間管理機構を通じた農地の集積・集約化を支援している。さらに令和3年度予算案においても、農地の確保・有効利用に必要な経費を計上した。
このように、政府は人材確保と農地の利用集積に重点を置き、荒廃農地の発生防止と農地の維持・確保に向けた取組を推進している。

キーワード

農業経営基盤強化促進法、農地中間管理事業の推進に関する法律、農地中間管理事業、「食料・農業・農村政策の新たな展開方向」に基づく具体的な施策の内容

実験2 生成結果

政府としては、農地の確保・有効利用を図るため、以下の取組を進めている。農業経営基盤強化促進法に基づき、農地の権利移転を計画的に進めるとともに、農地中間管理事業の推進に関する法律に基づき、農地中間管理機構による農地の借り入れ・集積と、担い手農家等への集積を一体的に実施し、農地の集約化を図っている。また、「食料・農業・農村政策の新たな展開方向」に基づき、担い手への農地集積の加速化、遊休農地発生の防止、農地利用の最適化のための施策を総合的に講じている。こうした取組を通じ、荒廃農地の発生防止や再生利用を図るとともに、優良農地を確保し、力強い農業を実現することを目指している。

（参考）集計フォーム

会期	議院	番号	質問者	質問ID	質問	模範解答	実験1回答	実験1評価	キーワード	実験2回答	実験2評価
213	衆	10	原口一博君	10	我が国の農地面積がピーク時の昭和三十六年から三割減少していることに関連して、政府は、荒廃農地が発生する主な要因として、「高齢化、病気」や「労働力不足」により適切な農業生産活動を行うことが困難となっていることが挙げられると説明している。また、非農業用途等への転用が発生する原因として、例えば、民間企業が住宅や工場を建設する場合に土地の価格等の観点から農地が選好されやすいことが考えられると説明している。このような状況等を踏まえ、政府としては農地の維持・確保のためにどのような対策を講じようとしているのか明らかにされたい。	お尋ねの「農地の維持・確保」のための対策については、荒廃農地（現に耕作に供されておらず、耕作の放棄により荒廃し、通常の農作業では作物の栽培が客観的に不可能となっている農地をいう。以下同じ。）の発生に関しては、農業経営基盤強化促進法（昭和五十五年法律第六十五号）第六条第一項に規定する基本構想を定めている市町村で令和七年三月までに定めるものとされている同法第十九条第一項に規定する地域計画（以下「地域計画」という。）において、その区域において農業を担う者ごとに利用する農用地等（同法第四条第一項に規定する「農用地等」をいう。以下同じ。）を地図に表示し、これを特定した上で、農地中間管理事業の推進に関する法律（平成二十五年法律第百一号）第十七条第二項の規定に基づき、当該区域において農地中間管理事業（同法第二条第三項に規定する農地中間管理事業をいう。以下同じ。）を重点的に行い、農業の担い手への農地の	政府は、荒廃農地の発生要因として指摘のとおり、高齢化や労働力不足による生産活動の困難さ、非農業利用のための農地転用など、様々な課題を認識している。これらを踏まえ、政府としては以下の対策を講じている。 まず、人材確保の面では、令和2年度補正予算において、就農相談体制の整備や、農業研修支援、農業法人への雇用支援などに必要な経費を計上。農業従事者の高齢化対策と担い手の確保に努めている。 次に、農地の利用集積を進めるため、「人・農地プラン」の実質化に向け、集落営農組織の育成や農地中間管理機構を通じた農地の集積・集約化を支援している。さらに令和3年度予算案においても、農地の確保・有効利用に必要な経費を計上した。 このように、政府は人材確保と農地	×	農業経営基盤強化促進法、農地中間管理事業の推進に関する法律、農地中間管理事業、「食料・農業・農村政策の新たな展開方向」に基づく具体的な施策の内容	政府としては、農地の確保・有効利用を図るため、以下の取組を進めている。 農業経営基盤強化促進法に基づき、農地の権利移転を計画的に進めるとともに、農地中間管理事業の推進に関する法律に基づき、農地中間管理機構による農地の借り入れ・集積と、担い手農家等への集積を一体的に実施し、農地の集約化を図っている。また、「食料・農業・農村政策の新たな展開方向」に基づき、担い手への農地集積の加速化、遊休農地発生の防止、農地利用の最適化のための施策を総合的に講じている。こうした取組を通じ、荒廃農地の発生防止や再生利用を図るとともに、優良農地を確保し、力強い農業を実現することを目指して	○

実験結果のまとめ

集計結果

質問ID	実験1評価	実験2評価	評価
1	×	○	向上
2	×	○	向上
3	△	△	変化なし
4	△	○	向上
5	△	○	向上
6	×	○	向上
7	×	×	変化なし
8	○	○	変化なし
9	△	△	変化なし
10	×	○	向上

	○	△	×	総計
○	1			1
△	2	2		4
×	4		1	5
総計	7	2	1	10

考察

過去の答弁データからの生成精度については限定的な結果に留まった

- 事実・データを問う質問に対して、正しく回答を生成することは難しい
  - 例: No.1 中規模・小規模の飲食店の倒産件数等の経営状況
  - 例: No.7 出入力管理施設の常勤医師の在籍状況
- 政府として把握していないものを、把握していないと答えるのは難しい
  - 例: No.6 上陸特別許可の件数

追加データとして、回答作成の担当者が考えているキーワードを与えることで、回答精度の向上が望める

- 事実となるファクト
  - 例: No.2 システム改修の完了見込み時期
- 理由のキーワード
  - 例: No.4 買春に罰則が設けられていない理由
- 根拠となる法令
  - 例: No.10 過去の農地対策の根拠となる法令

今後の方向性としては、2つが考えられる

- 担当者の文章作成支援用ツールとして、あくまで答弁キーワード(回答の核となる概念)は職員が与えるものとする
- 答弁キーワードを抽出できるよう、学習データの範囲を拡大する
  - 例: 各省庁のWebサイト、首相・大臣や政府関係者の会見 等

検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

E-1.ソースコードの言語変換

ユースケース概要

前提	ソースコードの言語変換に生成AIを活用できる可能性を探るため、 簡易的なソースコードの変化が可能かを確認することを主眼を置いて検証した
課題	言語の違いによる、構文やライブラリ、コードスタイルの違いに生成AIがどれだけ対応できるかを検証する必要がある

生成AIによるソースコードの言語変換

今回検証したケース

ケース1)

javaのプログラムをpythonのコードに変換

ケース2)

COBOLのプログラムをjavaのコードに変換

ケース1(Java→Python)　作業ステップ

No	検証段階	アクション	アクションの理由(簡潔に)
1	仮検証	500行程度のjavaのプログラムをGaiXerによって生成	検証に利用するプログラムを生成するため
2		1に続けてリクエストを送り、javaのソースから処理概要を生成	言語変換にあたり、処理概要を整理するため
3		2に続けてリクエストを送り、処理概要からプログラム設計書を生成	言語変換にあたり、設計内容を整理するため
4		3に続けてリクエストを送り、javaのプログラムにコメントを付与	処理内容を変更せずに、言語変換ができているかを確認しやすくするため
5		4に続けてリクエストを送り、コメントを付与したプログラムを可読性を上げてコードを変更（処理の内容を変えずに、javaのコードを変更）	言語変換をしやすくするため
6		5に続けてリクエストを送り、javaのプログラムをpythonのコードに変換	ソースコードを別の言語に変換するため

E-1.ソースコードの言語変換

ケース2(COBOL→Java)　作業ステップ

No	検証段階	アクション	アクションの理由(簡潔に)
1	仮検証	500行程度のCOBOLのプログラムをGaiXerによって生成	検証に利用するプログラムを生成するため
2		1に続けてリクエストを送り、COBOLのソースから処理概要を生成	言語変換にあたり、処理概要を整理するため
3		2に続けてリクエストを送り、処理概要からプログラム設計書を生成	言語変換にあたり、設計内容を整理するため
4		3に続けてリクエストを送り、COBOLのプログラムにコメントを付与	処理内容を変更せずに、言語変換ができているかを確認しやすくするため
5		4に続けてリクエストを送り、コメントを付与したプログラムを可読性を上げてコードを変更（処理の内容を変えずに、COBOLのコードを変更）	言語変換をしやすくするため
6		5に続けてリクエストを送り、COBOLのプログラムをjavaのコードに変換	ソースコードを別の言語に変換するため

ケース1(Java→Python)　プロンプトの試行過程(1/6)

LLM	Claude 3 Sonnet

プロンプト#1

500行のjavaのプログラムコードを生成してください。

生成結果

ここに500行のJavaプログラムコードを生成しました。このコードは、簡単なコンソールベースのToDoリストアプリケーションを実装しています。

```
import java.util.ArrayList;
import java.util.Scanner;

class ToDoItem {
    private String description;
    private boolean completed;

    public ToDoItem(String description) {
        this.description = description;
        this.completed = false;
    }

    public String getDescription() {
        return description;
    }
```

・・・
中略
・・・

```
    private static void markItemAsCompleted() {
        System.out.print("Enter item index to mark as completed: ");
        int index = getIntInput("");
        list.markItemAsCompleted(index - 1);
    }

    private static void unmarkItemAsCompleted() {
        System.out.print("Enter item index to unmark as completed: ");
        int index = getIntInput("");
        list.unmarkItemAsCompleted(index - 1);
    }
}
```

このプログラムでは、ToDoItemクラスがToDoリストの各項目を表し、ToDoListクラスがToDoリストの管理を行います。ToDoListAppクラスがメインクラスとなり、コンソールベースのユーザーインターフェースを提供しています。

ユーザーは、メニューから項目の追加、削除、完了/未完了の切り替え、リストの表示などの操作を選択できます。プログラムはユーザーの入力を処理し、ToDoリストを適切に更新します。

ケース1(Java→Python)　プロンプトの試行過程(2/6)

LLM	Claude 3 Sonnet

プロンプト#2

このプログラムの処理概要を簡潔に生成してください

生成結果

このプログラムの処理概要は以下の通りです。

ToDoItemクラスでToDoリストの各項目を表す。項目の説明文と完了状態を保持する。

ToDoListクラスでToDoリストを管理する。項目の追加、削除、完了/未完了の切り替え、リストの表示などの操作を提供する。

ToDoListAppクラスがメインクラスとなる。コンソールベースのユーザーインターフェースを提供する。

メインメソッドでメニューを表示し、ユーザーの入力を受け取る。

ユーザーの選択に応じて、ToDoリストへの操作を実行する。

項目の追加
項目の削除
項目の完了/未完了の切り替え
リストの表示
プログラムの終了
ToDoリストの操作はToDoListクラスのメソッドを呼び出して実行される。

プログラムはユーザーの入力を適切に処理し、ToDoリストを更新する。

要約すると、このプログラムはToDoリストの管理機能を提供するコンソールベースのアプリケーションです。ユーザーはメニューから操作を選択し、ToDoリストを編集できます。

ケース1(Java→Python)　プロンプトの試行過程(3/6)

LLM	Claude 3 Sonnet

プロンプト#3

この処理概要からプログラム設計書を生成してください

生成結果

プログラム設計書を以下に示します。

1. システム概要

このシステムはコンソールベースのToDoリストアプリケーションである。
ユーザーは、ToDoリストの項目の追加、削除、完了/未完了の切り替え、リストの表示などの操作を行うことができる。
2. クラス設計

2.1. ToDoItem クラス

ToDoリストの各項目を表すクラス
フィールド
description (String): 項目の説明
completed (boolean): 項目の完了状態
メソッド
getDescription() / setDescription(String): 項目の説明を取得/設定する
isCompleted() / setCompleted(boolean): 項目の完了状態を取得/設定する
toString(): 項目の文字列表現を返す
2.2. ToDoList クラス

ToDoリストを管理するクラス
フィールド
items (ArrayList<ToDoItem>): ToDoリストの項目のリスト
メソッド
addItem(ToDoItem): 新しい項目をリストに追加する
removeItem(int): 指定のインデックスの項目をリストから削除する
markItemAsCompleted(int) / unmarkItemAsCompleted(int): 指定のインデックスの項目の完了状態を切り替える
displayList(): リストの全項目を表示する
2.3. ToDoListApp クラス

メインクラス
フィールド
list (ToDoList): ToDoリストのインスタンス
scanner (Scanner): ユーザー入力を受け取るためのスキャナー
メソッド
main(String[]): プログラムのエントリポイント
displayMenu(): メニューを表示する
getIntInput(String): ユーザーから整数入力を受け取る
handleChoice(int): ユーザーの選択に応じて適切な処理を行う
addItem() / removeItem() / markItemAsCompleted() / unmarkItemAsCompleted(): 対応する操作を実行する
3. ユースケース

3.1. 項目の追加

ユーザーがメニューから"1. Add item"を選択する
プログラムが項目の説明の入力を求める
ユーザーが項目の説明を入力する
プログラムが新しいToDoItemインスタンスを作成し、ToDoListに追加する
3.2. 項目の削除

ユーザーがメニューから"2. Remove item"を選択する
プログラムが削除する項目のインデックスの入力を求める
ユーザーが項目のインデックスを入力する
プログラムがToDoListから指定のインデックスの項目を削除する
3.3. 項目の完了/未完了の切り替え

ユーザーがメニューから"3. Mark item as completed"または"4. Unmark item as completed"を選択する
プログラムが対象の項目のインデックスの入力を求める
ユーザーが項目のインデックスを入力する
プログラムがToDoListの指定のインデックスの項目の完了状態を切り替える
3.4. リストの表示

ユーザーがメニューから"5. Display list"を選択する
プログラムがToDoListの全項目を表示する
3.5. プログラムの終了

ユーザーがメニューから"6. Exit"を選択する
プログラムが終了する

ケース1(Java→Python)　プロンプトの試行過程(4/6)

LLM	Claude 3 Sonnet

プロンプト#4

生成したプログラムにコメントを追加してください。

生成結果

プログラムにコメントを追加しました。

```
import java.util.ArrayList;
import java.util.Scanner;

// ToDoリストの各項目を表すクラス
class ToDoItem {
    private String description; // 項目の説明
    private boolean completed; // 項目の完了状態

    public ToDoItem(String description) {
        this.description = description;
        this.completed = false; // 新しい項目は未完了状態で作成される
    }

    public String getDescription() {
        return description;
    }


・・・
中略
・・・

// 項目を完了状態にする
    private static void markItemAsCompleted() {
        System.out.print("Enter item index to mark as completed: ");
        int index = getIntInput("");
        list.markItemAsCompleted(index - 1); // インデックスは0から始まるため、1を引く
    }

    // 項目を未完了状態にする
    private static void unmarkItemAsCompleted() {
        System.out.print("Enter item index to unmark as completed: ");
        int index = getIntInput("");
        list.unmarkItemAsCompleted(index - 1); // インデックスは0から始まるため、1を引
く
    }
}
```

このコードには、各クラスとメソッドの役割や処理の流れを説明するコメントが追加されています。また、特定の処理の詳細や注意点についても、適切にコメントが記述されています。

ケース1(Java→Python)　プロンプトの試行過程(5/6)

LLM	Claude 3 Sonnet

プロンプト#5

生成したプログラムをリファクタリングし、リファクタリング後のプログラムを全文出力してください。
その際、処理内容は変更しないでください。

生成結果

リファクタリングを行いました。主な変更点は以下の通りです。

メソッド名の改善
冗長なコードの削減
コードの可読性の向上
リファクタリング後のプログラムは以下の通りです。

```
import java.util.ArrayList;
import java.util.Scanner;

// ToDoリストの各項目を表すクラス
class ToDoItem {
    private String description;
    private boolean isCompleted;

    public ToDoItem(String description) {
        this.description = description;
        this.isCompleted = false;
    }
```

・・・
中略
・・・

```
    private static void markItemAsIncomplete() {
        System.out.print("Enter item index to mark as incomplete: ");
        int index = getIntInput("");
        list.markAsIncomplete(index - 1);
    }
}
```

主な変更点は以下の通りです。

1.ToDoListクラスでitemsフィールドをfinalで初期化し、コンストラクタを削除した。
2.markItemAsCompletedとunmarkItemAsCompletedメソッドを、より一般的なmarkAsCompletedとmarkAsIncompleteメソッドに置き換えた。
3.isValidIndexとtoggleCompletionStatusメソッドを追加し、重複コードを削減した。
4.ToDoListAppクラスで、handleChoiceメソッドが実行の継続/終了を制御するように変更した。
5.addItemメソッドでscanner.nextLine()を使用し、残りの改行文字を適切に消費するようにした。
6.コーディングスタイルを一貫させ、可読性を向上させた。

これらの変更により、コードの冗長性が減り、可読性と保守性が向上しています。しかし、プログラムの機能や処理の流れは変更されていません。

ケース1(Java→Python)　プロンプトの試行過程(6/6)

LLM	Claude 3 Sonnet

プロンプト#6

生成したプログラムをpythonのコードに変換してください

作成者の評価

段階を踏むことで、JavaからPythonへの変換ができることを確認。
今後、LLMにテストコードを生成させ、プログラムとしての同値性検証ができると尚可。

生成結果

JavaからPythonのコードに変換しました。

```
class ToDoItem:
    def __init__(self, description):
        self.description = description
        self.is_completed = False

    def __str__(self):
        return f"{'[X]' if self.is_completed else '[ ]'} {self.description}"

class ToDoList:
    def __init__(self):
        self.items = []

    def add_item(self, item):
        self.items.append(item)

    def remove_item(self, index):
        if self.is_valid_index(index):
```

・・・
中略
・・・

```
def main():
    todo_list = ToDoList()
    running = True
    while running:
        display_menu()
        choice = get_int_input("Enter your choice: ")
        running = handle_choice(choice, todo_list)

if __name__ == "__main__":
    main()
```

このPythonコードは、Javaのコードと同様の機能を提供しています。主な変更点は以下の通りです。

・クラスの定義方法がJavaとは異なります。
・__str__メソッドを使用して、オブジェクトの文字列表現を定義しています。
・リストの操作は、Pythonのリスト機能を使用しています。
・入力の取得には、input関数を使用しています。
・例外処理を使用して、不正な入力を適切に処理しています。
・メインロジックはmain関数に集約されています。

メインの処理の流れはJavaのコードと同様ですが、Pythonの言語機能に合わせて適切に変換されています。

ケース2(COBOL→Java)　プロンプトの試行過程(1/6)

LLM	Claude 3 Sonnet

プロンプト#1

cobol言語の500行程度のコードを生成してください。

生成結果

はい、以下に500行程度のCOBOLコードのサンプルを示します。このコードはサンプル用ですので、実際の業務で使用する際は内容を確認する必要があります。

```
IDENTIFICATION DIVISION.
PROGRAM-ID. SAMPLE-PROGRAM.

ENVIRONMENT DIVISION.
INPUT-OUTPUT SECTION.
FILE-CONTROL.
   SELECT CUSTOMER-FILE ASSIGN TO CUSTOMER-FILE
      ORGANIZATION IS LINE SEQUENTIAL.
   SELECT REPORT-FILE ASSIGN TO REPORT-FILE
      ORGANIZATION IS LINE SEQUENTIAL.

DATA DIVISION.
FILE SECTION.
FD  CUSTOMER-FILE.
01  CUSTOMER-RECORD.


・・・
中略
・・・

4000-PRINT-CUSTOMER-DETAIL.
   READ CUSTOMER-FILE
      AT END
         MOVE 'Y' TO WS-END-OF-FILE
      NOT AT END
         MOVE SPACES TO REPORT-RECORD
         STRING CUST-ID DELIMITED BY SIZE
               ' ' DELIMITED BY SIZE
               CUST-NAME DELIMITED BY SIZE
               ' ' DELIMITED BY SIZE
               CUST-ADDRESS DELIMITED BY SIZE
               ' ' DELIMITED BY SIZE
               CUST-CITY DELIMITED BY SIZE
               ' ' DELIMITED BY SIZE
               CUST-STATE DELIMITED BY SIZE
               ' ' DELIMITED BY SIZE  CUST-ZIP DELIMITED BY SIZE
               ' $' DELIMITED BY SIZE
               CUST-BALANCE DELIMITED BY SIZE
               INTO REPORT-RECORD
         WRITE REPORT-RECORD AFTER ADVANCING 1 LINE.
```

このコードは、顧客ファイルを読み込み、顧客レポートを生成するサンプルプログラムです。メインラインでは、顧客ファイルを開いて顧客レコードを読み込み、合計顧客数と合計残高を集計します。次に、レポートファイルを開いてレポートヘッダーを書き込みます。その後、もう一度顧客ファイルを開き、各顧客の詳細をレポートに書き込んでいます。

このサンプルでは、ファイル入出力、レコード読み書き、文字列操作、日付関数の使用などの基本的な処理を含んでいますが、実際の業務システムではさらに複雑な処理が必要になる可能性があります。

ケース2(COBOL→Java)　プロンプトの試行過程(2/6)

LLM	Claude 3 Sonnet

プロンプト#2

あなたはCOBOL言語に詳しいエンジニアです。
このプログラムの処理内容を簡潔に示してください。

生成結果

はい、このCOBOLプログラムの処理内容を簡潔に説明します。

顧客ファイル(CUSTOMER-FILE)と報告ファイル(REPORT-FILE)を開きます。

顧客ファイルから1レコードずつ読み込み、以下の集計を行います。

顧客数(WS-CUST-COUNT)
顧客残高合計(WS-TOTAL-BALANCE)
報告ファイルにレポートヘッダーを出力します。

レポート名と日付
顧客数と残高合計
顧客ファイルを再度開き、1レコードずつ読み込んで報告ファイルに出力します。

顧客ID
顧客名
住所
市区町村
州
郵便番号
残高
顧客ファイルと報告ファイルを閉じます。

要約すると、顧客ファイルからデータを読み込み、集計を行った後、報告ファイルにレポートを出力するプログラムとなっています。
典型的な伝統的なCOBOLプログラムの構造で、ファイル入出力と報告生成の基本的な処理を含んでいます。

ケース2(COBOL→Java)　プロンプトの試行過程(3/6)

LLM	Claude 3 Sonnet

プロンプト#3

上記の処理概要から、プログラム設計書を生成してください。

生成結果

はい、処理概要からプログラム設計書を作成します。

プログラム名: SAMPLE-PROGRAM

【目的】
顧客ファイルから顧客データを読み込み、顧客残高の集計と顧客詳細のレポートを出力する。

【入力ファイル】

CUSTOMER-FILE
レコード構造
CUST-ID (顧客ID)
CUST-NAME (顧客名)
CUST-ADDRESS (住所)
CUST-CITY (市区町村)
CUST-STATE (州)
CUST-ZIP (郵便番号)
CUST-BALANCE (残高)
【出力ファイル】

REPORT-FILE
レポートヘッダー
レポート名
作成日付
顧客数
残高合計
顧客詳細
顧客ID
顧客名
住所
市区町村
州
郵便番号
残高
【処理手順】

初期化
1.1 カウンター(顧客数、残高合計)を0にする
1.2 EOFスイッチをNoにする

顧客ファイル読み込み
2.1 顧客ファイルから1レコードを読み込む
2.2 EOF時は手順3へ
2.3 顧客数をカウントアップ
2.4 残高を合計に加算
2.5 手順2.1へ戻る

レポート出力
3.1 レポートヘッダーを出力

レポート名、日付、顧客数、残高合計 3.2 顧客ファイルを開く 3.3 1レコードずつ読み込んで顧客詳細を出力
顧客ID、名前、住所、市区町村、州、郵便番号、残高 3.4 EOF時は手順4へ 3.5 手順3.3へ戻る
終了処理
4.1 顧客ファイルと報告ファイルを閉じる

【備考】

処理中に予期せぬエラーが発生した場合の回復手順が必要
以上がプログラム設計書の概要になります。処理手順の詳細、データ構造の詳細、例外処理などを追記する必要があります。

ケース2(COBOL→Java)　プロンプトの試行過程(4/6)

LLM	Claude 3 Sonnet

プロンプト#4

生成したCOBOLのプログラムに、日本語でコメントをつけてください。

生成結果

はい、COMBOLプログラムにコメントを追加します。

```
 IDENTIFICATION DIVISION.
 PROGRAM-ID. SAMPLE-PROGRAM.
*>プログラム名: SAMPLE-PROGRAM
*>目的: 顧客ファイルから顧客データを読み込み、顧客残高の集計と顧客詳細のレポートを出力する

 ENVIRONMENT DIVISION.
 INPUT-OUTPUT SECTION.
 FILE-CONTROL.
    SELECT CUSTOMER-FILE ASSIGN TO CUSTOMER-FILE
       ORGANIZATION IS LINE SEQUENTIAL.
*>入力ファイル: 顧客ファイル(CUSTOMER-FILE)
*>レコード構造: 顧客ID、顧客名、住所、市区町村、州、郵便番号、残高
    SELECT REPORT-FILE ASSIGN TO REPORT-FILE
       ORGANIZATION IS LINE SEQUENTIAL.
*>出力ファイル: レポートファイル(REPORT-FILE)

・・・
中略
・・・

 *>顧客詳細出力
 4000-PRINT-CUSTOMER-DETAIL.
    READ CUSTOMER-FILE
       AT END
          MOVE 'Y' TO WS-END-OF-FILE
*>ファイル終了時にスイッチをセット
       NOT AT END
          MOVE SPACES TO REPORT-RECORD
          STRING CUST-ID DELIMITED BY SIZE
                 ' ' DELIMITED BY SIZE
                 CUST-NAME DELIMITED BY SIZE
                 ' ' DELIMITED BY SIZE
                 CUST-ADDRESS DELIMITED BY SIZE
                 ' ' DELIMITED BY SIZE
                 CUST-CITY DELIMITED BY SIZE
                 ' ' DELIMITED BY SIZE
                 CUST-STATE DELIMITED BY SIZE
                 ' ' DELIMITED BY SIZE
                 CUST-ZIP DELIMITED BY SIZE
                 ' $' DELIMITED BY SIZE
                 CUST-BALANCE DELIMITED BY SIZE
                 INTO REPORT-RECORD
*>顧客詳細(ID、名前、住所、市区町村、州、郵便番号、残高)を出力
          WRITE REPORT-RECORD AFTER ADVANCING 1 LINE.
```

上記のように、プログラムの目的、処理の概要、変数の意味などを日本語でコメントとして追加しました。
コメントによりプログラムの理解が深まり、保守性が向上すると考えられます。

ケース2(COBOL→Java)　プロンプトの試行過程(5/6)

LLM	Claude 3 Sonnet

プロンプト#5

生成したコードをリファクタリングし、リファクタリング後のコードを全文出力してください。
ただし、コードの処理内容は、絶対変更しないでください。

生成結果

はい、コードをリファクタリングします。処理内容は変更せずに、可読性とメンテナンス性を高めるようにコードを整理しました。

```
 IDENTIFICATION DIVISION.
 PROGRAM-ID. SAMPLE-PROGRAM.
*>プログラム名: SAMPLE-PROGRAM
*>目的: 顧客ファイルから顧客データを読み込み、顧客残高の集計と顧客詳細のレポートを出力する

 ENVIRONMENT DIVISION.
 INPUT-OUTPUT SECTION.
 FILE-CONTROL.
    SELECT CUSTOMER-FILE ASSIGN TO CUSTOMER-FILE
       ORGANIZATION IS LINE SEQUENTIAL.
*>入力ファイル: 顧客ファイル(CUSTOMER-FILE)
*>レコード構造: 顧客ID、顧客名、住所、市区町村、州、郵便番号、残高
    SELECT REPORT-FILE ASSIGN TO REPORT-FILE
       ORGANIZATION IS LINE SEQUENTIAL.
*>出力ファイル: レポートファイル(REPORT-FILE)

・・・
中略
・・・

*>終了処理
4000-TERMINATE.
    CLOSE CUSTOMER-FILE
          REPORT-FILE.
*>ファイルをクローズ
```

リファクタリングの内容は以下の通りです。

・メイン処理の流れを明確にするため、プロシージャを分割しました。
　・初期化処理(1000-INITIALIZE)
　・顧客データ処理(2000-PROCESS-CUSTOMER-FILE)
　・レポート出力処理(3000-PRODUCE-REPORT)
　・終了処理(4000-TERMINATE)
・読み込み処理と集計処理を分離しました。
　・顧客レコード読み込み(2100-READ-CUSTOMER-RECORD)
　・顧客データ集計(2200-ACCUMULATE-CUSTOMER-DATA)
・レポート出力処理をさらに分割しました。
　・レポートヘッダー出力(3100-WRITE-REPORT-HEADER)
　・顧客詳細出力(3200-WRITE-CUSTOMER-DETAILS)
　・顧客詳細フォーマット(3220-FORMAT-CUSTOMER-DETAIL)
・コーディングスタイルを統一しました。
　・インデントを整理
　・無駄な空白行を削除
　・コメントを適切に配置
　・88レベル項目の記述を簡潔にしました。

このようにリファクタリングすることで、コードの可読性とメンテナンス性が向上し、将来的な修正や機能拡張が容易になると考えられます。

ケース2(COBOL→Java)　プロンプトの試行過程(6/6)

LLM	Claude 3 Sonnet

プロンプト#6

リファクタリングを行ったコードをjavaのコードに変換してください。

作成者の評価

段階を踏むことで、COBOLからjavaへの変換ができることを確認。
Python変換と同様の課題に加え、COBOL特有の命令を含む場合等、更なる調査余地あり。

生成結果

はい、COBOLコードをJavaに変換します。

```
/**
 * プログラム名: SAMPLE-PROGRAM
 * 目的: 顧客ファイルから顧客データを読み込み、顧客残高の集計と顧客詳細のレポートを出力する
 */
import java.io.*;
import java.text.SimpleDateFormat;
import java.util.ArrayList;
import java.util.Date;
import java.util.List;

public class SampleProgram {
    private static final String CUSTOMER_FILE = "customer.txt";
    private static final String REPORT_FILE = "report.txt";]

・・・
中略
・・・

    private static void writeCustomerDetails(BufferedWriter writer, List<Customer> customers) throws IOException {
        for (Customer customer : customers) {
            writer.write(String.format("%07d  %-20s  %-30s  %-15s  %-2s  %05d  $%9.2f", customer.customerId, customer.name, customer.address, customer.city, customer.state, customer.zipCode, customer.balance));
            writer.newLine();
        }
    }

    private static void terminate() {
        // ファイルをクローズ
    }
}
```

上記のJavaコードは、COBOLコードの機能を維持しつつ、Javaのプログラミングスタイルに合わせて変換しています。
主な変更点は以下の通りです。

・クラスとメソッドの構造に変更しました。
・顧客データをCustomerクラスとして表現しました。
・ファイル入出力をBufferedReader/BufferedWriterを使って実装しました。
・顧客データのリストを作成し、ストリーム操作で集計を行うようにしました。
・レポートの出力をStringのフォーマット化して行うようにしました。

COBOLのような手続き型言語からオブジェクト指向言語に変換する際は、データ構造とプログラムの流れを適切にクラスとメソッドに分割することが重要になります。また、Java特有の機能や構文を活用することで、より読みやすく保守性の高いコードになります。
ただし、このコードにはまだ改善の余地があり、例えばファイル入出力の例外処理やカスタマイズされた例外クラスの導入などが考えられます。実際のシステムでは、さらなる要件に応じてコードを拡張する必要があります。

総括

総括　：　Claude 3 Sonnet にてソースコードの言語変換を行うことは可能

考察　：

・本検証では500行という短いプログラムでの言語変換を実施したため、長いプログラムにおいても変換が可能かはさらなる検証が必要。
・実用上は、言語変換を行った際に元のプログラムと同じ処理をコーディングできているかを確認するステップが必要であり、
そのステップにも生成AIの活用余地があると考えられる。

検証ユースケースの全体像

検証ユースケース	最適な言語モデル	開発したプロンプトの性質 （どういったプロンプトタスクであるか）
A-1. パブリックコメント対応品質の向上	GPT-4 Turbo	長文を意図理解し、そのエッセンスを要約/意見分類
A-2. 調達仕様書のラベリング	GPT-4 Turbo	調達仕様書全文を入れることで調達仕様書のラベリングを行う
B-1. デジタル社会推進実践ガイドブック	Claude 2.1	質問の意図を理解し、長大な文章情報から適切な箇所を抜粋して回答を行う
B-1'. デジタル社会推進実践ガイドブック 自治体向けbot運用	Claude 3 Sonnet	生成速度が速く、違和感のない日本語の生成を行う
B-2. 人事院規則	Claude 3 Sonnet	適切な箇所から正確に情報を抽出し、簡潔な文章を生成する
B-3. デジタル庁問い合わせ窓口	A-1と類似ケースのため、A-1の検証に統合	
C-1. 用例による法令検索	GPT-3.5	入力したシステムプロンプトがシンプルであるため、速度重視し3.5を選択
C-2. 用語集作成業務の効率化	GPT-4 Turbo	ドキュメントから指定の条件を満たす用語リストを抽出し、その定義説明を生成
D-1. 質問主意書に対する答弁ドラフト生成	Claude 3 Sonnet	学習セットで登録されたデータを正確に読み解き、質の高い答弁を作成する
E-1. ソースコードの言語変換	Claude 3 Sonnet	プログラムから処理概要・設計書を作成し、ソースコードの言語を変換する
F-1. ドキュメントの機密度判定	GPT-4 Turbo	情報を入力するとその情報の機密度を判定し、判定した根拠と判定結果を回答する

ユースケース概要

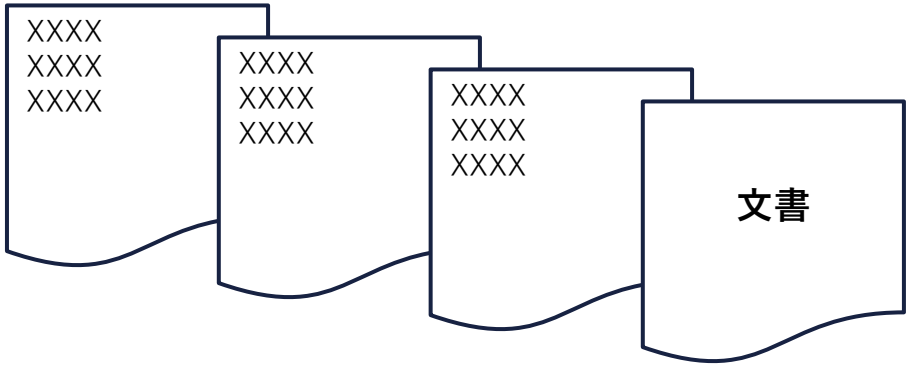
前提	行政で取り扱う情報は機密性に基づき分類がなされており、2以上にあたる情報は、その取扱いに当たり注意が必要 誤った運用や漏えいは、国民の権利侵害や行政業務への支障となり、国民への信頼を大きく損なうリスクとなりかねない
課題	行政業務で生成AIを活用するうえで、機密性の高い情報を取り扱う場合のリスク管理体制を整える必要がある

「政府機関の情報セキュリティ対策のための統一基準(第 4 版)」

格付けの区分	分類の基準
機密性３情報	行政事務で取り扱う情報のうち、秘密文書に相当する機密性を要する情報
機密性２情報	行政事務で取り扱う情報のうち、秘密文書に相当する機密性は要しないが、漏えいにより、国民の権利が侵害され又は行政事務の遂行に支障を及ぼすおそれがある情報
機密性１情報	機密性２情報又は機密性３情報以外の情報

機密性3情報に相当する秘密文書を入力することはできないため、
本検証においては、機密性1情報と機密性2情報の判定を取り扱った

生成AIによる機密度判定



生成AIによる判定

機密性１情報 or 機密性2情報

F-1.ドキュメントの機密度判定

ユーザーの作業分解とプロンプトの関連性

目的	・ 政府統一基準に則り、情報/文書の機密性を判定する。将来的には機密性の高い情報の生成AI入力を未然に防ぐ機能を開発する

作業工程

1. 文書の漏洩により、国民の権利が侵害されないかを確認する
2. 文書の漏洩により、行政業務の遂行に支障を及ぼすかを確認する
3. 文書の機密度を判定する

プロンプト

文書中の個人情報の有無と、
行政業務への影響を判定結果から、
文書の機密度を判定

検証アプローチ

・個人情報、機密性情報の定義をもとに判定するプロンプティング

・個人情報、機密性情報に関連する法律をプロンプトに組み込むことで、判定の精度向上を図る

検証結果

機密性情報の分類は高精度でできていたものの、100%の精度ではないため、未然に防ぐ機能として活用することは難しい。一方で、既に入力してしまった情報に、機密度の高い情報が含まれていないかを自動的に集計する機能としては活用が見込まれる。

＜正答率の内訳＞

- 個人情報判定、業務に支障があるかの判定が適切か：90点
- 機密性情報分類が適切か：98点
- 機密性情報分類した根拠が適切か：80点

生成AI導入による効果

▶ 行政における機密度判定を防止的にチェックするメカニズムが導入された場合、政府のオペレーションにおける情報漏えいリスクを大きく削減するとともに、業務における機密度判定を人手で行う実務的懸念や工数が削減されるため、効率化への恩恵も期待できる

効果	効果内容
情報漏えいリスクの削減と運用効率化	自動で走るチェック機能のようなものを通すプロセスを設ける事で、機密度判定のチェックや該当箇所の指摘が自動で行われる。これにより、未然に情報漏えいリスクをヘッジできる
	また、本仕組みの導入により、自動的な判定ツールによる、チェックプロセスそのもののシステム化、効率化が期待できる
対応品質の向上	現状、機密度判定はルールとして制定されつつ、職員単位での知見による判定に依存しており、そのチェックプロセスや対応品質は標準化されていない
	本ユースケースのような防止プロセスがもし機能する場合、行政におけるそうしたチェックプロセスの標準化、並びに対応品質向上が実現できる

評価基準と評価要件

ポイント1)

機密性1情報と機密性2情報の定義の違いに基づいて、個人情報判定と業務への支障を評価基準として採用した

ポイント2)

分類が適切かという基準に加え、根拠を基準とすることで、分類のプロセスの評価が可能となるようにした

ポイント3)

他のユースケースにおいては、部分点を採用しているケースもあるものの、
機密度判定においては、正しい/誤っているの2択であるため、評価点数も0点か1点の2択とした

評価基準	要件定義	評価(点数)
回答の正しさ1:個人情報判定、業務に支障があるかの判定が適切か	どちらも正しく判定できている	1
	どちらか片方しか正しく判定できていない	0
	両方とも間違っている	0
回答の正しさ2:機密性情報分類が適切か	適切に分類できている	1
	分類が不適切である	0
回答の正しさ3:機密性情報分類した根拠が適切か	根拠が全て適切である	1
	根拠は適切だが、個人情報定義や機密性2情報詳細が間違っている	0
	根拠が不適切である	0

F-1.ドキュメントの機密度判定

作業/評価ステップ

No.	検証段階	アクション	アクションの理由(簡潔に)
1	クイック検証	GPT-4 Turboをベースとしたプロンプトを開発	簡単なプロンプトを使った場合の生成内容を確認するため
2	クイック検証	テストケースを20個作成(機密性を含むもの、そうでないもの、個人情報を含むもの、そうでないもの)	主要質問タイプに対しての対応性能を試験する為
3	本検証	精度改善のために定義を調査：「行政機関の保有する情報の公開に関する法律」から機密性2情報の詳細を、「個人情報の保護に関する法律」から個人情報の定義を探す	2で法令の詳細定義がないと精度が上がらないと判定したため
4	本検証	プロンプトを修正：法律における定義を生成AIで要約させ、元の意味内容を維持しつつ、コンパクトな形式でプロンプト本文にfew-shot rulesとして繁栄した	3に基づく改善案を反映し本検証を完了させるため
5	本検証	評価基準/要件の整備	機密度判定するサービスの期待値と評価ポイントを改めて明確化する為
6	総括	ユーザー面談実施	5の評価基準を基に相談
7	総括	レポート作成	-

テストケースに関する詳細

テストケース作成方針	機密度判定の根拠となる「個人情報の有無（特定の個人を特定できるものとできないもの）」「行政業務への影響」の組み合わせのパターンでテストケースを作成した

テストケース

質問パターン	質問数
特定の個人を特定できる情報を含む、行政業務への影響がある、機密性2情報	14
特定の個人を特定できる情報を含む、行政業務への影響がない、機密性1情報	1
特定の個人を特定できない情報を含む、行政業務への影響がある、機密性2情報	2
特定の個人を特定できない情報を含む、行政業務への影響がない、機密性1情報	2
個人情報を含まない、行政業務への影響がある、機密性2情報	22
個人情報を含まない、行政業務への影響がない、機密性1情報	8
個人情報・行政への影響を判定するには情報が不足している	1
合計	50

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行1回目

LLM	GPT-4 Turbo

学習セット
無し

プロンプト

#命令
以下の手順に従って回答を生成してください

#手順
1.入力された情報から個人情報を特定し、回答する

個人情報を特定できた場合
2.機密性2情報、機密性3情報の中から該当する分類を選択する
3.分類した根拠を回答する

個人情報を特定できなかった場合
2.機密性1情報、機密性2情報、機密性3情報の中から該当する分類を選択する
3.分類した根拠を回答する

#出力形式
個人情報判定
<特定できた個人情報があれば回答する>

分類結果
個人を特定できた場合
<機密性2情報、機密性3情報の中から選んで回答する>
分類した根拠
<分類した根拠を回答する>

個人を特定できなかった場合
<機密性1情報、機密性2情報、機密性3情報の中から選んで回答する>
分類した根拠
<分類した根拠を回答する>

#前提知識
a."機密性"とは、情報に関して、アクセスを認められた者だけがこれにアクセスできる状態を確保することをいう。
b."機密性３情報"とは、行政事務で取り扱う情報のうち、秘密文書に相当する機密性を要する情報をいう。特定の職員だけが知り得る状態を確保する必要のある情報で秘密文書に相当するものなどが挙げられる。
c."機密性２情報"とは、行政事務で取り扱う情報のうち、秘密文書に相当する機密性は要しないが、その漏えいにより、国民の権利が侵害され又は行政事務の遂行に支障を及ぼすおそれがある情報をいう。職員だけが知り得る状態を確保する必要がある情報（職員のうち特定の職員だけが知り得る状態を確保する必要がある情報を含む）。
d."機密性１情報"とは、機密性２情報又は機密性３情報以外の情報をいう。公表・公開又はそれを前提として作成した情報及び職員以外が知り得ても問題のない情報（公表・公開しても問題のない情報）

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行1回目

LLM	GPT-4 Turbo

学習セット
無し

プロンプト

#命令
以下の手順に従って回答を生成してください

#手順
1.入力された情報から個人情報を特定し、回答する

個人情報を特定できた場合
2.機密性2情報、機密性3情報の中から該当する分類を選択する
3.分類した根拠を回答する

個人情報を特定できなかった場合
2.機密性1情報、機密性2情報、機密性3情報の中から該当する分類を選択する
3.分類した根拠を回答する

以下略

作成者の評価

22通りの情報を入力し、
機密性情報分類が適切に判定できたのは91%。
それなりに高い精度であるものの改善が必要

課題#1

- 個人情報と機密性情報の判定の精度について改善の余地がある

課題#2

- 「機密性情報」だけでなく、「個人情報」を判断した根拠を後で確認できるように、個人情報の考察結果を追加する

F-1.ドキュメントの機密度判定

試行1回目の生成結果

質問例

不正入国者による事件の増加が国民の安全に影響を及ぼす可能性があるため、治安部門が詳細な対策を検討中で、
その内容は特定の警察官達だけが把握している

生成結果例

個人情報判定
特定の個人情報は入力されていません。

分類結果
機密性3情報

分類した根拠
入力された情報は、治安部門が検討中の「不正入国者による事件の増加への対策」の内容に関するものであり、
これは特定の警察官だけが把握しているとされています。この情報は秘密文書に相当し、
特定の職員だけが知り得る状態を確保する必要がある情報であるため、機密性3情報に分類されます。
なぜなら、この情報の漏えいは国民の安全に重大な影響を及ぼす可能性があり、行政事務の遂行にも支障が出るおそれがあるためです。

評価項目

回答の正しさ1:分類が正しいかどうか	0~2点
回答の正しさ2:分類を判定した理由が適切か	0~2点
分かりやすさ:判定理由で参照した箇所が適切か	0~2点

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行2回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト(1/2)

#命令
以下の手順に従って回答を生成してください

#手順
1."個人情報定義"と"個人情報例"を参考にして、入力された情報から個人情報を特定し、回答する
2.入力された情報が公開予定の情報かどうかについて考察を行う

個人情報を特定できた場合
3.機密性2情報を選択する
4.分類した根拠を回答する
5.個人情報定義の何番に該当したのかを回答する

個人情報を特定できなかった場合
3."前提知識"と"機密性2情報詳細"を参考にして、機密性1情報、機密性2情報の中から該当する分類を選択する
4.分類した根拠を回答する
機密性2情報を選択した場合
5."機密性2情報"の何番に該当したのかを回答する

#出力形式
個人情報判定
<特定できた個人情報があれば回答する>
考察結果
<公開予定かどうか考察した結果を回答する>

個人情報を特定できた場合
分類結果<機密性2情報と回答する>
分類した根拠
<分類した根拠を回答する>
個人情報定義
<"個人情報定義"の何番に該当したかを回答する>

個人情報を特定できなかった場合
<機密性1情報、機密性2情報の中から選んで回答する>
分類した根拠
<分類した根拠を回答する>
機密性2情報を選択した場合
機密性2情報詳細
<"機密性2情報詳細"の何番に該当したのかを回答する>

改善方針#1

- 以下の情報をGaiXerに要約させ、プロンプトに組み込む
  - 「行政機関の保有する情報の公開に関する法律 第5条」に記載されている機密性情報の詳細
  - 「個人情報の保護に関する法律 第2条」に記載されている個人情報定義(次頁)

#前提知識
a."機密性"とは、情報に関して、アクセスを認められた者だけがこれにアクセスできる状態を確保することをいう。
b."機密性２情報"とは、行政事務で取り扱う情報のうち、秘密文書に相当する機密性は要しないが、その漏えいにより、国民の権利が侵害され又は行政事務の遂行に支障を及ぼすおそれがある情報をいう。職員だけが知り得る状態を確保する必要がある情報（職員のうち特定の職員だけが知り得る状態を確保する必要がある情報を含む）。
c."機密性１情報"とは、機密性２情報以外の情報をいう。公表・公開又はそれを前提として作成した情報及び職員以外が知り得ても問題のない情報（公表・公開しても問題のない情報）

#機密性2情報詳細
1.特定の個人を識別できる情報、または個人の権利利益を害する恐れがあるが、公開されるべき情報（公務員の職務に関する情報、公にされている情報、人の安全を保護するために必要な情報などを除く）。
2.法人や事業を営む個人に関する情報で、その権利や競争上の地位などを害する恐れがあるもの、または行政機関からの要請によって非公開の条件付きで提供された情報。
3.国の安全や国際関係、交渉上の不利益を招く恐れがある情報。
4.犯罪の予防、捜査、公訴の維持などに支障をきたす恐れがある情報。
5.国や地方公共団体の内部審議や協議に関する情報で、公開により意見交換の自由や意思決定の中立性を損なう、混乱を招く、不公平を生じさせる恐れがあるもの。
6.国や地方公共団体が行う事務や事業に関する情報で、公開することによりその適正な遂行に支障をきたす恐れがあるもの（監査、検査、契約、交渉、争訟、人事管理など）。

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行2回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト(2/2)

#個人情報定義
1.「個人情報」: 生存する個人に関する情報で、特定の個人を識別できるもの。これには、氏名や生年月日などの記述、個人識別符号が含まれる。
2.「個人識別符号」: 特定の個人を識別するための文字、番号、記号などで、政令で定められるもの。
3.「要配慮個人情報」: 人種、信条、社会的身分、病歴などの情報で、差別や不利益を避けるために特別な配慮が必要な個人情報。
4.「本人」: 個人情報によって識別される特定の個人。
5.「仮名加工情報」: 特定の個人を照合せずに識別できないように加工された個人情報。
6.「匿名加工情報」: 特定の個人を識別できないように加工され、元の個人情報を復元できない情報。
7.「個人関連情報」: 個人情報、仮名加工情報、匿名加工情報に該当しない、生存する個人に関する情報。

#個人情報例
1.個人情報:
氏名、住所、電話番号などが書かれた名簿
生年月日や性別が含まれる履歴書
顔写真や指紋データなどの生体情報
2.個人識別符号:
パスポート番号
運転免許証の番号
社会保障番号
3.要配慮個人情報:
民族や宗教に関する情報
性的指向を示すデータ
HIV感染状況などの医療情報
4.本人:
個人情報に含まれる氏名「山田太郎」によって識別される特定の人物
運転免許証の写真から識別される人物
5.仮名加工情報:
氏名がイニシャルに置き換えられた顧客リスト（例: T.Y.）
電話番号の一部が伏せられたデータ（例: 090-XXXX-1234）

改善方針#2

- 個人情報例もGaiXerに生成させ、few-shot-promptとしてプロンプトに組み込む

6.匿名加工情報:
年齢層や性別のみが示された集計データ（個々の人物を特定できない）
個々の顧客の購買履歴を集計し、個人を特定できないようにした統計データ
7.個人関連情報:
住んでいる地域や通っている学校といった情報（ただし、これ単独では特定の個人を識別できない）
インターネット上の匿名の投稿（個人が特定されない限り）

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行2回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト

#命令
以下の手順に従って回答を生成してください

#手順
1."個人情報定義"と"個人情報例"を参考にして、入力された情報から個人情報を特定し、回答する
2.入力された情報が公開予定の情報かどうかについて考察を行う

個人情報を特定できた場合
3.機密性2情報を選択する
4.分類した根拠を回答する
5.個人情報定義の何番に該当したのかを回答する

個人情報を特定できなかった場合
3."前提知識"と"機密性2情報詳細"を参考にして、機密性1情報、機密性2情報の中から該当する分類を選択する
4.分類した根拠を回答する
機密性2情報を選択した場合
5."機密性2情報"の何番に該当したのかを回答する
以下略

評価項目

回答の正しさ1:分類が正しいかどうか	0~2点
回答の正しさ2:分類を判定した理由が適切か	0~2点
分かりやすさ:判定理由で参照した箇所が適切か	0~2点

作成者の評価

50通りの情報を入力し、
機密性情報分類が適切に判定できたのは96%
個人情報定義を基に考えると、個人情報判定結果が誤っている場合があることと、
行政業務への影響有無を考慮できていないことが課題

課題#3

- 個人情報の中には、「特定の個人を識別できる個人情報」と、「特定の個人を識別できない個人情報」があるものの、現状では、特定の個人を識別できる個人情報のみを個人情報と判定してしまっている

課題#4

- 機密性2情報の定義では、行政業務に影響があるかを判定する必要があるがその判定ステップが含まれていない

課題#5

- 機密度判定は正しい/誤っているの2択であるため、部分点の付与は不適当

F-1.ドキュメントの機密度判定

試行2回目の生成結果

質問例

不正入国者による事件の増加が国民の安全に影響を及ぼす可能性があるため、治安部門が詳細な対策を検討中で、
その内容は特定の警察官達だけが把握している

生成結果例

個人情報判定
特定の個人情報は含まれていません。

考察結果
この情報は公開予定ではないと考えられます。この情報は治安部門が検討中の対策の内容であり、
特定の警察官のみが把握しているとされています。したがって、これは機密性が要求される情報であり、公開されるべきではないと思われます。

分類結果
機密性2情報

分類した根拠
この情報は治安部門が検討中の対策に関するものであり、漏えいすると公共の安全に影響を及ぼす可能性があるため、
機密性2情報に分類されます。特に、犯罪の予防や捜査などの支障をきたす恐れがあるため、この情報は公開すべきではなく、
限られた職員だけが知り得る状態を確保する必要があります。

機密性2情報詳細
4.犯罪の予防、捜査、公訴の維持などに支障をきたす恐れがある情報。

評価項目

回答の正しさ1:分類が正しいかどうか	0~2点
回答の正しさ2:分類を判定した理由が適切か	0~2点
分かりやすさ:判定理由で参照した箇所が適切か	0~2点

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行3回目

LLM	GPT-4 Turbo

学習セット

無し

プロンプト(1/2)

#命令
以下の手順に従って回答を生成してください

#手順
1."個人情報定義"と"個人情報例"を参考にし、"情報"から特定の個人を識別できる個人情報を特定して回答する
2."個人情報定義"と"個人情報例"を参考にし、"情報"から特定の個人を識別できない個人情報を特定して回答する
3."前提知識"と"機密性2情報詳細"を参考にし、"情報"を公開することによって行政業務に支障をきたす恐れがあるか推測する
4."情報"が公開予定のものかどうか考察を行う

特定の個人を識別できる個人情報が含まれている場合
5.機密性2情報と回答する
6."個人情報定義"の何番に該当するか回答する
7."機密性2情報詳細"の何番に該当するか回答する

特定の個人を識別できない個人情報が含まれている場合
8."個人情報定義"の何番に該当するか回答する

業務に支障をきたす恐れがある場合
9.機密性2情報と回答する
10.行政業務に支障をきたすとそれがあると推測した根拠を回答する
11."機密性2情報詳細"の何番に該当するか回答する

上記のどれにも該当しない場合
12.機密性1情報と回答する
13.機密性1情報と判定した根拠を回答する

#出力形式
特定の個人を識別できる個人情報
<"情報"から特定の個人を識別できる個人情報を特定して回答する>
特定の個人を識別できない個人情報
<特定の個人を識別できない個人情報を特定して回答する>
公開することによって行政業務に支障をきたす恐れがあるか
<"情報"を公開することによって行政業務に支障をきたす恐れがあるか推測した結果を回答する>
機密性情報分類結果
<機密性情報分類結果を回答する>
個人情報定義
<"個人情報定義"の何番に該当するか回答する>
機密性2情報詳細
<"機密性2情報詳細"の何番に該当するか回答する>
分類した根拠
<分類した根拠を回答する>

改善方針#3

- 個人情報を判定する処理を、「特定の個人を識別できる個人情報を判定するステップ」、「特定の個人を識別できない個人情報を判定するステップ」の2つに分けてそれぞれで判別するようプロンプトを修正する

改善方針#4

- 機密性2情報の定義に基づき、行政業務に影響があるかを判定するステップをプロンプトに追加し、その結果を機密性分類の判定に使うようプロンプトを修正する

F-1.ドキュメントの機密度判定

プロンプトの発展過程　試行3回目

LLM	GPT-4 Turbo

学習セット
無し

プロンプト(2/2)

#前提知識
a."機密性"とは、情報に関して、アクセスを認められた者だけがこれにアクセスできる状態を確保することをいう。
b."機密性２情報"とは、行政事務で取り扱う情報のうち、秘密文書に相当する機密性は要しないが、その漏えいにより、国民の権利が侵害され又は行政事務の遂行に支障を及ぼすおそれがある情報をいう。職員だけが知り得る状態を確保する必要がある情報（職員のうち特定の職員だけが知り得る状態を確保する必要がある情報を含む）。
c."機密性１情報"とは、機密性２情報以外の情報をいう。公表・公開又はそれを前提として作成した情報及び職員以外が知り得ても問題のない情報（公表・公開しても問題のない情報）

#機密性2情報詳細
1.特定の個人を識別できる情報、または個人の権利利益を害する恐れがあるが、公開されるべき情報（公務員の職務に関する情報、公にされている情報、人の安全を保護するために必要な情報などを除く）。
2.法人や事業を営む個人に関する情報で、その権利や競争上の地位などを害する恐れがあるもの、または行政機関からの要請によって非公開の条件付きで提供された情報。
3.国の安全や国際関係、交渉上の不利益を招く恐れがある情報。
4.犯罪の予防、捜査、公訴の維持などに支障をきたす恐れがある情報。
5.国や地方公共団体の内部審議や協議に関する情報で、公開により意見交換の自由や意思決定の中立性を損なう、混乱を招く、不公平を生じさせる恐れがあるもの。
6.国や地方公共団体が行う事務や事業に関する情報で、公開することによりその適正な遂行に支障をきたす恐れがあるもの（監査、検査、契約、交渉、争訟、人事管理など）。

#個人情報定義
1.「個人情報」: 生存する個人に関する情報で、特定の個人を識別できるもの。これには、氏名や生年月日などの記述、個人識別符号が含まれる。
2.「個人識別符号」: 特定の個人を識別するための文字、番号、記号などで、政令で定められるもの。
3.「要配慮個人情報」: 人種、信条、社会的身分、病歴などの情報で、差別や不利益を避けるために特別な配慮が必要な個人情報。
4.「本人」: 個人情報によって識別される特定の個人。
5.「仮名加工情報」: 特定の個人を照合せずに識別できないように加工された個人情報。
6.「匿名加工情報」: 特定の個人を識別できないように加工され、元の個人情報を復元できない情報。
7.「個人関連情報」: 個人情報、仮名加工情報、匿名加工情報に該当しない、生存する個人に関する情報。

#個人情報例
1.個人情報:
氏名、住所、電話番号などが書かれた名簿
生年月日や性別が含まれる履歴書
顔写真や指紋データなどの生体情報
2.個人識別符号:
パスポート番号
運転免許証の番号
社会保障番号
3.要配慮個人情報:
民族や宗教に関する情報
性的指向を示すデータ
HIV感染状況などの医療情報
4.本人:
個人情報に含まれる氏名「山田太郎」によって識別される特定の人物
運転免許証の写真から識別される人物
5.仮名加工情報:
氏名がイニシャルに置き換えられた顧客リスト（例: T.Y.）
電話番号の一部が伏せられたデータ（例: 090-XXXX-1234）
6.匿名加工情報:
年齢層や性別のみが示された集計データ（個々の人物を特定できない）
個々の顧客の購買履歴を集計し、個人を特定できないようにした統計データ
7.個人関連情報:
住んでいる地域や通っている学校といった情報（ただし、これ単独では特定の個人を識別できない）
インターネット上の匿名の投稿（個人が特定されない限り）

試行3回目の生成結果

質問例

FIXERのホームページにアクセスした人のIPアドレス

生成結果例

特定の個人を識別できる個人情報
IPアドレス

特定の個人を識別できない個人情報
なし

公開することによって行政業務に支障をきたす恐れがあるか
IPアドレスの公開は、そのIPアドレスを使用している人物のプライバシーを侵害したり、セキュリティ上のリスクを高めたりする可能性があるため、行政業務に支障をきたす恐れがあります。

機密性情報分類結果
機密性2情報

個人情報定義
1

機密性2情報詳細
1

分類した根拠
IPアドレスは、特定の個人を識別できる情報に該当し（個人情報定義1）、その漏洩によって個人の権利が侵害される恐れがあるため
（機密性2情報詳細1）、機密性2情報と分類しました。
また、IPアドレスはインターネット上の特定のデバイスを識別するためのアドレスであり、通常、
これだけで個人を特定することは難しいですが、他の情報と組み合わせることで、特定の個人を識別できる場合があります。
したがって、これを公開することは行政業務に支障をきたす恐れがあると判断しました。

試行3回目の評価項目

改善方針#5

- **機密度判定は正しい/誤っているの2択であるため、評価点数も0点か1点の2択とする**

評価基準	要件定義	評価(点数)
回答の正しさ1:個人情報判定、業務に支障があるかの判定が適切か	どちらも正しく判定できている	1
	どちらか片方しか正しく判定できていない	0
	両方とも間違っている	0
回答の正しさ2:機密性情報分類が適切か	適切に分類できている	1
	分類が不適切である	0
回答の正しさ3:機密性情報分類した根拠が適切か	根拠が全て適切である	1
	根拠は適切だが、個人情報定義や機密性2情報詳細が間違っている	0
	根拠が不適切である	0

F-1.ドキュメントの機密度判定

試行3回目の評価結果

- ▶ 機密性情報の分類得点率は98%であり、100%の精度が要求される入力防止の機能に活用することは現時点では困難
- ▶ 一方で、既にLLMへ入力した情報の中に機密性情報が含まれていたのかを確認する機能においては、十分に活用可能な水準

ポイント1)

機密性情報の分類の得点率は98%と高かったものの、1問で誤りがあったため、100%が要求される入力防止の機能に活用することは現時点では難しい
一方で、既にLLMへ入力した情報の中に機密性情報が含まれていたのかを確認する機能においては、十分に活用可能な水準といえる

ポイント2)

分類した根拠に関する得点率が80%とやや低いものの、
「根拠は適切だが、個人情報定義や機密性2情報詳細が間違っている」場合を誤りとしたことによるためであり、根拠自体は適切な割合は94%であった

評価基準	妥当	不適当	質問数	得点率
回答の正しさ1: 個人情報判定、業務に支障があるかの判定が適切か	45	5	50	90%
回答の正しさ2: 機密性情報分類が適切か	49	1	50	98%
回答の正しさ3: 機密性情報分類した根拠が適切か	40	10	50	80%
合計点	134	0	150	89%

総括

結論　：　機密性分類２判定は100%精度での回答は困難。よって防止策的な利用は非推奨。
　　但し、事後的にログなどを監査する目的でのバッチ処理は有効と考えられる。

考察　：

・誤り判定が出やすいケース
- 特定の個人を識別できない個人情報が含まれている場合に個人情報の判定が誤ることがあり、その影響で判定が誤りやすい（LLMは個人を特定できる情報のみを個人情報と判定する傾向があるため）
- 住所が含まれている場合に個人情報判定が揺れる（その住所が個人宅なのか公共施設なのか判別できないため）
- 状況や情勢次第で機密情報判定が変わるような場合に回答が揺れやすい（ LLMは最新情報を学習していないため）

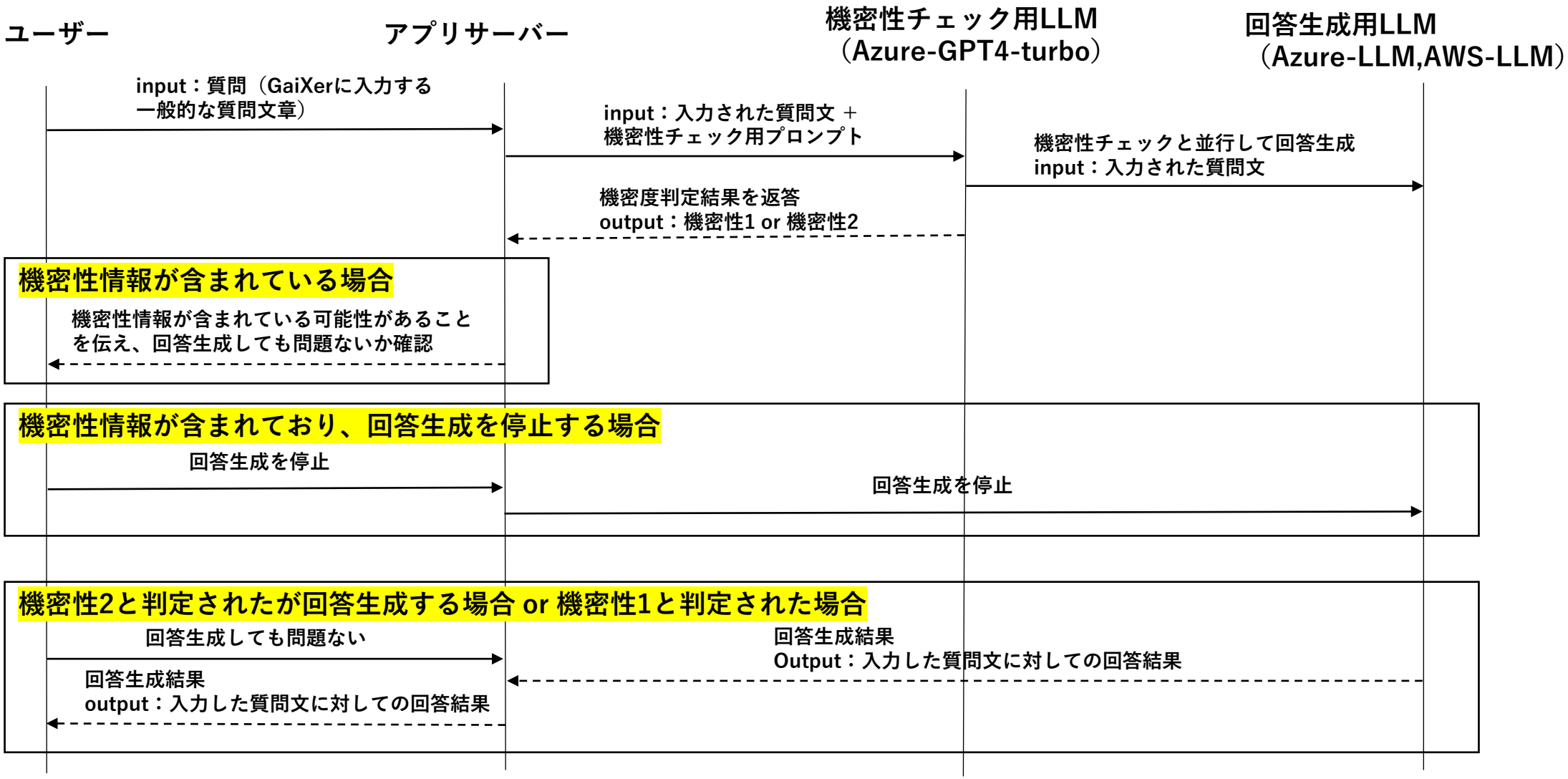
・品質評価の注意点
- 100%の精度が求められるユースケースであるが、機密度判定の結果だけではなくその根拠にも言及する事が重要の為、機密度判定過程をプロンプトに含め、それらを評価する事が重要。

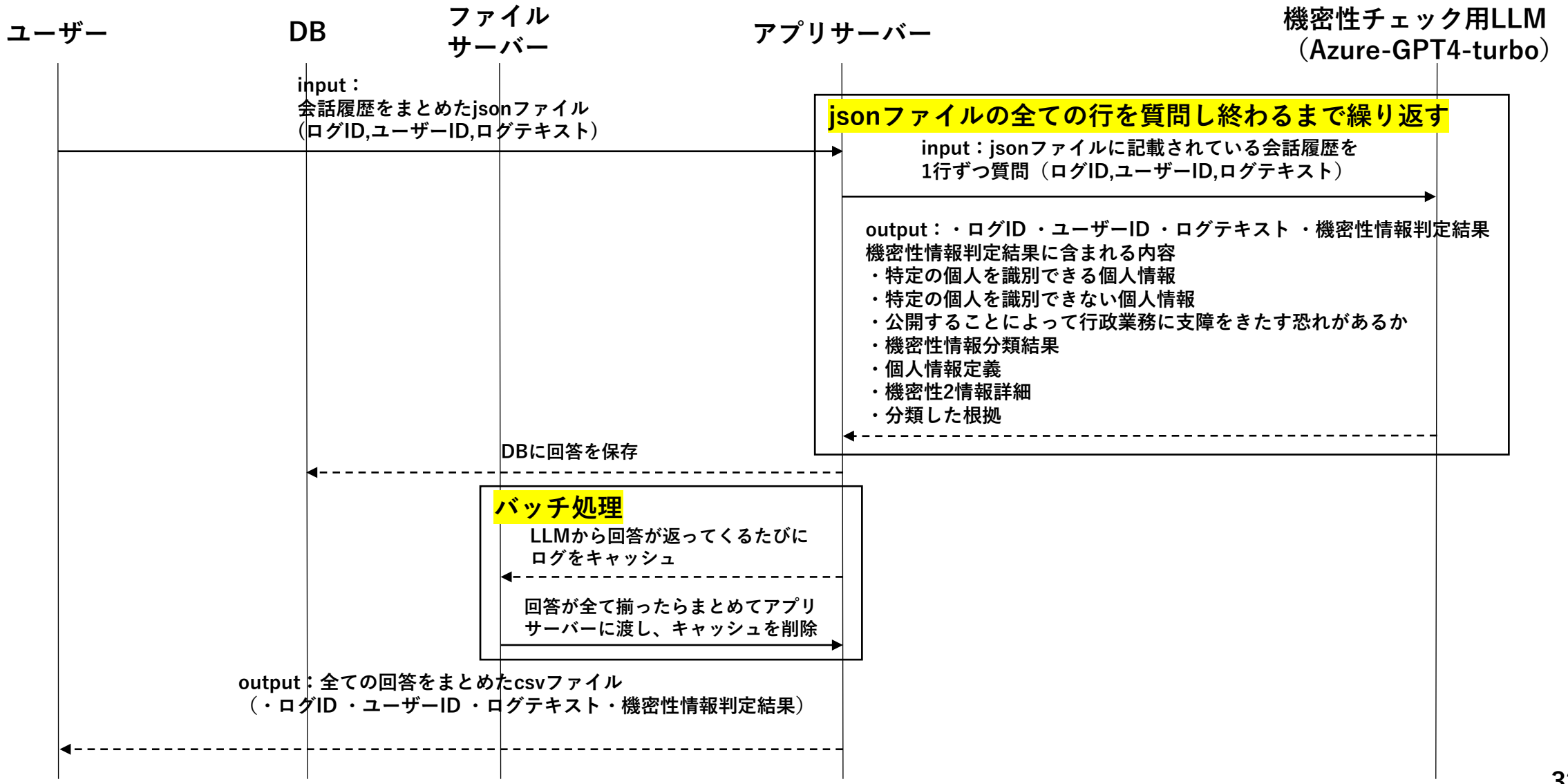
実装方法

実装レベル	実装形式	コスト構造	利点	課題
検証	JSON形式化されたFAQをRAGで実装し、個別質問で受けた問い合わせをFAQとして一般化し、新たなQAを追加。個別質問を受けた際には、生成AIでFAQから回答を参照して正確に抽出	SaaS利用料(トークン量込)	本利用に向けた本格検証が可能。プロンプトの組織管理/共有が可能なため、業務利用も可能。	FAQから忠実に回答する正確性に対応できるLLMの選定(GPT-4-TurboやClaude 2.1等では一定精度までが限界だが、今後のLLM性能に応じてその限りでない)
将来像	前述のシーケンス図に従い、Slack環境にLLM、Chatbot、FAQ DBを実装し、個別質問への回答と共に、FAQナレッジベースの拡充ができるサービスとして実装	・システム実装コスト ・チャットボット運用コスト ・トークン利用料 ・インフラ運用コスト	チャットボットとして搭載の為決められた回答のみを出力。生成AIは質問の受け手として機能の為、ハルシネーションリスクがヘッジできる。	FAQのレビューや回答内容の確認ステップなど、全自動運用は難しいため、将来のLLM性能の向上に応じたLLM対応領域の拡大が求められる

シーケンス図

■機密性情報が含まれている場合は回答生成用LLMへの入力を未然に防ぐケース(検証環境で実装済み)



シーケンス図

■既に回答生成用LLMに入力した情報の中に機密性情報が含まれていないかを確認するケース