クローズド環境下での生成AIのオンプレミス利用環境の構築・運用に係る調査研究事業

検証報告書

令和 7 年 3 月 21 日

株式会社 RUTILEA

株式会社コア

目次

報告書番号	1
項目	背景と目的

近年、生成 AI 技術の発展により、業務の合理化・効率化が進んでいる。警察庁においても文書作成、翻訳、プログラムコード生成等、様々な業務に生成 AI を活用することで、業務の合理化・効率化が期待されている。

しかし、警察庁が保有する情報は極めて機密性が高く、外部ネットワークと接続された環境での生成 AI 利用にはセキュリティ上のリスクが伴う。そのため、外部インターネットに一切接続しないクローズド環境下で生成 AI を活用する必要がある。

本事業の目的は、クローズド環境下での生成 AI のオンプレミス利用環境を構築し、警察庁が現にクローズド環境下で実施している各種業務の合理化・効率化等に係る検証、評価及び課題抽出を行い、それら業務における生成 AI の導入・活用の有用性を確認するとともに、当該環境の継続的な利用において必要となる機能・性能等の精査を行うことによって生成 AI のオンプレミス利用環境の本格導入につなげ、警察庁における業務全般の合理化・効率化に寄与することである。

報告書番号	2
項目	調査研究の方法

本事業では、警察庁のクローズド環境に生成 AI のオンプレミス利用環境を構築し、業務適用の可能性を検証する。

まず、2024 年 3 月時点で最高性能の GPU である NVIDIA H100 SXM を計 36 枚導入し、高い計算能力を確保することで、大規模かつ高性能な AI モデルの活用を可能にする。また、Infiniband スイッチを導入し、GPU サーバ間の高速通信を実現し、大規模な AI モデルの学習も実現できる環境を構築する。これらのハードウェアを早期に導入し、生成 AI 利用環境を使って利用者が評価検証する期間を確保する。

ソフトウェア面においては、最新のオープンモデルを速やかに実装・評価し、最適な性能を確保するとともに、AI モデルの性能向上に合わせた柔軟な運用を可能にするため、容易に AI モデルを入替えできるソフトウェアアーキテクチャを構築する。

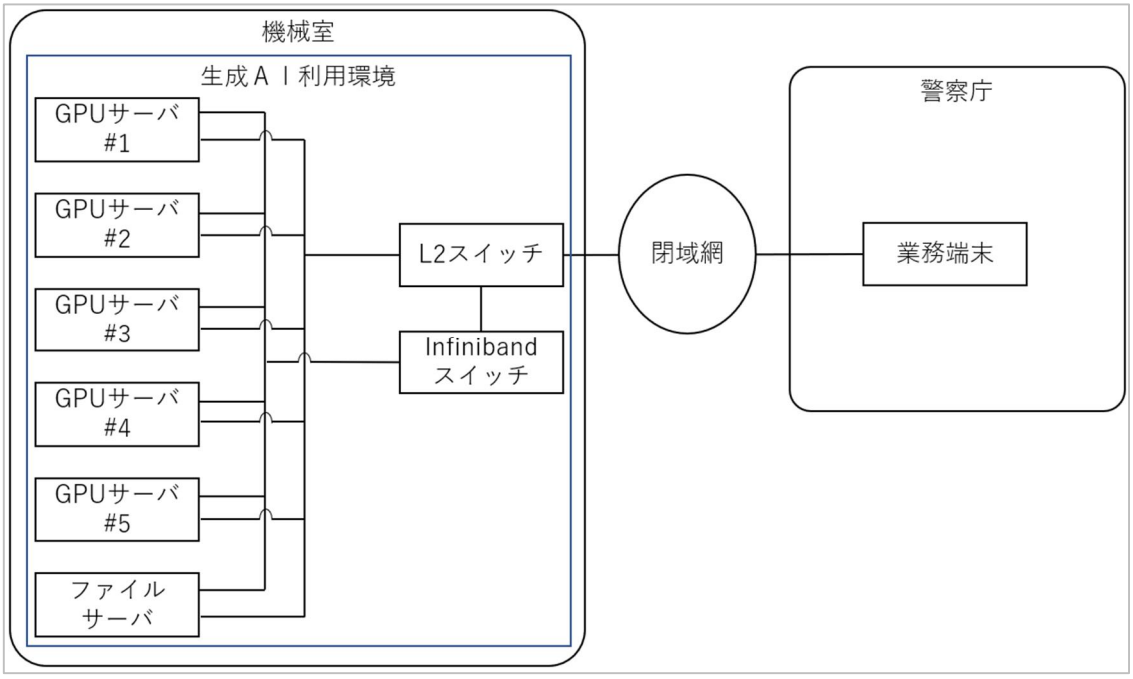
また、ワークショップを開催し、現場のニーズをヒアリングしながら生成 AI のプロトタイプを用いた評価を実施することにより、利用者による評価検証を実施し、課題を抽出するとともに、モデルやワークフローの最適化を繰り返し行う。これにより、業務適用に必要な要件を明確化し、フィードバックをもとに改善を重ねる。また、来年度以降、警察庁内での継続的な運用に必要なスキルを習得できるようサポートし、警察庁の業務全般の合理化・効率化への貢献を目指す。

報告書番号	3
項目	ハードウェア

クローズド環境において、生成 AI を活用するために大規模な AI モデルの運用・学習が可能な基盤を構築した。

- GPU サーバとして HGX-H100 を 1 台、DGX-H100 を 4 台導入し、合計 36 枚の NVIDIA H100 SXM による高い計算能力を確保した。
- 将来的に大規模な AI モデルの学習等を可能とするため、Infiniband スイッチを導入し、サーバ間の高速通信を実現した。
- ストレージ容量 1PB のファイルサーバを導入し、大量のデータを保存・活用できる設計とした。

図表 1:ハードウェア全体構成図

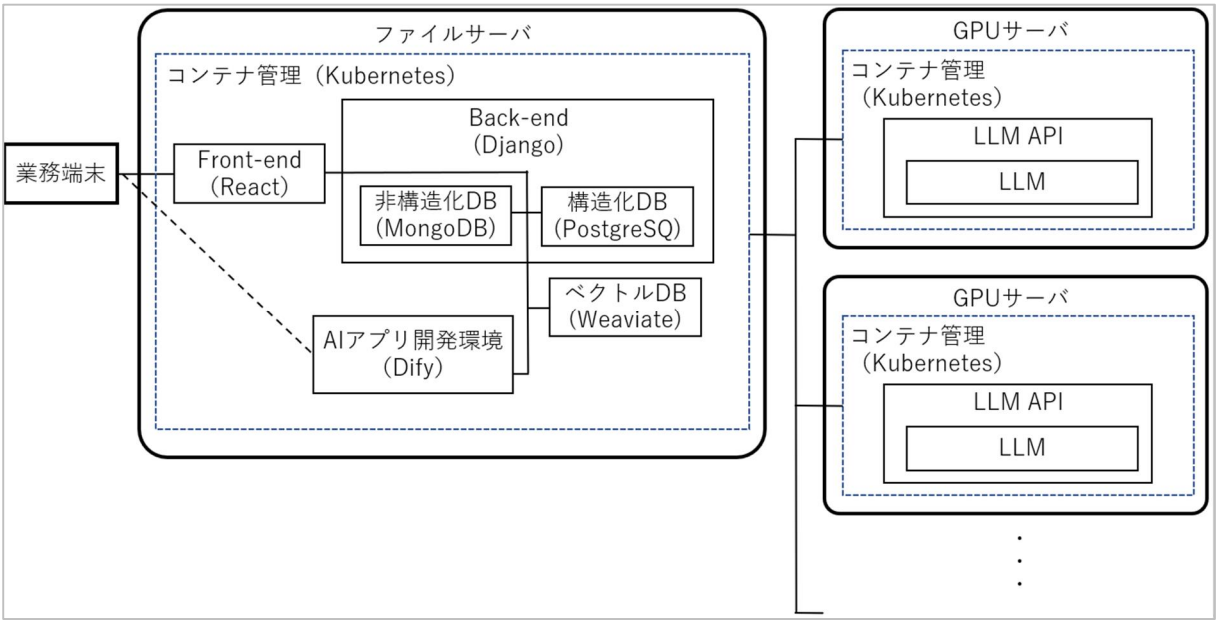


報告書番号	4
項目	ソフトウェア、システム

生成 AI の活用において、最新のオープンモデルを実装し、継続的なモデル更新が可能なソフトウェア環境を構築した。

- 外部インターネットに接続せずに生成 AI を利用可能なシステムを構築した。
- Llama3.3（70B）等の最新のオープンモデルを実装した。
- AI モデルは容易に入替可能なアーキテクチャを採用し、今後の技術進化にも対応できる柔軟な運用を実現した。
- Dify を活用してワークフローを実装し、業務適用を容易にした。
- コンテナ技術を活用し、システム全体をコンテナ化することで、環境の柔軟性を向上させた。コンテナオーケストレーションには Kubernetes を導入し、運用の効率化とスケーラビリティを確保した。

図表 2:ソフトウェア全体構成図



※オープンソースのライブラリ、フレームワーク、プラットフォーム等を使用

報告書番号	5
項目	実装機能

生成 AI 利用環境に実装した主要な機能と特徴について、図表 3 に示す。これらの機能は第 7 項記載のワークショップを実施する中で改善点や要望をヒアリングし、調整を繰り返したものである。

図表 3:実装した主要機能と特徴一覧

機能	特徴
チャット	・jtd、pdf、docx、xlsx、pptx 等の文書ファイル形式及び jpg 等の画像ファイル形式の入力に対応 ・日本語で最大 6 万字の入力が可能 ・AI モデルを選択可能
ナレッジデータ活用	・指定したナレッジデータを参照して、根拠を明示した回答を生成可能 ・根拠の出典となるファイルのダウンロードが可能
音声文字起こし	・wav 等の音声ファイル形式及び mp4 等の動画音声ファイル形式に対応 ・補助機能として、文字起こしテキストの整形・話者分離機能及び質問回答部分抽出機能を実装
翻訳	・指示文の入力を必要とせず翻訳前及び翻訳後の言語を指定するだけで翻訳可能 ・英語だけでなく、中国語、フランス語、ドイツ語等の複数言語に対応
プログラミング	・Python、JavaScript、C++等の多様なプログラミング言語に対応 ・生成したコードを安全な仮想環境内で実行可能
データ分析	・csv、txt、pdf 等の様々な形式のデータに対応 ・RAG 活用により非構造データの分析が可能

報告書番号	6
項目	検証評価の概要

本事業では、生成AIの急速な進化に対応するため、アジャイルなプロジェクト管理手法を採用した。実際に動作するプロトタイプをたたき台として関係者との議論を重ね、改善を進める方法で検証評価を行った。

具体的には、後述するワークショップにおいて、生成 AI の出力を確認しながら、改善点や要望をヒアリングし、得られたフィードバックを基にワークフロー等の機能を調整し、再度ワークショップで検証するというサイクルを繰り返した。このプロセスにより、現場のニーズや要望を迅速に反映するとともに、参加者の生成AIに対する理解を深めながら、警察庁の業務に適した形で生成AI利用環境を調整することが可能となった。

また、本事業の効果測定と客観的評価のため、以下の 3 つの検証評価を実施した。

- **ユースケース評価:**
  ワークショップを開催し、ユースケースごとに生成 AI 出力の質、正確性、有用性等の評価を実施した。
- **アンケート評価:**
  ワークショップ終了後に参加者を対象とした生成 AI 利用に関するアンケート調査を実施し、生成 AI の業務適用に関する意見、適用事例、課題等を収集した。
- **AI モデル及び RAG 性能評価:**
  警察庁から提供された警察業務に関連する択一問題を活用し、本事業で実装したベースとなる AI モデル及び RAG の性能評価を実施した。

これらを通じて、生成 AI 利用環境に実装した機能の警察庁の実務に対する有用性を評価し、課題の抽出を行った。各評価のまとめは、P20 の報告書番号9:評価結果にまとめを記載。

報告書番号	7
項目	ワークショップの概要

警察庁内での生成 AI の活用可能性を検証するため、それぞれ 5 名程度で構成される 6 つのチームを編成し、ワークショップを実施した。そのうち、5 つのチームでは警察庁が選定したユースケースの評価・検証を行い、残り 1 つのチームは、次年度以降の運用も見据え、生成 AI 技術への理解を深めるテクニカルチームとして、運用に必要な知識・スキルの習得に取り組んだ。

ワークショップは主に 2024 年 7 月下旬から 12 月下旬にかけて、約 3 週間おきに実施し、各チームがユースケースの検証評価を進めた。

ワークショップ終了後には、参加者によるユースケースごとの生成 AI 評価及び生成 AI 利用に関するアンケートを実施し、生成 AI の業務適用に関する意見、生成 AI を活用した事例や課題を収集した。

図表 4:ワークショップで取り組んだユースケース及びワークショップ開催回数

ユースケース	チーム	開催回数
国会答弁案等の作成	評価検証チーム①	8
都道府県警察からの質疑応答に係る回答案作成		
訓示案の作成		
伺い文等の案文作成	評価検証チーム②	8
申報の検索		
用例検索		
音声記録の文字起こし、議事録作成	評価検証チーム③	8
翻訳 （日本語から英語、英語から日本語、マイナー言語から日本語）		
匿名・流動型犯罪グループ対策に係るデータ分析	評価検証チーム④	9
プログラミングコードの生成、分析、ミス等の発見・修正等	評価検証チーム⑤	8
生成 AI 改善プロセスに係る技術の習得	テクニカルチーム	9

報告書番号	8
項目	各ユースケースの実施概要

本項では、ワークショップにおいて取り扱った各ユースケースの検証内容及び評価結果の概要についてユースケースごとに記載する。

1.ユースケース「国会答弁案等の作成」の実施内容

本ユースケースでは、プロンプトエンジニアリングや、ファインチューニングの一種である LoRA を活用し、国会答弁案等の作成における生成 AI の活用可能性について検証・評価を行った。

(1) 検証方法

Few-shot learning、又は約 1,100 件の過去国会における質問-答弁データセットを用いて LoRA により学習したモデルを使用して国会答弁案作成を検証した。

(2) 検証結果

今回の検証では、国会答弁独特の言い回しの習得手法として、LoRA を実施したモデル使用よりも Few-shot learning が適しているという結果が得られた。LoRA を実施したモデルは、「まず」「次に」等の国会答弁案でよく見られる言い回しで出力される場面が増えたものの、求める言い回しの出力の再現性は十分に得られなかった。

また、作成した答弁案に対して、関連資料等とともに追加の修正指示を与えることによって答弁案の内容を補強できることも確認した。

(3) 評価結果

生成 AI を用いて作成した答弁案と過去に作成した実際の答弁案（正解データ）を比較し、採点した。5 人による評価の結果、採点平均は 3.2（5 点満点）となった。これは「内容の方向性は良いが修正はそれなりに必要（概ね 50%）」というレベルに相当する。

主な課題としては以下の点が挙げられている。

- 初回生成された内容への依存度が高く、その後の修正指示による改善が困難
- 国会答弁特有の言い回しや文末表現の適切な使用ができていない

現段階では初期のたたき台作成と生成 AI との対話による修正プロセスの活用が期待できる一方で、国会答弁案に適した内容や表現の調整といった点は、人による修正作業が必要であるという結果となった。

2.ユースケース「都道府県警察からの質疑応答に係る回答案作成」

本ユースケースでは、モデルケースとして「古物営業法に関する問い合わせに対する回答案作成」を取り上げ、RAG を活用し、質疑応答に係る回答案作成における生成 AI の活用可能性について検証・評価を行った。

(1) 検証方法

RAG を用いて、入力した質問に対して適切な回答案及びエビデンスが出力されるかに主眼を置いて検証を実施した。ナレッジデータは警察庁が準備した古物営業法に関する Q&A 集、関連する通達、質問主意書等を使用した。

(2) 検証結果

本検証では RAG の 2 段階検索フローの有効性が確認できた。まず Q&A 集を優先検索することで、既存の確立された回答を活用できた。さらに、Q&A 集に該当する回答が見つからない場合に、二次検索を行い、その情報を参照して回答案を生成することによって検索精度の向上が可能であることが示された。

(3) 評価結果

出力された回答案及びエビデンスの正確性を採点した。5 人による評価の結果、エビデンスの正しさについては採点平均

が 3.8(5 点満点)となり、概ね正確な引用ができていると評価された。回答案の正しさについては採点平均が 3.8(5 点満点)であり、引用箇所の内容を概ね正確に反映できていることが示されている。回答の表現の適切さについては採点平均が 3.6 であり、修正がある程度必要という評価となっている。

主な課題としては以下の点が挙げられている。

- 正しく引用できている場合とできていない場合の差が大きい
- 回答生成時の参考資料の引用箇所が分かりにくい

簡単な問合せについては、生成 AI による資料の引用と回答案作成により作業時間の短縮が期待できること、最終的な人手によるチェックは必要であるが、たたき台作成の効率化に寄与することが確認された。

3.ユースケース「訓示案の作成」

本ユースケースでは、プロンプトエンジニアリングを活用し、訓示案の作成における生成 AI の活用可能性について検証・評価を行った。

(1) 検証方法

訓示案の関連情報となる通達や、訓示案の骨子を用いて訓示案作成を検証した。複数の AI モデルによる生成結果の比較を行うとともに、Few-shot learning による出力の言い回しや文章構成の矯正を検証した。

(2) 検証結果

検討初期段階では、AI モデルの入力可能な文字数が少なかったことから内容を補足する追加情報や Few-shot learning 用の例示を活用することが難しかったが、Llama3.1(70B)では入力文字数上限が約 6 万字に拡大されたことから、より詳細な指示や参照例を含めた入力が可能になった。

(3) 評価結果

生成 AI で出力した訓示案を過去に作成した実際の訓示案(正解データ)と比較し、採点した。5 人による評価の結果、採点平均は 2.8(5 点満点)となった。これは「内容の方向性は良いが修正はそれなりに必要(概ね 50%)」と「内容が不適切で大幅な修正が必要」の中間に位置する評価である。

主な課題としては以下の点が挙げられている。

- 指示内容の適切な反映が不完全。特に初回生成内容への依存度が高い。
- 修正には大量の詳細な指示が必要

骨子が大まかなほど修正作業が増大すること、LLM への詳細な指示による修正より、早期の手作業への移行が効率的であることが指摘された。

4.ユースケース「伺い文等の案文作成」

本ユースケースでは、プロンプトエンジニアリングを活用し、伺い文等の案文作成における生成 AI の活用可能性について検証・評価を行った。

(1) 検証方法

実際の依頼メールと伺い文案を用いて対話形式での作成支援を検証した。複数の AI モデルによる生成結果の比較を行うとともに、LLM と対話形式で Chain-of-Thought (CoT)や Few-shot learning を活用して伺い文案を作成することを検証した。

(2) 検証結果

対話形式での伺い文作成支援を複数の AI モデルで比較し、Llama3.1 等の最新モデルで対話性能の向上を確認した。CoT 手法を用いた出力改善では、現状を点数化し改善点を列挙させる指示を追加することで効果的な修正が可能になった。また、Few-shot learning の活用により伺い文特有の形式や表現を反映した文案が作成できた。

(3) 評価結果

生成 AI で出力した伺い文案を正解データ(過去事例)と比較し、採点した。4 人による評価の結果、採点平均は 2.5(5 点満点)となった。これは「内容の方向性は良いが修正はそれなりに必要(概ね 50%)」と「内容が不適切で大幅な修正が必要」の間に位置する評価である。

主な課題としては以下の点が挙げられている。

- 重複箇所が多く、文書全体としての読みやすさに欠ける
- 概要と個別項目の記載バランスが不適切等、文書スタイルの対応が困難
- 修正指示による望ましい改善が困難

伺い文案作成については、大幅な修正が必要であるなど、現時点では実務で使用できる結果とはならなかった。生成 AI のみで完璧な伺い文案を作成しようとした場合、複数回のやり取りと試行錯誤による修正が必要となるため、「最初から手動で作成する方が時間効率的なケースもある」というコメントがあった。概要の項目を生成 AI に作らせ、その他の項目は担当者が作成するなど、一部の項目を効率的に作成する場合や断片的な追加指示によりブラッシュアップを行う場合には効果的であるという意見もあった。

5.ユースケース「申報の検索」

本ユースケースでは、申報の検索プロセスに生成 AI を活用し、関心に応じた効率的な検索機能の実現を目的として、表記揺れや複数ファイル検索に対応する RAG の有効性を検証した。

(1) 検証方法

RAG を活用した検索手法を検討し、検索精度の向上と根拠提示機能の改善を図った。ナレッジデータは警察庁が準備した申報ファイル約 50 件を使用した。

(2) 検証結果

RAG を活用した申報検索では、文書の内容に基づく検索が可能となり、表記揺れへの対応が実現した。さらに、根拠となる文章の表示機能や根拠元ファイルのダウンロード機能を実装し、生成 AI が検索した結果を人が容易に確認できる仕組みを構築した。

また、検索したファイルの内容を要約して出力することでユーザはファイルを開くことなく内容を把握できるようになった。

(3) 評価結果

出力された回答案及びエビデンスの正確性を採点した。4 人による評価の結果、引用箇所の正しさについては採点平均が 4.3 となり、複数の引用箇所が正しく、一部漏れはあるものの高い精度を示している。回答案の正しさについては採点平均が 4.3 であり、引用箇所の内容をほぼ正しく反映していると評価された。回答の表現の適切さについては採点平均が 4.5 であり、修正が少し必要と評価された。

主な課題としては以下の点が挙げられている。

- 抽出情報の欠落をユーザが判別することが困難
- 回答内容と参照文献の整合性が取れていない事例の存在

複数の申報に目を通してまとめるプロセスは時間削減できる可能性があること、既に存在が分かっている特定の申報を探し出す場合には非常に有用であること、糸口を掴むために活用できることが指摘されている。本検索フローは申報以外のファイル検索にも応用可能であり、横展開が期待できる。

6.ユースケース「用例検索」

法令や文書から特定の言い回しや表現を効率的に検索するため、RAG を活用した用例検索を検証した。

(1) 検証方法

ナレッジデータは e-Gov 法令検索からダウンロードした法令データを使用した。これらのデータは検証環境に保存され、事前にデータベース化されている。RAG を活用した用例検索を検証するとともに、正規表現や検索ワードを LLM に生成させ、ファイルシステム検索と組み合わせる手法も並行して検証した。

(2) 検証結果

法令を検証環境でデータベース化し、用例検索を行うに当たり、特定の表現パターンを検索するための正規表現を生成 AI が自動生成し、ファイルシステム検索と組み合わせる手法を検証した。その結果、特に「第〇条に規定する～に限る」といった定型表現の検索において有効性が確認された。

また、文言の表現揺れを許容する検索の実現に向け、LLM を用いて検索ワードを拡張することで、文言の完全一致に加え、部分的に可変な表現の検索が可能となり、用例検索の適用範囲が拡大した。一方で当初想定していた RAG を活用した検索手法では、検索精度が低いという結果が得られた。

(3) 評価結果

正規表現を生成 AI に作成させファイルシステム検索と組み合わせて法令データから用例を検索した結果を採点した。4 人による評価の結果、採点平均は 4.5（5 点満点）となった。これは「表示された結果が全て正しい」に近い評価である。

課題としては、例えば、「「あって」と「あつて」等を同一視できていない」点が指摘されている。

本ユースケースで適用した範囲の用例検索においては、生成 AI の活用による効率化の効果が高く、実務での活用が期待できる。当初想定していた RAG を活用した検索手法では、検索精度が低いという結果が得られた原因としては、埋め込みモデルの日本語対応が十分でないことに加え、法令特有の表現や構造に適した分割が適切に行われていない可能性が考えられる。

さらに、本ユースケースと生成 AI による意味認識の相性の問題も明らかになった。具体的には、文章構造が一致するものを検索したいという本ユースケースに対して、AI では意味的な類似性に基づく検索を行うため、むしろ構造を省略してしまう傾向があった。このため、本検証ではユースケースの要件に即した処理方法に変更することで対応した。

7.ユースケース「音声記録の文字起こし、議事録作成」

会議音声の文字起こしと議事録作成業務の効率化のため、音声認識による高速な文字起こし、話者分離、整形、要約機能を段階的に組み合わせることで、議事録作成プロセスの効率化を検証した。

(1) 検証方法

音声データの文字起こしから議事録作成までの作業工程において、文字起こしデータの整形や話者分離等の基本機能を実装し、段階的な検証と改良を繰り返した。具体的には、まず音声データから文字起こしデータを作成し、構造化されたテキストへの変換プロセスを検討した。次に、用途に応じた様々な議事録形式への変換プロセスを検証し、ワークショップ参加者からのフィードバックに基づいて機能改善を実施した。

(2) 検証結果

文字起こしのほかに議事録作成の補助機能として、整形・話者分離機能や質問・回答抽出機能を実装し、人が一定程度手直しする前提の議事録案が短時間で作成することができた。

(3) 評価結果

音声ファイルを文字起こしし、補助機能も活用し、議事録案を作成し、正解データ(過去事例)と比較し採点した。4 人による評価の結果、採点平均は 3.3(5 点満点)となった。これは「出力されたテキストの修正作業に一定程度の時間が必要」というレベルに相当する。

主な課題としては以下の点が挙げられている。

- 発言の一部が文字起こしされない
- 固有名称の認識精度が不十分

議事録作成の基礎資料としては一定の有用性があり、整形機能は文書の可読性向上に寄与するものの、元音声との整合性確認作業が必要で、追加的な工数が発生することが指摘されている。音声記録の文字起こし、議事録作成においては生成 AI を活用することで一定の効率化が期待できるが、認識精度の向上が求められるという結論に至っている。

本評価後の対応として、音声データを文字起こしする際に、発話内容の一部が抜け落ちるという課題に対して、評価時に使用していた音声認識モデルを新しいモデルに変更することで、文字起こし時の欠落を防止することができることを確認した。

8.ユースケース「翻訳（日本語から英語、英語から日本語、マイナー言語から日本語）」

翻訳業務に生成 AI を活用し、「日本語から英語」、「英語から日本語」、「マイナー言語から日本語」の複数パターンの翻訳の効率化と精度向上を目指す検証を実施した。

(1) 検証方法

プロンプトによる自然な言い回しへの改善試行、マイナー言語の翻訳に強いと言われる AI モデルと検証時最新の AI モデルである Llama3.1 の比較等を実施した。

(2) 検証結果

生成 AI を活用した翻訳において、検証時最新の AI モデルである Llama3.1 を使用することで、一般的な英語だけでなくマイナー言語についても高い精度の翻訳が期待される結果となった。LLM と対話形式で翻訳後の文章に対して修正指示を出すことで、適切な表現への調整が可能であることを確認した。

(3) 評価結果

翻訳の 3 つのパターン（「日本語から英語」、「英語から日本語」、「マイナー言語から日本語」）それぞれに対して、翻訳前の文章を入力し、翻訳後の文章を正解データ（過去事例）と比較し採点した。4 人による評価の結果、採点平均は日本語➡英語が 3.5 点、英語➡日本語が 3.8 点、マイナー言語➡日本語が 4.0 点（いずれも 5 点満点）となった。

主な課題としては以下の点が挙げられている。

- 全文を一括で翻訳した場合に部分的な翻訳漏れが発生する
- マイナー言語については言語によって翻訳精度に差がある

実務上の有用性としては、文書内容の概要把握には有効に活用可能であること、個別の部分を辞書や翻訳ソフトで翻訳するよりも時間が大幅に短縮できる可能性があることが挙げられる。また、マイナー言語の翻訳は、まず何語であるかを理解する必要があるが、そのプロセスを削減できることが確認された。ただし、正式な資料作成時には翻訳漏れの確認作業が必要であり、翻訳精度の高い言語に限定して使用することで効率化が期待できるという結論に至っている。

全体として、生成 AI を活用した翻訳は一定の効率化をもたらすものの、精度向上のためにはセンテンス単位での処理や翻訳結果の確認作業が必要であることが示されている。翻訳した文章の比較については、ワークショップ以降に翻訳前後の比較テンプレート機能を実装し、ユーザが翻訳結果を確認しやすい環境を整備している。

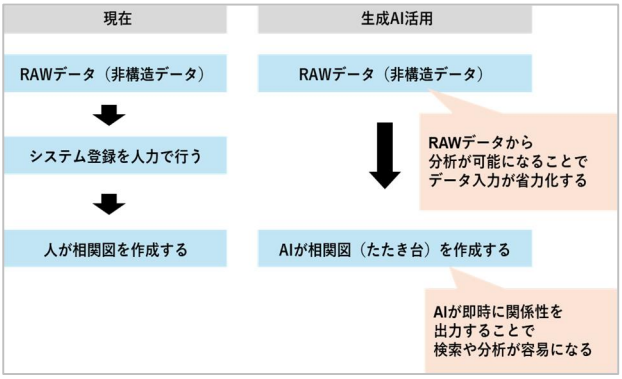
9.ユースケース「匿名・流動型犯罪グループ対策に係るデータ分析」

既存システムに登録されている構造データ・非構造データを活用した情報分析業務の合理化・効率化を目指し、GraphRAG をはじめとする生成 AI 技術の適用可能性を検討した。

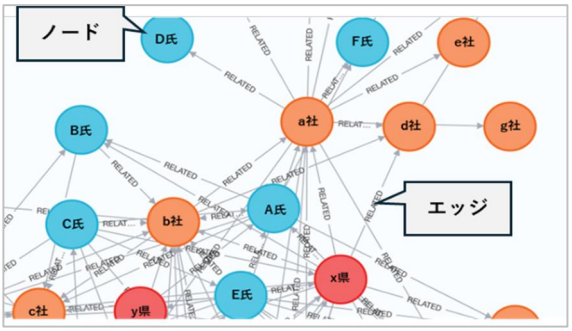
(1) 検証方法

事業者による技術提案→課題抽出→簡易検証→デモ→機能要望→改良検討→実務者評価→具体化検討のステップを踏んで検証を行った。ワークショップ参加者から現状の課題や生成 AI に対する課題をヒアリングした結果、非構造データの分析の高度化が業務効率の向上に寄与するのではないかという視点で議論が活性化し、簡易デモを含め様々な検証を進めた。その中でも、GraphRAG が有効ではないかと期待された。GraphRAG は、ナレッジベースをグラフデータベースとして構築し、エンティティをノード、関係性をエッジとして表現する。これにより、従来の RAG のような単純なテキストマッチングではなく、エンティティ間の関連性や階層構造を考慮した検索が可能となる。

図表 5: 課題ヒアリング結果に基づく技術提案

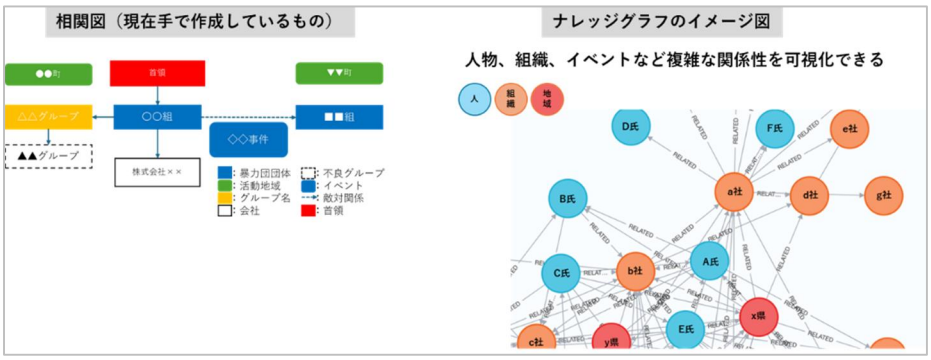


図表 6:ナレッジベースのグラフ構造



相関図作成の詳細フローをヒアリングし、中間生成物としてのグラフデータをビューワで表示する機能を検証した結果、警察庁から提供された 30 のファイルから約 1000 個のノードを抽出し、その分析結果を出力することで、担当者が手作業により作成していた相関図（検証用）を作成するために必要な情報の整理が半自動的にされ、視覚的に参照することが出来る形で表示できることが確認された。

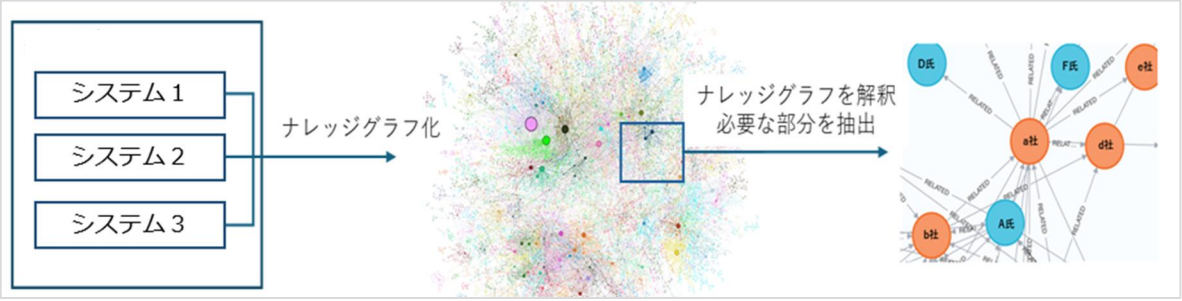
図表7:ナレッジグラフ活用案



(2) 検証結果

実務者からは、情報の信頼性表示や時系列データによる関係性変化の表現等の具体的な機能要望も挙げられ、これらを取り入れた GraphRAG 実装により、業務効率の大幅な向上が期待できるという意見が得られた。

図表8：既存システムとのデータ連携イメージ



10.ユースケース「プログラミングコードの生成、分析、ミス等の発見・修正等」

本ユースケースでは、プログラム開発・保守関連の取組とソースコード解析関連の取組に分けて結果を記載する。

10-1.プログラム開発・保守関連

業務用ソフトウェア開発・保守プロセス全般における生成 AI 活用可能性について検証・評価を行った。

(1) 検証方法

検証環境での初期試行として、ソースコード解析や改善案生成について、SQL パフォーマンス改善やテストコード生成等を事例として取り上げて検証した。

次に RAG 検索機能の実装を進め、大量のソースコードから効率的に必要情報を検索する検証を実施した。

これらの成果を実務へ適用し、実際の業務システムでの SQL 改善効果の確認や、画面設計書からテスト仕様を作成しテストコードを生成する一連のプロセス検証、障害対応支援等の幅広い活用方法を検証した。また運用支援への応用として、障害ログからの一次対応案生成や CPU 使用率の傾向分析による予兆検知の可能性も試行した。

(2) 検証結果

本ワークショップを通じて、業務用ソフトウェア開発・保守における生成 AI の活用可能性について複数の領域で有効性が確認された。

- SQL パフォーマンス改善においては、処理が遅くなっていた既存クエリに対し、生成 AI が最適化案を提示し、実際の業務システムにおいて処理時間が 1/5 に短縮されるという顕著な効果を示した。
- テスト支援分野では、ソースコード全体ではなく個別メソッドに分けてテストコード生成を依頼する手法が効果的であり、statement カバレッジ及び branch カバレッジが大幅に向上する例が確認された。また、画面設計書からテスト仕様を作成し、そこからテストコードを生成するという一連のプロセスも実現可能なことが確認された。
- ソースコード検索においては RAG の意味検索とキーワード検索を組み合わせることで、大量のソースコード群から必要な箇所を効率的に特定できるようになった。検索結果の出力文字数上限拡大や参照ソースコードのダウンロード機能追加によって実用性も向上した。
- 上流工程支援では、システム構築時の管理項目や画面一覧、画面イメージ作成等に活用され、業務部門との打合せ用たたき台として機能した。
- 運用支援においては、障害ログからの一次対応案生成や過去事例を含めた回答作成、CPU 使用率分析による予兆検知の可能性が示唆された。

これらの結果から、生成 AI は業務用ソフトウェア開発のライフサイクル全体にわたって有効に活用できることが確認された。特に、定型的なコード生成やテスト支援、パフォーマンス改善といった領域で高い効果を発揮した。

10-2.ソースコード解析関連

ソースコードや逆アセンブル結果等の技術的内容の解析における生成 AI 活用可能性について検証・評価を行った。

(1) 検証方法

検証対象として、インターネット上に公開されている一般的なソースコード群などのファイルを使用し検証を行った。

検証は、生成 AI 利用環境の機能実装状況に応じて段階的に行い、初期段階ではソースコード単体の解析から始め、環境の拡充に合わせて、逆アセンブル結果の解析や複数ファイルから構成されるプログラムの解析へと検証範囲を拡大していった。

(2) 検証結果

プログラムコード解析における生成 AI 活用の検証結果を①単体ファイルの解析、②複数ファイルの解析、③逆アセンブルコードの解析の観点で整理する。

①単体ファイルの解析

- LLM はプログラミング言語の文法（インデント、改行位置等）や関数名を理解し、解釈することが可能。文法だけでなく、ソースコードのバグチェックも実施することができる。
- LLM への入力トークン上限を超え、RAG 処理に切り替わると内容が分割されるため、重要情報が欠落し、不正確な回答を生成する可能性が高くなる。
- ファイル名や変数名に、特定の単語が含まれていると、たとえコード自体が無害であっても、不正プログラムとして誤解釈されるケースがある。

②複数ファイルの解析

- 解析に必要なファイルを全て入力することで、ファイル間の関連性を考慮した総合的な解釈が可能。
- 単体ファイルの解析と同様に、入力トークン上限を超えて RAG 処理に切り替わると不正確な回答を生成する可能性が高くなる。複数ファイルの解析時には、入力トークン数が大きくなる傾向があり、本事象が発生しやすい。

③逆アセンブルコードの解析

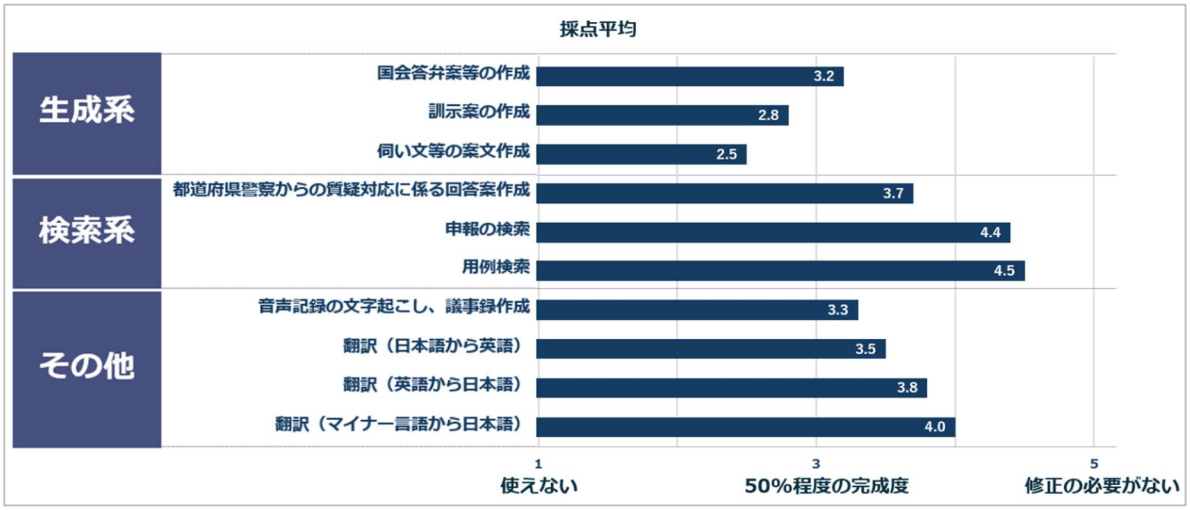
- 逆アセンブルコードは文字数が大きい傾向にあり、他の解析ケースと比べて、入力トークン上限を超え、RAG 処理に切り替わる可能性が高くなる。また、逆アセンブルコードはコード全体を見なければ処理内容を正確に把握することが難しいため、RAG 処理では不正確な回答を生成する可能性が高くなる。
- 逆アセンブルコードを解析する場合は、入力トークン上限を超えない範囲で処理ごとに分割することにより、正しい回答が得られる可能性を高めることができる。

報告書番号	9
項目	評価結果

1.ワークショップ参加者によるユースケースごとの評価

出力の採点が可能なユースケースについては、ワークショップ参加者による 5 段階評価を実施した。各ユースケースにおいて平均で 2 点台後半から 4 点台後半の評価が得られ、業務の自動化には至らないものの、たたき台として使用できる水準であることが示された。

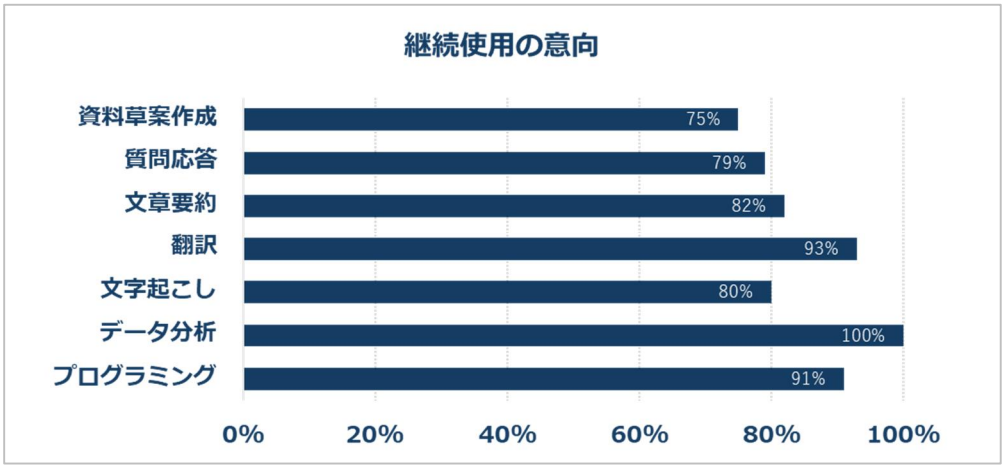
図表9:ユースケースごとの採点評価結果(平均)



2.アンケート評価

ワークショップの成果や生成 AI の業務適用に関する意識を把握するため、ワークショップ参加者を対象にアンケートを実施した。生成 AI の業務適用について、「業務で使用する機会があれば使いたい」と回答した割合が 7 つの機能すべてで 70%以上に達した。このアンケート結果は、生成 AI が警察庁の業務において一定の有用性を持つことを示している。

図表 10:アンケートにおける各機能の継続使用意向調査結果



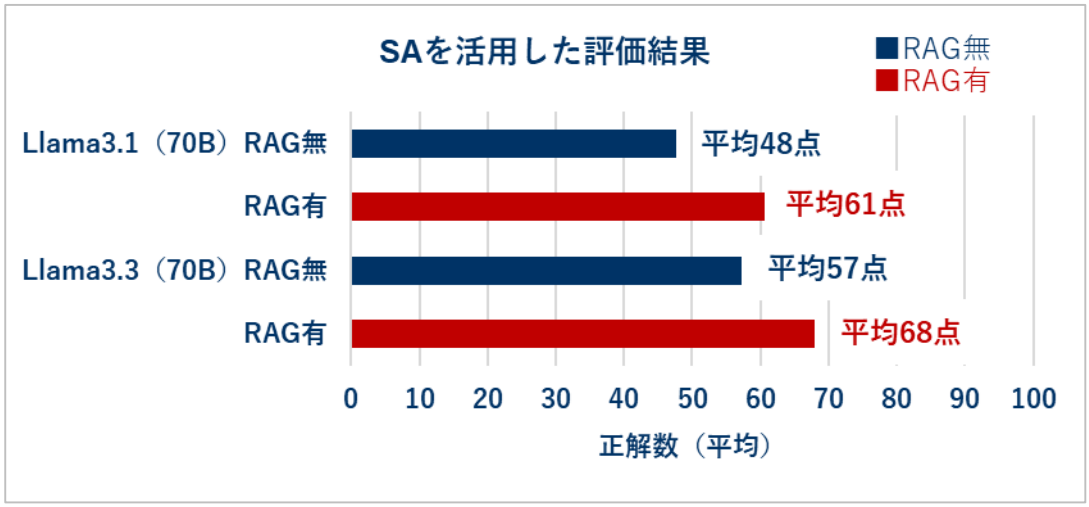
3.技術的性能評価:SA を活用した AI モデル及び RAG の性能評価

検証環境に実装した AI モデル及び RAG の性能を評価するため、警察庁から提供された警察業務に関する択一問題(ショートアンサー、SA)100 問を活用して、2 つの LLM を対象に、RAG を使用する場合と使用しない場合それぞれにおいて正解率を算出し、比較評価を行った。

AI モデル評価においては、より新しい AI モデル(Llama3.3(70B))の方が、正答数が多く、AI モデルを更新していく妥当性が確認できた。

また、RAG 性能評価においては、いずれのモデルでも RAG 無しより RAG 有りの方が、正答数が多く、知識獲得手法としての RAG の有用性が示された。

図表 11:SA を活用した評価結果



報告書番号	10
項目	クローズド環境下における構築課題

近年、生成 AI の活用が急速に進んでいるが、セキュリティ要件の厳しい業界や機密情報を扱う組織では、インターネットから完全に隔離されたクローズド環境での運用が求められる。

このようなクローズド環境では、通常の AI 開発やクラウドネイティブ技術の導入が大きな技術的ハードルとなる。特に生成 AI 基盤は、その性質上、大規模なモデルファイル、複雑な依存関係、頻繁なアップデートを前提としており、インターネット接続を前提とした設計が一般的である。加えて、生成 AI は新しい分野かつコミュニティの更新が頻発しており、バージョンの整合性がつかない等の問題が生じる。多くの場合これらの問題は、インターネット環境下であれば自動的に解消されるが、クローズド環境では手動で対応することが必要となる。

このような背景のもと、本事業では、インターネットに接続できないクローズド環境において、オープンソースソフトウェア(OSS)を活用した生成 AI 利用環境を構築した。その際、コンテナオーケストレーションによって柔軟性と拡張性を確保しつつ、様々な技術的障壁を克服する必要があった。特に依存関係の解消は、インターネット環境では自動的に行われるところを手動で実施しなければならない点が大きな課題であった。

本項では、本事業で直面した具体的な課題と対応策について整理する。

1.依存関係の解決

(1) パッケージ管理システムの制限

通常、pip、npm、apt 等のパッケージマネージャーはインターネット接続を前提としており、オフライン環境では機能が著しく制限される。

(2) 依存関係の手動解決

依存関係の解決はインターネット環境では自動的に行うことができるが、クローズド環境下では、すべての依存パッケージとその依存関係（推移的依存関係）を事前に特定し、手動でダウンロードして持ち込む必要がある。

(3) バージョン互換性の問題

あるライブラリが別のライブラリの特定バージョンに依存している場合、その関係性を正確に把握して解決する必要がある。

上記の課題に対しては、パッケージの性質に応じて以下の方法により対応した。

- ホスト OS に直接インストールが必要なパッケージの場合すべての依存パッケージ及びバージョンを確認のうえ、オンライン環境で同一の環境を再現し、パッケージマネージャーを用いて必要な構成を構築した。その後、オフライン環境での動作を確認したうえで、当該パッケージ及び必要なファイルを持ち込み、オフラインインストールを実施した。

- AI 等のホスト OS に直接インストールが不要なパッケージの場合オンライン環境にて Docker コンテナ内でパッケージマネージャーを利用して必要なパッケージをインストール・構築し、オフライン環境での動作確認を行った後、そのコンテナイメージをオフライン環境に持ち込んだ。

2.モデル取得の課題

(1) 大規模モデルファイルの取得

生成 AI モデルは数 GB～数百 GB になることもあり、これらを適切にダウンロードし、クローズド環境に持ち込む手段が必要である。

(2) モデルハブへのアクセス不能

HuggingFace、TensorFlow Hub 等のモデルリポジトリへの直接アクセスができない。

本項に関連する不具合事例及びその対応として以下が挙げられる。

- LLM の重みを Zip 化して検証環境に転送したところ、解凍後に LLM が正常に動作せずエラーが発生した。一方で、Zip 化せずにそのまま転送した場合にはエラーは発生せず、正常に動作した。ファイルの圧縮・展開処理による不整合が原因と推定される。
- モデルのダウンロード時間が長時間に及ぶため、インターネット接続が不安定な環境下では途中でダウンロードが途切れることがあった。一般的には途中から再ダウンロードが可能なはずであるが、そのようにして取得したファイルが破損していた事例が確認されている。

3.技術文書・コミュニティサポートの制限に係る課題

(1) リアルタイムの情報アクセス制限

エラー発生時に Stack Overflow や GitHub イシュー等のリソースを参照できない。

(2) 最新ドキュメントへのアクセス制限

オフライン環境では、最新の技術文書やガイドラインにアクセスできないため、あらかじめ読み込んだ上で作業する必要がある。

上記の課題に対しては以下の方法で対応した。

- 生成 AI 利用環境の LLM にエラーの内容を提示し、対策案を出力させ、試行する。
- 上記で解決しない場合は、エラーの内容を記録し、オンライン環境で調査する。

4.構築・運用上の課題

(1) コンテナイメージの事前準備

使用するすべてのコンテナイメージを事前にダウンロードし、プライベートレジストリに配置する必要がある。

(2) ソースコードの調査と改変

エラー発生時にソースコードを調査し、クローズド環境に適合するよう修正能力と知識が求められる。

5.テスト環境の課題

テスト環境がインターネットに接続されている場合、クローズド環境で発生する問題を事前に検出できない可能性がある。また、本事業の計算リソースの規模でテスト環境を準備することの難しさもある。

このような環境では、事前準備が極めて重要であり、発生する可能性のあるエラーや障害を予測し、あらかじめ GitHub 等で情報収集する等の対応策を用意しておくことが求められる。また、開発チームには高い技術力と問題解決能力が必要とされ、多くの場合「手探り」での対応を余儀なくされる。

告書番号	11
項目	課題及び技術的提案

本事業を通じて、生成 AI は警察庁の業務に一定の活用が可能であることが確認されたものの、更なる業務の合理化・効率化にはいくつかの課題が残ることが明らかになった。特に、性能向上と業務への適用範囲の拡大という 2 つの観点で、改善の余地があると考えられる。

1.性能向上の課題

現在のモデルは一定の業務に活用できる水準には達しているが、自動化には至っておらず、更なる性能向上が求められる。特に、生成 AI の基盤モデル（ベースモデル）の改良と、警察庁独自データを活用したドメイン特化学習の 2 点が重要な課題として挙げられる。

オープンな AI モデルは今後も進化が期待されるため、最新のモデルを継続的に実装することで、性能の向上が見込まれる。また、ドメイン特化学習については、本事業では国会答弁のデータを用いた学習を試行したが、データ量や網羅性が不足しており、十分な効果を得るには至らなかった。今後、質の高いデータを大量に収集し、網羅的な学習を行うことで、警察庁全般の知識を獲得した AI モデルが作成できる可能性がある。

2.業務適用範囲の拡大

生成 AI の利活用に関して、業務の更なる効率化には他システムとの連携が鍵となる。API 連携等を活用し、既存システムと直接連携できる仕組みを整備することで、AI の活用範囲を拡大できる可能性がある。他システムとのデータのやり取りをスムーズにすることで、AI の適用分野が広がり、より実用的な業務支援が可能になると考えられる。

3.警察庁独自データを活用したドメイン特化学習の提案

警察庁におけるドメイン特化型 AI モデルの構築に向けて、警察庁内の各部署が保有するデータを活用することが理想的であるが、データ収集にはいくつかの課題が存在する。

まず、警察庁内では組織や部署ごとに業務システムが分かれており、データが分散している。加えて、情報セキュリティの観点から、各データへのアクセス権限が厳格に管理されており、異なる部署間でのデータ共有は容易ではない。そのため、従来の手法では警察庁全体で統一的な学習データを収集することが難しい。そこで、「連合学習（Federated Learning）」を活用することで、各部署のデータを直接共有することなく、特化型の AI モデルを構築する方法を提案する。

4.連合学習による特化型 AI モデルの構築手法

(1) 各部署・組織単位でのトレーニング

各部署が保有する機密性の高いデータに対して、個別に AI モデルを学習し、それぞれのデータ特性を反映した AI を構築する。この際、各データを識別するための識別モデルを併せて構

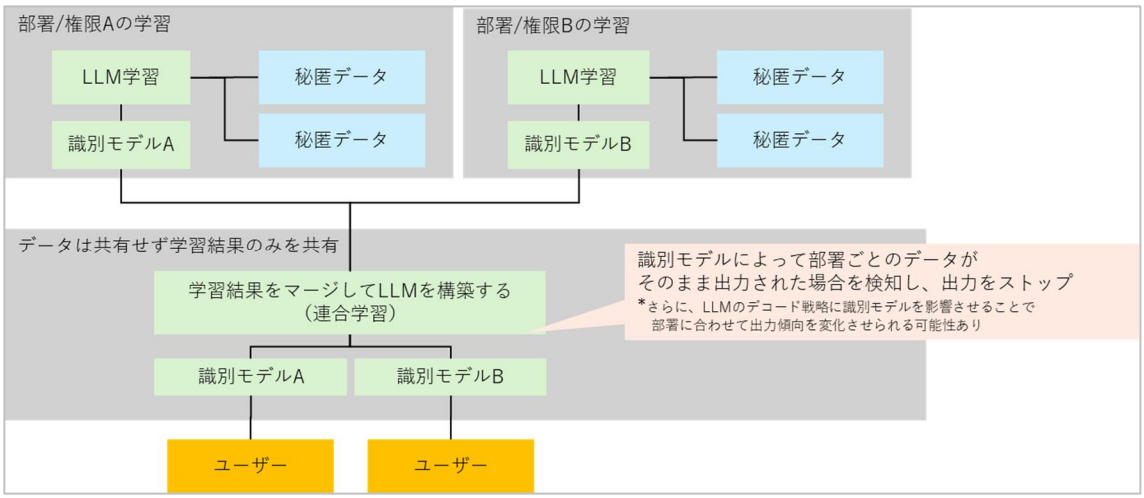
築し、部署ごとのデータの特徴を保持できるようにする。

(2) 各部署で学習した重みを統合

個別に学習された AI モデルの勾配(パラメータ)を集約し、統合することで、全体として精度の高い AI モデルを生成する。この方法により、データそのものを共有することなく、各部署のデータ特性を反映したドメイン特化型 AI を構築できる。

また、推論時には、識別モデルを活用し、閲覧可能な範囲のみを出力する制御を行うことで、情報の適切な管理を実現する(本手法は研究開発要素あり)。

図表 12:連合学習の概要図



報告書番号	12
項目	総括

1.クローズド環境下での生成 AI オンプレミス利用環境の構築

本事業では、警察庁のクローズド環境において、最新のオープンモデルを利用可能なオンプレミス環境を構築し、安全かつ効率的に生成 AI を活用できる基盤を整備した。高性能な GPU である NVIDIA H100 SXM 及び Infiniband スイッチを使用することで、大規模で高性能な AI モデルの運用や学習が可能となった。

また、ワークショップを通じて現場のニーズをヒアリングし、プロトタイプを活用した実証と改善を繰り返すことで、業務の合理化・効率化を実現する生成 AI 利用環境を構築した。

2.検証結果（概要）

本事業を通じて、クローズド環境において生成 AI 利用環境を構築するに当たっての課題を明らかにするとともに、評価結果から、生成 AI は警察庁の各種業務において活用可能であり、一定の有効性を持つことが確認された。特に以下の点が明らかとなった。

(1) 警察庁業務への生成 AI 適用

ユースケース評価では、各種業務において、業務の完全な自動化には至らないものの、生成 AI を積極的に活用することで、業務効率化に一定の効果が期待できることが確認された。また、ワークショップ参加者へのアンケート調査においても、全ての機能について「業務で使用する機会があれば使いたい」と回答した割合が 70%以上となり、特に使用経験者からは高い継続使用意向が示された。

(2) RAG の有効性

ユースケース評価及び SA を活用した評価結果から、RAG は業務知識の獲得に有効であることが確認された。

(3) AI モデルの進化

新しい AI モデル（Llama3.3）は旧モデル（Llama3.1）と比較して性能が向上しており、AI モデルを更新していく妥当性が確認された。

3.今後の展望

AI モデルの進化に合わせた継続的な更新による更なる性能向上が期待される。LLM だけでなく、音声認識モデルや埋め込みモデルについても同様の技術的進展が見込まれる。

また、生成 AI の導入効果を最大化するためには、ユーザが積極的に活用できる環境を整え、業務の中で実際に試しながら活用の幅を広げることが重要である。

なお、本事業で構築したクローズド環境における生成 AI 利用環境は、秘匿性の高いデータを取り扱う中央省庁等のほか、高度なセキュリティが求められる民間組織においても有効であると考えられる。クラウドベースの生成 AI の利用が難しい場合、本事業と同様に、オンプレミス環境への GPU サーバの導入等によって、データ流出リスクを低減しつつ、最新の AI 技術を活用していくことが可能となる。