テクニカルホワイトペーパー

Check if the document is available in the language of your choice.

オンプレ版セキュア GAI FOR HPE

セキュアな環境で社内の独自知識を AI で適切に共有し、企業の競争優位性に

目次

1.0 はじめに

2022 年 11 月 30 日に OpenAI 社が ChatGPT を公開して以降、生成 AI に対する関心と需要が一気に高まりました。この技術分野の発展はまさに日進月歩で、毎週のように新しいサービスや技術が発表されています。

その様な中、生成 AI の業務活用のニーズも急速に高まっており、その中でも特にニーズが高いのが RAG と呼ばれる文書検索の仕組みです。社内の規則や業務プロセス、提案書、レポートなど、あらゆる企業には書類という形で知見が溜まっていますが、十分に全社員に共有されている企業はいまだに少ない状況です。

そのため、欲しい情報を探すのに、同僚に聞いたり、先輩社員に聞いたり、担当部署と思われるところへ行って聞いたりと、「探す」ためだけに意外に時間を浪費しているのが実情です。ある研究結果によると、1 人あたり、ものや情報を探すのに年間 150 時間も費やしているとのことです。

また、過去の勝ちパターンを同僚に十分に共有されていないために、「車輪の再発明」に工数をかけ、書類という形で蓄積された知見が、企業の競争力につながっていない状況も良く発生します。

それを解決するのが RAG（Retrieval Augmented Generation）と呼ばれる仕組みで、企業で作成された書類を AI に学習させておくことで、チャットで書類に関する質問をすれば AI が回答してくれる仕組みを構築することができます。

しかし、ここで重要となるのが情報セキュリティです。社内の書類をシステムへアップロードする必要があるため、特に機密性の高い書類に関しては、外部が提供する RAG のクラウドサービスを使用できません。

その様なときに、本「オンプレ版セキュア GAI for HPE 」が必要になります。本ソリューションは、RAG の仕組みを外部のネットワークから隔絶されたオンプレミス環境に構築することで、情報セキュリティを担保しつつ、社内の情報共有を円滑にします。ハードウェアの選定からソフトウェアのセットアップまで一気通貫で対応できるため、お客様は RAG に必要なハードウェア要件を学習する必要がありませんし、RAG のソフトウェアを時前で構築する必要もありません。すぐに利用できる RAG を手に入れることができます。

本ホワイトペーパーでは、本ソリューションについて実際に動作した検証結果を解説し、様々なユースケースでの事例を紹介します。

2.0 協業パートナー情報

本ソリューションは、バイタリフィグループのバイタリフィ社とスクーティー社とで共同検証し、共同販売します。

2.1 会社情報

会社名	株式会社バイタリフィ
住所	東京都渋谷区恵比寿西 1-9-6 アストゥルビル 8F
事業内容	・アプリケーション及び WEB システムの受託開発 ・ベトナムオフショア開発 ・生成 AI を活用した SaaS ・自社サービスプロモーション支援

会社名	株式会社スクーティー
住所	東京都渋谷区恵比寿西 1-9-6 アストゥルビル 8F
事業内容	・ベトナムオフショア開発・ラボ型開発 ・生成 AI を活用した SaaS ・生成 AI コンサルティング ・ベトナム視察ツアーアテンド

2.2　主要ビジネス概要

・ラボ型開発サービス

日本の IT エンジニア不足に対するソリューションとして「ラボ型開発サービス」を提供しています。お客様専属の優秀なエンジニアチームをベトナムで確保します。

・生成 AI コンサルティングサービス

生成 AI を活用して業務効率化をしたい、あるいは、新たなサービスを始めたいお客様に対し、市場調査、業務調査からプロトタイプ実装、システム開発、生成 AI 導入支援などを提供します。

・セキュア GAI（後述）

AI チャットで社内の文書検索を超高速で行うことができるクラウドサービスです。

2.3　オンプレソリューション概要

本ソリューションでは、協業パートナーが提供する「セキュア GAI」というアプリケーションを、ハードウェアを購入した企業に対して提供します。セキュア GAI はクラウドサービスであり、Microsoft Azure 上で動作しますが、本ソリューションでは、オンプレミス環境のサーバー上で動作するように構築するものになります。

2.3.1　機能一覧

- ユーザー画面
  - ○ チャット機能
  - ○ 文書検索機能（RAG 機能）
  - ○ 使用法に関する問い合わせ
- 管理画面
  - ○ ログインユーザー管理
  - ○ 検索対象となるファイル管理、アップロード
  - ○ チャット履歴
  - ○ チャット Bot 埋め込み設定（外部サイトに本アプリケーションと同じ機能をもったチャット Bot を埋め込むことができる機能）

○　質問例管理（文書検索画面の初期表示する質問）

2.3.2　画面イメージ

図 1　ユーザー画面側：チャット機能の画面

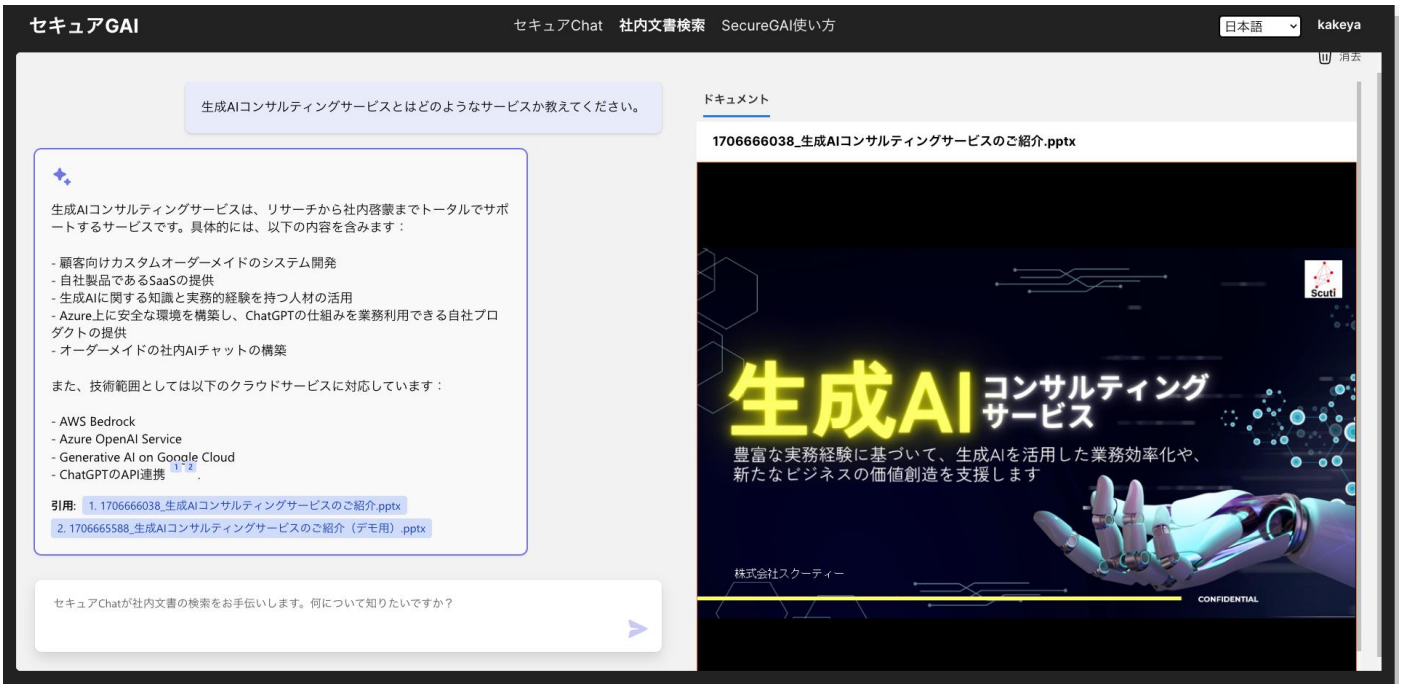

図 2　ユーザー画面側：社内文書検索の画面　右側ペインで検索で参照したドキュメントをプレビューできます

図 3　ユーザー画面側：使用法に関する問い合わせ画面　本アプリケーションの使用方法に関して質問できます

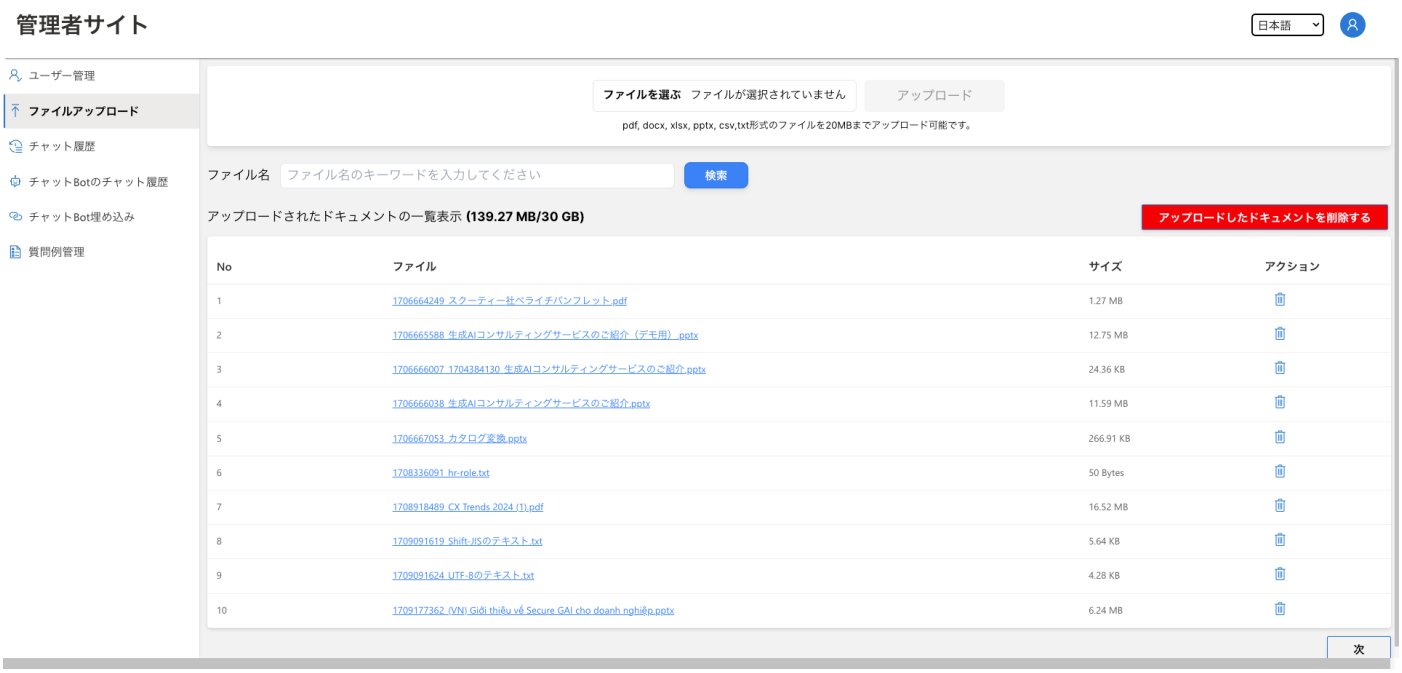

図 4　管理画面側：ファイル管理機能

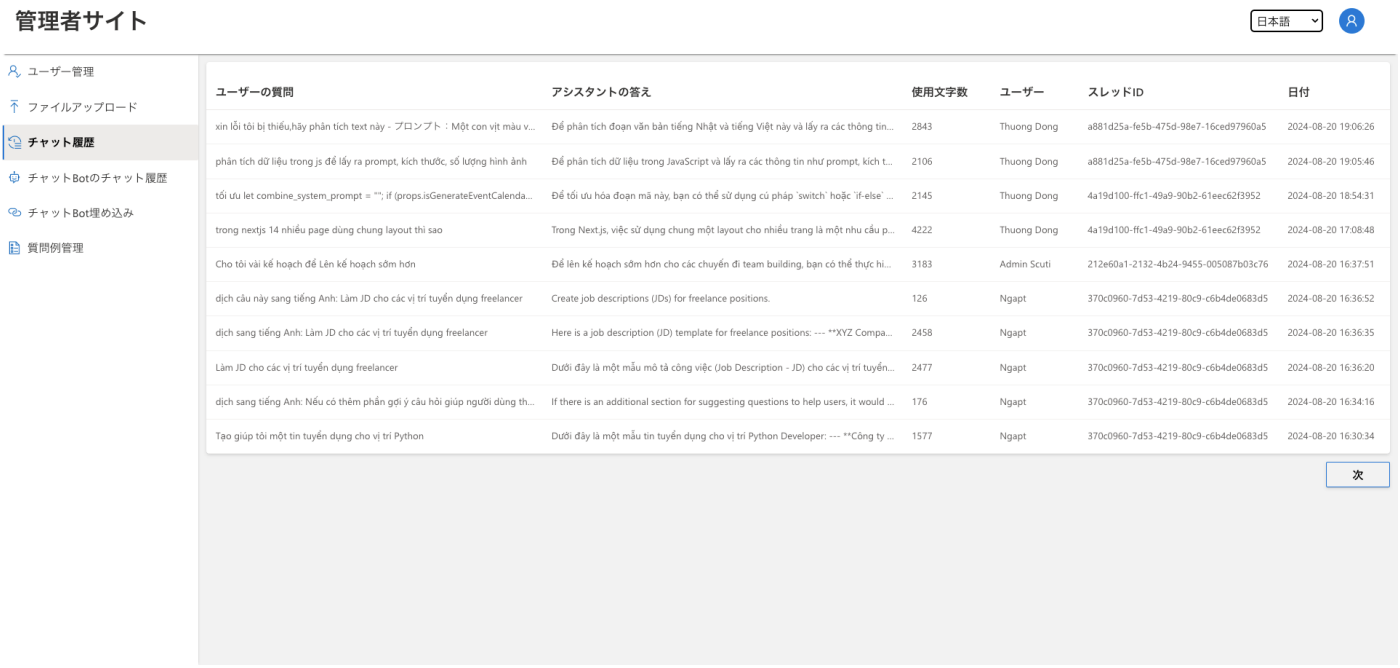

図 5　管理画面側：チャット履歴画面　管理者のみが、誰がどういうチャットをしているかを確認できます

2.3.3　使用している言語モデルとその概要

セキュア GAI は Microsoft Azure のクラウド環境上で動作しているため、Azure の機能の一つである、Azure OpenAI Service を利用し、GPT-4o の言語モデルを使用しています（2024 年 8 月現在）。本ソリューションでは、前述した通りオンプレミス環境のサーバー上で動作する言語モデルを使用する必要があるため、Meta 社の Llama3 を採用しました。Llama3 の概要は以下の通りです。

・性能

Llama3 は、Meta 社が開発した大規模言語モデル（LLM）であり、以下のような特徴を持っています：

- パラメータ数: Llama3 には 8B（80 億）と 70B（700 億）の 2 つのモデルがあります。パラメータ数が多いほど、より複雑な言語の特徴を学習できます。
- ベンチマーク結果: Llama3 70B モデルは、ChatGPT-3.5 や Claude 3 Sonnet を超える性能を有しているとされています。特に、MMLU（Massive Multitask Language Understanding）ベンチマークで高いスコアを達成しています。
- 多言語対応: Llama3 は英語を含む複数の言語に対応していますが、日本語については最適化されていないため、句読点が多すぎるなどの問題があることがあります。

・ライセンス

Llama3 はオープンソースライセンスのもとで提供されています。これにより、誰でも自由にモデルを利用、改変、再配布することが可能です。具体的なライセンス条件は以下の通りです：

- 再配布: モデルの再配布が許可されています。

- 微調整: モデルの微調整が可能です。
- 派生作品: 派生作品の作成も許可されています。
- 商用利用: 商用利用も可能ですが、月間アクティブユーザー数が 7 億人を超える製品やサービスで利用する場合は、Meta から個別にライセンスを取得する必要があります。

・ハードウェア要件

Llama3 を実行するためには、高性能なハードウェアが必要です。以下に主要な要件を示します：

- メモリ要件: 例えば、70B モデルの場合、FP16 精度で実行するには約 140GB のメモリが必要です。8B モデルの場合は約 16GB のメモリで実行可能です。
- GPU サーバー: 大規模な GPU サーバーが必要となります。特に 70B モデルでは、大規模なメモリを搭載した GPU サーバーが推奨されます。

・その他特徴

Llama3 には以下のような特徴があります：

- オープンソース: Llama3 はオープンソースとして公開されており、誰でも自由にアクセスして利用できます。これにより、多くの開発者や研究者が Llama3 を基盤として新しい技術やアプリケーションを開発することが可能です。
- コミュニティサポート: Meta 社はオープンソースコミュニティからのフィードバックを重視しており、早期かつ頻繁にリリースを行うことでコミュニティとの連携を強化しています。
- 多様なユースケース: Llama3 はチャットボットや検索エンジン、画像生成など、多様なユースケースに対応しています。例えば、Meta AI として Facebook や Instagram などで利用されており、ユーザーとのインタラクションを支援しています。

なお、本ホワイトペーパー執筆時点では、最新バージョンの Llama3.1 が公開されており、更に高い性能を示しています。Llama3.1 は Llama3 の上位互換であるため、本ソリューションでそのまま Llama3.1 を使用することができ、本検証よりも良い結果を示すものと思われます。

3.0 ソリューションの使い方と精度・性能評価

本ソリューションの性能を評価するために、3.1 に記載する検証用のアプリケーションを構築、使用し、いくつかの項目について検証を行いました。評価項目と、各評価方法に関しては、「付録 1 パフォーマンス検証方法」を参照ください。

3.1　検証用アプリの仕様

今回の検証に使用した検証用アプリケーションは、セキュア GAI と全く同じアプリケーションではなく、検証に最低限必要なセキュア GAI の機能の一部のみを簡易的に実装したアプリケーションになります。

OS、Web サーバ	Ubuntu 22.04.4 LTS、Streamlit、Cuda 12.5
使用プログラミング言語	Python 3.11

使用言語モデル	Meta Llama 3-8B
機能（チャット）	チャット機能、RAG 機能（PDF ファイルのみを最大 5 つまでアップロード可能）
機能（管理画面）	なし

表 1　チャット機能の画面

セキュア GAI と大きく異なるのは使用言語モデルです。
セキュア GAI は Azure OpenAI Service を使用していますが、オンプレミス環境では使用できないため、サーバー上で動作できる言語モデルを選択する必要があります。今回は、検証開始時点で最も高い性能をもっていた Meta 社の Llama3 を採用しました。

3.1.1 チャット機能

言語モデルが既に学習済みの知識のみを元に、AI がユーザーからの入力に対して応答する機能です。
世の中に一般的に公開されている情報に関して、回答したり、要約することができます。

図 6　チャット機能の画面

3.1.2 RAG 機能

ファイルをアップロードし、そのファイルの内容に関して回答したり、要約する機能です。

図 7　RAG 機能の画面

3.2　精度評価

今回の検証では、精度検証として回答精度と日本語精度を検証し、付随して回答速度を計測しています。

回答精度は、チャット機能に関しては、一般的な事項に関して質問し、事実として正しい回答を出力されたかどうか、RAG 機能に関しては、アップロードしたファイルに関して質問し、その内容に合致した回答を出力されたかどうかを 5 段階で評価しています。評価基準に関しては、巻末の「付録 1 パフォーマンス検証方法」をご覧ください。

また、RAG 機能の検証時には、検索対象とするドキュメントの種類や検索方法など様々な条件を変えて検証しています。各々の条件でどの様な違いが出て、それがなぜなのかを以下に説明します。

3.2.1 チャット vs RAG

本検証は、チャット機能と RAG 機能間での精度と回答速度の違いを比較しています。本検証結果が、本ソリューションの言語モデルの性能や、アプリケーションの文書検索能力の全体感を示していると解釈できます。

まず回答精度に関してはチャット機能と RAG 機能はほぼ同等ですが、RAG 機能のほうが検索処理がある分、処理ステップ数が多くなるため、チャット機能に比べてやや低い結果となっています。共に 3 と 4 の間という評価のため、回答の中に誤りが含まれており、網羅性も高くはないという評価になっています。実運用のためには少なくとも 4 を超える精度になるまでチューニングする必要はあるでしょう。

日本語精度に関しては高い評価結果となっており、概ね４から５の間という評価です。これは、日本語の文法としてほぼ誤りがなく、ネイティブが見ても、ほぼ違和感のない文章を生成しているということです。今回検証に使用した言語モデルである Llama3 は、公式には日本語をサポートしていないモデルであるため、日本語で正しく回答をしてもらうためにプロンプトのみで若干のチューニングをしています。それでも日本語精度に関しては非常にいいと言える結果を出しています。

回答速度に関しては、平均速度としてはチャット機能も RAG 機能も、一桁秒で回答を出力するという結果です。ただし、チャット機能は「検索」の処理がないため、RAG 機能よりも圧倒的に高速です。

本ソリューションの特徴として、ユーザーが何かを入力してから回答が得られるまで、外部システムへの通信が発生しないため、処理速度は速いです。もちろん、入力条件によっては回答に時間がかかることもありますが、基本的にはほぼストレスを感じない速度で回答をできる点は、日常的に使用するシステムとしては重要な点と言えます。

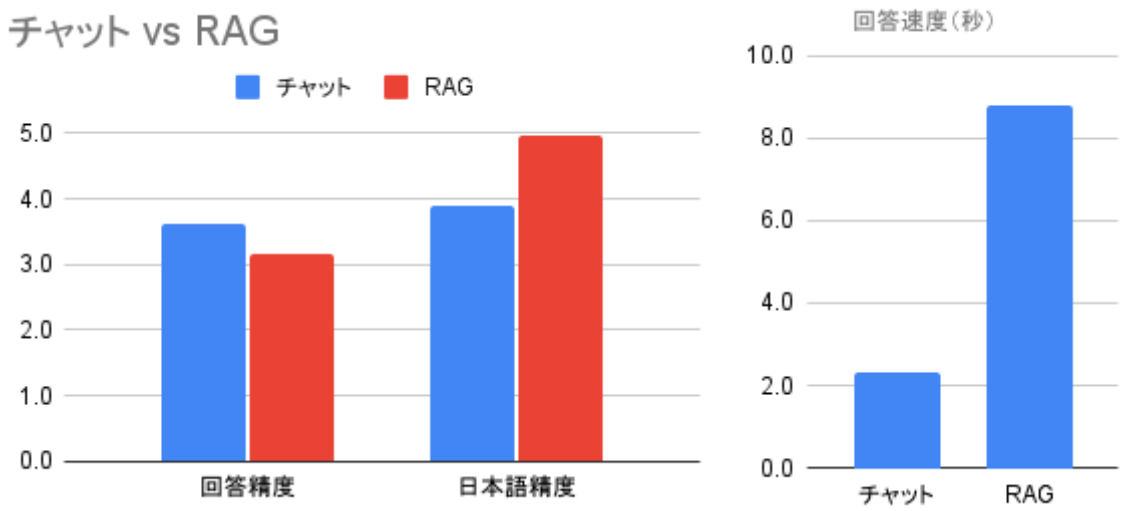

図 8　チャット機能と RAG 機能の比較　回答精度と日本語精度（左）、回答速度（右）

3.2.2 日本語資料 vs 英語資料

本検証は、RAG 機能の検索対象とするドキュメントとして、日本語のものを対象とした場合と、英語のものを対象とした場合で比較しています。結論としては、両者で有意な違いはないと言って良い結果となっています。

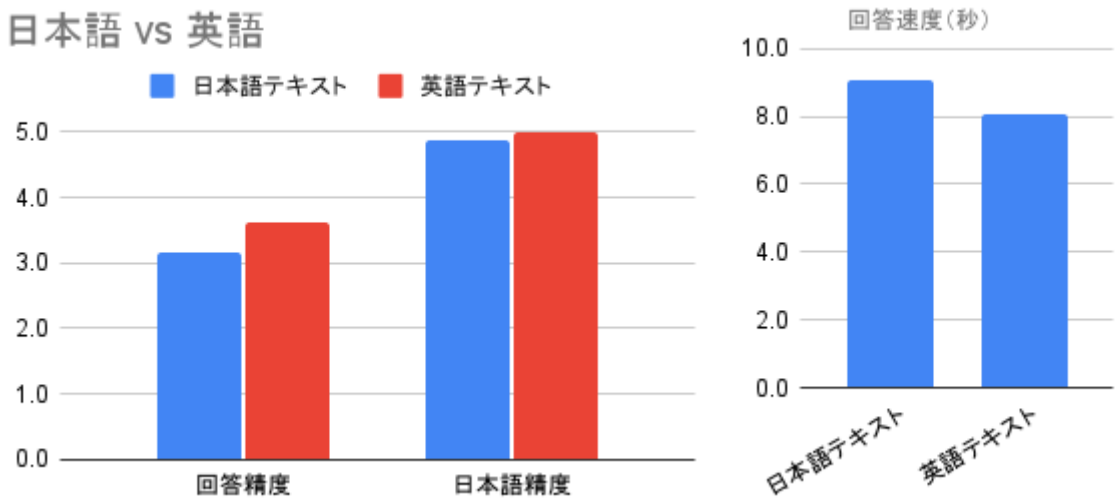

図 9　日本語資料と英語資料の比較　回答精度と日本語精度（左）、回答速度（右）

3.2.3 テキスト資料 vs プレゼン資料

本検証は、RAG 機能の検索対象とするドキュメントとして、テキスト形式の資料を対象とした場合と、プレゼン資料を対象とした場合で比較しています。

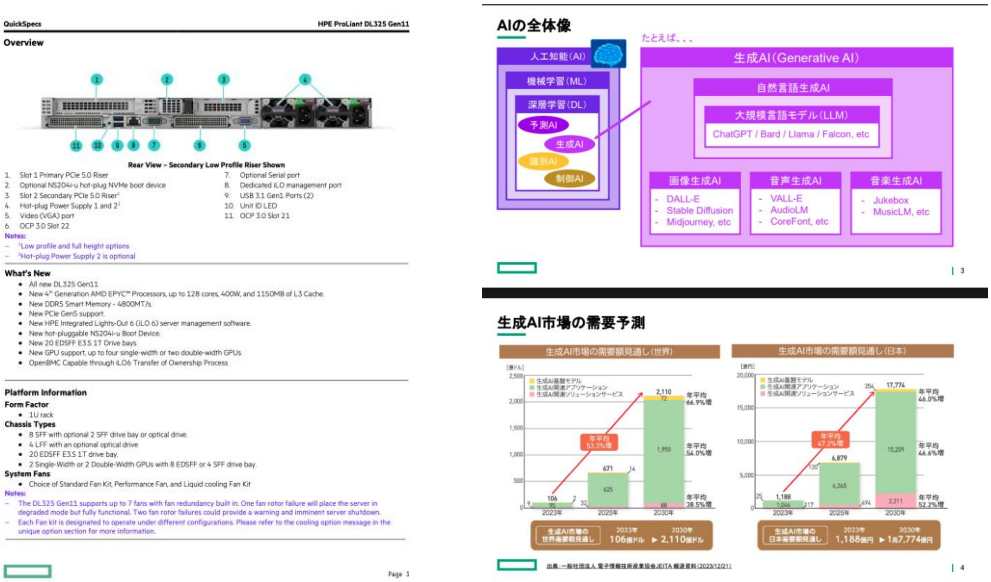

図 10　検証で使用したドキュメントのサンプル　テキスト形式（左）、プレゼン形式（右）

回答精度と日本語精度には有意な違いは見られませんが、回答速度は大きくことなり、プレゼン資料のほうが回答速度が 3 秒程度遅いという結果になっています。

RAG 機能は検索対象がテキストのみで、画像やグラフ内にあるデータは読み込めず、検索対象になりません。

従って、検索対象となるドキュメントの構成は両者で異なるものの、対象となるデータは同様で、両者ともテキストデータとなります。プレゼン資料のほうが資料内に写真や図、グラフなどが多いため、テキスト間の文脈が見えづらく、検索により時間が長くかかっているものと推測されます。

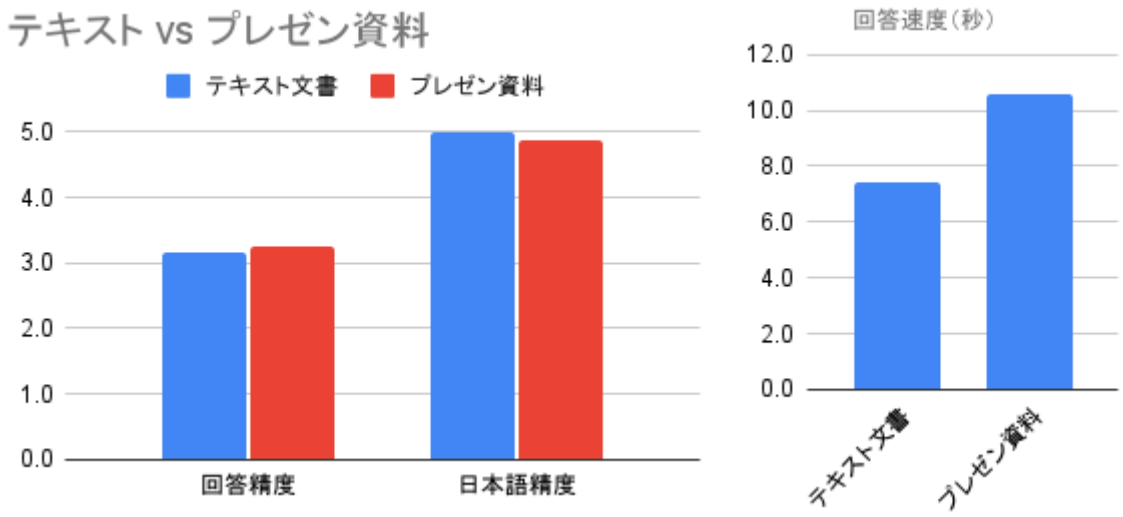

図 11　テキスト資料とプレゼン資料の比較　回答精度と日本語精度（左）、回答速度（右）

3.2.4 N to 1 vs Multi path

本検証は、RAG 機能の検索対象とするドキュメントが 1 つのみで回答を生成できる質問のしかたと（N to 1）、複数のドキュメントを参照しないと回答を生成できない質問のしかた（Multi path）を比較しています。
あまり大きな違いは見られませんでしたが、回答精度は Multi path のほうがやや落ちる傾向にありました。
Multi path のほうが回答速度がやや速い結果になっていますが、これはたまたま、あるいは他の要因によるものと推測されます。

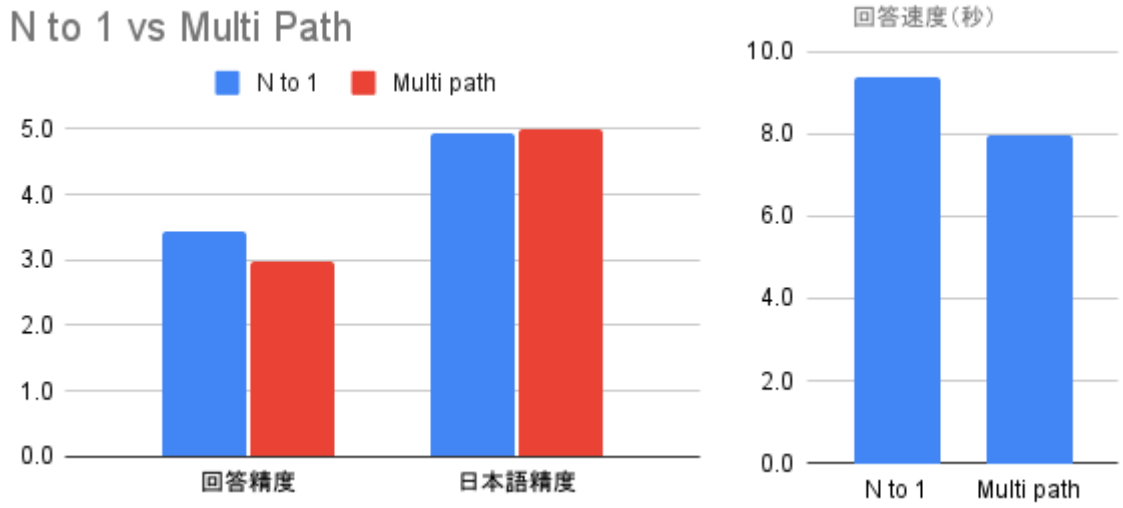

図 12　N to 1 と Multi path の比較　回答精度と日本語精度（左）、回答速度（右）

3.2.5 検索方法

本検証は、RAG 機能のいくつかの検索方法で回答精度を比較しています。
検索方法とは、本ソリューションの RAG 機能が、アップロードされたドキュメント内のテキストから、ユーザーが入力した内容に最適な返答を探すためのアルゴリズムです。今回は以下の 3 つの方法に対して検証しています。

・ベクトル検索
テキストをベクトル（多次元の数値）に変換し、意味的な類似性を基に検索する方法。文脈を理解し、類似した意味を持つ文書を高精度で検索できます。意味的な類似性を重視するため、ユーザーの意図をより深く理解した結果が返されます。ただし、正確な単語一致が必要な場合は適さないことがあります。

・全文検索
文書内のすべての単語をインデックス化し、ユーザーのクエリと一致するテキストを検索する方法。単語の正確な一致に基づいて検索を行うため、特定のキーワードに対する精度が高いです。クエリと正確に一致する単語が含まれる文書を見つけやすいですが、文脈や意味的な類似性を考慮しないため、類似表現に対応できないことがあります。

・ハイブリッド検索
ハイブリッド検索は、ベクトル検索と全文検索を組み合わせることで、双方の長所を活かします。

具体的な組み合わせ方：

- ベクトル検索：まず、ユーザーのクエリをベクトルに変換し、意味的な類似性に基づいて関連する文書を検索します。
- 全文検索：同時に、クエリのキーワードをインデックス化された文書から検索し、キーワード一致に基づいて関連する文書を検索します。
- 結果の統合：ベクトル検索とフルテキスト検索の結果を統合し、両方の結果から最適なものを選び出します。

結果としてはハイブリッド検索が一番いい回答精度を出しています。
ハイブリッド検索はベクトル検索と全文検索のいいとこ取りで精度向上を目指した検索方法ですが、それが今回の検証でも実証された形となります。

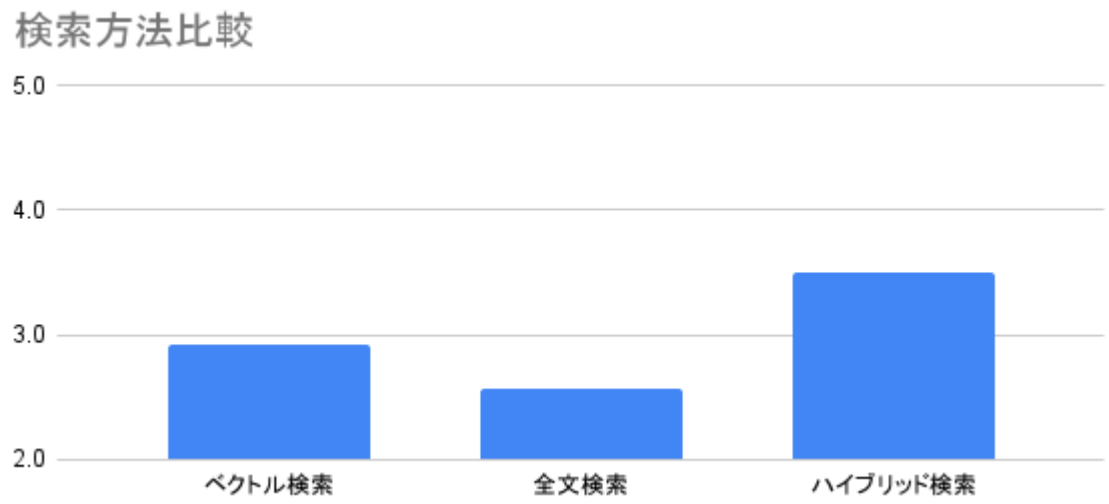

図 13　検索方法ごとの回答精度比較

3.2.6 ファイル量

本検証は、RAG 機能の検索対象とするドキュメントの量（合計ファイルサイズ）が検索精度にどの様な影響を与えるかを評価しています。結論としては、両者で有意な違いはないと言って良い結果となっています。
5MB（ファイル量が一番少ないケース）が最も精度が悪い結果となっていますが、これはテストケースの数が一番多いことが影響していると考えらます。従って回答精度は、ファイル量に依存しないという見方をするのが妥当であり、ファイル量が増えるほど検索精度が下がるという傾向は見られませんでした。

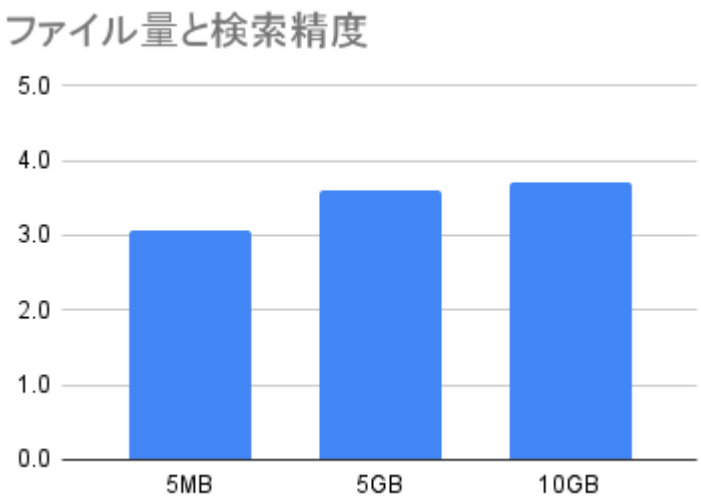

図 14　検索対象のファイル量と検索精度の関係

3.3　性能評価と多重度（複数ユーザー利用の観点）

このセクションでは応答速度に着目した性能評価および、リソース使用量に基づいた多重度の考え方を記載します。

ハードウェアの性能を評価するテストケースとしては、1、5、10 ユーザーが同時にファイルアップロード操作と RAG 機能の質問をした場合（このユーザー数を CCU: Concurrent Users と定義します）の、CPU 使用率、RAM 使用率、GPU 使用率の変化を観測しています。

ネットワークに関しては、同様に 3 パターンのユーザー数に対して、帯域幅を 1Mbps、10Mbps、100Mbps で制限するケースと制限しないケースで、チャットで質問をし、通信エラーが出るかを観測しました。

・ファイルアップロード操作

10 ユーザーまで増やしても有意な差は見られませんでした。

ユーザー数（CCU）	RAM 使用率	CPU 使用率	GPU 使用率
1 CCU	0.3%	6.0%	71.9%
5 CCU	0.3%	6.0%	71.9%
10 CCU	0.3%	6.0%	71.9%

表 2　同時接続数とハードウェアリソースの状況（ファイルアップロード操作時）

・ファイル検索が伴う質問操作

こちらも 10 ユーザーまで増やしても有意な差は見られませんでした。

ユーザー数（CCU）	RAM 使用率	CPU 使用率	GPU 使用率
1 CCU	1.3%	6.0%	84.5%
5 CCU	2.1%	6.0%	84.5%
10 CCU	2.2%	6.0%	84.5%

表 3　同時接続数とハードウェアリソースの状況（質問時）

・帯域制限

こちらも 10 ユーザーまで増やしても有意な差は見られませんでした。

ユーザー数（CCU）	帯域制限：1Mbps	帯域制限：10Mbps	帯域制限：100Mbps	帯域制限：なし
1 CCU	エラーなし	エラーなし	エラーなし	エラーなし
5 CCU	エラーなし	エラーなし	エラーなし	エラーなし
10 CCU	エラーなし	エラーなし	エラーなし	エラーなし

表 4　同時接続数とネットワークの状況

以上の結果から、今回の検証環境においては、少なくとも 10 CCU まではハードウェアにもネットワークにもほとんど影響なく動作することがわかります。10 CCU は 2,000 MAU 程度のゲームシステムに相当すると言われています。また、RAG は企業内で近年普及は進んでいるものの、どんなに RAG の導入に力を入れている企業でも日常的に使用している社員は 20%程度と言われています。これらの数字から推測すると、少なくとも 1 万人規模以下の企業に関してはストレスなく使用できるシステムと言えます。

3.4　評価結果から想定されるユースケース

ここまでの検証で得られた結果を以下に簡単にまとめます。

- ユーザーの入力に対する返答は数秒～十数秒程度の、ほぼストレスを感じない速さで得られる。
- RAG 機能の回答精度は総じて「回答の中に誤りが含まれており、網羅性も高くはない」程度だが、改善の余地は多い（後述の注記参照）。
- ただし上記は、アップロードファイル内のテキストのみが対象であり、ファイル内の画像やグラフなどテキスト以外に含まれる情報は検索できない。
- 日本語の精度に関しては全く問題がない。
- 10 CCU 程度の同時接続であれば、ハードウェアへの負荷やネットワーク帯域への影響はなく、問題なく本ソリューションを利用できる。

上記から、以下のようなユースケースにおいて、本ソリューションを効果的にご利用いただけます。

- 就業規則や情報セキュリティガイドラインなど、全社で共有すべき情報の共有。
    - 「同僚や担当部署に聞く」という手段の代替。年間で 1 人あたり 150 時間と言われる「探す」コストを低減できる。
- カスタマーサポートの、顧客へ回答する前のマニュアルの確認。
    - 「顧客からの問い合わせに対してカスタマーサポートがマニュアルから該当部を探す」という手段の代替。顧客へ直接回答するチャットボットではなく、社内のマニュアル検索手段として使用する。
- 新規に配属されたメンバーへの受け入れ教育
    - 特定の部署で共有されている資料の一覧から、新しいメンバーが最初に知っておくべき資料を AI にリストアップさせ、受け入れ教育のカリキュラムを作成させる。

- ○ 新しいメンバー自身は、資料を隅から隅まで読むのではなく、重要な点を AI に要約させることで学習コストを抑える。
- 提案書や申請書などのドラフトを自動生成

逆に、以下のようなユースケースにはあまり効果的ではありません。

- 画像や動画の生成（本ソリューションにはその様な機能が備わっていません）
- 最新の情報に関する調査

注記

本検証では、何もチューニングしていない、初期状態の Llama3 を使用しています。また、本検証完了直後の 2024 年 7 月 23 日に Llama3.1 がリリースされ、言語モデルの精度が大きく向上しています。したがって、本検証結果よりも精度が高くなる余地はまだあり、顧客へ導入したあとの実運用時にはより高い精度でご利用いただけます。

4.0 構成のベストプラクティス

このセクションでは HPE のサーバーポートフォリオとその特徴、およびスクーティ―社の AI ソリューションの推奨構成を提示します。

4.1　HPEサーバーポートフォリオとその特徴

HPE では、汎用的なラックマウント、タワー型のサーバーシリーズである HPE ProLiant サーバーや、ブレード型の HPE Synergy、エッジ型の HPE Edgeline、High Performance Computing の Cray シリーズ、ミッションクリティカルの Superdome シリーズ、無停止コンピュータシリーズの Non S top、仮想化専用機のハイパーコンバージドインフラなど、幅広いラインナップをご用意しています。

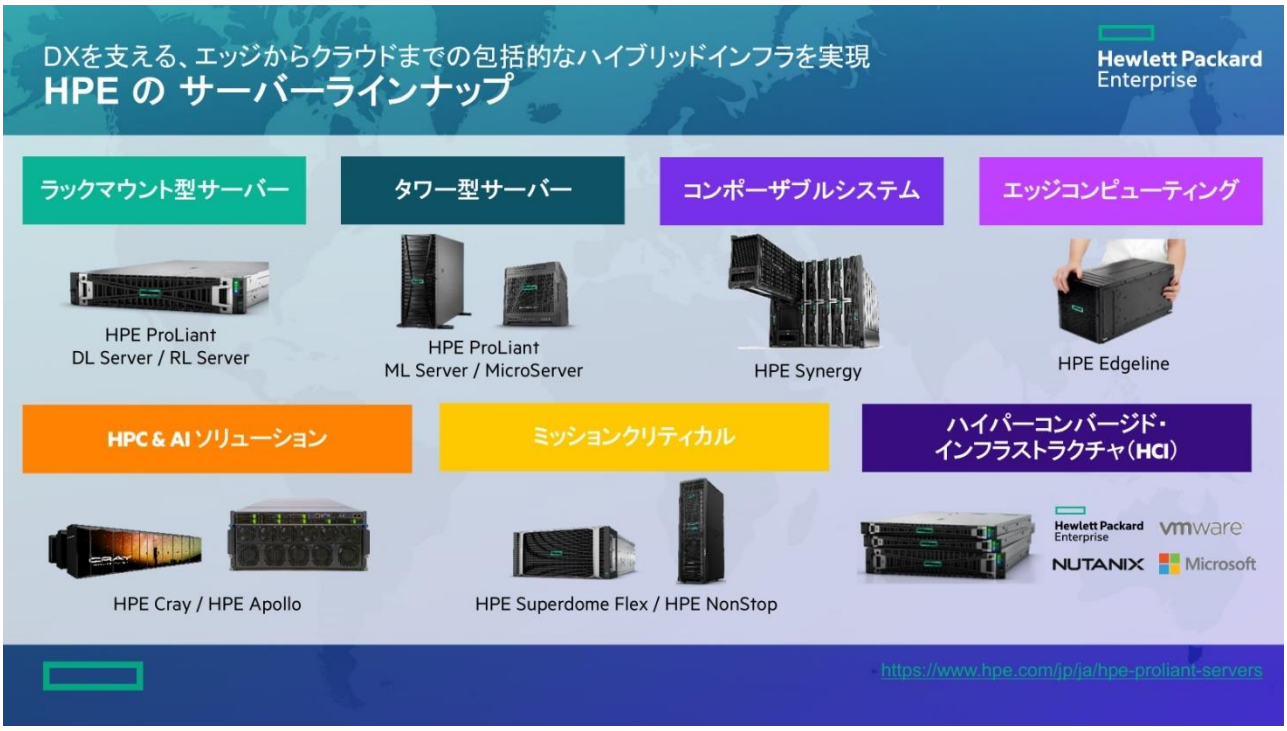

図 15　HPE サーバーラインナップ

エンタープライズ向けの汎用サーバーとしては、ラック型、タワー型を中心に、搭載できる CPU タイプ、CPU 数および、筐体サイズによって分かれます。

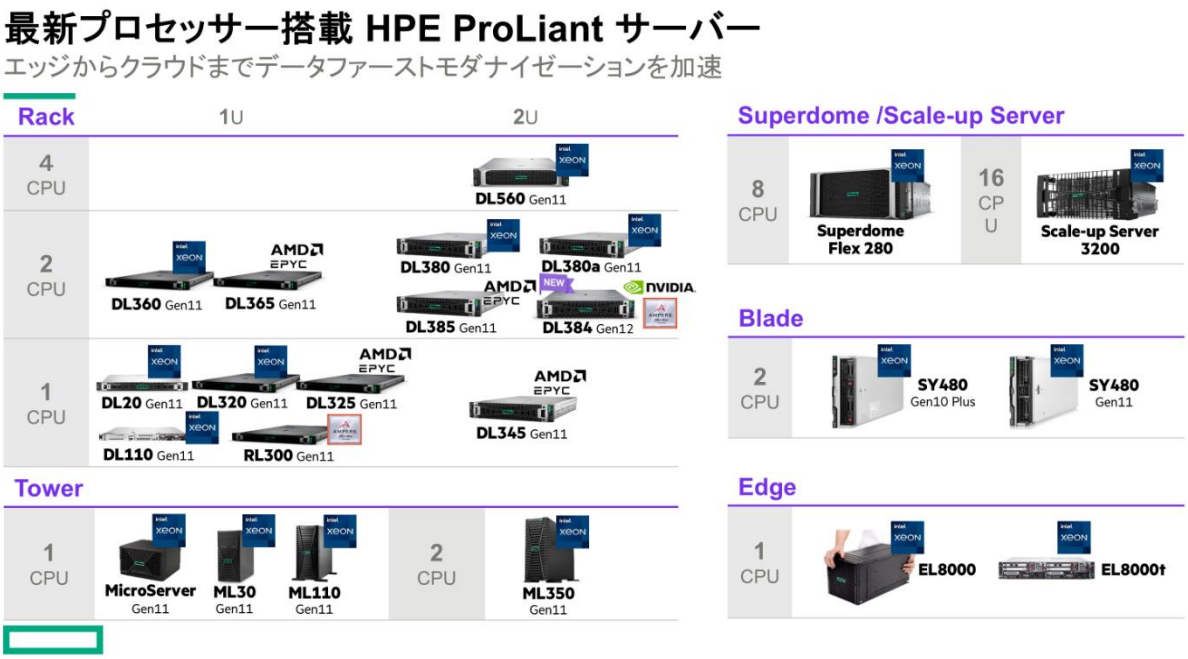

図 16. 最新プロセッサー搭載 HPE ProLiant サーバーシリーズ

HPE サーバーの特徴として、“直感的”、“安心”、“最適化”を謳っており、オンプレでもクラウド型の監視・管理が可能な SaaS サービスの提供や、ファームウェア改ざん対策をはじめとしたセキュリティ機能、サービスの提供、中でも AI サーバーとしてご利用いただけるよう最適なサーバー設計として GPU の搭載枚数の拡充や、排熱量低減、省電力化を意識したデザインとなっています。

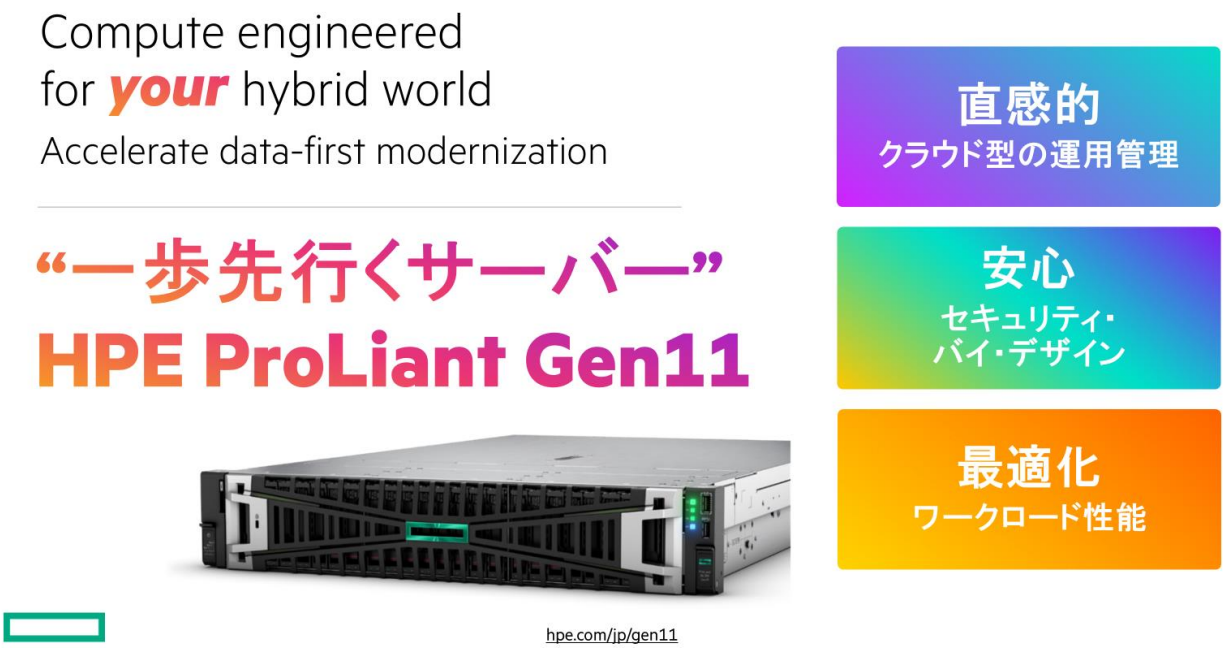

図 17. HPE ProLiant サーバーの特徴

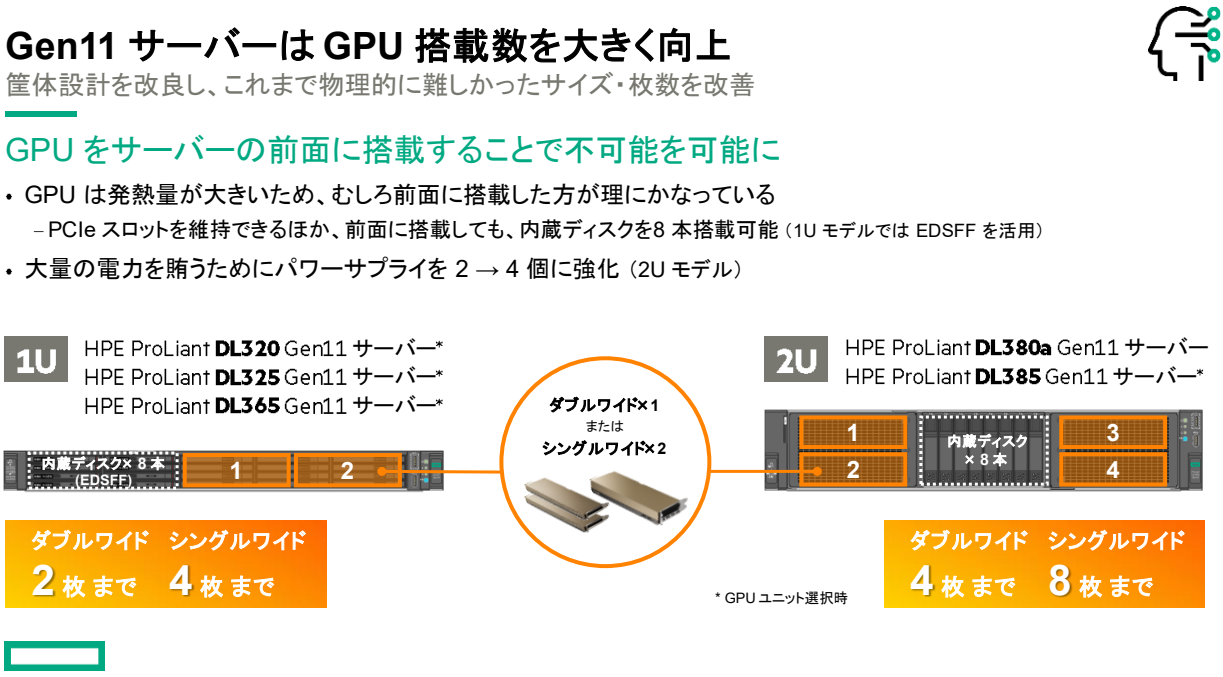

図 18. GPU 搭載枚数の拡充

図 19. 発熱量・消費電力を抑える工夫

4.2　検証システム構成

4.2.1 サーバースペック

スクーティー社のオンプレ版セキュア GAI FOR HPE ソリューションをホストするサーバーとして、2.3.3 に示す通り、必要なサーバースペックを図 20 に示します。

	Model	Qty	Spec, Options, Notes
GPU サーバー	 HPE ProLiant **DL380** Gen11	1	**Model:** HPE ProLiant DL380a Gen11 4 Double Wide **CPU:** Intel Xeon-Gold 6426Y 2.5GHz 16-core 185W Processor x2 **MEM:** 512GB / HPE 64GB (1x64GB) Dual Rank x4 DDR5-4800 x8 **Disk:** HPE 1.92TB SATA 6G Mixed Use SFF BC Multi Vendor SSD x2 **Array:** HPE MR216i-o Gen11 x16 Lanes without Cache OCP x1 **OS Boot:** HPE NS204i-u Gen11 NVMe Hot Plug Boot Optimized Storage Device **Network:** Broadcom BCM57416 Ethernet 10Gb 2-port BASE-T OCP3 Adapter x1 **PS:** HPE 1800W-2200W Flex Slot Titanium Hot Plug Power Supply Kit x4 **iLO:** iLO Advanced ライセンス
GPU	 NVIDIA **L40S**	1	**GPU:** NVIDIA L40S 48GB PCIe Accelerator

図 20.　スクーティー社オンプレ版セキュア GAI FOR HPE 推奨構成例

2U サイズの 2CPU サーバーで世界で最も人気の HPE ProLiant DL380 Gen11 サーバーに、Ada Lovelace アーキテクチャーの汎用万能型 GPU の NVIDIA L40S を 1 枚搭載しています。CPU、GPU をはじめとするマシンスペックを増強することで、さらに高性能な AI 環境をご利用いただくことも可能です。

4.2.2 ソフトウェア要件

本検証時点でのオンプレ版セキュア GAI FOR HPE のソフトウェア要件は以下となります。

OS、Web サーバ	Ubuntu 22.04.4 LTS、Nginx-1.25.4、 Cuda 12.5
プログラミング言語	Python 3.11（API）、Next JS（FE）
使用言語モデル	Meta Llama 3-8B
データベース	MongoDB(NoSQL), MySQL (RDB)
キャッシュ	Redis

4.2.3 ソリューションスタック

本検証時点でのオンプレセキュア GAI のソリューション全体のスタック図は以下となります。

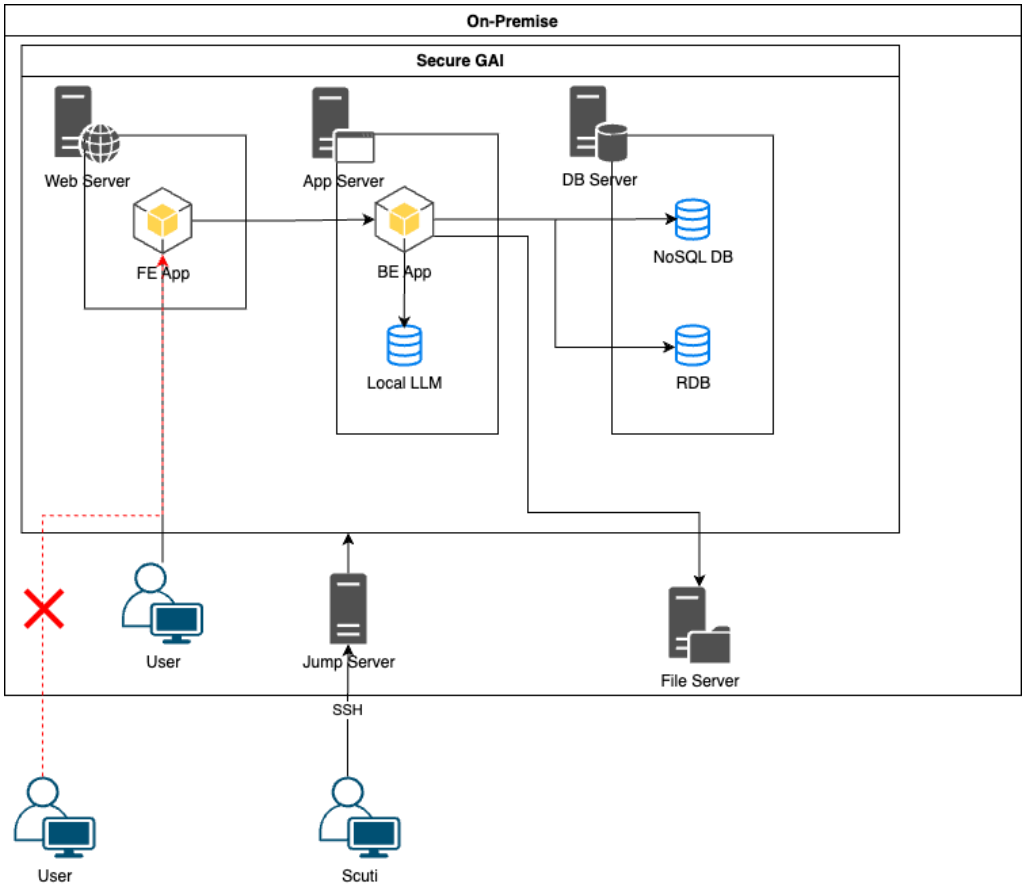

図 21.　オンプレ版セキュア GAI FOR HPE ソリューションスタック図

注記

2024 年 7 月の検証時点でのソリューションスタック図となります。あくまで参考例としてご参照ください。予告なく変更する可能性がございます。

5.0 お問い合わせ先

「オンプレ版セキュア GAI for HPE」に関するお問い合わせ：
株式会社バイタリフィ営業部 - sales_com@vitalify.jp

サーバーハードウェアに関するお問い合わせ：
お問い合わせ - ヒューレット・パッカード エンタープライズ - HPE - セールスへのお問い合わせとサポート | HPE 日本

6.0 その他のリソース

Scuti 社ホームページ

VITALIFY 社ホームページ

HPE ProLiant DL380 Gen11 サーバー

付録 1 パフォーマンス検証方法

パフォーマンス検証において、回答精度と日本語精度を 5 段階で評価しています。
以下に回答精度と日本語精度の 5 段階評価の指標とその指標を記しました。

A1.1 回答精度

スコア	正確性
5	適切な箇所の情報を参照し、内容を網羅して回答できている。(80 ~ 100% 正解)
4	ほとんど正確な回答ができているが、内容があまり網羅できていない。(60 ~ 80% 正解)
3	正解の回答があるが、不正解のものも混ざっている。網羅性も低い。(40 ~ 60% 正解)
2	正解の回答より不正解の回答の方が多い。内容も網羅できていない。(20 ~ 40% 正解)
1	ほとんどが不正解、内容の網羅性も低い。 (0 ~ 20% 正解)

A1.2 日本語精度

スコア	評価基準
5	- 完全に正確で文法的誤りなし - 非常に自然でネイティブのよう - 完全に一貫して文脈に合う
4	- ほぼ正確でわずかな誤り - 自然でネイティブに近い - 概ね一貫して文脈に合う - 文脈に多少の不適合があるが適切
3	- いくつかの誤りがあり理解に影響 - 一部不自然でネイティブとは異なる - 一貫していない部分もある - 文体が一貫せず文脈に適していないことがある
2	- 多くの誤りがあり理解に大きく影響 - 不自然でネイティブとは思えない - 一貫性に欠け文脈に合わない - 文体が不適切で文脈に反する
1	- ほとんど誤りで理解不能 - 非常に不自然でネイティブとは思えない - 完全に一貫せず文脈に合わない - 文体が完全に不適切

A1.3 回答速度

Google Chrome Developer tool を利用してプロンプトを打ってから応答、回答結果が完全に返ってくるまでの待ち時間を測定しています。具体的には、TTFT と TPOT の合計時間を計測しました。

<TTFT と TPOT の定義>

・TTFT(Time To First Token)：ユーザーがクエリを入力してから、どれだけ早くモデルの出力を見始めるか。

・TPOT(Time Per Output Token)：システムを照会する各ユーザーに対して出力トークンを生成する時間。

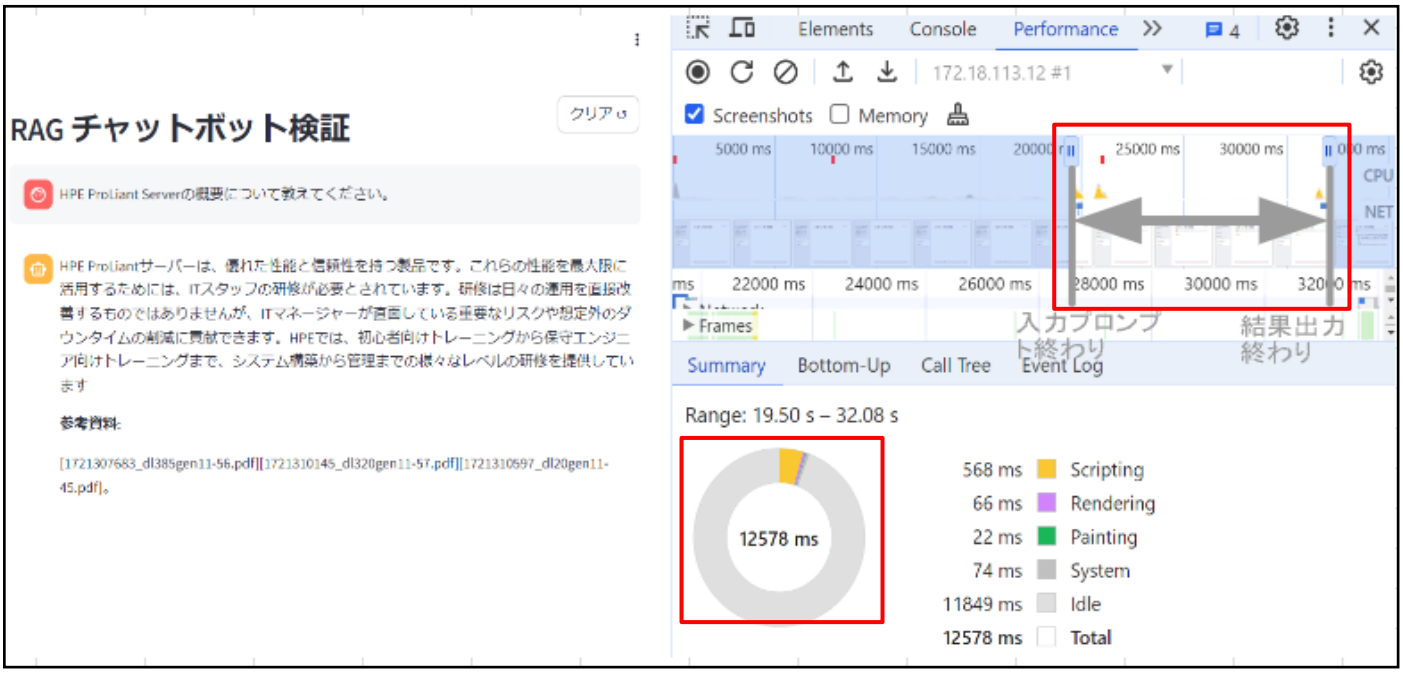

図 22.　Google Chrome Developer tool を利用した測定例

上記の場合、TTFT と TPOT の合計は 12578ms となります。

A1.4 性能評価

性能評価においては、ファイルアップロード時・RAG チャットボットへ質問時に CCU（同時接続者数）毎のリソース使用率（CPU、RAM、GPU）や帯域制限をかけてパフォーマンスの測定も行なっています。

CCU はブラウザを同時に複数立ち上げることで環境を作っています。例えば、CCU が 10 での検証の場合、10 のブラウザを同時に立ち上げて任意のアクション（ファイルアップロードやチャット）を行いました。

リソース使用率に関しては、任意のアクション（ファイルアップロードやチャット）が行われた際に使用されたリソース使用率を計測しました。例えば、10CCU でファイルアップロードした際に使用したリソースを測定する場合、ファイルアップロード後に使用したリソースを測定して計算しています。

帯域制限に関しては、任意のアクション時に帯域を 1Mbps, 10Mbps, 100Mbps、制限なしと指定して回答までの時間とエラーの有無を確認しました。

お客様のニーズに最適な製品をお選びください。

HPE 営業にお問い合わせください。

Get updates

2024 年 7 月

Hewlett Packard Enterprise