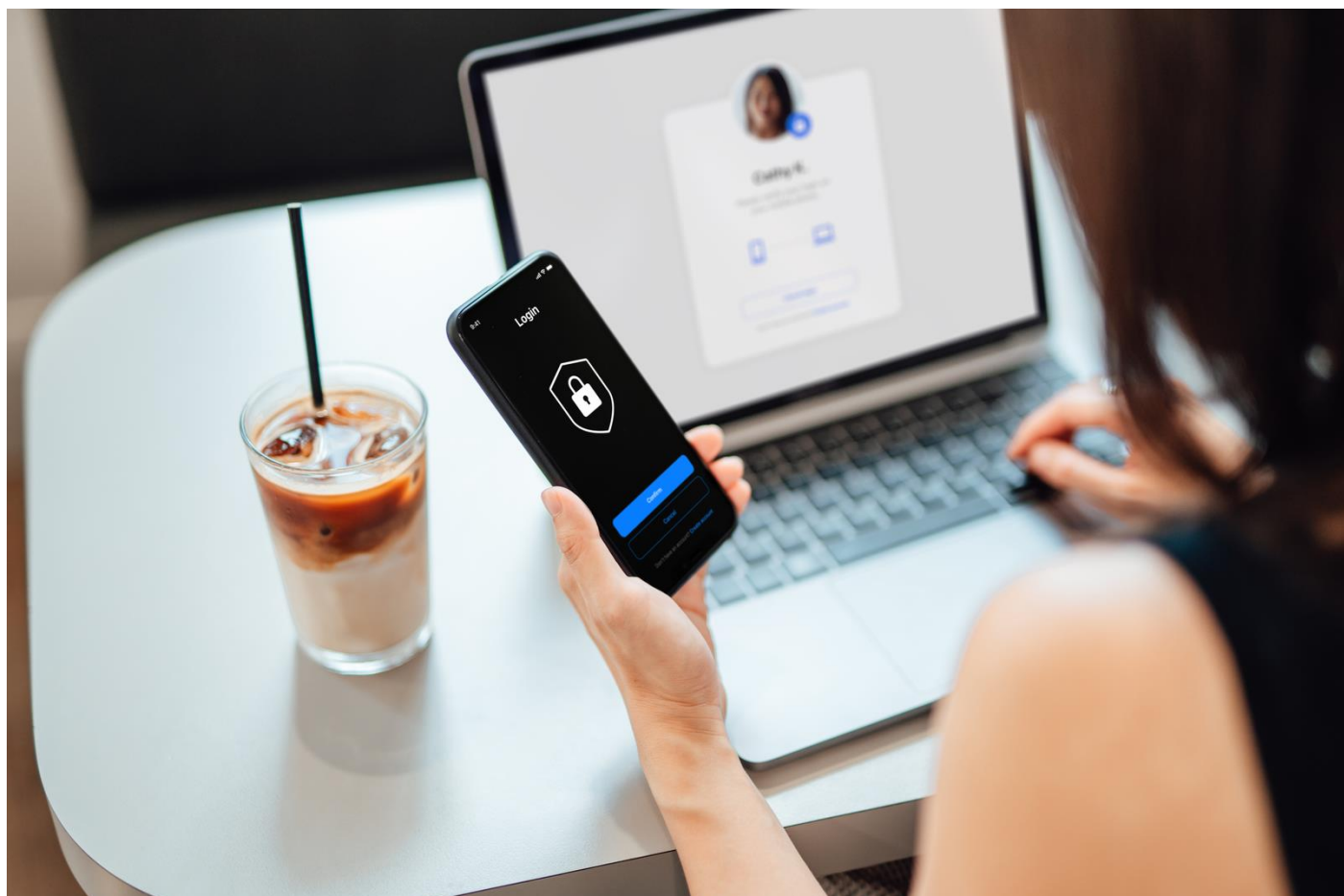




GPTBase (オンプレミス)

データセキュリティを強化したオンプレミスチャットボットソリューション



目次

1.0 はじめに 3

2.0 協業パートナー情報..... 3

3.0 ソリューションの使い方と精度観点..... 6

4.0 構成のベストプラクティス 17

5.0 まとめ 23

6.0 その他のリソース..... 24



1.0 はじめに

近年、生成 AI（Generative AI）の技術革新が加速する中、日本国内においてもその活用が急速に広がりを見せています。しかし、ビジネスの現場においてはデータ保護やセキュリティ、法規制への対応が重要視されるため、クラウドベースのソリューションに対する懸念も根強く残っています。このため、オンプレミスでの生成 AI 導入希望を求める声が高まっています。

HPE 社と Sparticle 社は、こうした日本市場のニーズに応える為、国内における生成 AI のオンプレミス対応の提供において協業する事を決定致しました。本協業の目的は、企業がデータを安全に管理しながら、生成 AI の先進的な利点を最大限に活用できる環境を柔軟に導入可能にすることです。

本取り組みでは、以下の点を重視しております。

- データセキュリティの強化：日本の規制や企業のデータガバナンスに則った、安全で信頼性の高いオンプレミスの構築。
- カスタマイゼーション：業界や業務プロセスごとのニーズに応じて最適化された生成 AI を提供。それぞれのビジネスの特性にあった運用を実現。
- コスト効率が良い導入：中堅企業やスタートアップ企業でも導入可能な生成 AI を実現し、広範なビジネスシーンでの活用を促進します。大規模な予算を持ってない企業にも生成 AI へのアクセスを提供します。

この競合におけるターゲットは、製造業、金融機関、政府機関など、データセキュリティに特に敏感な業界です。これらの分野では、生成 AI を用いた、効率的な業務プロセスの策定や新たなビジネス価値の創造には、オンプレミス環境での導入が非常に重要です。

2.0 協業パートナー情報

Sparticle 株式会社は 2019 年 8 月に創業した、生成 AI 分野におけるリーディングカンパニーです。生成 AI の活用時に課題となっている、それらしい嘘をついてしまう問題（ハルシネーション）の改善のため、日本では 2024 年初旬ごろから注目されている生成 AI の精度向上技術である“RAG 技術”について、2023 年の初旬より着手し研究開発を行ってきました。1 年ほど先行する RAG 技術を既に多数の自社製品にて実際に運用し、生成 AI 活用についての幅広い知見と技術を有しています。その中でも、Sparticle 株式会社にて独自開発を行なった日本語対応の独自 LLM により力を入れ、製品展開を行なっています。

2.1 会社情報

社名：Sparticle 株式会社

代表者：代表取締役 金田 達也

設立：2019 年 7 月：資本金：9,000 万円

事業内容：

- ・ AI コンサルティング事業
- ・ AI 研究開発事業
- ・ AI プロダクト事業
- ・ AI ソフトウェア設計およびソフトウェア開発に関連するサービス
- ・ クラウドサービス導入支援



2.2 主要ビジネス概要

1:GPTBase

GPTBase は RAG 技術※1 を活用し、生成 AI に自社データなど特定のデータを学習させることにより、自社データに基づいて回答を生成するオリジナルな生成 AI ツールです。社内外のデータに基づいて生成 AI を製作、それを社内または社外に公開することが可能です。カスタマーサポートや社内の問い合わせ対応、ナレッジ共有など幅広い用途でご活用頂けます。

学習可能なデータの幅が広いこと、回答精度が良いことを高く評価頂いており、読解難易度が高い図面データなどを含む製造業のお客様や、コンプライアンス観点から正確な情報を伝えることが必要な金融業のお客様など、さまざまな業種業界のお客様に採用いただいています。



図 1. GPTBase のサンプル画面

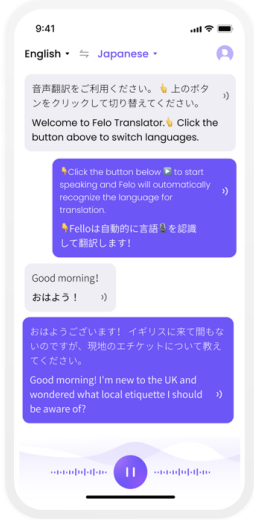
※1: RAG (Retrieval-Augmented Generation) 技術は、情報検索と生成モデルを組み合わせた手法です。具体的には、まず関連する情報を外部データベースから取得し、その情報を基に自然言語生成を行います。これにより、より正確で具体的な回答やコンテンツを提供できるため、対話システムや質問応答システムなどでの応用が期待されています。

2:Felo 字幕・Felo 瞬訳

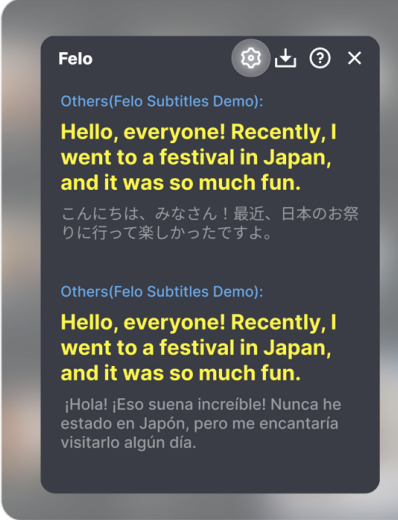


Felo 瞬訳・Felo 字幕は生成 AI を活用した自動同時通訳ツールです。15 の言語に対応しており、ユーザーの発言をリアルタイムで正確に翻訳を進めることが可能です。IOS および Android のスマートフォンアプリ、Chrome 拡張機能、Windows アプリケーションに対応しており、対面でのコミュニケーションから、オンライン会議、大人数の会議・イベントでの同時通訳など、外国語翻訳が必要になる様々なシーンでご活用頂けます。

Felo瞬訳 (スマホ版)



Felo字幕 (パソコン版)



1：高精度＋多言語の文字起こし技術

- Open AIの音声認識モデルである“Whisper”を独自手法でチューニング。日本語の高い文字起こし精度を実現。
- 話者の言語を自動認識、翻訳対象の言語に自動的に翻訳。話者の言語が変わっても、スムーズな会話を実現。

2：逐語訳＋修正翻訳で速さと正確性を担保

- 性質の異なる2種類のAIを駆使。会話の流れを途絶えさせない速度の逐語訳、翻訳の正確性を担保する修正翻訳の2度の翻訳により“同時通訳”を実現。

3：対面やオンライン会議等に対応

- 対面で海外の方と話す場面ではスマホのFelo瞬訳、オンライン会議で海外の方と話し場面ではパソコンのFelo字幕
- どのようなシーンでも各デバイスでご利用頂けます。

図 2. Felo 翻訳・Felo 字幕のサンプル画面

2.3 オンプレソリューション概要

本 GPTBase オンプレミスソリューションは、企業が自社内で高度な AI 機能を安全に活用できるよう設計しています。データのセキュリティを最優先とし、プライベートな環境での LLM と GPU 一体型システムによって、高度な処理能力を提供します。

基盤として採用している Llama3.1-70B モデルは、RAG システムにおいて 80.3%という高精度を達成しています。この精度は、市場で高く評価されている GPT-4o の 91.1%には及ばないものの、多くの実務用途では十分な性能です。

Sparticle では、従来の 4 ビット量子化技術を活用とこの基盤技術を一層強化すべくいくつか追加作業も実施いたしました。まず、KV キャッシュ (Key-Value Cache) ※³のさらなる効率的な量子化を行い、これによってデータアクセス速度が向上し、一貫したパフォーマンスが確保されました。また、新たなプレフィックスキャッシュ※⁴機能も導入し、大規模データ処理時にもスムーズなレスポンスを実現しました。加えて、多様なタスク間で最適化されたスケジューラーなど、複数の技術的改善策を講じることで、更なる効率性向上と資源利用最適化を図りました。その結果、モデルの vRAM 使用量を大幅に削減致しました。この成果により、高精度な RAG システムをわずか 2 枚の NVIDIA L40S 構成でも 100 qps※²及び 16k コンテキストサイズに対応を実現致しました。また、4 枚の NVIDIA L40S 構成では 300 qps※² (queries per second) 及び 16k コンテキストサイズという高負荷条件に対応可能です。



今後、さらに大型かつ高性能な Llama3.1 405B への量子化対応や、新たな混合量子化手法（4bit+16bit）の導入も視野に入れて開発が進められる予定です。この混合量子化アプローチでは、一部重要パラメーターのみ 16bit 表現することで、更なる精度向上と効率的資源利用とのバランス改善が期待できます。

こうした技術革新によって、お客様固有の業務要件や産業特性にも柔軟かつ迅速に応じることのできるカスタマイズ性・拡張性ある AI ソリューション(RAG だけでなく、他 APP 対応可能)をご提案いたします。また、自社サーバー内完結型なので外部依存なく安心して運営いただけます。

※2：QPS（queries per second）は、システムが 1 秒間に処理できるクエリ（問い合わせ）の数を示す指標です。この指標は、特にデータベースや検索エンジン、API サーバーなどのパフォーマンスを評価する際に用いられます。

※3：KV キャッシュ（Key-Value Cache）は、データをキーと値のペアとして保存し、高速なデータアクセスを実現するキャッシュシステムです。

※4：プレフィックスキャッシュは、キーに特定のプレフィックス（接頭辞）を付けることで、キャッシュデータを整理・管理する手法です。この方法は、特に大規模なキャッシュシステムや複数の異なるデータセットを扱う場合に有効です。

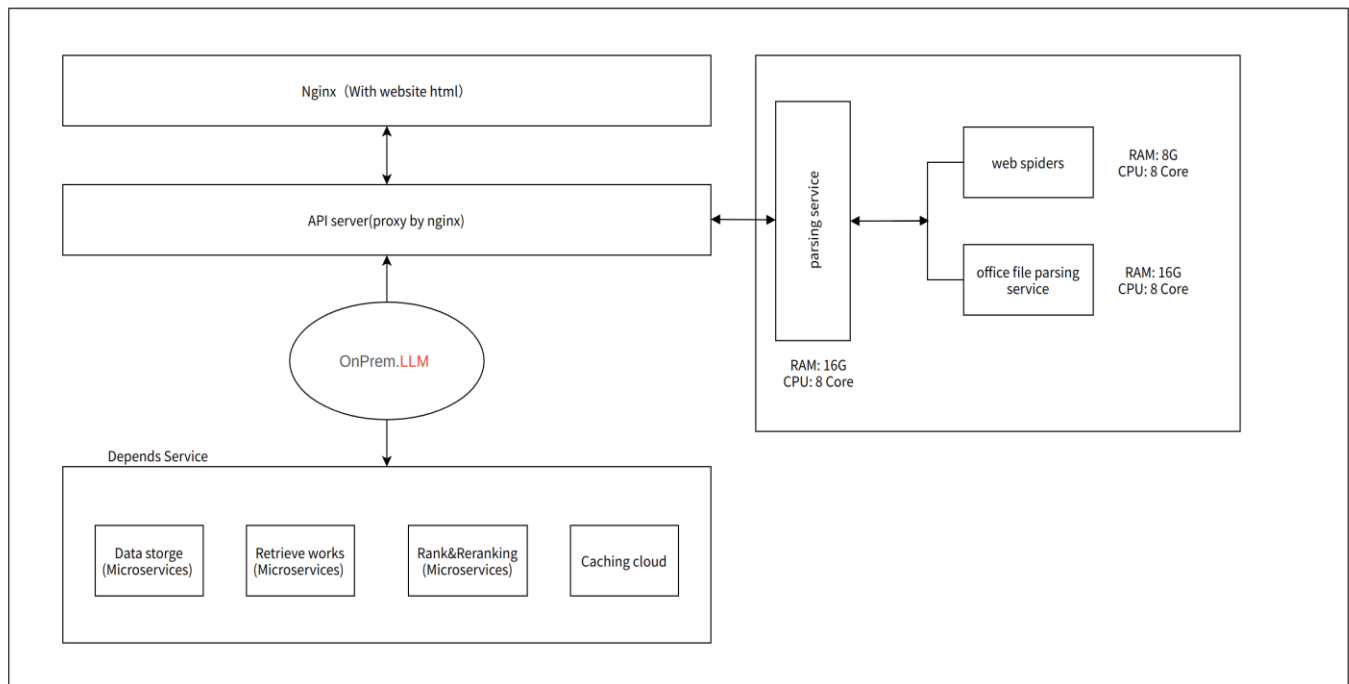


図 3. GPTBase(オンプレミス)のアーキテクチャー

3.0 ソリューションの使い方と精度観点

3.1 ソリューションの使い方

GPTBase の管理者側の操作方法としては以下の流れになります。

下記画像にて GPTBase のデータ学習方法、チャット利用方法、ダッシュボードの活用方法についてご紹介します。



1：ユーザー企業ごとのナレッジの作成

まずは、AI アシスタントを作成から新規ボットを作成していきます。

その際には企業さまごとに学習させたいデータ（URL、Word、Excel、PDF）をご用意いただき、GPTBase に学習させます。

• ボット作成方法

GPTBaseへのナレッジ学習

ログイン後、まず初めにマイボットから“①AIアシスタントを作成”をクリックします。
その後、右側の画面に遷移しますので、URLやファイルなどの学習を行っていきましょう。
※Git bookのデータ追加につきましてはご相談下さい。

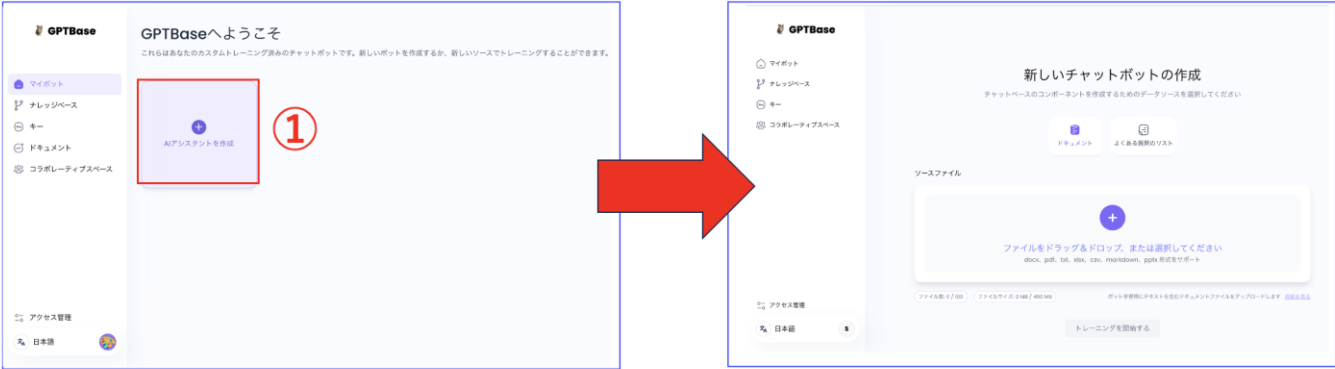


図 4. ユーザー企業ごとのナレッジの作成-ナレッジの学習設定①

• ボット作成方法

ドキュメントを学習させる場合

学習させたいデータを選択し学習を行います。
Word/Excel/CSV/PDFなど幅広いファイルに対応しています。
(PowerPointについてはご相談下さい。)

”①トレーニングを開始“をクリックし、終了後に表示される“チャット開始”をクリックで
利用が開始されます。開始後に学習データの編集も可能です。



図 5. ユーザー企業ごとのナレッジの作成-ナレッジの学習設定②

• ボット作成方法

よくある質問のリストを学習させる場合

よくある質問のリストをクリックします。
 ファイルを選択をクリックし、学習させたいQAリストを選択して下さい。
“①例となるExcelドキュメントをダウンロード”をクリックしダウンロード。
 QAリスト追加の際に、弊社にて推奨しているフォーマットをご確認下さい。
 その他のファイル・形式でも可能ですので必要に応じてお試しください。

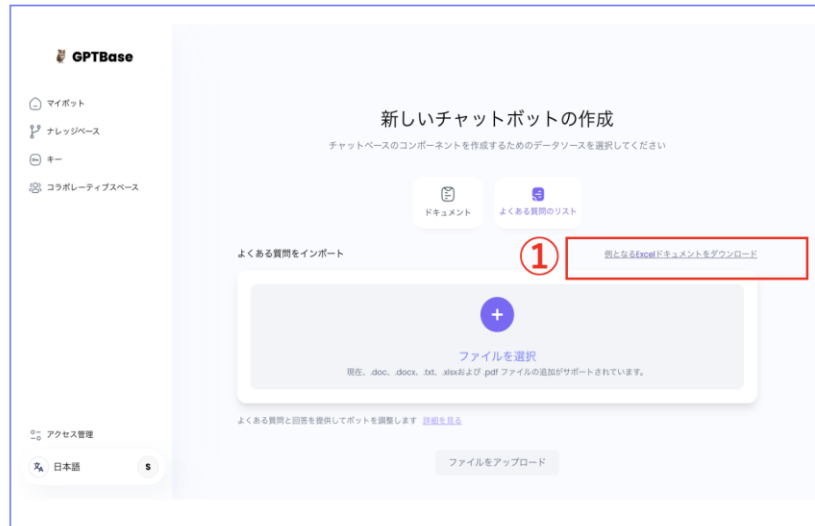


図 6. ユーザー企業ごとのナレッジの作成-ナレッジの学習設定③

• ボット管理手順

チャット

GPTBase作成後に、管理画面での管理を行なっていきます。
 学習させたデータについて、問題なく回答が返ってくるかどうか確認できます。
 企業様ごとに求める回答の方向性が異なりますので、ここで精度の確認をお願いします。
 精度が求める水準に達していない場合には、管理画面での精度改善を行しましょう。
 その他、“①共有”から管理者以外へのチャット共有などが実施可能です。



図 7. チャット内容の確認-チャット機能の確認①

2：チャット内容の確認

チャット画面にて問題なく学習させた内容に基づいた回答を出しているか、精度を確認します。

回答に問題がある場合にはチャットで質問した内容を、ダッシュボードにて直接 QA 設定することが可能です。

• ボット管理手順

チャット共有

共有をクリックすることで、下記チャットページの共有リンクを取得できます。
管理画面へのアクセスやHPへの埋め込みなしで、GPTBaseをご活用いただけます。



Powered by GPTbase.ai
© SPARTICLE.COM

Sparticle

図 8. チャット内容の確認-チャット機能の確認②

3：ダッシュボードの活用

ダッシュボードでは GPTBase が利用された件数や利用時間帯など、詳細な利用状況の確認ができます。

またチャット履歴からは、GPTBase が対応したチャット内容の確認、回答できなかった質問の QA 設定など、GPTBase の利用状況と修正作業を行うことができます。Sparticle での精度向上サポートを実施させていただく場合には、チャット履歴から回答内容のスコアリングを実施いただくとより効果的です。

• ポット管理手順

ダッシュボード



図 9. ダッシュボードの活用-記載内容

• ポット管理手順

ダッシュボード

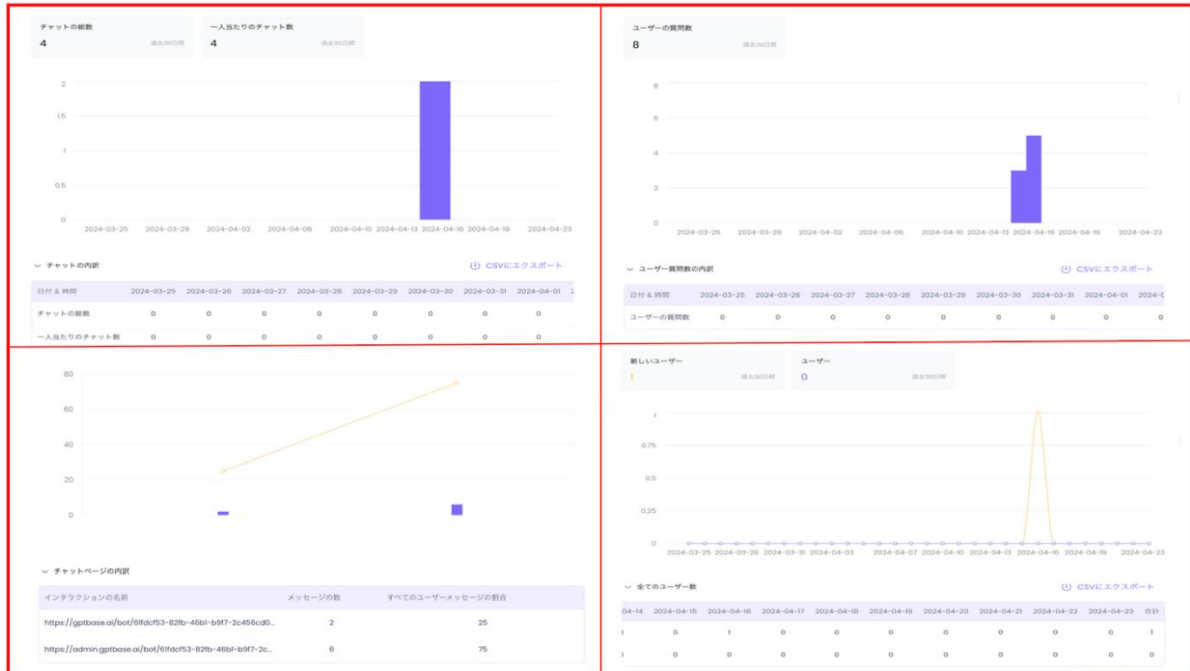


図 10. ダッシュボードの活用-グラフィメー



図 11. ダッシュボードの活用-返答修正

チャット履歴からチャットのログ確認やスコアリング、編集を行います。

チャットの総数や修正数、回答の有効率などが確認できますので、ここで GPTBase のチャットに関して管理し、精度の向上を行っていきましょう。チャット履歴より下記の項目を確認し、編集作業を行なって下さい。

- ①ユーザーの利用状況の確認
- ②質問と回答内容・状態の確認
- ③質問と回答内容の修正・編集



図 12. ダッシュボードの活用-ユーザーフィードバック

・ ポット管理手順

回答できなかった質問

- ・ 未回答の質問では、GPTBaseが回答できなかった質問一覧を確認できます。

チャット履歴

ユーザーフィードバック

未回答の質問

未回答の質問の記録

2024-04-09 - 2024-04-16

ログをエクスポート

最後のチャット時間	タイトル	タイプ	質問	評価理由	ユーザーフィードバックの詳細	アクション
2024-04-16 10:39:53	ここでは何が？	管理バックエンド	Q 株価について教えて？	-		
2024-04-15 16:48:45	株価について教えて・	管理バックエンド	Q 株価について教えて・	-		

合計 2 項目

<

1

>



図 13. ダッシュボードの活用-未回答の質問

・ ポット管理手順

スコアリング・編集

チャットのオリジナルテキスト

AIの応答をチャットログに注釈付けし、応答率を向上させます。

ユーザー 2024-04-16 10:39

ここでは何が？

AI応答 2024-04-16 10:39

私はSPARTICLE INC. Sparticle株式会社のサポートカスタマーチャットボットです。会社や製品に関する質問に答えたり、サポートを提供したりすることができます。また、製品やサービスに関する情報を提供することもできます。

ユーザー 2024-04-16 10:39

ニュースについて教えて？

AI応答 2024-04-16 10:39

申し訳ありませんが、SPARTICLE INC. Sparticle株式会社のお知らせに関する情報は、私の知識には含まれていません。他にお手伝いできることがございましたらお知らせください。

ユーザー 2024-04-16 10:40

製品の特徴は？

AI応答 2024-04-16 10:40

チャットの回答を評価してください

Q&Aを追加

参照ナレッジベースを選択する*

SPARTICLE INC. Sparticle株式会社の dataset

質問*

株価について教えて？

類似の質問 + 追加

類似の質問を生成

回答入力方法を選択*

☒ 直接回答入力

正しい答えを覚悟の場合は、直接入力してください。ロボットがこの回答を使ってユーザーの質問に答えます。

申し訳ありませんが、SPARTICLE INC. Sparticle株式会社のお知らせに関する情報は、私の知識には含まれていません。他にお手伝いできることがございましたらお知らせください。

☐ 回答源の引用

推奨される質問 + 追加



図 14. ダッシュボードの活用-スコアリング・編集

チャット履歴より、スコアリングをクリックすると下記の画面が表示されます。

チャットにカーソルを合わせ①「訂正」をクリックすると、右側の画面へと遷移します。



画面下部“②評価”を行うことで、Sparticle 社側にて GPTBase の精度向上にご評価を活用させていただきます。
右画面では直接、QA として質問の回答を追加することができます。

3.2 精度評価 開発 と利用ユースケース

最先端の研究開発に基づいた RAG 技術を駆使し、社内外の様々なデータやナレッジに基づいた回答を LLM に生成させることが可能です。社内外からの問い合わせに対応するヘルプデスクとしてや、専門的な情報を取り扱う社内のナレッジ管理などさまざまな利用シーンでご活用いただけます。

弊社 GPTBase は特に回答精度面での評価が高く、2024 年 6 月に実施した競合他社製品とのベンチマークでは 3 倍以上の正答率を記録、他社製品からの切り替えを含め、RAG 技術を活用したシステムの改善などのご要望を頂くこともございます。



図 15. GPTBase のベンチマーク結果

- 製品を使用するお客様から、製品の調子が悪いとマニュアルを読まず、すぐに連絡が来る。このようなユースケース時に、製品マニュアルや顧客対応マニュアル等を学習した AI が音声や LINE などに対応します。
- AI に学習させるマニュアルは日本語のみで可能です。かつ外国人・海外ユーザーが母国語で質問すると同言語で答えます。
- お客様がマニュアルを紛失した場合でも、本体貼り付けの QR コードを読み込むことでマニュアルを確認可能です。

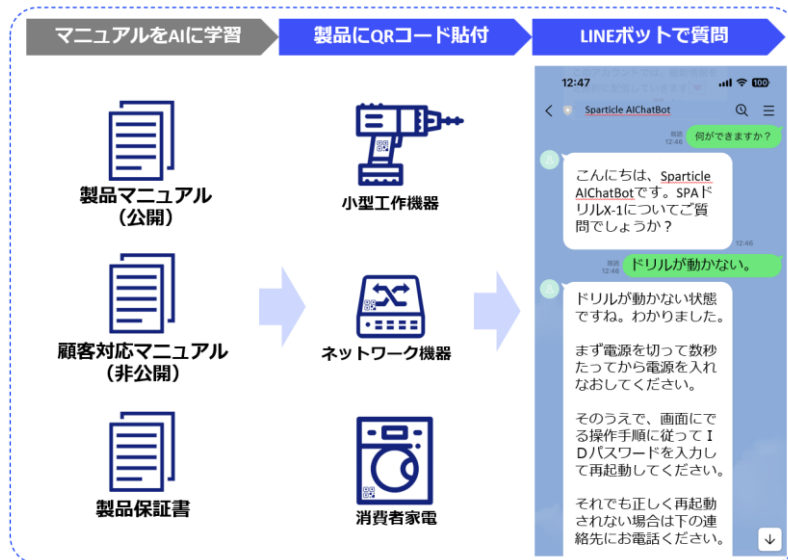


図 16. BTOC 製品のサポート、ヘルプデスクへの活用

- 保守メンテナンスのプロも、マニュアルどおりに問題を解決しているとは限りません。
- 優秀なスタッフの知識をマニュアルにプラス、それを AI が覚えておけば、ベテラン社員のナレッジも共有できます。
- 人による技量のバラツキを無くし、ベテラン社員の経験を次の世代に引き継ぐことができます。

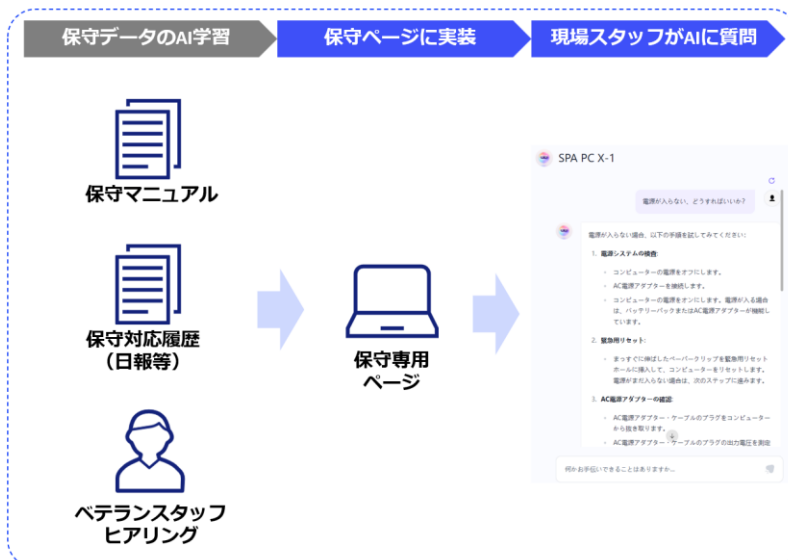


図 17. メーカー保守・点検での活用

- 展示会・美術館・テーマパークなどでは展示説明や音声ガイダンスによる一方的な来訪者へのコミュニケーションが中心だったものを、AI に展示品の説明や関連情報を事前に学習させることにより、来訪者の興味に応じて情報を提供することが可能になります。

- スマホをかざすと、他の言語に変換してくれるサービスがありますが、QR コードをスマホで読み込むことで、多くの情報を多言語対応で得ることができます。

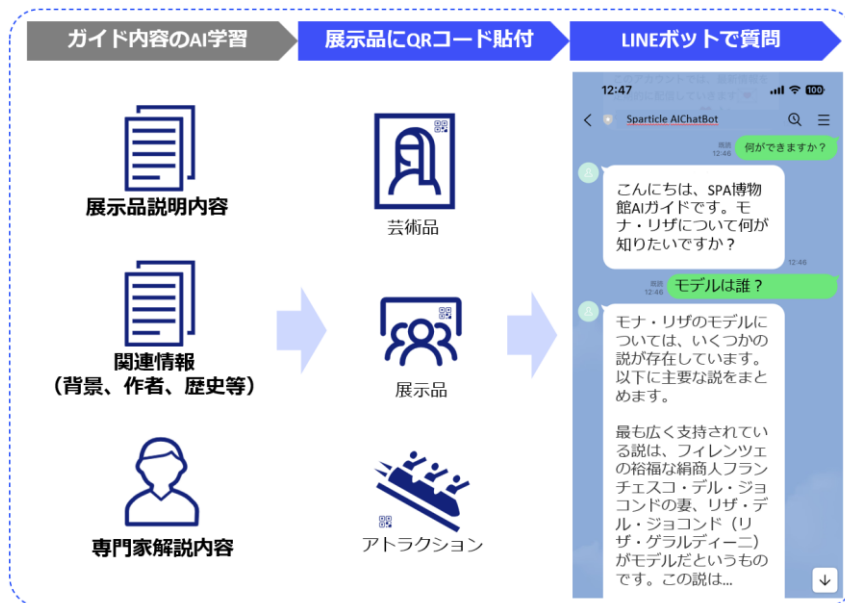


図 18 . 展示会・美術館・テーマパークにおける活用

- 若手社員が社内の情報を調べるときは、①知りたいことがどの資料に書いているか、②その資料はどこにあるか、③どのページに書いているか、④他に詳しく書かれた資料はないか等、社内の資料が散在している場合など非常に面倒で煩雑なプロセスが発生します。
- また年に1回しか発生しないイレギュラーな業務など、担当部署に問い合わせても確認に時間がかかるケースもあります。
- 社内の文章を AI に学習させることで、散在した情報から横断的に回答を作成。各個人がさまざまなプラットフォームで検索する手間が軽減できたり、先輩社員に遠慮なく新人は何度も質問したり、新しいことを学ぶことができます。

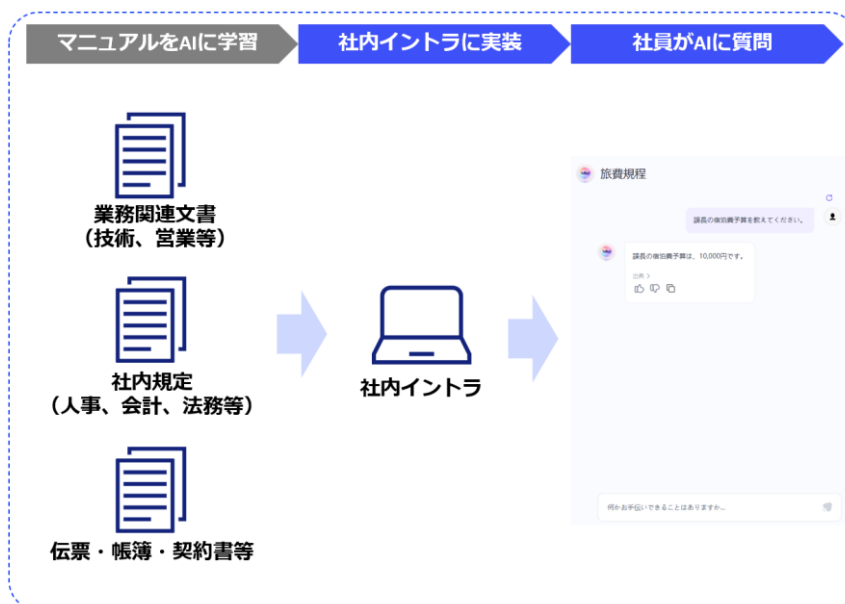


図 19. 社内における活用

- コールセンターではオペレーターの人手不足が深刻、採用や教育に多くのコストと工数を割かざるを得ず、省人化を伴う業務効率化が強く求められています。
- GPTBase では製品マニュアルや顧客対応マニュアルなどの資料に加え、過去の顧客対応履歴などから経験学習的に回答を生成、オペレーターからのフィードバックなどを基に、回答の幅と精度を向上
- AI による自動応答からオペレーターへの切り替えも設定可能、オペレーターは AI の受け答え履歴を確認した上で対応

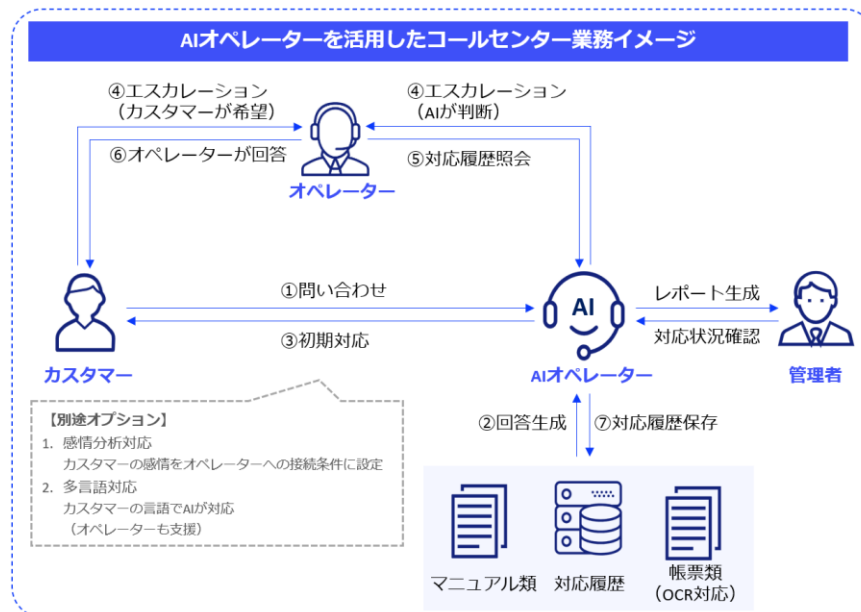


図 20. コールセンターにおける活用

3.3 複数ユーザー利用の観点

本セクションでは、生成 AI のオンプレミス対応による複数ユーザー同時利用の観点から、応答速度、リソース使用量、多重度、およびサイジングに関する考え方を整理いたします。企業が生成 AI を効果的に活用するためには、これらの要素が重要な指標となります。

1. 応答速度に着目した性能評価

生成 AI の応答速度は、ユーザー体験に直接影響します。複数のユーザーが同時に AI モデルを利用する場合、それぞれのリクエストが適切に処理され、迅速なレスポンスが求められます。以下の指標に基づいて性能評価を行います。

- **レスポンスタイム**: 各ユーザーからのリクエストに対する応答時間を測定します。一般的に、300 ミリ秒以下の応答が理想的とされ、ユーザーが快適に利用できる基準とされます。
- **レイテンシ**: システムへのリクエスト送信から応答を受け取るまでの時間を測定し、特にピーク時の遅延を分析します。複数ユーザー同時利用時のボトルネックを特定し、遅延を最小限に抑えるための改善策を検討します。
- **スループット**: 一定時間内に処理できるリクエスト数を示し、特にピーク時の処理能力を評価します。高いスループットを維持できることで、多数のユーザーが同時にシステムを使用する際の負荷に耐えることができます。

2. リソース使用量に基づいた多重度の考え方



生成 AI のオンプレミス展開では、システムのリソース（GPU、CPU、メモリ、ストレージなど）の適切な管理が重要です。多重度とは、一台のサーバーが同時に何人のユーザーリクエストを処理できるかを示します。以下の観点を考慮します。

- **メモリ使用量:** 各ユーザーリクエストにおけるメモリ消費量を測定し、複数ユーザー利用時に必要なメモリ量を算出します。一般的な AI モデルでは、リクエストごとに数百 MB から数 GB のメモリを消費することが多いです。
- **CPU 使用率:** 同時リクエストに対する CPU 負荷を分析し、最適なサーバー台数を決定します。CPU 使用率が 80%以上になると、高負荷状態となり、応答速度が低下する可能性があります。
- **GPU リソースの分割と管理:** 高性能な GPU を選定し、必要な計算能力を確保します。GPU リソースを効率的に分割し、複数のジョブを同時に実行できるようにします。

3. サイジングの参考になる目安

サイジングの際には、想定されるユーザー数やその利用頻度に基づいた具体的な基準を設けることが重要です。以下は目安として参考にできるポイントです。

- **小規模システム:** ユーザー数: 5~15 人 推定メモリ要件: 32GB~64GB 推定 CPU コア要件: 4~8 コア 量子化対応: モデルサイズを省メモリにするための動的量子化対応が推奨されます。
- **中規模システム:** ユーザー数: 20~50 人 推定メモリ要件: 128GB~256GB 推定 CPU コア要件: 8~16 コア 量子化対応: 静的量子化とモデル圧縮を適用することで、メモリ効率とレスポンス性能を向上させることが可能です。
- **大規模システム:** ユーザー数: 100 人以上 推定メモリ要件: 512GB~1TB 推定 CPU コア要件: 16 コア以上 量子化対応: より大規模なモデルに対して量子化を利用し、リアルタイム処理ができるように最適化します。特に、大規模データセットを扱う場合は、効率的なメモリ管理が求められます。

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 Sparticle 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な Superdome シリーズ、無停止コンピュータシリーズの Non Stop、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。



DXを支える、エッジからクラウドまでの包括的なハイブリッドインフラを実現
HPE の サーバーラインナップ

Hewlett Packard Enterprise

ラックマウント型サーバー  HPE ProLiant DL Server / RL Server	タワー型サーバー  HPE ProLiant ML Server / MicroServer	コンポーザブルシステム  HPE Synergy	エッジコンピューティング  HPE Edgeline
HPC & AI ソリューション  HPE Cray / HPE Apollo	ミッションクリティカル  HPE Superdome Flex / HPE NonStop	ハイパーコンバージド・インフラストラクチャ (HCI)  Hewlett Packard Enterprise, VMware, NUTANIX, Microsoft	

<https://www.hpe.com/jp/ja/hpe-proliant-servers>

図 21. 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ります。

最新プロセッサ搭載 HPE ProLiant サーバー

エッジからクラウドまでデータファーストモダナイゼーションを加速

Rack	1U	2U	Superdome /Scale-up Server
4 CPU		DL560 Gen11	8 CPU Superdome Flex 280
2 CPU	DL360 Gen11, DL365 Gen11	DL380 Gen11, DL380a Gen11, DL385 Gen11, DL384 Gen12	16 CPU Scale-up Server 3200
1 CPU	DL20 Gen11, DL320 Gen11, DL325 Gen11, DL110 Gen11, RL300 Gen11	DL345 Gen11	
Tower			Blade
1 CPU	MicroServer Gen11, ML30 Gen11, ML110 Gen11	2 CPU ML350 Gen11	2 CPU SY480 Gen10 Plus, SY480 Gen11
			Edge
			1 CPU EL8000, EL8000t

図 22. 最新プロセッサ搭載 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。

Compute engineered
for **your** hybrid world
Accelerate data-first modernization

“一歩先行くサーバー” HPE ProLiant Gen11



hpe.com/jp/gen11

直感的
クラウド型の運用管理

安心
セキュリティ・
バイ・デザイン

最適化
ワークロード性能

図 23. HPE ProLiant サーバーの特徴

Gen11 サーバーは GPU 搭載数を大きく向上

筐体設計を改良し、これまで物理的に難しかったサイズ・枚数を改善



GPU をサーバーの前面に搭載することで不可能を可能に

- GPU は発熱量が大きいため、むしろ前面に搭載した方が理にかなっている
– PCIe スロットを維持できるほか、前面に搭載しても、内蔵ディスクを 8 本搭載可能（1U モデルでは EDSFF を活用）
- 大量の電力を賄うためにパワーサプライを 2 → 4 個に強化（2U モデル）



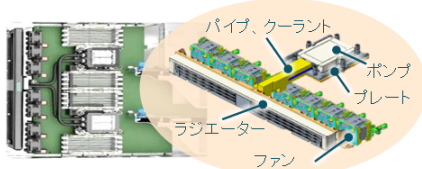
図 24. GPU 搭載枚数の拡充

最新サーバーの高性能がゆえの発熱量・消費電力を抑える工夫

電力効率のよい機種や筐体、パーツの選択

液冷ソリューションの “民主化”

液冷＋空冷で省電力



HPE スマートリキッドクーリング

- ✓ ハイブリッド型内部冷却方式 (Closed-loop Liquid Cooling)
- ✓ クルマのように、サーバー内部に小型ラジエーターを搭載し、空冷ファンによる風で冷却水を冷やす構造
- ✓ 特別な工事や外部ユニット不要

電源周りの効率化

PlatinumからTitaniumに変更することで、省電力



- ✓ 最大96%の高効率電源ユニット

- ✓ 500w ~ 2200w までの出力



- ✓ x4電源ユニット搭載可能 (x2 main board + x2 GPUs)
- ✓ Ex. DL380a / DL385

筐体サイズの最適化

1Uから2Uに変更することで、省電力



1Uサーバー



2Uサーバー



1Uサーバーの
ヒートシンク



2Uサーバーの
ヒートシンク

図 25. 発熱量・消費電力を抑える工夫

4.2 推奨システム構成

4.2.1 サーバースペック

Sparticle 社の GPTBase（オンプレミス）ソリューションをホストするサーバーとして、必要なスペック例を図 26 に示します。

	Model	Qty	Spec, Options, Notes
GPU サーバー	 HPE ProLiant DL380 Gen11	1	Model: HPE ProLiant DL380 Gen11 8SFF CPU: Intel Xeon-Gold 5415+ 2.9GHz 8-core 150W 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 x2 iLO: iLO Advanced ライセンス
GPU	 NVIDIA L40S	2	GPU: NVIDIA L40S 48GB PCIe Accelerator

図 26. Sparticle 社 GPTBase（オンプレミス）サンプル構成

DL380 は 2U サイズの 2CPU サーバーで世界で最も人気のあるサーバーモデルです。

Ada Lovelace アーキテクチャーの汎用万能型 GPU の NVIDIA L40S を搭載した構成です。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

ベアメタル環境

1.オペレーティングシステム

Linux(Ubuntu 20.04 LTS 推奨)

2.Python バージョン

Python3.8 以上

3.依存ライブラリ

CUDA 12.0 以上 (NVIDIA GPU 利用時)

PyTorch 1.10 以上

Transformers ライブラリ (Hugging Face)

Faiss (Facebook AI Similarity Search) ライブラリ

SQLAlchemy (データベース接続用)

4.メモリ要件

CPU メモリ 40GB 以上

コンテナ環境 (オプション)

1.オペレーティングシステム

Linux ディストリビューション (Ubuntu 20.04、CentOS 8 など)

コンテナランタイム環境 (Docker、Podman などのコンテナエンジンを推奨)

2.コンテナオーケストレーション

Kubernetes または Red Hat OpenShift (大規模デプロイメント用)

3.コンテナイメージ

Llama3.1-70B モデルと GPTBase(RAG)を含んだ専用コンテナイメージ (事前ビルドされたモデルと依存ライブラリを保持)

PyTorch 1.9 以上 (モデル実行用)

4.ハードウェア依存ライブラリ

CUDA 12.0 以上 (NVIDIA GPU 利用時)

cuDNN 8.2 以上

5.ネットワークとインフラ

モデル通信用の restful/API サービス

6.ロギングとモニタリング

Prometheus や Grafana (メトリクス収集と表示)

ELK スタック (OpenSearch、Logstash、Kibana)



4.2.3 ソリューションスタック概略図

本 GPTBase（オンプレミス）のアーキテクチャーの概略を開設します。大きく 5 つのコンポーネントをもとにしています。

1.LLAMA3.1-70B モデル

自然言語処理(NLP)を活用した生成 AI モデル。

テキスト生成、質問応答、サマリー生成などをサポート。

2.RETRIEVAL-AUGMENTED GENERATION (RAG)

クエリに対し関連情報を迅速に取り出し AI 生成結果を強化。

高精度の情報提供とユーザー体験向上を実現。

3.データストレージとインデックス

顧客データとドキュメントのための効率的なデータ管理システム。

ELASTICSEARCH などを用いたリアルタイム検索とインデックス機能。

4.フロントエンド処理

ユーザーフレンドリーなインターフェースを構築。

リクエストの受付、レスポンス表示、ユーザーインタラクション管理。

5.インフラとセキュリティ

HPE の高度なオンプレミスサーバーインフラを活用。

スケーラブルな計算リソース管理とデータ保護機能。

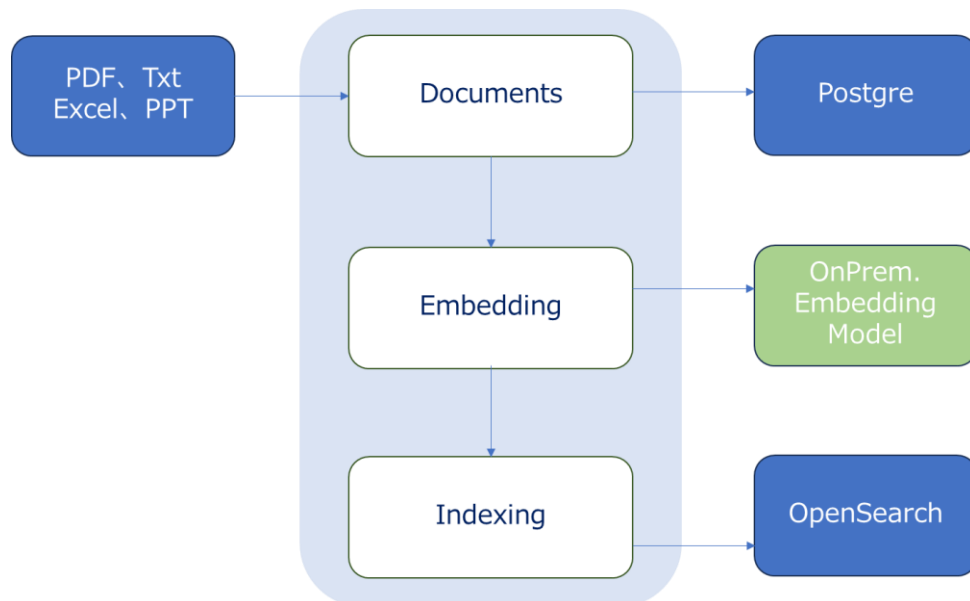


図 27. ソリューションスタック図（学習）



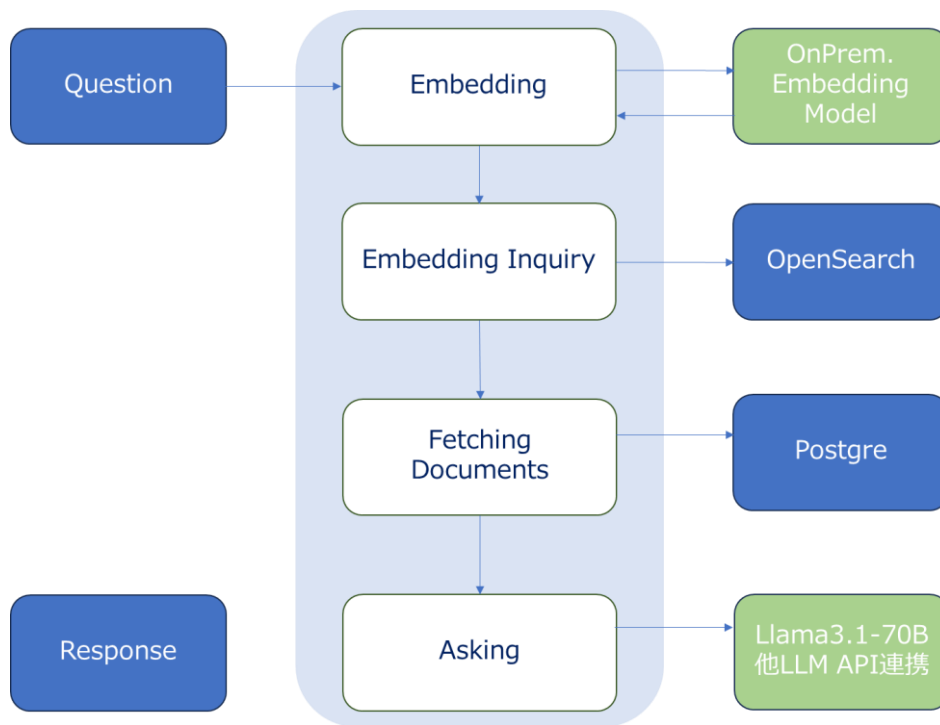


図 28. ソリューションスタック図（質問）

注記

2024 年 9 月の 本ドキュメントリリース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あります。

5.0 まとめ

HPE社とSparticle社は日本国内での生成AIのオンプレミス対応に関する協業を通じて、企業の競争力を向上させるための包括的なソリューションを提供することを目指しています。生成AIの導入により、企業は効率的なデータ分析、プロセス自動化、クリエイティブコンテンツ生成を実現し、迅速な意思決定を行える環境を整えることができます。オンプレミスでの運用はセキュリティやプライバシーの観点からも強固であり、特にデータがセンシティブな情報を含む業界では必要不可欠な選択肢です。HPE社の強力なインフラストラクチャとSparticle社の生成AIに関する専門知識を組み合わせることで、企業は統合されたソリューションをコスト効率良く得ることができます。

今後の展望

1. 継続的な技術革新

成長するAI市場に対応するため、新しい技術やアルゴリズムの研究開発を続け、生成AIの性能を向上させる取り組みを進めます

2. 業種別のカスタマイズ:

異なる業界や企業のニーズに応じたカスタマイズソリューションを提供することで、市場への適応力を高めます。例えば、製造業、金融業、ヘルスケアなど、各業界に特化した用途での生成AI活用法を模索し、成功事例を提供することを目指します。



6.0 その他のリソース

[Sparticle 社ホームページ](#)

[HPE ProLiant DL380 Gen11 サーバー](#)



お客様のニーズに最適な製品をお選びください。

HPE のプリセールススペシャリスト
にお問い合わせください。



Chat



Email



Call



Get updates