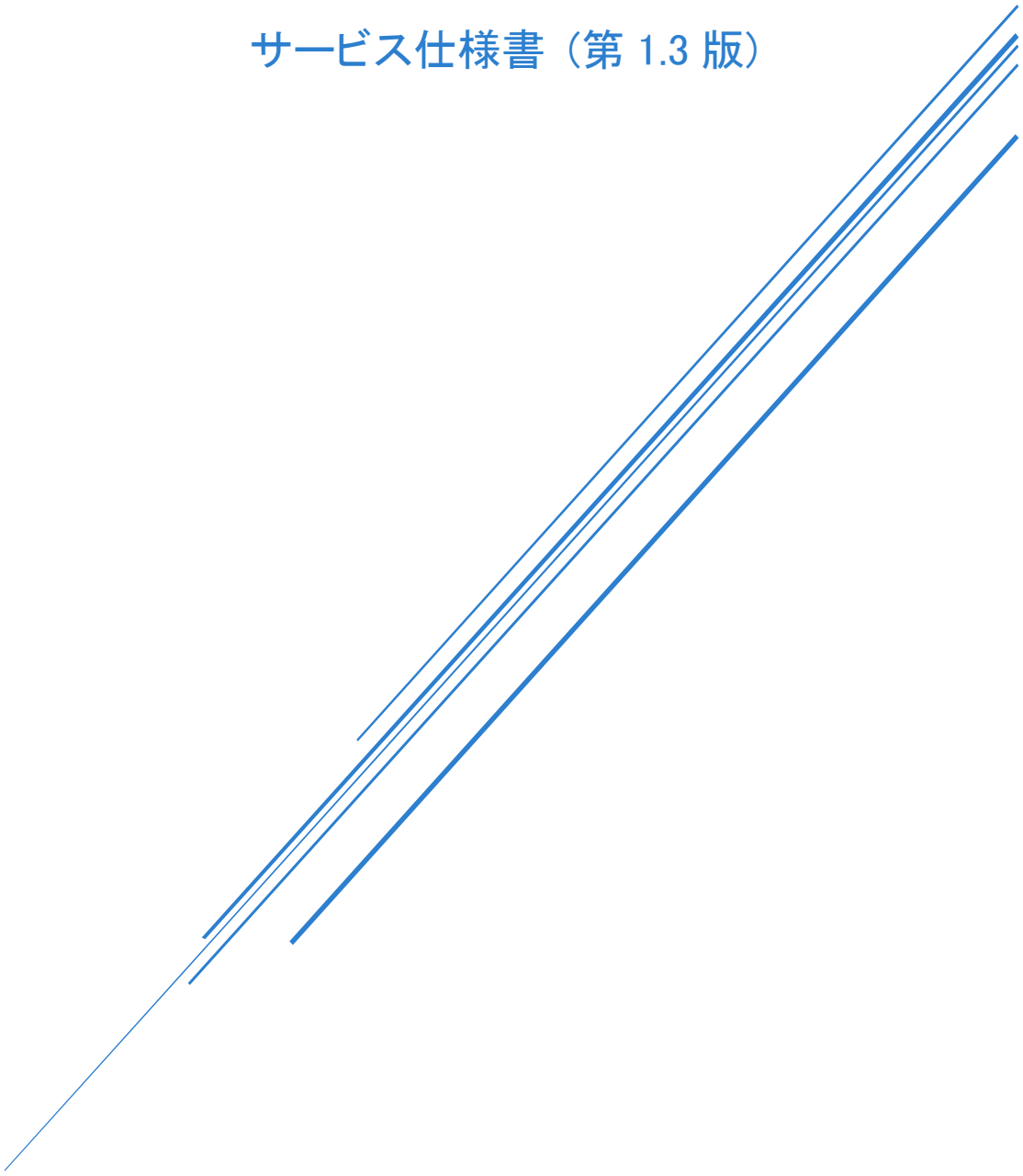
LLM/RAG シリーズ代行インストールサービス

サービス仕様書（第 1.3 版）



プロメテック・ソフトウェア株式会社

AI/HPC プラットフォーム事業開発本部

改版履歴

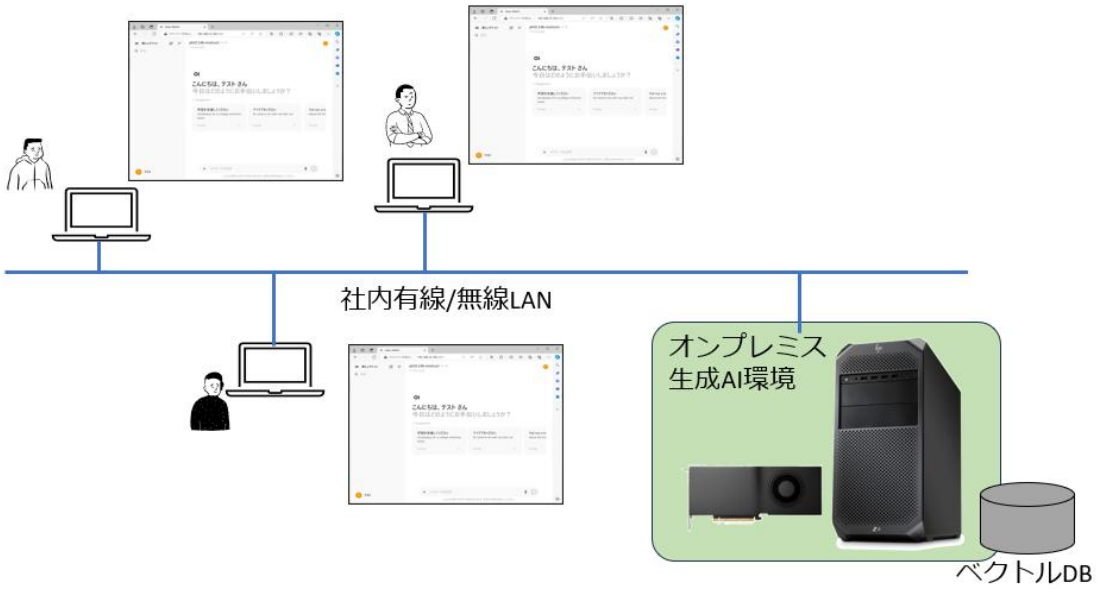
版数	作成日	作成者	内容
1.0	2025/3/11	プロメテック・ソフトウェア株式会社	新規
1.1	2025/5/12	プロメテック・ソフトウェア株式会社	モデルの削除
1.2	2025/5/23	プロメテック・ソフトウェア株式会社	モデルの削除
1.3	2025/7/11	プロメテック・ソフトウェア株式会社	トライアルサービスの追加に伴い、サービス総称を追加

内容

1. サービス概要

1.1 サービスの説明

LLM/RAG シリーズ代行インストールサービスは、当社が提供する「LLM/RAG トライアルサービス」、「LLM/RAG スタートアップサービス V2」および「LLM/RAG 業務活用サービス」の総称です。「LLM/RAG トライアルサービス」および「LLM/RAG スタートアップサービス V2」では、GPU サーバー/ワークステーションにオープンソースの生成 AI モデル/Embedding モデル/Reranking モデル/AI モデル実行ツール/Web ユーザーインターフェース・AI アプリケーション開発ツール等の構築を行います。一方、LLM/RAG 業務活用サービスでは、これらに加えて RAG 回答精度改善・運用支援ツール G-RAGon を AI アプリケーション開発ツールに設定します。社内 LAN に接続した GPU サーバー/ワークステーションに、Windows PC から Web ブラウザで接続することで、オンプレミス環境において生成 AI を活用した文章生成や、RAG の構築や RAG を活用した AI アプリケーションの検証を行うことが可能です。



1.2 主なシステム構成

（1）LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2 エントリー/ミッドレンジ

オンプレミス生成 AI 環境のソフトウェア構成

No	コンポーネント	詳細仕様
1	Web ユーザーインターフェース	Open WebUI https://openwebui.com/

2	AI アプリケーション開発ツール	Dify https://dify.ai/jp
3	AI モデル実行ツール	Ollama https://github.com/ollama/ollama
4		Xinference https://inference.readthedocs.io/en/latest/
5	生成 AI モデル	Ollama でサポートされているモデルをインストールします https://ollama.com/library
6	Embedding モデル	Ollama でサポートされているモデルをインストールします https://ollama.com/library
7	Reranking モデル	Xinference でサポートされているモデルをインストールします https://inference.readthedocs.io/en/latest/models/index.html

オンプレミス生成 AI 環境の前提となるソフトウェア構成

No	コンポーネント	詳細仕様
1	OS	Ubuntu https://jp.ubuntu.com/download
2	NVIDIA ドライバー	NVIDIA Accelerated Linux Graphics Driver https://www.nvidia.com/ja-jp/drivers/
3	Docker	Docker Engine - Community: https://docs.docker.com/engine/install/ubuntu/

上記の他、gcc、make、python3 などがインストールされています。

その他管理ソフトウェア構成

No	コンポーネント	詳細仕様
1	デスクトップ	GNOME デスクトップ https://www.gnome.org/
2	サーバー管理ソフトウェア	Cockpit https://cockpit-project.org/
3	アンチウィルススキャン	ClamAV https://www.clamav.net/

出荷時にアンチウイルススキャンソフトでフルスキャンし納入します。

（2）LLM/RAG 業務活用サービス　エントリー/ミッドレンジ

(1) の LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2 に下記のソフトウェアを追加したものになります。

No	コンポーネント	詳細仕様
1	RAG 回答精度改善・運用支援ツール	G-RAGon Chunking 自動最適化機能、Dify ナレッジバックアップ機能など

その他、G-RAGon の実行に必要な Python パッケージや samba などがインストールされています。

1.3　責任範囲

本サービスの責任範囲は下記の通りとなります。

No	レイヤー	責任範囲	責任
1	GPU サーバー／ワークステーションのオンプレミス生成 AI 環境の初期設定	AI アプリケーションフレームワークの代行インストール	当社
2	GPU サーバー／ワークステーション社内 LAN 接続環境	お客様社内 LAN への物理的接続	お客様
3		IP アドレス割り当て・設定	お客様
4	生成 AI 利用環境	利用者の Windows 端末・ブラウザ設定	お客様

1.4　品質保証制度(SLA)

「LLM/RAG トライアルサービス」、「LLM/RAG スタートアップサービス V2」および「LLM/RAG 業務活用サービス」では構築したソフトウェアについて、品質保証制度（SLA）を規定致しません。

本サービス仕様書に沿って、ソフトウェアのインストールを実施し、動作確認を実施します。構築したソフトウェアは主にオープンソースソフトウェアであり、お客様がそれぞれのオープンソースソフトウェアの利用条件をご了解頂いた上で本構築サービスを利用いただく必要があります。以下に主な利用条件を示します。

No	ソフトウェア	利用条件（ライセンス）
1	OS (Ubuntu)	https://ubuntu.com/legal/open-source-licences
2	NVIDIA ドライバー	NVIDIA ソフトウェアお客様使用ライセンス https://www.nvidia.com/ja-jp/drivers/nvidia-license/
3	Docker Engine	https://docs.docker.com/engine/#licensing
4	Web ユーザーインターフ	MIT ライセンス

	ェース(Open WebUI)	https://github.com/open-webui/open-webui/blob/main/LICENSE
5	AIアプリケーション開発ツール(Dify)	Dify オープンソースライセンス(Apache ライセンス 2.0) https://github.com/langgenius/dify/blob/main/LICENSE
6	AI モデル実行ツール(Ollama)	MIT ライセンス https://github.com/ollama/ollama/blob/main/LICENSE
7	AI モデル実行ツール(Xinference)	Apache ライセンス 2.0 https://github.com/xorbitsai/inference/blob/main/LICENSE
8	生成 AI モデル	Llama3 生成 AI モデルライセンス https://llama.meta.com/llama3/license/ Gemma 生成 AI モデルライセンス https://ai.google.dev/gemma/terms Phi4 生成 AI モデルライセンス https://ollama.com/library/phi4/blobs/fa8235e5b48f Qwen2.5 生成 AI モデルライセンス https://ollama.com/library/qwen2.5/blobs/832dd9e00a68 QwQ 生成 AI モデルライセンス https://ollama.com/library/qwq/blobs/b87250e4478f
9	Embedding モデル	nomic-embed-text Embedding モデルライセンス (Apache ライセンス 2.0) https://ollama.com/library/nomic-embed-text/blobs/c71d239df917
10	Rerank モデル	bge-reranker-v2-m3 Rerank モデルライセンス (Apache ライセンス 2.0) https://huggingface.co/BAAI/bge-reranker-v2-m3

一方、RAG 回答精度改善・運用支援ツール「G-RAGon(ジー・ラグ・オン)」およびパートナーアプリ「blueqat-RAG」はプロメテック・ソフトウェアおよび blueqat が独自に開発したツールです。これらのツールについては、ライセンサーの許可なく他のワークステーションやデバイスに複製、配布、再インストール、または再利用を禁止しています。また本ツールの複製、改変、再使用許諾、リバースエンジニアリング、逆アセンブル、逆コンパイル、その他ライセンサーの知的財産権を侵害するおそれのある行為を行わないでください。

2. サービス体系および詳細

2.1 サービス一覧

サービス名	サービス内容
LLM/RAG トライアルサービス LLM/RAG スタートアップサービス V2 LLM/RAG 業務活用サービス	LLM/RAG トライアルサービスおよび LLM/RAG スタートアップサービス V2 では、AI アプリケーション開発ツール、Web ユーザーインターフェース、AI モデル実行ツール、生成 AI モデル、Embedding モデル、Reranking モデルを GPU サーバー／ワークステーションにインストールし、基本的な設定を実施します 一方で、LLM/RAG 業務活用サービスでは、これらに加えて RAG 回答精度改善・運用支援ツールをインストールします
Q&A チケット	構築したオンプレミス生成 AI 環境の利用方法等について問い合わせをすることができます LLM/RAG スターターセット V2 および LLM/RAG 業務活用セットには Q&A チケットが 3 枚付属されています

2.2 サービス詳細

(1) LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2 エントリー

オンプレミス生成 AI 環境の前提となるソフトウェアとその他管理ソフトウェアのインストール・設定を行います。

作業項目	作業内容
OS インストール	Ubuntu OS をインストールします
オンプレミス生成 AI 環境の前提となるソフトウェアインストール	以下のソフトウェアのインストールを行います - NVIDIA ドライバー - Docker Engine
管理ソフトウェアインストール	以下のソフトウェアのインストールを行います - GNOME デスクトップ - サーバー管理ソフトウェア Cockpit - ウイルススキャンソフトウェア ClamAV

Ollama のホームページ https://github.com/ollama/ollama の手順に沿って AI モデル実行ツールである Ollama のインストール・設定を行います。

作業項目	作業内容
Ollama のインストール	Ollama の Docker コンテナをインストールします
Ollama の設定	以下の初期設定をおこないます - 生成 AI、Embedding で使用する GPU の設定 - 生成 AI モデル(LLM)、Embedding モデルの設定

以下の生成 AI モデルと Embedding モデルを利用できるように設定します。

生成 AI モデル(ABC 順)	説明
gemma2: 9B・Q4	Google 社が開発・公開している 90 億パラメタ・4bit 量子化モデル
llama3: 8B・Q4	Meta 社が開発・公開している 80 億パラメタ・4bit 量子化モデル
phi4: 14B・Q4	Microsoft 社が開発・公開している 140 億パラメタ・4bit 量子化モデル
Qwen2.5: 14B・Q4	アリババ社が開発・公開している 140 億パラメタ・4bit 量子化モデル

Embedding モデル	説明
nomic-embed-text	1.37 億パラメタの Embedding モデル

Xinference のホームページ https://inference.readthedocs.io/en/latest/の手順に沿って AI モデル実行ツールである Xinference のインストール・設定を行います。

作業項目	作業内容
Xinference のインストール	Xinference の Docker コンテナをインストールします
Xinference の設定	以下の初期設定をおこないます - Rerank で使用する GPU の設定 - Rerank モデルの設定

以下の Rerank モデルを利用できるように設定します。

Rerank モデル	説明
bge-reranker-v2-m3	多言語対応の軽量 Rerank モデル

Dify のホームページ https://dify.ai/jp の手順に沿って AI アプリケーション開発ツールである Dify のインストール・設定を行います。

作業項目	作業内容
Dify のインストール	Dify の Docker コンテナをインストールします

Dify の設定	以下の初期設定をおこないます - 管理者アカウントの登録 - 日本語、タイムゾーンの設定 - 生成 AI モデル(LLM)、Embedding モデル、Rerank モデルの設定 - サンプルデータの登録

Open WebUI のホームページ https://github.com/open-webui/open-webui の手順に沿って Web ユーザーインターフェースである Open WebUI のインストール・設定を行います。

作業項目	作業内容
Open WebUI のインストール	Open WebUI の Docker コンテナをインストールします
Open WebUI の設定	以下の初期設定をおこないます - 管理者アカウントの登録 - Web 画面の日本語設定 - Chat テンプレートの日本語設定 - 生成 AI モデル(LLM)の設定

利用方法について、弊社が作成したマニュアル類を提供します。

(2) LLM/RAG スタートアップサービス V2 ミッドレンジ

(1)LLM/RAG スタートアップサービス V2 エントリーに下記生成 AI モデルを追加したものになります。

生成 AI モデル(ABC 順)	説明
llama3: 70B・Q4	Meta 社が開発・公開している 700 億パラメタ・4bit 量子化モデル
gemma2: 27B・Q4	Google 社が開発・公開している 270 億パラメタ・4bit 量子化モデル
Qwen2.5: 32B・Q4	アリババ社が開発・公開している 320 億パラメタ・4bit 量子化モデル
QwQ: 32B・Q4	アリババ社が開発・公開している 320 億パラメタ・4bit 量子化モデル

(3) LLM/RAG 業務活用サービス エントリー

(1)LLM/RAG スタートアップサービス V2 エントリーに RAG 回答精度改善・運用支援ツール G-RAGon を AI アプリケーション開発ツールに設定します。

作業項目	作業内容
G-RAGon のインストール	G-RAGon の Docker コンテナをインストールします

G-RAGon の設定	Dify に G-RAGon を組み込みます

利用方法について、弊社が作成したマニュアル類を提供します。

(4) LLM/RAG 業務活用サービス ミッドレンジ

(2)LLM/RAG スタートアップサービス V2 ミッドレンジに RAG 回答精度改善・運用支援ツール G-RAGon を AI アプリケーション開発ツールに設定します。

作業項目	作業内容
G-RAGon のインストール	G-RAGon の Docker コンテナをインストールします
G-RAGon の設定	Dify に G-RAGon を組み込みます

利用方法について、弊社が作成したマニュアル類を提供します。

(5) LLM/RAG シリーズ Q&A チケット

LLM/RAG スタートアップサービス V2 および LLM/RAG 業務活用セットで構築した LLM 等(Ollama/Xinference/Open WebUI/Dify)の利用方法について、弊社が提供しているマニュアル類に記載された範囲についてのみ問い合わせをすることができます。

例:

- 利用者の登録方法
- 生成 AI モデルを追加したい
- ドキュメントをナレッジ登録したい

チケット 1 枚につき、1 つの疑問が解決するまで質問が可能です。ただし、異なる内容の質問と弊社で判断した場合には追加でチケットを消費したものとみなします。

Q&A チケットで質問した内容が、LLM/RAG スタートアップサービス V2 および LLM/RAG 業務活用セットの範囲外である場合、その旨ご回答し、Q&A チケットは消費されません。

以下、LLM/RAG シリーズ Q&A チケットの範疇外の例です。

例:

- ハードウェア/ドライバー等が原因と考えられる不具合
  例)サーバー/ワークステーションの電源が入らない、OS が起動しない

- LLM 等プレインストールしたソフトウェアの不具合
- お客様が追加でドライバー/ソフトウェアをインストールした場合の不具合
- Ubuntu の基本的な使い方に関する質問
- セキュリティパッチの適用方法およびセキュリティインシデント発生に関する質問

また、Q&A チケット 2 枚で 1 時間/1 回のコンサルテーションミーティングを受けられます。コンサルテーションミーティング実施前に確認したい項目をご連絡いただきます。

3. サービス利用のイメージ

(1) LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2、LLM/RAG 業務活用サービス（単独）のお申し込み手順

納入前				納入後
△お問合せフォーム お申込み	△ヒアリング実施 △お見積書のご提供	△申し込みの意思表明 △申込書のご提供	△構築作業実施 △検収作業 △サービスご提供	△問い合わせ

(2) LLM/RAG トライアルキット、LLM/RAG スターターセット、LLM/RAG 業務活用セットご購入時のお申し込み手順

納入前			納入後
△お見積書のご提供	△ヒアリング実施	△LLM/RAG シリーズセット納入	△問い合わせ

4. 禁止事項

利用者は、本サービスを利用して第三者または当社に損害を与える行為や、法律に違反する行為、公序良俗に反する行為等、ならびに「LLM/RAG シリーズ代行インストールサービス約款」に規定された禁止事項を行ってはならないものとします。

5. サービスに関するサポート窓口

LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2 および LLM/RAG 業務活用サービスのお問い合わせには、下記の 2 つの種類があります。

① LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2 および LLM/RAG 業務活用サービスの不具合に関するお問い合わせ

LLM/RAG トライアルサービス（もしくは LLM/RAG トライアルキット）、LLM/RAG スタートアップサービス V2（もしくは LLM/RAG スターターセット V2）および LLM/RAG 業務活用サービス（もしくは LLM/RAG 業務活用セット）でプレインストールした状態で、弊社が提供するマニュアル類の通りに動作しない場合など、明らかにサービスご提供時の不具合と考えられるもの

② LLM/RAG トライアルサービス、LLM/RAG スタートアップサービス V2 および LLM/RAG 業務活用サービスの使用方法

LLM 等のソフトウェアの使用方法など、LLM/RAG トライアルサービス（もしくは LLM/RAG トライアルキット）、LLM/RAG スタートアップサービス V2（もしくは LLM/RAG スターターセット V2）および LLM/RAG 業務活用サービス（もしくは LLM/RAG 業務活用セット）で弊社が提供するマニュアル類の内容に関するご質問に、Q&A チケットを使用して行うお問い合わせ

どちらも下記メールによるお問い合わせとなります。

(1) メールによるお問い合わせ

メールアドレス： llm_support@prometech.co.jp

受付時間帯： 平日（祝祭日、年末年始その他弊社サポート休業日を除く）10 時～17 時

問い合わせ件数はご契約に応じて以下の回数が条件となります。

契約内容	Q&A チケット利用した場合の問い合わせ回数上限
LLM/RAG スターターセット（Q&A チケット x3 付き）	3 回
LLM/RAG 業務活用セット（Q&A チケット x3 付き）	3 回

なお、LLM/RAG トライアルキットには Q&A チケットは付属されていません。

以上