

# Open Platform for Enterprise AI (OPEA) を 活用した生成 AI の利用基盤の構築

エクセルソフト株式会社  
2025年6月26日

# はじめに

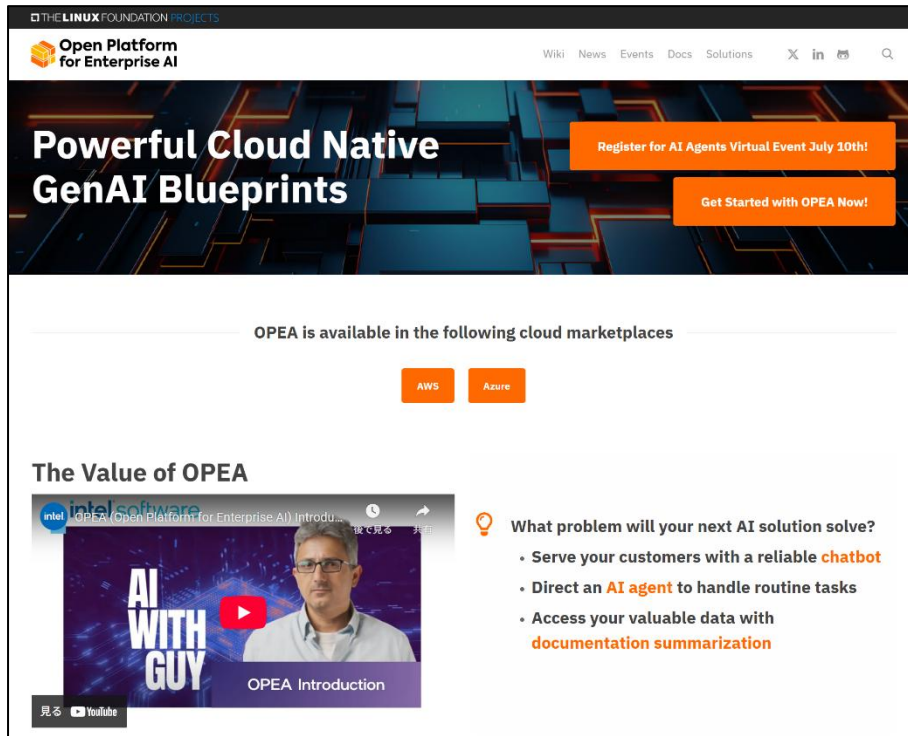
- 本資料は 2025年6月25日時点の情報を基にしています
  - ✓ 将来に変更される可能性があります

# ご紹介したいこと

- OPEA を利用した生成 AI 利用基盤の構築
  - ✓ Open Platform for Enterprise AI (OPEA) とは
  - ✓ OPEA アーキテクチャーについて
  - ✓ OPEA に基づいたチャットボットのサンプル

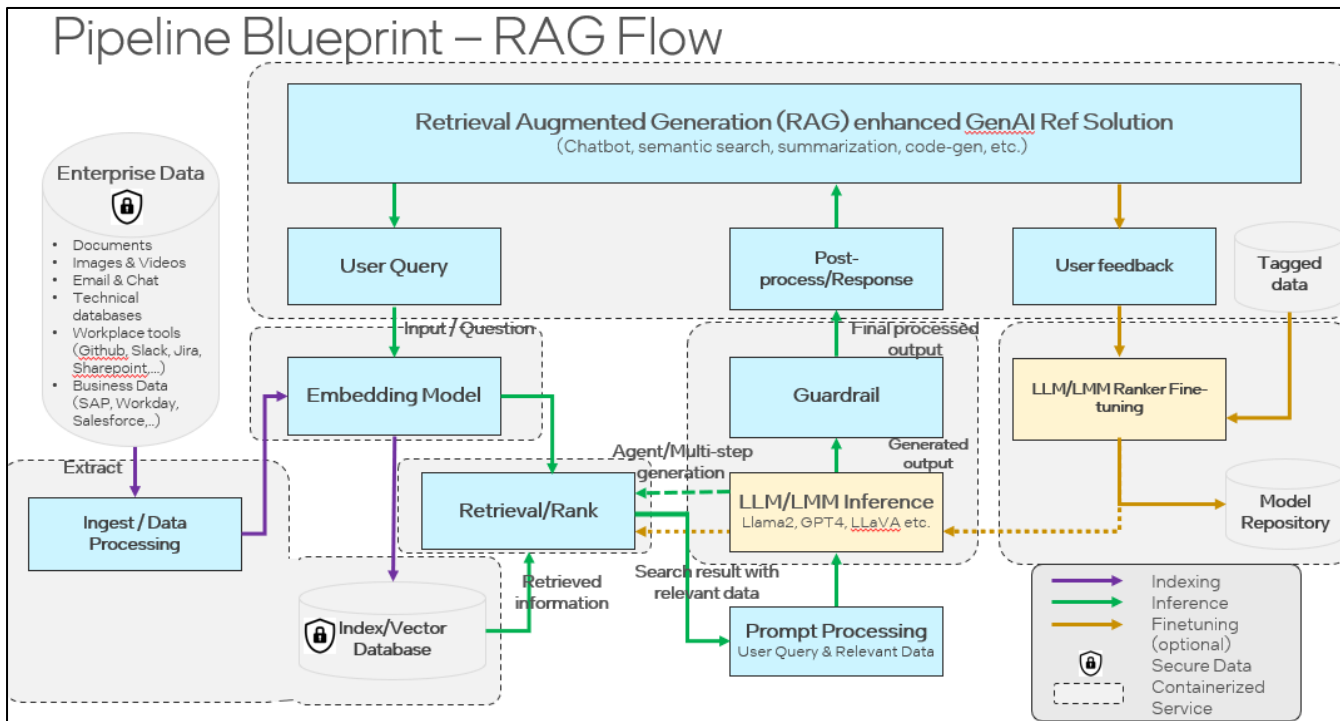
# Open Platform for Enterprise AI

- Open Platform for Enterprise AI
  - ✓ <https://opea.dev> (英語)
- LF AI & Data 傘下のプロジェクト
  - ✓ 2024年4月16日発表
- 複雑化しやすい生成 AI アプリケーションの実装を標準化するための取り組み
  - ✓ 現状では特に Agent AI や RAG の導入に向けた実装を整備しています



The screenshot shows the Open Platform for Enterprise AI website. The header includes the Linux Foundation Projects logo and navigation links for Wiki, News, Events, Docs, and Solutions. The main banner features the text "Powerful Cloud Native GenAI Blueprints" with two orange buttons: "Register for AI Agents Virtual Event July 10th!" and "Get Started with OPEA Now!". Below the banner, a section titled "OPEA is available in the following cloud marketplaces" displays buttons for AWS and Azure. The "The Value of OPEA" section includes a video player for "OPEA Introduction" by Intel Software, featuring a man speaking. To the right of the video, a lightbulb icon introduces a list of questions: "What problem will your next AI solution solve?" with bullet points: "Serve your customers with a reliable chatbot", "Direct an AI agent to handle routine tasks", and "Access your valuable data with documentation summarization".

# OPEA が提案する会社/組織での運用を想定した RAG を使った生成 AI アプリケーションのフロー

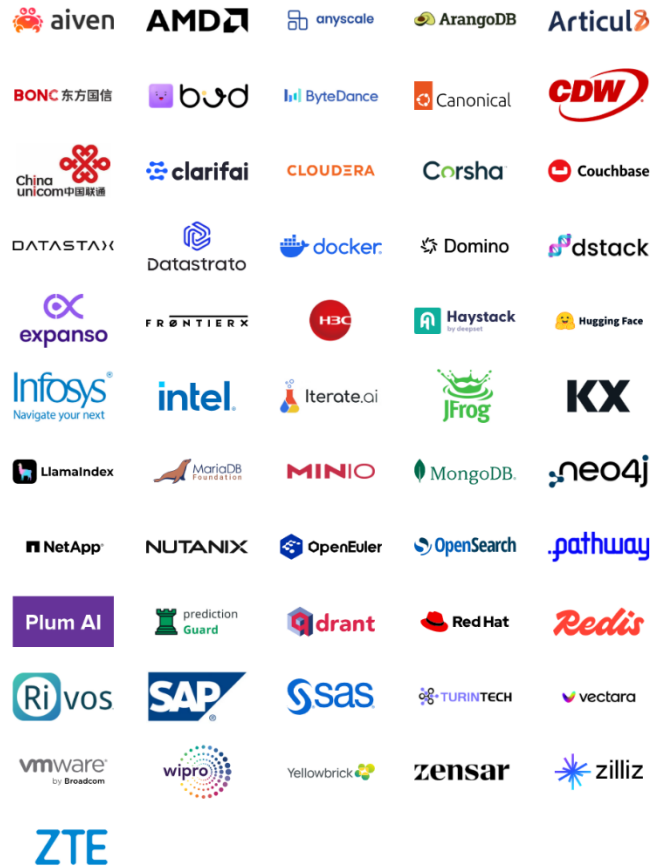


出典: [Open Platform For Enterprise AI](#) (英語)

# OPEA の主要な取り組みと コントリビューター

- 2025年6月時点のメンバー
- OPEA が取り組んでいるもの
  - ✓ エンドツーエンドのワークフローを  
提供するためのアーキテクチャーの整備
  - ✓ アーキテクチャーをもとにした  
生成 AI アプリケーションの  
リファレンス実装
  - ✓ 生成 AI アプリケーションを、  
「パフォーマンス、機能、信頼性、対応性」  
の面から評価するための仕組み作り

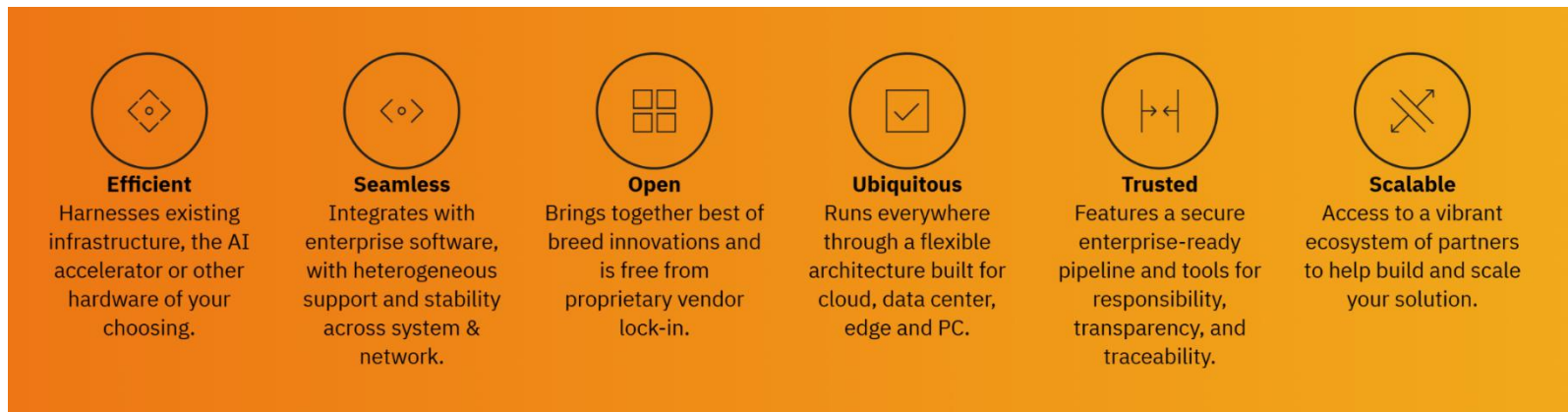
## OPEA Partners



出典: [Open Platform For Enterprise AI](#) (英語)

# OPEA の特長

- 既存の仕組み/ソフトウェアを活用します
- 柔軟な構造によりあらゆる環境で利用できます
- 高負荷時においてもパフォーマンスを維持するための拡張性



出典: [Open Platform For Enterprise AI](#) (英語)



# OPEA アーキテクチャー

## マイクロサービス、メガサービス、ゲートウェイ

- モジュラー式の構造を採用することで、  
実行環境に依存しないメンテナンス性の高いサービスの実装
- 3つの主要な構造により実現
  - ✓ Microservices  
生成 AI アプリケーションを構成するために基本的な要素を提供する小規模なサービス
  - ✓ Megaservices  
1つ以上のマイクロサービスで構築された包括的なソリューション
  - ✓ Gateways  
ユーザーがメガサービスにアクセスするためのインターフェイス  
ユーザー要件によってカスタマイズされます



# マイクロサービス

## ■ 各マイクロサービスはコンテナ化された自己完結型の小規模なサービス

### ✓ Embeddings Microservice

文字列をベクトルデータに変換

### ✓ Retrievers Microservice

ベクトルデータベース内の検索

### ✓ LLMs Microservice

LLM のデプロイ

### ✓ Rerankings Microservice

入力クエリーに対して関連性に基づいた優先付け

#### Microservices Table of Contents

- Agent Microservice
- Animation Microservice
- Asr Microservice
- Chathistory Microservice
- Cores Microservice
- Dataprep Microservice
- Embeddings Microservice
- Feedback\_management Microservice
- Finetuning Microservice
- Guardrails Microservice
- Image2image Microservice
- Image2video Microservice
- Language\_detection Microservice
- Llms Microservice
- Lvms Microservice
- Prompt\_registry Microservice
- Rerankings Microservice
- Retrievers Microservice
- Router Microservice
- Struct2graph Microservice
- Text2cypher Microservice
- Text2graph Microservice
- Text2image Microservice
- Text2kg Microservice
- Text2sql Microservice
- Third\_parties Microservice
- Tts Microservice
- Web\_retrievers Microservice

出典: [GenAI Microservices](#) (英語)

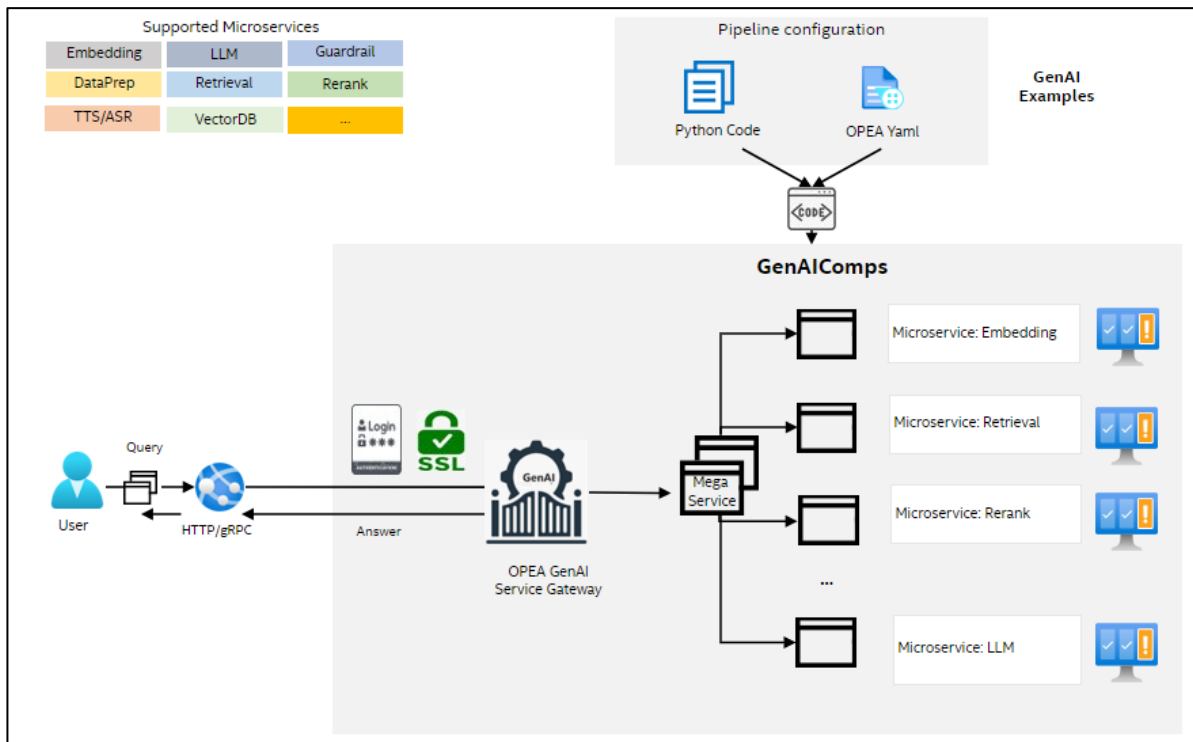
### ✓ Guardrails Microservice

LLM の出力に対して有害検出やバイアス検出を実施

### ✓ Fine-tuning Microservice

LLM は Instruction Tuning Reranking や Embedding モデル

# OPEA の構造を利用した生成 AI システム



出典: [opea-project/GenAIComps](https://opea-project.github.io/GenAIComps/) (英語)

# OPEA リポジトリ

■ Github : [OPEA \[Open Platform for Enterprise AI\]](#) (英語)

■ 例:

✓ GenAIComps

各種マイクロサービスの実装

Docker hub へ公開されています → [OPEA | Docker Hub](#) (英語)

✓ GenAIExamples

GenAIComps のマイクロサービスを利用したソリューション・レベルの  
サンプル・アプリケーション

チャットボット、AI アバター、音声応答など 23 種類を公開

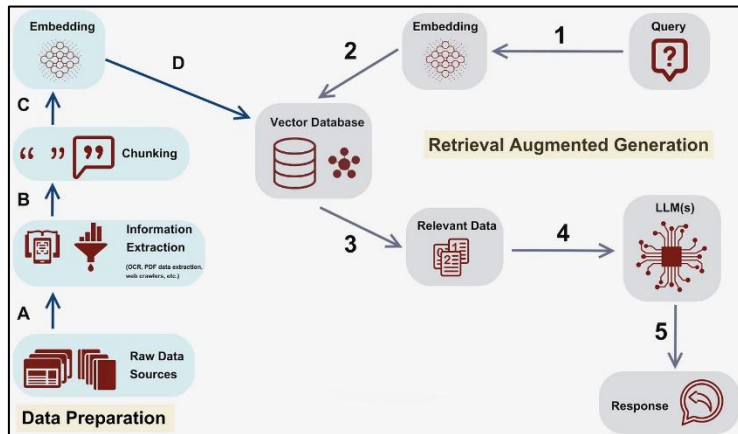
✓ GenAIEval

テキスト生成のモデルを評価するためのベンチマーク

# GenAExamples

## ChatQnA アプリケーション

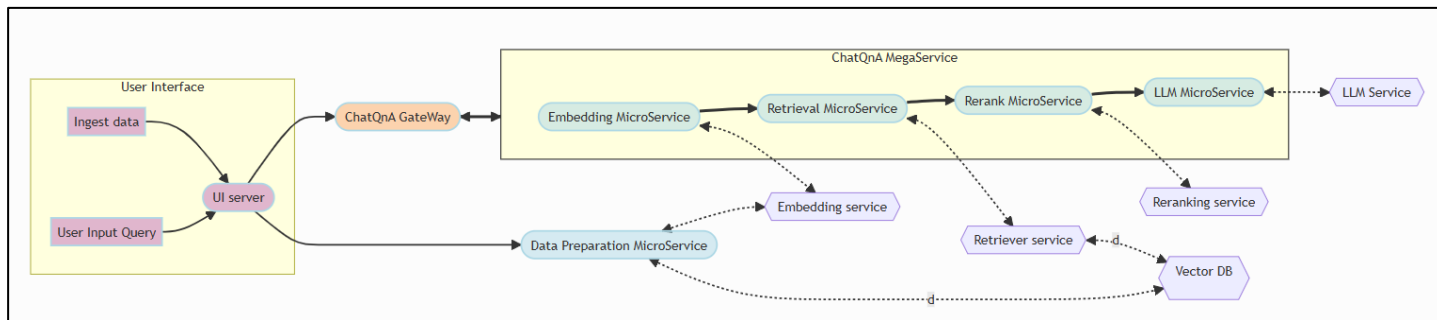
- RAG によるチャットボットを使った質問回答サービス
- インテル® Gaudi® AI アクセラレーター、  
インテル® Xeon® スケーラブル・プロセッサー、  
NVIDIA GPU、AMD GPU のサポート
- 各種クラウドサービス上で動作します
  - ✓ AWS、GCP、IBM Cloud、Azure、  
Oracle Cloud、ITAC



出典: [ChatQnA Application](#) (英語)

# ChatQnA フローチャート

- ChatQnA メガサービスは Embedding、Retrieval、Rerank、LLM マイクロサービスで構成
  - ✓ 個々のマイクロサービスは各サービス(機能)を提供
- RAG が参照するデータは DetaPrep マイクロサービスによりベクトル・データベースに登録



出典: [ChatQnA Application](#) (英語)

# ChatQnA の導入と実行

- ChatQnA を構成するマイクロサービスのイメージを作成、Docker Compose にてデプロイ、各マイクロサービスを統合するためのメガサービスを構築します
- デプロイオプション
  - ✓ Docker Compose、Kubernetes (Helm Chart の提供)
  - Terraform (AWS、Azure、GCP など)
- [Getting Started with OPEA](#) (英語) にて各クラウドサービス向けの構築手順を提供しています
- このセッションはインテル® Tiber™ AI クラウドのインスタンスを利用して ChatQnA をデプロイします

# ITAC インスタンスに ChatQnA を手動デプロイ

## 前提条件

1. Docker、Docker Compose のインストール
2. GenAIEamples リポジトリのクローン

```
> git clone https://github.com/opea-project/GenAIEamples.git
```



# ITAC インスタンスに ChatQnA を手動デプロイ

## 3. ディレクトリーの移動

```
cd GenAIEamples/ChatQnA/docker_compose/intel/cpu/xeon/
```

- ✓ このセッションではインテル® Gaudi® AI アクセラレーターを利用するため GenAIEamples/ChatQnA/docker\_compose/intel/hpu/gaudi/ にアクセスしています

## 4. 環境変数の設定

```
> source ./set_env.sh
```

- ✓ デフォルトでは Hugging Face Hub に公開されているモデルを使用します
- ✓ 対話式となっておりモデルは任意に入力できます
  - EMBEDDING\_MODEL\_ID → BAAI/bge-base-en-v1.5 (デフォルト)
  - RERANK\_MODEL\_ID → BAAI/bge-reranker-base (デフォルト)
  - LLM\_MODEL\_ID → meta-llama/Llama-3.1-8B-Instruct

# ITAC インスタンスに ChatQnA を手動デプロイ

## 5. 各サービスを立ち上げて ChatQnA サービスを起動

```
> docker compose up -d
> docker ps -a
```

```
sdp@gaudi3-01:~/GenAIEamples/ChatQnA/docker_compose/intel/hpu/gaudi$ docker ps -a
```

CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS	NAMES
a1df5ae77ddd	opea/nginx:latest	"/docker-entrypoint..."	20 minutes ago	Up 3 minutes	0.0.0.0:80->80/tcp, [::]:80->80/tcp	chatqna-gaudi-nginx-server
fd7b5bcc7fbc	opea/chatqna-ui:latest	"docker-entrypoint.s..."	20 minutes ago	Up 3 minutes	0.0.0.0:5173->5173/tcp, [::]:5173->5173/tcp	chatqna-gaudi-ui-server
2ef07d23c2c1	opea/chatqna:latest	"bash entrypoint.sh"	20 minutes ago	Up 3 minutes	0.0.0.0:8888->8888/tcp, [::]:8888->8888/tcp	chatqna-gaudi-backend-server
30f18a489cad	opea/retriever:latest	"python opea_retriev..."	20 minutes ago	Up 20 minutes	0.0.0.0:7000->7000/tcp, [::]:7000->7000/tcp	retriever-redis-server
a93b0db4f4e6	opea/dataprep:latest	"sh -c 'python \$( [ ..."	20 minutes ago	Up 20 minutes (healthy)	0.0.0.0:6007->5000/tcp, [::]:6007->5000/tcp	dataprep-redis-server
5d6f35d51499	ghcr.io/huggingface/tei-gaudi:1.5.0	"text-embeddings-rou..."	20 minutes ago	Up 20 minutes	0.0.0.0:8808->80/tcp, [::]:8808->80/tcp	tei-reranking-gaudi-server
752598887cf0	opea/vllm-gaudi:latest	"python3 -m vllm.ent..."	20 minutes ago	Up 20 minutes (healthy)	0.0.0.0:8007->80/tcp, [::]:8007->80/tcp	vllm-gaudi-server
0a18b549c036	ghcr.io/huggingface/text-embeddings-inference:cpu-1.5	"text-embeddings-rou..."	20 minutes ago	Up 20 minutes	0.0.0.0:8090->80/tcp, [::]:8090->80/tcp	tei-embedding-gaudi-server
834945a6c9e5	redis/redis-stack:7.2.0-v9	"/entrypoint.sh"	20 minutes ago	Up 20 minutes (healthy)	0.0.0.0:6379->6379/tcp, [::]:6379->6379/tcp, 0.0.0.0:8001->8001/tcp, [::]:8001->8001/tcp	redis-vector-db

# チャットボットとの会話

- ブラウザーから ChatQnA にアクセスして OPEA について質問します

OPEA とは何ですか

OPEAとは、オペレーティング・プライベート・エンタープライズ・アーキテクチャ（Operating Private Enterprise Architecture）です。

# チャットボットとの会話

## ■ OPEA の説明が書かれた PDF ファイルをアップロードして再度質問します

✓ [what\\_is\\_opea.pdf](#)

Getting Started の手順内にダウンロード・リンクがあります

OPEA とは何か

OPEA (Open Platform for Enterprise AI)は、企業向けのオープンなプラットフォームです。AIソリューションの作成と評価を支援するフレームワークです。

Latest » OPEA Overview

### OPEA Overview

OPEA (Open Platform for Enterprise AI) is a framework that enables the creation and evaluation of open, multi-provider, robust, and composable generative AI (GenAI) solutions. It harnesses the best innovations across the ecosystem while keeping enterprise-level needs front and center.

OPEA simplifies the implementation of enterprise-grade composite GenAI solutions, starting with a focus on Retrieval Augmented Generative AI (RAG). The platform is designed to facilitate efficient integration of secure, performant, and cost-effective GenAI workflows into business systems and manage its deployments, leading to quicker GenAI adoption and business value.

The OPEA platform includes:

- Detailed framework of composable microservices building blocks for state-of-the-art GenAI systems including LLMs, data stores, and prompt engines
- Architectural blueprints of retrieval-augmented GenAI component stack structure and end-to-end workflows
- Multiple micro- and megaservices to get your GenAI into production and deployed
- A four-step assessment for grading GenAI systems around performance, features, trustworthiness and enterprise-grade readiness

### OPEA Project Architecture

OPEA uses microservices to create high-quality GenAI applications for enterprises, simplifying the scaling and deployment process for production. These microservices leverage a service composer that assembles them into a megaservice thereby creating real-world Enterprise AI applications.

### Microservices: Flexible and Scalable Architecture

The GenAI Microservices documentation describes a suite of microservices. Each microservice is designed to perform a specific function or task within the application architecture. By breaking down the system into these smaller, self-contained services, microservices promote modularity, flexibility, and scalability. This modular approach allows developers to independently develop, deploy, and scale individual components of the application, making it easier to maintain and evolve over time. All of the microservices are containerized, allowing cloud native deployment.

# まとめ

- OPEA は複雑化しやすい生成 AI アプリケーションの実装を簡素化することを目的としています
  - ✓ 現時点では RAG に注目した実装や整備が行われています
- OPEA はマイクロサービスによるモジュラー式を採用して環境に依存しないアプリケーションを構築します

GenAIEamples リポジトリにはチャットボット以外にも多くのサンプルが公開されています  
ぜひお試しください

[Examples - OPEA documentation](#) (英語)

お問い合わせはこちらまで  
<https://www.xlsoft.com/jp/qa>

Intel、インテル、Intel ロゴ、その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。

\* その他の社名、製品名などは一般に各社の表示、商標または登録商標です。

製品および性能に関する情報: 性能は、使用状況、構成、その他の要因によって異なります。詳細については、<http://www.intel.com/PerformanceIndex/> (英語) を参照してください。

© 2025 Intel Corporation. 無断での引用、転載を禁じます。

XLsoft のロゴ、XLsoft は XLsoft Corporation の商標です。Copyright © 2025 XLsoft Corporation.