

インテル® Tiber™ AI クラウドと 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1月17日

はじめに

- ご紹介内容は 2025/01/09 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 と ITAC を利用した生成 AI 利用基盤の構築
 - ✓ Open Platform for Enterprise AI (OPEA) とは
 - ✓ インテル® Tiber™ AI クラウドと OPEA の利用
 - ✓ 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

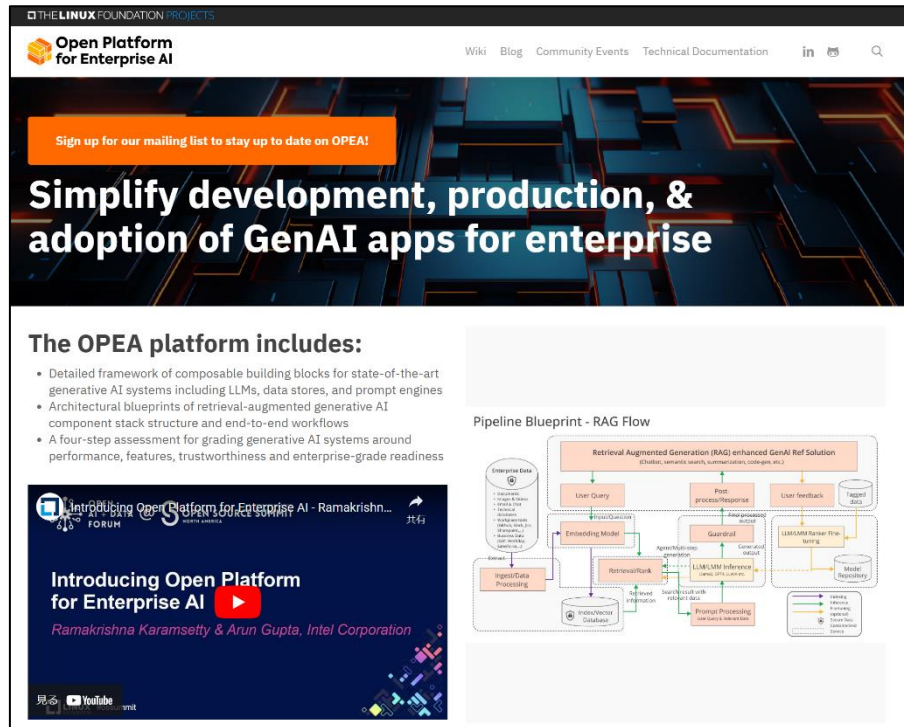
✓ <https://opea.dev> (英語)

■ 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 シンプルにする取り組み

✓ 現状では特に RAG の導入に向けた
実装を整備しています



The screenshot shows the website for the Open Platform for Enterprise AI, a project under The Linux Foundation. The main headline reads "Simplify development, production, & adoption of GenAI apps for enterprise". Below this, it lists the platform's features: a detailed framework of composable building blocks for state-of-the-art generative AI systems, architectural blueprints for retrieval-augmented generative AI, and a four-step assessment for grading generative AI systems. A video thumbnail titled "Introducing Open Platform for Enterprise AI" is also visible.

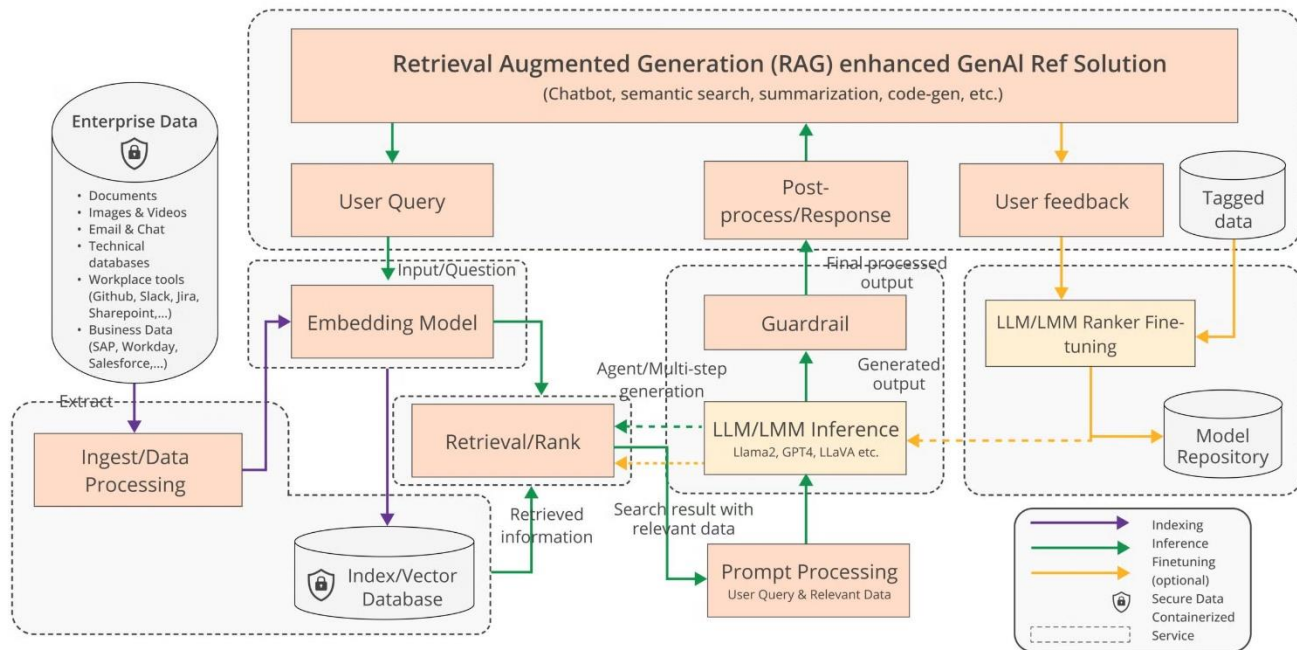
Pipeline Blueprint - RAG Flow

The diagram illustrates the RAG flow process:

- Input/Query Processing:** User Query is processed into a User Query.
- Retrieval:** The query is used to search a Vector Store (Index/Vector Database) and a Knowledge Base (Retrieval/Vector Store).
- Generation:** The retrieved information is processed by an LLM/LLM Inference (GenAI) component, which also receives input from an Embedding Model and a Prompt Processor.
- Output:** The final output is a User Feedback loop, which is used to refine the system.

例: 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

Pipeline Blueprint - RAG Flow



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

OPEA プラットフォームと コントリビューター

- 2025年1月時点のメンバー
- OPEA が取り組んでいるもの
 - ✓ 生成 AI システムのためのフレームワークの実装
 - ✓ 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
 - ✓ 生成 AI システムをパフォーマンス、機能、信頼性、対応性で評価するための仕組み

Contribute to OPEA

We invite like-minded industry peers to contribute to the development and standardization of enterprise-grade Retrieval Augmented Generative AI.



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



Efficient

Harnesses existing infrastructure, the AI accelerator or other hardware of your choosing.



Seamless

Integrates with enterprise software, with heterogeneous support and stability across system & network.



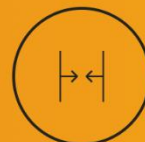
Open

Brings together best of breed innovations and is free from proprietary vendor lock-in.



Ubiquitous

Runs everywhere through a flexible architecture built for cloud, data center, edge and PC.



Trusted

Features a secure enterprise-ready pipeline and tools for responsibility, transparency, and traceability.



Scalable

Access to a vibrant ecosystem of partners to help build and scale your solution.

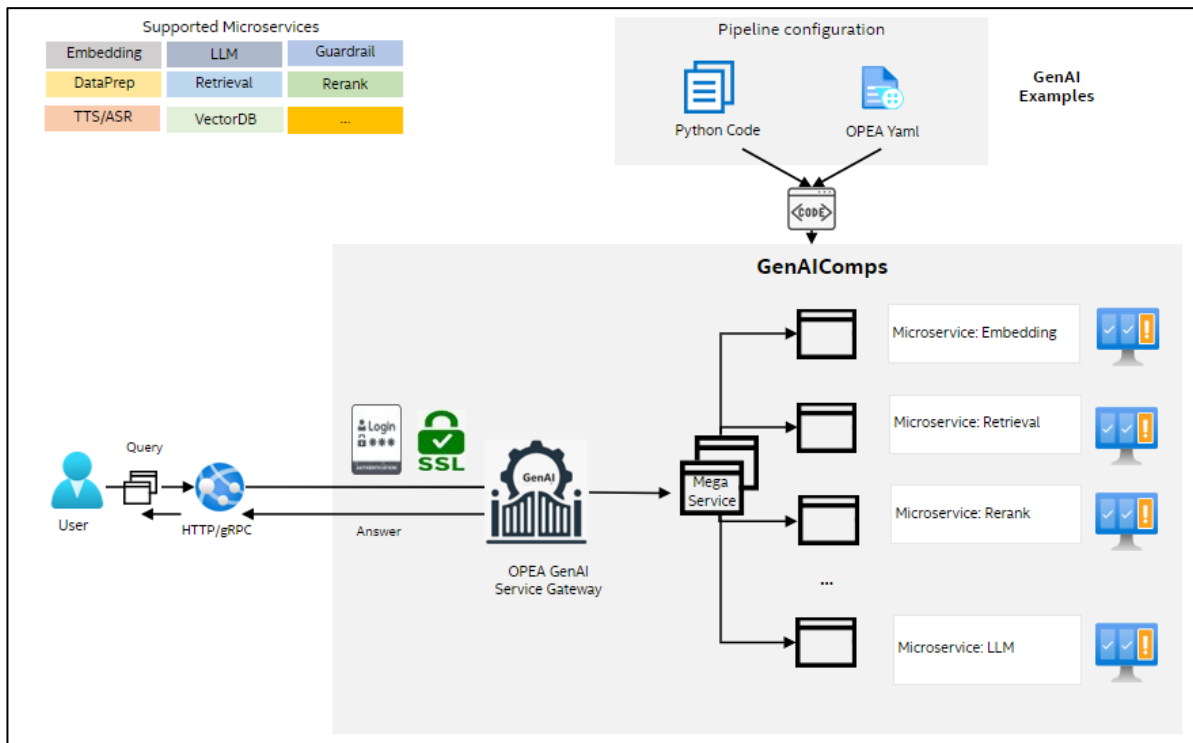
出典: [Open Platform For Enterprise AI](#) (英語)

OPEA の構造

メガサービス、マイクロサービス、ゲートウェイ

- モジュール式の構造を採用することで、
実行環境に依存せず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
 - ✓ 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
 - ✓ Microservices
生成 AI システム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
 - ✓ 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

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)

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ており 自己完結型の小規模なサービスに分解

✓ 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

✓ Guardrails Microservice

LLM の出力に対して有害検出や
バイアス検出を実施

✓ Avatar Animation Microservice

音声と唇の動きを同期した顔画像のビデオ
の生成

✓ Chathistory Microservice

チャットボットとの会話を保存、取得、管理

Microservices Table of Contents

- 3rd_parties Microservice
- 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
- Intent_detection Microservice
- Lims Microservice
- Lvms Microservice
- Prompt_registry Microservice
- Ragas Microservice
- Rerankings Microservice
- Retrievers Microservice
- Text2image Microservice
- Text2sql Microservice
- Tts Microservice
- Vectorstores Microservice
- Web_retrievers Microservice

出典: [GenAI Microservices — OPEA 1.1 documentation](#) (英語)

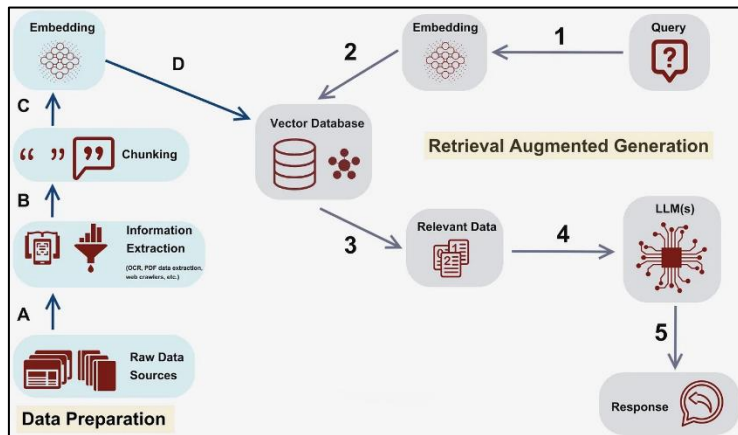
OPEA リポジトリ

- Github : [OPEA \[Open Platform for Enterprise AI\]](#) (英語)
 - ✓ GenAIComps (Generative AI Components)
各種マイクロサービスの実装
 - ✓ GenAIExamples
GenAIComps のコンポーネントを利用したソリューション・レベルのサンプル
チャットボット、AI アバター、音声応答など 11 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

GenAI Examples

ChatQnA アプリケーション

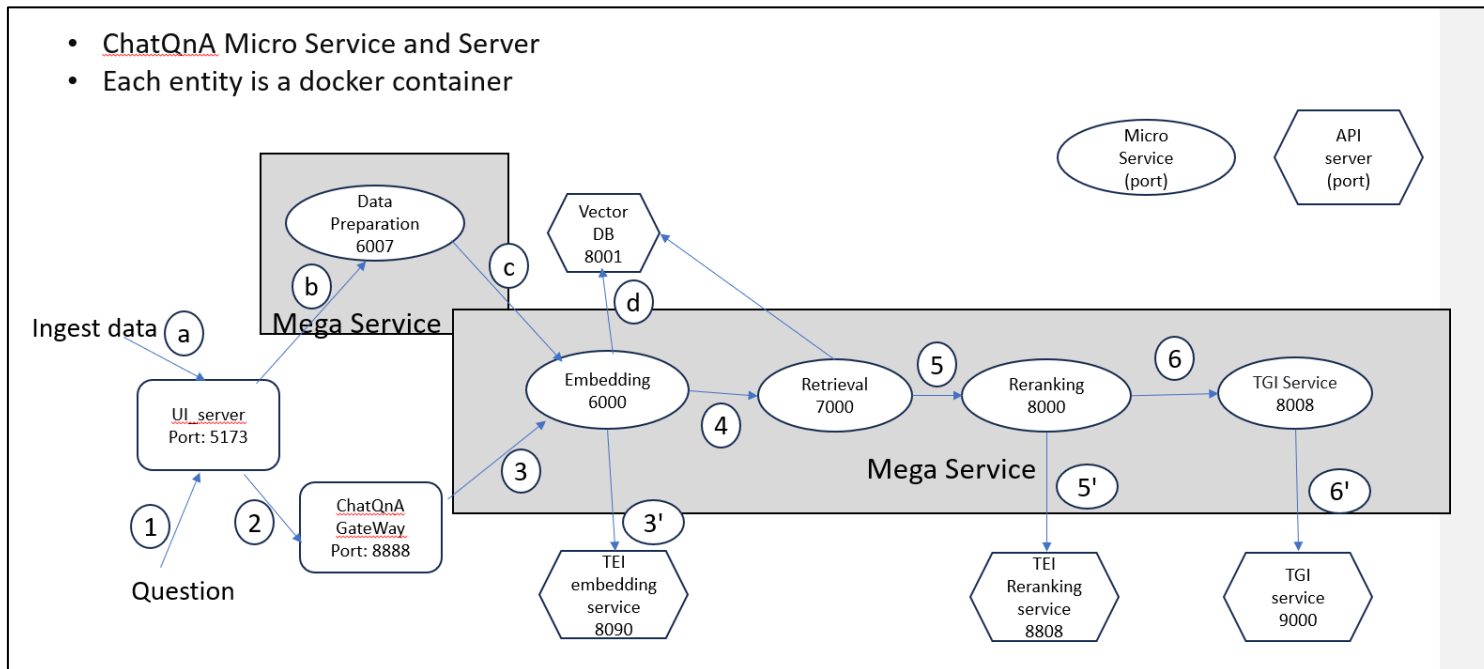
- RAG によるチャットボットを使った質問回答サービス
- インテル® Gaudi® 2 AI アクセラレーター、
インテル® Xeon® スケーラブル・プロセッサ、
NVIDIA GPU のサポート
- サービスはコンテナ化されているため
各種クラウドサービス上で動作します
 - ✓ AWS、GCP、IBM Cloud、Azure、
Oracle Cloud、ITAC



出典: [ChatQnA Sample Guide — OPEA 1.1 documentation](#) (英語)

ChatQnA の構造

- [ChatQnA Micro Service and Server](#)
- Each entity is a docker container



出典: [ChatQnA Sample Guide — OPEA 1.1 documentation](#) (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サービスを実行します
 - ✓ Embedding、retriever、rerank、tgi マイクロサービスを利用します
- ITAC 上の CPU インスタンスを利用して ChatQnA をデプロイします



intel® tiber™ AI Cloud

- インテルの CPU、GPU、AI アクセラレーターを搭載したサーバーを利用できるクラウド・コンピューティング・サービス
 - ✓ AlaaS を構築するためのコンピューティング・リソースを提供
ベアメタルおよび VM、k8s サーバー
 - ✓ + AI ソフトウェア開発者向けの勉強コンテンツ
SYCL や生成 AI などのオンライン・コンテンツおよびチュートリアルの利用
↑を実行するための無償利用枠ハードウェア・リソースの提供

インテルの AI ハードウェア・ポートフォリオ

インテル® Tiber™
AI クラウドで利用可能

ディープラーニング・アクセラレーション	 <p>インテル® Gaudi® 製品: 専用のディープラーニングのトレーニングと推論</p>	
汎用 アクセラレーション	 <p>クラウドゲーム、VDI、メディア・アナリティクス、リアルタイム高密度ビデオ</p>	 <p>並列コンピューティング、HPC、AI (HPC 向け)</p>
	  <p>リアルタイム、中スループット、低レイテンシー、スパース推論</p> 	  <p>中小規模のトレーニングと微調整</p>  
汎用	    <p>エッジおよびネットワーク AI 推論</p>  	   <p>クライアント AI の使用</p>

ITAC インスタンスに ChatQnA を手動デプロイ

[Getting Started with OPEA — OPEA 1.1 documentation](#) (英語)

1. GenAIEamples リポジトリのクローンと ChatQnA ディレクトリーに移動

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

2. Docker のインストール ※未インストールの場合

```
> cd GenAIEamples/ChatQnA/docker_compose/  
> ./install_docker.sh
```

✓ apt 経由で下記をインストール

docker-ce, docker-ce-cli, containerd.io, docker-buildx-plugin, docker-compose-plugin

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 GenAIExamples/ChatQnA/docker_compose/intel/cpu/xeon/
```

4. 環境変数の設定

- ✓ `set_env.sh` の実行と追加の環境変数を設定します

```
> export host_ip="ホスト側のIPアドレス"  
> export HUGGINGFACEHUB_API_TOKEN="Hugging Face のアクセストークン"
```

- ✓ 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
- ✓ 利用するモデルは `set_env.sh` に記述されています

```
EMBEDDING_MODEL_ID="BAAI/bge-base-en-v1.5"
```

```
RERANK_MODEL_ID="BAAI/bge-reranker-base"
```

```
LLM_MODEL_ID="Intel/neural-chat-7b-v3-3"
```

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
```

CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS
3a65ff9e16bd	opea/nginx:latest	`/docker-entrypoint.\...	14 hours ago	Up 14 hours	0.0.0
7563b2ee1cd9	opea/chatqna-ui:latest	`docker-entrypoint.s\...	14 hours ago	Up 14 hours	0.0.0
9ea57a660cd6	opea/chatqna:latest	`python chatqna.py`	14 hours ago	Up 14 hours	0.0.0
451bacaac3e6	opea/retriever-redis:latest	`python retriever_re\...	14 hours ago	Up 14 hours	0.0.0
c1f952ef5c08	opea/dataprep-redis:latest	`python prepare_doc_\...	14 hours ago	Up 14 hours	0.0.0
2a874ed8ce6f	redis/redis-stack:7.2.0-v9	`/entrypoint.sh`	14 hours ago	Up 14 hours	0.0.0
ac7b62306eb8	ghcr.io/huggingface/text-embeddings-inference:cpu-1.5	`text-embeddings-rou\...	14 hours ago	Up 14 hours	0.0.0
521cc7faa00e	ghcr.io/huggingface/text-generation-inference:sha-e4201f4-intel-cpu	`text-generation-lau\...	14 hours ago	Up 14 hours	0.0.0
9faf553d3939	ghcr.io/huggingface/text-embeddings-inference:cpu-1.5	`text-embeddings-rou\...	14 hours ago	Up 14 hours	0.0.0

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 とはなんですか？

OPEA (オペア) は、オペラの略称です。オペラは音楽、歌、舞踏などの芸術を組み合わせた表演の形式で、その中で歌劇を表現しています。

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 とは何ですか？

OPEA (Open Platform for Enterprise AI) は、企業向けのAI (人工知能) の開発と評価を支援するフレームワークです。このプラットフォームは、最先端のイノベーションを活用しつつ企業レベルのニーズを優先して、複合型の生成AI (GenAI) ソリューションを作成および評価することを目的としています。OPEA は、特に Retrieval Augmented Generative AI (RAG) を中心に、企業システムに簡単に GenAI ワークフローを統合し、そのデプロイメントを管理することで、GenAI の採用とビジネス価値の高速化を図ります。プラットフォームには、最新の GenAI システムの構築と評価に必要な複数のマイクロサービスが含まれています。

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ール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
- インテル® Tiber™ AI クラウドはインテル製の最新 CPU、GPU、AI アクセラレーターを提供するクラウドサービスです

GenAIExamples 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

補足: GenAIEXamples

- OPEA Project Documentation の Getting Started
 - ✓ [Getting Started with OPEA — OPEA 1.1 documentation](#) (英語)
- マイクロサービスの一覧
 - ✓ [GenAI Microservices — OPEA 1.1 documentation](#) (英語)
- それぞれのサンプルでサポートされるモデル一覧
 - ✓ [Supported Examples — OPEA 1.1 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.