



Open Platform for Enterprise AI

<https://github.com/oapea-project>

zhiwei.zhang@intel.com



GenAI is **emerging rapidly**,
but enterprises are
struggling to **realize GenAI
value** in production.

**Open Source AI Inferencing microservices
and reference solution to simplify
enterprise Generative AI adoption and
reduce the Time To Market**


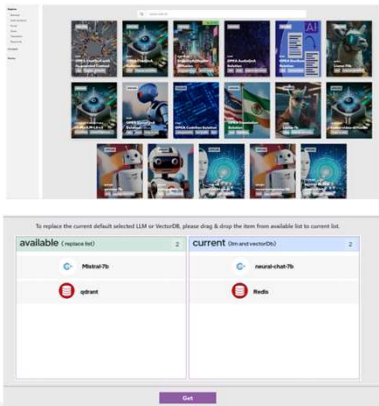


OPEA Partners

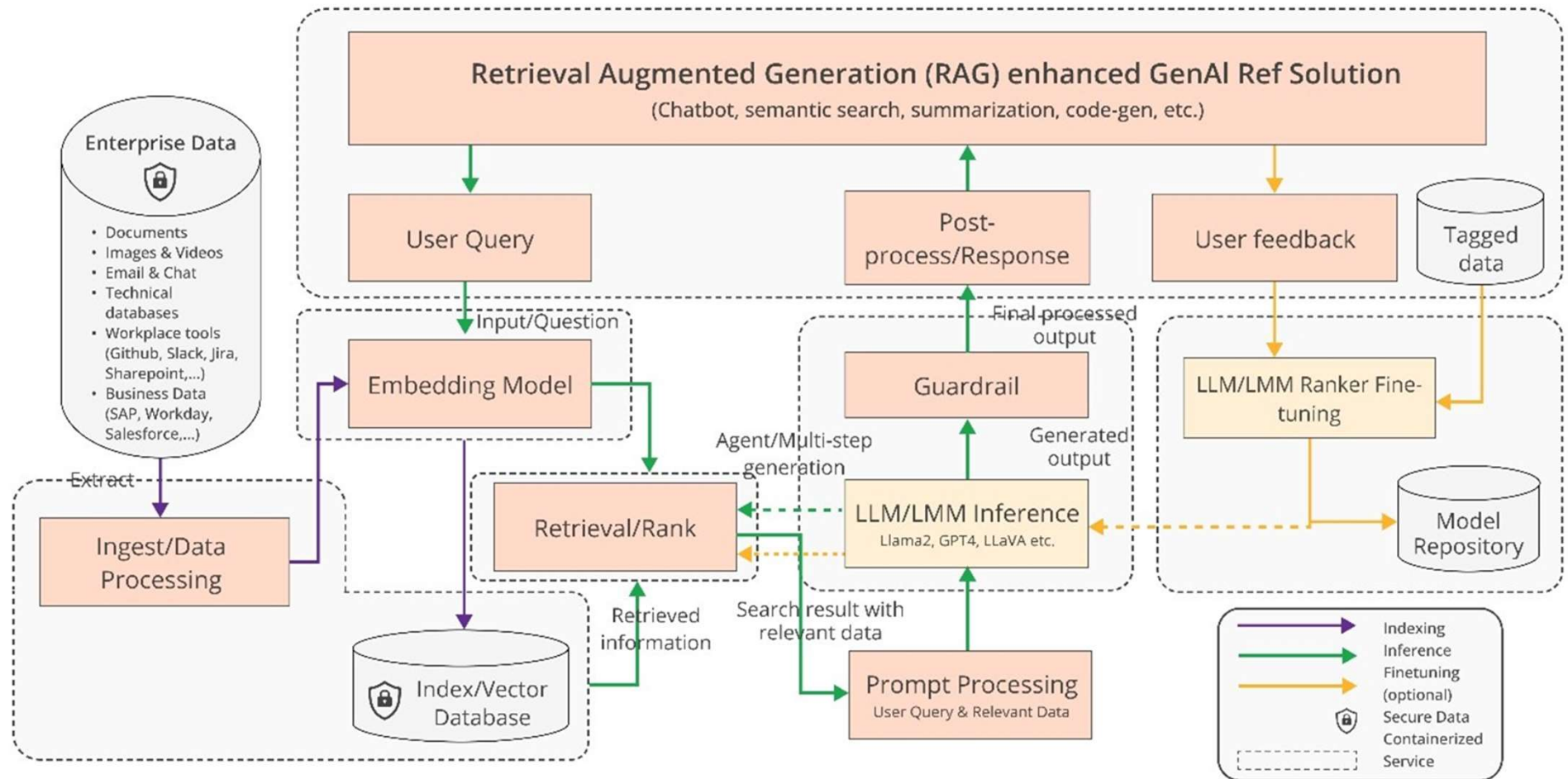


Partners as of August 7

OPEA: convert customer requirement into “Easy to use” reference solutions to increase adoption rate at Enterprise

Customer requirement	OPEA component	OPEA reference solution
<ul style="list-style-type: none">▪ Data security▪ Easy to use▪ Cost and ROI▪ Which model▪ User experience▪ Evaluation▪ Various deployment scenarios<ul style="list-style-type: none">▪ On Prem▪ Hybrid▪ On Cloud	<ul style="list-style-type: none">▪ Embeddings▪ VectorDB▪ Retrieval and Reranking▪ popular LLM & LVM▪ Model serving (vLLM, TGI)▪ LLM Frameworks▪ Knowledge Graph/GraphRAG▪ Finetuning▪ AI Agent▪ ...	<ul style="list-style-type: none">▪ E2E solution w/ composable and configurable Micro Service▪ Q&A, DocuSum, CodeGen▪ Easy to use <div></div> <div></div>

Technical Complexity



OPEA core intent

Construction of GenAI solutions, including retrieval augmentation:

- Set of building blocks to be used in the composition of GenAI solutions.
 - GenAI models – Large Language Models (LLMs), Large Vision Models (LVMs), etc.
 - System components – e.g., Embedding Models; Vector DB; Ranking, Prompt processing, and more
- Set of compositional capabilities for building AI agents & creating full end-to-end GenAI flows
- Tools for fine-tuning, customizing and optimizing, including for datacenter/on-prem settings
- A variety of validated, ready for deployment end-to-end reference flows

Evaluation of GenAI solutions, including retrieval augmentation:

- Means and services to fully evaluate and grade components and end-to-end GenAI solutions
 - **Assessment** – Detailed tests done for particular modules or attributes of the end-to-end flow.
 - **Grading** - Aggregation of the individual assessments to a grade per each of the four domains -
 - Performance
 - Features
 - Trustworthiness
 - Enterprise-readiness
 - **Certification** (if offered) - meeting a minimum level of grading on all four domains.

OPEA Roadmap

approved by TSC

		May 2024	June 2024	July 2024	Aug 2024	Sep 2024	Q4 2024	Q1 2025
contribution	Components	ASR, Data Prep, Embedding, Guardrails, LLM (Gaudi TGI), Rerank, retrieval, TTS, vectorDB	LLM (Xeon vLLM & Ray, Ollama), OVMS, prompting, user feedback management, Mega Component (M16 RAG service)	LVM (Gaudi vLLM & Ray), vectordb (svs), Gateway guardrail, Auth Z/N	Documentation Test automation script, Telemetry	Microservice for Image and Video	Finetuning E2E pipeline Knowledge Graph	more Microservice request from community Confidential Container
	Use Cases/Examples	ChatQnA, CodeGen, CodeTrans	DocSum, SearchQnA,	FAQGen	Documentation Test automation script	Text to Image generation, Image to Video generation Playground (composable and configurable)	Finetuning (Lora) AI Agent (single Agent with text and Audio as user interface) Closed source LLM GraphRAG	AI Agent (Multi Agent) Finetuning (Adaptive) Long context window (>1M) GenAI Studio
	Cloud Native	OneClickOPEA: ChatQnA, CodeGen GenAI microservice connector,	OneClickOPEA for 2 more examples GMC with switch support (dynamic pipelines) Helm charts/templates for custom yamls (refactoring)	OpenShift enablement for OPEA, OneClickOPEA for 3 more examples. Security (Service Mesh, guardrails)	Demok8s resource management, Documentation on autoscaler analysis.		Static tuning on Resource management for deployment	Dynamic tuning on Resource management through K8s
	Evaluation + Others	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples), lm-eval-harness, bigcode-eval-harness RAGAS evaluation service 	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples) Gaudi (2) and CPUs in CICD cluster! 	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples) 	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples) 	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples) 	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples) 	<ul style="list-style-type: none"> CICD & Validation Eval: E2E (comps + examples)
AI models		LLM: llama2 (7b, 13b, 70b), llama3 (8b, 70b), code-llama, Llama guard Embedding: BGE-base	LLM: mistral-7B, mixtral-8x7B, Embedding: E5-mistral-7b-instruct, all-mpnet-base-v2	LLM: Phi, Gemma Embedding: all-MiniLM-L6-v2, paraphrase-albert-small-v2	Vision: llava Mixtral-8x22B	Diffusion model: Stable Diffusion XL Stable Diffusion 3M Stable Video Diffusion	LLM Close: GPT3.5/4/4o, Claude 3/3.5, AWS Bedrock endpoint	LLM: SetFit More to be defined
AI tools Integration		VectorDB: Chroma Framework: Langchain,	VectorDB: Pinecone, Redis Framework: Llamaindex, Haystack	VectorDB: PGVector, Qdrant	VectorDB: Milvus	VectorDB: Weaviate	Knowledge graph: Neo4j Agent: LangGraph	AutoGen, CrewAI
Deployment type		On Prem, IDC (XEON, Gaudi)	On Prem, IDC (XEON, Gaudi)		Public Cloud AWS (XEON CPU + NV GPU)		Public Cloud (Azure, GCP, Oracle, AWS) AI PC (Intel)	Public Cloud (tier2 CSP) AI PC (others)

OPEA partners in general

Potential OPEA partners on opea.dev, can [send us](#) a logo and a quote. A quote involving potential partner's GenAI vision & plan and possible intersect w/ OPEA vision, plan and roadmap.

Partner and OPEA can work towards one/more of the following paths

- Contributing partner
 - Contribute to any of the four OPEA repos
 - Contribute components to github.com/opea-project
 - Contribute complete E2E use cases for RAG, Fine-tuning adhering to OPEA specification
 - Contribute to multi-agent use cases
- Utilizing partner
 - Create custom POCs or use cases for internal use utilizing OPEA repos
 - Create customer driven solutions that can land hosted/managed services with OPEA repos
 - Harden, Certify, productize GenAI microservices/ use cases as partner deployment ready
 - Utilize details, documentation and Infra setup scripts to setup a node and/or a Kubernetes cluster on-prem or CSPs
- Co-Development partner
 - Co-develop **new** custom E2E GenAI solutions (not in repos today) as per market needs w/ OPEA
 - Provide these E2E GenAI solutions as hosted services on CSP Marketplaces

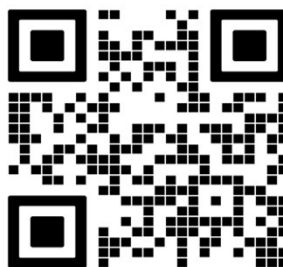
Participate in OPEA

Take Action

Learn More



OPEA Project
on [GitHub](#)



[OPEA](#)
Mailing Lists



Visit
[OPEA.dev](#)

Contribute

Join a Working Group

Bring Enterprise AI use cases

Contribute code, docs,
projects, blueprints, & more

Provide feedback

Evangelism and Promotion
of OPEA in YOUR
communities, events



Open Platform for Enterprise AI