Intel® AI for Enterprise Inference

Open-source, native LLM serving stack designed for Intel® Xeon® processors and Intel® Gaudi® AI accelerators, deployable in cloud and on-premises environments

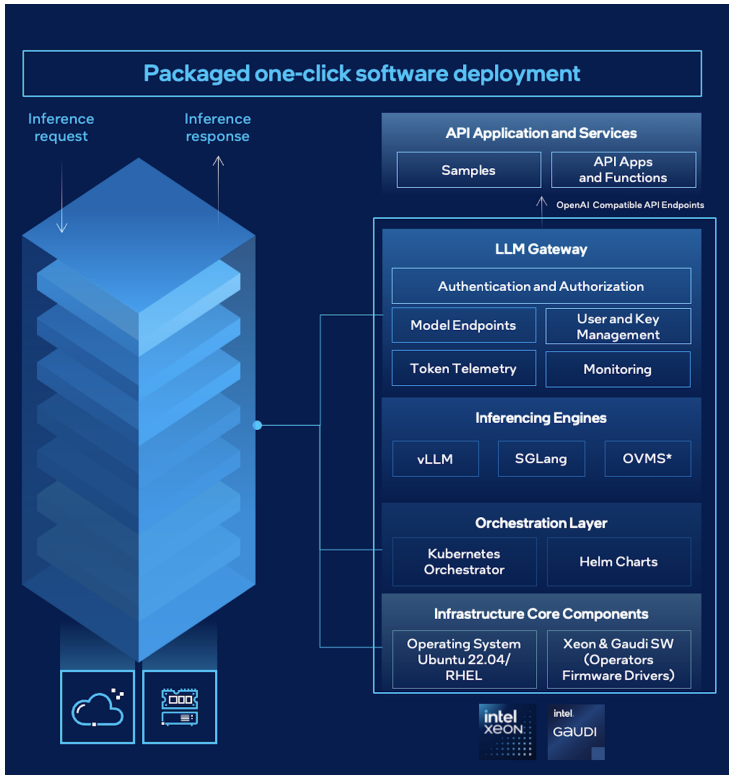
Overview

Intel® AI for Enterprise Inference is an open-source packaged software stack that empowers developers and IT teams to deploy large language models (LLMs) with Intel® Xeon® processors or Intel ® Gaudi® AI accelerators.

The open-source solution offers simple, transparent APIs for inference compatible with common generative AI (GenAI) API patterns and can be run on cloud or on-premesis infrastructure. It supports production-ready inference for various enterprise use cases, reducing setup time, simplifying deployment, and enhancing performance through Intel hardware-accelerated execution. All microservices and deployment assets are open source and available on GitHub.

System capabilities

- Quickly launch and manage LLM inference workloads with Helm on any Kubernetes or Red Hat® Open Shift® environment.
- Provision open and compatible inference endpoints using standard APIs and token-based access.
- Seamlessly integrate into applications without DevOps customization or workflow retraining.
- Choose Intel Gaudi AI accelerators for high throughput workloads and Intel Xeon processors for low-latency and orchestration-heavy tasks.
- Select and deploy open-source models or use a bring-your-own-model (BYOM) approach, with support for versioning and model swapping.
- Manage access, scale via Kubernetes, and monitor API -activity using built-in orchestration and observability tools.
- Deploy with a single click across any environment, cloud service providers (CSPs) or on-premises, ensuring fast, consistent rollout and scalability.
- Designed for enterprise compatibility, meeting the expectations of IT Ops, site reliability engineer (SRE), and developer workflows via CI/CD support and GitHub-based deployment pipelines.



Performance and optimization[1]

- Use Intel Xeon processors for latency-sensitive tasks (e.g., chat, summarization).
- Use Intel Gaudi accelerators for high-throughput workloads (e.g., batch inference, image generation).
- Optimize memory usage using vLLM-based architecture, supporting concurrent inference and faster model serving.

92% more cost-efficient than GPU competition on Llama 3.1-405B-Instruct-FP8.

43% more tokens/second than GPU competition on IBM Granite-3.18B-Instruct.

36% more tokens/second than GPU competition on Llama 3.1-405B-Instruct-FP8.

1. https://signal65.com/research/ai/intel-gaudi-3-accelerates-ai-at-scale-on-ibm-cloud/

Model compatibility and flexibility

- Run pre-integrated models such as LLama, Mistral, Mixtral, Falcon, DeepSeek or BYOM.
- Swap, version, and test models through a standardized API interface without modifying the infrastructure.

Integration with enterprise ecosystems

- Connect to enterprise platforms (Slack, SharePoint, Jira, custom apps) using standard REST APIs.
- Integrate with CI/CD pipelines, logging/monitoring stacks (e.g., Prometheus, Grafana), and DevOps workflows.
- Maintain compatibility with GPT-based orchestration tools, including LangChain and RAG pipelines.

Main use cases

- Chatbots and co-pilots integrated into enterprise workflows
- Internal document summarization and classification
- Retrieval-augmented generation (RAG) for internal knowledge bases

Key Benefits

Leverage transparent, modular, and flexible setup and avoid opaque APIs that limit visibility, customization, or cost control.

Engineering: Access familiar APIs, customize models, and iterate quickly.

IT operations: Deploy secure, auditable services and manage orchestration at scale. Scale up or down using Kubernetes-native auto-scaling and resource control.

Observability: Monitor endpoint performance, resource usage, and inference activity using built-in metrics and logging. Integrate with existing observability stacks for uptime, error tracking, and compliance audits.

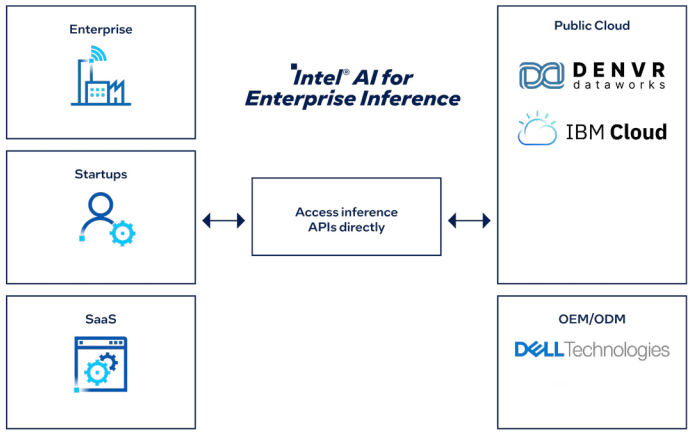
Product management: Launch AI-powered features with minimal setup and ongoing flexibility.

Security and compliance

- Build hybrid environments while ensuring compliance with features like API Key Security, Model level access control and encrypted, role-based access.
- Support tenant isolation, token-based permissions, and encrypted communication between components.
- Comply with enterprise security and deployment policies across internal, hybrid, and regulated environments.

Deployment options

- CSPs
  - Denvr Dataworks: Fully managed API services (coming Q3 2025).
  - IBM Cloud: Enterprise-ready integration with IBM-native infrastructure.
  - Azure Marketplace: Solution based on 4th Gen Intel Xeon processors and above.
  - AWS Marketplace: Solution based on 4th Gen Intel Xeon processors and above.
- On-premises
  - Self-hosted: Private, customizable Kubernetes or RedHat Open-Shift environments with full control.



How to Get Started

The open-source microservices powering the solution can be self-hosted or modified for custom deployments using the publicly available GitHub repository. In addition, you can make your request for endpoints. After receiving credentials, you can begin testing inference performance using your preferred tools.

Add QR code with a link to - https://github.com/opea-project/Enterprise-Inference

intel ai

© Intel Corporation. Intel, the Intel logo and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.