Qualcomm

AI Inference Suite

for cloud and on-prem deployments

The Qualcomm® AI Inference Suite for Cloud and for On-Prem enables the deployment of AI models and applications with a single click, supporting efficient and scalable AI, while eliminating the need for complex infrastructure management.

Highlights

- Powered by Qualcomm® Cloud AI roadmap of accelerators
- Robust APIs, OpenAI compatible
- Ready to use gen AI applications
- Configurable for any use case
- Supports 1000s of models
- Supports multi-tenancy

Unlock AI with ease

Seamless one-click deployment. Easily swap or add your own models as needed, including gen AI, computer vision, and natural language processing. Build custom apps with common frameworks.

Deploy your way

Choose your preferred deployment: On-prem or cloud, powered by Qualcomm® Cloud AI roadmap of accelerators.

Top performance, future-proofed

Maximize performance and cost efficiency with powerful inference accelerators, embedded optimization techniques, and state-of-the-art models.

Run with confidence

High-availability and strict data privacy; no storage of model inputs or outputs. Designed and pressure tested by enterprise, for enterprise.

Accelerate gen AI development with ready-to-use applications and agents

- Chatbot
- AI agents
- Retrieval-augmented generation (RAG)
- Summarization
- Image generation
- Real-time translation
- Code development
- Real-time transcription
- Your next use case

Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.