Edge AI inference: small models on devices, offline and in real time

26 October 2026 · 5 min read · blogs.iictp.org/edge-ai-inference-slms-on-devices

Edge inference runs a model on or near the device that produces the data: a phone, a gateway, a vehicle, a factory machine. It cuts delay, reduces data sent over the network and keeps working when the connection drops.

TL;DR The short version

- Edge inference runs a model on or near the device that produces the data, such as a phone, gateway, vehicle or machine.
- It gives real-time responses, keeps sensitive data on the device and keeps working when the connection drops.
- You give up capability and work within tight memory, power and cooling, so use a small task-specific model and test quantization loss.

Read online, listen to it, or change the text size: https://blogs.iictp.org/edge-ai-inference-slms-on-devices

Why run at the edge

- Responses in real time, without a network round trip.
- Sensitive data stays on the device.
- Operation continues offline.
- Less data sent upstream.

What you give up

- Smaller models and less capability.
- Limited memory, power and cooling.
- Harder updates across many devices.

Making it fit

- Use a small model suited to one task.
- Quantize, and test the quality loss.
- Use the device’s accelerator if its software supports it.
- Plan updates, rollback and monitoring from the start.

Related reading

See small language models explained and edge computing explained.

> Related credential: Local LLM Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. IICTP. Local LLM Practitioner credential page. iictp.org/certifications/llmp.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/edge-ai-inference-slms-on-devices. More guides: blogs.iictp.org.