Local LLM vs cloud AI: cost, privacy and latency compared

26 October 2026 · 5 min read · blogs.iictp.org/local-llm-vs-cloud-ai-cost-privacy-latency

The honest answer to “local or cloud?” is a comparison you do with your own numbers. Here is how to set it up, so you do not rely on anyone else’s claims.

TL;DR The short version

- Local versus cloud AI is a comparison to run with your own numbers, not anyone else's claims.
- Compare the whole bill: tokens and commitments in the cloud against hardware, power, hosting and staff time locally.
- Local keeps data inside your boundary; cloud offers stronger models, so run both on 50 real examples and score quality.

Read online, listen to it, or change the text size: https://blogs.iictp.org/local-llm-vs-cloud-ai-cost-privacy-latency

Cost: compare the whole bill

- **Cloud:** price per token or request, times your expected volume, plus any minimum commitments.
- **Local:** hardware, power, hosting, staff time to run and patch it, and the cost of replacing it.
- A local deployment pays off when use is steady and high; a cloud service often wins when use is low or spiky.

Privacy

- Local keeps prompts and documents inside your boundary.
- Cloud can still meet strict needs through contracts and regional options; read the terms.

Latency

- Local removes a network round trip but is limited by your hardware.
- Cloud offers more powerful models but adds network time and can queue under load.

How to compare

1. Pick a realistic task and 50 real examples.
2. Run both options and score quality.
3. Measure time to first token and total time.
4. Cost out a month of expected use for each.

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/local-llm-vs-cloud-ai-cost-privacy-latency. More guides: blogs.iictp.org.