LLMOps: running AI in production, covering cost, monitoring and versioning

26 October 2026 · 6 min read · blogs.iictp.org/llmops-running-ai-in-production

LLMOps is the operational side of language-model applications: keeping them reliable, affordable and improvable once real users depend on them. Much of it is ordinary operations with a few new failure modes.

TL;DR The short version

- LLMOps keeps language-model features reliable, affordable and improvable once real users depend on them.
- Monitor latency, errors, token use and cost per feature, and version prompts, models and settings.
- Use the smallest model that passes your evals, cache repeated work, and roll out model upgrades like any release.

Read online, listen to it, or change the text size: https://blogs.iictp.org/llmops-running-ai-in-production

What to monitor

- Latency and error rates, as with any service.
- Token use and cost per request and per feature.
- Quality signals: user feedback, refusals, retries, escalations.
- Drift: answer quality changing after a model or data change.

What to version

- Prompts and prompt templates.
- The model and its settings.
- The retrieval index and its source data.
- The evaluation set.

Controlling cost

- Use the smallest model that passes your evals.
- Cache repeated work.
- Trim the context you send.
- Set per-user and per-feature limits.

Change safely

Treat a model upgrade as a release: run the eval set, roll out to a small share first, compare, and keep a way back. Log enough to reproduce a bad answer, while keeping personal data out of logs that do not need it.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Google Cloud. “MLOps: Continuous delivery and automation pipelines in machine learning.” cloud.google.com/architecture.

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