Tanimawo AI Academy

ADVANCED

AI Operations: Testing, Monitoring and Costs

Run AI features reliably and affordably in production

Free guide

About 16 minutes to read

tanimawoai.com

Free AI training every Sunday
with Dr. Daniel Tanimawo

TANIMAWO ACADEMY
tanimawoai.com
Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.

Why AI needs operations

AI features can fail quietly: a model update changes answers, costs creep up, or a prompt breaks on unusual input. Good operations catch problems **before users do**.

1. Evaluate before you ship (evals)

Build a small **test set**: 30 to 100 realistic inputs with what a good answer must contain.

Check	Example
Correctness	Answer includes the right class time
Format	Valid JSON / under 100 words / has 3 bullets
Safety	Refuses to reveal private data
Tone	Friendly, no jargon

Run the test set on every change (prompt, model, settings) and compare scores. Some checks can be automatic (format, keywords); others need a human or a second model as a judge.

2. Monitor in production

Log for every request (never store passwords or unnecessary personal data):

- Time, feature name, model
- Input and output token counts and estimated cost
- Latency and errors
- Whether the answer came from a cache

Set **alerts** for spikes in errors, latency or daily spend.

TANIMAWO AI ACADEMY
tanimawoai.com · Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.

3. Control costs

Technique	Typical effect
Right-size the model	Use small models for simple tasks, large ones only where needed
Set max_tokens per feature	Stops long, expensive answers
Cache repeated answers	Identical common questions cost nothing after the first
Prompt caching	Providers can discount a repeated, unchanging prompt prefix
Trim context	Summarise old conversation turns instead of resending everything
Batch processing	Non-urgent bulk jobs are often cheaper through batch APIs
Budgets and rate limits	Per user, per IP and global daily caps

Simple cost estimate: requests per day × (average input tokens × input price + average output tokens × output price).

4. Reliability

- **Timeouts and retries** with increasing delay for temporary errors
- **Fallbacks:** a smaller model or a friendly message when the main model fails
- **Graceful degradation:** the rest of the app keeps working if AI is down
- **Version pinning:** know which model version you use and test before switching

5. Guardrails

- Validate and limit inputs (length, type)
- Check outputs before showing them (format, banned content, personal data)
- Keep system prompts free of secrets; assume users may try to reveal them
- Treat retrieved web pages and documents as untrusted input

6. Feedback loop

Add a thumbs up / down to AI answers. Review the thumbs-down answers weekly, add them to your test set, fix the prompt, and re-run the evals.

TANIMAWO AI ACADEMY
tanimawoai.com · Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.

> **The goal:** you should be able to answer three questions at any time: Is it working? Is it good? What does it cost?

Launch checklist

- [ ] Test set passes at an agreed score
- [ ] Logging and cost tracking switched on
- [ ] Daily budget cap and per-user rate limits set
- [ ] Fallback message when the AI is unavailable
- [ ] Someone owns reviewing feedback each week

Keep learning, free

Join the free live AI class every Sunday at 8 PM Nigeria time with Dr. Daniel Tanimawo. Register at **tanimawoai.com**, watch past classes on the Replays page, and download more free guides.

Share this guide with a friend who wants to learn AI.

TANIMAWO AI ACADEMY
tanimawoai.com · Dr. Daniel Tanimawo

© 2026 Dr. Daniel Tanimawo · Tanimawo AI Academy · tanimawoai.com · Free to download and share with credit. Selling, rebranding or removing this notice is not allowed.