Cost of LLM in Production (Thai Edition) / [redacted] LLM [redacted]

Framework + checklists for budgeting, measurement, and optimization

Disclaimer: This document is a practical template and does not constitute legal, compliance, or financial advice. | [redacted]/[redacted]

TH ([redacted]): [redacted] LLM [redacted]

EN: A practical cost framework for running LLM systems in production.

1) Cost categories / [redacted]

- API usage (tokens), embeddings, reranking, tool calls
- Infrastructure: gateways, vector DB, caching, observability
- People: prompt/eval, ops/on-call, content owners

2) Measurement / [redacted]

- Track tokens/request, tokens/user/day, cost per successful outcome
- Separate experimental traffic from production traffic

3) Optimization levers / [redacted]

- Reduce context size (better retrieval + chunking)
- Cache where safe; batch requests; streaming
- Use smaller models for simpler tasks; route by complexity

4) Guardrails / [redacted]

- Budgets, rate limits, user tiers, abuse prevention

5) Typical pitfalls / [redacted]

- Not measuring 'cost per outcome' → optimize the wrong thing
- Letting prompts grow without regression tests

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■:

Cost of LLM in Production (Thai Edition) / ■■■■■■ LLM ■■■■■■■■■

Framework + checklists for budgeting, measurement, and optimization

Notes / ■■■■■■■■: