

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]

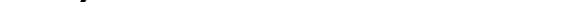
- Reduce context size (better retrieval + chunking)
- Cache where safe; batch requests; streaming
- Use smaller models for simpler tasks; route by complexity

4) Guardrails / [REDACTED]

- Budgets, rate limits, user tiers, abuse prevention

5) Typical pitfalls / [REDACTED]

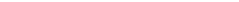
- Not measuring 'cost per outcome' → optimize the wrong thing
- Letting prompts grow without regression tests

Cost of LLM in Production (Thai Edition) /  LLM 

Framework + checklists for budgeting, measurement, and optimization

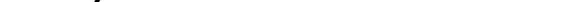
Notes / :

Cost of LLM in Production (Thai Edition) /

LLM  3

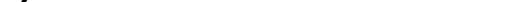
Framework + checklists for budgeting, measurement, and optimization

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

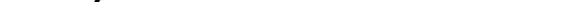
Notes / :

Cost of LLM in Production (Thai Edition) /

LLM 3

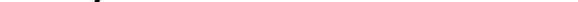
Framework + checklists for budgeting, measurement, and optimization

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 3/10

Framework + checklists for budgeting, measurement, and optimization

Notes / :

Cost of LLM in Production (Thai Edition) /

Framework + checklists for budgeting, measurement, and optimization

Notes / :

Cost of LLM in Production (Thai Edition) /

Framework + checklists for budgeting, measurement, and optimization

Notes / :

Cost of LLM in Production (Thai Edition) /

Framework + checklists for budgeting, measurement, and optimization

Notes / :

Cost of LLM in Production (Thai Edition) /

Framework + checklists for budgeting, measurement, and optimization

Notes / :

Cost of LLM in Production (Thai Edition) /

Framework + checklists for budgeting, measurement, and optimization

Notes / :