LLM Evaluation Matrix

Use this matrix to compare multiple LLMs against your requirements. Assign a weight to each pillar, then score each model from 1 (poor) to 5 (excellent).

Step 1: Define Weights per Pillar

Adjust these weights based on what matters most for this project.

- Performance & quality (P): _____ / 100
- Latency & scalability (L): _____ / 100
- Cost & total cost of ownership (C): _____ / 100
- Security & compliance (S): _____ / 100
- Vendor support & ecosystem (V): _____ / 100

Step 2: Scoring Scale

Use the same scale for all criteria:

- 1 = Poor / does not meet requirements
- 3 = Acceptable / meets minimum requirements
- 5 = Excellent / strongly exceeds requirements

Step 3: List Candidate Models

- **Model A:** _____
- **Model B:** _____
- **Model C:** _____

Step 4: Score Each Model by Pillar

For each criterion below, score Model A, Model B, and Model C from 1–5.

1. Performance & Quality (P)

How well does each model handle your tasks and domain?

Criterion	Weight (P)	Model A (1–5)	Model B (1–5)	Model C (1–5)
P1. Accuracy on your test set	____	____	____	____
P2. Reasoning on complex cases	____	____	____	____
P3. Handling domain-specific jargon	____	____	____	____

P4. Reliability / consistency of outputs	____	____	____	____

2. Latency & Scalability (L)

Can the model meet your speed and scale requirements?

Criterion	Weight (L)	Model A (1–5)	Model B (1–5)	Model C (1–5)
L1. Average latency at expected load	____	____	____	____
L2. Behaviour at peak traffic	____	____	____	____
L3. Rate limits vs. your needs (TPS / QPS)	____	____	____	____
L4. Uptime / availability guarantees	____	____	____	____

3. Cost & Total Cost of Ownership (C)

Look beyond per-token price to total cost over time.

Criterion	Weight (C)	Model A (1–5)	Model B (1–5)	Model C (1–5)
C1. Input token price	____	____	____	____
C2. Output token price	____	____	____	____
C3. Estimated monthly run-rate at your usage	____	____	____	____
C4. Extra infra / ops cost (if self-hosted)	____	____	____	____
C5. Integration & engineering effort	____	____	____	____

4. Security & Compliance (S)

Ensure the model fits your risk, data, and regulatory needs.

Criterion	Weight (S)	Model A (1–5)	Model B (1–5)	Model C (1–5)

S1. Data usage policy (training / logs)	____	____	____	____
S2. Data residency options match needs	____	____	____	____
S3. Private networking / VPC / on-prem options	____	____	____	____
S4. Logging & audit capabilities	____	____	____	____
S5. Fit with your specific regulations	____	____	____	____

5. Vendor Support & Ecosystem (V)

Consider maturity, support quality, and ecosystem strength.

Criterion	Weight (V)	Model A (1–5)	Model B (1–5)	Model C (1–5)
V1. Provider maturity & stability	____	____	____	____
V2. Documentation & SDK quality	____	____	____	____
V3. Support responsiveness / SLAs	____	____	____	____
V4. Roadmap alignment (12–24 months)	____	____	____	____
V5. Community, ecosystem & integrations	____	____	____	____

Step 5: Summary & Decision

After scoring, you can compute a rough weighted score per model (manually or in a spreadsheet) and capture your decision notes here.

- **Model A – total score & notes:** _____
- **Model B – total score & notes:** _____
- **Model C – total score & notes:** _____

Final decision and rationale: _____