LLM FinOps Infrastructure Maturity Scorecard

Maturity Dimension	Weight	Self-Score (1-5)	Key Audit Checklist Criterion
1. Hardware Amortization	20%	3 / 5	GPU instances scaled dynamically; spotVM nodes active.
2. API Routing Proxy	25%	4 / 5	Dynamic routing proxy triages queries by complexity.
3. Caching & Hit Rates	25%	2 / 5	System prompts structured; prefix caching active.
4. Output Integrity	15%	3 / 5	Automated CI/CD assertion checks active on model swaps.
5. FinOps Lifecycle	15%	2 / 5	Token telemetry dashboard maps real-time cost attribution.

How to Compute Your Score: Multiply your score on each row (1 to 5) by the weighting percentage. Sum the resulting products to get your overall weighted score. An overall score above 3.5 denotes solid managed capability, while scores below 2.5 indicate urgent optimization needs (e.g. system prompts sending raw redundant tokens).

Authored by Vatsal Shah. For help analyzing your cost structures, contact contact@shahvatsal.com. All deliverables, including the TCO Excel workbook, are available for download on shahvatsal.com.