Sample AI FinOps Audit Report

Illustrative structure for a ML Mind audit. Use with stakeholders to understand deliverables and savings categories.

Executive summary

The audit maps hidden AI waste across the request path and separates discovery-only savings from control-level savings.

Area	What to review
RAG/context	Identify irrelevant chunks, stale sources and oversized context.
Retries	Find repeated failures, tool loops and blind reruns.
Routing	Map requests to the cheapest safe model by risk and complexity.
Cache	Find verified repeated answers that can be served without new inference.
GPU/serving	Detect idle replicas, low utilization, OOM loops and inefficient batching.
Integrity	Protect numbers, dates, citations, policies and trust-sensitive facts.

Recommended output

A validated opportunity map, deployment-level recommendation, safe savings estimate and next-step pilot plan.

Area	What to review
Deliverable	Savings map by waste source, risk notes and implementation sequence.
Success metric	Integrity-adjusted savings: cost reduction that preserves answer quality and trust.