Case Study —

TL;DR

In one sentence, state the problem, your approach, and a measurable outcome. Example: Built a retrieval■augmented QA for support docs; improved first■contact resolution **+12.4%** at **p95=78ms** and **–37%** cost/1k queries.

Quick Facts

Role	<ML Engineer / Researcher / Solo>
Timeline	<Dates / weeks>
Stack	<PyTorch/JAX, vLLM/Triton, Ray/Spark, Feast/MLflow, Docker/K8s>
Data	<size, sources, license>
Model	<baseline → current; params; adapters/quant>
SLOs	<p95/p99 latency, availability>
Business Metric	<conversion, AHT, FCR, incidents↓>

Problem

Who has the problem and why now? Define success (e.g., “reduce average handle time by 15% without hurting CSAT”). List constraints: privacy, compute budget, latency/cost targets.

Users & Stakeholders

Primary users and adjacent teams. Pain points, workflows, and success metrics per stakeholder.

Solution Overview

Explain the approach in a short paragraph for a non■expert, then add a technical note. Call out key trade■offs (quality vs. latency vs. cost).

Architecture (high level)

[Client/UI] → [Gateway] → [Retrieval] → [Model] → [Post■process] → [Metrics/Logs] (↔ Feature Store)

Data & Methods

Data sources, preprocessing, versioning (DVC/LakeFS), splits and leakage checks. Baselines and current approach (adapters/quantization). Repro: seeds, env pinning.

Evaluation & Benchmarks

Task■appropriate metrics (AUC/F1/MAE/BLEU/ROUGE/MMLU/etc.), offline protocol, and—if applicable—online A/B. Include error analysis and guardrail evals (toxicity/bias/jailbreak).

Reliability, Observability & Cost

SLOs and dashboards (Prometheus/Grafana). Canary/rollback plan. Cost: GPU hours and $/1k requests; wins from caching/routing/quant.

Impact & Results

User/business outcomes with numbers. Short stakeholder quote. Call out what changed because **you** were there (your specific role).

Next Steps

Concrete next experiments or product improvements you would prioritize.

Ethics, Safety & Privacy

Data licensing/consent, known limitations/biases, and safety mitigations. Add a one■line disclosure: “Drafted/edited with AI; all results verified by the author.”

Links: • • •

Contact: • •