Memorandum

To: Engineering Team
From: Your Name
Date: November 13, 2026
Re: Migration to the new inference serving stack

Purpose

This memo outlines the plan to migrate model serving from our current home-grown stack to the new Triton-based inference platform, targeting completion by the end of next quarter.

Action Items

- Audit all currently deployed models for compatibility with the new serving format.
- Stand up a staging environment mirroring production traffic patterns.
- Run a two-week shadow-traffic comparison before cutover.
- Update on-call runbooks to reflect the new deployment process.

Timeline

Week 1–2	Compatibility audit
Week 3–4	Staging environment setup
Week 5–6	Shadow traffic comparison
Week 7	Production cutover

Please direct questions to the infrastructure channel. We will hold a kickoff meeting this Friday at 10am.