AI Orchestration Pipeline: A 10-Part Guide

This guide walks you through a complete AI orchestration system designed for teams managing large codebases — the kind that exceed every AI's context window. **Part 1** tackles the core problem: how to make a 42MB codebase searchable and feed only the relevant files to AI. **Part 2** introduces multi-AI orchestration, running Claude, GPT-4, and AWS Bedrock with identical context in a single API call. **Part 3** automates the git workflow — one command creates a branch, commits files, and opens a PR. **Part 4** adds cost transparency with per-call tracking across all providers.

Parts 5-6 focus on error handling: diagnose production errors in 60 seconds, or go further with auto-fix that creates a ready-to-merge PR. **Part 7** solves the IAM problem — AI now reads your Terraform files and flags missing permissions before deployment fails. **Part 8** introduces rollback capability, tracking every AI-generated PR so you can revert with one click when things go wrong.

Part 9 brings it all together with a complete walkthrough showing how each pain point maps to a specific command. **Part 10** is your quick-start playbook — cheat sheets, copy-paste commands, and everything needed to get running today.

Bonus: Persistence Layer — The hidden infrastructure that makes everything work. Your 42MB codebase gets indexed once and stored in ChromaDB; server restarts load instantly instead of re-embedding for 5 minutes. Every API call, every cost, every AI-generated PR is tracked automatically. Session data persists across restarts for monthly reporting and budget alerts. The PR registry maintains a complete audit trail, enabling one-click rollbacks weeks after a PR was merged. Index once, search forever.

The result: tasks that took 2+ hours now take 5 minutes, errors that required 5+ manual steps now need one command, and rollbacks that took 26+ minutes happen in 30 seconds — all for about $0.08 per workflow.

Introduction:
https://claude.ai/public/artifacts/d71ab0ba-ac3d-4f08-a5c3-b33b21a2ef4b

Part 1: The Context Problem - 42MB Codebase
https://claude.ai/public/artifacts/905afa5f-051e-437d-9d3f-f932126fe00c

Part 2: Multi-AI Orchestration 3 AIs, One Command
https://claude.ai/public/artifacts/8cfb85dc-1bbe-4829-99f5-fe35217083ec

Part 3: GitHub Workflow Automation (Branch → Commit → PR)
https://claude.ai/public/artifacts/f3ad5e2d-7d17-46ca-bc83-0bfb16873f59

Part 4: Cost Tracking
https://claude.ai/public/artifacts/7ea9b604-d5a8-4db6-9e1e-6a41b7c8e8f4

Part 5: Error Diagnosis
https://claude.ai/public/artifacts/3997da16-e8f5-4df7-9850-feb65a66faf7

Part 6: Diagnose & Auto-Fix
https://claude.ai/public/artifacts/6c4492ba-a54d-4a30-8c8f-63be2e1b066a

Part 7: IAM-Aware Orchestration
https://claude.ai/public/artifacts/a1a94397-2974-4024-afd5-3a75fa1c52a1

Part 8: Rollback Capability
https://claude.ai/public/artifacts/436d528a-fd3d-460d-b02d-4b0307603601

Part 9: Pain Points Solved
Https://claude.ai/public/artifacts/02ad0d09-45aa-4fde-9a41-bf0e5aeeab82

Part 10: Getting Started Guide
https://claude.ai/public/artifacts/a263aec7-bec8-4e23-8ebf-06db76b4323a

Bonus: Persistence Layer
https://claude.ai/public/artifacts/6d9dcdb5-3f6f-45f9-af7e-ce2cd1cb51e8