LOOPS

Case Study — Context Engineering with Multimodal Image Understanding (E-commerce Shoes, Vertex AI, GCP)

1. Overview

Your task is to build a Cloud Run API that analyzes shoe product images together with structured context (catalog specs, brand rules, policies) and produces grounded, verifiable outputs. The system must accept image inputs from GCS URIs, compress context into a Context Card, query Vertex AI Multimodal (Gemini), and return strict JSON with answers, citations, structured attributes, and compliance results.

2. Dataset Provided

This package includes:
• 16 SKU images (SKU-001 … SKU-016)
• spec_catalog.json / .csv — structured specs (material, color, closure, style, outsole, etc.)
• brand_rules_shoes_v1.json — brand voice + rules (e.g., no orthopedic claims, color must match catalog)
• policy_shoes_v1.json — listing compliance checks (title must include brand + category)
• eval_samples.jsonl — sample prompts for automated scoring

Images are synthetic but spec-aligned, designed to test context engineering and grounded reasoning.

3. API Specification

POST /analyze
Input: image_gcs_uri, question, context_ids, metadata
Output: JSON with answer, grounding (citations + visual refs), structured (attributes + compliance), diagnostics (latency, tokens, model).

4. Tasks

T1 — Visual Q&A;: Return answer if clear, else 'insufficient visual evidence' (cite R-102).
T2 — Attribute Extraction: Extract color, closure, style, outsole; reconcile with spec.
T3 — Policy Compliance: Evaluate R-201..R-204 and provide pass/fail with evidence.

T4 — Context Adaptation: Swap v1 → v2 brand rules with only context_ids change.
T5 — Edge Cases: Prefer uncertainty/flagging over hallucinations when data conflicts.

5. System Architecture (GCP)

Vertex AI Multimodal (Gemini) — inference.
Cloud Run (FastAPI) — REST API.
Cloud Storage — images + context packs.
Optional: BigQuery (logging), Pub/Sub (batch processing).

6. Context Engineering Requirements

Context Packs must be modular and versioned.
Context Card: compact (≤2K tokens), deterministically citable (IDs like R-101), filtered rules, spec attributes included, JSON schema instruction appended.

7. Evaluation & Scoring

Automated: grounded accuracy, visual fidelity, compliance precision/recall, JSON validity, latency.
Human: clarity, tone, uncertainty handling.
Rubric (100 pts): Context engineering design 25, grounded correctness 25, structured outputs 20, robustness 15, production readiness 15. Pass ≥ 75.

8. Deliverables

• Cloud Run API endpoint (/analyze)
• Repo: context loader/composer, prompt templates, Vertex client wrapper, eval scripts
• Report: metrics (accuracy, compliance, latency) + observations on conflicts/edge cases.