Update Notes for "000-original plan-v2.md"

Date: 2025-10-29

Drift Definition & Thresholds

- Unify to weighted `drift_score ∈ [0,1]` (four guards).
- Online decisions: <0.5 ALLOW, 0.5–0.8 WARN, ≥0.8 ROLLBACK.
- Quality labels (post-hoc): HIGH <0.2, MEDIUM 0.2–0.35, LOW ≥0.35.

Run Instructions

- Replace `runner.sh` with actual scripts:
  - `demo/generate_predictions.py`, `demo/batch_generate_predictions.py`
  - `demo/batch_generate_with_q1_metrics.py`, `demo/compute_drift_from_predictions.py`

Tech Stack (Current vs Plan)

- Current: Rule-based Guards (Scope/Plan/Test/Evidence), SimpleBedrockAgent (Claude via Bedrock).
- Planned in Q2: Vector index + (optional) ML ranker. Move GPT-4o/Qwen/Chroma to "planned" not "present".

Metrics Claims (SOTA)

- Remove or footnote closed-source % numbers until verified with sources. Keep placeholders or cite exact links/dates.

Q2 Retrieval Metric

- Avoid fixed "similarity ≥ 0.7"; define as top-k retrieval + adoption evidence (e.g., Δdrift ≤ 0 or step/template overlap ≥ τ).

Status Banner (Add to Top of Plan)

- Q1 P0 ≈ 85%: rules, monitoring, evaluation, batch tools done; evaluator (Docker) pending.
- Predictions: 408 + 15 done; compute drift on existing logs; finish remaining ~77.
- Q2: Option B (quality-labeled patterns: resolved ∧ drift<0.2).

Hints Usage

- `hints_text` optional for agent/retrieval; exclude from Q1 scoring and evaluation signals.