Project A — AI Agent Plan

Agent name: Overseer (Overture Semantic & Error Reasoning Agent)
Tagline: “Trust every monthly release.”

Overview

Overseer reads the release metrics (row counts, geometry/area stats, attribute coverage, change types, etc.), learns stable patterns across past releases, and flags anomalies with explanations. It compares against a rule-based baseline, writes a comparative analysis, and produces a triage report for humans-in-the-loop.

Scope

Input: CSV metrics + changelog summaries + README/rules (plain text).
Task: Detect anomalies and explain them (global + per-theme + per-type + per-source).
Output: Pass/flag decisions with explanations.
Test: Compare precision/recall/FPR against rule-based systems.

Data Model

Organize each release as a fact table with hierarchical keys.
Metrics: volume, geometry integrity, attribute coverage, and dataset mix.
Normalization per area or population. Exclude removed features for stability.

Architecture

Pipeline:
Ingest → Normalize → Feature Store → Baselines (rules) → Time-Series & Distribution Tests → LLM Reasoning → Severity Scoring → Reports.

Components: Ingestion, Feature Engineering, Rule Baseline, Statistical Tests, LLM Explanations, Severity Scoring, Reporting.

Technical Stack

Python 3.11, Polars, DuckDB, PyArrow, Great Expectations, Evidently, Scikit-learn, Statsmodels, LightGBM, Prophet, MLflow, Plotly, Jinja2, OpenAI API, Dagster, Docker, GitHub Actions, Poetry, Ruff, Black, Mypy.

Implementation Steps

1. Ingestion: parse metrics CSVs, normalize schema.
2. Rule Baseline: define and apply Great Expectations rules.
3. Statistical Tests: z-scores, EWMA, PSI/KL divergence.
4. LLM Reasoning: context-aware explanations.
5. Severity Scoring: combine deviation and context.
6. Reporting: Jinja2 + Plotly HTML reports.

Evaluation

Label anomalies from historical releases.
Metrics: precision, recall, FPR.
Acceptance: ≥20% fewer false positives or ≥15% more true positives.

Roadmap

M1: Ingestion + baseline
M2: Statistical layer
M3: LLM reasoning
M4: Evaluation & dashboard

CLI

overseer validate --release 2025-08-20.1 --metrics ./metrics --out ./reports
overseer backfill --releases-file releases.txt