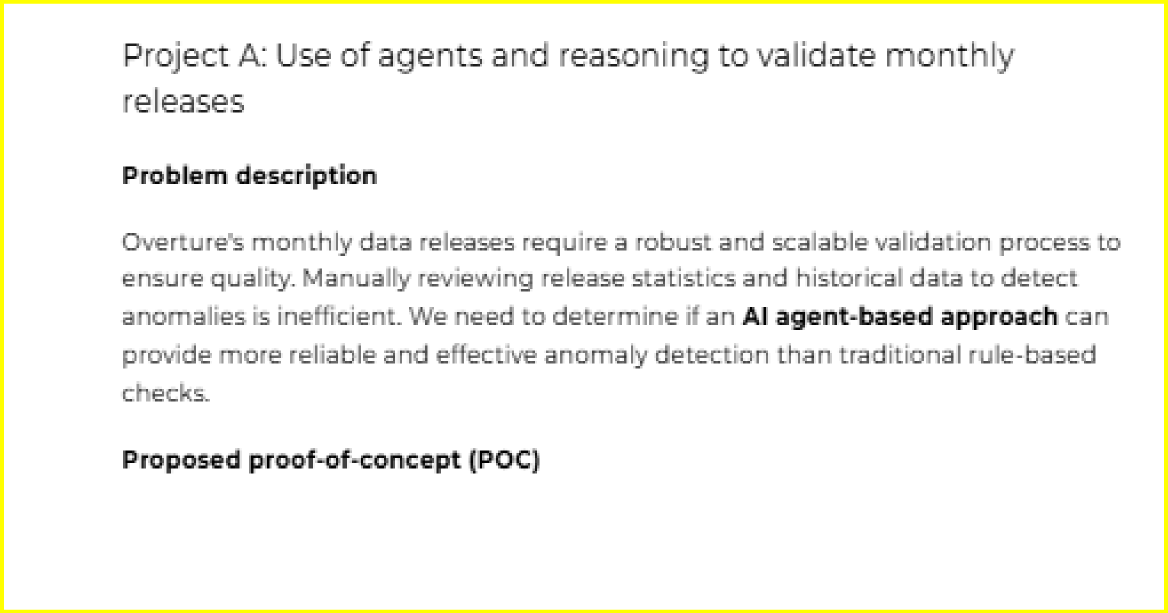
Kartikeya Kumaria
CRWN 102
Oct 14, 2025
OKR Draft

OKR for Terraforma Project A

Project A Description:

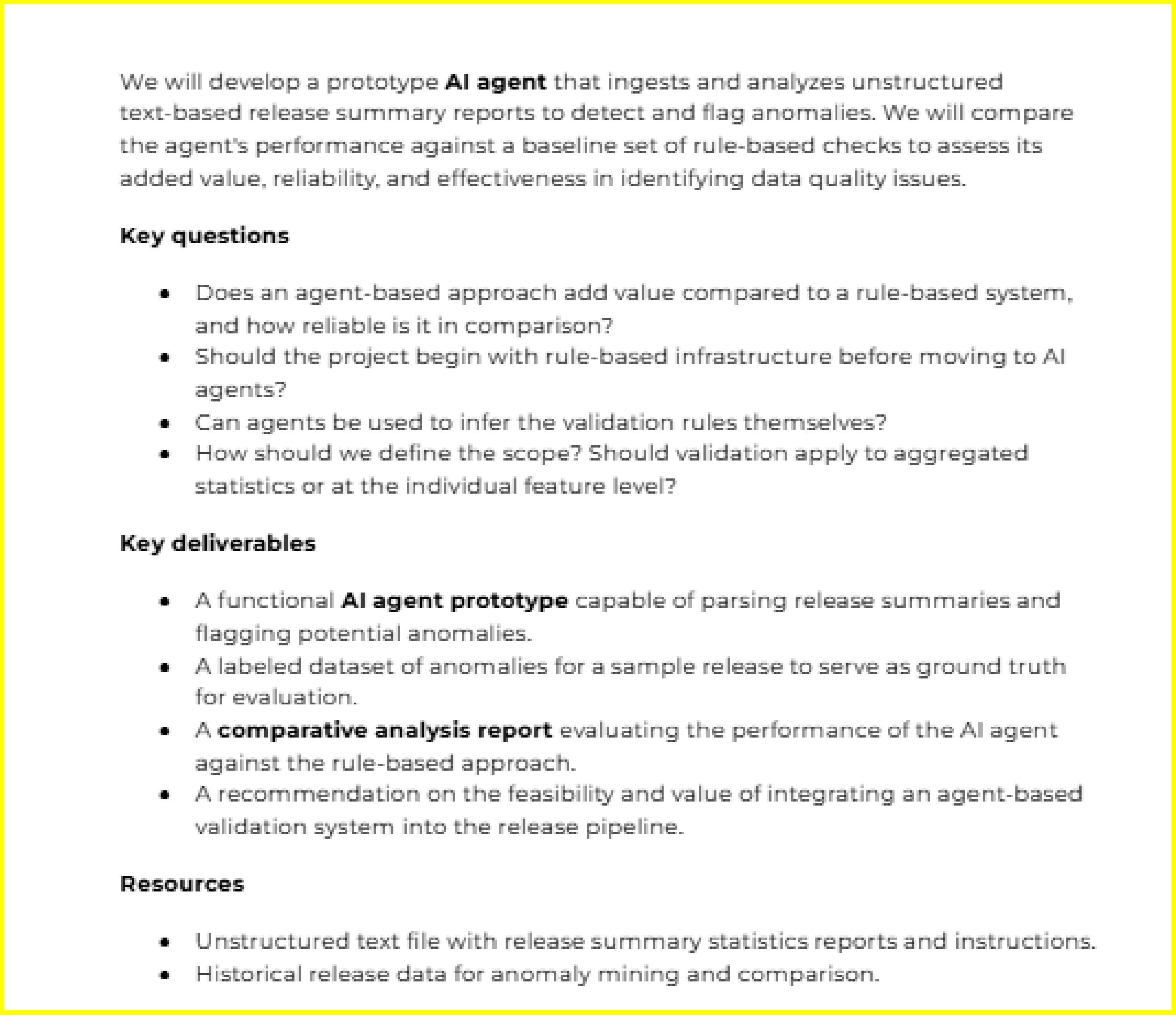


Project A: Use of agents and reasoning to validate monthly releases

Problem description

Overture's monthly data releases require a robust and scalable validation process to ensure quality. Manually reviewing release statistics and historical data to detect anomalies is inefficient. We need to determine if an **AI agent-based approach** can provide more reliable and effective anomaly detection than traditional rule-based checks.

Proposed proof-of-concept (POC)



We will develop a prototype **AI agent** that ingests and analyzes unstructured text-based release summary reports to detect and flag anomalies. We will compare the agent's performance against a baseline set of rule-based checks to assess its added value, reliability, and effectiveness in identifying data quality issues.

Key questions

- Does an agent-based approach add value compared to a rule-based system, and how reliable is it in comparison?
- Should the project begin with rule-based infrastructure before moving to AI agents?
- Can agents be used to infer the validation rules themselves?
- How should we define the scope? Should validation apply to aggregated statistics or at the individual feature level?

Key deliverables

- A functional **AI agent prototype** capable of parsing release summaries and flagging potential anomalies.
- A labeled dataset of anomalies for a sample release to serve as ground truth for evaluation.
- A **comparative analysis report** evaluating the performance of the AI agent against the rule-based approach.
- A recommendation on the feasibility and value of integrating an agent-based validation system into the release pipeline.

Resources

- Unstructured text file with release summary statistics reports and instructions.
- Historical release data for anomaly mining and comparison.

Objectives:

1. Develop a reliable AI agent prototype to automate anomaly detection in Overture's monthly release validation process.

2. Evaluate the AI agent's performance, reliability, and value versus traditional rule-based checks.

Key Results:

For Objective 1: Prototype Development

1. **Deliver a functional AI agent prototype capable of ingesting and parsing at least three months of release summaries.**
   - Must successfully ingest and parse ≥ 90 % of release summaries without errors.
   - End-to-end processing time ≤ 10 minutes per release cycle.
   - Successful run verified and documented by QA.
2. **Enable the agent to detect anomalies in both aggregated statistics and feature-level metrics.**
   - Achieve ≥ 80 % detection accuracy on a labeled validation subset.
   - Cover ≥ 95 % of monitored features.
   - Maintain false-positive rate ≤ 15 %.
   - Validation results logged automatically and reviewed in two evaluation runs.

3. **Build a labeled dataset of at least 500 anomalies to serve as ground truth for model evaluation.**

4. **Integrate a rule-based baseline system to compare against the AI agent's performance.**

- Implement a baseline model that reproduces existing rule-based validation checks with ≥ 95% functional coverage.
- Ensure baseline execution time ≤ 10 minutes per release cycle.
- Record precision and recall metrics for both systems on the same labeled dataset (target: baseline precision ≥ 80 %, recall ≥ 60 %).
- Document comparative performance results in the evaluation report.

5. **Ensure the prototype runs end-to-end without manual data preprocessing and demonstrates stable performance across multiple releases.**
   - The prototype must complete end-to-end execution successfully on ≥ 95 % of release datasets.
   - Processing time should remain consistent within ±20 % variance across runs.
   - The system must automatically handle data ingestion, validation, and anomaly output without human intervention.
   - Log all execution metrics (runtime, errors, data integrity checks) for each run.

For Objective 2: Performance and Evaluation

6. **Demonstrate consistent anomaly detection improvements over the rule-based baseline across multiple release periods.**
   - Conduct an initial exploratory analysis to establish baseline precision, recall, and false-positive rates.
   - Show measurable improvement in at least two consecutive release cycles for either precision, recall, or both (target improvement ≥ 10 % over baseline).
   - Document comparative performance curves and confidence intervals for each run.
   - Summarize key insights on trade-offs between precision and recall in the final evaluation report.

7. **Conduct at least 10 structured evaluation runs using different release data sets to ensure robustness.**

8. **Produce a comparative analysis report summarizing key findings, limitations, and recommendations.**
    - Include quantitative comparison of AI agent vs. rule-based baseline on at least five key metrics (e.g., precision, recall, F1-score, latency, and coverage).
    - Highlight a minimum of three identified improvement areas and two documented limitations.
    - Ensure report completeness and clarity score ≥ 4 / 5 from peer review feedback.
    - Deliver the final version within two weeks of completing the tenth evaluation run.

9. **Present results to stakeholders (data engineering + release QA) and gather at least two rounds of feedback.**
    - Conduct two formal presentation sessions (one initial, one follow-up after revisions).
    - Collect ≥ 80 % attendance from invited stakeholders in each session.
    - Capture all feedback items in a shared tracking document and implement or address ≥ 90 % of actionable comments.
    - Confirm stakeholder sign-off on final evaluation results within one week of the second session.