Exercise C — Production Monitoring Without Ground Truth (Advanced)

Goal

In production, you won't have gold labels. Build a **monitoring and evaluation framework** that detects model/prompt drift and unsafe behavior **without ground truth**, and demonstrates how you would operate a journaling parser safely at scale.

This exercise tests whether you can design **production-grade guardrails** around a non-deterministic model.

Ashwam processes free-form women's health journals across symptoms, food, emotion, and mind domains.

Key constraints:

- **No canonical labels** for symptoms, food, emotion, or mind
- LLM outputs may drift over time (prompt changes, model updates, distribution shift)
- Safety and restraint matter more than recall

Input Provided

You will receive a synthetic, production-like dataset containing:

- A batch of journal texts
- Parser outputs for:
  - **Day 0** (baseline behavior)
  - **Day 1** (drift / breakage behavior)
- A small **canary set** with evidence-grounded gold (no canonical labels)

You may treat these as two consecutive production runs.

Tasks & What You Must Implement

You must build a tool that, given journal texts and parser outputs, computes **invariants, proxy drift metrics, and canary results**, and explains how these would be used in production.

1) Invariants (hard checks — must never fail)

Define and implement **hard rules** that always apply in production.

Your system must compute at least:

- **Schema validity rate**(% of outputs that conform to the expected JSON schema)
- **Evidence span validity rate** (% of extracted items whose `evidence_span` appears verbatim in the journal text)
- **Hallucination rate** (% of extracted items not supported by the source text)
- **Contradiction rate** (same evidence span extracted with conflicting polarity)

For each invariant:

- explain why it exists
- explain what risk it mitigates
- explain what action would be taken if it fails

2) Proxy drift metrics (no labels)

Design and compute **proxy metrics** to monitor system health over time.

Your system must compute and compare (Day 0 vs Day 1):

- **Extraction volume** number of extracted items per journal (distribution)
- **Uncertainty rate** proportion of `unknown` / `uncertain` polarity or buckets
- **Intensity / arousal drift** change in proportion of `high` intensity or `high` arousal items

- **Domain mix drift**
  - distribution across symptom / food / emotion / mind
  - detect sudden surges (e.g., over-parsing mind/emotion)

For each metric:

- define what “normal” looks like
- define what constitutes drift or breakage

---

3) Canary & audit strategy

Implement a **canary runner** using the provided labeled subset.

Your canary logic must:

- run on a fixed, small dataset
- compute a minimal, stable metric set (you choose which)
- trigger:
  - alerts
  - rollback
  - or human review

Explain:

- how often the canary runs
- why the thresholds are chosen
- how this prevents silent degradation

---

4) Human-in-the-loop design

Explain how humans fit into this system:

- What humans review (exactly)
- What humans explicitly **do not** review
- How often reviews occur

- How this scales with limited analyst or clinician time

Restraint and prioritization matter more than coverage.

5) Explainability (role-based)

Describe how you would explain a parsing decision to:

- **A PM** system health, trends, risk signals
- **A clinician** evidence grounding, uncertainty, limitations
- **A user** high-level, non-alarming, trust-preserving explanation

You may answer this in bullets or short paragraphs.

Output Requirements

Your submission must include:

- A **CLI entrypoint**, for example:

  ```
  python -m ashwam_monitor run --data ./data --out ./out
  ```

- Output artifacts such as:
  - `out/invariant_report.json`
  - `out/drift_report.json`
  - `out/canary_report.json`

What We’re Evaluating

- Production realism
- Monitoring and alerting maturity
- Safety-first thinking
- Ability to operate without perfect data

- Clear reasoning and tradeoffs

Data Package - What’s inside

- `data/journals.jsonl` — 20 synthetic “production-like” journals (English + Hinglish)
- `data/parser_outputs_day0.jsonl` — baseline parser outputs
- `data/parser_outputs_day1.jsonl` — drift/breakage outputs with intentional issues:
  - evidence spans not found in text (hallucination/invariant failures)
  - missing required fields on some items (schema validity degradation)
  - domain mix drift (mind/emotion surge)
  - high-arousal rate drift
  - contradiction cases (same evidence span with conflicting polarity)
- `data/canary/`
  - `journals.jsonl` (5 journals)
  - `gold.jsonl` (evidence-grounded labels; no canonical labels)