

Exercise B — Run-to-Run Variance & Stability Analysis

Goal

LLM outputs vary across runs. Build a stability framework that measures whether the system is safe and consistent **without canonical labels**.

You must demonstrate how you would **measure and control variance** when outputs change across runs.

Input Provided

You will receive:

- **5 synthetic journals**
 - For each journal: **3 LLM outputs** from the same prompt (different runs)
 - Use the provided `llm_runs/` (3 runs per journal) and the journals.
-

Tasks

1) Define “stability” formally

- What does it mean for two outputs to be “the same” if the text differs?
- Which fields must be stable vs allowed to drift?

2) Design a matching algorithm

- How do you match semantic objects across runs?
- Primary signal: **evidence span overlap**
- Fallbacks (optional): semantic similarity / heuristic matching / etc.

3) Propose stability metrics (quantitative)

Include at least 3 metrics and define what “good” vs “bad” looks like:

- Agreement rate (matched objects / union)
- Polarity flip rate (present \leftrightarrow absent) — **high-risk**
- Bucket drift rate (intensity/arousal/time changes)

4) Risk framing

- Which variance is acceptable?
- Which variance is dangerous in a women's health context, and why?

5) Production implications

Explain how instability impacts:

- downstream nudges
 - user trust
 - auditability
-

What you must implement

1. A **deterministic matching algorithm** aligning objects across runs:
 - evidence overlap first
 - semantic similarity fallback (optional)
2. A **stability report** computing at least:
 - Agreement rate across runs
 - Polarity flip rate
 - Bucket drift rate

Bonus (optional)

Produce a **single stable final output** from 3 runs:

- majority vote for stable fields
 - abstain / mark uncertain on disagreements
-

What We're Evaluating

- Systems thinking
- Comfort with non-determinism
- Safety awareness
- Practical ML judgment

Data Package - What's inside

- `data/journals.jsonl` — 5 journals
- `data/llm_runs/` — 3 JSON outputs per journal:
 - synonyms/paraphrases in `text`
 - missing objects in some runs
 - bucket drift (intensity/arousal/time)
 - **one deliberate polarity flip on a negated emotion case** (to test risk metrics)