A) Data / logging questions (blockers if unanswered)

1. **What exactly can you record per failing test run?**

- I have exception type, as well as aborted, disabled, or failed (with the Throwable cause as a String).
- For the test harness, I guess I could, for the given test run, just run all of the tests that were run with that code version, since I have the code.

2. **How do you uniquely identify tests across runs?**

- Stable test ID = `className#methodName#parameterCount`

3. **Can you compute diffs deterministically from snapshots?**

- The compute diffs are actually what's stored, and the snapshots can be constructed from them
- I can safely assume no new files are created for the assignment

4. **Do you have per-test runtime (or at least per-suite runtime)?**

- I'll ask the https://www.timecomplexity.ai/ what the run time of the code solution is, at least for the time being

5. **Are timestamps trustworthy enough for episode splitting?**

- I'm using System.currentTimeMillis(), but everyone is in the same timezone, and I'm really more concerned about relative time

B) Phase 1 product behavior questions (must decide)

6. **What is the exact unit of "episode" in Phase 1?**

- Time-gap + focus-window category shift
- "Wrong" episode boundaries are fine as long as the timeline is roughly right

7. **When you say "tests improved/worsened", what's the baseline?**

- vs episode start and vs very first run
- The run-level delta vs previous run and episode-level delta vs episode start

8. **How will you pick the "representative failing test" per episode (even without traces)?**

- I’m going to have there be multiple (any that fluctuated between passing and failing, so long as it ended up failing)

9. **What’s your Phase 1 root-cause label policy when evidence is weak?**

- I’ll allow “Unknown / insufficient evidence”

10. **What are the confidence levels and what triggers them?**
Simple rubric (MVP-friendly), e.g.:

    - High: direct signal (NPE → null handling; IndexOutOfBounds → boundary)
    - Medium: assertion expected/actual pattern + category match
    - Low: LLM inference based on sparse assertion text only

11. **Will test case categories be manual or LLM-generated for MVP?**

- LLM-first and then adjusted manually

12. **How will you compute diff categories in MVP?**

- purely LLM-based classification on the code changes

C) Engineering / system questions (important for not burning the week)

13. **What is the canonical data schema you’ll store and replay?**

- Please give me schemas for Run, TestResult, DiffSummary, Episode

14. **Where will you compute things: offline batch or on-demand?**

- Batch compute after submission

15. **What is your caching strategy for LLM calls?**

- Probably none, since I won’t be sending the same query twice

16. **How do you handle missing / inconsistent evidence gracefully in the UI?**

- compile error: if no tests ran, it will state that there was a compile error
- crash before assertions: similar to with compile errors, will just state the test failed
- partial runs: if only some tests are executed, then only they will be displayed