

Internal Evals: The First 90 Days

From escalation mining to a CI-gated regression suite — the checklist.

AI Agents · Evals · LLM · Companion to <https://aabee.tech/blog/internal-eval-infrastructure-from-failures-to-regression-suite/> · October 2026

First 90 days

1. **Weeks 1–2:** mine escalations → 10–20 named failure patterns, traces attached
2. **Weeks 3–4:** cheapest mechanical check per case; validate any LLM judge on ~100 human labels
3. **Month 2:** suite in CI; no release on regression; replay 1% of traffic nightly
4. **Month 3:** new escalations become cases within a week; dashboard: score by case, trend by release

Grading ladder (cheapest that works)

- ☐ Exact string / ID match
- ☐ JSON schema validation
- ☐ Unit-test the computation
- ☐ LLM judge — only for tone/coverage, calibrated first

Statistics discipline

- ☐ Small n: report ranges, not point scores
- ☐ Bimodal results: split the population before averaging
- ☐ Every case links to its production trace

The product is decision speed: a new model version gets a per-failure-mode verdict within a day.

aabee.tech — AI agent engineering and software performance engineering, delivered as measurable outcomes. IT Park Uzbekistan Resident.

Questions or a project where this applies? b2b@aabee.tech · aabee.tech