

Evaluating AI Systems: Why Testing Never Stops

FREE EDUCATIONAL BRIEF · LEARN AI. UNCOVER THE DATA. PROTECT THE TRUTH.

Traditional software either passes a test or fails it. AI systems answer slightly differently every time, on inputs you never anticipated, using models that change under you when providers ship updates. Evaluation therefore isn't a phase before launch — it is a permanent operating function, and it's what separates dependable AI from a demo that decays.

THE EVALUATION TOOLKIT

» Golden sets

A curated collection of real inputs with known-good answers, re-run on every change to catch regressions.

» Rubric scoring

Grading outputs against explicit criteria — accuracy, tone, format, safety — by humans, or by a stronger model whose grades are spot-audited.

» Adversarial testing

Deliberate attacks: prompt injection attempts, off-topic bait, edge-case inputs, hostile phrasing.

» Production monitoring

Live sampling of real traffic, user feedback signals, cost and latency tracking, drift alarms.

THE SILENT REGRESSION PROBLEM

The most dangerous failure is the quiet one: a provider updates a model, a prompt edit shifts behavior, a document update poisons retrieval — and quality drops with no error message. Golden sets on a schedule are the smoke detector. Teams that skip them discover regressions from customer complaints instead of dashboards.

A RIGHT-SIZED PRACTICE

You don't need a research lab. Fifty representative test cases, a simple rubric, a weekly automated run, and a named owner will catch the majority of quality drift for a typical business system. SMI builds this into automation deliveries because an unmeasured system is an unmanaged one.

INVESTIGATOR'S TAKEAWAYS

- AI quality is a distribution — measure it, don't assert it.
- Golden sets re-run on every change are your regression alarm.
- Test adversarially before your users (or attackers) do.
- Monitor production continuously; models change beneath you.
- Even 50 good test cases with an owner beats a thousand vibes.