SINCE 1994

Evaluation *before* shipping.

HOW TO TEST AN AI APPLICATION BEFORE IT HITS PRODUCTION · REX BLACK, INC.

"The demo worked" is *not* a release gate.

Unit tests catch what the author thought of.
Production catches what the author *did not*.

Generative systems break the binary.
Non-determinism, open output space, open failure modes.

Five dimensions. Score each *independently.*

DIMENSIONS

Correctness. Grounded accuracy on the golden set.
Groundedness. For any RAG or tool user, can the answer be traced?
Safety. Refusal rate for off-policy, jailbreak survival.

DIMENSIONS (CONT.)

Cost & latency. Per-decision cost, P50 and P95.
Regression stability. Does yesterday's score hold after a prompt edit or model upgrade?

The golden set.

Small. *Ruthless*. Maintained.

SIZE

60 to 150 examples per feature.
Below 60: too noisy.
Above 150: stops getting looked at.

COMPOSITION

Happy path **~40%**.
Edge cases **~30%**.
Adversarial **~20%**.
Known regressions **~10%**, grows every escape.

LLM-as-judge.
Useful. And *dangerous*.

Three failure modes to design around:

Position bias. Judges prefer the first answer shown. Score both orderings.
Length bias. Judges prefer longer answers. Put this in the rubric.
Self-preference. Same-family judge inflates by measurable margin. Cross-family judges where cost allows.

The release gate.

A *named boolean.*

RULES

Every gate has a named owner.
Gates block merges, not deploys.
Thresholds are explicit.

TYPICAL THRESHOLDS

Correctness floor **85%**.
Groundedness floor **95%**.
Adversarial refusal **90%**.
Hallucination ceiling **2%**.
P95 latency **3s**.

Rollout.

Canary. *1%. 10%. 100%.*

STAGES

Canary. Internal. 48 to 72 hours.
1% cohort. Real customers, randomized.
10% cohort. Broader, balanced.
100%. Hold for two weeks before declaring stable.

DISCIPLINE

Each stage has exit criteria.
Criteria are *observable* before promotion.
Do not telescope. Canary is not production.

Once it ships.
The question *flips*.

"Does it still work?" replaces *"is it ready?"*

Four drift sources: **model, context, prompt, upstream.**

The defense: **weekly regression run.** CI job, golden set, dashboard, Slack. Twenty lines of code. Catches slow rot.

What this changes about your *QA org*.

Scripted acceptance testing **does not disappear.** It shifts.

The new work: **rubric design, golden-set curation, release-gate ownership.**
The new QA engineer: authors 100 examples, specifies a judge rubric, owns the gate thresholds.

This week.

Write down the current release gate for your AI feature.
Build a 60-example golden set for the highest-blast-radius feature.
Schedule the weekly regression run, even if the rubric is rough.

Shipping the rhythm is more valuable than perfecting the artifacts.

REX BLACK, INC.

Evaluation is the new QA.

REXBLACK.COM/RESOURCES/WRITING/AI-EVALUATION-BEFORE-SHIPPING