

Evaluation Charter

Purpose

The Evaluation Charter declares what a GROW-built system must prove before it is allowed to act on behalf of a user, a constituent, or an operating business. It is the standing policy that separates a working demo from an operational system. Until a system has cleared the evaluation gate defined here, it is unfit for production traffic, regulator-visible output, or any decision with downstream cost.

What "Evaluated" Means in GROW

A system is evaluated when, and only when, all four conditions hold:

- 1. Coverage of failure modes. Every entry in the consumed s1-failure-mode-register with severity critical or high has at least one safety test traceable by failure_id. Medium and low entries have either a test or a documented monitoring signal.
- 2. Reproducibility. Every eval run produces a provenance record conforming to s2-audit-trail-schema. Re-running the same inputs against the same artifact version produces the same provenance shape (not necessarily identical text, but identical decision_origin path for deterministic branches).
- 3. Calibrated scoring. Pass thresholds in s2-scoring-system are tied to a stated risk tolerance, not to whatever number the model happens to produce.
- 4. Named evaluators. Each test family in s2-test-architecture is signed off by a named role from s2-evaluator-roster. Sign-off is recorded, not implicit.

Demo vs. Operational Quality

Dimension	Demo Quality	Operational Quality
Inputs tested	Curated happy path	Sampled from real traffic plus adversarial fixtures
Failure visibility	"It worked when I tried it"	Logged provenance for every run, queryable by failure_id
Sign-off	Builder's own judgment	Named evaluator role with authority scope
Regression posture	None	Comparison harness per s2-regression-discipline
Recovery	Restart the demo	Defined fallback, escalation, and override paths

A GROW system may ship at demo quality only for internal exploration. Anything touching a constituent, a regulator, a paying guest, a podcast publication pipeline, or money must be at operational quality.

High-Stakes Decision Map

Evaluation rigor scales with decision stakes. A decision is high-stakes if any of the following is true:

- It is regulator-visible (public comment, FASD filing, permit response, DEP/HOA correspondence).
- It is financially binding (rate quote, refund, payout, contract acceptance).
- It is irreversible within 24 hours (sent email, published episode, posted vote).
- It involves personal data of a constituent, guest, or applicant.
- It substitutes for a statutorily required human judgment.

High-stakes decisions require: critical-severity failure-mode coverage, a HITL gate before commit, source_confidence of high on every retrieval claim cited, and a named human override path in the provenance record.

Risk Mapping

Every test in s2-test-architecture is tagged with the risk(s) it retires:

- User risk — harm, confusion, exclusion, or loss to the end user.
- Regulatory risk — non-compliance with Florida statutes, sunshine law, ADA, fair-housing, GDPR/CCPA, IRS, or grant covenants.
- Operational risk — downtime, cost spike, data loss, contractual breach.

A test family that retires none of these is not an eval; it is telemetry.

Worked Example: Civic Permit Triage Agent

A South Walton special-district agent that triages incoming dock, dune-walkover, and stormwater permit inquiries.

- High-stakes? Yes — regulator-visible, statutorily framed by FS § 161 and FS § 373.
- Demo behavior: classifies inquiries into routing buckets.
- Operational evaluation requires:
 - Safety tests seeded from s1-failure-mode-register covering misrouting a wetland inquiry to a non-jurisdictional reply (critical), and citing a statute that does not apply (critical).
 - HITL gate before any reply that cites a statute number; gate captured in decision_trace with decision_origin = human-override when commissioner staff modifies the draft.
 - Source_confidence of high required on every statute citation; medium or below routes to escalation per s2-evaluator-roster.
 - Regression harness fires when the routing prompt changes or when FS revisions land.

Authority

This Charter is binding within the GROW library. Builders may propose amendments via the regression-discipline change log. Builders may not ship past the evaluation gate without the artifacts named in this Charter being present, populated, and signed.