

Failure-Mode Register

Canonical taxonomy of agent failure modes for the GROW library. Course 2 (S2) generates eval test seeds directly from `failure_id` and `detectable_signal`. Course 3 (S3) consumes the same ids in provenance traces. Do not rename or repurpose `failure_id` values without a coordinated bump across S1/S2/S3.

Schema (contract C1):

```
- failure_id: <kebab-case>
  name: <short title>
  description: <one to three sentences>
  severity: <critical | high | medium | low | info>
  detectable_signal: <observable telemetry or output pattern>
  expected_behavior: <what the agent must do when this is detected>
  recovery_path: <state transition or operator action>
```

Register

```
- failure_id: tool-timeout
  name: Tool or connector timeout
  description: A tool call exceeds its per-call timeout budget without returning a status.
  Common with slow GIS, CRM, or county-records connectors.
  severity: high
  detectable_signal: tool_call.duration_ms > tool.timeout_ms AND no terminal status frame
  received
  expected_behavior: Cancel the call, mark the step as degraded, attempt the deterministic
  fallback before any retry.
  recovery_path: fallback -> bounded-retry -> escalate

- failure_id: connector-auth-failure
  name: Connector authentication or authorization failure
  description: OAuth token, API key, or scope is invalid, expired, or revoked.
  severity: high
  detectable_signal: HTTP 401/403 from connector OR auth-refresh exception
  expected_behavior: Stop the run for this connector, do not retry with stale credentials,
  surface a structured escalation.
  recovery_path: escalate-to-operator -> credential-refresh -> resume

- failure_id: unsafe-action-attempted
  name: Unsafe or out-of-scope action attempted
  description: Agent proposes a tool call that crosses the irreversible-impact boundary
  defined in the Operating Context Canvas.
  severity: critical
  detectable_signal: planned_action matches irreversible_impact_boundary list OR fails
  policy filter
  expected_behavior: Hard stop. Refuse the action. Do not attempt a softer variant. Emit
  override event with decision_origin=fallback.
  recovery_path: terminate-run -> HITL gate -> postmortem

- failure_id: low-confidence-routing
  name: Low-confidence routing decision
  description: Agent selects a route with confidence_band=low or unknown when the operating
  canvas requires medium or higher.
  severity: medium
  detectable_signal: routing_confidence < threshold OR confidence_band in {low, unknown}
  expected_behavior: Escalate to HITL queue with the top two candidate routes and evidence.
  recovery_path: HITL-review -> resume-with-human-decision

- failure_id: false-success-report
  name: False success report
  description: Agent reports success but the verifying check (write-confirm, schema
```

```
validation, downstream ack) did not pass.
  severity: critical
  detectable_signal: agent.status=success AND post_condition_check.passed=false
  expected_behavior: Override the success report, mark the run as failed, emit an evidence-
backed correction.
  recovery_path: rollback-if-possible -> escalate -> add to S2 regression suite

- failure_id: looping-retry
  name: Looping retry without progress
  description: Same tool call or planning step is repeated more than N times without state
delta. Indicates planner-tool deadlock.
  severity: high
  detectable_signal: same step_hash count > retry_cap within window OR stagnant state hash
  expected_behavior: Break the loop, emit a stop event, route to fallback or HITL.
  recovery_path: terminate-loop -> fallback -> escalate

- failure_id: state-corruption
  name: State corruption
  description: Agent memory, scratchpad, or tool-state references an entity that no longer
matches the system of record.
  severity: high
  detectable_signal: id_mismatch on read-back OR version-stamp drift > tolerated_age
  expected_behavior: Discard local state, re-hydrate from system of record, log the
corruption event.
  recovery_path: rehydrate -> retry-once -> escalate

- failure_id: schema-drift-input
  name: Input schema drift
  description: An expected input field is missing, renamed, or has changed type relative to
the Operating Context Canvas.
  severity: high
  detectable_signal: schema_validation.failed=true OR unexpected_keys_count > 0
  expected_behavior: Halt the run for this input, do not coerce silently, notify the schema
owner.
  recovery_path: quarantine-input -> escalate -> schema-update-task

- failure_id: prompt-injection-detected
  name: Prompt injection or instruction override detected
  description: Untrusted content (email body, scraped page, applicant note) attempts to
alter agent instructions or exfiltrate data.
  severity: critical
  detectable_signal: injection-classifier score > threshold OR pattern match against known-
injection rules
  expected_behavior: Quarantine the input, do not act on injected directives, log raw
payload to provenance.
  recovery_path: quarantine -> HITL-review -> add to adversarial suite

- failure_id: rate-limit-exceeded
  name: Rate limit or budget exceeded
  description: Connector, model, or per-tenant budget cap is hit mid-run.
  severity: medium
  detectable_signal: HTTP 429 OR token/cost ledger > cap
  expected_behavior: Pause the run, emit a deferred-execution event, do not silently fail
the user-visible step.
  recovery_path: backoff -> resume-or-defer -> notify-owner

- failure_id: hallucinated-citation
  name: Hallucinated citation or evidence pointer
  description: Agent emits a citation (statute, parcel id, grant number, URL) that does not
resolve in the source of record.
  severity: high
  detectable_signal: citation-resolver.resolved=false OR hash-of-cited-doc not in evidence
store
  expected_behavior: Strip the citation, mark the output as evidence-incomplete, route to
HITL if the output is user-facing.
  recovery_path: re-ground -> HITL-if-needed -> regression-seed

- failure_id: stale-data
  name: Stale data used in decision
```

```
description: Agent relies on a cached or upstream record older than the freshness budget
for this decision class.
severity: medium
detectable_signal: record.age > freshness_budget for the routing class
expected_behavior: Force a refresh; if refresh fails, demote confidence_band and re-route.
recovery_path: refresh -> re-evaluate -> escalate-if-stale

- failure_id: pii-leak-risk
  name: PII or restricted-data leak risk
  description: Output or tool call routes restricted data (applicant PII, ePHI, sealed
records) to a connector not authorized for that class.
  severity: critical
  detectable_signal: data_class in {pii, ephi, sealed} AND destination not in
authorized_sinks
  expected_behavior: Block the egress, redact, emit an unsafe-action event.
  recovery_path: block -> redact -> HITL + compliance review

- failure_id: ambiguous-intent
  name: Ambiguous user or upstream intent
  description: The triggering request maps to two or more valid agent objectives with
comparable confidence.
  severity: low
  detectable_signal: top1_intent_score - top2_intent_score < margin
  expected_behavior: Ask one clarifying question OR escalate; do not pick silently.
  recovery_path: clarify -> resume

- failure_id: silent-degradation
  name: Silent quality degradation
  description: Output quality drops (eval score, override rate) without a triggering
exception. Often model or prompt drift.
  severity: medium
  detectable_signal: rolling_override_rate > baseline + 2 sigma OR S2 eval delta < threshold
  expected_behavior: Open a drift incident, freeze deploys for this agent, route to S3
provenance review.
  recovery_path: freeze -> investigate -> rollback-or-rerelease
```

Add new entries by appending; never re-use a failure_id. Severity changes require a version bump and a note in the S2 regression seed manifest.