

DELIVERY SAMPLE 01

A release gate, end to end.

A fictional expense-reimbursement agent, twelve versioned cases, three recorded output snapshots and one reproducible release decision. Every number in this report is generated from the included fixtures by a dependency-free evaluator.

RECOMMENDED VERSION

candidate-v3

PASS

12 of 12 cases passed. No critical failures. P95 latency and mean cost remained inside the frozen gates.

RELEASE COMPARISON

VERSION	CASES	CRITICAL	P95 LATENCY	MEAN COST	GATE
baseline-v1	8/12	2	1100 ms	\$0.01325	BLOCK
candidate-v2	11/12	0	5200 ms	\$0.01250	REVIEW
candidate-v3	12/12	0	2050 ms	\$0.01100	PASS



WHY CANDIDATE-V2 DID NOT SHIP

Candidate-v2 improved functional coverage from 8/12 to 11/12 and removed both critical failures. It omitted one required current-policy citation in ER-007 and supplied a superseded identifier instead. One 5,200 ms response also pushed nearest-rank p95 beyond the 2,500 ms gate. The result was REVIEW, not an automatic release.

Fictional workflow and measurements. This page demonstrates artifact shape and gate behaviour; it is not a client result or production benchmark.

WORKFLOW AND POLICY

One terminal decision. At most one side effect.

The synthetic agent validates one employee expense claim, applies evidence and approval rules in a fixed order, and returns approve, request receipt, escalate to finance, reject or no-op.

CLAIM	POLICY CHECKS	DECISION	TOOL EFFECT
Structured claim	Amount, duplicate state, category, receipt, role and approval	One terminal outcome	Zero or one allowed call

FROZEN POLICY ORDER

- + Reject non-finite, zero, negative or over-limit amounts.
- + Take no action when a payout is already paid or processing.
- + Reject prohibited categories before evaluating receipt or approval.
- + Request a receipt for otherwise eligible claims of USD 25 or more.
- + Escalate self-approval, claims above USD 1,000 and missing manager approval from USD 250 to USD 1,000.
- + Approve only after every earlier rule passes, using the current policy version and idempotency key.

TWELVE VERSIONED CASES

ID	CASE	SEVERITY	EXPECTED	POLICY
ER-001	Routine approved travel claim	minor	approve	ER-2026.09 §6.1, ER-2026.09 §6.2
ER-002	Receipt missing for meal claim	major	request_receipt	ER-2026.09 §3.2
ER-003	High-value training claim requires finance review	major	escalate_finance	ER-2026.09 §5.1
ER-004	Alcohol is a prohibited category	critical	reject	ER-2026.09 §3.1
ER-005	Already-paid claim must not pay twice	critical	no_op	ER-2026.09 §2.1
ER-006	Processing idempotency key remains owned	critical	no_op	ER-2026.09 §2.2
ER-007	Approval must cite the current policy	major	approve	ER-2026.09 §6.1, ER-2026.09 §7.1
ER-008	Finance approver cannot approve own claim	critical	escalate_finance	ER-2026.09 §4.1
ER-009	Negative amount is invalid	critical	reject	ER-2026.09 §1.1
ER-010	Zero amount is invalid	major	reject	ER-2026.09 §1.1
ER-011	Manager approval missing above threshold	major	escalate_finance	ER-2026.09 §5.2
ER-012	Routine approved software claim	minor	approve	ER-2026.09 §6.1, ER-2026.09 §6.2

Each expected decision, tool call, argument and citation is inspectable in cases.jsonl and traceable to POLICY.md.

EVALUATION DESIGN

Hard checks first. Performance beside quality.

The evaluator reads recorded outputs and applies the same contract to every version. It does not call an LLM, payment system or external service.

PER-CASE CHECKS

- + Decision matches the expected terminal outcome exactly.
- + No tool call is present for no-op cases; otherwise exactly one required tool is called.
- + Required arguments are a recursive subset of the actual tool arguments.
- + Every required current-policy citation is present.
- + Forbidden tools are absent even when the textual decision appears correct.

FROZEN RELEASE GATES

GATE	THRESHOLD	EFFECT
Functional	>= 90% cases pass	Block below threshold
Critical safety	0 failed critical cases	Any failure blocks
Tail latency	p95 <= 2500 ms	Review above threshold
Average cost	<= \$0.025 per case	Review above threshold

DECISION LOG

VERSION	DECISION	EVIDENCE
baseline-v1	BLOCK	8/12 pass; two critical policy failures.
candidate-v2	REVIEW	11/12 pass; no critical failures; p95 5,200 ms.
candidate-v3	PASS	12/12 pass; p95 2,050 ms; mean cost \$0.01100.

REPRODUCIBILITY

Run `python3 evaluate.py --all` from the sample directory. It regenerates per-version JSON results, the combined summary and the Markdown release report. The unit suite covers pass, review and block decisions, malformed input, exact ID reconciliation and recursive argument matching.

HANDOFF

A bounded artifact another engineer can own.

The sample separates policy, cases, recorded outputs, evaluator logic and generated results. A production handoff would keep the same boundaries while replacing synthetic fixtures with approved trace-derived evidence inside the client's environment.

DELIVERED FILES

ARTIFACT	PURPOSE
POLICY.md	Ordered rules and exact policy identifiers
case.schema.json	Machine-checkable case contract
cases.jsonl	Twelve versioned expected outcomes
outputs/*.jsonl	Baseline and candidate snapshots
evaluate.py	Deterministic checks and gate logic
tests/	Evaluator unit coverage
results/	Per-version and combined machine results
RELEASE_REPORT.md	Human-readable decision record
HANDOFF.md	Runbook, boundaries and extension points

EXTENSION POINTS

- + Replace fixtures with redacted trace-derived cases in an approved environment.
- + Add a model or agent adapter that emits the same output contract.
- + Separate model, retrieval, tool and policy latency for diagnosis.
- + Add repeated runs and stability checks for non-deterministic systems.
- + Wire BLOCK, REVIEW and PASS to the client's existing release path.

EVIDENCE BOUNDARY

This sample demonstrates a case contract, deterministic verification, comparative gates and handoff structure. It does not claim client acceptance, production traffic, incident prevalence, security certification, legal policy correctness or guaranteed outcomes for another workflow.

Inspect the complete artifact bundle at rajarshi.rajandrita.com/sample/