

The Golden Gate: Correct Abstention and Reproducible Semantic Control for Agent Systems

John Komkov, Independent Researcher

July 2026

FRONTIER RECORD — ACCESSION GATE OPEN

Adjacent evidence, not an accession to the Technical Spine. Unmet accession gate: Reviewer-originated replay, independent implementation parity, and external adjudication remain open.

LIFECYCLE active-draft | EVIDENCE not-assessed | RECORD /research/golden-gate
BUILT FROM research-status.yaml @ 2026-07-22

Abstract. Agent protocols frequently face disagreements that are neither ordinary execution errors nor questions a model should silently resolve. A seam may lack a definition in the shared language, depend on absent authority, remain open under the declared model class, admit several institutionally inequivalent minima, or exhaust a bounded search. We present Golden Gate, a methods framework for evaluating whether a semantic-control system accepts, refuses, requests evidence, preserves choice, or abstains for the correct checkable reason. The framework separates machine, property, and adjudication oracles; distinguishes participant-blind author-known tests from reviewer-originated blindness; places oracle material under threshold reviewer custody; requires clean-room reproduction; and combines metamorphic relations, explicit state-space coverage, mutation analysis, provenance cards, and non-scalar burden reporting. The protocol is designed so that incomplete external participation yields a typed blocker rather than an internally manufactured result. This paper specifies the method. External results are intentionally absent until a frozen candidate is evaluated under reviewer control.

1. The evidence problem

A system that issues receipts cannot establish the value of stranger-verifiable receipts solely by checking its own artifacts. Captive evaluation remains useful: it removes shallow defects, maps boundaries, and makes scarce reviewer time productive. It does not provide distributional independence, incentive independence, unknown-unknown discovery, third-party attestation, or the recursive product demonstration in which a stranger verifies without trusting the issuer.

Golden Gate treats external control as a property of an evidence-producing process. It is not a label applied to a hidden file. The process binds a candidate before case generation, gives case and oracle control to role-disjoint curators, binds both implementation outputs before reveal, and preserves disagreement rather than laundering it into an oracle.

2. Correct abstention

Conventional selective classification often frames abstention as a response to low predictive confidence. Semantic control has several categorically different reasons not to accept:

- a fixed-language countermodel establishes non-definability;
- operative or supersession authority is missing;
- the closure warrant is open or has changed;
- independently warranted constraints conflict;
- several inequivalent choices lack authored priority;
- bounded enumeration is incomplete;
- constitutionally permitted evidence could resolve the residue.

Collapsing these outcomes to a generic reject option destroys operational information. Golden Gate retains typed exits and scores them only after confirmed unsafe acceptance, unsafe refusal, and unauthorized governance selection. `ESCALATE`, `CHOICE_REQUIRED`, and `INDETERMINATE` remain separate. Evidence requests are evaluated separately from terminal abstentions.

3. Oracle strata

`MACHINE` cases have a canonical result or certificate that an independent checker can decide. `PROPERTY` cases are decided by a trace or state invariant. `ADJUDICATION` cases lack an internally manufactured correct answer; the two primary judgments and any disagreement are themselves evidence.

This separation prevents a benchmark author from disguising institutional choice as ground truth. It also makes clean-room parity interpretable: implementation agreement is required on machine/property cases, while adjudication cases measure whether the system safely preserves unresolved authority and judgment.

4. Blindness and custody

Golden Gate distinguishes `AUTHOR_KNOWN_PARTICIPANT_BLIND` from `REVIEWER_ORIGINATED_BLIND`. The latter requires three curator-custodians, 36 cases generated after candidate freeze, three reviewer-held encryption key groups with threshold two, no implementation-team decryption key, receipt-bound submissions from Bulla and an independent implementation, and a post-submission reveal.

High-entropy per-case nonces prevent brute forcing low-entropy exit commitments. The transcript records conflicts, compensation, organizational ties, and prior exposure. Compensation is not reinterpreted as incentive independence.

5. Found data and adjudication

The found-data lane uses schema artifacts authored by MCP owners for their own purposes. It provides partial distributional independence without pretending that candidate producer-consumer edges are demonstrated deployments. Every case has a provenance card binding capture, content, transformations, redistribution status, question, and adjudication state.

Two blinded primary ratings are required per case. A third diagnostic rating may explain disagreement but cannot overwrite it. “Direct public harvest” is reserved for the preregistered minimum of redistributable direct captures and independent owners; otherwise the evidence remains a seed-corpus result.

6. Strong internal evidence

Metamorphic testing declares exact invariants and forcing changes before execution. Canonical byte equivalence demands equal hashes. Logical equivalence demands equal semantic exits and protected consequences, not equal package hashes. Authority, closure, protected-signature, ambiguity, conflict, and evidence transformations must force their declared fail-closed result.

The economic lane replaces randomized-trace volume with a finite abstract model at reserve guard boundaries. Reports expose reachable states, accepted and rejected transitions, guard pairs, causes, terminal phases, shortest witnesses, and fairness assumptions. A separate two-commitment model tests shared collateral, double pledge, verifier failure, and correlated insolvency. Mutation analysis requires a non-equivalence witness for every denominator exclusion.

7. Reproduction and portability

The clean-room implementation receives the profile and blinded vectors, not Bulla code or expected outputs. Portability is an observation: a configured CI cell without an archived run is `BLOCKED_NOT_OBSERVED`. Each observed cell binds OS, architecture, locale, timezone, runtimes, dependencies, memory, and artifact hashes. Raw solver proofs may vary only if they normalize to the same independently checked certificate.

8. Limitations

Golden Gate cannot convert a bounded closure warrant into open-world safety. Found schemas are not production traffic. An abstract reserve model is not custody or collectibility. Independent technical reproduction is not independent legal or domain validation. Typed abstention can preserve a decision for an authorized forum but cannot supply the missing authority. Finally, an external gate can remain blocked indefinitely; the method treats that as an evidence state, not a reason to relax the constitution.

9. Relation to prior work

Selective classification and reject-option systems already study risk-coverage tradeoffs and context-dependent abstention. Golden Gate does not claim abstention as a new category. Its narrower object is the reproducible distinction among proof-bearing non-definability, missing authority, open closure, institutional non-uniqueness, and bounded resource exhaustion. Stateful coverage-guided fuzzing similarly supplies the scheduler precedent; the new object here is a receipted semantic/finality transition system with explicit authority and closure causes. Threshold custody uses existing SOPS key-group recovery rather than a new cryptographic protocol. Build provenance uses GitHub’s artifact-attestation mechanism rather than a Bulla-specific supply-chain format.

References:

- Gangrade, Kag, Saligrama, “Selective Classification via One-Sided Prediction”, AISTATS 2021.
- Tayebati et al., “CAP: Conformalized Abstention Policies for Context-Adaptive Risk Management for LLMs and VLMs”, ACML 2025.
- Ba et al., “Stateful Greybox Fuzzing”, USENIX Security 2022.
- SOPS threshold key groups.
- GitHub artifact attestations.

10. Status

The methods, types, internal suites, and role packets are implemented experimentally. External cases, ratings, custody, reveal, clean-room parity, and non-local portability observations are pending real participants. Results will be added only after the preregistered reveal.