

Critics-as-Sensors

Abstention-Gated Selection for Verifier-Free Discovery

Synthesis note / bridge across the regulation series

Tyson Jeffreys

Independent Researcher

tyson@staygolden.dev

Version 0.1 — February 2026

Bridge note. This short note operationalizes the “verifier gap” as a concrete selection-and-governance layer for non-verifiable tasks (research synthesis, strategy, writing, design tradeoffs). It connects three existing claims from the regulation series: (i) analysis-layer artifacts reduce time-to-analysis by exposing causal structure and falsifiers, (ii) concept containers regulate representation-level commitments as reusable compressed structure, and (iii) trustworthy discovery requires band-limited optimization with explicit posture control and rollback discipline.

Abstract

Many high-impact tasks lack crisp verifiers: there is no reliable ground-truth checker that can be applied during reasoning or selection. In these domains, systems often fail not because information is missing, but because selection among competing causal stories is unconstrained, leading to thrash, cycling, or premature compression.

We propose a missing operational layer: **critics-as-sensors**. Learned judges/critics can be useful, but they must be treated as *noisy telemetry* rather than authorities. The core primitive is **abstention-gated selection**: bounded candidate generation plus bounded pairwise selection (tournament-style), with **tie/abstain mass** acting as a first-class uncertainty signal that triggers evidence collection rather than further synthesis. We then specify minimal critic governance (versioning, drift checks, and a replay suite) to prevent silent preference rotation and exploitation.

The result is a falsifiable blueprint for verifier-free discovery under regulation: selection becomes a controlled procedure, and commitments (containers, summaries, plans) become gated commits rather than free-running narratives.

1. Introduction

Most “research assistants” accelerate retrieval and summarization. But the bottleneck in real work is often different: selecting and committing to a defensible structure when correctness is not directly checkable.

This note isolates that condition as the **verifier gap** and proposes an operational response that fits inside a regulated agent stack.

2. The verifier gap

A **verifier** is any procedure that can reliably determine correctness (or assign a stable reward) for candidate outputs. Many tasks do not provide such a procedure at decision time:

- research synthesis (conflicting sources, incomplete evidence)
- strategy and prioritization (counterfactual, value-laden)
- analytical writing (argument quality, framing)
- design tradeoffs (multi-objective and preference-dependent)

In verifier-free domains, systems face a selection problem: they can generate many plausible stories, but they cannot cleanly *verify* which story is best. Without regulation, optimization pressure often produces:

- **thrash**: repeated recomputation and reversals
 - **cycling**: exploitation of selection quirks; oscillating preferences
 - **premature compression**: committing to a slogan-like container too early
-

3. Critics as sensors (not authorities)

A learned critic/judge can rank alternatives, but it is itself a fallible model that can drift, be gamed, or collapse into degenerate behaviors (always win / always tie). Therefore:

Use critics as telemetry sources, not as authorities.

In practice, this means critic outputs are treated like any other sensor stream:

- they are **versioned** and monitored for drift
 - they contribute to posture control (tighten or loosen budgets)
 - they do not directly authorize irreversible commitments without gating
-

4. Abstention-gated selection

We define a selection primitive suitable for verifier-free domains.

4.1 Bounded candidate generation

Generate a small set of candidates (K) where each candidate is an **analysis artifact**: it must include levers, predictions, falsifiers, and uncertainty boundaries. The goal is not “more text,” but a small set of competing, testable causal stories.

4.2 Bounded pairwise selection

Select among candidates with bounded pairwise comparisons (tournament-style or capped round-robin). Each comparison returns:

- $\text{win}(A)$
- $\text{win}(B)$
- tie/abstain

Pairwise selection is favored over isolated scoring because it reduces drift in absolute reward scales and keeps selection anchored to relative preference.

4.3 Abstention mass as an uncertainty signal

Define **abstention mass** as the fraction of comparisons returning tie/abstain (or equivalently, the entropy/margin of the pairwise preference distribution).

High abstention mass should not trigger more synthesis. It should trigger evidence.

If abstention mass exceeds a threshold, the system must route control flow into an evidence-gathering step (stronger sources, discriminating falsifiers, or minimal interventions) before re-running selection.

5. Commit semantics across the regulation series

The selection primitive becomes the “missing joint” across the trilogy:

- **Time-to-Analysis:** analysis artifacts are the unit of action readiness, but verifier-free selection requires a bounded protocol.
 - **Concept Containers:** container writes are commits. High abstention mass gates writes to prevent over-compression.
 - **Regulatory Ground:** abstention mass feeds posture control: tighten budgets, restrict actions, require stronger provenance, and enable rollback.
-

6. Critic governance: versioning, drift checks, replay suite

Verifier-free systems become brittle when critic behavior changes invisibly. Minimal governance:

- 1) **Version pinning:** record critic version, prompt template, and calibration settings for every selection run.
- 2) **Drift checks:** maintain a small fixed set of canonical comparisons (a replay suite). Track preference flip rate, abstention drift, and paraphrase sensitivity across versions.
- 3) **Rollback semantics:** if drift spikes or the critic collapses (always-win or always-tie), downgrade posture and revert critic configuration.

This is the same spirit as regulated tool use: selection itself is a high-leverage control surface and must be governed.

7. Metrics

Add verifier-free selection metrics to the series’ evaluation family:

- **Selection budget:** candidates K , number of comparisons, number of cycles.
 - **Abstention mass:** tie/abstain rate; preference entropy; win margin.
 - **Synthesis gate rate:** frequency of “refuse to synthesize; request evidence” actions.
 - **Commit stability:** reversal rate for committed artifacts (containers, decisions, plans).
-

8. Falsifiable experiments

8.1 Verifier-free decision tasks

Compare:

- baseline: retrieve + summarize + single synthesis
- regulated: K candidates + tournament selection + abstention gating + evidence escalation

Measure time-to-analysis, compute-to-analysis, reversals, and post-hoc expert grading.

8.2 Container write discipline

Test whether abstention-gated container writes reduce over-compression:

- lower reversal rate
- higher falsifier quality
- better transfer under surface-form shift

8.3 Critic drift and gaming

Perturb critic prompts or swap judge models; measure replay-suite flip rates and whether governance triggers rollback and posture tightening.

9. Limitations

- **Critic bias:** critics can be systematically biased or preference-misaligned.
 - **Degenerate abstention:** always-tie behavior can block progress; governance must detect collapse.
 - **Judge quirks:** tournament selection can amplify quirks unless calibration and drift checks exist.
 - **Threshold tuning:** abstention thresholds and escalation ladders are domain dependent.
-

10. Conclusion

Verifier-free discovery is not solved by “more reasoning.” It requires a regulated selection-and-commit discipline. Treat critics as sensors, keep selection bounded, use abstention mass as an uncertainty control signal, and govern critic drift with versioning, replay suites, and rollback semantics.

This note proposes a concrete missing layer that completes the trilogy’s arc: baseline posture control, reusable structure, and regulated synthesis are not sufficient unless verifier-free selection itself is operationalized and governed. **Implementation pointer.** A small replayable CI gate accompanies this note to operationalize tournament selection + abstention gating as a concrete harness. The goal is not to “prove correctness,” but to enforce commit discipline, injection resistance, and stability under perturbation as minimal requirements for verifier-free discovery systems. **Retrieval-gate analogue.** The same discipline applies to retrieval systems: deterministic scoring where needed, tie-mass telemetry as uncertainty, and abstention-gated commit policy for

downstream writes. The target is not identical prose on every run, but stable commit/withhold decisions, explicit rationale, and bounded variation under replay.

Appendix A: Minimal selection protocol (pseudocode)

Input: query/task Q, candidate budget K, abstention threshold tau

- 1) Generate K analysis artifacts A_i (each must include levers + falsifiers)
- 2) Run bounded pairwise comparisons between candidates using critic C
- 3) Compute $\text{abstention_mass} = (\# \text{ tie/abstain comparisons}) / (\text{total comparisons})$
- 4) If $\text{abstention_mass} > \tau$:
 gather evidence (retrieve stronger sources, produce discriminating falsifiers,
 or run a minimal intervention) and return to step (1)
- Else:
 select winner via tournament/majority vote and synthesize a final artifact
- 5) Record critic version, prompts, and replay-suite results for drift monitoring

Appendix B: Replay suite checklist

Maintain a small fixed set of canonical comparisons and track across versions:

- preference flip rate
- abstention rate drift
- sensitivity to paraphrase
- sensitivity to added irrelevant detail

Note on authorship and tools:

This work was developed through iterative reasoning, modeling, and synthesis. Large language models were used as a collaborative tool to assist with drafting, clarification, and cross-domain translation. All conceptual framing, structure, and final judgments remain the responsibility of the author.