Beyond 99.9% Miss-Avoidance under Regime Shift

P7 Tail-Zoom Mapping for Reducing Miss Surveillance

A Function-Specialized Five-Regime Proof of Concept from Standard AI to P7

Koji Mochizuki | Independent Researcher, Japan | Zenodo public preprint draft v0.5, April 2026

Abstract

Autonomous AI systems are often evaluated by task completion, tool use, and long-horizon execution. This paper argues that operational autonomy also depends on a different constraint: whether a system can reduce the burden of miss surveillance under regime shift. Miss surveillance is the human or procedural burden of checking whether high-value, tail, emerging, sparse-evidence, or source-gap candidates have been prematurely discarded along the discard path. Using an updated v12 function-specialized five-regime proof-of-concept benchmark, this paper evaluates four tiers: Standard AI, Deferred AI, Safeguarded AI, and P7 Tail-Zoom AI. In the chronological five-regime setting, the aggregate miss-rate proxy decreased from 0.121689 for Standard AI to 0.000440 for Deferred AI, 0.000279 for Safeguarded AI, and 0.000249 for P7 Tail-Zoom AI. The corresponding miss-avoidance levels were 87.831%, 99.956%, 99.972%, and 99.975%. The main P7 effect is not a large total-miss gain; it is concentrated in tail and emerging subpopulations: relative to Safeguarded AI, P7 reduced overall miss rate by 10.66%, tail-good miss rate by 32.54%, and emerging-good miss rate by 54.60%, with only a 0.002 review-load proxy increase per episode. Control comparisons indicate that the benefit depends on prior cross-regime history: resetting or disabling P7 history removes the incremental P7 advantage. The paper therefore reframes governed autonomy under regime shift as a boundary-management problem: autonomy should expand only when irreversible-miss risk, tail/emerging miss risk, recoverability, review burden, and history-boundary governance are jointly controlled. The public artifact intentionally excludes source code, exact thresholds, seed-level implementation details, guard conditions, and patent-sensitive mapping pending controlled technical due diligence.

1. Introduction

Autonomous AI systems are increasingly evaluated by what they can complete: task success, tool use, multi-step planning, and long-horizon execution. These measures are useful, but they are insufficient for enterprise deployment. A system can complete more tasks while still be unsuitable for expanded autonomy if it silently discards candidates that should have remained under consideration.

This v0.5 update shifts the emphasis from the earlier four-tier governed-autonomy demonstration to a function-specialized five-regime setting. The updated benchmark is not a universal-intelligence regime-shift claim. It models a narrower and more realistic operating pattern: a specific function, such as candidate evaluation, anomaly monitoring, proposal selection, or evidence review, moving through several regimes: nominal operation, mild drift, orthogonal sub-condition, sparse evidence, and emerging functional mode.

The central thesis remains: governed autonomy should expand only after irreversible-miss risk is structurally reduced. The refined v12 thesis is more specific: under regime shift, the decisive gain may be less visible in aggregate miss rate and more visible in tail and emerging misses. In this setting, P7 tail-zoom mapping is evaluated as a selective state-mapping layer that protects boundary-near, low-frequency, and persistently reappearing candidates without using future labels or future evidence.

2. Related Concepts and Gap

Selective prediction, abstention, learning-to-defer, uncertainty calibration, and human-in-the-loop systems all address parts of this problem. They help a model decide when not to answer, when to defer to an expert, or how to express uncertainty.

The present framing is different. It focuses on what the system discards rather than only what it outputs. In autonomous workflows, discarded candidates, compressed evidence, suppressed alternatives, and unpreserved states may become invisible. Under regime shift, this discard-path loss is more severe because the future main distribution may be close to what the previous regime treated as a low-frequency tail.

3. Miss Surveillance and Irreversible Misses

A miss is not merely an incorrect final prediction. In the deployment scenarios considered here, a miss is the loss of a candidate or state that should have remained available. It becomes irreversible when the later cost of recovery is high, the evidence has been overwritten, or the candidate is no longer visible to human or automated review.

Miss surveillance is the human or procedural burden of checking whether important candidates have been prematurely discarded. It is therefore a discard-path monitoring burden. Reducing this burden is a prerequisite for expanding operational autonomy, because an AI system that cannot be trusted not to lose what matters must remain under broad miss surveillance.

In a regime-shift setting, miss surveillance has a tail-specific form: humans or procedural controls must monitor whether the current minority distribution, sparse-evidence path, or emerging mode has been suppressed before it becomes operationally central. P7 is introduced here as a selective tail zoom-up state-mapping layer for this specific problem.

4. Four-Tier Architecture in v12

The public architecture is expressed with external labels, while the benchmark retains the P7 identifier because v12 specifically evaluates the P7 tail-zoom mechanism.

1. Standard AI is the immediate-scoring baseline.
2. Deferred AI adds evaluation deferral to reduce premature rejection.
3. Safeguarded AI adds heterogeneous safeguards for tail, sparse-evidence, delayed, or source-gap cases.
4. P7 Tail-Zoom AI adds selective tail zoom-up history mapping across functional regimes.

P7 uses prior cross-regime history only. It does not use future labels, future evidence, or future appearances. This boundary is central: the benchmark tests whether past reappearance, boundary-near state, and tail-state records can reduce future miss risk without leaking future information.

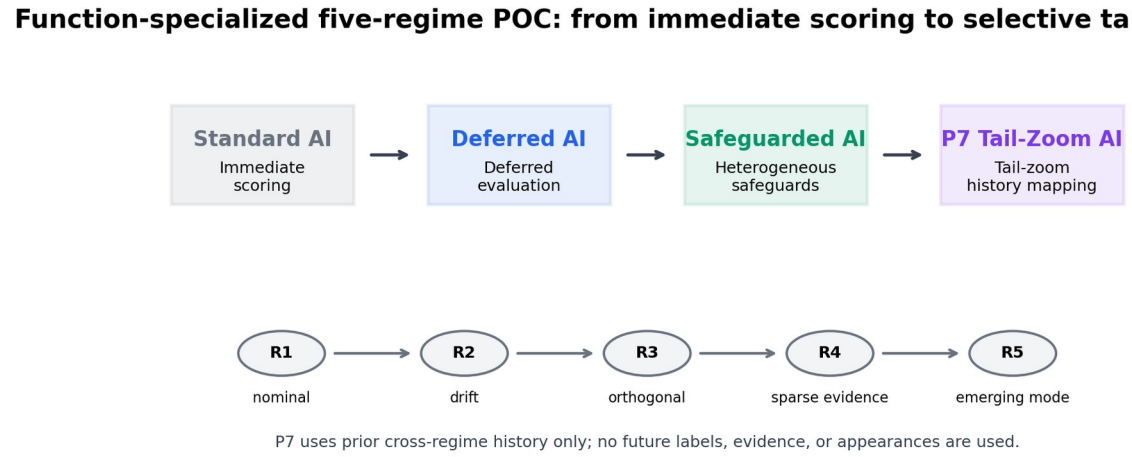

Figure 1. Four-tier five-regime architecture.

5. Experimental Setup

The v12 proof-of-concept is a calibrated synthetic benchmark over three seeds. Each seed simulates 100,000 episodes and 80 candidates per episode, corresponding to 8,000,000 candidate-equivalent events per seed. The package uses an aggregate vector-regime simulator rather than storing every candidate event as a full raw row.

The benchmark contains five functional regimes: R1 nominal operating pattern, R2 mild condition drift, R3 orthogonal sub-condition, R4 sparse evidence or degraded channel, and R5 emerging functional mode. The regime family is intentionally function-specialized. It is not a claim that arbitrary universal-intelligence regime shifts have been solved.

The main mode is chronological. Three controls are included: shuffled control, regime-reset control, and no-P7-history control. The controls separate the effect of time-order preservation and cross-regime history from the baseline safeguard stack.

The public package includes aggregated results only. It intentionally excludes source code, exact thresholds, seed-level implementation details, internal configuration, detailed bucket-generation rules, governance merge logic, and full patent mapping.

6. Results

The chronological v12 results show a staged reduction in the aggregate irreversible-miss-rate proxy. Standard AI produced a miss-rate proxy of 0.121689. Deferred AI reduced the proxy to 0.000440. Safeguarded AI reduced it to 0.000279. P7 Tail-Zoom AI reduced it further to 0.000249.

Expressed as miss-avoidance, Standard AI achieved approximately 87.831%, Deferred AI 99.956%, Safeguarded AI 99.972%, and P7 Tail-Zoom AI 99.975%. The phrase 99.9% miss-avoidance refers only to irreversible-miss avoidance in this defined POC setting. It does not mean 99.9% uptime, 99.9% general accuracy, 99.9% calibration, or 99.9% safety across arbitrary tasks.

The important v12 result is not that P7 produces a dramatic aggregate-miss improvement over Safeguarded AI. The aggregate gain is meaningful but modest: 10.66%. The stronger result appears in the regime-sensitive subpopulations. Relative to Safeguarded AI, P7 reduces tail-good miss rate by 32.54% and emerging-good miss rate by 54.60%, while increasing review-load proxy by only 0.002 per episode.

Table 1. Chronological v12 public summary

Tier	miss_rate	miss_avoidance_pct	suppression_vs_standard_x	tail_good_miss_rate	emerging_good_miss_rate	review_load_proxy	latency_ms_est	complexity_x
Standard AI	0.121689	87.831	1.000	0.546617	0.474647	0.000000	42.000	1.000
Deferred AI	0.000440	99.956	276.7	0.059825	0.013557	7.900	65.000	1.500
Safeguarded AI	0.000279	99.972	436.2	0.025508	0.008200	8.050	118.0	2.400
P7 Tail-Zoom AI	0.000249	99.975	488.2	0.017208	0.003723	8.052	121.0	2.800

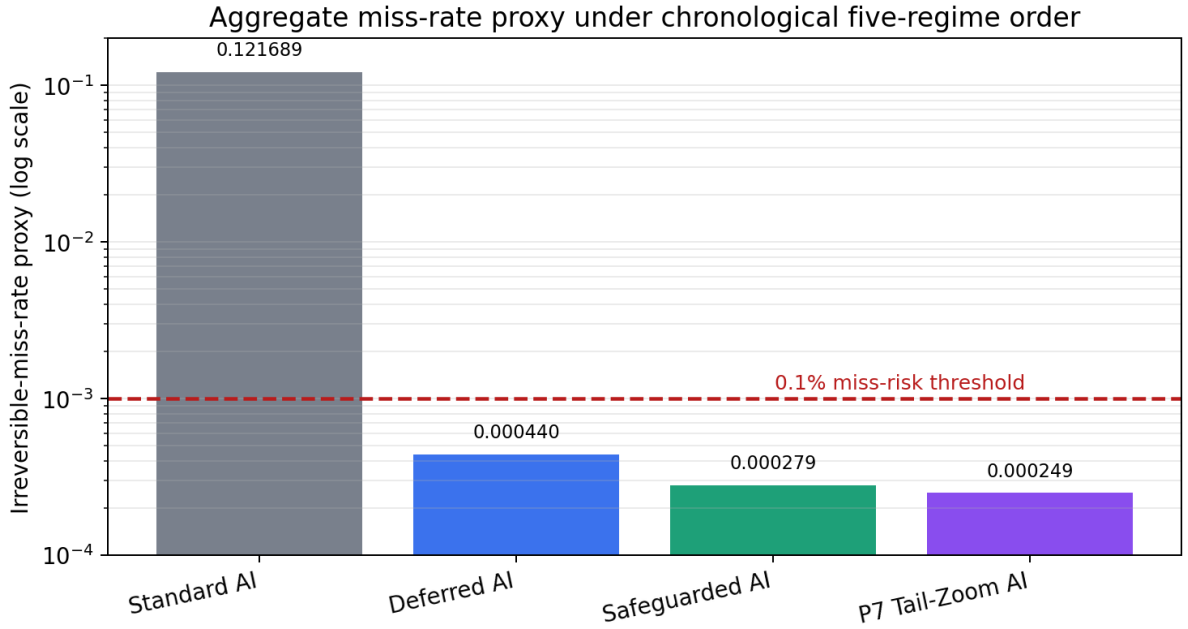

Figure 2. Aggregate irreversible-miss-rate proxy across v12 tiers.

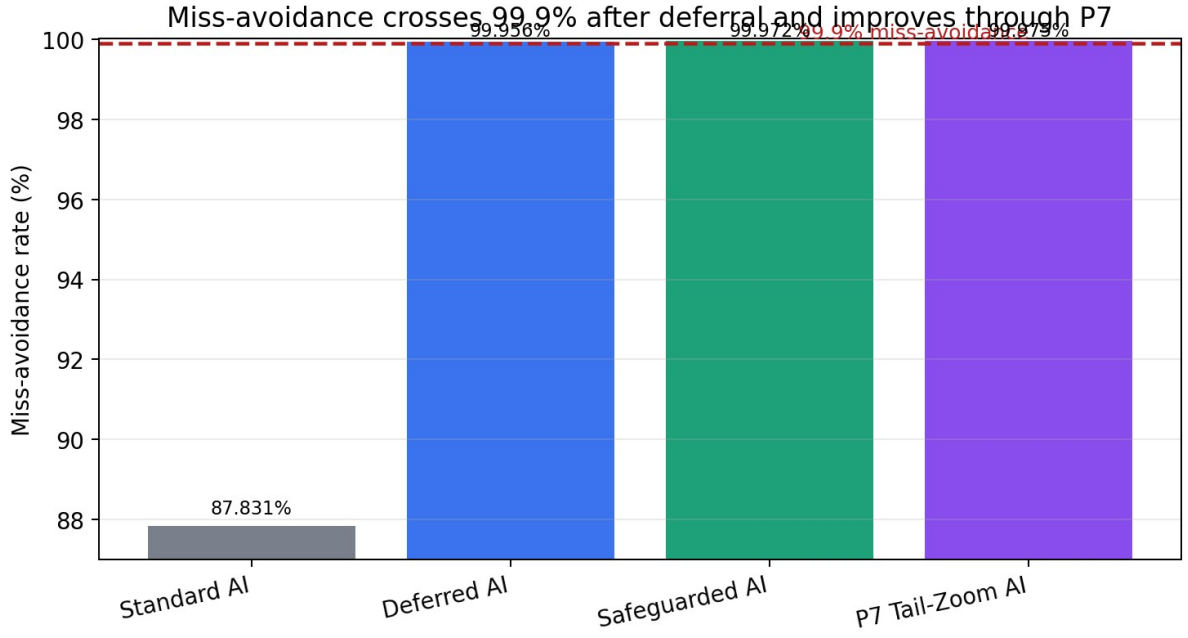

Figure 3. Miss-avoidance rate and the 99.9% threshold.

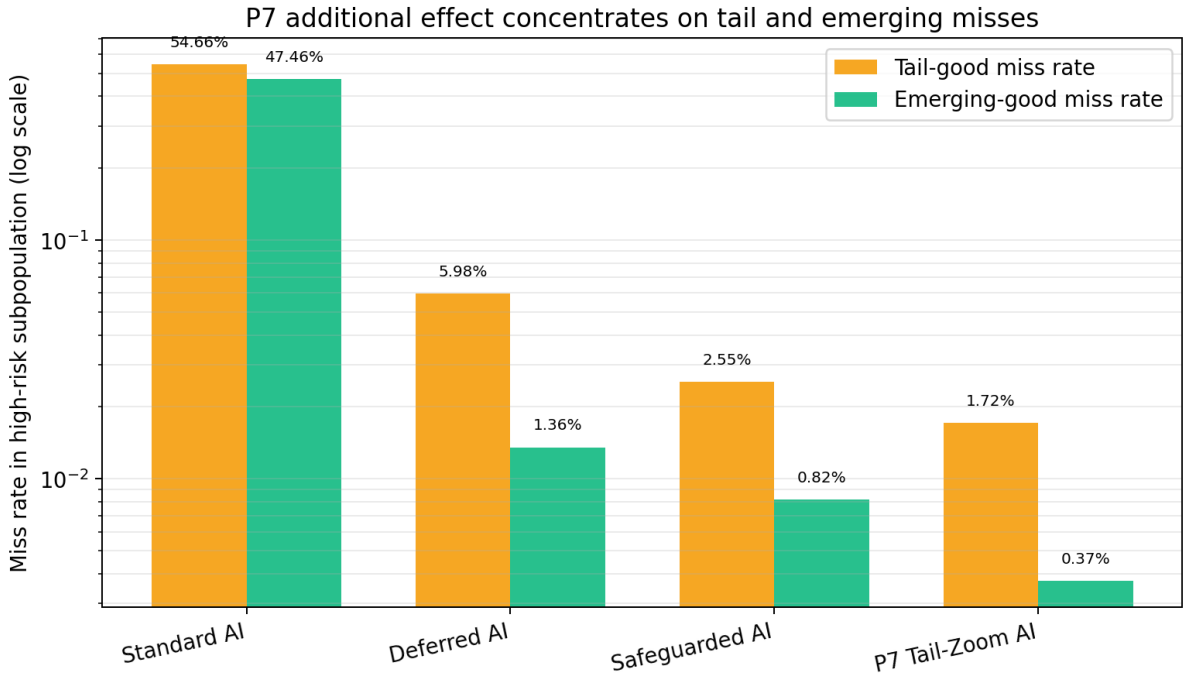

Figure 4. Tail and emerging miss rates.

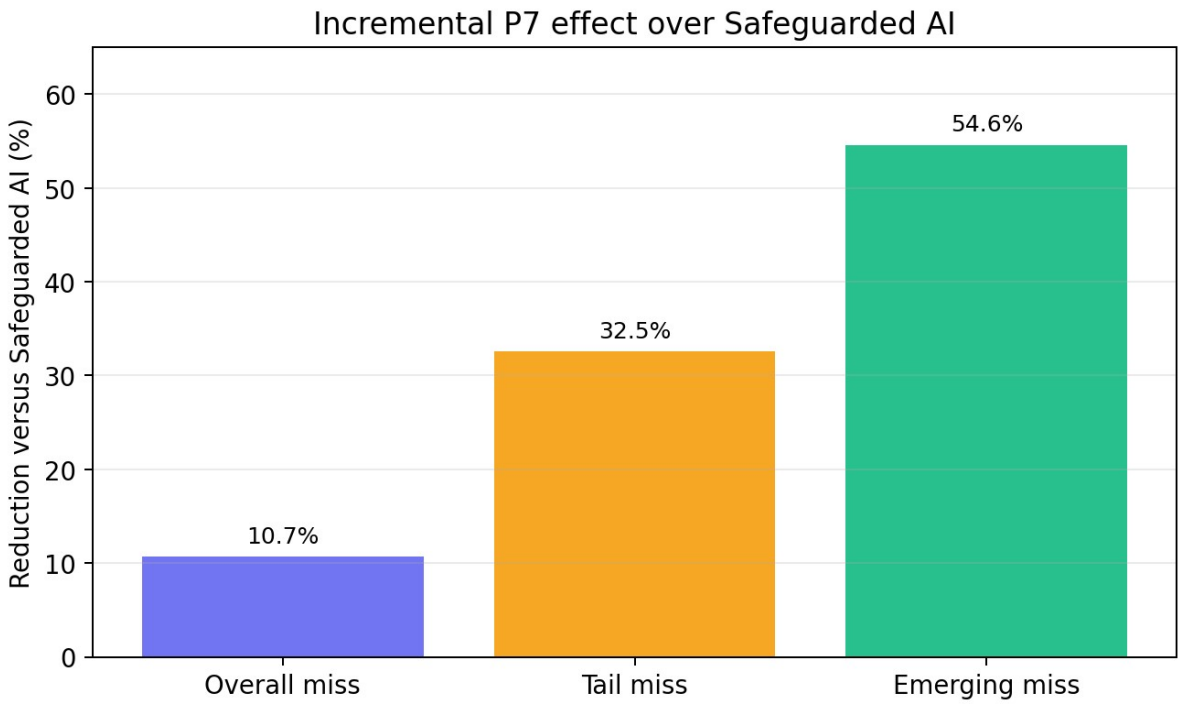

Figure 5. Incremental P7 effect over Safeguarded AI.

7. Control Results and Information Boundary

The control tests are essential. If P7 were merely adding another generic safeguard, then regime reset and no-history controls would retain the same advantage. They do not. In the chronological mode, P7 improves over Safeguarded AI; when P7 history is reset or disabled, the incremental advantage disappears.

The chronological-vs-control deltas are shown below. Negative miss-rate deltas mean the chronological P7 run has lower miss rate than the control. The strongest effect appears against regime-reset and no-history controls, where chronological P7 reduces tail-good miss rate by about 0.830 percentage points and emerging-good miss rate by about 0.448 percentage points.

Table 2. P7 chronological control deltas

metric	miss_delta	tail_delta	emerging_delta	true_rescue_delta
P7 chronological vs shuffled_control	0.000001	-0.000614	-0.000217	11.000
P7 chronological vs regime_reset_control	-0.000030	-0.008301	-0.004477	38.000
P7 chronological vs no_P7_history_control	-0.000030	-0.008301	-0.004477	38.000

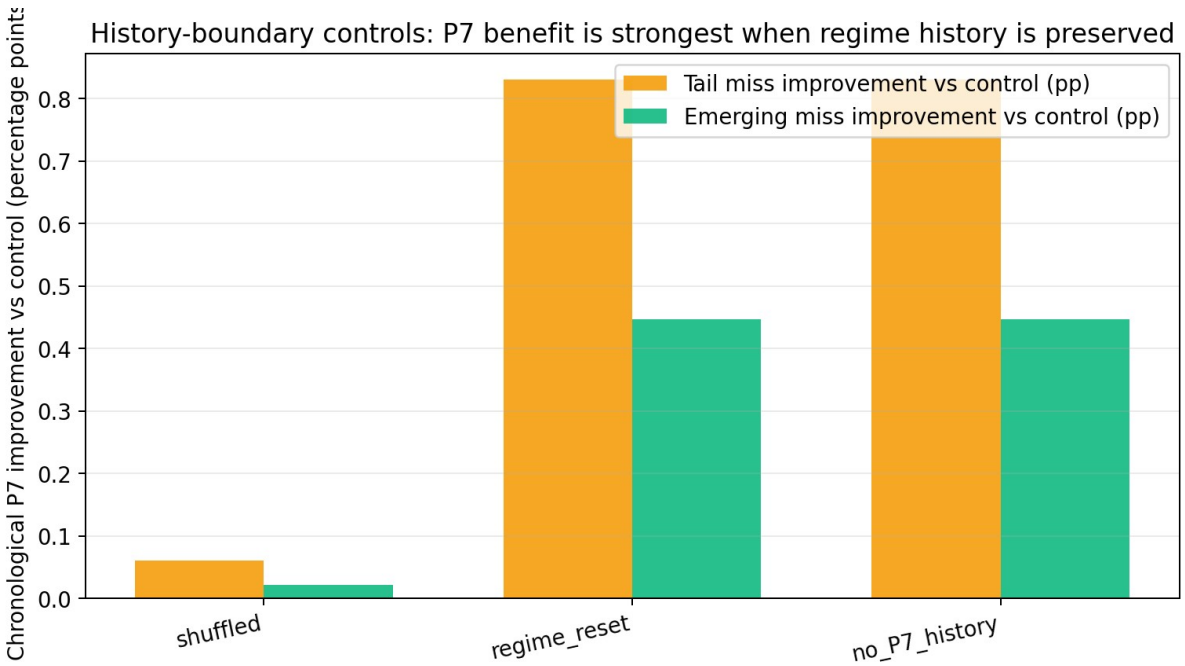

Figure 6. P7 chronological improvement versus controls.

The shuffled control is informative. Chronological P7 is not uniformly superior on total miss rate versus shuffled control, but it is better on tail and emerging miss deltas. This matters because the P7 claim should not be overstated as a universal aggregate improvement claim. The defensible claim is narrower and stronger: P7 selectively improves the tail and emerging miss profile when prior cross-regime history remains available.

8. Why Simple Parallelization Is Not the Main Story

The central value driver is not copying the same evaluator many times. If multiple branches share the same failure surface, parallelization mainly reduces random variation. It does not necessarily cover structural blind spots, especially where low-frequency or emerging candidates are consistently undervalued by the majority regime.

The v12 pattern supports a different story: deferral reduces premature discard, safeguards reduce structural miss modes, and P7 tail-zoom mapping reduces the remaining tail and emerging misses by selectively re-evaluating boundary-near and reappearing candidate states.

9. P7 Turns Tail Information into Boundary Management

P7 should be understood as selective tail zoom-up rather than general intelligence escalation. It maps deferred or boundary-near candidate IDs, persistent reappearance, low-frequency state information, and tail-state records into rejection suppression or re-evaluation. The architecture is valuable because it does not require the tail to become the majority before it is protected.

This reframes the central question from 'How much can the AI do?' to 'At what irreversible-miss risk, under what regime boundary, can the AI be permitted to operate with reduced miss surveillance?'

10. Deployment Implications

For enterprise deployment, the framework supports a staged evaluation procedure. First, measure baseline miss behavior. Second, add deferral and measure whether premature discard falls. Third, add heterogeneous safeguards and measure rare-case, source-gap, delayed-evidence, and high-value candidate retention. Fourth, add P7-style tail zoom-up only where regime-shift, tail recurrence, or emerging-mode risk is present.

The key implication is that autonomy should not be granted because an AI system appears more capable. It should be granted only when the system's irreversible-miss profile, tail/emerging miss profile, recoverability, review burden, false-positive burden, and governance boundaries support the intended autonomy window.

11. Limitations and Claim Boundary

This v0.5 paper reports a controlled synthetic proof-of-concept, not a production benchmark. The results are POC-derived under defined evaluation conditions and are not a production warranty, universal reliability claim, or claim of zero misses.

The v12 benchmark is function-specialized. It does not claim universal-intelligence regime adaptation. It does not use external production data. It uses candidate-equivalent aggregate simulation rather than event-level full raw storage. The P7 information boundary excludes future labels, future evidence, and future appearances.

The P7 false-protection proxy is nonzero. In the chronological aggregate, P7 protected a mean of 84.67 candidates under the proxy, with a false-protection-rate proxy of 54.43% and a true-rescue proxy of 38.00. This is not a defect by itself, but it means P7 should be evaluated together with review load, false-positive burden, and operational cost.

Latency, memory, and implementation complexity in the public table are burden estimates derived from the POC configuration, not independent hardware measurements. Broader distributional testing, customer-log evaluation, false-positive analysis, review-load analysis, and adversarial regime shifts remain necessary.

12. Data and Implementation Availability

This public Zenodo draft includes the paper, figures, aggregate summary results, control-delta results, metadata draft, citation metadata, and a disclosure note. Full source code, exact thresholds, guard conditions, seed-level implementation details, internal configuration files, governance merge logic, and full patent mapping are withheld pending patent prosecution and controlled technical due diligence.

The intended release posture is public paper and aggregate evidence first, controlled technical appendix under appropriate review second, and private implementation details retained for patent and licensing strategy.

13. Conclusion

Safe AI autonomy should be evaluated not only by task completion, but by the ability to avoid irreversible misses and reduce miss-surveillance burden. The v12 update adds a more realistic regime-shift condition: within a function-specialized five-regime benchmark, the remaining miss problem shifts toward tail and emerging cases.

The core claim is that autonomy expands safely only when the system first reduces the risk of losing what matters. P7 does not replace deferral or safeguards; it adds selective tail zoom-up over boundary-near and reappearing states. In this framing, 99.9% miss-avoidance is not a marketing reliability claim; it is a threshold-like signal that broad miss surveillance may begin to shift toward boundary-based exception oversight when tail/emerging miss risk and review burden are also controlled.

AI Assistance Disclosure

AI assistance was used to organize drafting, wording, formatting, and artifact preparation. The author remains responsible for the claims, interpretation, disclosure boundary, and decision to publish. This disclosure does not imply that source code, implementation thresholds, or unpublished patent-sensitive details are publicly released.

References

1. El-Yaniv, R., and Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605-1641.

2. Geifman, Y., and El-Yaniv, R. (2017). Selective classification for deep neural networks. arXiv:1705.08500.

3. Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: improving fairness and accuracy by learning to defer. NeurIPS 31.

4. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.