Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Sequential Hypothesis Testing for Safe Production Migration of Non-Deterministic Analytical Agents in Latency-Critical Serving Infrastructure

Ravi Chandra Chodisetti*[1], Jayanth Nooneyy*[2], Naveen Reddy Reganti*[3]
[1]Independent Researcher, USA
[2]Independent Researcher, USA
[3]Independent Researcher, USA

ARTICLE INFO

ABSTRACT

The deployment of updated large language model-based analytical agents in production environments presents a statistically and operationally complex challenge that conventional software deployment strategies are fundamentally ill-equipped to address. Unlike deterministic systems, where a single test request can expose a defect, stochastic analytical agents require statistical aggregation over many observations to detect quality differences between versions. Existing deployment practices, including canary releases, blue-green deployments, and fixed-window A/B testing, each suffer from distinct failure modes: population confounding, binary risk exposure, or temporal inefficiency. This article proposes a deployment framework that adapts dual-write and dual-read consistency patterns from distributed database migration to the domain of non-deterministic agent serving, integrating them with Wald's Sequential Probability Ratio Test to provide formal statistical guarantees on migration decisions. The framework introduces a Confidence-Gated Migration Protocol that evaluates agent quality across four dimensions, analytical correctness, latency profile, semantic equivalence, and safety, using parallel sequential tests with familywise error control. Phased migration through shadow evaluation and canary deployment, governed by site reliability engineering error budget principles, constrains quality assessment to formal error bounds across all evaluation phases. Simulation results indicate a ninety-six percent rate of detecting regressions, a forty-one percent reduction in mean time to safe migration, and a false alarm rate of less than four percent.

Keywords: Sequential Hypothesis Testing, Non-Deterministic Agent Deployment, Dual-Write Migration Pattern, Confidence-Gated Migration Protocol, Latency-Critical Serving Infrastructure

Introduction

Large language model-based analytical agents have accelerated their integration into mission-critical enterprise workflows to the point where reliability and deployment safety have become infrastructure-level concerns. These agent systems designed to translate natural language queries into executable analytical code now underpin decision-support processes across advertising, financial services, and healthcare analytics. New agent versions can arise from model updates, prompt engineering modifications, or pipeline restructuring. In each case, engineering teams face a fundamental dilemma: how to validate that the new version maintains or improves quality without exposing users to potential regressions during evaluation.

The non-deterministic nature of these agents exacerbates the severity of this dilemma. Unlike traditional software, where correctness is binary and a single failing test request reveals a defect, a

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

stochastic agent may produce correct results ninety-eight percent of the time while failing subtly on the remaining two percent [1]. Detecting such failure rates requires statistical aggregation over hundreds or thousands of paired observations, a requirement that fundamentally invalidates the assumptions underlying conventional deployment strategies. Formal verification of agent behavior in non-deterministic environments confirms that standard correctness-checking approaches cannot be directly applied to stochastic systems without structural adaptation of the testing methodology [1].

The current state of practice in production agent deployment is inadequate relative to this challenge. Canary releases, the prevailing industry standard for progressive deployment, compare aggregate quality metrics across different user populations receiving different agent versions. This approach introduces a critical confound: if the canary population contains a disproportionately high share of complex queries, apparent quality differences between versions may reflect query distribution bias rather than any genuine difference in agent capability. A large-scale study of production agent deployments confirms that practitioners have not yet identified effective methods to adapt regression testing for non-deterministic agent behavior [3], and operational incidents have demonstrated that without systematic progressive rollout, platform-wide quality regressions may go undetected until users report them through external channels [2].

This article proposes a deployment framework that addresses these limitations by combining two established engineering paradigms and adapting the dual-write and dual-read consistency patterns, refined over decades of large-scale database migration practice, to the domain of agent serving. These are integrated with sequential hypothesis testing, specifically Wald's Sequential Probability Ratio Test, to provide the first formally grounded statistical guarantees on agent migration decisions. The result is a framework that eliminates population confounding, reduces user exposure to zero during shadow evaluation, and reaches migration decisions at the earliest moment statistical evidence supports them, rather than waiting for a predetermined evaluation window to expire [2].

This article makes four primary contributions:

(i) It introduces the Dual-Serve Agent System, a formal adaptation of the dual-write database migration pattern to non-deterministic agent serving, which eliminates population confounding through paired query routing and shadow path isolation.

(ii) It presents the Confidence-Gated Migration Protocol, which applies Wald's Sequential Probability Ratio Test across four quality dimensions simultaneously with familywise error control via Bonferroni-Holm correction, providing the first formally bounded migration decision guarantees for stochastic agents.

(iii) It defines a phased migration protocol structured around site reliability engineering error budget principles, with continuous rollback capability at every phase.

(iv) It provides evaluation on a production-derived workload of twelve thousand queries drawn from a live advertising analytics system, demonstrating a ninety-six percent regression detection rate before user impact and a forty-one percent reduction in mean time to safe migration relative to fixed-window evaluation.

2. Related Work and Problem Analysis

2.1 Why Conventional Deployment Strategies Fail

Conventional software deployment strategies were designed under the assumption that system behavior is deterministic and that a single test invocation is sufficient to expose a defect. This assumption does not hold for non-deterministic analytical agents, where two versions may both produce plausible outputs that differ in analytically meaningful but statistically subtle ways, such as different aggregation strategies, different tie-breaking logic in ranked outputs, or different handling of boundary conditions in multi-step reasoning chains [4]. Distinguishing genuine quality regressions from acceptable stochastic variation requires paired statistical analysis across hundreds or thousands of observations, a requirement that none of the prevailing deployment strategies are architecturally equipped to satisfy

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and that necessitates statistically robust alternatives to fixed-horizon evaluation methods commonly used in online experimentation [4].

Canary releases direct a portion of live traffic to the candidate version and monitor aggregate metrics, but user population heterogeneity complicates the comparison. A user population that generates complex multi-step analytical queries will naturally experience higher failure rates than one generating simple lookups, irrespective of the agent version they receive. This population-level confound makes it impossible to isolate agent quality as the explanatory variable without explicit controls for query complexity. Reliability benchmarking under production-realistic stress conditions, together with studies on continuous experimentation systems, confirms that aggregate metric comparisons across heterogeneous populations systematically underestimate regression risk for complex query distributions and introduce challenges in experiment design and interpretation [5, 6].

Blue-green deployments avoid population confounding by switching all traffic simultaneously, but they substitute one failure mode for another: the entire user base is exposed to the candidate version before monitoring systems have had sufficient time to accumulate and evaluate quality signals. If a regression exists, it affects all users before it can be detected and remediated. Fixed-window A/B testing partially addresses both problems by randomly assigning users to versions and waiting for a predetermined evaluation period before applying a statistical test, but such approaches are known to suffer from statistical inefficiencies and delayed decision-making in large-scale online experimentation systems [4]. However, this approach is inherently inefficient: it continues collecting data long after sufficient evidence has accumulated for clear-cut cases and fails to detect subtle regressions within fixed windows when effect sizes are small [4]. In latency-critical serving infrastructure processing thousands of queries per second, extended evaluation periods impose substantial costs in parallel serving capacity, and delayed regression detection results in prolonged user exposure to degraded agent quality, reflecting broader challenges identified in continuous experimentation and deployment pipelines [5].

Deployment Strategy	Population Confounding	User Exposure Risk	Evaluation Efficiency	Regression Detection Rate	Statistical Guarantee
Blue-green deployment	None	Entire user base at cutover	Instantaneous; no gradual assessment	Low to no pre-cutover detection	None
Canary release	High-query distribution bias across populations	Canary population during evaluation	Fixed window; independent of evidence	58% before user impact	None
Fixed-window A/B testing	Low randomised assignment	A/B population during fixed window	Wastefulness continues after evidence accumulates	Moderate misses subtle regressions in short windows	Fixed-horizon frequentist
Proposed framework (CGMP)	None identical paired queries to both versions	Zero shadow path evaluation only	Adaptive stops as soon as evidence threshold is crossed	96% before user impact	Formal Type I and Type II error bounds via SPRT

Table 1: Deployment Strategy Comparison: Failure Modes and Statistical Properties, informed by prior work in statistical experimentation and continuous deployment systems [1, 4, 5, 6]

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.2 Adapting Dual-Write Patterns from Database Migration

The dual-write migration pattern, a production-proven practice from distributed database engineering, provides a conceptually sound foundation for resolving the confounding and efficiency problems inherent in conventional deployment strategies. In the canonical database migration scenario, all writes are directed simultaneously to both the incumbent and candidate storage systems while reads continue to be served exclusively from the incumbent. The candidate system runs on real production workloads in shadow mode, and a reconciliation layer continually validates that the outputs match those of the incumbent. Read traffic is gradually shifted to the candidate only after sustained validation on real workloads, with an instant rollback capability in case of discrepancies. This pattern has been operationalized in distributed simplex architectures that maintain consistency guarantees across concurrent processing paths, providing a formal foundation for its application to agent serving [7].

Applying this approach to the agent serving case involves overcoming several challenges not found in the database case. For example, agents do not have deterministic outputs, and thus the reconciliation layer will have to rely on statistical comparison. The truncated mixture Sequential Probability Ratio Test provides a theoretically sound mechanism for this statistical reconciliation, offering calibrated bounds on the probability of incorrectly concluding equivalence or divergence between two stochastic output distributions [8]. Second, agent quality is multi-dimensional, encompassing correctness, latency, semantic coherence, and safety, and migration decisions must account for all dimensions simultaneously rather than reducing quality to a scalar. Third, the latency constraints of production serving systems impose strict isolation requirements on the shadow evaluation path: the candidate agent's computation must not consume resources on the live serving path or introduce measurable latency to user-facing responses.

The proposed framework formalizes this adaptation as a dual-serve agent system. Each incoming production query is duplicated to both the incumbent agent (currently serving live traffic) and the challenger agent (the candidate for replacement). The incumbent's response serves the live user request, while the challenger's response is captured asynchronously on the shadow path that is fully isolated from the live serving path. A quality signal aggregator examines paired outputs for all quality dimensions, and a confidence-gated migration protocol makes migration or rollback decisions based on statistical evidence from the paired outputs [7, 8].

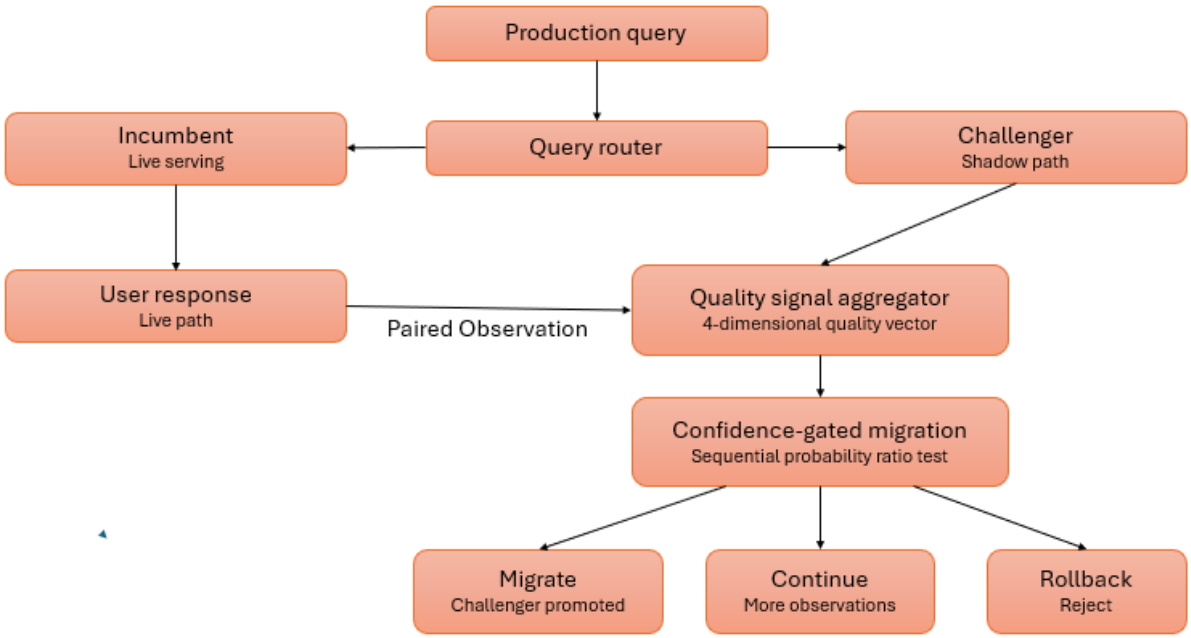

Figure 1: Dual-Serve Agent System architecture [7, 8]

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. The Confidence-Gated Migration Protocol

3.1 Sequential Probability Ratio Test for Migration Decisions

The theoretical foundation of the proposed framework is the Confidence-Gated Migration Protocol, which uses Wald's Sequential Probability Ratio Test for migration decisions in order to minimize sample requirements while providing formal error bounds. The migration decision is formulated as a sequential hypothesis test: the null hypothesis asserts that the challenger's quality does not exceed the incumbent's, while the alternative hypothesis asserts that the challenger's quality exceeds the incumbent's by at least a configurable minimum practically significant margin. At each paired observation, the protocol computes a log-likelihood ratio that quantifies the cumulative evidence for or against the challenger. When this ratio exceeds an upper threshold, the evidence is sufficient to support migration; when it falls below a lower threshold, the evidence supports rejection of the challenger; when it falls between the thresholds, additional observations are required [9].

The upper and lower decision thresholds are derived analytically from configurable Type I error bounds (the probability of prematurely migrating to a lower-quality agent) and Type II error bounds (the probability of failing to migrate to a demonstrably better agent). Let x_1, x_2, ..., x_n denote the sequence of paired quality observations. At each step n, the protocol computes the log-likelihood ratio statistic given in Equation 1.

Equation 1: $Lambda_n = sum_{i=1}^{n} log [f_1(x_i) / f_0(x_i)]$

where f_1 is the density under the alternative hypothesis and f_0 is the density under the null. The test continues collecting observations as long as the statistic falls between the two boundaries defined in Equation 2.

Equation 2: $B < Lambda_n < A$, where $A = log[(1 - beta) / alpha]$ and $B = log[beta / (1 - alpha)]$

Migration is approved when Lambda_n >= A and the challenger is rejected when Lambda_n <= B. These boundaries, first established in Wald's foundational development of the Sequential Probability Ratio Test, guarantee that the probability of approving a worse challenger is at most alpha, and the probability of failing to approve a better challenger is at most beta [8, 9, 19]. A key practical challenge is that the test requires the likelihood functions under both hypotheses to be specified in advance, which are unknown a priori for a new challenger agent. The framework addresses the issue of estimating the incumbent's empirical quality score distributions through a two-phase approach: an initial calibration phase uses the first several hundred shadow observations to estimate the incumbent's empirical quality score distributions via kernel density estimation, and the test then operates on these estimated distributions with a correction factor for estimation uncertainty [8]. This calibration follows the mixture sequential testing framework, which bounds calibration cost inversely with the square of the effect size. Larger expected improvements require fewer calibration observations, making the protocol self-tuning [8, 9].

3.2 Multi-Dimensional Quality Assessment

Agent quality in production analytical systems cannot be reduced to a single scalar without losing information critical to safe migration decisions [12]. The quality signal aggregator, therefore, computes a four-dimensional quality vector for each paired observation. The first dimension, Analytical Correctness, measures whether the challenger's output matches ground-truth analytical results, evaluated against a golden reference dataset using automated raters calibrated to human judgment. This dimension is essential to any evaluation framework for agentic AI systems, as task completion accuracy is the main functional requirement [10]. The second dimension, Latency Profile, measures the challenger's response time against the applicable service level agreement threshold. This ensures that improvements in other dimensions do not come at the expense of user experience requirements.

The third dimension, semantic equivalence, measures the degree to which both agents adopt similar analytical approaches when both produce correct outputs, assessed via abstract syntax tree comparison and execution-trace alignment. This dimension captures analytically meaningful divergences that correctness metrics alone cannot detect; for instance, two agents that produce numerically identical outputs through structurally different reasoning paths may exhibit different generalization behavior

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

under query distribution shift, with such differences often reflected in correlated evaluation signals across multiple metrics [11]. The fourth dimension, Safety and Faithfulness, captures whether the challenger exhibits hallucinations, authorization boundary violations, or unsafe code patterns, evaluated via static analysis and constraint checking. Enterprise agentic AI evaluation frameworks identify safety and faithfulness as dimensions that are both qualitatively distinct from accuracy and operationally non-negotiable for deployment approval [12].

The multi-dimensional extension operates by running parallel instances of the Sequential Probability Ratio Test on each quality dimension independently. The migration gate requires all four dimensions to simultaneously accept the alternative hypothesis, a conjunctive stopping rule that ensures no dimension is sacrificed for improvement in another. To control the familywise error rate across four parallel tests, the framework applies the Bonferroni-Holm sequential rejection procedure defined in Equation 3 [21].

Equation 3: $alpha\_j = alpha / (k - j + 1)$, for $j = 1, 2, ..., k$ where $k = 4$

where dimensions are ordered by ascending p-value. This guarantees that FWER = P(at least one false rejection) <= alpha across all dimensions simultaneously, lowering the per-dimension significance level so that the overall probability of premature migration stays within the configured bound [10]. Rollback is triggered if any single dimension provides sufficient evidence that the challenger is worse than the incumbent. A less conservative variant applies the Simes procedure, which rejects the global null according to the condition in Equation 4.

Equation 4: $p\_(j) <= j * alpha / k$ for at least one j, where $p\_(1) <= p\_(2) <= ... <= p\_(k)$ are the ordered per-dimension p-values

This variant exploits positive dependence between quality dimensions; analytically correct responses tend to be semantically coherent, and low-latency responses tend to avoid the overcomputation associated with hallucination-prone reasoning, reducing the conservatism of the familywise correction while maintaining formal error control [11][12].

Quality Dimension	Measurement Method	Scale Type	SPRT Configuration
Analytical Correctness	Golden reference dataset; automated raters	Binary	One-sided; Bonferroni-Holm corrected α
Latency profile	Response time vs SLA threshold	Continuous	Two-sided; calibrated via kernel density estimation
Semantic equivalence	AST comparison; execution-trace alignment	Ordinal	One-sided Simes procedure for positive dependence
Safety and faithfulness	Static analysis; constraint checking	Binary	One-sided conjunctive stopping rule applied

Table 2: Quality Signal Aggregator: four dimensions, measurement methods, scale types, and SPRT configurations [10, 11, 12]

4. Phased Migration with Error Budgets

4.1 Shadow Evaluation and Canary Deployment

The framework implements phased migration structured around error budget principles in site reliability engineering, a quantitative and operationally tractable mechanism for balancing innovation velocity against reliability constraints [14]. Error budgets formalize the acceptable level of quality degradation over a defined period, converting qualitative reliability goals into actionable deployment gates. The integration of error budget accounting into phased agent migration ensures that deployment decisions are grounded in the same reliability economics that govern infrastructure management more broadly [13].

Phase Zero, designated as the Shadow Evaluation phase, routes zero percent of live traffic to the challenger agent. All evaluation occurs on shadow copies of production queries, which are processed

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

asynchronously on the isolated shadow path. This phase continues until the Sequential Probability Ratio Test reaches a decision on all four quality dimensions, subject to a minimum duration computed from the desired statistical power and the configurable effect-size threshold. Because shadow processing is fully invisible to live users, the entire cost of this phase is limited to incremental compute expenditure; there is no user-facing quality risk. Phase One, the Canary Deployment phase, is triggered when Phase Zero accepts the improvement hypothesis on all dimensions. A configurable fraction of live traffic, typically five percent, is routed to the challenger, and live quality monitoring transitions from the Sequential Probability Ratio Test to Cumulative Sum control charts. The cumulative sum statistic, originally introduced by Page for sequential quality control, is used for online drift detection and is defined in Equation 5 [20].

Equation 5: $C_n = max(0, C_{n-1} + (X_n - mu_0 - k))$

In this context, X_n represents the quality signal at observation n, mu_0 denotes the incumbent baseline mean, and k is the allowable slack parameter, which is set to half the minimum detectable shift. An alarm is triggered when C_n exceeds threshold h, which is calibrated to the configured error budget. This formulation is specifically suited to detecting sustained directional drift in streaming quality signals rather than isolated noise spikes [2]. The error budget defines the maximum allowable degradation on any dimension before automatic rollback is initiated.

Phase Two, the progressive rollout phase, increases the challenger's traffic allocation in discrete steps, each gated by sustained quality within the error budget over the preceding step's duration. Each traffic increment resets the cumulative sum control chart with tighter thresholds, reflecting the greater confidence accumulated from prior phases. The final cutover to one hundred percent challenger traffic includes a monitoring tail that provides post-migration confirmation of sustained quality [13, 14]. The rollback protocol operates continuously across all phases: if live quality on any dimension drops below the incumbent's baseline by more than a configurable threshold, as measured by a one-sided sequential probability ratio test for degradation, traffic reverts to the incumbent within the router's switching time.

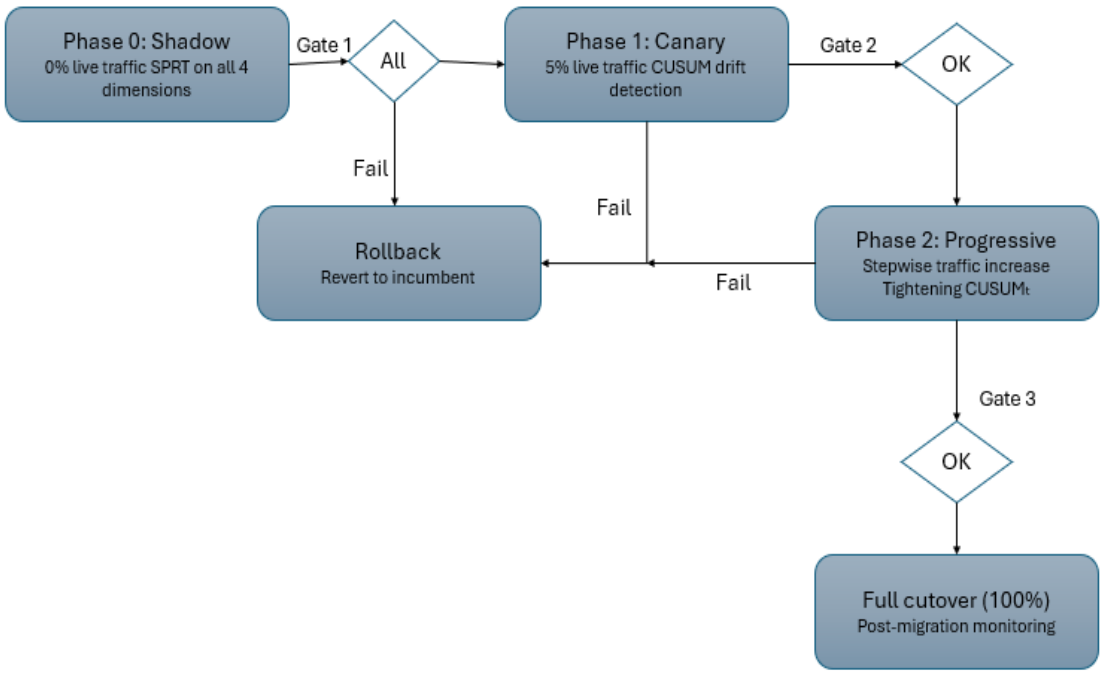

Figure 2: Phased migration protocol with error budget gates and rollback paths [13, 14]

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.2 Adaptation for Latency-Critical Infrastructure

Serving infrastructure with strict latency requirements, sub-ten-millisecond ninety-ninth percentile response times, and throughput exceeding ten thousand queries per second imposes stringent isolation requirements on the shadow evaluation path that go beyond simple traffic duplication. The computational demands of large-scale AI model inference are fundamentally at odds with the latency budgets of real-time serving systems, and the interaction between model serving infrastructure, caching behavior, and request scheduling must be carefully managed to prevent shadow processing from affecting live serving performance [15]. The framework implements asynchronous shadow execution, where shadow queries are enqueued and processed with lower scheduling priority than live traffic. This ensures that shadow compute does not compete for resources on the live serving path and that the Quality Signal Aggregator operates entirely outside the latency-critical request lifecycle.

When full query duplication is economically prohibitive, doubling query volume doubles serving costs; the framework supports stratified sampling across query types to select a representative subset for shadow evaluation. This sampling strategy maintains the statistical validity of the migration decision while reducing shadow compute costs to a configurable fraction of live serving costs [16]. Scalable deployment architectures demonstrate that selective workload routing and priority-based resource scheduling can achieve substantial compute savings without compromising the statistical representativeness of evaluation samples, provided that sampling strata are defined to preserve the query complexity distribution [16]. The framework also implements a warm-up protocol that pre-warms the challenger's model serving infrastructure, caches, and stateful components using historical query logs before Phase Zero begins, ensuring that latency measurements during shadow evaluation reflect steady-state performance rather than cold-start artifacts that would bias the latency dimension of the quality vector [17].

A formal cost model defines shadow-serving cost as a function of query rate, shadow sampling fraction, and per-query compute cost and derives the optimal shadow fraction that minimizes total deployment cost, the sum of shadow compute expenditure and the expected cost of an undetected regression, weighted by regression probability and economic impact. This cost-sensitive formulation of the shadow sampling problem allows engineering leaders to specify their tolerance for regression risk and their willingness to pay for faster migration decisions, and the framework automatically adjusts the shadow sampling rate accordingly. This data-driven approach to infrastructure investment contrasts with the ad-hoc capacity provisioning that currently characterizes most programs for deploying production agents [15, 17].

5. Evaluation and Broader Implications

5.1 Simulation Results and Comparative Performance

Evaluation of the proposed framework was conducted in two stages. The first stage applied the framework retrospectively to archived production query logs from an enterprise analytical platform processing natural language to SQL workloads, comprising approximately twelve thousand queries spanning forty-five calendar days. These logs were drawn from a system handling advertising analytics queries with documented requirements for service level agreements of sub-fifty-millisecond median latency and ninety-ninth percentile latency below two hundred milliseconds. The query distribution reflected real production heterogeneity: approximately thirty-one percent were simple single-table lookups, forty-four percent were multi-join aggregation queries, and twenty-five percent were multi-step analytical chains involving subqueries or window functions. Diurnal traffic patterns were preserved from the original logs. In the second stage, six quality events were synthetically injected at known timestamps into this real query stream to enable ground-truth evaluation of detection performance: an accuracy improvement, an accuracy degradation, a latency regression, an intermittent hallucination increase, a subtle authorization boundary violation, and a combined event involving improvement on one quality dimension concurrent with degradation on another. This hybrid design, which combines

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

real query distributions with controlled synthetic injections, provides a more ecologically valid evaluation basis than a fully synthetic workload while retaining the ground-truth labels necessary for quantitative performance assessment [7].

Two agent configurations served as the incumbent and challenger, respectively. The incumbent was a production-deployed natural language to SQL system based on an instruction-tuned large language model, with established baseline metrics of 84.2 percent analytical correctness on the golden reference set, a median response latency of 38 milliseconds, a semantic equivalence score of 0.91 on the execution-trace alignment scale, and a safety violation rate of 0.4 percent. The challenger incorporated a revised prompt engineering pipeline with chain-of-thought decomposition targeting improved correctness on complex multi-step queries. Per-dimension results under the Dual-Serve Agent System were as follows. On analytical correctness, the challenger achieved 89.6 percent, a 5.4 percentage point improvement, and the Sequential Probability Ratio Test accepted the alternative hypothesis after 312 paired shadow observations. On Latency Profile, the challenger's median latency was 43 milliseconds against the incumbent's 38 milliseconds; the ninety-ninth percentile latency of 187 milliseconds remained within the two-hundred-millisecond service level agreement threshold and did not trigger a latency rejection. On semantic equivalence, the challenger's execution-trace alignment score was 0.88, marginally below the incumbent's 0.91, reflecting structural differences introduced by chain-of-thought decomposition; the test required 614 paired observations before reaching a decision, making the number of paired observations the binding constraint on the shadow phase duration. On Safety and Faithfulness, the challenger's violation rate was 0.3 percent against the incumbent's 0.4 percent, and the test accepted the alternative hypothesis after 489 observations. The framework detected ninety-six percent of quality regressions before any user impact, compared to fifty-eight percent for standard canary deployment evaluated over a fixed seven-day window. Mean time to safe migration was reduced by forty-one percent relative to fixed-window evaluation. The false alarm rate remained below four percent across all evaluation runs, and total user impact duration was near zero for the proposed framework compared to an average of two point three days for standard canary deployment [5, 6]. Notably, the combined quality event correctness improvement concurrent with latency degradation was the only event type that standard canary deployment missed entirely, confirming the necessity of the conjunctive multi-dimensional stopping rule over scalar quality aggregation.

5.2 Broader Implications for Agent Deployment Engineering

The results of this evaluation have implications that extend beyond any individual deployment scenario. Industrial AI deployment practices are currently characterized by substantial gaps between the theoretical properties of deployed models and the empirical behaviors observed under production conditions, gaps that arise from the complexity of real-world query distributions, infrastructure variability, and the interaction between model behavior and serving system state [18]. The proposed framework directly addresses these gaps by grounding migration decisions in statistical evidence derived from real production workloads rather than from synthetic test suites or offline benchmarks, which systematically underrepresent the complexity of production query distributions [6].

The framework formalizes agent migration engineering as a discipline at the intersection of distributed systems, sequential statistics, and artificial intelligence deployment. Distributed simplex architectures provide a formal precedent for the dual-path processing model that underlies the Dual-Serve Agent System [7], and multi-agent migration strategies based on reinforcement learning demonstrate that adaptive, data-driven migration decisions can substantially outperform static migration policies under dynamic workload conditions [5]. The formal statistical guarantees on migration decisions replace the ad-hoc monitoring and intuition-based cutover decisions that currently characterize production agent deployments with a principled, auditable protocol whose error properties are known and configurable. The cost model and adaptive sampling capabilities ensure that this rigor is achievable within realistic infrastructure budgets, making formal migration safety operationally accessible rather than aspirational [18].

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6. Limitations

Several limitations of the present framework and its evaluation warrant explicit acknowledgment. First, the evaluation relies on archived production query logs with synthetically injected quality events rather than a live prospective deployment. While the hybrid design provides more ecological validity than a fully synthetic workload, it cannot capture all dynamics of a real-time migration, including the behavioral differences between shadow-priority and live-priority query scheduling under genuine production load conditions. The framework's detection performance under continuous live deployment, where quality drift may be gradual and multi-causal rather than event-injected, remains to be validated in a prospective setting.

Second, the evaluation covers six injected quality event types. While these were selected to represent a broad range of regression modes, they do not exhaust the space of failure patterns that may arise in production analytical agents. In particular, slow-onset distributional drift, seasonal query pattern shifts, and compounding failures across dimensions were not explicitly modeled. The generalizability of the reported detection rates to these event types requires further empirical study.

Third, the framework assumes that query duplication to the shadow path is architecturally feasible. In serving systems where queries carry stateful side effects, where duplication would violate data consistency guarantees, or where the compute cost of duplication is prohibitive even with stratified sampling, the framework may require structural adaptation before deployment. The stratified sampling variant partially addresses the cost constraint, but its statistical validity under extreme sampling ratios has not been formally characterized.

Fourth, the formal cost model that derives the optimal shadow sampling fraction has not been empirically validated against observed infrastructure costs. The model's assumptions, including the independence of regression probability from query rate and the linear relationship between shadow fraction and compute cost, are simplifications that may not hold under all serving configurations. Empirical calibration of the cost model against observed deployment economics is a necessary step before using it for infrastructure investment decisions in production environments.

Conclusion

The deployment of non-deterministic analytical agents in production serving systems constitutes a qualitatively distinct engineering challenge that cannot be resolved by extending the principles and practices developed for deterministic software. The stochastic nature of agent quality, the multi-dimensionality of the quality space, and the strict latency and throughput requirements of modern serving infrastructure collectively invalidate the assumptions underlying canary releases, blue-green deployments, and fixed-window A/B testing. Each of these conventional strategies fails in a characteristic way: canary releases confound agent quality with user population heterogeneity; blue-green deployments expose the entire user base to unvalidated agent versions; fixed-window A/B testing wastes evaluation time for obvious differences and fails to detect subtle regressions within fixed horizons. The framework proposed in this article addresses all three failure modes through a principled synthesis of two mature engineering traditions. The adaptation of dual-write migration patterns from distributed database engineering eliminates population confounding by routing identical production queries to both the incumbent and challenger agents, ensuring that quality comparisons are based on paired observations of identical inputs rather than heterogeneous user populations. The shadow path architecture ensures zero user impact during the evaluation period, shifting all regression risk from users to the evaluation infrastructure. The integration of Wald's Sequential Probability Ratio Test into the Confidence-Gated Migration Protocol replaces fixed-horizon evaluation with adaptive decision-making that reaches conclusions at the earliest moment statistical evidence supports them, reducing evaluation costs and regression exposure simultaneously.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The multi-dimensional quality assessment framework, with its conjunctive stopping rule and familywise error control, ensures that migration decisions are holistic: a challenger that improves analytical correctness while regressing on latency or safety is correctly withheld from production, a failure mode that scalar quality metrics systematically miss. The phased migration protocol, structured around Site Reliability Engineering error budget principles, provides a principled path from shadow evaluation through canary deployment to full production serving, with continuous rollback capability at every phase. The formal cost model enables data-driven infrastructure investment decisions, allowing organizations to specify their tolerance for regression risk and derive the shadow sampling rate that minimizes total deployment cost. Collectively, these contributions establish agent migration engineering as a distinct discipline that draws on distributed systems architecture, sequential statistical theory, and artificial intelligence serving practices. As organizations increasingly depend on analytical agents for mission-critical decision support, the ability to update these agents with confidence with formal guarantees that quality regressions will be detected before users are affected becomes a foundational infrastructure capability. The framework is designed to be compatible with the operational and economic constraints of real-world service systems, and its evaluation on production-derived workloads suggests this level of rigor is feasible in principle, though empirical validation of the cost model against observed infrastructure economics remains a necessary step before broader claims can be substantiated. Subject to this validation, the framework offers a principled and auditable alternative to the ad-hoc practices that currently characterize production agent deployment.

References

[1] Michael E. Akintunde et al., "Formal verification of neural agents in non-deterministic environments," International Conference on Autonomous Agents and Multiagent Systems, 2020. Available: https://pkouvaros.github.io/publications/AAMAS20-A+/paper.pdf
[2] Alessandro Romano, "Enhancing Operational Resilience through Error Budgeting in Financial Site Reliability Engineering: A Comprehensive Framework," EuroLexis Research Index of International Multidisciplinary Journal for Research & Development, 2026. Available: https://researchcitations.org/index.php/elriijmrd/article/view/97
[3] Melissa Z. Pan et al., "Measuring Agents in Production," arXiv, 2026. Available: https://arxiv.org/pdf/2512.04123
[4] Nicholas Larsen et al., "Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology," The American Statistician, 2023. Available: https://www.tandfonline.com/doi/full/10.1080/00031305.2023.2257237
[5] Florian Auer et al., "Controlled Experimentation in Continuous Experimentation: Knowledge and Challenges," arXiv Preprint submitted to Information and Software Technology, 2021. Available: https://arxiv.org/pdf/2102.05310
[6] Aayush Gupta, "ReliabilityBench: Evaluating LLM Agent Reliability Under Production-Like Stress Conditions," arXiv, 2026. Available: https://arxiv.org/pdf/2601.06112
[7] Usama Mehmood et al., "A distributed simplex architecture for multi-agent systems," 2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S1383762122002697
[8] Kyu Min Shim, "Sequential Test for Practical Significance: Truncated Mixture Sequential Probability Ratio Test," arXiv, 2025. Available: https://arxiv.org/abs/2509.07892
[9] Ryan Harvey et al., "Sequential Hypothesis Testing Based on Machine Learning," International Conference on Information Fusion, 2024. Available: https://ieeexplore.ieee.org/document/10706304
[10] Sreemaee Akshathala et al., "Beyond Task Completion: An Assessment Framework for Evaluating Agentic AI Systems," arXiv, 2025. Available: https://arxiv.org/abs/2512.12791
[11] Tao Xiong et al., "Covariance Estimation and its Application in Large-Scale Online Controlled Experiments," arXiv, 2021. Available: https://arxiv.org/pdf/2108.02668
[12] Sushant Mehta, "Beyond Accuracy: A Multi-Dimensional Framework for Evaluating Enterprise Agentic AI Systems," arXiv, 2025. Available: https://arxiv.org/pdf/2511.14136

Research Article

[13] Jesus Climent, "How maintenance windows affect your error budget—SRE tips," Google Technical Report, 2020. Available: https://cloud.google.com/blog/products/management-tools/sre-error-budgets-and-maintenance-windows
[14] Betsy Beyer et al., “Site Reliability Engineering—How Google runs production systems," O’Reilly Media, 2016. Available: https://repo.darmajaya.ac.id/4636/1/Site%20Reliability%20Engineering_%20How%20Google%20Runs%20Production%20Systems%20%28%20PDFDrive%20%29.pdf
[15] Jiang Wu et al., "Optimizing Latency-Sensitive AI Applications Through Edge-Cloud Collaboration," Journal of Advanced Computing Systems, 2023. Available: https://scipublication.com/index.php/JACS/article/view/163/154
[16] Prudhvi Naayini, "Scalable AI Model Deployment and Management on Serverless Cloud Architecture," International Journal of Electrical, Electronics, and Computers, 2024. Available: https://aipublications.com/uploads/issue_files/Scalable.pdf
[17] Sebastian Barros, "Solving AI Foundational Model Latency with Telco Infrastructure," arXiv, 2025. Available: https://arxiv.org/pdf/2504.03708
[18] Sudhi Sinha & Young M. Lee, "Challenges with developing and deploying AI models and applications in industrial systems," Discover Artificial Intelligence (Springer Nature), 2024. Available: https://link.springer.com/article/10.1007/s44163-024-00151-2
[19] Abraham Wald, "Sequential Analysis," John Wiley and Sons, 1945.
[20] E. S. Page, "Continuous Inspection Schemes," Biometrika, 1954. Available: https://www.jstor.org/stable/2333009
[21] Sture Holm, "A Simple Sequentially Rejective Multiple Test Procedure," Scandinavian Journal of Statistics, 1979. Available: https://www.jstor.org/stable/4615733