International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

A Reliability-Driven Framework for Service Level Governance and Error Budget Optimization in Large-Scale Language Model Inference Systems

Christopher L. Davenport
University of Queensland, Australia

Abstract: The unprecedented growth of large-scale language model (LLM) inference systems has introduced a new generation of cloud-native digital services that operate at massive scale while being subject to stringent reliability and performance expectations. As these systems increasingly support mission-critical workloads such as intelligent assistants, enterprise automation, and real-time decision support, the need for structured governance mechanisms that balance innovation velocity with service stability has become paramount. Site Reliability Engineering (SRE), originally developed within hyperscale web companies, has emerged as a leading paradigm for reconciling this tension through the formalization of Service Level Objectives, error budgets, and continuous operational learning. In parallel, the LLM inference ecosystem has rapidly evolved through innovations in batching, scheduling, and performance tuning that seek to optimize throughput-latency tradeoffs under volatile demand. Despite the convergence of these two domains, existing literature has largely treated reliability engineering and LLM inference optimization as separate research trajectories, leaving an unresolved theoretical and methodological gap regarding how SRE principles can be systematically embedded into large-scale model-serving infrastructures.

This study develops a comprehensive, reliability-driven framework for Service Level governance in LLM inference systems by synthesizing error budget management concepts from classical SRE theory with state-of-the-art inference optimization techniques. Building on the foundational work of Dasari (2025), which articulates how error budgets serve as the operational fulcrum between innovation and stability in large-scale systems, this article argues that error budgets can be reinterpreted as first-class control variables within LLM-serving platforms. By aligning batching strategies, request scheduling, and performance tuning with dynamic reliability budgets, providers can move beyond static Service Level Agreements toward adaptive, self-regulating systems capable of maintaining user-perceived quality of experience under fluctuating workloads.

The methodology of this research is interpretive and design-oriented, integrating cross-domain literature from operating systems research, cloud economics, and causal inference to construct a conceptual architecture for reliability-aware inference. Prior studies on throughput-latency tradeoffs, generation length prediction, and SLO-oriented tuning are critically examined to demonstrate how their performance-centric objectives can be reframed in reliability terms. At the same time, the article draws upon quality of experience models and service level objective frameworks to show how user-facing metrics can be causally linked to internal error budget consumption. Through this synthesis, the study proposes a multi-layer governance model in which error budgets guide operational decisions at the level of infrastructure, inference engines, and application services.

The results of this conceptual analysis indicate that reliability-driven optimization produces qualitatively different system behaviors than purely performance-driven tuning. When error budgets are treated as finite, exhaustible resources, system designers are incentivized to

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

allocate computational capacity, batching depth, and admission control in ways that maximize long-term service sustainability rather than short-term throughput. This shift has significant implications for cloud economics, as it enables more predictable cost structures and more transparent tradeoffs between user experience and infrastructure expenditure. Moreover, by embedding causal reasoning into reliability management, operators can more accurately diagnose the origins of service degradation and target corrective actions with minimal disruption.

The discussion situates these findings within broader debates about the future of cloud and edge intelligence, highlighting how reliability-aware governance can support the scaling of LLM-powered services across heterogeneous and distributed environments. Limitations related to the absence of empirical deployment data are acknowledged, and directions for future research are outlined, including the integration of Bayesian reliability models and automated SLO negotiation mechanisms. Overall, this study contributes a theoretically grounded and practically relevant framework that advances the state of knowledge at the intersection of SRE and large-scale AI system engineering.

Keywords: Site reliability engineering, error budget management, service level objectives, large language model inference, cloud performance governance, quality of experience

INTRODUCTION

The contemporary digital economy is increasingly shaped by large-scale language model inference systems that act as the computational substrate for intelligent assistants, conversational agents, and a wide array of knowledge-driven applications. These systems, exemplified by modern families of transformer-based models, have transformed how users interact with information and services, enabling natural language interfaces that operate in real time across global networks (Dong et al., 2023). At the same time, the operational complexity of delivering such services at scale has grown dramatically, as providers must simultaneously satisfy stringent latency requirements, maintain high availability, and manage the immense computational cost associated with model execution (Agrawal et al., 2024). This confluence of technological ambition and operational constraint has elevated reliability from a peripheral engineering concern to a central strategic imperative.

Historically, the challenge of balancing system reliability with rapid innovation has been addressed through the discipline of Site Reliability Engineering, which reframes operational management as a problem of probabilistic risk and controlled experimentation (Dasari, 2025). Within this paradigm, Service Level Objectives and error budgets function as quantitative representations of acceptable failure, allowing organizations to trade a bounded amount of unreliability for the freedom to deploy new features and optimizations. In large-scale web services, this approach has proven effective in aligning development teams and operations teams around shared, user-centric goals, thereby reducing the cultural and technical friction that traditionally accompanies system growth (Eccleston, 2012). However, the application of these principles to the domain of LLM inference remains underdeveloped, even as such systems now rival or exceed the scale of classical web platforms.

eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

The literature on LLM inference has, until recently, been dominated by a focus on performance optimization, particularly the mitigation of throughput-latency tradeoffs that arise when serving heterogeneous workloads (Agrawal et al., 2024). Techniques such as chunked prefills, piggybacked decodes, and adaptive batching have been proposed to increase hardware utilization while preserving acceptable response times for end users (Agrawal et al., 2023). Similarly, recent work on generation length prediction and SLO-oriented tuning has sought to make inference engines more responsive to workload variability by anticipating the computational cost of individual requests and adjusting scheduling policies accordingly (Cheng et al., 2024; Cheng et al., 2025). While these contributions have significantly advanced the state of the art in inference efficiency, they are typically evaluated in terms of performance metrics such as latency percentiles and throughput, rather than in terms of reliability budgets and long-term service sustainability.

This performance-centric orientation reflects a deeper theoretical assumption within systems research: that optimizing average or tail latency is a sufficient proxy for user satisfaction and service quality. Yet decades of research in quality of experience have demonstrated that human perceptions of service reliability are shaped not only by instantaneous delays but also by patterns of variability, expectation, and trust over time (Egger et al., 2012). When an intelligent assistant intermittently fails to respond or produces excessively delayed outputs, the resulting erosion of user confidence can be more damaging than a modest, consistent slowdown. From this perspective, the absence of explicit reliability governance in LLM inference frameworks represents a significant gap in both theory and practice.

The work of Dasari (2025) provides a critical foundation for addressing this gap by articulating how error budgets can be used to manage operational risk in large-scale systems. Rather than treating failures as anomalies to be eliminated at all costs, error budgets define a permissible envelope of unreliability within which innovation can safely occur. This concept is particularly salient for LLM inference, where the deployment of new model versions, optimization algorithms, and hardware accelerators inherently carries the risk of performance regressions and service disruptions. By embedding error budgets into the control plane of inference systems, providers can create feedback loops that regulate the pace of change based on observed reliability outcomes.

At the same time, the cloud computing literature has long emphasized the economic and architectural challenges of delivering high-performance services on shared, elastic infrastructure (Kilcioglu et al., 2017; Hwang et al., 2016). Public cloud platforms exhibit significant heterogeneity in hardware performance, even within nominally identical instance types, leading to unpredictable execution times and resource contention (Ou et al., 2012). For LLM inference, which is highly sensitive to memory bandwidth, compute throughput, and interconnect latency, such variability can translate directly into user-visible performance fluctuations. Traditional performance tuning approaches struggle to accommodate this uncertainty, whereas a reliability-driven framework that incorporates error budgets and Service Level Objectives offers a more robust mechanism for coping with infrastructural volatility.

The emergence of modeling languages and frameworks for cloud and multicloud environments further underscores the need for formalized governance structures

eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

(Goncalves et al., 2011; Guillen et al., 2013; Rossini et al., 2017). These tools enable architects to describe resource requirements, service dependencies, and deployment constraints in a declarative manner, facilitating automated provisioning and adaptation. However, they rarely incorporate explicit representations of reliability budgets or SLO compliance, focusing instead on functional and performance attributes. As a result, they provide limited support for the kind of continuous, reliability-aware decision making envisioned by SRE theory (Dasari, 2025).

The integration of causal inference into systems management offers another promising avenue for bridging this gap. By modeling the relationships between configuration changes, workload characteristics, and observed service outcomes, causal frameworks enable operators to distinguish correlation from causation and to predict the impact of interventions (Pearl, 2009). In the context of LLM inference, causal models could be used to estimate how adjustments to batching policies or hardware allocations affect error budget consumption and user experience. This perspective aligns with recent work on Bayesian and probabilistic approaches to anomaly detection and resilience assessment in distributed systems (Odiathevar et al., 2022; Yazdi et al., 2022), which emphasize the importance of uncertainty-aware reasoning in complex, stochastic environments.

Despite these converging lines of research, there remains a lack of a unified theoretical framework that connects SRE error budget management with the specific operational dynamics of large-scale LLM inference. Existing studies tend to focus either on high-level reliability governance in generic cloud systems or on low-level performance optimization in model-serving engines, without articulating how these layers interact. This fragmentation limits the ability of practitioners and researchers to reason holistically about the tradeoffs involved in deploying and evolving AI-powered services.

The present study seeks to address this literature gap by developing a reliability-driven framework for Service Level governance in LLM inference systems. Drawing on the principles articulated by Dasari (2025) and integrating them with recent advances in inference optimization, cloud performance modeling, and causal analysis, the article proposes a conceptual architecture in which error budgets function as the central coordinating mechanism across infrastructure, platform, and application layers. By situating performance tuning within a broader reliability context, this framework aims to reconcile the competing imperatives of responsiveness, scalability, and stability that define the modern AI service landscape.

In articulating this framework, the study also engages with broader debates about the future of intelligent assistants and distributed computing. As LLM-powered services increasingly migrate toward the edge and operate across heterogeneous networks, the challenges of maintaining consistent quality of experience will intensify (Casamayor Pujol et al., 2023; Nastic et al., 2020). A reliability-centric approach offers a principled way to navigate these complexities by making tradeoffs explicit and by grounding operational decisions in user-centric metrics. Through an extensive theoretical and critical analysis of the relevant literature, this article aims to contribute a foundational perspective that can inform both academic inquiry and industrial practice in the evolving domain of large-scale AI system engineering.

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

METHODOLOGY

The methodological orientation of this research is interpretive, integrative, and design-oriented, reflecting the complex and interdisciplinary nature of reliability management in large-scale language model inference systems. Rather than relying on controlled experiments or proprietary operational data, which are often inaccessible in the context of hyperscale AI services, this study employs a structured synthesis of existing scholarly and technical literature to construct a coherent conceptual framework. This approach is grounded in the recognition that many of the most significant advances in systems engineering and cloud governance have emerged from theoretical integration and analytical generalization across diverse empirical contexts (Hwang et al., 2016; Kilcioglu et al., 2017).

At the core of the methodology lies a comparative analysis of two bodies of work that have historically evolved in parallel: Site Reliability Engineering and large-scale LLM inference optimization. SRE literature, as exemplified by Dasari (2025), provides normative and descriptive models for how organizations define, measure, and manage service reliability through constructs such as Service Level Objectives and error budgets. In contrast, the LLM inference literature focuses on algorithmic and architectural techniques for improving performance, including batching, scheduling, and resource allocation (Agrawal et al., 2024; Cheng et al., 2025). By systematically mapping the concepts, assumptions, and objectives of these two domains onto one another, the study seeks to identify points of alignment, tension, and potential synthesis.

The first stage of the methodology involves a thematic coding of the reference corpus to extract key constructs related to reliability, performance, and governance. Works on throughput-latency tradeoffs and inference engine design are analyzed to identify how they conceptualize service quality and user experience, even when these terms are not explicitly invoked (Agrawal et al., 2023; Cheng et al., 2024). In parallel, studies on quality of experience and Service Level Objectives are examined to elucidate how user perceptions of reliability are shaped by system behavior over time (Egger et al., 2012; Nastic et al., 2020). This dual coding process enables the identification of conceptual gaps, such as the absence of error budget reasoning in performance optimization research or the lack of fine-grained computational models in SRE-oriented frameworks.

The second stage employs a form of analytical triangulation, drawing on cloud modeling languages and economic analyses to situate LLM inference within broader infrastructural and organizational contexts. Modeling frameworks such as CloudML, CAMEL, and Stratus ML are used as reference points for understanding how complex cloud services are specified, deployed, and evolved (Goncalves et al., 2011; Hamdaqa and Tahvildari, 2015; Rossini et al., 2017). These frameworks provide a vocabulary for describing multi-layered systems, which is essential for articulating how error budgets might propagate from user-facing applications down to underlying hardware resources. Economic analyses of public cloud usage and performance heterogeneity further inform the discussion by highlighting the financial and technical constraints under which inference systems operate (Kilcioglu et al., 2017; Ou et al., 2012).

The third stage integrates causal inference and probabilistic reasoning to support the design of reliability-aware control mechanisms. Drawing on Pearl's (2009) foundational work on causal models, the

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

methodology conceptualizes error budget consumption as an outcome variable influenced by a network of operational decisions, workload characteristics, and infrastructural conditions. This perspective is reinforced by studies that apply Bayesian and probabilistic models to anomaly detection and resilience assessment in distributed systems (Odiathevar et al., 2022; Yazdi et al., 2022). By incorporating these insights, the framework can account for uncertainty and partial observability, which are inherent features of large-scale, distributed AI services.

Throughout this methodological process, particular attention is paid to the principle of theoretical saturation, ensuring that the integration of concepts from different domains is sufficiently deep to support robust analytical claims (Dasari, 2025). Rather than treating references as isolated contributions, the study seeks to weave them into a coherent narrative that explains how reliability governance can be operationalized in the specific context of LLM inference. This involves not only identifying complementarities but also critically engaging with contradictions and limitations in the existing literature. For example, while SLO-oriented tuning mechanisms promise to align system behavior with user expectations, they may conflict with aggressive throughput maximization strategies when resources are scarce (Cheng et al., 2025; Agrawal et al., 2024).

The design-oriented dimension of the methodology culminates in the construction of a conceptual architecture for reliability-driven inference. This architecture is not presented as a concrete software implementation but as a set of interrelated principles and control loops that could, in principle, be instantiated in a variety of technological environments. By articulating how error budgets, SLOs, and performance metrics interact across system layers, the framework provides a basis for both analytical reasoning and practical experimentation.

A key limitation of this methodology is its reliance on secondary sources and theoretical reasoning rather than on direct empirical validation. While the referenced studies provide rich insights into specific aspects of LLM inference and cloud reliability, the absence of unified datasets or benchmarks that explicitly measure error budget dynamics in AI services constrains the ability to draw quantitative conclusions (Agrawal et al., 2024; Dasari, 2025). Nevertheless, the interpretive and integrative nature of the approach is well suited to the exploratory goals of this research, which seeks to establish a conceptual foundation for future empirical and engineering work.

By grounding its analysis in a diverse and rigorously selected body of literature, the methodology ensures that the resulting framework is both theoretically informed and practically relevant. In doing so, it advances a form of scholarship that recognizes the co-evolution of technology, organization, and user experience as central to the reliability of modern AI-powered services.

RESULTS

The synthesis of Site Reliability Engineering theory and large-scale LLM inference research yields several important conceptual results that illuminate how reliability-driven governance can reshape the operation of AI services. One of the most significant findings is that error budgets, when interpreted as dynamic control variables rather than static thresholds, provide a unifying metric that can align diverse optimization strategies across system layers (Dasari, 2025). In traditional SRE practice, error budgets quantify the

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

amount of unreliability that a service can tolerate over a given period without violating its Service Level Objectives. When this concept is mapped onto LLM inference, it becomes possible to treat every instance of latency violation, failed request, or degraded output quality as a drawdown from a finite reliability reserve.

This reconceptualization has profound implications for how throughput-latency tradeoffs are managed. Performance-oriented techniques such as those proposed in Sarathi-Serve and related frameworks aim to maximize hardware utilization by batching and scheduling requests in ways that minimize idle time and amortize overheads (Agrawal et al., 2024; Agrawal et al., 2023). While these methods are highly effective in boosting throughput, they can also introduce tail-latency spikes when workloads are bursty or when long-generation requests block shorter ones. Under a reliability-driven regime, such tail-latency events would be explicitly accounted for as error budget consumption, forcing the system to balance aggressive batching against the risk of exhausting its reliability allowance (Dasari, 2025).

Another key result concerns the role of prediction and foresight in reliability management. Techniques for generation length prediction and SLO-oriented tuning have demonstrated that inference engines can anticipate the computational cost of individual requests and adjust scheduling policies accordingly (Cheng et al., 2024; Cheng et al., 2025). When integrated with error budget accounting, these predictive capabilities enable a form of proactive reliability control. For example, if the system detects that a sequence of long-generation requests is likely to push latency beyond acceptable thresholds, it can throttle or defer certain jobs to preserve error budget for higher-priority interactions. This moves reliability management from a reactive posture, in which violations are merely recorded after the fact, to a proactive one in which potential failures are mitigated before they occur.

The analysis also reveals that user-perceived quality of experience is more closely aligned with error budget trajectories than with isolated performance metrics. Studies on waiting times and quality of experience show that users are sensitive not only to average delays but also to the consistency and predictability of service behavior (Egger et al., 2012). By tracking error budget consumption over time, operators gain a holistic view of how often and how severely the service deviates from its intended performance envelope. This temporal perspective allows for more nuanced interpretations of user satisfaction, supporting decisions such as when to roll out new model versions or when to prioritize stability over innovation (Dasari, 2025).

From an infrastructural standpoint, the results highlight the importance of accounting for hardware heterogeneity and cloud economics in reliability governance. Public cloud environments exhibit significant variability in performance, even among instances of the same nominal type, which can lead to unpredictable inference times (Ou et al., 2012). When error budgets are used as the primary metric of service health, this variability becomes directly visible in reliability accounting, incentivizing providers to allocate workloads in ways that minimize the risk of budget depletion. This may involve preferentially assigning critical or latency-sensitive requests to more stable hardware or reserving a portion of capacity as a reliability buffer, even if doing so reduces peak throughput (Kilcioglu et al., 2017; Dasari, 2025).

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

The integration of causal inference further strengthens the reliability-driven framework by enabling more precise attribution of error budget consumption to specific operational factors. By modeling the causal relationships between configuration changes, workload patterns, and service outcomes, operators can identify which interventions are most likely to improve reliability without unnecessarily constraining performance (Pearl, 2009). This aligns with probabilistic approaches to anomaly detection and resilience assessment, which emphasize the value of uncertainty-aware diagnostics in complex distributed systems (Odiathevar et al., 2022; Yazdi et al., 2022).

Finally, the synthesis underscores the potential for modeling languages and declarative frameworks to incorporate reliability constraints alongside functional and performance requirements. While existing tools such as CloudML, CAMEL, and Stratus ML provide rich abstractions for describing cloud services, they lack native constructs for error budgets and SLO compliance (Goncalves et al., 2011; Hamdaqa and Tahvildari, 2015; Rossini et al., 2017). By extending these frameworks to include reliability metrics, architects could design and deploy LLM inference systems that are inherently aligned with SRE principles, enabling automated enforcement of reliability-driven policies across the service lifecycle (Dasari, 2025).

Together, these results suggest that a reliability-driven approach to LLM inference is not merely a conceptual overlay on existing performance optimization techniques but a fundamentally different way of structuring operational decision making. By elevating error budgets to a central role in system governance, providers can achieve a more sustainable balance between innovation, cost efficiency, and user trust in the rapidly evolving landscape of AI-powered services.

DISCUSSION

The implications of integrating Site Reliability Engineering principles into large-scale LLM inference systems extend far beyond the technical mechanics of batching and scheduling, touching on fundamental questions about how AI services should be governed, evaluated, and evolved over time. At the heart of this discussion lies the concept of the error budget, which, as articulated by Dasari (2025), serves as both a quantitative constraint and a cultural instrument that aligns diverse stakeholders around a shared understanding of acceptable risk. When applied to LLM inference, error budgets become a lens through which performance optimization, user experience, and organizational priorities can be simultaneously interpreted.

One of the most salient theoretical contributions of this reliability-driven perspective is its challenge to the prevailing performance-centric paradigm in systems research. Much of the literature on LLM inference optimization is rooted in the assumption that maximizing throughput and minimizing latency are the primary objectives of system design (Agrawal et al., 2024; Cheng et al., 2025). While these metrics are undeniably important, they provide an incomplete picture of service quality when considered in isolation. A system that achieves low average latency but exhibits frequent, unpredictable spikes may satisfy benchmark criteria while undermining user trust and long-term adoption. By contrast, a reliability-driven framework evaluates performance through the cumulative impact of deviations from SLOs, thereby capturing the temporal and experiential dimensions of service quality

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

emphasized in quality of experience research (Egger et al., 2012; Dasari, 2025).

This shift in evaluative focus has significant implications for how optimization algorithms are designed and deployed. Techniques such as piggybacked decodes and adaptive batching are often justified in terms of their ability to improve hardware utilization and reduce mean response times (Agrawal et al., 2023). However, when viewed through the lens of error budgets, these same techniques must be assessed for their potential to introduce correlated failures or tail-latency outliers that could rapidly deplete the reliability allowance. This creates a new set of design tradeoffs, in which incremental gains in throughput must be weighed against the risk of eroding the service's reliability posture (Dasari, 2025).

The integration of SLO-oriented tuning mechanisms further complicates this landscape. On the one hand, such mechanisms promise to align system behavior with explicit performance targets by dynamically adjusting parameters in response to observed workload conditions (Cheng et al., 2025). On the other hand, the existence of multiple, potentially competing SLOs for different classes of users or applications raises questions about how error budgets should be allocated and prioritized. For example, should an enterprise customer with a premium contract be granted a larger share of the error budget than a casual user of a free-tier intelligent assistant? Addressing such questions requires not only technical solutions but also normative and economic considerations, echoing the insights of cloud usage and pricing research (Kilcioglu et al., 2017; Ghrada et al., 2018).

The cloud and edge computing context adds another layer of complexity to this discussion. As LLM-powered services increasingly operate across distributed and heterogeneous environments, the sources of unreliability multiply, ranging from network congestion and hardware variability to software misconfigurations and model drift (Casamayor Pujol et al., 2023; Ou et al., 2012). A reliability-driven framework provides a principled way to aggregate these diverse risks into a single, actionable metric, but it also demands more sophisticated monitoring and diagnostic capabilities. Probabilistic and Bayesian approaches to anomaly detection and resilience assessment offer promising tools in this regard, enabling operators to quantify uncertainty and to update their beliefs about system health in light of new evidence (Odiathevar et al., 2022; Yazdi et al., 2022).

Causal inference plays a particularly important role in this evolving governance model. Without a clear understanding of how specific interventions affect error budget consumption, operators risk making changes that inadvertently exacerbate reliability problems (Pearl, 2009). For instance, scaling up a cluster of GPUs might reduce latency under certain workloads but could also introduce synchronization overheads or contention that increase the likelihood of SLO violations under others. By constructing causal models that link configuration choices to reliability outcomes, organizations can move beyond trial-and-error optimization toward a more scientifically grounded form of operational management (Dasari, 2025).

The discussion also highlights the organizational and cultural dimensions of reliability-driven governance. In classical SRE practice, error budgets serve as a shared contract between development and operations teams, mediating conflicts between the desire to deploy new features and the need to maintain stability (Dasari, 2025). In the context of LLM inference, this

International Journal of Modern Medicine
VOLUME05 ISSUE01
PUBLISHED DATE: - 31-01-2026
Page 70-81
eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

mediating function becomes even more critical, as model updates, optimization algorithms, and hardware upgrades are often driven by different teams with distinct incentives. By making the consumption of reliability a visible and quantifiable phenomenon, error budgets create a common language for negotiating tradeoffs and coordinating action across these organizational boundaries.

Nevertheless, the adoption of a reliability-driven framework is not without its challenges and limitations. One concern is the potential for over-conservatism, in which organizations become so focused on preserving error budget that they stifle innovation and fail to capitalize on opportunities for performance improvement (Agrawal et al., 2024). This risk underscores the importance of setting appropriate SLOs and error budgets that reflect both user expectations and competitive pressures. Another limitation lies in the difficulty of measuring and attributing error budget consumption in complex, multi-tenant environments, where failures may be caused by interactions between multiple services and infrastructure components (Goncalves et al., 2011; Rossini et al., 2017).

Future research can address these limitations by developing more granular and context-aware reliability metrics, as well as by exploring automated mechanisms for negotiating and adjusting SLOs in response to changing conditions. The integration of modeling languages with real-time monitoring data offers one promising direction, enabling the creation of digital twins that simulate the impact of potential interventions on error budgets and service quality (Hamdaqa and Tahvildari, 2015; Dasari, 2025). Similarly, advances in edge intelligence and distributed computing open new avenues for decentralizing reliability management, allowing local nodes to make context-sensitive decisions that contribute to global service stability (Casamayor Pujol et al., 2023; Nastic et al., 2020).

In sum, the reliability-driven framework articulated in this study represents a significant departure from conventional performance-centric approaches to LLM inference. By foregrounding error budgets and Service Level Objectives as the primary instruments of governance, it offers a more holistic and sustainable way to manage the complex tradeoffs inherent in delivering AI-powered services at scale. While many practical and theoretical questions remain, the synthesis of SRE principles with modern inference optimization provides a fertile ground for future innovation and scholarly inquiry.

CONCLUSION

The rapid expansion of large-scale language model inference systems has fundamentally altered the landscape of cloud-based digital services, elevating reliability from a secondary operational concern to a central determinant of user trust and organizational viability. Through an extensive theoretical and integrative analysis, this study has demonstrated that Site Reliability Engineering principles, and in particular the concept of error budget management articulated by Dasari (2025), offer a powerful framework for governing the complex tradeoffs between performance, cost, and stability in AI-powered services. By reconceptualizing error budgets as dynamic control variables that guide operational decisions across infrastructure, platform, and application layers, the research provides a coherent lens through which the diverse challenges of LLM inference can be understood and addressed.

The findings suggest that a reliability-driven approach enables more sustainable

eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

optimization strategies than those focused solely on throughput and latency. By explicitly accounting for the cumulative impact of service degradations on user experience, error budgets encourage a form of governance that prioritizes long-term service quality over short-term performance gains (Agrawal et al., 2024; Egger et al., 2012). Moreover, the integration of predictive and causal techniques into reliability management supports more proactive and scientifically grounded interventions, reducing the likelihood of unintended consequences and enabling continuous learning in complex, distributed environments (Pearl, 2009; Odiathevar et al., 2022).

While the framework proposed in this article is conceptual in nature, it lays a foundation for future empirical and engineering work aimed at operationalizing reliability-aware inference in real-world systems. As LLM-powered services continue to proliferate across cloud and edge infrastructures, the need for principled, user-centric governance mechanisms will only intensify. By bringing together insights from SRE, cloud economics, quality of experience, and inference optimization, this study contributes a holistic perspective that can guide both scholarly research and industrial practice in the next generation of intelligent digital services.

REFERENCES

1. Agrawal, Amey, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, and Ramachandran Ramjee. 2023. SARATHI: Efficient LLM Inference by Piggybacking Decodes with Chunked Prefills. arXiv:2308.16369.
2. Pearl, Judea. 2009. Causal inference in statistics: An overview. Statistics Surveys, 3, 96–146.
3. Goncalves, Glauco Estacio, Patricia Endo, Marcelo Santos, Djamel Sadok, Judith Kelner, Bob Melander, and Jan Erik Mangs. 2011. CloudML: An Integrated Language for Resource, Service and Request Description for D-Clouds. CloudCom.
4. Agrawal, Amey, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. In 18th USENIX Symposium on Operating Systems Design and Implementation.
5. Kilcioglu, Cinar, Justin M. Rao, Aadharsh Kannan, and R. Preston McAfee. 2017. Usage Patterns and the Economics of the Public Cloud. In Proceedings of the 26th International Conference on World Wide Web.
6. Dasari, H. 2025. Site Reliability Engineering Practices for Error Budget Management in Large-Scale Systems. International Journal of Applied Mathematics, 38(5s), 991–1001.
7. Egger, Sebastian, Tobias Hossfeld, Raimund Schatz, and Markus Fiedler. 2012. Waiting times in quality of experience for web based services. In Fourth International Workshop on Quality of Multimedia Experience.
8. Cheng, Ke, Zhi Wang, Wen Hu, Tiannuo Yang, Jianguo Li, and Sheng Zhang. 2025. SCOOT: SLO-Oriented Performance Tuning for LLM Inference Engines. arXiv:2408.04323.
9. Ou, Zhonghong, Hao Zhuang, Jukka K. Nurminen, Antti Yla Jaaski, and Pan Hui. 2012. Exploiting Hardware Heterogeneity within the Same Instance Type of Amazon EC2. In Hot Topics in Cloud Computing.
10. Rossini, Alessandro, Kiriakos Kritikos, Nikolay Nikolov, Jorg Domaschka, Frank Griesinger, Daniel Seybold, Daniel Romero, Michal Orzechowski, Georgia

eISSN: 2156-5198 pISSN: 2156-518X

RESEARCH ARTICLE

Kapitsaki, and Achilleas Achilleos. 2017. The Cloud Application Modelling and Execution Language CAMEL. Technical Report, Universitat Ulm.

11. Cheng, Ke, Wen Hu, Zhi Wang, Peng Du, Jianguo Li, and Sheng Zhang. 2024. Enabling Efficient Batch Serving for LMaaS via Generation Length Prediction. arXiv:2406.04785.
12. Casamayor Pujol, Victor, P. K. Donta, A. Morichetta, I. Murturi, and Schahram Dustdar. 2023. Edge Intelligence Research Opportunities for Distributed Computing Continuum Systems. IEEE Internet Computing.
13. Nastic, Stefan, A. Morichetta, T. Pusztai, S. Dustdar, X. Ding, D. Vij, and Y. Xiong. 2020. SLOC: Service Level Objectives for Next Generation Cloud Computing. IEEE Internet Computing.
14. Hamdaqa, Mohammad, and Ladan Tahvildari. 2015. Stratus ML: A Layered Cloud Modeling Framework. IEEE International Conference on Cloud Engineering.
15. Hwang, Kai, Xiaosong Bai, Yihua Shi, Min Li, W. G. Chen, and Y. Wu. 2016. Cloud Performance Modeling with Benchmark Evaluation of Elastic Scaling Strategies. IEEE Transactions on Parallel and Distributed Systems, 27(1), 130–143.
16. Odiathevar, M., W. K. Seah, and M. Frean. 2022. A Bayesian Approach to Distributed Anomaly Detection in Edge AI Networks. IEEE Transactions on Parallel and Distributed Systems.
17. Yazdi, M., F. Khan, R. Abbassi, and N. Quddus. 2022. Resilience assessment of a subsea pipeline using dynamic Bayesian network. Journal of Pipeline Science and Engineering, 2(2), 100053.
18. Dong, Xin Luna, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards next-generation intelligent assistants leveraging LLM techniques. Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
19. Ghrada, Nadir, Mohamed Faten Zhani, and Yehia Elkhatib. 2018. Price and Performance of Cloud-hosted Virtual Network Functions: Analysis and Future Challenges. PVE-SDN.