Reliability-Aware Error Budget Governance and Resource Orchestration for Large-Scale Language Model Serving Infrastructures

[1] Kaniel Verhoeven
[1] Department of Computer Science, Delft University of Technology, Netherlands

Received: 22th Nov 2025 | Received Revised Version: 16th Dec 2025 | Accepted: 27th Jan 2026 | Published: 31st Jan 2026

Volume 08 Issue 01 2026 |

Abstract

The rapid industrialization of large-scale language models has transformed contemporary digital services into reliability-critical socio-technical systems whose failure modes propagate across economic, institutional, and cognitive domains. While traditional Site Reliability Engineering (SRE) frameworks have long governed the stability of web-scale platforms, the arrival of transformer-based generative models introduces unprecedented variability in workload, latency, cost, and quality. This article develops a unified theoretical and methodological framework that integrates classical error budget management with the emerging operational realities of large language model (LLM) serving infrastructures. Anchored in the reliability-centered principles articulated by Dasari (2025), this study expands the concept of error budgets from static service-level constructs into adaptive, learning-driven governance mechanisms capable of handling bursty inference traffic, heterogeneous hardware topologies, and stochastic model behaviors. Drawing on recent advances in transformer architectures, distributed inference engines, resource-aware scheduling, and LLM evaluation frameworks, the article constructs a multi-layered analytical model that explains how reliability objectives can be decomposed, monitored, enforced, and renegotiated across the computing continuum. The methodology combines systems-theoretic reasoning, socio-technical risk analysis, and interpretive synthesis of contemporary research to derive a set of reliability patterns for AI-native infrastructures. The results demonstrate that error budgets, when reinterpreted through probabilistic service envelopes and adaptive feedback control, can serve as the central coordination primitive between offline training, online inference, and user-facing service-level objectives. The discussion situates these findings within broader debates on cloud servitization, cyber-physical monitoring, causal modeling, and digital transformation, revealing how reliability engineering becomes the institutional backbone of trustworthy AI. The article concludes by outlining a future research agenda in which reliability budgets evolve into market-like governance instruments mediating trade-offs between performance, cost, sustainability, and societal trust.

Keywords: Site reliability engineering, error budget management, large language model serving, distributed inference, service level objectives, adaptive scheduling, digital infrastructure governance

Cite This Article: Kaniel Verhoeven. (2026). Reliability-Aware Error Budget Governance and Resource Orchestration for Large-Scale Language Model Serving Infrastructures. The American Journal of Interdisciplinary Innovations and Research, 8(01), 155–160. Retrieved from https://theamericanjournals.com/index.php/tajiir/article/view/7437

1. Introduction

The modern digital economy is increasingly mediated by algorithmic systems whose performance is no longer merely a matter of convenience but of structural dependence. From financial transactions and healthcare diagnostics to creative production and public communication, large-scale computational services have become embedded within the fabric of social life. In this context, the concept of reliability has shifted from a narrow engineering metric into a foundational requirement of institutional trust. The field of Site Reliability Engineering emerged in response to this transformation by offering a systematic framework for quantifying, governing, and continuously improving the operational stability of complex services. At the heart of this framework lies the concept of the error budget, which defines how much unreliability a system can tolerate while still meeting its service-level objectives. Dasari (2025) formalized this principle within large-scale systems by demonstrating how error budgets provide a disciplined way to balance innovation and stability through quantifiable risk. However, the rise of large language models and their deployment as real-time services introduces a new layer of complexity that strains traditional interpretations of error budgets and service reliability.

Large language models, built on transformer architectures and trained on massive corpora, differ fundamentally from earlier web services in their computational and epistemic properties. Unlike static databases or deterministic application servers, these models produce probabilistic outputs whose latency, accuracy, and cost vary with prompt structure, model size, and resource availability. Foundational work on attention-based architectures has shown that the internal mechanics of these systems involve high-dimensional interactions that scale nonlinearly with input length and model depth (Vaswani et al., 2023). Open foundation models such as those described by Touvron et al. (2023) have further expanded the range of applications, making LLMs the backbone of conversational agents, search systems, and decision-support tools. Yet this very generality creates operational unpredictability that challenges classical reliability engineering, which historically assumes more stable workload patterns and failure modes.

The infrastructure that supports LLMs is itself a layered ecosystem of distributed inference engines, heterogeneous accelerators, and adaptive schedulers. Systems such as Orca have demonstrated that transformer-based models can be served at scale through sophisticated orchestration of GPU clusters and request batching (Yu et al., 2022), while newer platforms such as DistServe and BlendServe propose disaggregated pipelines that separate prefill and decoding phases to optimize throughput and latency (Zhong et al., 2024; Zhao et al., 2024). These innovations, while technically impressive, introduce additional degrees of freedom that complicate the mapping between low-level resource behavior and high-level service reliability. In traditional SRE, a service-level objective such as ninety-nine point nine percent availability could be translated into thresholds on server uptime or request success rates. In LLM-serving environments, however, the quality of a response, its latency, and its cost are intertwined in ways that resist simple thresholding.

This article argues that the error budget paradigm articulated by Dasari (2025) remains conceptually valid in the era of AI-native services but requires a fundamental re-theorization to accommodate probabilistic computation and adaptive resource allocation. Rather than treating error budgets as fixed tolerances for failure, we propose that they be understood as dynamic envelopes within which learning systems can explore, adapt, and optimize. This shift aligns with broader movements in cyber-physical system monitoring, where real-time data streams and anomaly detection frameworks continuously update the perceived state of a system (Canizo et al., 2019; Habeeb et al., 2019). In such environments, reliability becomes not a static property but an emergent outcome of ongoing feedback loops between measurement, control, and decision-making.

The literature on service-level objective decomposition provides a useful starting point for this reconceptualization. Chen et al. (2007) showed that high-level service guarantees can be translated into system-level thresholds through formal decomposition, enabling automated management of complex infrastructures. When applied to LLM-serving systems, however, this decomposition must account for the stochastic nature of model outputs and the adaptive behavior of schedulers that respond to workload fluctuations (Zhang et al., 2024). Error budgets thus become probabilistic contracts that specify not just how often a service may fail but how much deviation in latency, accuracy, or cost is acceptable across a population of requests.

At the same time, the economic stakes of unreliability have grown dramatically. Studies on the cost of downtime in enterprise environments have shown that even brief service disruptions can have cascading financial and reputational consequences (Elliot, 2014; Melo et al., 2017). In the context of AI services, these costs are amplified by the centrality of models to user workflows and the opacity of their internal decision processes. A delayed or degraded LLM response can propagate errors into downstream systems, affecting everything from logistics planning to medical triage. Dasari (2025) emphasized that error budget management provides a governance mechanism that forces organizations to make explicit trade-offs between innovation velocity and operational risk, a principle that becomes even more critical when AI systems are involved.

The challenge, then, is to integrate the rich body of research on distributed LLM serving, adaptive scheduling, and model evaluation into a coherent reliability engineering framework that preserves the normative clarity of SRE while embracing the technical realities of AI. Recent technical reports on systems such as Yi-Lightning and Qwen2.5 illustrate how model architectures and serving strategies continue to evolve, creating moving targets for reliability governance (Wake et al., 2024; Yang et al., 2025). Meanwhile, evaluation frameworks such as MT-Bench and Chatbot Arena highlight the difficulty of defining quality in a domain where human judgment and contextual appropriateness play central roles (Zheng et al., 2024). These developments suggest that reliability in AI services cannot be reduced to uptime alone but must encompass multidimensional notions of performance and trustworthiness.

This introduction has established the conceptual terrain on which the present study is situated. By weaving together SRE theory, distributed systems research, and the operational challenges of LLM deployment, we identify a critical literature gap: the absence of a unified framework that connects error budget management with the adaptive, probabilistic nature of AI inference services. While Dasari (2025) provides a rigorous account of error budgets in large-scale systems, and the LLM serving literature offers sophisticated technical solutions for performance optimization, there remains a lack of integration between these domains. The remainder of this article seeks to fill this gap by developing a comprehensive model of reliability-aware resource orchestration for LLM infrastructures, grounded in both theoretical reasoning and interpretive synthesis of contemporary research.

2. Methodology

The methodological foundation of this study is rooted in systems-theoretic and interpretive analytical approaches that are particularly well suited to the study of complex digital infrastructures. Unlike experimental sciences that rely on controlled manipulation of variables, the investigation of large-scale LLM serving systems and their reliability governance requires a form of methodological pluralism that can integrate heterogeneous sources of evidence and theoretical insight. This study therefore adopts a qualitative-analytical methodology that synthesizes technical literature, reliability engineering theory, and socio-technical systems analysis to construct an integrative framework. The guiding principle of this approach is the recognition that reliability in AI-native infrastructures is not merely a technical attribute but an emergent property of interacting organizational, computational, and economic processes, a view that resonates with the holistic perspective advocated by Dasari (2025).

The first methodological pillar is systematic literature integration. The reference corpus encompasses research on transformer architectures and foundation models (Vaswani et al., 2023; Touvron et al., 2023), distributed LLM serving systems (Yu et al., 2022; Zhong et al., 2024; Wu et al., 2024), adaptive scheduling and co-location of workloads (Zhang et al., 2024; Wang et al., 2025), and evaluation frameworks for generative models (Zheng et al., 2024). It also includes foundational work on service-level objectives, cyber-physical monitoring, and digital infrastructure reliability (Chen et al., 2007; Canizo et al., 2019; Habeeb et al., 2019). By treating these diverse contributions as components of a single conceptual landscape, the study avoids the fragmentation that often characterizes interdisciplinary research.

The second pillar is conceptual modeling. Drawing on the error budget framework articulated by Dasari (2025), the study constructs an abstract model of reliability governance that maps service-level objectives onto resource-level behaviors through adaptive feedback loops. This model is not expressed in mathematical equations, in accordance with the constraints of this article, but in descriptive theoretical terms that articulate the relationships between workload variability, scheduling decisions, model performance, and user experience. The goal is to provide a language in which engineers, managers, and researchers can reason about reliability trade-offs in LLM systems without reducing them to oversimplified metrics.

A third methodological component is socio-technical interpretation. Large language model infrastructures are not purely technical artifacts; they are embedded within organizational processes, regulatory environments, and user communities. The literature on digital servitization and cloud-based service systems highlights how technical architectures co-evolve with business models and governance structures (Poniszewska-Maranda et al., 2020; Kryvinska and Bickel, 2020). By incorporating these perspectives, the study interprets error budgets not only as engineering tools but as institutional mechanisms that mediate accountability and risk distribution across stakeholders, a theme that is central to Dasari’s (2025) treatment of SRE practices.

The methodological limitations of this approach must also be acknowledged. Because the study relies on interpretive synthesis rather than empirical experimentation, its conclusions are necessarily contingent on the quality and representativeness of the existing literature. While the selected references cover a wide range of technical and theoretical perspectives, they cannot capture the full diversity of industrial practices or emerging proprietary systems. Moreover, the rapidly evolving nature of LLM technologies means that any static model risks obsolescence. Nevertheless, by grounding its analysis in enduring principles of reliability engineering and systems theory, as articulated by Dasari (2025) and others, the study aims to provide insights that remain relevant even as specific technologies change.

3. Results

The analytical synthesis conducted in this study yields a set of interrelated findings that illuminate how error budget management can be adapted to the operational realities of large-scale LLM serving infrastructures. The first major result is the identification of error budgets as dynamic coordination mechanisms rather than static tolerances. In traditional web services, an error budget might specify that a service can fail for a certain number of minutes per month without violating its service-level objective. Dasari (2025) showed how this quantification enables teams to allocate risk and prioritize engineering work. In LLM-serving systems, however, the notion of failure must be expanded to include not only outages but also degraded response quality, excessive latency, and unsustainable cost. By interpreting error budgets as multidimensional envelopes, operators can make nuanced trade-offs between these factors, a necessity in environments where workload characteristics shift rapidly (Zhong et al., 2024; Wang et al., 2025).

A second result concerns the role of distributed inference architectures in shaping reliability outcomes. Systems such as Orca and DistServe demonstrate that separating the stages of inference and distributing them across heterogeneous resources can dramatically improve throughput and latency (Yu et al., 2022; Zhong et al., 2024). From a reliability perspective, this disaggregation creates new failure modes, as bottlenecks or misallocations in one stage can propagate through the pipeline. However, when governed by an error budget framework, these architectures also provide more levers for corrective action. If decoding latency begins to exceed acceptable bounds, for example, schedulers can reallocate resources or adjust batching strategies to bring the system back within its reliability envelope, an operationalization of the adaptive control envisioned by Dasari (2025).

The third result highlights the importance of evaluation and benchmarking in reliability governance. Tools such as MT-Bench and Chatbot Arena provide comparative assessments of model performance across tasks and contexts (Zheng et al., 2024). While these benchmarks are often used to guide model selection and training, they also have implications for error budget management. A model with higher average performance but greater variance may consume error budget more rapidly than a slightly less capable but more stable model. This insight reinforces Dasari's (2025) argument that reliability engineering requires continuous measurement and feedback, extending it into the realm of model quality assessment.

Finally, the synthesis reveals that error budgets can serve as a unifying language between offline and online components of LLM ecosystems. Offline training, fine-tuning, and evaluation determine the statistical properties of a model, while online serving systems manage the real-time execution of inference requests. By expressing both domains in terms of their contribution to reliability risk, organizations can align their investments in model development and infrastructure with their service-level commitments, a governance alignment that echoes the socio-technical integration advocated by Kryvinska and Bickel (2020) and formalized in SRE practice by Dasari (2025).

4. Discussion

The results presented above invite a deeper theoretical and practical examination of how reliability engineering must evolve in response to the rise of AI-native services. At a theoretical level, the reconceptualization of error budgets as dynamic, multidimensional constructs challenges the traditional view of reliability as a binary or scalar property. In classical systems engineering, reliability was often treated as the probability that a system would function correctly over a given period. SRE expanded this view by embedding reliability within organizational processes and decision-making structures, as Dasari (2025) so clearly articulated. In the context of LLM serving, this evolution continues as reliability becomes a negotiated outcome between stochastic computation, adaptive scheduling, and human expectations.

One of the central debates in the literature concerns the trade-off between performance optimization and operational stability. Distributed inference systems such as BlendServe and Echo are designed to maximize throughput and minimize cost through sophisticated batching and co-scheduling strategies (Zhao et al., 2024; Wang et al., 2025). Critics argue that such complexity increases the risk of cascading failures, as tightly coupled components may amplify small perturbations. Proponents counter that without such optimization, the economic viability of large models would be compromised. An error budget framework provides a way to mediate this debate by making explicit how much instability is acceptable in pursuit of performance gains, a normative stance that aligns with Dasari's (2025) emphasis on disciplined risk-taking.

Another scholarly debate revolves around the meaning of quality in generative AI systems. Unlike traditional services, where correctness can often be objectively verified, LLM outputs are evaluated through human judgment and contextual appropriateness (Zheng et al., 2024). This subjectivity complicates the integration of quality into reliability metrics. However, by treating quality deviations as a form of error that consumes error budget, operators can incorporate human-centered evaluation into their reliability governance. This move echoes the broader trend in digital transformation research toward integrating user experience and trust into technical performance metrics (Poniszewska-Maranda et al., 2020; Kryvinska and Bickel, 2020).

The limitations of the proposed framework must also be acknowledged. The reliance on interpretive synthesis means that the model may oversimplify certain technical nuances or organizational dynamics. Moreover, the rapid pace of innovation in LLM architectures, as evidenced by the continual release of new technical reports (Wake et al., 2024; Yang et al., 2025), means that specific operational strategies may soon be outdated. Nevertheless, the underlying principles of error budget management, as articulated by Dasari (2025), provide a stable foundation on which future adaptations can be built.

Looking forward, several avenues for future research emerge. One promising direction is the integration of causal modeling and Bayesian networks into reliability management, allowing operators to infer the root causes of reliability degradation and allocate error budgets more intelligently (Kitson et al., 2023; Ankan and Textor, 2023). Another is the exploration of active inference and equilibrium models in the computing continuum, which could enable self-regulating infrastructures that dynamically balance reliability and performance (Sedlak et al., 2024). These directions suggest that reliability engineering for AI systems is poised to become an even more interdisciplinary and theoretically rich field.

5. Conclusion

This article has developed a comprehensive framework for understanding and managing reliability in large-scale language model serving infrastructures through the lens of error budget governance. By grounding its analysis in the principles articulated by Dasari (2025) and integrating insights from distributed systems, model evaluation, and digital infrastructure research, the study demonstrates that error budgets can serve as the central coordination mechanism for AI-native services. In an era where generative models mediate ever more aspects of social and economic life, the ability to quantify, allocate, and govern unreliability is not merely a technical concern but a cornerstone of institutional trust. The reconceptualization of error budgets as dynamic, multidimensional envelopes offers a path forward for organizations seeking to balance innovation, performance, and stability in the age of artificial intelligence.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need. arXiv:1706.03762.
2. Kryvinska, N., and Bickel, L. (2020). Scenario-Based analysis of IT enterprises servitization as a part of digital transformation of modern economy. Applied Sciences, 10, 1076.
3. Dasari, H. (2025). Site reliability engineering practices for error budget management in large-scale systems. International Journal of Applied Mathematics, 38(5s), 991–1001.
4. Yu, G. I., Jeong, J. S., Kim, G. W., Kim, S., and Chun, B. G. (2022). Orca: A distributed serving system for transformer-based generative models. Proceedings of the USENIX Symposium on Operating Systems Design and Implementation.
5. Zhong, Y., Liu, S., Chen, J., Hu, J., Zhu, Y., Liu, X., Jin, X., and Zhang, H. (2024). Distserve: Disaggregating prefill and decoding for goodput-optimized large language model serving. arXiv:2401.09670.

6. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971.

7. Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36.

8. Wang, Z., Li, S., Li, X., Zhou, Y., Zhang, Z., Wang, Z., Gu, R., Tian, C., Yang, K., and Zhong, S. (2025). Echo: Efficient co-scheduling of hybrid online-offline tasks for large language model serving. arXiv:2504.03651.

9. Zhao, Y., Yang, S., Zhu, K., Zheng, L., Kasikci, B., Zhou, Y., Xing, J., and Stoica, I. (2024). BlendServe: Optimizing offline inference for auto-regressive large models with resource-aware batching. arXiv:2411.16102.

10. Wake, A., Chen, B., Lv, C. X., Li, C., Huang, C., Cai, C., Zheng, C., Cooper, D., Zhou, F., Hu, F., Wang, G., Ji, H., Qiu, H., Zhu, J., Tian, J., Su, K., Zhang, L., Li, L., Song, M., Li, M., Liu, P., Hu, Q., Wang, S., Zhou, S., Yang, S., Li, S., Zhu, T., Xie, W., He, X., Chen, X., Hu, X., Ren, X., Niu, X., Li, Y., Zhao, Y., Luo, Y., Xu, Y., Sha, Y., Yan, Z., Liu, Z., Zhang, Z., and Dai, Z. (2024). Yi-Lightning Technical Report. arXiv:2412.01253.

11. Yang, A., et al. (2025). Qwen2.5 Technical Report. arXiv:2412.15115.

12. Canizo, M., Conde, A., Charramendieta, S., Minon, R., Cid-Fuentes, R. G., and Onieva, E. (2019). Implementation of a large-scale platform for cyber-physical system real-time monitoring. IEEE Access, 7, 52455–52466.

13. Habeeb, R. A. A., Nasaruddin, F., Gani, A., Hashem, I. A. T., Ahmed, E., and Imran, M. (2019). Real-time big data processing for anomaly detection: A survey. International Journal of Information Management, 45, 289–307.

14. Chen, Y., Iyer, S., Liu, X., Milojicic, D., and Sahai, A. (2007). SLA decomposition: Translating service level objectives to system level thresholds. Proceedings of the International Conference on Autonomic Computing.

15. Poniszewska-Maranda, A., Matusiak, R., Kryvinska, N., and Yasar, A. U. H. (2020). A real-time service system in the cloud. Journal of Ambient Intelligence and Humanized Computing, 11, 961–977.

16. Sedlak, B., Casamayor Pujol, V., Donta, P. K., and Dustdar, S. (2024). Markov blanket composition of SLOs. Proceedings of the IEEE International Conference on Edge Computing and Communications.

17. Sedlak, B., Casamayor Pujol, V., Donta, P. K., and Dustdar, S. (2024). Equilibrium in the computing continuum through active inference. Future Generation Computer Systems.

18. Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., and Chobtham, K. (2023). A survey of Bayesian network structure learning. Artificial Intelligence Review, 56, 8721–8814.

19. Ankan, A., and Textor, J. (2023). pgmpy: A Python toolkit for Bayesian networks.