International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

Journal Website: http://sciencebring.com/index.php/ijasr

Research Article

Adaptive Reliability and Error Budget Governance for Large-Scale Language Model Inference Systems

Submission Date: January 01, 2026, **Accepted Date:** January 15, 2026,
Published Date: January 31, 2026

Dr. Stefan Baumann
Department of Computer Science, University of Zurich, Switzerland

ABSTRACT

The rapid industrialization of large-scale language model inference has transformed the operational landscape of digital services, pushing traditional reliability engineering paradigms into unprecedented complexity. Contemporary artificial intelligence platforms are no longer monolithic computational services; they are distributed, heterogeneous, and deeply interwoven with user experience, data gravity, and real-time quality-of-service constraints. This article develops a comprehensive theoretical and empirical framework for understanding how site reliability engineering practices, particularly error budget management, must evolve to remain effective in the era of large-scale language model serving. Building on recent advances in large language model systems engineering, cloud-native service-level objective orchestration, and error-budget-driven reliability governance, the paper articulates a unified perspective that integrates infrastructural elasticity, memory management, and inference routing into a single reliability economics model.

The study is grounded in a detailed synthesis of recent research on language model inference pipelines, including GPU and CPU offloading strategies, scheduling architectures, long-context processing, and streaming quality of experience. These technical developments are examined through the lens of site reliability engineering, where error budgets serve not merely as operational constraints but as strategic instruments for balancing innovation velocity with user trust. A central conceptual contribution of this work is the reinterpretation of error budgets as multidimensional governance constructs that encompass latency, availability, accuracy, fairness, and contextual coherence rather than only uptime. This conceptualization is directly aligned with contemporary reliability thinking in large-scale systems,

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

particularly the error budget management paradigm articulated in modern site reliability engineering literature (Dasari, 2025).

Methodologically, the paper employs a qualitative systems synthesis approach, integrating architectural analyses, operational theory, and service-level objective modeling to derive an interpretive framework for adaptive reliability. Rather than relying on numerical simulation, the study develops descriptive and inferential arguments that map how inference pipelines, scheduling strategies, and routing mechanisms collectively determine the consumption and replenishment of error budgets. The results demonstrate that advanced scheduling and memory management techniques can be interpreted as implicit reliability controls that redistribute error budget expenditure across time, users, and workloads.

The discussion extends these findings into a broader theoretical debate about the future of reliability engineering in AI-driven infrastructures. It argues that traditional binary notions of failure are inadequate for generative systems whose outputs are probabilistic, contextual, and socially embedded. By positioning error budgets as socio-technical contracts between service providers and users, the article offers a foundation for reliability governance that is both technically rigorous and ethically responsive. The paper concludes by outlining implications for cloud-native architecture, regulatory compliance, and the design of next-generation service-level objectives in AI platforms.

Keywords

Site reliability engineering, error budget management, large language models, service-level objectives, inference scheduling, cloud-native systems

Introduction

The evolution of digital infrastructure over the last two decades has been defined by an accelerating interplay between computational scale, user expectations, and organizational accountability. Early web services were primarily concerned with basic availability and throughput, but the rise of cloud computing and software-as-a-service introduced a more nuanced vocabulary of reliability that included latency, consistency, and cost-efficiency (Lakshminarayanan et al., 2013). As digital services became embedded in education, healthcare, finance, and governance, reliability was no longer a purely technical metric but a socio-economic guarantee that shaped trust and adoption (Benta et al., 2015; Moreno and Mayer, 2007). This transformation laid the foundation for site reliability engineering as a discipline that integrates software engineering, operations, and risk management into a coherent practice of service stewardship.

Within this broader historical trajectory, the emergence of large language models represents a profound inflection point. Unlike traditional transactional systems, language model services are characterized by probabilistic inference, massive parameter spaces, and context-dependent outputs (Kwon et al., 2023). These properties fundamentally challenge classical reliability paradigms that were designed for deterministic or near-deterministic systems. In a language model inference pipeline, a "successful" response may still

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

ISSN-2750-1396

be semantically incorrect, biased, or misaligned with user intent, complicating the very definition of failure (Jaech et al., 2024). Consequently, the reliability of such systems cannot be reduced to uptime or error rates alone; it must encompass qualitative dimensions of output quality and user experience (Liu et al., 2024).

At the same time, the operational complexity of large-scale language model serving has increased dramatically. Contemporary deployments rely on heterogeneous hardware, including GPUs, CPUs, and edge devices, orchestrated through sophisticated scheduling and routing mechanisms (Jiang et al., 2024; Kossmann et al., 2025). Memory management techniques such as paged attention and offloading are essential for supporting long-context inference without exhausting scarce accelerator resources (Kwon et al., 2023; Jiang et al., 2024). These technical innovations, while necessary for scalability, introduce new failure modes, performance variability, and cost trade-offs that must be governed through reliability engineering.

Within this context, the concept of the error budget has emerged as a central organizing principle of modern site reliability engineering. An error budget represents the permissible amount of unreliability that a service can tolerate over a given period, enabling organizations to balance stability with innovation. Dasari (2025) argues that in large-scale systems, error budgets function as both operational metrics and strategic levers, guiding decisions about deployment velocity, architectural change, and incident response. This dual role is particularly salient for language model services, where rapid iteration and experimentation are critical for model improvement, yet failures can have immediate and visible impacts on users.

However, existing formulations of error budget management were largely developed in the context of conventional web services and microservice architectures. They assume relatively clear definitions of success and failure, as well as relatively stable workload patterns. Large language model inference violates these assumptions in multiple ways. Workloads are highly bursty and context-dependent, user expectations are shaped by subjective perceptions of response quality, and the underlying models evolve continuously through retraining and fine-tuning (Li et al., 2024). These dynamics raise a fundamental research question: how can error budget management be adapted to govern the reliability of large-scale language model systems in a way that is both technically sound and socially meaningful?

The literature on large language model serving provides valuable insights into the technical dimensions of this problem. Scheduling architectures such as One Queue Is All You Need seek to mitigate head-of-line blocking and improve throughput fairness (Patke et al., 2024), while phase-splitting approaches like Splitwise optimize energy efficiency and performance (Patel et al., 2023). RouteLLM introduces preference-based routing to allocate requests across models based on cost and quality trade-offs (Ong et al., 2025). These systems implicitly manage reliability by shaping how and when requests are served, yet they are rarely framed in terms of error budgets or service-level objectives.

Similarly, research on service-level objective orchestration in cloud and edge environments has produced languages and schedulers for expressing

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

and enforcing complex reliability goals (Pusztai et al., 2021; Pusztai et al., 2022). These frameworks emphasize elasticity, topology awareness, and cross-site coordination, offering a conceptual toolkit for managing distributed services under uncertainty (Cao, 2023; Cardellini et al., 2018). However, they have not been extensively applied to the unique challenges of generative AI, where the notion of a service-level objective must encompass not only performance but also semantic and ethical dimensions.

The literature on quality of experience in streaming and interactive services further complicates the picture. Liu et al. (2024) demonstrate that user satisfaction with language model-based text streaming depends on subtle factors such as perceived responsiveness, coherence, and continuity. These subjective dimensions of experience are not easily captured by traditional metrics, yet they directly influence the consumption of error budgets insofar as degraded experience can be interpreted as a form of service failure. In healthcare and self-management applications, decentralization and privacy considerations add another layer of reliability complexity, as data locality and regulatory compliance become integral to service quality (Montagna et al., 2023; Alfian et al., 2018).

Against this backdrop, the present study seeks to synthesize these disparate strands of research into a unified theory of adaptive reliability for large-scale language model inference systems. The central thesis is that error budget management, as articulated in contemporary site reliability engineering, provides a powerful but underutilized framework for integrating technical, economic, and experiential dimensions of AI service reliability. By reconceptualizing error budgets as multidimensional constructs that encompass latency, availability, accuracy, and contextual integrity, it becomes possible to align infrastructure design, scheduling policy, and user-facing quality metrics within a single governance regime (Dasari, 2025; Pujol and Dustdar, 2023).

The literature gap addressed by this work lies in the absence of a comprehensive theoretical model that connects the micro-level mechanisms of inference serving with the macro-level goals of reliability governance. While individual studies have examined GPU scheduling, routing, or quality of experience in isolation, few have explored how these elements collectively determine the consumption and replenishment of error budgets over time (Kossmann et al., 2025; Liu et al., 2024). Moreover, the normative implications of reliability management for trust, accountability, and ethical deployment remain under-theorized in the context of generative AI (Jaech et al., 2024; Montagna et al., 2023).

By addressing these gaps, this article aims to contribute not only to the technical literature on language model serving but also to the broader field of socio-technical systems engineering. Reliability, in this view, is not merely a property of machines but a negotiated relationship between providers, users, and institutions. Error budgets become the formal expression of this relationship, encoding how much deviation from ideal service is acceptable in exchange for innovation, affordability, and scalability (Dasari, 2025; Walter et al., 2017).

The remainder of this article develops this argument in depth. The methodology section outlines the interpretive and analytical approach

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

used to integrate insights from systems engineering, reliability theory, and service management. The results section presents a detailed synthesis of how contemporary inference architectures implicitly shape error budget dynamics. The discussion section offers a critical examination of these findings, situating them within ongoing debates about AI governance, cloud-native design, and the future of site reliability engineering. Through this extended analysis, the article seeks to provide a robust intellectual foundation for adaptive reliability management in the age of large-scale language models.

Methodology

The methodological approach of this study is rooted in qualitative systems analysis and interpretive synthesis rather than empirical experimentation or numerical simulation. This choice is motivated by the nature of the research question, which concerns the conceptual and operational integration of error budget management with large-scale language model inference systems. Such integration cannot be meaningfully captured through isolated benchmarks or performance metrics alone, as it involves multi-layered interactions between architectural design, scheduling policy, and user experience (Kossmann et al., 2025; Liu et al., 2024). Instead, the methodology seeks to construct a coherent explanatory framework that accounts for these interactions in a holistic manner.

At the core of the methodology lies a systematic literature synthesis of recent research on site reliability engineering, large language model serving, and service-level objective orchestration. The works of Dasari (2025) provide the foundational lens for understanding error budget management as a strategic and operational construct. This lens is then extended through engagement with studies on GPU scheduling, memory management, and inference routing, which offer detailed accounts of how modern AI services are engineered in practice (Jiang et al., 2024; Kwon et al., 2023; Ong et al., 2025). The synthesis also incorporates research on quality of experience, decentralization, and cloud-edge orchestration to capture the broader socio-technical context in which these systems operate (Montagna et al., 2023; Pusztai et al., 2021).

The methodological process unfolds in three interrelated stages. First, the study identifies key architectural and operational mechanisms that characterize large-scale language model inference. These include request routing, memory management, scheduling, and streaming, as documented in the contemporary systems literature (Patke et al., 2024; Patel et al., 2023). Each mechanism is analyzed in terms of how it influences latency, throughput, and resource utilization, which are traditional dimensions of reliability in distributed systems (Fedushko et al., 2020; Belforte et al., 2010).

Second, these mechanisms are mapped onto the conceptual framework of error budget management. Following Dasari (2025), error budgets are understood not merely as tolerances for downtime but as quantitative expressions of acceptable service deviation across multiple dimensions. The mapping process involves interpreting how architectural choices, such as CPU offloading or queue management, effectively allocate and consume portions of the error budget

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

by trading off performance, cost, and quality. This interpretive mapping is informed by service-level objective theory, which provides formal languages and models for expressing reliability goals in cloud-native environments (Pusztai et al., 2021; Walter et al., 2017).

Third, the study situates these technical and conceptual insights within a broader theoretical discourse on socio-technical reliability. Drawing on research in e-learning, healthcare systems, and real-time data processing, the analysis considers how user trust, regulatory compliance, and ethical considerations shape the meaning of reliability in practice (Benta et al., 2015; Alfian et al., 2018; Montagna et al., 2023). This stage recognizes that error budgets are ultimately negotiated constructs that reflect organizational values and societal expectations as much as engineering constraints.

Throughout this process, the methodology maintains a reflexive stance toward the sources of knowledge it integrates. Rather than treating any single architectural model or scheduling algorithm as definitive, the analysis emphasizes the plurality of approaches and the contingent nature of design decisions. For example, the comparative discussion of One Queue Is All You Need and RouteLLM highlights how different routing and scheduling philosophies imply different distributions of reliability risk across users and workloads (Patke et al., 2024; Ong et al., 2025). These differences are not evaluated solely in terms of technical efficiency but also in terms of how they align with error budget governance.

A critical methodological limitation of this approach is its reliance on secondary sources and theoretical reasoning rather than direct empirical observation. While this allows for a broad and integrative perspective, it also means that the conclusions are necessarily interpretive and subject to revision as new empirical data emerge (Kossmann et al., 2025). However, this limitation is mitigated by the depth and diversity of the literature consulted, which spans multiple domains of systems engineering and service management.

Another limitation lies in the abstraction required to integrate heterogeneous research traditions. Studies of GPU scheduling, for instance, often employ highly specialized metrics and experimental setups that do not easily translate into the language of error budgets and service-level objectives (Jiang et al., 2024; Kwon et al., 2023). The methodology addresses this challenge by focusing on conceptual correspondences rather than direct quantitative equivalence, thereby preserving the richness of each domain while enabling cross-disciplinary synthesis (Dasari, 2025; Pujol and Dustdar, 2023).

Despite these limitations, the chosen methodology is well suited to the exploratory and theory-building aims of the study. By weaving together technical, operational, and socio-technical perspectives, it provides a robust foundation for understanding how error budget management can be reimagined for the age of large-scale language model inference. This integrative approach aligns with contemporary calls for more holistic and ethically informed systems engineering in the deployment of artificial intelligence (Jaech et al., 2024; Montagna et al., 2023).

Results

The synthesis of contemporary literature on large-scale language model inference and site reliability engineering reveals a complex and often implicit

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

relationship between architectural design and error budget dynamics. One of the most salient findings is that many of the techniques developed to improve performance and efficiency in language model serving can be reinterpreted as mechanisms for redistributing reliability risk across time, users, and computational resources (Dasari, 2025; Kossmann et al., 2025).

A first area where this relationship becomes evident is in memory management and resource allocation. Techniques such as paged attention and CPU offloading are designed to alleviate GPU memory constraints and enable the serving of larger models or longer contexts without incurring prohibitive costs (Kwon et al., 2023; Jiang et al., 2024). From a reliability perspective, these techniques effectively trade off latency and throughput for expanded capacity. When inference requests are partially processed on CPUs, response times may increase, but the system avoids outright failures due to memory exhaustion. This trade-off can be understood as a deliberate consumption of a portion of the latency error budget in order to preserve availability and functional correctness, aligning with the strategic use of error budgets described by Dasari (2025).

Scheduling and queuing architectures further illustrate how reliability is managed implicitly through system design. The One Queue Is All You Need approach seeks to eliminate head-of-line blocking by unifying request handling, thereby improving fairness and reducing extreme latency outliers (Patke et al., 2024). This architectural choice has direct implications for error budgets because it smooths the distribution of latency across requests, reducing the likelihood that any single user will experience a severe service degradation. In effect, the system reallocates the error budget more evenly, prioritizing predictability over peak throughput, which is consistent with service-level objective frameworks that emphasize tail latency and user-perceived quality (Walter et al., 2017; Liu et al., 2024).

Routing mechanisms such as RouteLLM introduce another dimension to error budget management by explicitly associating different models with different cost-quality trade-offs (Ong et al., 2025). By learning to route requests based on user preferences and workload characteristics, such systems can allocate more reliable or higher-quality models to critical requests while directing less sensitive workloads to cheaper or less performant models. This stratification effectively creates multiple tiers of error budgets within a single service, allowing organizations to optimize resource utilization without violating overall reliability commitments (Dasari, 2025; Pujol and Dustdar, 2023).

The analysis of streaming and quality-of-experience research further underscores the multidimensional nature of reliability in language model services. Liu et al. (2024) show that user satisfaction depends not only on absolute response times but also on the continuity and coherence of streamed text. Interruptions or inconsistencies, even if brief, can significantly degrade perceived quality. From an error budget perspective, this implies that small technical deviations can have disproportionately large experiential impacts, effectively consuming more of the "perceived reliability" budget than traditional metrics would suggest (Dasari, 2025; Moreno and Mayer, 2007).

Decentralization and edge computing add yet another layer of complexity. In applications such as

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

chronic disease self-management, data locality and privacy are integral to service quality (Montagna et al., 2023; Alfian et al., 2018). Systems that distribute inference across cloudlets and edge devices must balance latency, security, and regulatory compliance. Failures in any of these dimensions can be interpreted as breaches of reliability, even if the core inference engine remains operational. Consequently, error budgets in such contexts must encompass legal and ethical constraints in addition to technical performance (Cardellini et al., 2018; Cao, 2023).

Taken together, these findings indicate that error budgets in large-scale language model systems are not consumed in a single dimension but across a complex space of performance, quality, and trust. Architectural and operational decisions continuously shift how this budget is allocated and expended, often in ways that are not explicitly recognized by engineers or managers (Dasari, 2025; Kossmann et al., 2025). The result is a form of implicit reliability governance, where system behavior reflects a set of unarticulated priorities and trade-offs.

The results also reveal a gap between the sophistication of technical mechanisms and the formality of reliability governance. While advanced schedulers and routers dynamically manage workloads and resources, they are rarely integrated into a coherent error budget framework that makes these trade-offs transparent and accountable (Pusztai et al., 2021; Walter et al., 2017). This disconnect suggests an opportunity for more explicit and adaptive reliability management that aligns system design with organizational and societal goals (Dasari, 2025; Pujol and Dustdar, 2023).

Discussion

The findings of this study invite a rethinking of how reliability is conceptualized and governed in the context of large-scale language model inference. Traditional site reliability engineering emerged in an era of relatively deterministic services, where failures could be clearly identified and measured in terms of downtime or error rates (Belforte et al., 2010; Fedushko et al., 2020). In contrast, generative AI systems operate in a probabilistic and context-sensitive domain, where the boundary between success and failure is often ambiguous (Jaech et al., 2024; Li et al., 2024). This ambiguity challenges the applicability of conventional error budget models and calls for a more nuanced, multidimensional approach.

Dasari (2025) provides a critical starting point by framing error budgets as instruments for balancing reliability and innovation in large-scale systems. However, the present analysis suggests that in the realm of language model services, error budgets must be expanded beyond their traditional scope. Latency, availability, and correctness remain important, but they must be complemented by measures of semantic accuracy, contextual coherence, and user trust. This expansion aligns with research on quality of experience, which demonstrates that subjective perceptions of service quality are central to user satisfaction (Liu et al., 2024; Moreno and Mayer, 2007).

One of the most significant implications of this expanded view is that architectural choices become normative decisions about how reliability is distributed. For example, routing systems that prioritize high-value users or tasks may improve overall efficiency but risk creating inequities in

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

service quality (Ong et al., 2025; Montagna et al., 2023). From an error budget perspective, this raises questions about who is entitled to consume more of the reliability budget and under what conditions. These questions cannot be answered solely through technical optimization; they require engagement with ethical and regulatory considerations (Jaech et al., 2024; Cardellini et al., 2018).

Similarly, memory management and offloading strategies illustrate how reliability trade-offs are embedded in system design. By shifting computation to CPUs or edge devices, systems can avoid catastrophic failures due to resource exhaustion, but they may introduce variability in response times and output quality (Jiang et al., 2024; Kwon et al., 2023). These trade-offs are often justified in terms of cost and scalability, yet they also represent decisions about how much unreliability users should tolerate. Explicitly framing these decisions in terms of error budgets could make them more transparent and accountable (Dasari, 2025; Pujol and Dustdar, 2023).

The discussion also highlights the need for more sophisticated service-level objective frameworks that can accommodate the unique properties of generative AI. Languages such as SLO Script and schedulers like Polaris provide mechanisms for expressing and enforcing complex reliability goals in cloud-native environments (Pusztai et al., 2021; Pusztai et al., 2022). However, these tools were not originally designed to capture semantic or experiential dimensions of service quality. Extending them to include metrics of coherence, bias, or user satisfaction would be a significant but necessary step toward truly adaptive reliability management (Walter et al., 2017; Liu et al., 2024).

Another important theoretical implication concerns the temporal dynamics of error budgets. In language model services, model updates, fine-tuning, and prompt engineering can rapidly change system behavior, effectively resetting or reallocating reliability risks (Li et al., 2024; Jaech et al., 2024). This dynamism suggests that error budgets should be managed not only over fixed time windows but also across model lifecycles and deployment phases. Dasari (2025) hints at this need for temporal flexibility, but the present analysis underscores its urgency in the context of rapidly evolving AI systems.

The limitations of the current study also warrant discussion. As a qualitative synthesis, it cannot provide precise quantitative estimates of how specific architectural choices affect error budget consumption. Future research could address this gap through empirical studies that correlate scheduling policies, routing strategies, and memory management techniques with user-perceived reliability (Kossmann et al., 2025; Liu et al., 2024). Additionally, the ethical and regulatory dimensions of reliability governance in AI systems remain underexplored and require interdisciplinary collaboration (Montagna et al., 2023; Jaech et al., 2024).

Despite these limitations, the theoretical framework developed here offers a valuable lens for understanding and guiding the evolution of site reliability engineering in the age of large-scale language models. By positioning error budgets as multidimensional, adaptive, and socio-technical constructs, it becomes possible to align technical innovation with the broader goals of trust, fairness,

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

and sustainability (Dasari, 2025; Pujol and Dustdar, 2023).

Conclusion

This article has argued that the reliability of large-scale language model inference systems cannot be adequately governed through traditional, one-dimensional notions of uptime and failure. Instead, it requires a reimagined framework of error budget management that integrates technical performance, user experience, and socio-ethical considerations. Building on the principles articulated by Dasari (2025), the study has shown how contemporary architectures for memory management, scheduling, and routing implicitly allocate and consume reliability resources across a complex landscape of trade-offs.

By synthesizing research on inference systems, service-level objectives, and quality of experience, the article has demonstrated that error budgets in generative AI services are inherently multidimensional and dynamic. They encompass not only latency and availability but also semantic coherence, privacy, and trust. Recognizing and formalizing this complexity is essential for developing adaptive and accountable reliability governance in AI-driven infrastructures (Pusztai et al., 2021; Liu et al., 2024).

Ultimately, the future of site reliability engineering in the era of large language models will depend on its ability to bridge the gap between technical optimization and societal expectations. Error budgets, when properly understood and implemented, offer a powerful means of achieving this balance, enabling innovation while safeguarding the integrity of the services on which modern life increasingly depends (Dasari, 2025; Jaech et al., 2024).

References

1. Patke, A., Reddy, D., Jha, S., Qiu, H., Pinto, C., Cui, S., Narayanaswami, C., Kalbarczyk, Z., and Iyer, R. One Queue Is All You Need: Resolving Head-of-Line Blocking in Large Language Model Serving. arXiv:2407.00047.
2. Benta, D., Bologa, G., Dzitac, S., and Dzitac, I. University Level Learning and Teaching via E-Learning Platforms. Procedia Computer Science, 55, 1366–1373.
3. Dasari, H. Site Reliability Engineering Practices for Error Budget Management in Large-Scale Systems. International Journal of Applied Mathematics, 38(5s), 991–1001.
4. Montagna, S., Ferretti, S., Klopfenstein, L. C., Florio, A., and Pengo, M. F. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. Proceedings of the ACM Conference on Information Technology for Social Good, 205–212.
5. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. Proceedings of the 29th Symposium on Operating Systems Principles, 611–626.
6. Pusztai, T., Morichetta, A., Pujol, V. C., Dustdar, S., Nastic, S., Ding, X., Vij, D., and Xiong, Y. SLO Script: A Novel Language for Implementing Complex Cloud-Native Elasticity-Driven SLOs. IEEE ICWS, 21–31.
7. Liu, J., Wu, Z., Chung, J.-W., Lai, F., Lee, M., and Chowdhury, M. Andes: Defining and Enhancing Quality of Experience in LLM-Based Text Streaming Services. arXiv:2404.16283.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

8. Jiang, X., Zhou, Y., Cao, S., Stoica, I., and Yu, M. NEO: Saving GPU Memory Crisis with CPU Offloading for Online LLM Inference. arXiv:2411.01142.
9. Jaech, A., et al. OpenAI o1 System Card. arXiv:2412.16720.
10. Pujol, V. C., and Dustdar, S. Towards a Prime Directive of SLOs. IEEE International Conference on Software Services Engineering, 61–70.
11. Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., and Stoica, I. RouteLLM: Learning to Route LLMs with Preference Data. arXiv:2406.18665.
12. Kossmann, F., Fontaine, B., Khudia, D., Cafarella, M., and Madden, S. Is the GPU Half-Empty or Half-Full? Practical Scheduling Techniques for LLMs. arXiv:2410.17840.
13. Moreno, R., and Mayer, R. Interactive multimodal learning environments. Educational Psychology Review, 19, 309–326.
14. Lakshminarayanan, R., Kumar, B., and Raju, M. Cloud Computing Benefits for Educational Institutions. Second International Conference of the Omani Society for Educational Technology.
15. Patel, P., Choukse, E., Zhang, C., Shah, A., Goiri, I., Maleki, S., and Bianchini, R. Splitwise: Efficient generative LLM inference using phase splitting.
16. Cardellini, V., Galinac Grbac, T., Nardelli, M., Tankovic, N., and Truong, H.-L. QoS-Based Elasticity for Service Chains in Distributed Edge Cloud Environments.
17. Pusztai, T., Nastic, S., Morichetta, A., Pujol, V. C., Raith, P., Dustdar, S., Vij, D., Xiong, Y., and Zhang, Z. Polaris Scheduler: SLO- and Topology-aware Microservices Scheduling at the Edge.
18. Walter, J., Okanovic, D., and Kounev, S. Mapping of Service Level Objectives to Performance Queries. Proceedings of the ACM/SPEC International Conference on Performance Engineering Companion, 197–202.
19. Alfian, G., Syafrudin, M., Ijaz, M. F., Syaekhoni, M. A., Fitriyani, N. L., and Rhee, J. A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. Sensors, 18, 2183.
20. Fedushko, S., Ustyianovych, T., and Gregus, M. Real-time high-load infrastructure transaction status output prediction using operational intelligence and big data technologies. Electronics, 9, 668.
21. Cao, Y. Better Orchestration for SLO-Oriented Cross-site Microservices in Multi-tenant Cloud and Edge Continuum. Proceedings of the International Middleware Conference.
22. Belforte, S., Fisk, I., Flix, J., Hernandez, M., Klem, J., Letts, J., Magini, N., Saiz, P., and Sciaba, A. The commissioning of CMS sites: Improving the site reliability. Journal of Physics, 219, 062047.
23. Li, J., Wang, M., Zheng, Z., and Zhang, M. LooGLE: Can Long-Context Language Models Understand Long Contexts? arXiv:2311.04939.
24. Sedlak, B., Pujol, V. C., Donta, P. K., and Dustdar, S. Designing Reconfigurable Intelligent Systems with Markov Blankets. Service-Oriented Computing.
25. Sedlak, B., Casamayor Pujol, V., Donta, P. K., and Dustdar, S. Controlling Data Gravity and Data Friction: From Metrics to Multidimensional Elasticity Strategies. IEEE SSE.
26. Guan, S., and Boukerche, A. Intelligent Edge-Based Service Provisioning Using Smart Cloudlets, Fog and Mobile Edges. IEEE Network, 36(2), 139–145.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 01 Pages: 179-190
OCLC – 1368736135

27. Wang, C., Wang, L., Chen, H., Yang, Y., and Li, Y. Fault Diagnosis of Train Network Control Management System Based on Dynamic Fault Tree and Bayesian Network. IEEE Access, 9, 2618–2632.

28. Cegan, L., and Filip, P. Advanced web analytics tool for mouse tracking and real-time data processing. IEEE International Scientific Conference on Informatics.