

Digital Twin-Based Performance Analysis of Large Language Model Inference Infrastructure

Taru Kaikkonen

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
hellotaru@missouri.edu

Abstract

Large language model inference infrastructure has become a critical layer of contemporary artificial intelligence systems, mediating between increasingly capable foundation models and the socio-technical environments in which they operate. As deployment scales from single-node experimentation to geographically distributed, multi-tenant serving platforms, the performance of inference infrastructure can no longer be understood solely through isolated latency or throughput measurements. It must be examined as a dynamic property of coupled computational, network, energy, governance, and human-organizational systems. This paper develops a digital twin-based framework for performance analysis of large language model inference infrastructure. The framework treats the digital twin not as a static simulation artifact but as a continuously synchronized analytical instrument that supports capacity planning, fault diagnosis, sustainability accounting, fairness auditing, and policy evaluation. The paper first reviews conceptual foundations of digital twins and their extension to artificial intelligence serving systems. It then examines the architectural layers of inference infrastructure, including request routing, batching, memory management, model parallelism, quantization, and autoscaling. It argues that digital twins can make visible the structural trade-offs among latency, cost, energy consumption, robustness, and equitable access that are otherwise obscured by vendor abstractions and aggregate service metrics. The discussion also considers governance implications, especially when digital twins are used to audit inference systems for reliability, fairness, transparency, and environmental impact. Through case illustrations and cross-domain comparisons, the paper shows how digital twin-based analysis can support more accountable and sustainable large language model deployment. The conclusion proposes a research agenda for validating, standardizing, and democratizing digital twin methods for inference infrastructure.

Keywords

digital twin, large language models, inference infrastructure, performance analysis, sustainability, robustness, fairness, governance.

1. Introduction

Large language models have moved rapidly from research prototypes to infrastructural components of search, software development, education, customer service, scientific assistance, and public administration. This transition has shifted technical attention from training alone toward inference, the continuous process by which models generate responses for users, applications, and autonomous agents. Inference is now a major determinant of cost, latency, energy demand, and user experience. Unlike training, which can be scheduled as a large batch workload, inference is often interactive, bursty, geographically distributed, and subject to strict service expectations. These characteristics make performance analysis of

inference infrastructure a problem of systems design rather than merely model optimization. Foundational work on transformer architectures established the computational patterns that now dominate generative artificial intelligence [5], while later research demonstrated that scaling language models produces emergent few-shot capabilities that encourage broad deployment [6]. The resulting foundation model ecosystem has intensified demand for serving infrastructure that can support many concurrent users, heterogeneous tasks, and variable context lengths [7].

Performance analysis in this setting faces a structural difficulty. The behavior of an inference service emerges from interactions among model architecture, hardware accelerators, memory hierarchies, network topology, request schedulers, software runtimes, energy markets, and organizational policies. A benchmark that reports average tokens per second or median latency captures only a narrow slice of this space. It may obscure tail latency, queueing instability, thermal throttling, carbon intensity variation, and the unequal distribution of service quality across regions or user groups. The problem is compounded by rapid hardware refresh cycles, changing model versions, and opaque commercial serving stacks. Researchers and operators therefore need analytical instruments that can represent the infrastructure as a living system rather than a frozen testbed.

Digital twin concepts offer a promising direction. A digital twin is a virtual representation of a physical or socio-technical asset that is connected to data streams and used for simulation, prediction, and decision support [1]. In manufacturing and industrial systems, digital twins have been applied to monitor machine behavior, predict maintenance needs, and optimize production processes [2]. Systematic reviews have refined the concept by distinguishing digital models, digital shadows, and true digital twins according to the direction and frequency of data exchange [3], [4]. Extending this logic to large language model inference infrastructure requires attention to both computational and institutional dimensions. The twin must represent hardware and software, but it must also represent service-level objectives, energy contracts, access policies, and accountability arrangements.

This paper argues that digital twin-based performance analysis can improve the design, operation, and governance of large language model inference infrastructure. It can help operators anticipate congestion before it occurs, evaluate the carbon consequences of routing decisions, test fairness interventions without disrupting production, and provide regulators with evidence about system behavior under stress. At the same time, digital twins introduce new risks. They can create false confidence if their models drift from reality, centralize sensitive operational data, or privilege metrics that are easy to measure over those that matter for public accountability. The sections that follow examine these tensions through an interdisciplinary systems perspective. They consider architecture, workload modeling, sustainability, robustness, fairness, policy, and future research directions.

2. Conceptual Foundations of Digital Twins for Inference Infrastructure

The digital twin literature has evolved from simple virtual representations toward complex, connected systems that support real-time decision-making. Early formulations emphasized the pairing of a physical asset with a virtual counterpart to mitigate unpredictable or undesirable emergent behavior [1]. Subsequent industrial research broadened the concept to include lifecycle management, simulation, predictive maintenance, and integration with enterprise systems [2]. A key conceptual distinction concerns the degree of connectivity. A digital model is a standalone representation that is not automatically updated by the physical system. A digital shadow receives data from the physical system but does not actively influence it. A

digital twin supports bidirectional interaction, where data from the physical system informs the virtual representation and insights from the virtual representation inform control or operational decisions [3], [4]. For inference infrastructure, the relevant asset is not a single machine but a socio-technical assemblage of accelerators, network links, schedulers, model artifacts, energy sources, and service agreements.

Applying digital twins to large language model inference requires representing multiple layers of abstraction simultaneously. At the hardware layer, the twin must capture accelerator utilization, memory bandwidth, thermal conditions, power draw, and interconnect behavior. At the runtime layer, it must represent request queues, batching strategies, cache states, model parallelism, and quantization configurations. At the service layer, it must model application programming interfaces, rate limits, tenant priorities, and service-level objectives. At the organizational layer, it must encode cost budgets, sustainability targets, audit requirements, and fairness commitments. This multi-layer representation is necessary because inference performance is not determined by any single layer. A change in batch size can improve throughput while increasing tail latency. A change in routing policy can reduce energy use in one region while shifting load to a region with weaker privacy protections. A change in quantization can lower memory pressure while degrading output quality for certain languages or domains.

The digital twin concept also invites comparison with related approaches such as observability platforms, discrete-event simulation, and machine learning-based performance prediction. Observability platforms collect logs, metrics, and traces from production systems, but they are primarily descriptive. They show what happened, not necessarily what could happen under alternative configurations. Discrete-event simulation can explore counterfactual scenarios, but it often relies on simplified assumptions about workload arrivals and service times. Machine learning predictors can estimate latency or throughput from telemetry, but they may lack interpretability and can fail under distribution shift. A digital twin can integrate these approaches by using production observability data to calibrate a simulation model, using machine learning to fill gaps in complex subcomponents, and using simulation to generate synthetic scenarios for stress testing. This integration is especially important for inference infrastructure, where workload patterns are nonstationary and hardware environments change frequently.

Furthermore, digital twins for inference infrastructure must account for the fact that large language models are not ordinary software components. Their behavior depends on prompt content, decoding strategy, context length, and sampling parameters. The same model can exhibit very different computational profiles for a short classification request, a long document summarization task, and a multi-turn dialogue. The digital twin must therefore represent workload semantics as well as workload volume. It must distinguish between prefill and decode phases, between cache hits and misses, and between latency-sensitive interactive sessions and throughput-oriented batch jobs. Without this semantic layer, the twin may optimize for aggregate efficiency while harming the user experiences that matter most. This concern connects to broader calls for evaluating artificial intelligence systems in context rather than through isolated technical benchmarks [7].

3. Architecture of Digital Twin-Enabled Large Language Model Inference Systems

A digital twin-enabled inference infrastructure can be conceptualized as a layered architecture that spans data acquisition, representation, simulation, prediction, decision support, and governance. The data acquisition layer gathers telemetry from hardware counters, operating

system metrics, container orchestration platforms, network monitors, application logs, and energy sources. It must also collect service-level data such as request arrival rates, response times, error rates, and user feedback where appropriate. Because inference systems are often multi-tenant, data acquisition must respect privacy, contractual confidentiality, and security boundaries. The representation layer organizes these data into a model of the infrastructure that can be queried and simulated. This model may include graph representations of network topology, queueing models of request schedulers, state machines of model loading and unloading, and cost models of energy consumption.

The simulation and prediction layer uses the representation to answer operational and strategic questions. For example, it can estimate how a new batching policy would affect tail latency during peak hours, how a regional power outage would propagate through failover mechanisms, or how a change in model quantization would alter energy use and output quality. Machine learning can be used within this layer to predict inference delay from telemetry, as recent work on delay prediction demonstrates [12]. Such predictors can be embedded in the digital twin to provide fast approximations when full simulation is too costly. However, the digital twin should not reduce to a black-box predictor. It should maintain enough structural transparency to explain why a predicted outcome occurs and which assumptions drive the result. This is essential for operators who need to make high-stakes decisions about capacity, reliability, and compliance.

The decision support layer translates simulations and predictions into recommendations, alerts, and automated actions. In a mature digital twin, some decisions may be automated, such as adjusting autoscaling thresholds or rerouting requests to lower-carbon regions. Other decisions may remain human-led, such as changing service-level agreements, deploying a new model version, or altering access policies. The governance layer establishes who can access the twin, what data can be used, how models are validated, and how decisions are audited. Without a governance layer, digital twins can become instruments of opaque control rather than accountable analysis. This concern is particularly acute when inference infrastructure supports public services, healthcare, education, or civic discourse. The architecture must therefore include mechanisms for documentation, versioning, access control, and independent review.

Within the runtime architecture of large language model serving, several technical components are especially important for digital twin representation. Request scheduling determines how incoming queries are ordered, batched, and assigned to model replicas. Memory management determines how key-value caches are allocated and evicted, which strongly affects latency and throughput. Model parallelism distributes layers or tensors across multiple accelerators, introducing communication overhead and synchronization delays. Quantization reduces numerical precision to lower memory and compute costs, but may affect output quality. Serving systems such as those designed for efficient memory management and distributed generative inference illustrate the complexity of these interactions [13], [14]. Research on scaling transformer inference and high-throughput generation further shows that performance depends on the joint configuration of hardware, batching, and parallelism [15], [16]. Quantization techniques such as eight-bit matrix multiplication add another dimension of trade-off between efficiency and fidelity [17]. A digital twin can help operators navigate this multidimensional design space by simulating combinations that would be expensive or risky to test in production.

4. Performance Analysis and Workload Modeling

Performance analysis of large language model inference must begin with a clear articulation of objectives. Latency, throughput, cost, energy, quality, fairness, and robustness are not interchangeable. A system optimized for maximum throughput may produce unacceptable tail latency for interactive users. A system optimized for minimum energy may route requests to regions with weaker data protection. A system optimized for cost may use aggressive quantization that degrades performance for low-resource languages. Digital twins can make these trade-offs explicit by allowing operators to define multiple objectives and observe how different configurations affect each one. They can also support scenario analysis, where hypothetical changes in demand, hardware availability, energy prices, or regulatory constraints are tested before implementation.

Workload modeling is a central challenge. Inference workloads are often described through arrival rates and token counts, but these summary statistics hide important structure. Interactive chat sessions have different temporal dependencies than batch translation jobs. Retrieval-augmented generation introduces variable context lengths and external service calls. Agentic workflows may involve multiple model invocations, tool use, and iterative planning. The digital twin must represent these patterns at a granularity sufficient for the decisions at hand. For capacity planning, coarse-grained arrival models may suffice. For tail latency analysis, detailed models of session interarrival times and context length distributions are necessary. For sustainability analysis, the twin must connect workload scheduling to regional energy mix and cooling efficiency. For fairness analysis, it must track performance across user groups, languages, and task types.

A digital twin can also support performance analysis under uncertainty. Hardware failures, network congestion, software bugs, and demand spikes are inevitable in large-scale infrastructure. The twin can simulate failure scenarios and evaluate the effectiveness of redundancy, failover, and degradation strategies. It can identify single points of failure that are not visible in normal operation. It can test whether autoscaling policies respond quickly enough to sudden load increases. It can evaluate whether graceful degradation preserves essential services or disproportionately harms certain users. This form of analysis is valuable because production systems rarely experience the full range of possible failures except during costly incidents. The twin provides a controlled environment for learning from imagined failures without exposing real users to harm.

Recent advances in machine learning-based delay prediction can enhance digital twin performance analysis. By learning from historical telemetry, predictive models can estimate how changes in request characteristics, queue depth, and resource utilization affect response time [12]. These predictions can be integrated into the twin to provide rapid feedback during interactive scenario exploration. However, predictive models must be carefully validated. They may perform well on average but fail on rare or emerging workloads. They may inherit biases from the data used to train them. They may become stale as hardware and software evolve. The digital twin should therefore include mechanisms for continuous validation, uncertainty quantification, and drift detection. Performance analysis should not treat predictions as ground truth but as one source of evidence among many, to be combined with simulation, expert judgment, and controlled experimentation.

5. Deployment, Scaling, and Sustainability Trade-Offs

Deployment of large language model inference infrastructure involves choices about centralized versus distributed serving, public cloud versus private data centers, and specialized accelerators versus general-purpose hardware. Each choice has performance,

economic, and environmental consequences. Centralized serving can exploit economies of scale and simplify management, but it may increase network latency for distant users and concentrate energy demand in specific regions. Distributed serving can reduce latency and improve resilience, but it complicates model synchronization, data governance, and capacity utilization. Digital twins can model these alternatives by representing network topology, energy availability, cooling constraints, and regulatory boundaries. They can help operators identify configurations that meet service-level objectives while respecting sustainability and equity goals.

Sustainability is a particularly important dimension. Training large neural networks has received significant attention for its carbon footprint [8], and inference adds a continuous operational burden that may exceed training costs over time. Research on energy and policy considerations in natural language processing has highlighted the need to report energy use and consider environmental impacts throughout the model lifecycle [9]. The Green AI movement argues for efficiency metrics that go beyond accuracy and include computational cost [10]. A digital twin can operationalize these concerns by tracking energy consumption at the request, tenant, and service levels. It can simulate the effect of shifting workloads to times or regions with lower carbon intensity. It can evaluate whether caching, batching, or quantization reduces total energy use without shifting burdens elsewhere. It can also expose rebound effects, where efficiency improvements lead to increased demand and higher total consumption.

Scaling inference infrastructure also raises questions about resource allocation and fairness. When capacity is limited, scheduling policies determine who receives fast service and who experiences delays. A digital twin can simulate different allocation rules, such as first-come-first-served, priority-based, fair-share, or market-based mechanisms. It can reveal how these rules affect different user groups, applications, and regions. For example, a policy that prioritizes paying customers may improve revenue but degrade service for public-interest applications. A policy that prioritizes short requests may improve average latency but disadvantage users who need long-form analysis. A policy that routes requests to the nearest data center may reduce latency but increase energy use if that region relies on carbon-intensive power. By making these trade-offs visible, digital twins can support more deliberative infrastructure governance.

From a deployment perspective, digital twins can also assist with capacity planning and lifecycle management. They can estimate when additional accelerators will be needed, when older hardware should be retired, and how model updates will affect resource requirements. They can evaluate the risks of vendor lock-in by simulating migration to alternative serving stacks or hardware platforms. They can assess the resilience of supply chains by modeling scenarios in which certain components are unavailable. These analyses are not purely technical; they intersect with procurement, finance, and public policy. A digital twin that represents only technical metrics will be insufficient for these decisions. It must also represent costs, contracts, regulations, and organizational constraints.

6. Robustness, Fairness, and Governance

Robustness in large language model inference infrastructure refers to the ability to maintain acceptable service under disruption, adversarial input, and unexpected load. Digital twins can contribute to robustness by simulating stress scenarios and identifying weaknesses before they cause outages. They can model the effects of malicious prompts designed to consume excessive resources, such as very long context requests or recursive agent loops. They can

evaluate rate limiting, anomaly detection, and isolation strategies. They can test whether failover mechanisms preserve state and consistency. They can also assess whether degradation policies are transparent and equitable. A system that preserves service for high-priority users while silently dropping requests from others may be robust in a narrow technical sense but unjust in a broader sense.

Fairness is a complex property of inference infrastructure because it emerges from the interaction of model behavior, service quality, and access conditions. Two users may receive the same model output but experience very different latency, reliability, or cost. A digital twin can track performance across dimensions such as geography, language, device type, and tenant category. It can simulate how changes in scheduling or routing affect these dimensions. It can help operators detect disparities that are invisible in aggregate dashboards. However, fairness cannot be reduced to a single metric. It requires normative choices about what constitutes equitable service and who has standing to define it. Digital twins can inform these choices, but they cannot make them legitimate on their own.

Governance of digital twin-based inference analysis must address accountability, transparency, and power. Internal algorithmic auditing frameworks have been proposed to define responsibilities across the artificial intelligence lifecycle [18]. These frameworks emphasize documentation, testing, and oversight, all of which can be supported by digital twins. At the same time, principles alone cannot guarantee ethical artificial intelligence; institutional mechanisms, enforcement, and contestability are necessary [19]. A digital twin that is controlled solely by an infrastructure provider may reproduce existing power asymmetries. A more accountable approach would involve independent researchers, civil society, and regulators in defining scenarios, validating models, and interpreting results. It would also require clear rules about data access, intellectual property, and security. The twin should not become a black box that legitimizes decisions without meaningful scrutiny.

Digital twins also raise privacy and security concerns. To be accurate, they require detailed telemetry about requests, users, and system state. This data can reveal sensitive information about individual behavior, organizational priorities, or national infrastructure. Aggregation and anonymization can reduce some risks, but they may also reduce the twin's usefulness for fairness and robustness analysis. Differential privacy and secure computation are potential technical responses, but they introduce trade-offs in accuracy and complexity. Governance must therefore balance the benefits of detailed modeling against the risks of surveillance and data misuse. This balance is not purely technical; it depends on legal frameworks, ethical norms, and public trust.

7. Case Illustrations and Cross-Domain Comparisons

Digital twin approaches in manufacturing and industrial systems offer useful lessons for inference infrastructure. In manufacturing, digital twins are used to monitor equipment, predict maintenance, and optimize production schedules [2], [3]. These applications share with inference infrastructure a concern for real-time state estimation, resource constraints, and operational continuity. However, inference infrastructure differs in important ways. Manufacturing assets are often spatially bounded and governed by clear ownership. Inference infrastructure is frequently distributed across jurisdictions, owned by multiple providers, and used by diverse populations. The digital twin must therefore represent not only physical processes but also contractual and regulatory relationships. The distinction between digital models, shadows, and twins is relevant here: many current observability systems for inference

are digital shadows, not full twins, because they do not feed insights back into automated control [4].

Online prediction serving systems provide another point of comparison. Systems such as Clipper were designed to deliver low-latency predictions across multiple machine learning frameworks and hardware platforms [11]. They addressed challenges of model versioning, batching, and resource isolation that remain relevant for large language models. However, large language model inference introduces new challenges due to autoregressive decoding, long context windows, and massive memory requirements. Distributed serving systems for generative models have developed specialized scheduling and memory management techniques to address these challenges [13], [14]. Digital twins for current inference infrastructure must therefore integrate lessons from both classical prediction serving and modern generative model serving. They must represent not only request-response latency but also token generation dynamics, cache behavior, and quality degradation under quantization.

Cross-domain comparisons also highlight the potential of digital twins for policy evaluation. In energy systems, digital twins can simulate grid behavior under different demand and supply scenarios. In transportation, they can model traffic flows and infrastructure investments. In healthcare, they can support planning and resource allocation. Inference infrastructure is increasingly similar to these domains because it is a shared resource with public consequences. A digital twin could be used to evaluate the impact of policies such as energy-aware routing mandates, transparency requirements for service quality, or public procurement standards for artificial intelligence services. The Internet of Things context has produced architectural models for digital twins that emphasize interoperability, scalability, and security [20]. These models can inform the design of inference infrastructure twins, especially when serving spans cloud, edge, and device environments.

8. Policy Implications and Future Directions

The policy implications of digital twin-based performance analysis are significant. If digital twins become standard tools for managing inference infrastructure, they may shape what counts as acceptable performance and who gets to decide. Policymakers may require operators to maintain digital twins for auditing purposes, similar to financial reporting or environmental monitoring. Such requirements could improve transparency and accountability, but they could also impose burdens on smaller providers and entrench incumbents with greater resources. A thoughtful policy framework would need to specify minimum standards for data collection, model validation, and independent access while allowing flexibility for innovation. It would also need to address cross-border data flows, since inference infrastructure often serves users across jurisdictions.

Sustainability policy is another area where digital twins can contribute. Governments and regulators are increasingly interested in the energy and carbon impacts of artificial intelligence [8], [10]. Digital twins can provide granular data for carbon accounting, but they can also be used to obscure responsibility through creative allocation. Standardized methodologies are needed to ensure that energy and emissions reporting is comparable across providers. Digital twins could support these methodologies by automating measurement, verification, and scenario analysis. They could also help evaluate the effectiveness of policies such as carbon-aware scheduling or efficiency standards. However, the environmental benefits of such policies depend on broader energy system conditions and demand growth. Efficiency improvements alone may not reduce total emissions if inference demand continues to rise.

Fairness and accountability policy also requires attention. Internal auditing frameworks and ethical principles provide useful guidance, but they must be operationalized in specific infrastructure contexts [18], [19]. Digital twins can support operationalization by providing evidence about service quality, error rates, and resource allocation across groups. They can help regulators and auditors test claims about robustness and fairness. However, the legitimacy of these analyses depends on participation and contestability. Affected communities should have opportunities to shape the scenarios that digital twins simulate and to challenge results that affect their interests. This suggests a participatory approach to digital twin governance, where technical modeling is combined with deliberative processes.

Future research should focus on several directions. First, methods are needed for validating digital twins under nonstationary conditions. Inference workloads, hardware, and model versions change rapidly, so static validation is insufficient. Second, standards are needed for interoperability between digital twins from different providers and domains. Third, privacy-preserving techniques must be developed to enable detailed analysis without compromising sensitive data. Fourth, human-centered evaluation should examine how operators, auditors, and policymakers interpret digital twin outputs and make decisions. Fifth, economic and political analyses should investigate how digital twins affect market structure, competition, and public accountability. These directions require collaboration across computer science, engineering, social science, law, and policy.

9. Conclusion

Digital twin-based performance analysis offers a powerful approach to understanding large language model inference infrastructure as a complex, dynamic, and socially embedded system. By continuously integrating telemetry, simulation, prediction, and governance, digital twins can illuminate trade-offs that are hidden in conventional benchmarks and dashboards. They can support capacity planning, sustainability accounting, robustness testing, fairness auditing, and policy evaluation. They can help operators anticipate failures, reduce environmental harm, and allocate resources more equitably. At the same time, digital twins are not neutral instruments. Their design, data sources, validation methods, and governance arrangements shape what they can see and what they ignore. They can reinforce existing power asymmetries if they are controlled exclusively by infrastructure providers or used to legitimize opaque decisions.

A responsible research agenda for digital twin-based inference analysis must therefore combine technical rigor with institutional reflexivity. It must develop methods for validation, interoperability, privacy preservation, and uncertainty communication. It must create governance structures that enable independent scrutiny and public participation. It must connect performance metrics to normative questions about fairness, sustainability, and accountability. The future of large language model inference will depend not only on faster accelerators and more efficient algorithms but also on the collective ability to understand and govern the infrastructure that sustains them. Digital twins, if designed and governed well, can contribute to that ability by making the invisible workings of inference visible to those who must make decisions about its role in society.

References

1. Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In F.-J. Kahlen, S. Flumerfelt, & A. Alves (Eds.), *Transdisciplinary perspectives on complex systems* (pp. 85–113). Springer.

2. Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics*, 15(4), 2405–2415.
3. Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29, 36–52.
4. Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, 51(11), 1016–1022.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
7. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
8. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network research. *arXiv preprint arXiv:2104.10350*.
9. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
10. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63.
11. Crankshaw, D., Wang, X., Zhou, G., Franklin, M. J., Gonzalez, J. E., & Stoica, I. (2017). Clipper: A low-latency online prediction serving system. *Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation*, 613–627.
12. Zhang, J., & Gao, H. (2026, March). LLM inference delay prediction based on machine learning algorithms. In *2026 International Conference on Generative Artificial Intelligence and Information Security (GAIIS)* (pp. 37-41). IEEE.
13. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626.
14. Yu, G. I., Jeong, J. S., Kim, G. W., Kim, S., & Chun, B. G. (2022). Orca: A distributed serving system for Transformer-based generative models. *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation*, 521–538.

15. Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., Xiao, K., Agrawal, S., & Dean, J. (2023). Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5, 606–624.
16. Sheng, Y., Zheng, L., Yuan, B., Li, Z., Ryabinin, M., Chen, B., Liang, P., Ré, C., Stoica, I., & Zhang, C. (2023). FlexGen: High-throughput generative inference of large language models with a single GPU. *Proceedings of the 40th International Conference on Machine Learning*, 31094–31116.
17. Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35, 30318–30332.
18. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
19. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507.
20. Minerva, R., Lee, G. M., & Crespi, N. (2020). Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models. *Proceedings of the IEEE*, 108(10), 1785–1824.