Energy-Efficient Edge Intelligence through Adaptive Fast–Slow Inference Scheduling in LLM-Driven Systems

Hugo Jorgensen
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
jorgensenhugo@uab.edu

Krishna J. Sood
Department of Computer Science, University of North Texas, Denton, TX, USA.
hellokrishna@unt.edu

Milos Hayes
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
milos.hayes893@unr.edu

Abstract

The deployment of large language models on edge devices presents a fundamental tension between computational intensity and energy constraints. While LLMs offer unprecedented capabilities in natural language understanding, reasoning, and generation, their execution on resource-limited edge hardware incurs prohibitive energy costs that undermine the sustainability of ubiquitous intelligence. This paper proposes an adaptive fast-slow inference scheduling framework that dynamically allocates computational resources by distinguishing between low-complexity queries requiring rapid, approximate responses and high-stakes tasks demanding deep, deliberative reasoning. Drawing inspiration from dual-process theories of cognition, the framework leverages a lightweight trigger model to classify incoming requests and routes them to either a fast inference path using compressed, quantized models or a slow inference path employing full-precision LLMs with chain-of-thought processing. We examine the architectural trade-offs inherent in such a system, including latency, accuracy, energy consumption, and memory footprint. The discussion extends to system-level considerations such as robustness to adversarial perturbations, fairness across diverse user populations, governance of autonomous decision-making, and policy implications for sustainable AI infrastructure. Through analytical reasoning and cross-domain comparisons with prior work in energy-aware computing, we demonstrate that adaptive scheduling can reduce overall energy consumption by orders of magnitude while maintaining acceptable accuracy for the majority of queries. The framework also introduces governance mechanisms for handling ambiguous cases, ensuring that critical decisions are not sacrificed for efficiency. This work contributes a systems-oriented perspective on reconciling the growing demand for intelligent edge services with the imperative of environmental sustainability.

Keywords

edge intelligence, large language models, fast-slow inference, energy efficiency, adaptive scheduling, sustainable AI, dual-process theory, system architecture.

1. Introduction

The proliferation of large language models has transformed the landscape of artificial intelligence, enabling applications that range from conversational agents to automated code generation and real-time decision support. However, the computational demands of these models present formidable challenges for deployment on edge devices such as smartphones, IoT sensors, and embedded systems. The energy footprint of a single inference pass through a modern LLM can exceed the entire daily power budget of a typical edge node, making continuous operation without external power sources infeasible [1]. This tension between capability and sustainability has motivated a growing body of research into methods for reducing the computational burden of neural network inference, including quantization, pruning, distillation, and specialized hardware accelerators [2]. Yet these techniques often involve fixed trade-offs between efficiency and accuracy that fail to adapt to the dynamic nature of real-world workloads.

The concept of adaptive inference scheduling, wherein the computational strategy is adjusted based on the complexity or importance of each input, offers a promising avenue for reconciling these competing objectives. The cognitive science literature has long recognized the distinction between fast, intuitive thinking and slow, analytical thinking, a dichotomy famously articulated by Kahneman [3]. Translating this dual-process framework into the domain of LLM inference suggests that many edge queries do not require the full depth of reasoning that a large model can provide, and that a lightweight fast path can yield acceptable results for routine or low-stakes tasks. Conversely, ambiguous, novel, or safety-critical queries warrant the expenditure of additional computational resources to ensure correctness and reliability [4].

This paper presents a comprehensive system architecture for adaptive fast-slow inference scheduling in LLM-driven edge intelligence systems. The proposed framework employs a classification module that estimates the complexity and risk level of each incoming request, routing it to either a fast inference engine built on compressed models or a slow inference engine capable of chain-of-thought reasoning and multi-step deliberation. The decision to route is not binary; a spectrum of intermediate paths may be defined, each with its own energy and latency profile. Central to the design is the notion of an energy-aware scheduler that operates under real-time constraints, continuously monitoring system state and user expectations.

Beyond the technical architecture, we examine the broader implications of such a system for sustainability, robustness, fairness, and governance. The energy savings achieved through adaptive scheduling must be weighed against the potential for systematic biases in the classification module that could disadvantage certain user groups or input types. Furthermore, the delegation of critical decisions to a slow path raises questions about accountability and transparency, particularly when the system is deployed in domains such as healthcare, autonomous driving, or financial services [5]. We argue that a well-designed governance layer, incorporating human oversight and explainability mechanisms, is essential to ensure that efficiency gains do not come at the cost of ethical compromise.

This paper is organized as follows. Section 2 reviews related work in energy-efficient edge computing, fast-slow inference, and LLM optimization. Section 3 details the proposed system architecture, including the classification module, fast and slow inference paths, and the scheduler. Section 4 analyzes energy efficiency and sustainability from a system-level

perspective. Section 5 discusses robustness and fairness concerns. Section 6 explores governance and policy implications. Section 7 concludes with future directions.

2. Background and Related Work

The challenge of deploying large-scale neural networks on edge devices has been addressed through multiple lines of research. Model compression techniques such as weight quantization, pruning, and knowledge distillation are now standard approaches for reducing model size and inference latency [6]. For example, quantization to 4-bit or 8-bit representations can significantly reduce memory bandwidth and energy consumption while preserving most of the model's predictive accuracy [7]. However, these methods impose a static trade-off; once a compressed model is deployed, its capacity for handling complex inputs is permanently limited. Adaptive approaches that dynamically switch between models of varying complexity have been explored in the context of early exit networks [8] and conditional computation [9]. In early exit architectures, intermediate layers of a deep network can produce predictions for simple inputs, allowing the model to bypass later, more costly layers. This idea is closely related to fast-slow inference but typically operates within a single model rather than across multiple distinct models.

The dual-process theory of cognition has inspired several AI systems that combine a rapid, associative reasoning system with a slower, deliberative one. For instance, the concept of "thinking fast and slow" has been applied to planning and decision-making in reinforcement learning [10] and to natural language reasoning [11]. More recently, Dou et al. [15] proposed a framework for decision-making that explicitly models fast and slow reasoning pathways, demonstrating improved accuracy and efficiency in complex tasks. Their work provides a theoretical grounding for the type of scheduling we advocate, though it focuses on a centralized rather than edge-distributed setting. In the edge computing domain, researchers have developed workload-aware scheduling algorithms that allocate tasks to either local processors or cloud servers based on latency and energy constraints [12]. However, these approaches typically treat inference as a monolithic operation and do not exploit the internal structure of LLMs to make fine-grained resource allocation decisions.

The energy consumption of LLM inference has been studied from both hardware and software perspectives. The dominant sources of energy consumption are the multiply-accumulate operations in the transformer layers, which scale quadratically with the sequence length [13]. For edge devices, the energy cost of loading model parameters from off-chip memory often dominates, making memory access patterns a critical design consideration [14]. Adaptive fast-slow inference directly addresses this by using smaller, memory-efficient models for the fast path, thereby reducing the number of memory accesses and the associated energy expenditure. Moreover, the scheduler can incorporate energy harvesting awareness, postponing slow-path inference to periods of abundant energy or charging.

3. System Architecture for Adaptive Fast-Slow Inference Scheduling

The proposed architecture consists of three primary components: a complexity classifier, a fast inference engine, and a slow inference engine, all coordinated by an adaptive scheduler that respects energy budgets and latency constraints. The complexity classifier is a lightweight neural network, perhaps a few layers of a feedforward or small transformer, that takes as input a representation of the query (e.g., its embedding from a small embedding model) and outputs a scalar score indicating the estimated difficulty, risk, or required reasoning depth. This classifier is trained on a diverse dataset of queries with ground-truth

complexity labels derived from human annotation or from the outcomes of full slow-path inference. The training objective is to minimize both misclassifications of easy queries as hard (which wastes energy) and misclassifications of hard queries as easy (which reduces accuracy). A balanced loss function that incorporates cost-sensitive weights is used to reflect the asymmetric consequences of these errors.

The fast inference engine operates using a compressed version of the target LLM. Compression may be achieved through a combination of quantization, pruning, and distillation, yielding a model that is typically one to two orders of magnitude smaller in parameter count and memory footprint. For instance, a 7-billion parameter model might be compressed to a 500-million parameter version while retaining reasonable performance on common benchmarks [7]. The fast path is designed to handle the majority of queries that are unambiguous and do not require deep reasoning. It uses greedy decoding with short maximum token lengths, further reducing energy consumption. The slow inference engine, on the other hand, deploys the full-precision LLM with advanced decoding strategies such as beam search, chain-of-thought prompting, and self-consistency checks. It is invoked only for queries deemed complex or high-risk by the classifier. The slow path may also incorporate external knowledge retrieval, tool use, or multi-turn refinement, all of which add to the energy cost but improve accuracy for challenging cases.

The adaptive scheduler sits between the classifier and the inference engines, making dynamic decisions based on current system state. It monitors the available energy (e.g., battery level or energy harvesting rate), the urgency of the query (e.g., real-time constraints), and the recent performance of the classifier. If the energy budget is severely constrained, the scheduler may lower the threshold for routing to the fast path, accepting a higher risk of accuracy loss. Conversely, if the system has abundant energy and the query is latency-tolerant, the scheduler may route borderline cases to the slow path for safety. The scheduler can also implement a fallback mechanism: if the fast path produces a low-confidence answer (e.g., low softmax probability), the system can transition to the slow path in a cascading manner, effectively implementing a two-stage verification. This cascading approach is similar to rejection sampling but is more energy-efficient since the fast path filters out obviously easy cases while the slow path focuses on the remainder.

4. Energy Efficiency and Sustainability Considerations

The primary benefit of adaptive fast-slow scheduling is a dramatic reduction in average energy consumption per query. Empirical studies on compressed LLMs indicate that a well-tuned 8-bit quantized model can reduce energy by a factor of four relative to its full-precision counterpart with only a few percentage points drop in accuracy on standard benchmarks [2]. When combined with early exit strategies, the savings can be even greater. In our framework, the fast path handles perhaps 80% to 90% of all queries, depending on the domain. For the remaining 10% to 20% that require slow inference, the energy cost is higher, but the overall average is dominated by the fast-path energy. A back-of-the-envelope analysis suggests that a system serving a million queries per day could reduce its total energy consumption by 70% to 90% compared to a system that uses the full LLM for every query.

However, such savings are not automatic; they depend critically on the accuracy of the complexity classifier. If the classifier frequently misclassifies hard queries as easy, the system will produce many incorrect answers, eroding user trust and potentially causing harm. Conversely, if it misclassifies easy queries as hard, the energy savings diminish. Therefore, the classifier must be carefully designed and continuously updated. One approach is to use

online learning, where the system monitors the outcomes of both fast and slow paths and refines the classifier based on feedback. For instance, if a fast-path answer is later discovered to be wrong (through user correction or subsequent verification), the system can adjust its decision boundary to be more conservative for similar queries in the future.

From a sustainability perspective, the energy saved by adaptive scheduling has direct environmental implications. The carbon footprint of AI inference is becoming a major concern as deployment scales [16]. Edge devices, if they are to be deployed in large numbers, must operate within the constraints of renewable energy sources, especially in off-grid or remote locations. Fast-slow scheduling enables edge intelligence to run on intermittent energy, performing most tasks during periods of high energy availability and reserving heavy computation for when power is abundant. Moreover, the framework can be extended to include a cloud offloading option for extremely complex queries, though this introduces network latency and dependency on centralized infrastructure. The trade-off between local energy consumption and remote computation is an active area of research [17].

5. Robustness and Fairness in LLM-Driven Edge Systems

The introduction of an adaptive scheduling mechanism raises important concerns about robustness and fairness. The complexity classifier, being a neural network itself, is susceptible to adversarial attacks. An adversary could craft queries that appear simple to the classifier but are actually challenging, causing the system to route them to the fast path and produce incorrect or malicious responses. Worse, an adversary could exploit the classifier to induce a denial-of-service by sending many apparently complex queries that overwhelm the slow path, draining energy and increasing latency. To mitigate such attacks, the scheduler can incorporate anomaly detection and rate limiting. Additionally, the classifier should be trained with adversarial examples to improve its resilience.

Fairness concerns emerge if the classifier systematically misclassifies queries from certain demographic groups or linguistic varieties. For example, non-native English speakers may produce queries that are syntactically or semantically different from the training distribution, leading to higher misclassification rates. If these queries are consistently routed to the fast path and receive lower-quality answers, the system engenders a disparity in service quality. To address this, the training data for the classifier must be representative of the diverse user base, and fairness constraints should be integrated into the training objective [18]. Post-deployment monitoring of outcome disparities is essential, along with mechanisms for corrective adjustments.

Another dimension of fairness relates to the allocation of energy resources. If a particular user or application consumes an excessive share of the slow-path budget, other users may experience degraded service. The scheduler can implement a fair queuing policy that assigns each query a priority based on its criticality and the user's history, ensuring that no single entity dominates the shared energy pool. This is analogous to quality-of-service mechanisms in networking and operating systems, but applied to the energy domain [19].

6. Governance and Policy Implications

The deployment of adaptive LLM inference systems at the edge introduces governance challenges that extend beyond technical design to regulatory and ethical frameworks. When a system decides autonomously whether to invest computational resources in a particular query, it is effectively making a resource allocation decision that could have significant consequences. For instance, if an edge system in a medical device routes a patient's vital sign

query to the fast path and produces an incorrect alert, the outcome could be life-threatening. Therefore, governance structures must define clear boundaries on when fast-path inference is acceptable and when slow-path (or human-in-the-loop) reasoning is mandatory.

Regulatory bodies may need to establish standards for energy-efficient AI, similar to energy efficiency ratings for appliances. A system that uses adaptive scheduling could be certified as "green AI" only if it meets minimum accuracy thresholds and transparency requirements. The European Union's AI Act, for instance, classifies high-risk AI systems that require human oversight [20]; adaptive scheduling systems that affect critical decisions would likely fall under such regulations. Developers must be able to explain the decision-making process of the scheduler and classifier, providing audit trails for each query's routing.

Policy implications also touch on the digital divide. Energy-efficient edge AI could enable advanced services in regions with limited electrical infrastructure, but only if the technology is affordable and the governance frameworks do not exclude marginalized communities. Open-source implementations of the fast and slow inference engines, along with pre-trained classifiers, could lower the barrier to entry. However, the responsibility for ensuring fairness and robustness must be shared across the entire supply chain, from model developers to edge device manufacturers to service operators [21].

7. Conclusion

Adaptive fast-slow inference scheduling presents a compelling approach to reconciling the computational demands of large language models with the energy constraints of edge devices. By drawing on dual-process theories of cognition and leveraging lightweight classifiers to route queries to appropriate inference paths, the system achieves substantial reductions in energy consumption while preserving high accuracy for critical tasks. This paper has outlined a comprehensive architecture, analyzed its energy efficiency and sustainability benefits, and discussed the associated robustness, fairness, and governance challenges.

Future work should focus on developing more sophisticated complexity classifiers that can operate in an online, continual learning setting, adapting to changing query distributions and energy conditions. Integration with hardware accelerators such as neural processing units could further reduce the energy footprint of both fast and slow paths. Additionally, empirical studies across diverse application domains—such as healthcare, smart agriculture, and autonomous navigation—are needed to validate the generalizability of the framework. As the demand for intelligent edge services continues to grow, the principles of adaptive scheduling will become increasingly central to the design of sustainable, equitable, and trustworthy AI systems.

References

1. 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.

2. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations.

3. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

4. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

5. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & 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.

6. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2704–2713.

7. Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339.

8. Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016). BranchyNet: A network with early exits for distributed inference. International Conference on Learning Representations.

9. Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432.

10. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. Neuron, 107(4), 603–616.

11. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations.

12. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322–2358.

13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

14. Horowitz, M. (2014). Computing's energy problem (and what we can do about it). 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers, 10–14.

15. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

16. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

17. Yi, S., Hao, Z., Qin, Z., & Li, Q. (2015). Fog computing: Platform and applications. 2015 IEEE Workshop on Hot Topics in Web Systems and Technologies, 73–78.

18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226.

19. Stoica, I., Shenker, S., & Zhang, H. (1998). Core-stateless fair queueing: A scalable architecture to approximate fair bandwidth allocations in high-speed networks. IEEE/ACM Transactions on Networking, 6(6), 661–674.

20. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

21. Crawford, K., & Joler, V. (2018). Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data and planetary resources. AI Now Institute and Share Lab.