

VoltanaLLM: Feedback-Driven Frequency Control and State-Space Routing for Energy-Efficient LLM Serving

Jiahuan Yu

University of Illinois Urbana-Champaign
jiahuan2@illinois.edu

Junfeng Lin*

University of Illinois Urbana-Champaign
junfengl@illinois.edu

Aryan Taneja*

University of Illinois Urbana-Champaign
aryant2@illinois.edu

Minjia Zhang

University of Illinois Urbana-Champaign
minjiaz@illinois.edu

Abstract

Modern Large Language Model (LLM) serving systems increasingly support interactive applications, like real-time chat assistants, code generation tools, and agentic workflows. However, the soaring energy cost of LLM inference presents a growing challenge for sustainable and cost-effective deployment. This paper introduces VoltanaLLM, a system for SLO-aware, energy-efficient LLM serving, built from a control theory perspective. VoltanaLLM co-designs frequency scaling and request routing in emerging prefill/decode disaggregated architectures, leveraging their decoupled execution to enable fine-grained phase-specific control. It consists of a feedback-driven frequency controller that dynamically adapts GPU frequency for prefill and decode phases, and a state-space router that explores routing decisions across frequency-scaled instances to minimize energy under latency constraints. We implement VoltanaLLM in SGLang and evaluate its performance over multiple state-of-the-art LLMs and real-world datasets. The results demonstrate that VoltanaLLM achieves up to 36.3% energy savings while maintaining near-perfect SLO attainment rate, paving the way for sustainable and intelligent LLM serving.

1 Introduction

Large Language Models (LLMs) have emerged as a backbone of modern AI applications, serving billions of requests in applications ranging from chat assistants [33, 52, 60], code generation tools [8, 15], and increasingly, real-time agentic systems [26, 59]. Given LLMs are deployed in production at unprecedented scale, inference energy consumption significantly increases the total ownership of cost (TOC) and has become a major concern for sustainability [34]. Recent studies show that LLM inference can account for over 90% of AI infrastructure utilization for companies [10], significantly pressuring power delivery and thermal limits in data centers. For context, large datacenters are estimated to consume power for up to 2M households [7, 19, 24].

At the same time, LLM applications are increasingly latency-sensitive, especially in interactive and real-time settings. *Service Level Objectives* (SLOs), such as *Time-To-First-Token*

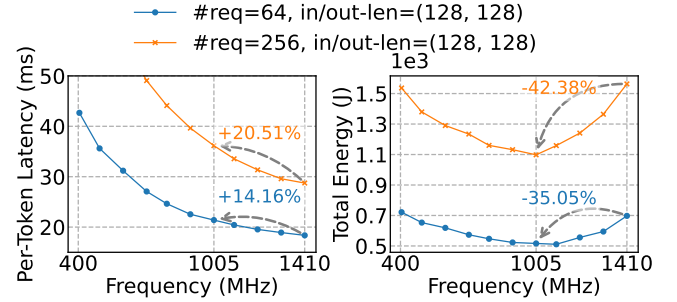


Figure 1. The decreasing per-token latency and the U-shaped energy-frequency curve across varying GPU frequencies. Results come from serving LLaMA-3.1-8B [4] on an A100-40G with SGLang [68].

(TTFT) and *Inter-Token Latency* (ITL), are critical [2, 31, 42], and violating these SLOs degrades not only user experience but also downstream system responsiveness such as agentic pipelines [65]. For example, agent planning and dialog continuation are highly sensitive to ITLs which can lead to degraded performance for many applications [13, 23, 56, 65, 66]. As LLMs are integrated into more interactive pipelines, SLO-aware serving is becoming very desirable. As such, it raises a key challenge: how can we serve LLMs under SLO constraints while reducing the system’s energy footprint?

While the conventional wisdom of dynamic frequency scaling-based energy control [1] holds that reducing frequency directly lowers power, the net effect on energy (time $\times$ power) is not straightforward, because the increase in execution time varies significantly depending on the workload’s characteristics. In particular, our empirical profiling of LLM inference reveals an interesting non-monotonic energy-frequency relationship. As shown in Fig. 1, while reducing GPU frequency from 1410 MHz to 1005 MHz (-28.7%) does increase execution time, the increase is sub-linear. Consequently, the total energy follows a U-shaped curve with respect to GPU frequency. This U-shaped trend suggests that at low frequencies, execution time dominates energy, whereas running at high frequencies, power dominates; in the middle lies an energy sweet point.

*equal contribution

While this observation indicates a clear optimization opportunity, its practical realization for LLM inference faces several challenges. First, many systems leverage token-level batching to amortize overheads and improve GPU utilization. However, batching prefill and decode tokens together makes it difficult to control energy without compromising phase-specific SLO targets (§ 3.1). Second, real-world workloads exhibit strong temporal variations in both arrival rates and request types, which lead to shifting prefill-decode throughput demand, calling for highly adaptive control policies (§ 3.2). Third, energy efficiency also depends on low-level execution. We find that batch size boundaries introduce subtle implication to LLM decode execution. When batch sizes cross specific thresholds (e.g., 128), GPU becomes underutilized, creating “staircases” in *Energy-Per-Output-Token* (Fig. 4). This interaction between execution and energy usage is not well-studied in existing literature (§ 3.3).

To address these challenges, we turn to a new architectural trend in LLM serving systems: prefill-decode (P/D) disaggregation [35, 69] and present VoltanaLLM, a novel system for SLO-aware and energy-optimized LLM inference via a control theory inspired co-design. Specifically, VoltanaLLM introduces two key strategies: (1) a lightweight feedback-driven frequency controller (§ 5.3), which independently adjusts GPU frequency for all instances at fine-grained per-iteration granularity to respond to instantaneous request load without violating SLOs; and (2) a state space navigation-based router (§ 5.4), which steers requests among decode instances by performing a “what-if” analysis in the state spaces of decode instances and navigating a state space to maximize energy efficiency. To support these strategies, VoltanaLLM also introduces a lightweight yet accurate latency predictor (§ 5.5) that estimates TTFT and ITL for candidate frequencies and routing decisions. Together, these techniques allow VoltanaLLM to jointly optimize for SLO attainment, dynamic load adaptation, and energy efficiency.

We implement VoltanaLLM on SGLang [68], a production-grade LLM inference engine. Our extensive evaluations over multiple state-of-the-art LLMs and real-world datasets show that VoltanaLLM achieves up to 36.3% GPU energy savings while preserving TTFT/ITL SLO attainment rates. In summary, we make the following contributions in this paper:

- To the best of our knowledge, we present the first formal formulation of energy efficiency optimization for LLM inference under a P/D disaggregated architecture, while highlighting its connection to control theory and phase-specific scheduling.
- We design and implement VoltanaLLM, an SLO-aware and energy-efficient LLM serving system that incorporates phase-specific feedback-driven frequency control, state space navigation-based router, and lightweight yet accurate latency prediction model.

- We conduct extensive evaluations across state-of-the-art LLMs and various workloads. Our results show that VoltanaLLM achieves up to 36.3% energy savings while preserving SLO attainment rates.

2 Background and Related Work

2.1 LLM Serving Systems

Numerous systems have been proposed to improve LLM serving efficiency, including advanced batching strategies for throughput optimization [11, 64], memory management techniques such as PagedAttention [25], CPU offloading [44, 61], and vAttention [37], and GPU kernel-level optimizations for accelerating attention and decoding [5, 9, 32, 40, 55]. To better utilize available resources and better scheduling, model parallelism and pipelining frameworks have been introduced [5, 28, 36], along with parameter sharing mechanisms to reduce memory and computation overhead [43]. Speculative decoding has emerged as a promising technique to reduce tail latency by predicting likely tokens ahead of time and verifying them in parallel [29, 47]. Additionally, systems like FastServe [58] explore preemptive scheduling policies to reduce job completion time (JCT) in multi-tenant serving environments. While these systems reduce energy consumption as a byproduct of latency or throughput improvements, VoltanaLLM directly targets energy efficiency through frequency and routing-aware scheduling, which has been relatively underexplored despite its critical implications for sustainable deployment.

2.2 Energy-Efficient LLM Serving

Several work have started to explore energy-efficient and power management frameworks for LLM inference [12, 27, 39, 48–50]. For example, DynamoLLM [50] shows benefits of GPU frequency control in LLM inference serving based on request characteristics (predicted input/output tokens) to yield energy savings. It proposes a hierarchical design to efficiently route incoming requests to respective GPU pools that vary by model sharding and frequency while changing frequency on a coarse-grained based on load characteristics. Similarly, μ -Serve [38] shows benefits of dynamic frequency scaling and proposed a model-serving framework to optimize for power consumption by co-serving multiple ML models. These works show early signals towards power-saving opportunities by coarse-grain control in LLM inference but do not explore the significant benefits of fine-grained control, especially in P/D disaggregated serving. Our work focuses on deeply understanding the fine-grained frequency and routing control for varying loads and SLOs, and showing the benefits of using them with P/D disaggregated LLM serving. EcoServe [27] focuses on reducing the operational and embodied carbon emissions throughout the hardware lifecycle. It proposes reusing, allocating and provisioning GPUs and CPUs based on offline/online inference, workload demand,

model-size etc. TAPAS [49] focused on routing requests to specific instances on row/aisle in GPU datacenter racks by exploiting available thermal and power slack to avoid high-power hotspots. Recently, Heron[39] proposed placing GPUs closer to renewable energy sources and adjust workloads for improved efficiency. These works are complementary to our contribution and underscore the importance of sustainable AI. We deeply study frequency-control under varying loads and VoltanaLLM is the first to explore energy-efficient LLM serving under P/D-disaggregated architecture systematically with SLO-aware feedback baked into the system.

2.3 P/D Disaggregation in LLM Inference

To better manage the different compute characteristics of the prefill and decode phases, recent work has proposed prefill-decode (P/D) disaggregation, which assigns the two phases to separate GPU nodes. Systems such as SplitWise [35], Tetri-Infer [18], Llumnix [51] and DistServe [69] show that this separation improves goodput and TTFT and ITL SLO attainment rate by avoiding phase-level contention and enabling distinct execution policies. As such, major open-source LLM inference libraries such as vLLM [25] and SGLang [68] have also added runtime support for this architectural pattern. However, these efforts focus primarily on improving metrics like latency and throughput. To our knowledge, they have neither explored the energy implications of P/D disaggregation nor exploited the new opportunities that this architectural separation enables for energy optimization, such as phase-specific fine-grained frequency control and energy-aware P/D routing policies.

3 Observations and Opportunities

To motivate the design of energy-efficient LLM serving systems, we characterize the performance and energy of representative workloads using LLaMA-3.1-8B [4] served by SGLang on an A100-40GB GPU. We profile key metrics including TTFT, ITL, and energy consumption (using pyNVM [17]). Furthermore, to ground our analysis in realistic scenarios, we analyze the Azure LLM Inference Trace 2024 dataset [50] to identify its temporal and structural patterns.

3.1 Observation 1: U-Shaped Energy-Frequency Relationship in P/D Disaggregated Serving

Our analysis in § 1 reveals a U-shaped energy-frequency curve in LLM inference, implying the existence of a sweet point for energy efficiency. We extend this analysis to P/D-disaggregated architecture. Both phases independently exhibit U-shaped energy-frequency relationships, but their characteristics differ due to different workload characteristics and compute/memory patterns. Fig. 2 presents the energy and latency for prefill and decode instances across different GPU frequencies, using requests with different fixed input/output lengths for each phase. Both phases exhibit distinct U-shaped

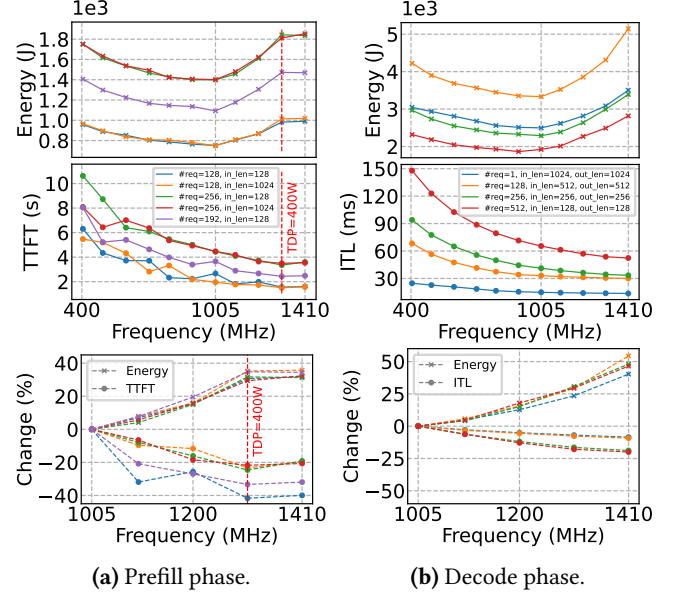


Figure 2. Impact of GPU frequency on energy and latency (TTFT/ITL) for P/D phases. Energy exhibits a characteristic U-shape curve, while latency decreases monotonically with frequency increase. Third row depicts relative percent change in energy/latency from 1005 MHz to 1410 MHz. During prefill, the GPU hits TDP limit (400 W) near 1305 MHz.

energy-frequency curves, with 1005 MHz consistently being the energy-optimal point before a sharp energy upsurge towards the maximum frequency (1410 MHz). However, they also show significant differences. First, the prefill phase is more compute- and power-intensive, hitting the GPU’s Thermal Design Power (TDP) limit at 1305 MHz, whereas the decode can sustain higher frequencies within the same TDP limit. Second, their sensitivities to frequency scaling differ. TTFT responds nearly proportionally to frequency increases, while ITL shows sub-linear diminishing returns: increasing the frequency from 1005 MHz to 1410 MHz incurs ~50% energy penalty for only ~20% reduction in ITL.

The differences between P/D phases can be explained by two underlying factors. First, for most semiconductors, their power (P) increases hyper-linearly with the frequency (f) when f is high (Eq. 1), due to the increased voltage (V) [46]. Second, the prefill phase is predominantly compute-bound and therefore can continuously benefit from high frequency. As formalized in Eq. 2, its execution time is proportional to f^{-1} and $P^{1/(1+\alpha)}$ in theory. In contrast, the decode phase is predominantly memory-bound. Consequently, its performance exhibits sub-linear returns from increased frequency, as execution becomes limited by memory bandwidth rather

than computation. This relationship is modeled in Eq. 3.

$$P \propto V^2 f \propto f^{1+\alpha} \quad (\alpha > 0, \text{ when } f \text{ is high}) \quad (1)$$

$$T^{\text{prefill}} \propto f^{-1} \propto P^{\frac{1}{1+\alpha}} \quad (2)$$

$$T^{\text{decode}} \propto f^{-(1-\beta)} \propto P^{\frac{1-\beta}{1+\alpha}} \quad (0 < \beta \leq 1) \quad (3)$$

This analysis is further supported by Fig. 3, which illustrates that as the batch size increases, decode phase gradually transits from memory-bound to compute-bound, while the ITL benefits from frequency scaling become larger. For LLaMA-3.1-8B served on A100-40G GPU, this turning point ~ 150 .

Insight #1: The non-monotonic, phase-specific energy-frequency relationship motivates dynamic, phase-aware control to minimize energy while preserving latency SLOs.

3.2 Observation 2: Temporal Variation in Prefill-Decode Demand

Real-world LLM workloads exhibits significant temporal variation across both request types and phases. We analyze the Azure LLM Inference Trace 2024 dataset, which contains requests sampled from LLM inference services in Microsoft Azure. The dataset categorizes requests into two types: *conversation* and *code*. As shown in Fig. 5a, the prefill throughput of conversation requests remains relatively stable in a day, while code requests exhibit a significant diurnal pattern, peaking in the afternoon and evening. Furthermore, code requests typically have shorter decode lengths. Consequently, as depicted in Fig. 5b, the overall variation of the decode throughput is much smaller than prefill. For a system handling both request types concurrently, this shifting workload composition leads to significant variation in the overall P/D throughput ratio throughout the day, as shown in Fig. 5c.

This dynamic workload pattern variation presents both a significant challenge and a unique opportunity for energy efficiency. On one hand, a unified frequency scaling strategy is problematic: optimizing GPU frequency solely for one phase can degrade the performance of the other. On the other hand, this pattern allows aggressive energy savings if independently scaling frequency for the two phases, outperforming the policy that treats them uniformly. The P/D disaggregated architecture for LLM serving perfectly matches this independent control policy.

Insight #2: Real-world workloads exhibit significant temporal variations in the P/D throughput ratio and distinct phase-specific characteristics. As a result, one-size-fits-all frequency policies lead to suboptimal energy efficiency.

3.3 Observation 3: Batch Size Boundary-Induced Energy Inefficiency

Modern LLM serving systems batch multiple requests for better throughput [25, 64]. However, our profiling reveals that implicit batch size boundaries plays a crucial role in both energy and latency characteristics, especially during

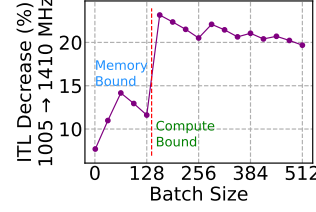


Figure 3. ITL decrease percentage vs. batch size when increasing frequency.

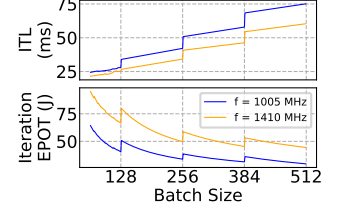


Figure 4. Batch size boundaries effect on ITL and EPOT of decode phase.

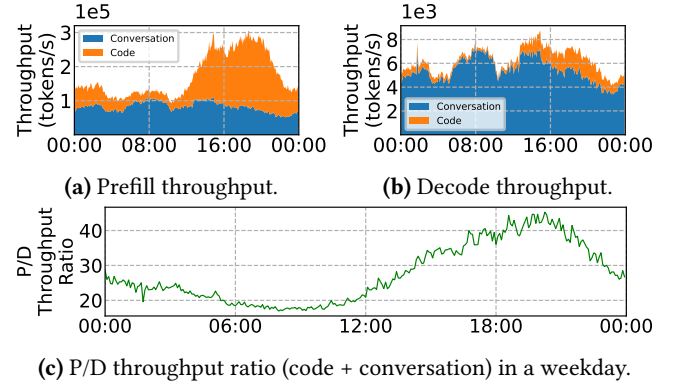


Figure 5. Temporal variations of P/D throughput in Azure LLM Inference Trace 2024 dataset. (Colors are stacked rather than overlapped).

the decode phase. We profile offline serving workloads by varying request numbers for different frequencies. As illustrated in Fig. 4, while larger batch sizes generally lead to lower EPOT and thus higher energy efficiency, this trend is disrupted at specific thresholds. For example, increasing the batch size from 128 to 129 (crossing a "boundary") incurs sudden performance drops, resulting the staircase-like ITL/EPOT relationship to batch size. Similar effect exists for the prefill phase, but gets diminished by relatively larger batch token numbers (see Appendix B).

This phenomenon can be attributed to hardware under-utilization known as the GPU tail effect [63]. A GPU processes large workloads by breaking them into sequential "waves" of thread blocks that run for a fixed processing cycle. When the last wave of a workload cannot fully occupy the GPU's Streaming Multiprocessors (SMs), the hardware is under-utilized. However, this partially filled wave still consumes a full processing cycle, leading to a disproportionate increase in latency and creating a "latency staircase" pattern rather than a smooth, linear improvement. In coupled architectures that handle prefill and decode on the same GPU, techniques like chunked prefill [3] can mitigate this issue by mixing prefill and decode tokens into same batches. However, in a P/D disaggregated architecture, the small number

of batched tokens in the decode instance makes it an unignorable problem. This is a critical inefficiency that is overlooked by popular LLM serving frameworks like SGLang and vLLM.

Insight #3: Certain Batch size boundaries have significant impacts on latency and energy efficiency.

4 Problem Formulation

We formulate the energy optimization for LLM serving system with P/D disaggregation as a constrained, multi-variable control problem. The system consists of N^P prefill instances and N^D decode instances, each with adjustable GPU frequencies (f^P and f^D). Denote M^P/M^D as the system status of an instance. A router dispatches requests to prefill and decode instances. The objective is to minimize total energy consumption while ensuring that SLO^{ttft} and SLO^{itl} are satisfied. We decompose the whole problem into two subproblems: *frequency control* and *request routing*.

For *frequency control*, assuming there are N^{iter} iterations totally for an instance. The optimization is written as:

$$\min_{f_{p,t}^P} \sum_{t=1}^{N^{\text{iter}}} E^P(f_{p,t}^P, M_{p,t}^P), \quad p = 1, \dots, N^P, \quad (4)$$

$$\min_{f_{d,t}^D} \sum_{t=1}^{N^{\text{iter}}} E^D(f_{d,t}^D, M_{d,t}^D), \quad d = 1, \dots, N^D, \quad (5)$$

$$\text{s.t. } T^{\text{ttft}}(f_{p,t}^P, M_{p,t}^P) \leq SLO^{\text{ttft}}, T^{\text{itl}}(f_{d,t}^D, M_{d,t}^D) \leq SLO^{\text{itl}}, \quad (6)$$

where E^P and E^D are iteration energy functions, T^{ttft} and T^{itl} are latency functions. It means to minimize the total energy consumption of each instance, by finding optimal sequence of frequencies $\{f_{p,t}^P\}$ and $\{f_{d,t}^D\}$.

For *request routing*, assuming that there are N requests in total. Denote $I_{p,n}^P \in \{0, 1\}$ as a binary indicator, where $I_{p,n}^P = 1$ means that n -th request is dispatched to p -th prefill instance. $I_{d,n}^D$ is defined similarly. Thus, the routing decision can be written as an assignment optimization problem:

$$\min_{I_{p,n}^P, I_{d,n}^D} \sum_{p=1}^{N^P} E(I_{p,1}^P, \dots, I_{p,n}^P) + \sum_{d=1}^{N^D} E(I_{d,1}^D, \dots, I_{d,n}^D), \quad (7)$$

$$\text{s.t. } \sum_{p=1}^{N^P} I_{p,n}^P = \sum_{d=1}^{N^D} I_{d,n}^D = 1, \quad n = 1, \dots, N, \quad (8)$$

where E is a complicated energy function of an instance which we cannot give analytical form. While these formulations can theoretically yield a global optimum, the problem is NP-hard and even practically intractable. We will simplify them from control theory perspective and design efficient algorithms that approximate the optimum.

5 System Design

Motivated by the insights in § 3 and the problem formulation in § 4, we propose VoltanaLLM, a novel LLM serving system based on P/D disaggregation architecture for SLO-aware and energy-efficient LLM inference. We elaborate the design of VoltanaLLM in the following subsections.

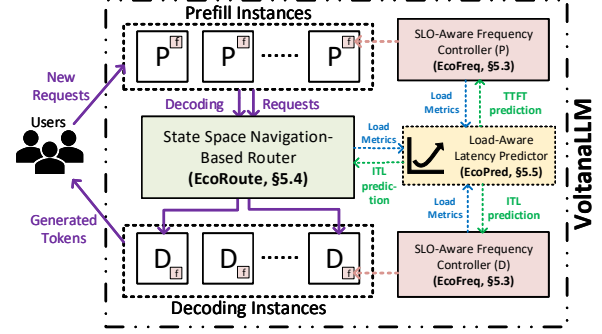


Figure 6. Overall architecture of VoltanaLLM: Incoming Requests are first handled by prefill instances (round-robin) to generate the first output token. EcoRoute then dispatches the decode requests to decode instance, where newly generated tokens are streamed back to the user. Within each instance, EcoFreq runs a separate process for continuous frequency control. Both EcoFreq and EcoRoute query EcoPred for ITL/TTFT predictions.

5.1 Overall Architecture

Fig. 6 illustrates the overall architecture of VoltanaLLM. It features three key novel components:

1. **Phase-Wise SLO-Aware Frequency Controller (EcoFreq):** A feedback-based frequency controller that performs frequency scaling in a lightweight, phase-specific, and fine-grained (per-iteration) manner, to maximize energy efficiency while preserving SLO attainment. Its decisions are guided by real-time load metrics and latency predictions from EcoPred.
2. **State Space Navigation-Based Router (EcoRoute):** A router specifically designed for prefill-to-decode request routing. It mitigates boundary-induced energy inefficiency by analyzing and navigating in the state spaces of decode instances. It also relies on accurate ITL predictions from EcoPred.
3. **Load-Aware Latency Predictor (EcoPred):** A lightweight yet accurate TTFT/ITL predictor. It utilizes collected profiling data and interpretable linear regression to estimate latencies.

5.2 Design Principle: A Control Theory Perspective

In § 4, we pose the energy-efficient frequency scheduling as a constrained optimization problem. While Eq. 4 - Eq. 8 can theoretically yield a global optimum, the problem is combinatorial and NP-hard. In principle, many control strategies are possible, ranging from simple heuristics to complex learning-based policies. However, for real-world LLM serving systems, where simplicity, transparency, and responsiveness are critical, we draw inspirations from classical control policies about feedback system [57] and state space [21, 22], and design *feedback-based* frequency control strategy (§ 5.3) and

state space navigation-based request routing algorithm (§ 5.4). They operate in continuous control loops and follow the simple 3-step principle: gathering system information, doing prediction or “what-if” analysis, and applying new control decisions, including new frequency and routing decision.

The primary advantage of these designs is their robustness and practicality. It obviates the need for potentially unreliable models for predicting request arrival times [20, 30] or output lengths [20, 50], ensuring the system can generalize across dynamic workload conditions and various configurations. Furthermore, by avoiding complex queuing models [30] or learning-based approaches [14, 20], our system remains lightweight and highly interpretable, requiring only a one-time offline profiling pass to calibrate for new served models.

5.3 EcoFreq: Lightweight and Feedback-Driven Phase-Specific Frequency Controller

Although modern GPUs are capable of dynamic frequency scaling, leveraging it for LLM serving is non-trivial. First, the inherent trade-off between energy and latency (§ 3.1) requires a control policy that is both adaptive to real-time load and meticulously managed to prevent SLO violations. Furthermore, temporal variation of P/D demands (§ 3.2) necessitates the phase-specific control policy that allows each phase to operate at its distinct optimal point.

However, a major obstacle is the non-negligible overhead of frequency setting itself. Recent studies such as DynamoLLM [50] report that invoking frequency changes via blocking `nvidia-smi` calls incurs ~50 ms of overhead per iteration, which is expensive in latency-sensitive engine loop. To mitigate this overhead, existing methods propose window-based approach by enlarging the control granularity (e.g., 5 seconds), but such coarse granularity lacks the necessary responsiveness for batch sizes that can fluctuate rapidly, especially for the prefill phase (§ 6.3).

To address these challenges, we introduce EcoFreq, a module that dynamically adjusts the GPU frequency in an SLO-aware, phase-specific, and lightweight manner. Fig. 7 shows how EcoFreq works with engine iterations of both P/D phases. It runs in a separate process to avoid potential overhead, and communicate with the engine via ZeroMQ [6], a high performance IPC framework. Upon scheduling a new batch, the engine sends the relevant metadata and load metrics to the controller. The controller then performs its frequency selection logic (<1 ms). The new frequency is then applied via `pyNVML` (~3 ms), avoiding the high overhead of `nvidia-smi`. Because the entire control cycle (<4 ms) is significantly shorter than the typical model execution time of both prefill (~500 ms) and decode iteration (~50 ms), our approach achieves per-iteration responsiveness with negligible performance overhead. A key difference between P/D phases is how the controller accounts for waiting time. As illustrated in Fig. 7, the TTFT of a prefill request contains two parts: *waiting time* and *execution time* (which can be adjustable via frequency).

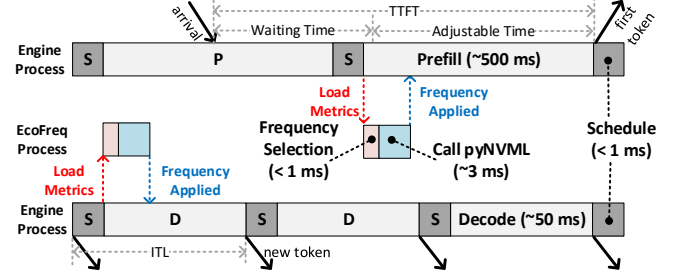


Figure 7. How EcoFreq works with P/D instances: EcoFreq runs in a separate process avoiding potential overhead. It communicates with the engine for load metrics and completes its control cycle (frequency selection and setting).

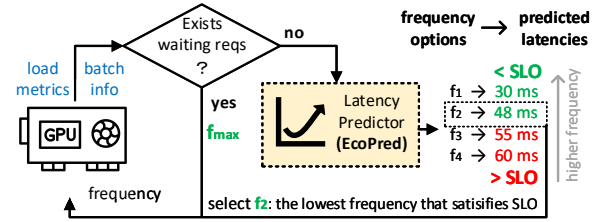


Figure 8. Algorithm of EcoFreq to select frequency. SLO and list of frequency options are given when deploy VoltanaLLM. Inputs are load metrics and batch information from the engine. Output is the selected frequency.

Therefore, to meet the TTFT SLO, EcoFreq subtracts the waiting time from total SLO budget, and use subtraction result as the target of frequency selection. This consideration is not applied for the decode phase, as new tokens are generated auto-regressively without any extra waiting time.

Fig. 8 details the frequency selection algorithm of the EcoFreq. It selects frequency from a list of frequency options, which are adjustable parameters when deploying VoltanaLLM. The design is a feedback-based, stateless decision procedure whose inputs are system load metrics and batch information. The algorithm begins by checking the waiting queue; if a backlog exists, it directly selects the maximum frequency as algorithm output. Otherwise, if the queue is clear, it performs the latency prediction: for all candidates in the list of frequencies, it queries EcoPred using frequency, system load metrics and batch information, and get a list of predicted latencies. Finally, it selects the lowest frequency in the list whose corresponding predicted latency satisfies the SLO target, and uses it as the algorithm output. The whole algorithm is lightweight and takes < 1 ms to finish.

These above designs allow EcoFreq to apply fine-grained and independent control for each phase and instance, which allows the system to achieve maximum energy efficiency while ensuring SLO attainment.

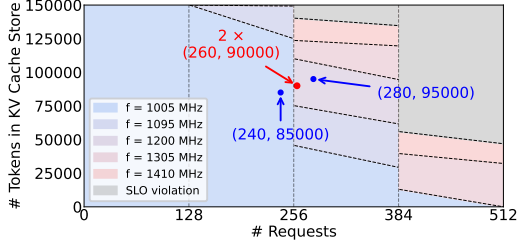


Figure 9. An example of the state space of a decode instance. Number of requests and tokens in KV cache storage are represented by two axes. Colors denote the frequency controlled by EcoFreq. A point in the figure denotes a feasible state of a decode instance, with the frequency control of EcoFreq.

5.4 EcoRoute: State Space Navigation-Based Router

EcoFreq shows how to adaptively control the frequency to improve the energy efficiency for each instance. For systems configured with multiple instances, another potential dimension to save energy is routing. While Eq. 7 – Eq. 8 define the global optimal routing, it is an NP-hard assignment problem and difficult to model accurately and analytically in a dynamic system. Instead of seeking a direct solution, we re-frame the problem from the perspective of state space.

Fig. 9 depicts the state space of a single decode instance. Within this framework, dispatching a request can be conceptualized as a state transition, or moving a point in Fig. 9. Optimization can be transformed into high frequency avoidance in the state space. For example, consider decode instances and 520 requests, where 256 is a batch size boundary described in § 3.3. A “symmetric” router dispatching 260 requests to each forces both of them cross the boundary, requiring a high frequency to meet the SLO. In contrast, an “asymmetric” router can assign fewer (240) requests to one instance, ensuring it remains within the boundary and can thus operate at a low frequency. The second instance then handles more (280) requests, only with cost of a small increase in ITL within the SLO. Consequently, this “asymmetric” router obtains higher energy efficiency.

We design EcoRoute to exploit this opportunity. Fig. 10 illustrates its routing decision process for two decode instances, whose frequencies are initially f_1 and f_2 . There are boundaries of frequencies in the figure. The core idea is to navigate the state space in a boundary-aware manner. It employs a “what-if” analysis in the state space—considers the frequency impact of hypothetically adding the request to all decode instances, and then evaluate new frequencies (f'_1 and f'_2). ①-⑤ are cases in Fig. 10:

1. If no instance across the boundaries ($f_1 = f'_1, f_2 = f'_2$):
 - a. (①) If one instance has a uniquely new low frequency ($f_1 < f_2$), it is the choice.
 - b. (②) If all instances are at the same frequency ($f_1 = f_2$), the router falls back to round-robin policy.

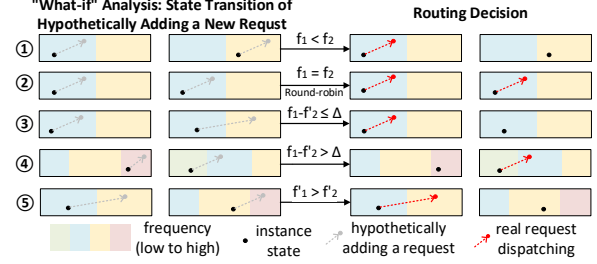


Figure 10. Illustration of routing decision for 2 decode instance by EcoRoute. Different colors denote frequencies from low to high. The point denotes the state of a instance in the state space. The frequency of two instances before routing are f_1 and f_2 , while f'_1 and f'_2 are frequencies after hypothetically adding a request. Left two columns are “what-if” analysis, right two columns are routing decisions.

2. If some instances would cross the boundaries ($f_2 \rightarrow f'_2$), others would not ($f_1 = f'_1$):
 - a. (③) After adding the request, if difference of highest and lowest frequency is below a threshold ($f_1 - f'_2 \leq \Delta$), the request is dispatched to the instance whose frequency would remain unchanged. This rule prioritizes avoiding unnecessary frequency increase.
 - b. (④) Otherwise ($f_1 - f'_2 > \Delta$): the request is dispatched to the instance with the lowest current frequency. This acts as a safeguard against severe load imbalances.
3. (⑤) If all instances would cross the boundaries ($f_1 \rightarrow f'_1, f_2 \rightarrow f'_2$), the request is dispatched to the one that will have the lowest resulting frequency ($f'_1 > f'_2$) after dispatching.

Fig. 11 shows a conceptual example on how EcoRoute obtains energy benefits by contrasting it with a standard round-robin policy. While the round-robin policy maintains a balanced load, often forcing both instances into an inefficient state that requires higher frequencies, EcoRoute uses an asymmetric routing to prevent unnecessary frequency increases. As a result, one instance (indicated by the orange line) spends significantly more time at the energy-efficient low frequency.

While EcoRoute is effective for decode instances, we just apply simple round-robin routing for prefill instances for two reasons. First, the prefill phase lacks the significant batch size boundary effect seen in decode due to its typically large number of batched tokens (§ 3.3 and Appendix B). Second, the prefill phase is largely stateless, with little correlation between the load metrics of consecutive batches (thus shows large fluctuation in Fig. 20). Consequently, the EcoRoute’s state space navigation approach is inapplicable.

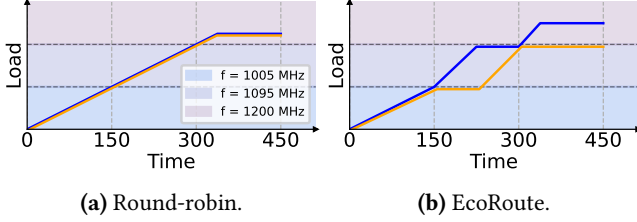


Figure 11. A conceptual example to comparing round-robin and EcoRoute with two decoding instances. Different line colors denote different instances. Area colors denote the targeted frequencies which maintain SLO targets. GPU instance (orange) spends more time at lower frequency with EcoRoute (right) vs. round-robin (left).

5.5 EcoPred: Load-Aware Latency Predictor

The effectiveness of EcoFreq and EcoRoute relies on estimating accurately the impact of GPU frequency on latency under dynamic system loads. VoltanaLLM achieves this by developing a lightweight and interpretable linear regression model to predict TTFT and ITL.

A key insight from our analysis is that TTFT and ITL in LLM inference exhibit highly predictable trends under varying system loads and GPU frequencies. Specifically, both TTFT and ITL show strong correlations with metrics such as batch size, and can be captured accurately via low-overhead models. This predictability enables fast, online control decisions without requiring expensive runtime profiling or black-box machine learning models. As an example, we conduct same profiling as § 3.1. For each request and engine iteration, we measure TTFT and ITL, and collect related system load metrics, including running request number (#reqs, N^{req}), batched token number (#batched tokens, N^{bt}), and number of tokens in KV cache storage (#tokens in KV, N^{kv}). Notably, for the decode phase, $N^{\text{bt}} = N^{\text{req}}$, as only one token’s KV cache is computed in an iteration per request.

Fig. 12a plots the relationship between T^{tft} and N^{bt} at several GPU frequencies, where each point represents a sampled model execution iteration. The data exhibits a strong linear relationship, allowing T^{tft} to be accurately predicted using a simple linear regression model. Fig. 12b illustrates the sampled iterations during the decode phase. Unlike prefill, T^{itl} is related to both N^{req} and N^{kv} , and shows “staircase-like” effects (§ 3.3). Based on the profile results, we can build the following latency prediction models:

$$T^{\text{tft}}(f, N^{\text{bt}}) = a_f^1 N^{\text{bt}} + c_f^1, \quad (9)$$

$$T^{\text{itl}}(f, N^{\text{req}}, N^{\text{kv}}) = a_f^2 N^{\text{req}} + b_f^2 N^{\text{kv}} + c_f^2 \text{ (each tile)}, \quad (10)$$

where a_f^i, b_f^i, c_f^i are coefficients related to f .

Unlike black-box models, our model is interpretable because it directly corresponds to the distinct computational characteristics of the prefill and decode phases. For the prefill

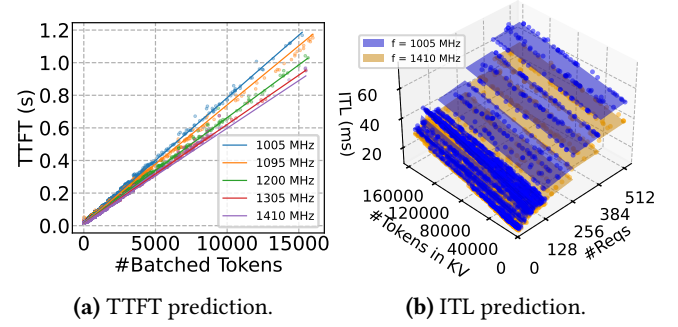


Figure 12. EcoPred visualization. Each point denotes the collected profiling metric of an engine iteration. Lines and planes denote the linear regression predictions.

phase, which is commonly compute-bound, the total computation scales linearly with N^{bt} in theory. This results in a highly linear relationship in Eq. 9 and Fig. 12a. Instead, for the decode phase, N^{req} determines the computation of the MLP layers [54], and N^{kv} determines the computation and memory access of attention layers [9]. Thus, the ITL can be accurately predicted as a linear combination of these two variables. This model also captures the transition of the decode phase from memory-bound to compute-bound shown in Fig. 3. As shown in Fig. 12b, when the batch size (N^{req}) is small, the ITL gap between 1005 MHz and 1410 MHz is also small. As the batch size grows, the workload becomes more compute-intensive, thus the latency gap widens significantly. At runtime, EcoFreq queries EcoPred to estimate latency. Since the models are lightweight, the decision process is fast and incurs negligible overhead (e.g., < 0.1 ms).

6 Evaluation

6.1 Evaluation Methodology

Implementation. We implement VoltanaLLM in Python on top of SGLang [68] (v0.4.7), a widely used LLM inference framework with production-level support for P/D disaggregation. We adopt pyNVML [17] to set frequency with minimal overhead. As described in § 5.3, the frequency control logic runs in a separate process to overlap overhead.

Models and Workloads. We evaluate VoltanaLLM over three models: Ministral-3B [53], LLaMA-3.1-8B [16] and Qwen3-32B [62]. We use ShareGPT [41] and LMSYS-Chat-1M [67] datasets with controlled request arrival rates (quantified by *Request-Per-Second*, RPS) and sampled arrival intervals that follow Poisson distribution. See Appendix A for length information of these two datasets.

Metrics. We evaluate VoltanaLLM using several key performance and efficiency metrics. Our latency evaluation focuses on TTFT and ITL. These metrics align with typical user-facing quality of service expectations. To quantify the effectiveness of our control mechanisms under these constraints, we report the attainment rate, defined as the percentage of

requests that satisfy the SLO thresholds for both TTFT and ITL. Beyond latency, we also examine end-to-end energy consumption, which is directly measured by pyNVML and reported in Joules.

Hardware Testbeds. All experiments are performed on NVIDIA A100-80G SXM4 GPUs connected via NVLink.

6.2 Main Result

We evaluate VoltanaLLM across all three models and all two datasets, using a 2P2D (2 prefill and 2 decode instances) configuration. For Qwen3-32B, 2-way Tensor Parallelism [45] is used for each instance. While DynamoLLM [50] and throttLL’eM [20] are the most relevant systems to our work, they are not included for comparison because of the lack of publicly available code. As such, we construct two baseline configurations based on SGLang: static GPU frequencies of 1005 MHz (the “sweet point” described in § 3.1). and 1410 MHz (maximum frequency, the default strategy of A100 [46]). SGLang uses round-robin as its routing policy. The frequency list of EcoFreq is set to [1005, 1410] MHz. Notably, frequency difference threshold Δ in § 5.4 is set to a large value so that both frequencies in the list can exist for different decode instances. TTFT/ITL SLOs are set to 200/20, 600/60, 1200/120 ms respectively for Ministral-3B, LLaMA-3.1-8B and Qwen3-32B. For each configuration, we report: (1) TTFT SLO attainment rate, (2) ITL SLO attainment rate, and (3) end-to-end energy consumption. Fig. 13 shows our results. We have the following key findings.

First, VoltanaLLM consistently achieves comparable latency SLO attainment rates to SGLang at maximum frequency (1410 MHz) across different models, datasets and request rates. It significantly outperforms the SGLang at fixed low frequency (1005 MHz), which is the energy sweet point but yields substantially lower SLO attainment. This ability to maintain high SLO attainment is attributed to the SLO-aware design of the EcoFreq. It adaptively selects the GPU frequency to ensure the execution time for each batch remains within the SLO budget. Second, while preserving latency SLOs, VoltanaLLM reduces energy consumption (third row in Fig. 13) by up to 36.3% compared to SGLang at 1410 MHz. Its effectiveness is most pronounced at lower request rates where static low frequency is sufficient to maintain SLO targets. In these scenarios, VoltanaLLM operates at low frequency most of the time. Conversely, at higher request rates, it dynamically increases frequency to maintain SLO attainment, resulting in smaller yet still significant energy benefits. These improvements validate that SLO-aware adaptive frequency control and request routing effectively optimizes energy consumption while preserving the quality of service. We also provide the throughput metrics in Appendix C.

6.3 Analysis Results

How does each module of VoltanaLLM bring benefits? To quantify the contribution of each components in

VoltanaLLM, we conduct experiments same to § 6.2. We evaluate two different system configurations: EcoFreq-only and a full version of VoltanaLLM (EcoFreq + EcoRoute), with ShareGPT dataset on LLaMA-3.1-8B.

Fig. 14 presents the results. First, the EcoFreq-only configuration achieves substantial energy savings for both prefill and decode instances compared to the static high-frequency baseline, validating the effectiveness of our SLO-aware frequency control policy. Fig. 15 provides an example of real-time frequency dynamics. It shows how EcoFreq reacts to different request rates: proportion of high-frequency time increases for higher request rates. Furthermore, the full version of VoltanaLLM (EcoFreq + EcoRoute), achieves additional energy reductions compared to EcoFreq-only for the decode instances. As EcoRoute is only applied for decode instances, it shows no benefit for the prefill instances. Fig. 16 compares round-robin and EcoRoute by showing the batch size and frequencies of two decode instances over time. It shows EcoRoute can restrict the batch size of one instance under a specific boundary (256) so it can operate at lower frequency. **How does VoltanaLLM perform under different latency SLO targets?** We follow the setting in § 6.2 to evaluate ShareGPT dataset on LLaMA-3.1-8B, but use three different TTFT/ITL SLO profiles: 400/40 (“low”), 600/60 (“medium”) and 800/80 (“high”). We report same metrics as § 6.2.

The results are presented in Fig. 17. Across all SLO profiles, VoltanaLLM consistently achieves SLO attainment rates comparable to that of the static maximum frequency baseline. Specifically, as the SLO constraints are relaxed, (i.e., moving from “low” to “high” profile), VoltanaLLM’s adaptive frequency control allows for a higher TTFT/ITL in exchange for lower energy consumption, demonstrating the latency-energy trade-off, which provides users the flexibility to define custom phase-specific SLOs to suit their application needs. For instance, they can prioritize lower ITL for responsiveness of chat applications, and higher prefill energy efficiency for better service costs. This also emphasize the necessity of the phase-specific design choice as we discussed in § 3.

Do more frequency levels bring the benefits? In § 6.2, we configure VoltanaLLM with 2-level frequency options. To analyze the potential of finer-grained levels, we now extend it to configurations with more levels. We follow the same settings in § 6.2 but configure the EcoFreq with more granular 5 levels: [1005, 1095, 1200, 1305, 1410] MHz, with frequency difference threshold $\Delta = 150$, and evaluate it with ShareGPT dataset on LLaMA-3.1-8B.

Fig. 18 presents the results of different frequency level granularities. For prefill instance, using more than two frequency levels yields negligible benefits. The decode instance, however, presents a trade-off: more granular levels offer slight energy savings but at the cost of slightly lower SLO attainment rates. This allows users to select an appropriate granularity based on their specific performance-energy priorities for real-world scenarios.

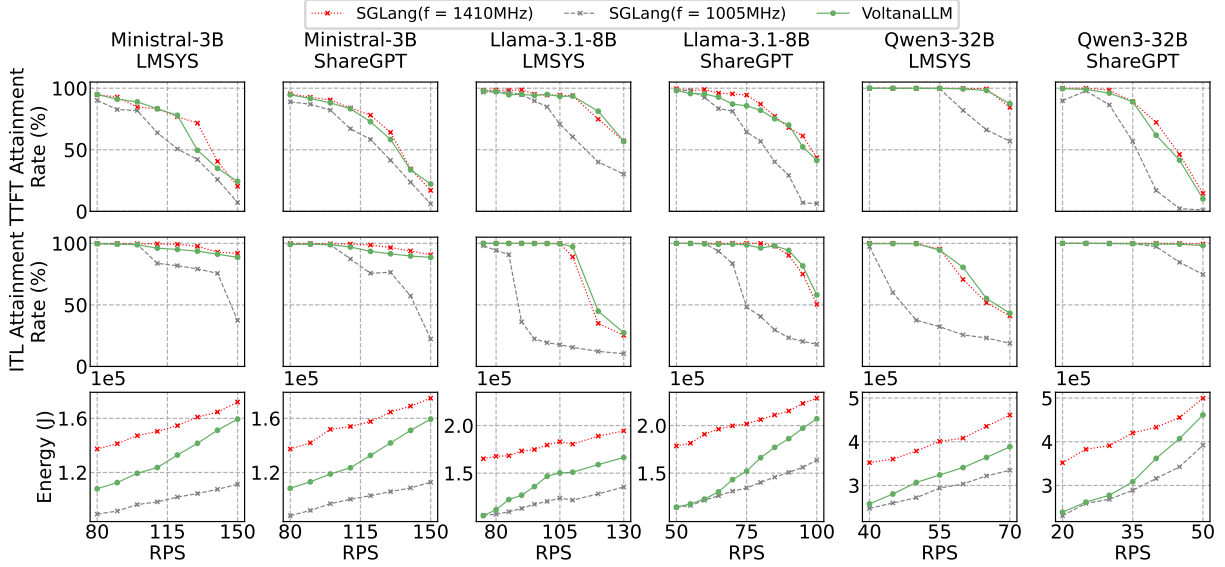


Figure 13. Our main result across various models and datasets. VoltanaLLM consistently achieves comparable SLO attainment rate to SGLang with static highest frequency, while obtaining significant energy saving.

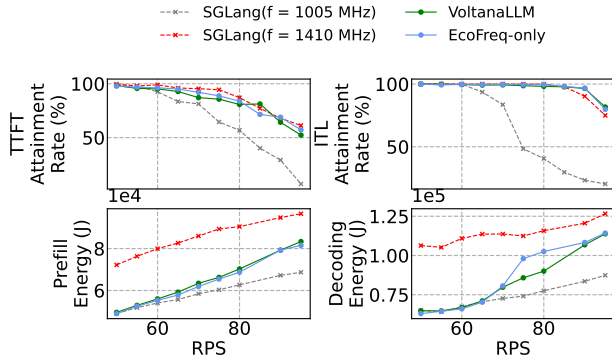


Figure 14. Comparing EcoFreq-only configuration and full version of VoltanaLLM. EcoFreq-only shows energy benefit for both prefill and decode instances, and EcoRoute achieves extra energy benefit for the decode instances.

How important is it to have per-iteration frequency control? Unlike prior approaches such as DynamoLLM [50], which adjusts frequency at window-based intervals (e.g., 5s), EcoFreq achieves per-iteration responsiveness. To verify its benefit, we adapt EcoFreq to operate at various fixed time intervals, evaluating it by ShareGPT dataset on LLaMA-3.1-8B under the 1P1D configuration (EcoRoute has no effect).

The results in Fig. 19 demonstrate that window-based frequency control degrades SLO attainment for both phases. Specifically, per-iteration frequency responsiveness is particularly critical for the prefill phase, where window-based control shows larger degradation. Due to dynamic batching, the total number of tokens processed in a prefill batch (and thus the optimal frequency) can vary significantly from one

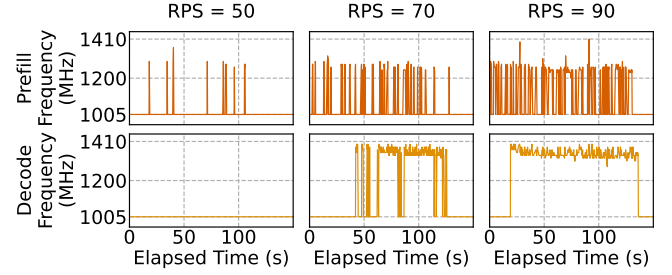


Figure 15. Real-time frequency dynamics of P/D instances managed by EcoFreq. At low request rate, both instances run mainly at low frequency. With higher request rates, the proportion of high-frequency time increases.

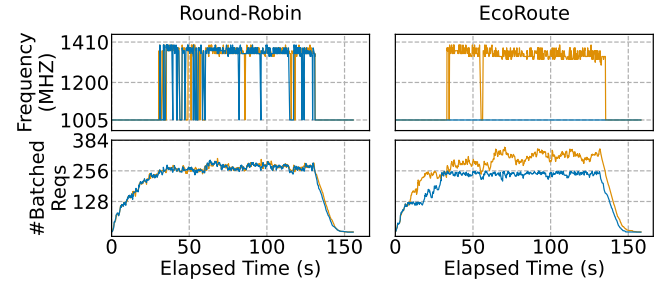


Figure 16. An example to show how EcoRoute dispatches requests differently compared to round-robin policy at a request rate of 75. Each line denotes a decode instance. EcoRoute keeps an instance running under the batch size boundary (256) to enable it runs on lower frequency.

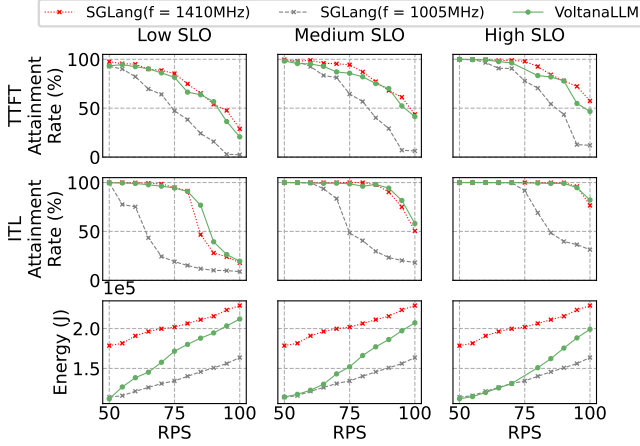


Figure 17. Comparing VoltanaLLM with baselines under different SLO profiles. VoltanaLLM can generalize to different SLO settings, all of them showing high SLO attainment rates and energy benefits.

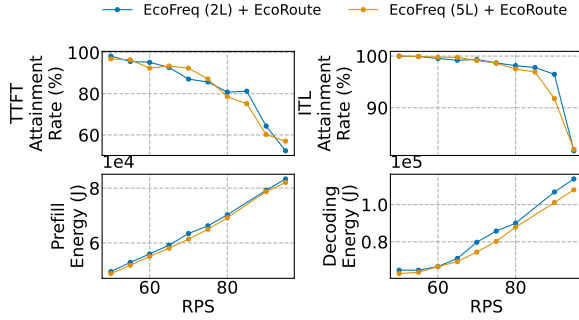


Figure 18. Comparing EcoFreq with different frequency levels. EcoFreq (2L) and EcoFreq (5L) refer to EcoFreq with 2-level and 5-level frequency list respectively.

iteration to the next, as Fig. 20 shows. A window-based approach is too coarse-grained to react to such rapid changes. In contrast, as shown in Fig. 16, the load of the decode instance show smoother change in batched requests, leading to relative insensitivity to window interval settings.

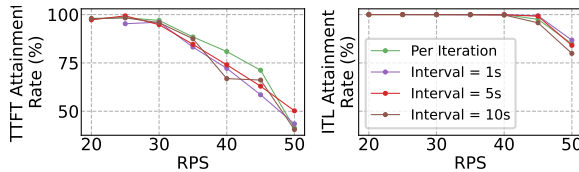


Figure 19. Latency attainment rates with different window interval settings of EcoFreq. Our per-iteration design shows the best responsiveness.

Is EcoPred accurate? We collect the empirical latency data in § 6.2 and compare them with predicted values of EcoPred. We quantify the accuracy using *Mean Absolute Error*

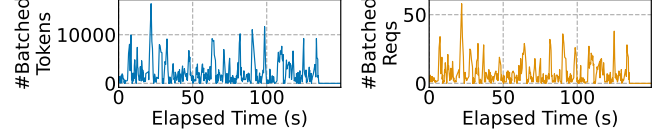


Figure 20. An example to show how prefill batched request number and batched token number fluctuate rapidly.

(MAE). For **TTFT** MAE was 11.02 ms on Mistral-3B, 6.90 ms on LLaMA-3.1-8B, and 14.14 ms on Qwen3-32B. For **ITL** MAE remained consistently small with 2.18 ms, 2.68 ms, and 2.58 ms for the 3B, 8B, and 32B models, respectively which emphasize the performance of EcoPred. We achieve these using a simple and lightweight linear regression, instead of complex ML models.

What happens in varying the P/D throughput demand ratio? As discussed in § 3.2, the temporal variation of P/D throughput demand ratio requires phase-specific frequency control. However, evaluating the whole Azure LLM Inference Trace dataset incurs significant GPU cost and time given it only shows fluctuation on the scale of hours. Hence, we use a synthetic dataset with P/D demand ratio fluctuating in 5 minutes. Fig. 21 shows the evaluation results with LLaMA-3.1-8B using the same configuration in § 6.2.

The results shown in Table 1 confirm the adaptiveness of VoltanaLLM. It can still maintain comparable SLO attainment rates to SGLang at maximum frequency, while obtain significant energy savings. Fig. 21 shows the frequency dynamics. When prefill throughput demand is high, the prefill instance operates at a higher frequency, while the decode instance remains at a low frequency. Conversely, when decode throughput demand is high, their roles reverse, shows highly adaptiveness.

Table 1. Performance and energy under a synthetic workload with varying P/D demand ratios. TSAR/ISAR refers to TTFT/ITL SLO attainment rate.

Metrics	TSAR (%)	ISAR (%)	Energy (J)
VoltanaLLM	96.05	91.74	289409
SGLang (1005 MHz)	77.86	36.78	257768
SGLang (1410 MHz)	97.85	88.92	405340

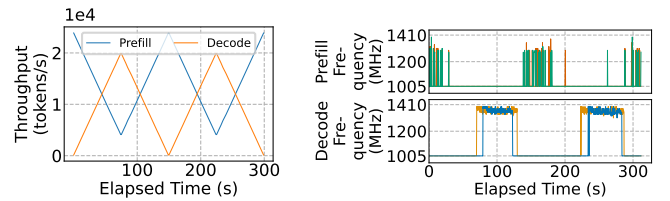


Figure 21. Left : P/D throughput with our synthetic dataset. Right : Frequency dynamics of all instances in VoltanaLLM.

7 Conclusions

In this work, we present VoltanaLLM, a system for energy-efficient and SLO-aware LLM inference built on prefill-decode (P/D) disaggregation. By exploiting the distinct characteristics of prefill and decode phases, VoltanaLLM introduces phase-specific frequency scheduling and state-space navigation-based routing to minimize energy consumption without violating latency SLO guarantees. Through comprehensive evaluation on state-of-the-art LLMs and real-world datasets, VoltanaLLM achieves up to 36.3% energy savings while preserving latency SLO attainment. These results highlight the potential of fine-grained, phase-aware co-design in advancing sustainable LLM serving systems.

8 Acknowledgments

We gratefully acknowledge Alaa Youssef, Vijay Naik, and Yu Chin Fabian Lim from IBM for their valuable discussions on energy-efficient LLM inference. We thank Kartik Ramesh for assistance with conducting preliminary experiments and providing feedback on the manuscript during the early stage of this project. This research was supported by the National Science Foundation (NSF) under Grant No. 2441601. The work utilized the Delta and DeltaAI system at the National Center for Supercomputing Applications (NCSA) and Jetstream2 at Indiana University through allocation CIS240055 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The Delta advanced computing resource is a collaborative effort between the University of Illinois Urbana-Champaign and NCSA, supported by the NSF (award OAC 2005572) and the State of Illinois. UIUC SSAIL Lab is supported by research funding and gift from Google, IBM, and AMD.

References

- [1] 2025. Dynamic frequency scaling. https://en.wikipedia.org/w/index.php?title=Dynamic_frequency_scaling
- [2] Amey Agrawal, Nitin Kedia, Anmol Agarwal, Jayashree Mohan, Nipun Kwatra, Souvik Kundu, Ramachandran Ramjee, and Alexey Tumanov. 2025. On Evaluating Performance of LLM Inference Serving Systems.
- [3] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024. Taming {Throughput-Latency} tradeoff in {LLM} inference with {Sarathi-Serve}. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*. 117–134.
- [4] Meta AI. 2024. Introducing Meta LLaMA-3. <https://ai.meta.com/blog/meta-llama-3/>.
- [5] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. 2022. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–15.
- [6] The ZeroMQ authors. [n. d.]. ZeroMQ. <https://zeromq.org/>
- [7] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd workshop on sustainable computer systems*. 1–7.
- [8] Cursor. 2023. Cursor. <https://www.cursor.com/> Accessed: 2025-05-15.
- [9] Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023).
- [10] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. 2023. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems* 38 (2023), 100857.
- [11] Jiarui Fang, Yang Yu, Chengduo Zhao, and Jie Zhou. 2021. TurboTransformers: an efficient gpu serving system for transformer models. In *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. 389–402.
- [12] Jared Fernandez, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Lucic, and Emma Strubell. 2025. Energy Considerations of Large Language Model Inference and Efficiency Optimizations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.
- [13] Yao Fu, Leyang Xue, Yeqi Huang, Andrei-Octavian Brabete, Dmitrii Ustiugov, Yuvraj Patel, and Luo Mai. 2024. ServerlessLLM: Low-Latency Serverless Inference for Large Language Models. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*.
- [14] Yichao Fu, Siqi Zhu, Runlong Su, Aurick Qiao, Ion Stoica, and Hao Zhang. 2024. Efficient llm scheduling by learning to rank. *Advances in Neural Information Processing Systems* 37 (2024), 59006–59029.
- [15] GitHub. 2021. GitHub Copilot. <https://github.com/features/copilot> Accessed: 2025-05-15.
- [16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [17] Alex Hirschefeld. [n. d.]. nvidia-ml-py: Python Bindings for the NVIDIA Management Library. <https://forums.developer.nvidia.com>
- [18] Cunchen Hu, Heyang Huang, Liangliang Xu, Xusheng Chen, Jiang Xu, Shuang Chen, Hao Feng, Chenxi Wang, Sa Wang, Yungang Bao, et al. 2024. Inference without interference: Disaggregate llm inference for mixed downstream workloads. *arXiv preprint arXiv:2401.11181* (2024).
- [19] International Energy Agency (IEA). 2025. Energy and AI. *International Energy Agency* (2025). <https://www.iea.org/reports/energy-and-ai> Licence: CC BY 4.0.

- [20] Andreas Kosmas Kakolyris, Dimosthenis Masouros, Petros Vavaroutsos, Sotirios Xydis, and Dimitrios Soudris. 2025. throtLL'eM: Predictive GPU Throttling for Energy Efficient LLM Inference Serving. In *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*.
- [21] Rudolf E Kalman. 1960. On the general theory of control systems. In *Proceedings first international conference on automatic control, USSR*.
- [22] Rudolf Emil Kalman et al. 1960. Contributions to the theory of optimal control. *Bol. soc. mat. mexicana* (1960).
- [23] Hao Kang, Qingru Zhang, Han Cai, Weiyan Xu, Tushar Krishna, Yilun Du, and Tsachy Weissman. 2025. Win Fast or Lose Slow: Balancing Speed and Accuracy in Latency-Sensitive Decisions of LLMs. (2025). arXiv:2505.19481
- [24] Emilie Kemene, Bart Valkhof, and Grégoire Greene-Dewasmes. 2024. AI and energy: Will AI help reduce emissions or increase power demand? (22 July 2024). <https://www.weforum.org/stories/2024/07/generative-ai-energy-emissions/> Accessed: 2025-05-15.
- [25] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*.
- [26] LangChain. 2023. LangChain Documentation. <https://python.langchain.com/> Accessed: 2025-05-15.
- [27] Yueying Li, Zhanqiu Hu, Esha Choukse, Rodrigo Fonseca, G Edward Suh, and Udit Gupta. 2025. Ecoserve: Designing carbon-aware ai inference systems. *arXiv preprint arXiv:2502.05043* (2025).
- [28] Zhuohan Li, Lianmin Zheng, Yinmin Zhong, Vincent Liu, Ying Sheng, Xin Jin, Yanping Huang, Zhifeng Chen, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. AlpaServe: Statistical Multiplexing with Model Parallelism for Deep Learning Serving. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*.
- [29] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. 2024. SpecInfer: Accelerating Large Language Model Serving with Tree-based Speculative Inference and Verification (*ASPLOS '24*).
- [30] Michael Mitzenmacher and Rana Shahout. 2025. Queueing, predictions, and llms: Challenges and open problems. *arXiv preprint arXiv:2503.07545* (2025).
- [31] Chengyi Nie, Rodrigo Fonseca, and Zhenhua Liu. 2024. Aladdin: Joint Placement and Scaling for SLO-Aware LLM Serving. arXiv:2405.06856
- [32] NVIDIA. 2021. FasterTransformer. <https://github.com/NVIDIA/FasterTransformer>. GitHub repository.
- [33] OpenAI. 2023. ChatGPT. <https://chat.openai.com/> Accessed: 2025-05-15.
- [34] Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, and Ricardo Bianchini. 2024. Characterizing Power Management Opportunities for LLMs in the Cloud. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS '24)*.
- [35] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*. IEEE.
- [36] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. Efficiently scaling transformer inference. *Proceedings of machine learning and systems* (2023).
- [37] Ramya Prabhu, Ajay Nayak, Jayashree Mohan, Ramachandran Ramjee, and Ashish Panwar. 2025. vAttention: Dynamic Memory Management for Serving LLMs without PagedAttention. arXiv:2405.04437
- [38] Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Saurabh Jha, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Tamer Başar, and Ravishanker K Iyer. 2024. Power-aware deep learning model serving with {μ-Serve}. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*.
- [39] Tella Rajashekhar Reddy, Palak, Rohan Gandhi, Anjaly Parayil, Chaojie Zhang, Mike Shepperd, Liangcheng Yu, Jayashree Mohan, Srinivasan Iyengar, Shivkumar Kalyanaraman, and Debopam Bhattacharjee. 2025. AI Greenfencing: Routing AI Inference to Green Modular Data Centers with Heron. arXiv:2505.09989
- [40] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems* 37 (2024).
- [41] ShareGPT Team. 2023. ShareGPT. <https://sharegpt.com/>
- [42] Haiying Shen and Tanmoy Sen. 2025. AccelGen: Heterogeneous SLO-Guaranteed High-Throughput LLM Inference Serving for Diverse Applications. arXiv:2503.13737
- [43] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. 2023. S-LoRA: Serving Thousands of Concurrent LoRA Adapters. *arXiv preprint arXiv:2311.03285* (2023).
- [44] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. Flexgen: High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*.
- [45] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).
- [46] Matej Špet'ko, Ondřej Vysocký, Branislav Jansík, and Lubomír Říha. 2021. Dgx-a100 face to face dgx-2—performance, power and thermal behavior evaluation. *Energies* (2021).
- [47] Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise Parallel Decoding for Deep Autoregressive Models. In *Advances in Neural Information Processing Systems*.
- [48] Jovan Stojkovic, Esha Choukse, Chaojie Zhang, Inigo Goiri, and Josep Torrellas. 2024. Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. arXiv:2403.20306
- [49] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Esha Choukse, Haoran Qiu, Rodrigo Fonseca, Josep Torrellas, and Ricardo Bianchini. 2025. Tapas: Thermal-and power-aware scheduling for LLM inference in cloud platforms. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*.
- [50] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, and Esha Choukse. 2025. Dynamollm: Designing llm inference clusters for performance and energy efficiency. In *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*.
- [51] Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin. 2024. Llumnix: Dynamic scheduling for large language model serving. In *18th USENIX symposium on operating systems design and implementation (OSDI 24)*.
- [52] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multi-modal models. *arXiv preprint arXiv:2312.11805* (2023).
- [53] Mistral AI team. 2024. Un Ministral, des Ministraux | Mistral AI. <https://mistral.ai/news/ministraux>
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017.

- Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [55] Xiaohui Wang, Ying Xiong, Yang Wei, Mingxuan Wang, and Lei Li. 2021. LightSeq: A High Performance Inference Library for Transformers. In *NAACL’21 Industry Papers*.
- [56] Zhibin Wang, Shipeng Li, Yuhang Zhou, Xue Li, Rong Gu, Nguyen Cam-Tu, Chen Tian, and Sheng Zhong. 2024. Revisiting SLO and Goodput Metrics in LLM Serving. *arXiv:2410.14257*
- [57] Norbert Wiener. 2019. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.
- [58] Bingyang Wu, Yinmin Zhong, Zili Zhang, Shengyu Liu, Fangyue Liu, Yuanhang Sun, Gang Huang, Xuanzhe Liu, and Xin Jin. 2024. Fast Distributed Inference Serving for Large Language Models. *arXiv:2305.05920*
- [59] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.
- [60] xAI. 2023. Announcing Grok. <https://x.ai/blog/grok> Accessed: 2025-05-15.
- [61] Yi Xu, Ziming Mao, Xiangxi Mo, Shu Liu, and Ion Stoica. 2024. Pie: Pooling cpu memory for llm inference. *arXiv preprint arXiv:2411.09317* (2024).
- [62] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).
- [63] Fuxun Yu, Zirui Xu, Tong Shen, Dimitrios Stamoulis, Longfei Shang-guan, Di Wang, Rishi Madhok, Chunshui Zhao, Xin Li, Nikolaos Karianakis, et al. 2020. Towards latency-aware DNN optimization with GPU runtime analysis and tail effect elimination. *arXiv preprint arXiv:2011.03897* (2020).
- [64] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. Orca: A distributed serving system for {Transformer-Based} generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. 521–538.
- [65] Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. 2025. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 23378–23386.
- [66] Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. 2024. Enhancing Zero-Shot Chain-of-Thought Reasoning in Large Language Models through Logic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 6144–6166.
- [67] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998* (2023).
- [68] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. 2024. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems* 37 (2024), 62557–62583.
- [69] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. 2024. {DistServe}: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*. 193–210.

A Length Distribution of ShareGPT and LMSYS Dataset

Table 2 summarizes the statistical properties of the ShareGPT and LMSYS datasets used in the evaluation, including the mean and standard deviation of their input and output lengths. The cumulative distribution functions (CDFs) for these length distributions are visualized in Fig. 22. Analysis of this data shows that:

- ShareGPT has longer prefill and decode length compared to LMSYS dataset.
- ShareGPT has larger prefill/decode length ratio.

Table 2. Length distribution of datasets used in our evaluation. Std refers to standard deviation.

Dataset	Prefill		Decode	
	Mean	Std	Mean	Std
ShareGPT	280.27	375.58	190.90	209.15
LMSYS	78.40	133.29	174.57	166.13

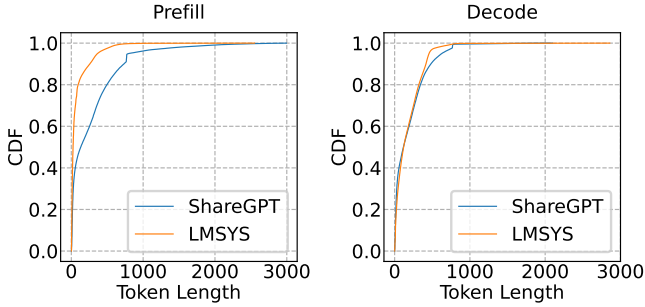


Figure 22. CDF of prefill/decode length for ShareGPT and LMSYS datasets.

B Batch Size Boundaries of Prefill Phase

Fig. 23, a zoomed-in view of Fig. 12a, focuses on the prefill phase with a small number of batched tokens. We can observe similar “staircase-like” boundary effect like decode phase (Fig. 12b). However, this effect only exist when number of batched tokens are relative small. When number of batched tokens goes over about 2000, this effect gradually becomes less significant.

C Additional Evaluation Results

We compute the throughput from the experimental results in § 6.2 using ShareGPT dataset on LLaMA-3.1-8B. As shown in Fig. 24, VoltanaLLM achieves slightly lower throughput than SGLang with static maximum frequency due to its opportunistic frequency scaling, but not at the cost of SLO violations. Furthermore, at high request rates where greater

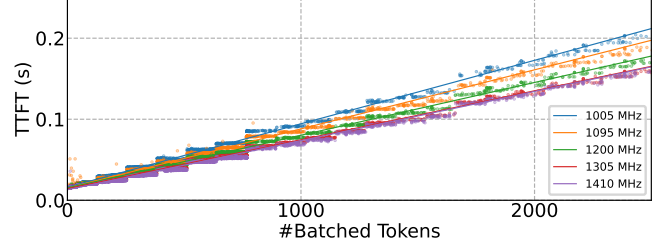


Figure 23. Relationship between TTFTs and batched token numbers based on profiling result from § 5.5.

throughput is required, VoltanaLLM can also adaptively achieve nearly the same throughput as SGLang with static maximum frequency.

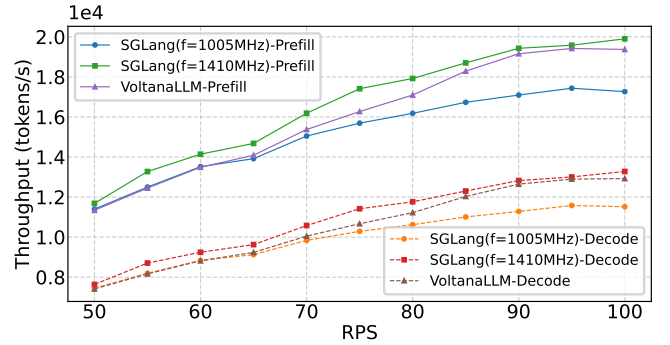


Figure 24. Comparing throughput of VoltanaLLM with baselines.