Algorithms for Context Engineering in LLM Inference: Optimization of Placement, Compression, and Scheduling

Teresa Zhang

Stanford University

Abstract

Scaling long-context and agentic LLMs is increasingly limited by memory capacity and bandwidth rather than FLOPs. I propose an algorithmic framework for context engineering that treats placement, compression, and scheduling as coupled optimization problems with explicit accuracy-efficiency trade-offs. I will develop (1) salience-aware retention/eviction with approximation guarantees; (2) tier-dependent compression with bounded distortion; and (3) probabilistic prefetching/scheduling that controls tail latency. Evaluation on long-context modeling and reasoning benchmarks will isolate each component and compare against strong heuristics under controlled bandwidth/capacity regimes. The goal is improved throughput and energy efficiency at near-baseline quality, enabling principled, hardware-aware inference without custom hardware.

Introduction

Generative AI is increasingly constrained not by FLOPs but by *memory capacity and bandwidth* at inference, as shown by IO-aware attention (Dao et al. 2022; Dao 2024) and memory-centric serving systems such as PagedAttention (Kwon et al. 2023) and FlexGen (Sheng et al. 2023). As long-context use cases (e.g., agents, tool use, multi-turn reasoning) expand the working set, memory pressure grows sharply, now recognized as a core efficiency bottleneck (Mei et al. 2025a; Tomar et al. 2025). Yet current systems rely on ad-hoc heuristics such as uniform eviction (Kwon et al. 2023), flat quantization (Frantar et al. 2023; Dettmers et al. 2023; Lin et al. 2024), and reactive prefetching (Ren et al. 2021a), leaving performance on the table under tight budgets. This motivates a unified, algorithmic treatment of context—selecting, compressing, and scheduling state within fixed capacity and bandwidth—that remains effective across attention variants and hardware regimes.

I propose **hardware-aware context engineering**: a unified framework for retention, compression, and deadline-aware scheduling under explicit memory constraints. The goals are to: (i) formalize *budgeted retention/eviction* with approximation targets to an oracle; (ii) design *tier-dependent compression* that bounds attention distortion; and (iii) develop *deadline-aware prefetching* that controls tail latency. Each component will include simple guarantees and be evaluated on long-context modeling and reasoning tasks under controlled bandwidth/capacity regimes.

This work builds on my prior research on bandwidth-efficient ANN search and compression-aware data structures. While hardware motivates the constraints, the *focus is algorithmic*, with validation via simulation and integration into inference frameworks. I view hardware-aware context engineering as a natural next step in my undergraduate research trajectory and a foundation for doctoral work on scalable, efficient AI systems.

Background

Inference at scale is increasingly limited by *memory capacity and bandwidth* rather than compute. Along the kernel path, FlashAttention (Dao et al. 2022; Dao 2024) makes attention I/O-aware. At the system level, vLLM's PagedAttention (Kwon et al. 2023), FlexGen (Sheng et al. 2023), and DeepSpeed-Inference (Ren et al. 2021b) manage weights and KV caches across heterogeneous tiers. Complementary work compresses model states: GPTQ (Frantar et al. 2023), AWQ (Lin et al. 2024), and QLoRA (Dettmers et al. 2023) reduce weight precision, while KVQuant (Hooper et al. 2024) targets KV caches. XQuant (Tomar et al. 2025) further shows that activation quantization with KV rematerialization can reduce memory use by an order of magnitude at near-baseline accuracy.

Concurrently, *context* itself is emerging as the next efficiency frontier. A recent survey (Mei et al. 2025b) summarizes methods for restructuring prompts, histories, and external knowledge, while long-context reasoning, retrieval-augmented generation, and agentic workflows continue to expand context footprints. Techniques such as linear and hybrid attention reduce *compute* cost but do not address how large, dynamic state should be *selected, compressed, placed, and prefetched* under memory bandwidth and capacity limits. Across these strands, most techniques are heuristic and *modularized*: kernel optimizations, cache placement, and compression are tuned independently. What is missing is a **unified, hardware-aware algorithmic framework** coupling (i) salience-aware retention, (ii) tiered compression with distortion control, and (iii) deadline-aware scheduling/prefetching, with simple guarantees under controlled capacity/bandwidth regimes. This proposal addresses that gap.

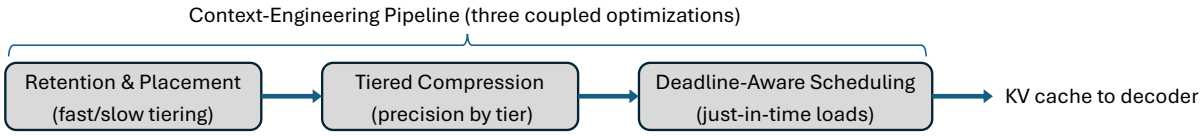

Figure 1: Unified context-engineering pipeline. Retention & placement assign context to memory tiers; tiered compression sets per-tier quantization/precision; deadline-aware scheduling prefetches KV to meet latency targets under bandwidth constraints.

Prior Work by the Applicant

Working primarily as an *independent, sole author*, I have advanced bandwidth-centric ANN search and compression-aware data structures (Zhang 2025a,b), alongside one collaborative applied-compression poster (Zhang, Scott, and Pauly 2025). These projects (including a *Best Student Presentation Award* (Zhang 2025a)) have not only established my independence as a researcher but also prepared me well to pursue the unified, hardware-aware context-engineering framework proposed here.

Approach

This project proceeds in three stages: *characterization*, *algorithm design*, and *prototyping*.

Characterization. I will quantify how growing context affects inference under heterogeneous memory. Using a controllable fast→slow hierarchy, I will sweep bandwidth and capacity and report tokens/s, p95/p99 latency, bytes moved, and task quality.

Algorithms. As shown in Fig. 1, placement, compression, and scheduling are posed as coupled **optimization problems** with accuracy-efficiency trade-offs. KV retention/eviction is modeled as budgeted selection using reuse/attention-decay estimates and submodular/knapsack-style approximations (Nemhauser, Wolsey, and Fisher 1978). Tiered compression follows rate–distortion principles, allocating precision by tier and connecting to quantization error analyses (Frantar et al. 2023; Lin et al. 2024). Deadline-aware prefetch draws on stochastic scheduling and queuing tail bounds to control miss probability.

Prototype & validation. Policies will be implemented in a trace-driven simulator and then a PyTorch/vLLM-style stack. Success is defined by improved throughput and p95/p99 latency, reduced KV bytes moved, and agreement between empirical and theoretical guarantees under fixed bandwidth/capacity budgets.

Evaluation

I will evaluate along two angles: *theoretical guarantees* and *empirical performance* under explicit memory budgets.

Theory. Provide (i) approximation factors for retention/placement versus an oracle; (ii) bounds linking tiered compression to attention distortion and task quality; (iii) probabilistic bounds on deadline-miss rates; and (iv) a simple composition bound for combined error.

Empirics. Validate on long-context language modeling (extended WikiText-2, C4) and long-context reasoning/comprehension (LongBench, RULER). Sweep per-tier bandwidth/capacity; report tokens/s, tail latency, bytes moved on fast tiers, and task quality. Baselines: Strong heuristics (uniform/LRU eviction, flat KV quantization, no prefetch) and a PagedAttention-style system.

Sensitivity & stress tests. Probe robustness under (i) highly skewed vs. uniform access, (ii) bursty vs. smooth reuse, and (iii) prefetch misprediction and bandwidth throttling. Check that empirical degradation matches the theoretical bounds.

Success. Demonstrate statistically significant speed/latency gains at comparable quality, reduced KV bytes moved, and observed miss rates consistent with theoretical limits.

Discussion

Hardware-aware context engineering can lower the resource cost of long-context inference without sacrificing quality. By casting salience-aware retention/placement, tiered compression, and deadline-aware scheduling as simple, analyzable optimizations, it elevates *context* to a core algorithmic dimension, complementary to quantization and sparsity, and orthogonal to attention variants (standard, linear, hybrid). Reducing bytes moved and pressure on fast tiers can translate to lower energy and enable long-context usage on modest hardware, while exposing explicit knobs to trade accuracy for efficiency under fixed budgets.

Limitations. Salience estimation and reuse prediction are imperfect; guarantees rely on model assumptions (e.g., monotone utility or light-tail delays). I will report failure modes (adversarial access, distribution shift) and ensure graceful fallback to safe baselines.

Broader impacts. Lowering memory traffic can reduce datacenter energy and enable on-device/private long-context use. Potential risks include fairness or safety shifts from retention policies; I will test on standard safety benchmarks and report mitigations.

Conclusion

This project targets the bottleneck of context growth under tight memory capacity and bandwidth and develops a unified, hardware-aware algorithmic framework with guarantees for retention, compression, and scheduling. I will validate on long-context benchmarks under controlled budgets, aiming for practical speed/latency gains at comparable quality and a principled foundation for context-centric inference. Longer-term, I hope to extend learning-based salience and adaptive budgeting, integrate with linear/hybrid attention and multimodal contexts, and explore on-device deployments where memory is the dominant constraint.

References

Dao, T. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *International Conference on Learning Representations (ICLR)*.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Frantar, E.; Ashkboos, S.; Kurtic, E.; Fawzi, A.; and Alistarh, D. 2023. GPTQ: Accurate Post-training Quantization for Generative Pre-trained Transformers. In *International Conference on Learning Representations (ICLR)*.

Hooper, C.; Kim, S.; Mohammadzadeh, H.; Mahoney, M. W.; Shao, Y. S.; Keutzer, K.; and Gholami, A. 2024. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*.

Lin, J.; Tang, C.; Tang, Z.; and Han, S. 2024. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. In *Conference on Machine Learning and Systems (MLSys)*.

Mei, L.; et al. 2025a. Context Engineering for Large Language Models: Survey and Vision. *arXiv preprint arXiv:2507.13334*.

Mei, L.; et al. 2025b. Context Engineering for Large Language Models: Survey and Vision. *arXiv preprint arXiv:2507.13334*.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An Analysis of Approximations for Maximizing Submodular Set Functions. *Mathematical Programming*, 14(1): 265–294.

Ren, S.; Rajbhandari, S.; Aminabadi, R. Y.; Ruwase, O.; Yang, M.; Rasley, J.; and He, Y. 2021a. DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. In *ICML Systems Workshop*.

Ren, S.; Rajbhandari, S.; Aminabadi, R. Y.; Ruwase, O.; Yang, M.; Rasley, J.; and He, Y. 2021b. DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. In *Proceedings of the ICML Systems Workshop*.

Sheng, Y.; Wang, Z.; Zhang, Y.; Jia, X.; Zhang, C.; Zhuo, D.; Cui, H.; and Cheng, J. 2023. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR.

Tomar, A.; Hooper, C.; Lee, M.; Xi, H.; Tiwari, R.; Kang, W.; Manolache, L.; Mahoney, M. W.; Keutzer, K.; and Gholami, A. 2025. XQuant: Breaking the Memory Wall for LLM Inference with KV Cache Rematerialization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, T. 2025a. Optimizing Memory Bandwidth for Efficient Approximate Nearest Neighbor Search. In *KDD Undergraduate Consortium (UMC)*. Toronto, Canada. Best Student Presentation Award.

Zhang, T. 2025b. Rethinking In-Memory Hash Table Design for CXL-Based Main Memory Compression. *IEEE Computer Architecture Letters*. Under 2nd-round review.

Zhang, T.; Scott, G.; and Pauly, J. 2025. Simple Universal Codes: Lossless Compression for Lower MRI Data Transmission Rates. In *International Society for Magnetic Resonance in Medicine (ISMRM) Annual Meeting*. Hawaii, USA. Poster.