

AI Economics

Engineering enterprise AI for performance, control and measurable business value

A research-led white paper on why enterprise AI economics depend on a combination of architecture, memory management, decoding strategy, workload design and governance and not simply token price.

SOLVEWORX

AI ECONOMICS

Why enterprise AI cost depends on **far more than token price.**

Research shows that inference architecture, memory management, decoding strategy, workload design and governance can materially change throughput, latency and infrastructure cost.

The graphic illustrates a stack of layers representing the AI architecture, from bottom to top: COMPUTE INFRASTRUCTURE, MODEL LAYER, INFERENCE ENGINE, RETRIEVAL & CONTEXT, AGENTS, APPLICATIONS & WORKFLOWS, and BUSINESS VALUE. To the right of the stack, five icons represent key benefits: a bar chart for HIGHER THROUGHPUT, a clock for LOWER LATENCY, a database cylinder for LOWER INFRASTRUCTURE COST (POTENTIAL), a shield for GREATER RELIABILITY, and a group of people for MEASURABLE BUSINESS VALUE.

Prepared by
Solveworx Digital Innovation

Date
October 2026

Important Note

The white paper has been prepared using peer-reviewed and institutional research findings in support of Solveworx own in-house data science research and findings which has in turn informed Solveworx product positioning. No claim is made that T-Flux Ultra *automatically* reproduces published benchmark results; realised gains depend on workload, hardware, model choice, implementation quality and measurement.

Executive summary

Enterprise AI economics are increasingly determined by how models are served, not simply by which model is selected or the nominal price of a token. The research base now shows that memory management, attention implementation, batching and scheduling, decoding strategy and system-level serving design can materially change throughput, latency and infrastructure utilisation. ^{[1][2][3][4]}

This matters because production inference is a recurring operating cost. The [FinOps Foundation](#) reports that inference can account for 80-90% of total GenAI spend in many use cases and that GPU under-utilisation can be substantial, creating a direct link between engineering efficiency and financial performance. ^[5]

At the same time, frontier model performance is converging on some widely used comparative measures. [Stanford's 2026 AI Index](#) reports that top providers were clustered closely on Arena-style Elo ratings as of March 2026, increasing the importance of cost, reliability and domain-specific performance as differentiators. ^[6]

CORE CONCLUSION

The defensible executive proposition is not that token prices are irrelevant. It is that token price is only one component of total AI economics. Architecture and operational discipline determine how much useful work the organisation obtains from each unit of compute, memory and model usage.

- Serving architecture can produce large performance differences. vLLM's PagedAttention reported 2-4x higher throughput at comparable latency than prior systems in evaluated workloads. ^[1]
- Memory is a first-class constraint. KV-cache design affects the ability to deliver low-latency, high-throughput serving and is the subject of active optimisation at token, model and system levels. ^{[4][7]}
- Decoding strategy can reduce latency. Speculative decoding drafts candidate tokens cheaply and verifies them in parallel, reducing the sequential bottleneck of autoregressive generation. ^[3]
- Low-level implementation matters. FlashAttention demonstrated that IO-aware attention can reduce memory traffic and deliver material end-to-end speedups in tested transformer workloads. ^[2]
- Financial governance must accompany technical optimisation. FinOps principles emphasise continuous measurement of cost, performance and business impact rather than infrastructure cost in isolation. ^[5]

1. The problem: AI cost is a systems question

A simplistic view of AI cost focuses on the model invoice: input tokens, output tokens or GPU-hours. That is necessary for budgeting, but insufficient for understanding total economics. Production AI is a chain of interdependent design choices: model size and architecture, prompt and context length, retrieval quality, batching, KV-cache behaviour, decoding strategy, GPU utilisation, observability and governance. Inefficiency at any layer can increase the amount of infrastructure required to deliver the same business outcome.

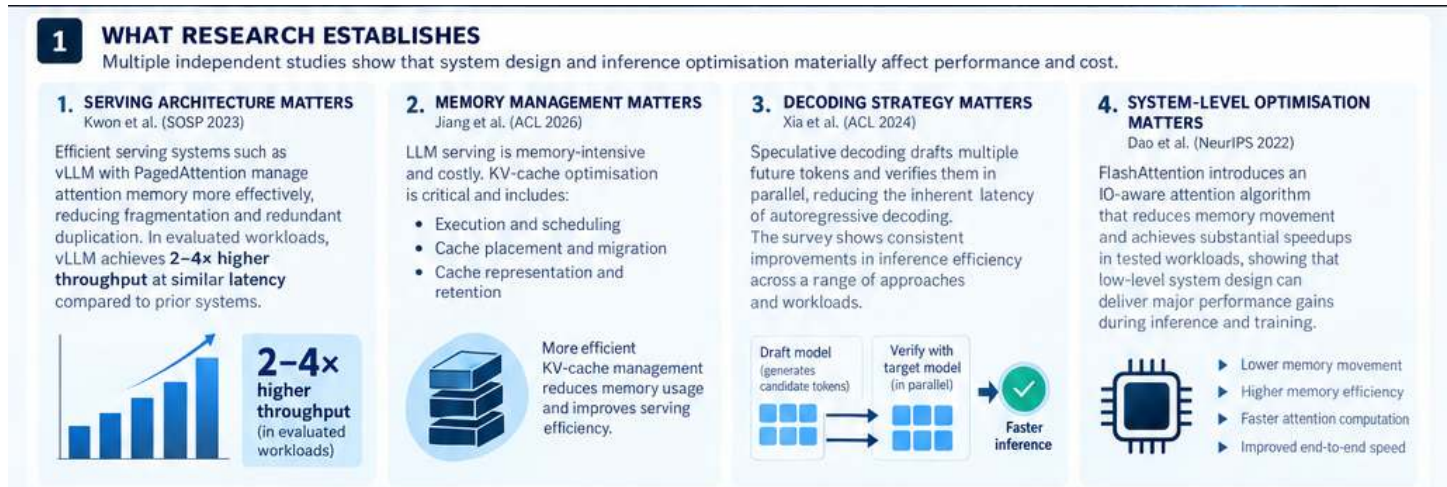


Enterprise AI cost is created across the serving stack, not at a single price point.

This systems framing is consistent with the research literature. Kwon et al. identify KV-cache fragmentation and duplication as constraints on batch size and throughput. Jiang et al. describe modern LLM serving as memory-intensive and costly, and organise KV-cache optimisation around execution and scheduling, placement and migration, and representation and retention. Li et al. similarly classify KV-cache optimisation at token, model and system levels.^{[1][4][7]}

The implication is operational: enterprise teams should measure cost per useful task, request or business outcome, not only cost per token. A low token price can still produce poor economics if prompts are oversized, retrieval is noisy, GPU capacity is idle, concurrency is poorly managed or the serving stack repeatedly recomputes avoidable work.

2. What the research establishes



research-backed mechanisms that materially influence LLM serving efficiency.

2.1 Serving architecture and memory management

In vLLM, Kwon et al. introduced PagedAttention, adapting virtual-memory concepts to KV-cache management. The system was designed to reduce fragmentation and redundant duplication so that more requests could be batched within available memory. In their evaluation, vLLM delivered 2-4x higher throughput at the same level of latency relative to the comparison systems, with larger gains for longer sequences, larger models and more complex decoding algorithms.^[1]

WHY THIS MATTERS TO EXECUTIVES

The published result is not a universal 2-4x cost reduction. It demonstrates that serving-system architecture alone can create multi-fold differences in usable throughput on the same underlying class of workload. That makes architecture an economic variable, not merely a technical implementation detail.

2.2 IO-aware attention

Dao et al. showed that transformer performance is constrained not only by arithmetic operations but by data movement between GPU memory hierarchies. FlashAttention reduces reads and writes between high-bandwidth memory and on-chip SRAM through tiling. The paper reported end-to-end wall-clock improvements including 15% on BERT-large, 3x on GPT-2 at sequence length 1K and 2.4x on Long Range Arena workloads.^[2]

For enterprise architecture, the important point is broader than FlashAttention itself: implementation choices that reduce memory movement can materially alter latency, throughput and the hardware footprint required for a given workload.

2.3 Speculative decoding

Autoregressive generation is sequential: each new token ordinarily depends on the preceding sequence. Xia et al. survey speculative decoding techniques that draft several future tokens using a lower-cost mechanism and then verify them in parallel with the target model. This approach can decode multiple accepted tokens per target-model step and thereby reduce inference latency.^[3]

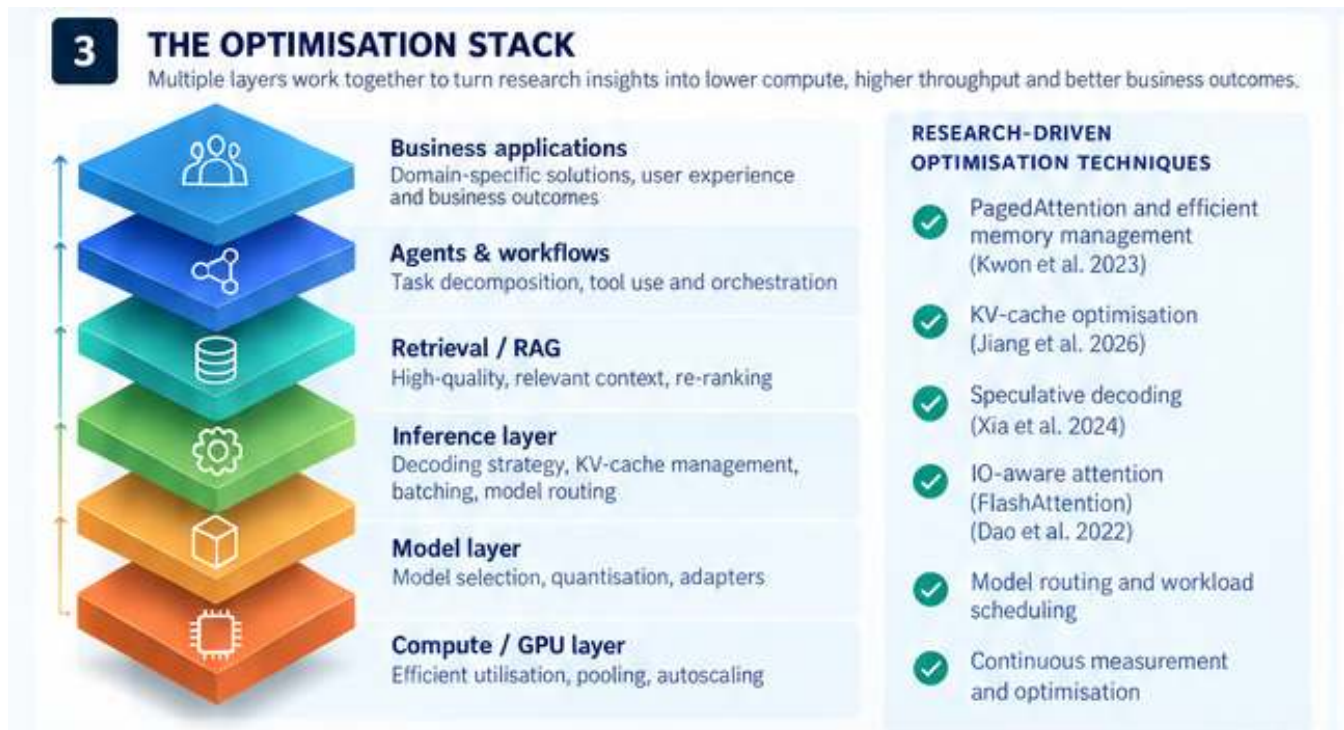
The magnitude of benefit is workload- and implementation-dependent. Drafter quality, verification strategy, model characteristics and hardware all affect realised speedup. The defensible claim is therefore that decoding strategy is a material optimisation lever, not that every workload will achieve a fixed percentage reduction.

2.4 KV-cache optimisation as a multi-layer discipline

The KV cache stores attention key and value tensors generated during autoregressive decoding. It avoids recomputing prior states, but its memory footprint grows with sequence length and concurrency. Jiang et al. describe the KV cache as crucial to low-latency, high-throughput serving and structure optimisation into three system behaviours: execution and scheduling, placement and migration, and representation and retention.^[4]

Li et al., published in Transactions on Machine Learning Research, provide a complementary taxonomy spanning token-level techniques such as selection, allocation, merging, quantisation and low-rank decomposition; model-level architectural techniques; and system-level memory management and hardware-aware scheduling.^[7]

3. The optimisation stack: from GPU to business application



Optimisation operates across layers: compute, model, inference, retrieval, agents and business applications.

The research points toward a layered optimisation model rather than a single “best model” decision. Each layer creates distinct opportunities and trade-offs:

- **Compute / GPU layer:** Utilisation, pooling, autoscaling and hardware fit determine how much paid capacity is productive. FinOps guidance explicitly calls out pooling and dynamic scaling as responses to under-utilisation.^[5]
- **Model layer:** Model size, quantisation and routing influence memory footprint, latency and quality. Smaller or specialised models may be preferable when they satisfy the task at lower resource cost.^[5]
- **Inference layer:** Batching, scheduling, KV-cache management, speculative decoding and memory-aware attention directly shape throughput and latency.^[1]
- **Retrieval / RAG layer:** Retrieval quality determines how much context must be processed and whether the model receives relevant evidence. Poor retrieval can increase tokens and degrade answer quality simultaneously.^[5]
- **Agents and workflows:** Tool selection, step count, retries and orchestration affect the number of model calls and the amount of repeated context. These are application architecture choices rather than model-provider prices.^[5]
- **Business application layer:** The optimisation objective should ultimately be expressed in service-level and business terms: cost per resolved case, report, decision, transaction or other measurable outcome.^[5]

The stack also exposes an important governance principle: optimisation should not be equated with minimising compute at any cost. The target is an efficient frontier between quality, latency, reliability, security and cost. A cheaper model or shorter context is economically inferior if it increases error rates or forces expensive human remediation.

4. AI FinOps: continuous optimisation rather than one-off tuning

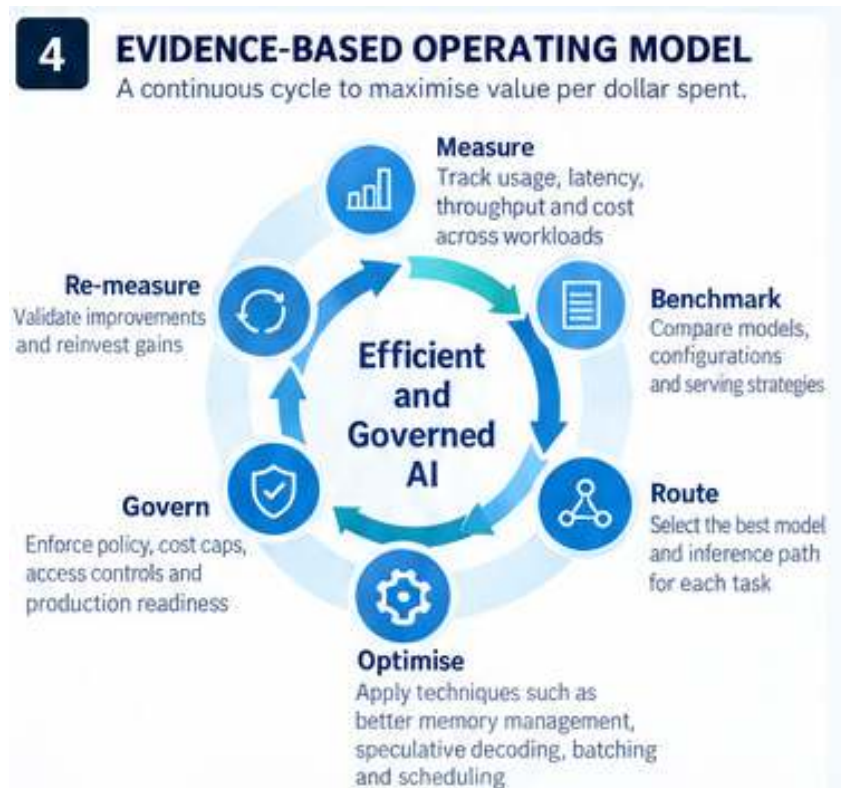


Figure 5. An evidence-based operating cycle: measure, benchmark, route, optimise, govern and re-measure.

FinOps provides a useful operating analogy. The FinOps Foundation argues that generative AI creates financial-management challenges because recurring inference spend, infrastructure configuration and product decisions interact. Its guidance recommends granular metrics that expose the trade-off between cost, performance and business impact.^[5]

What is FinOps?

FinOps is an operational framework and cultural practice which maximizes the business value of technology, enables timely data-driven decision making, and creates financial accountability through collaboration between engineering, finance, and business teams - *FinOps Foundation Technical Advisory Council. Updated: March 2026*

For enterprise AI, this can be operationalised as a continuous cycle:

1. **Measure:** Capture tokens, latency, GPU utilisation, cache hit rates, concurrency, quality and unit cost.
2. **Benchmark:** Compare models, configurations and serving paths using representative workloads rather than synthetic averages alone.
3. **Route:** Use the least expensive model and inference path that satisfies the task's quality, latency, privacy and reliability constraints.
4. **Optimise:** Tune batching, scheduling, cache management, decoding strategy, retrieval and context discipline.
5. **Govern:** Apply policy, access controls, approved configurations, observability and cost caps so optimisation is repeatable rather than ad hoc.
6. **Re-measure:** Validate gains in production and reinvest only where improvements are measurable.

MEASUREMENT DISCIPLINE

Published speedups should be treated as evidence that optimisation matters, not as guaranteed enterprise savings. Each organisation should establish its own baseline and measure cost, quality, latency and utilisation before and after changes.

5. Why model selection alone is becoming a weaker differentiator

Stanford's 2026 AI Index reports convergence among top frontier models on Arena-style Elo ratings, with leading providers clustered closely as of March 2026. Stanford explicitly notes that this shifts competitive pressure toward cost, reliability and domain-specific performance.^[6]

This should not be interpreted as saying that models are interchangeable. Capability varies by task, modality, reasoning depth, safety controls, deployment mode and licensing. The more defensible strategic conclusion is that organisations should avoid treating access to a frontier model as a durable competitive advantage. The advantage increasingly lies in the surrounding system: proprietary data, retrieval quality, workflow integration, governance, evaluation, optimisation and operational execution.

6. Implications for Solveworx and T-Flux Ultra



Figure 6. Research-aligned claims and the implementation implications for T-Flux Ultra.

The research supports a restrained positioning for T-Flux Ultra. The platform should not claim that it creates lower cost simply because it orchestrates multiple models or operates within a private environment. The stronger proposition is that an orchestration and governance layer can create economic value when it operationalises the optimisation techniques and control disciplines identified in the literature.

Accordingly, the evidence-aligned Solveworx proposition is:

- T-Flux Ultra provides a control plane through which model routing, governed retrieval, inference settings, workload scheduling, observability and cost policies can be operationalised.
- The optimisation service applies specialist expertise to tune those controls against representative workloads and explicit service-level objectives.
- The combination is intended to reduce avoidable computation, improve utilisation and make AI operating expenditure more predictable; the magnitude of any improvement must be demonstrated through customer-specific measurement.
- Fixed or cost-capped commercial models become credible only when the underlying platform has sufficient observability, governance and workload control to manage the cost envelope.
- Business value should be measured in terms of unit economics and outcomes, not simply token consumption.

This framing is consistent with Solveworx’s internal strategy notes, which position the company around “AI Economics”, continuous optimisation and predictable operating expenditure rather than unrestricted developer consumption.^[8]

Authors Note

The research sources cited in this paper validates research conducted by Solveworx data analysis, namely the importance of the mechanisms - memory management, scheduling, decoding, attention implementation, utilization and continuous measurement as contributory factors as controllable inputs to economic cost. These controllable cost inputs have been incorporated in the design and implementation of T-Flux Ultra however the findings in the research papers do not independently validate the performance of T-Flux Ultra. Solveworx encourages product-specific performance through reproducible benchmarks and customer measurements.

7. A practical measurement framework

A research-led AI economics programme should maintain a scorecard across four dimensions. This prevents local optimisation - for example, minimising token cost while degrading latency or accuracy - from being mistaken for enterprise value.

Dimension	Example measures	Why it matters	Primary evidence
Cost	Cost/request; cost/task; GPU-hour/unit; token spend	Connects engineering choices to unit economics	FinOps [5]
Performance	Throughput; time-to-first-token; tokens/sec; queue time	Tests serving architecture and decoding efficiency	Kwon [1]; Xia [3]
Resource efficiency	GPU utilisation; memory occupancy; KV-cache efficiency; batch size	Shows whether paid capacity is productive	Kwon [1]; Jiang [4]; Li [7]
Quality & reliability	Task success; factuality; retrieval precision; retry rate; SLA attainment	Prevents false savings that degrade outcomes	FinOps [5]; Stanford [6]

8. Conclusions

The academic and institutional evidence supports a clear reframing of enterprise AI economics. The cost of production AI is not reducible to the price of a model token. Serving architecture, memory management, attention implementation, decoding strategy, scheduling, retrieval and resource utilisation materially influence the amount of infrastructure and time required to deliver a workload.^{[1][2][3][4][7]}

For executives, this means AI should be managed as an operating system with measurable unit economics, not as an unlimited experimentation environment. For technology teams, it means optimisation must be continuous and workload specific.

For Solveworx, it supports an evidence-aligned position around AI optimisation and governance - provided product claims are benchmarked and expressed as measured outcomes rather than assumed benefits and are

References

- [1] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023). arXiv:2309.06180.
<https://arxiv.org/abs/2309.06180>
- [2] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. Advances in Neural Information Processing Systems (NeurIPS 2022). arXiv:2205.14135.
<https://arxiv.org/abs/2205.14135>
- [3] Xia, H., Yang, Z., Dong, Q., Wang, P., Li, Y., Ge, T., Liu, T., Li, W., & Sui, Z. (2024). Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding. Findings of ACL 2024, 7655-7671. DOI: 10.18653/v1/2024.findings-acl.456.
<https://aclanthology.org/2024.findings-acl.456/>
- [4] Jiang, J., Yang, P., Zhang, R., & Liu, F. (2026). Towards Efficient Large Language Model Serving: A Survey on System-Aware KV Cache Optimization. Findings of ACL 2026, 38450-38476. DOI: 10.18653/v1/2026.findings-acl.1916.
<https://aclanthology.org/2026.findings-acl.1916/>
- [5] FinOps Foundation. (2025/2026). Optimizing GenAI Usage: A FinOps Perspective on Cost, Performance, and Efficiency. Industry guidance on inference spend, GPU utilisation and cost/performance measurement.
<https://www.finops.org/wg/optimizing-genai-usage/>
- [6] Stanford Institute for Human-Centered Artificial Intelligence. (2026). AI Index Report 2026 - Technical Performance. Stanford University.
<https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance>
- [7] Li, H., Li, Y., Tian, A., Tang, T., Xu, Z., Chen, X., Hu, N., Dong, W., Li, Q., & Chen, L. (2025). A Survey on Large Language Model Acceleration Based on KV Cache Management. Transactions on Machine Learning Research.
<https://openreview.net/forum?id=z3JZzu9EA3>
- [8] Solveworx Digital Innovation. (2026). Internal AI Economics strategy notes: optimisation, predictable operating expenditure, AI FinOps and T-Flux positioning. Internal source supplied for this white paper.

Research note

The literature cited contains a mix of peer-reviewed conference/journal publications and respected institutional/industry guidance. Benchmark results are reported in the context of the original study and should not be generalised as guaranteed savings. Enterprise results should be established through controlled workload-specific testing.