Inference Avoidance and Harmonic Storage: A Quantitative Framework for Ultra-Efficient Artificial Intelligence

Robert Edward Grant
The Institute of Unified Mathematics
Irvine, California

August 5, 2026

Abstract

Frontier Large Language Models (LLMs) have demonstrated extraordinary capabilities by applying transformer-based neural inference to virtually every user request. This design philosophy has also resulted in unprecedented computational cost, energy consumption, and infrastructure requirements.

The Orion Architect proposes an alternative computational paradigm based upon two complementary principles.

The first principle—Inference Avoidance—reduces energy consumption by minimizing how frequently transformer-based neural inference is required. Rather than relying on a neural model for every request, Orion resolves most tasks through deterministic reasoning, symbolic computation, graph-based knowledge, structured memory, and algorithmic execution, invoking an embedded Small Language Model (SLM) only when probabilistic language reasoning is necessary.

The second principle—Harmonic Improvements to Storage and Resonance—proposes several additional mechanisms that may further reduce computational energy through alternative methods of information representation, retrieval, and contextual integration. These harmonic mechanisms are presented as research hypotheses requiring experimental validation and are intentionally excluded from the quantitative energy projections developed in this paper.

Taken together, these principles provide a roadmap for developing AI systems that may consume orders of magnitude less energy than contemporary frontier LLMs.

1. Introduction

Modern frontier AI systems generally activate transformer-based neural inference for nearly every request.

Whether retrieving information, performing logical reasoning, generating language, or solving mathematical problems, billions of neural parameters participate in the computation.

While this architecture has proven remarkably capable, it also represents one of the most computationally expensive approaches to artificial intelligence ever developed.

The Orion Architect begins from a different premise:

> The greatest reduction in AI energy consumption may come not from making transformer inference slightly cheaper, but from designing intelligent systems that rarely need transformer inference in the first place.

This represents an architectural rather than purely neural approach to intelligence.

2. Architectural Principle I: Inference Avoidance

Instead of treating transformer inference as the universal computational mechanism, Orion distributes computation across specialized deterministic systems.

These include:

- symbolic reasoning
- graph-based knowledge retrieval
- structured memory
- mathematical algorithms
- deterministic planning
- rule-based execution
- contextual memory
- orchestration and routing

These deterministic operations consume energy, but they generally require dramatically fewer floating-point operations, substantially less memory bandwidth, and significantly less computational overhead than transformer inference.

Only when deterministic methods cannot confidently resolve a problem does Orion invoke an embedded Small Language Model.

Consequently, transformer inference becomes a specialized capability rather than the default computational engine.

3. Energy Model

To compare Orion with frontier LLMs, one frontier transformer inference is normalized to 100% energy.

The Orion Architect is modeled as:

$$E_{\mathrm{O}} \;=\; D + (S \times F)$$

where:

- D = deterministic processing energy relative to one frontier LLM inference.
- S = energy required by one embedded SLM inference relative to one frontier LLM inference.
- F = fraction of requests requiring SLM inference.

Unlike frontier LLMs, Orion minimizes both S and F.

4. Engineering Assumptions

The following assumptions are illustrative engineering parameters rather than experimentally verified measurements.

Embedded Small Language Model

Projected energy:

- approximately 1–5% of the energy required for one frontier LLM inference.

Neural Inference Frequency

Illustrative operating scenarios:

- 20% of requests
- 10% of requests
- 2% of requests

The actual percentage will be determined through benchmark testing.

Deterministic Processing Energy

The energy consumed by deterministic processing remains an engineering parameter to be measured.

Because graph traversal, symbolic reasoning, routing, memory retrieval, and algorithmic execution generally require far fewer computational resources than transformer inference, the deterministic energy cost is expected to be substantially lower than that of a frontier transformer inference.

This paper therefore treats deterministic processing energy as a measurable variable rather than assigning it a fixed value.

5. Baseline Energy Projections

The following projections assume:

- Embedded SLM = 2% of frontier LLM inference energy.

Deterministic Energy	Neural Usage	Orion Total Energy	Improvement vs Frontier LLM
1.00%	20%	1.40%	71×
1.00%	10%	1.20%	83×
1.00%	2%	1.04%	96×
0.50%	20%	0.90%	111×
0.50%	10%	0.70%	143×
0.50%	2%	0.54%	185×
0.10%	20%	0.50%	200×
0.10%	10%	0.30%	333×
0.10%	2%	0.14%	714×
0.05%	2%	0.09%	1,111×
0.02%	2%	0.06%	1,667×
0.01%	2%	0.05%	2,000×

These projections arise entirely from the baseline Orion architecture.

They assume no contribution whatsoever from harmonic computational mechanisms.

6. Architectural Principle II: Harmonic Improvements to Storage and Resonance

Beyond inference avoidance, the Orion Architect proposes a second, independent research program investigating whether additional reductions in computational energy may be achieved through alternative methods of information representation and computation.

Collectively these proposed mechanisms are referred to as:

Harmonic Improvements to Storage and Resonance

Unlike the inference-avoidance architecture described above, these mechanisms remain research hypotheses and require independent theoretical development, engineering implementation, and empirical validation.

The principal research directions include:

Harmonic Information Encoding

Conventional AI stores information within distributed neural weight matrices.

The Orion hypothesis proposes representing information through coherent harmonic state relationships.

Potential advantages to be investigated include:

- reduced storage energy
- lower memory bandwidth
- more efficient information representation
- reduced heat dissipation

Resonance-Based Information Retrieval

Conventional transformers retrieve information through attention-guided token prediction.

The Orion hypothesis proposes that coherent state representations may permit resonance-based activation of relevant information.

Potential research questions include:

- reduced computational search
- lower retrieval energy
- reduced inference latency
- improved retrieval efficiency

Harmonic Computational Processing

Current AI systems perform sequential numerical tensor computations.

The Orion hypothesis proposes investigating coherent harmonic interactions as an alternative computational substrate.

Potential research areas include:

- reduced floating-point computation
- improved computational parallelism
- lower processing energy
- alternative computational substrates

Continuous Context Integration

Current AI systems generally adapt through retraining, fine-tuning, or retrieval-augmented generation.

The Orion hypothesis investigates whether coherent state representations may support continuous contextual adaptation while reducing repeated computation.

Potential research questions include:

- reduced retraining requirements
- lower adaptive learning energy
- persistent contextual coherence
- reduced computational overhead

7. Two Independent Sources of Energy Reduction

The Orion Architect therefore contains two independent mechanisms for improving energy efficiency.

Baseline Architectural Efficiency

Derived from:

- deterministic reasoning
- symbolic computation
- graph intelligence
- structured memory
- inference avoidance
- compact neural models

These mechanisms produce the quantitative projections presented in this paper.

Harmonic Improvements to Storage and Resonance

Potential future reductions through:

- harmonic information encoding
- resonance-based retrieval
- harmonic computational processing
- continuous context integration

These mechanisms are not included in the quantitative energy projections developed herein.

If experimentally validated, they would represent additional energy savings beyond the projected baseline improvements of up to approximately 2,000×.

8. Discussion

The Orion research program intentionally separates measurable engineering claims from theoretical research hypotheses.

The baseline architecture can be benchmarked today using established methodologies, including measurements of:

- energy per inference
- latency
- throughput
- hardware utilization
- deterministic processing energy
- neural inference frequency

Separately, Harmonic Improvements to Storage and Resonance constitute an independent research program whose contributions should be quantified only after experimental validation.

This separation preserves scientific rigor while providing a clear roadmap for future investigation.

9. Conclusion

The Orion Architect proposes that the future of energy-efficient artificial intelligence may depend upon two complementary innovations.

The first is an immediately testable architectural principle: minimizing dependence on transformer inference through deterministic reasoning, symbolic computation, structured memory, graph intelligence, and selective neural inference.

Under the engineering assumptions explored in this paper, this baseline architecture alone projects improvements ranging from approximately $70\times$ to as much as $2{,}000\times$ lower inference energy than frontier Large Language Models, depending on the measured efficiency of deterministic processing and the fraction of requests requiring neural inference.

The second is an independent research program investigating Harmonic Improvements to Storage and Resonance. These proposed mechanisms are intentionally excluded from the baseline quantitative projections. Should future research validate their effectiveness, they would represent additional energy reductions beyond the baseline architectural improvements presented here.

The central architectural insight is therefore straightforward:

> The greatest opportunity for reducing AI energy consumption is not simply building more efficient neural networks, but building intelligent systems that rarely need to invoke them at all.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[2] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[3] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 3645–3650.

[4] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon Emissions and Large Neural Network Training," arXiv:2104.10350, 2021.

[5] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," *Journal of Machine Learning Research*, vol. 24, no. 253, pp. 1–15, 2023.

[6] A. S. Luccioni, Y. Jernite, and E. Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

[7] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference," in *IEEE High Performance Extreme Computing Conference (HPEC)*, 2023.

[8] A. de Vries, "The Growing Energy Footprint of Artificial Intelligence," *Joule*, vol. 7, no. 10, pp. 2191–2194, 2023.

[9] M. Horowitz, "Computing's Energy Problem (and What We Can Do About It)," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 2014, pp. 10–14.

[10] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," in *International Conference on Learning Representations (ICLR)*, 2017.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv:1503.02531, 2015.

[12] M. Abdin et al., "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," arXiv:2404.14219, 2024.

[13] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[14] A. Hogan et al., "Knowledge Graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2021.

[15] A. d'Avila Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd Wave," *Artificial Intelligence Review*, vol. 56, pp. 12387–12406, 2023.

[16] G. Marcus, "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence," arXiv:2002.06177, 2020.