

Orbital Intelligence: A Vision for Sustainable Large Language Model Training and Inference in Space

Mohamed Khalil Brik¹

¹The American University in Cairo

September 2025

Abstract

The rapid expansion of Large Language Model (LLM) training has led to significant computational requirements and associated carbon emissions. For example, training models such as GPT-4 produces approximately 7,138 metric tons of carbon dioxide equivalent. This vision paper proposes the use of space-based infrastructure for LLM training and, critically, inference, to address these sustainability challenges by utilizing abundant orbital solar energy. Orbital computation offers a critical thermodynamic advantage, accessing solar irradiance that is four times more intense than the terrestrial average. We propose a hybrid approach that combines dedicated orbital data centers (ODCs) in Geostationary Earth Orbit (GEO) for foundational training with distributed edge computing resources across Low Earth Orbit (LEO) constellations for sustainable inference. This reframed strategy tackles the 90% of a model's lifecycle energy consumption attributed to usage. The paper addresses core challenges, including radiation hardening, thermal management using Loop Heat Pipes (LHP), the logistics of data migration, and the strategic mandate for a tenfold reduction in launch emissions, concluding that space represents the next necessary frontier for large-scale, sustainable AI development.

1 Introduction

The unprecedented scaling of Artificial Intelligence (AI) has established Large Language Models (LLMs) as the technological cornerstone of the modern digital economy [1]. However, this transformative growth is underpinned by an accelerating and globally significant environmental footprint. Training state-of-the-art LLMs demands computational resources that place immense strain on terrestrial energy grids and water supplies, necessitating a radical shift in infrastructure deployment.

The migration of high-performance computing (HPC) to space is not merely a novelty but a strategic response to a fundamental energy crisis. The orbital environment provides a decisive advantage: unparalleled access to solar radiation unattenuated by the Earth's atmosphere [2]. This continuous, zero-carbon power source offers the means to fundamentally decouple AI computation from the variable, carbon-intensive constraints of terrestrial grids [3].

This paper details the techno-economic and archi-

tectural feasibility of establishing Orbital Data Centers (ODCs). We advocate for a hybrid, multi-orbital architecture designed to optimally place AI workloads: synchronous, high-power training in the stable environment of GEO, and asynchronous, high-volume inference within distributed LEO constellations. We analyze the core engineering solutions required for hardware resilience and thermal management, and we address the strategic mandates related to data logistics, regulatory governance, and the critical challenge of launch sustainability.

2 The Environmental Imperative

The case for orbital deployment is rooted in the alarming quantification of LLM development costs across carbon, energy, and water consumption.

2.1 Quantifying the Terrestrial Footprint

The training of frontier LLMs results in energy and carbon costs that far exceed previous computational models. For example, training OpenAI's GPT-3, possessing 175 billion parameters, required approximately 1,287 megawatt-hours (MWh) of electricity [4] and resulted in an estimated 552 metric tons of CO₂ [5]. The environmental cost scaled dramatically with its successor, GPT-4, which was estimated to produce 7,138 metric tons of CO₂, a footprint comparable to the annual emissions of 1,550 US citizens [5, 6].

Furthermore, terrestrial LLM operations impose a profound strain on water resources. Training GPT-3 alone required over 700 kiloliters (kL) of water for cooling, a volume large enough to fill two-thirds of an Olympic-sized swimming pool [7]. The orbital shift, by utilizing the deep-space vacuum for passive heat rejection, completely eliminates this reliance on terrestrial water for cooling.

2.2 The Inference-Centric Strategy

Historically, sustainability discussions concentrated on the singular, massive training phase. However, this focus is strategically incomplete. Recent comprehensive benchmarking and analysis indicate that LLM *inference* (continuous, global usage) is the dominant factor in lifetime environmental costs, potentially accounting for up to 90% of a model's total lifecycle energy consumption [7–9].

This paradigm shift necessitates a re-evaluation of the orbital value proposition. While a GEO hub may handle the 10% training problem, the greatest environmental return on investment (EROI) lies in leveraging LEO con-

stellations as a zero-carbon, distributed **edge inference network** to address the dominant 90% problem [10]. By relying exclusively on clean, captured solar energy, orbital infrastructure operates with a localized operational carbon intensity factor (F_{em}) tending toward zero [9].

3 Thermodynamic Advantages of Orbital Placement

The feasibility of Orbital Intelligence relies entirely on the superior energy continuity and intensity provided by the space environment.

3.1 Solar Irradiance and Attenuation

The critical advantage of space is the solar constant, the intense, unattenuated solar radiation flux, nominally measured at 1367.5 W/m^2 outside the Earth's atmosphere [2]. This figure varies seasonally by approximately $\pm 3.5\%$ due to orbital eccentricity [2].

This contrasts sharply with the average solar radiation reaching the Earth's surface (insolation), which is dramatically attenuated by atmospheric scattering, reflection, and absorption, averaging only about 340 W/m^2 [11]. Orbital collection, therefore, confirms a foundational physics advantage, accessing a solar energy source that is four times more intense than the terrestrial average, maximizing the power generation potential per unit of solar collector mass.

3.2 Orbital Stability and Power Continuity

The stability required for continuous, synchronous LLM training dictates the choice of orbit.

Table 1: Comparative Orbital Power Continuity

Parameter	LEO	GEO
Altitude (km)	160–2,000 [12]	35,786 [13]
Eclipses per Year	$\approx 5,500$ [14]	≈ 84 [14]
Max Eclipse Duration	35 minutes [15]	72 minutes [14]
Battery DoD Stress	High (5,000+ cycles) [14]	Low (Seasonal only)
Best Workload	Inference / Edge Computing [16]	Foundational Training [17]

Geostationary Earth Orbit (GEO) is ideal for foundational training due to its near-continuous solar exposure, experiencing predictable eclipses only seasonally around the equinoxes [13, 14]. Conversely, LEO's high frequency of eclipses (approximately 15 per day) [14] subjects hardware and battery systems to severe thermal and battery cycling stress (over 5,000 cycles annually), rendering it fundamentally unsuitable for sustained, synchronous training operations [15].

4 Hybrid Architectural Framework

The proposed hybrid architecture must strategically match the computational workload to the optimal orbital environment, leveraging both dedicated infrastructure and repurposing existing assets.

4.1 Dedicated Orbital Data Centers

Dedicated AI satellites are already transitioning from concept to reality. China, for instance, has embarked on constructing a vast orbital supercomputer network, currently achieving 5 POPS (peta operations per second) with a long-term objective of 1,000 POPS [18]. These specialized satellites utilize onboard 8-billion parameter AI models, high-speed laser inter-satellite links (ISLs) up to 100 Gbps, and 30 terabytes of storage to process remote sensing data and bypass terrestrial bandwidth bottlenecks [18, 18].

- **GEO Cluster for Foundational Training:** Foundational LLM training requires high-performance Remote Direct Memory Access (RDMA) networking to perform constant, synchronous, all-to-all model weight reduction operations across distributed GPU clusters [17]. The required instantaneous interconnect bandwidth is estimated at approximately 1 Gbps per GPU [19]. Only the static, stable GEO environment can reliably provide the required stable, high-synchronicity links.

- **LEO Constellations for Distributed Inference:** LEO networks are ideally suited for asynchronous distributed inference [10, 16]. In this model, LEO satellites function as "requesters" and "workers," processing data near the source to reduce communication latency and bandwidth cost [16].

4.2 Constraints of Mission Augmentation

While the hybrid approach includes computational resources on existing missions, their contribution is constrained. The International Space Station (ISS) successfully prototypes AI, notably using platforms like HPE's Spaceborne Computer-3 [20] for edge AI tasks such as assisting astronauts with maintenance procedures [12] and dynamic targeting for Earth observation [21]. However, repurposing Global Navigation Satellite Systems (GNSS), such as GPS, for deep learning is severely limited. Legacy, radiation-tolerant processors on these platforms are several magnitudes below the computing power required for advanced deep learning algorithms [22]. Existing missions can only feasibly support small-scale, asynchronous edge inference, lacking the volume, power budget, and thermal management necessary for sustained foundational LLM training [3, 22].

5 Critical Engineering and Logistics

The deployment of commercial-grade HPC hardware in space is constrained by two fundamental physical challenges: radiation and heat dissipation in a vacuum.

5.1 Radiation Hardening for COTS Hardware

The space environment presents severe risks to modern Commercial Off-The-Shelf (COTS) components, which are essential for LLM performance but are designed for benign terrestrial operation. Radiation from galactic cosmic rays and solar events causes two primary failure modes: Total Ionizing Dose (TID), resulting in cumulative material degradation, and Single Event Effects (SEE), causing functional disturbances or catastrophic failure from single particle strikes [23, 24]. The Geosynchronous Transfer Orbit (GTO), a proxy for the GEO environment, subjects hardware to extreme TID, potentially up to 100K Rad per annum [25].

Since performance necessitates COTS GPUs/TPUs, the strategy relies on advanced mitigation:

- **System-Level Mitigation:** Employing good fault-tolerant design practices is crucial for mitigating SEE at the circuit and system level [24].
- **Hardware Redundancy:** Specialized manufacturers, such as AMD, offer components incorporating Triple Modular Redundancy (TMR) designed to mitigate radiation effects [26].
- **Real-Time Shielding:** Novel concepts like the COTS-Capsule propose using particle detector arrays to encapsulate CSEE-sensitive electronics, enabling real-time detection and characterization of particle strikes for immediate mitigation or error correction [27].
- **Co-Processing:** Leveraging radiation-hardened coprocessors, like the High Performance Data Processor (HPDP), as a dependable mathematical backend allows the execution of AI-driven workloads without relying solely on vulnerable COTS components [28].

5.2 Thermal Management in Vacuum

High-performance AI accelerators generate substantial heat loads [3]. The vacuum of space prohibits atmospheric convection, rendering traditional terrestrial cooling impossible [3].

The reliable solution is the **Loop Heat Pipe (LHP)**. LHPs are passive, two-phase heat transfer devices that use the evaporation and condensation of a working fluid and capillary forces to circulate the fluid, transferring heat from the GPU (evaporator) to a large radiator panel (condenser) that rejects the heat into the deep-space heat sink [29, 30]. This passive architecture offers significant advantages critical for orbital deployment: up to 10 times longer maintenance-free lifetime, lower power consumption, and up to 30% less mass compared to active systems [31].

5.3 Data Logistics: The Hidden Uplink Problem

LLM training is a synchronous, compute-bound workload [17, 19]. While existing space infrastructure focuses on solving the *downlink* bottleneck (space-to-ground data transmission), the initial training phase introduces the **Hidden Uplink Problem**: the necessity of migrating petabytes of clean, foundational data corpora from terrestrial ground stations to orbit. This initial data migration requires dedicated, high-capacity ground-to-space links and represents a vast capital expenditure and time investment.

Furthermore, the dynamic orbital motion of LEO satellites creates the **LEO Motion Liability**. LEO constellations, despite their advantages for inference, exhibit sustained latency spikes caused by dynamic routing and reliance on constantly shifting inter-satellite links (ISLs). These instabilities are catastrophic for the strict, all-to-all synchronization protocols required for large-scale distributed training, confirming the necessity of a static GEO architecture for the most compute-intensive workloads.

6 Strategic Viability and Governance

The success of Orbital Intelligence is ultimately measured by its economic justification and its ability to operate within the emerging, fractured space regulatory framework.

6.1 Comparative Economics and Launch Debt

Global CapEx on terrestrial data center infrastructure is projected to exceed \$1.7 trillion by 2030, driven largely by AI expansion [32]. The Horizon Europe-funded Ascend study concluded that space data centers could be economically viable, projecting a return on investment (ROI) of several billion by 2050, primarily by eliminating the operational costs of carbon energy and complex cooling [3].

However, this economic viability is critically vulnerable to the environmental cost of deployment. The launch phase presents the **Launch Sustainability Achilles' Heel**. The operational carbon savings will be nullified unless the system's launch debt is drastically reduced. The Ascend study mandate requires the **development of a launcher ten times less emissive over its entire lifecycle** compared to current vehicles [3, 33]. Without this next-generation, low-emissivity launch capability, the project risks merely shifting the environmental burden from electrical OpEx to launch CapEx.

6.2 The Regulatory and Sovereignty Void

The exponential proliferation of commercial megaconstellations (projected to reach up to 100,000 LEO satellites) is rapidly outpacing the international legal framework, which is currently centered on the 1967 Outer Space Treaty (OST) [34].

- **Data Sovereignty and Residency:** Processing personal data in non-sovereign space introduces significant legal risk. Since international space law is nascent regarding data privacy, orbital operators must preemptively design systems to comply with diverse, complex domestic and global privacy systems (e.g., GDPR).
- **Orbital Debris Mitigation:** Computational constellations must adhere rigorously to international guidelines, which typically mandate a 90% success rate for post-mission disposal (PMD) [35, 36]. The sheer number of objects in these mega-constellations increases the computational requirement for continuous, complex propagation and conjunction analyses over multi-century timescales [35].
- **Resource Governance:** Orbital resources and radio spectrum, currently managed by the International Telecommunication Union (ITU) under a "First Come, First Served" principle [37], are finite. New governance mechanisms are required for the allocation and coordination of these resources to avoid conflict and harmful interference [38, 39].

7 Conclusion

The vision of Orbital Intelligence is technologically sound and strategically necessary. By exploiting the thermodynamic advantages of space—zero-carbon solar power and passive deep-space cooling—we can establish a sustainable platform that addresses the escalating environmental footprint of frontier AI. The strategic roadmap must prioritize the deployment of LEO constellations as an immediate, sustainable *inference* network (the 90% problem), while aggressively developing the radiation-hardened and thermally optimized GEO infrastructure for foundational *training* (the 10% problem). The ultimate feasibility of this transition rests on a non-negotiable strategic mandate: the development and utilization of launch vehicle technology that can achieve a tenfold reduction in lifecycle emissions, thereby ensuring the orbital solution does not simply trade one carbon burden for another.

References

- [1] Wikipedia contributors, "Large language model," *Wikipedia, The Free Encyclopedia*, September 2025, accessed: [2025-09-28]. [Online]. Available: https://en.wikipedia.org/wiki/Large_language_model
- [2] NASA LLIS, "Solar constant, albedo, and earth radiation values," NASA Lesson Learned 693, 1995. [Online]. Available: <https://llis.nasa.gov/lesson/693>
- [3] Innovation News Network, "New study highlights the viability of space data centres," *Innovation News Network*, October 2024, citing Thales Alenia Space Ascend Study. [Online]. Available: <https://www.innovationnewsnetwork.com/space-data-centres-viability-study/>
- [4] ADASci, "How much energy do llms consume? unveiling the power behind ai," *ADASci Blog*, August 2024. [Online]. Available: <https://adasci.org/llm-energy-consumption/>
- [5] Shop Without Plastic, "The carbon cost of training large ai models," *Shop Without Plastic Blog*, March 2024. [Online]. Available: <https://shopwithoutplastic.com/carbon-cost-ai-models/>
- [6] Reddit User /u/throwaway-3413, "The carbon footprint of gpt-4 is the equivalent of powering more than 1300 homes for one year!" Reddit r/LocalLLaMA, April 2024. [Online]. Available: <https://www.reddit.com/r/LocalLLaMA/comments/carbon-footprint-gpt4/>
- [7] A. Lucioni, S. Varma, and A. Awe, "The environmental impact of large language models: Training and inference," *arXiv preprint arXiv:2505.09598*, May 2025. [Online]. Available: <https://arxiv.org/abs/2505.09598>
- [8] S. Poddar, P. Koley, J. Misra, N. Ganguly, and S. Ghosh, "Towards sustainable nlp: Insights from benchmarking inference energy in large language models," *arXiv preprint arXiv:2502.05610*, February 2025. [Online]. Available: <https://arxiv.org/abs/2502.05610>
- [9] S. Hugon and T. Giraud, "Llm inference methodology," Ecologits.ai Technical Note, November 2024. [Online]. Available: <https://ecologits.ai/methodology>
- [10] Eutelsat OneWeb, "Edge computing solutions with leo connectivity," Eutelsat OneWeb Solutions, 2024. [Online]. Available: <https://oneweb.net/solutions/edge-computing>
- [11] Wikipedia contributors, "Solar constant," *Wikipedia, The Free Encyclopedia*, September 2025, accessed: [2025-09-28]. [Online]. Available: https://en.wikipedia.org/wiki/Solar_constant
- [12] Booz Allen Hamilton, "Deploying a large language model in space," *Booz Allen Hamilton Insights*, July 2024. [Online]. Available: <https://www.boozallen.com/insights/deploying-llm-space.html>
- [13] Wikipedia contributors, "Geostationary orbit," *Wikipedia, The Free Encyclopedia*, September 2025, accessed: [2025-09-28]. [Online]. Available: https://en.wikipedia.org/wiki/Geostationary_orbit
- [14] N. Al-Falahy, "Advanced communications systems for 4th class students," *University of Anbar Engineering College*, 2019.
- [15] R. M. Mahendran, S. K. Narayanan, and T. V. Pradeep, "Computation of eclipse time for leo small satellites," *International Journal of Advances in Engineering and Architecture*, vol. 3, no. 2, pp. 45–52, 2019.

- [16] B. Lin, Z. Xu, R. Li, Z. Wang, M. Chen, and Q. Zhang, “Dynamic velocity-aware resource allocation for leo edge computing networks,” *arXiv preprint arXiv:2406.10856*, June 2024. [Online]. Available: <https://arxiv.org/abs/2406.10856>
- [17] A. Pira, “Interconnect needs for llm inference to drive networking bandwidth,” *650 Group Blog*, May 2024. [Online]. Available: <https://650group.com/llm-interconnect-bandwidth/>
- [18] Dig.Watch and Space News, “China launches first ai satellites in orbital supercomputer network,” *Digital Watch*, June 2024, based on reports from Space News and MojoAuth. [Online]. Available: <https://dig.watch/updates/china-ai-satellites-orbital-supercomputer>
- [19] NVIDIA Developer, “Turbocharge llm training across long-haul data center networks with nvidia nemo framework,” *NVIDIA Developer Blog*, March 2024. [Online]. Available: <https://developer.nvidia.com/blog/turbocharge-llm-training-nemo-framework/>
- [20] Y. Min *et al.*, “Hpe spaceborne computer-3: Enabling real-time ai and edge computing for space exploration,” *bioRxiv preprint bioRxiv:2025.01.14.633017*, January 2025. [Online]. Available: <https://www.biorxiv.org/content/10.1101/2025.01.14.633017>
- [21] NASA Jet Propulsion Laboratory, “New ai algorithms streamline data processing for space-based instruments,” *NASA ESTO News*, October 2022. [Online]. Available: <https://esto.nasa.gov/news/ai-algorithms-space-instruments>
- [22] S. Barmada, O. Koudelka, and M. Pacher, “Onboard processing for advanced ai/ml space applications: Challenges and prospects,” *Aerospace*, vol. 10, no. 2, p. 101, February 2023.
- [23] Curtiss-Wright, “Radiation mitigation in space: A cots approach,” White Paper, 2025. [Online]. Available: <https://www.curtiswright.com/products-solutions/technologies/radiation-mitigation>
- [24] NASA Engineering and Safety Center, “Radiation mitigation with good fault-tolerant design practices,” NASA Technical Bulletin No. 19-01-1, 2021. [Online]. Available: https://nesc.nasa.gov/NESC_TB_19-01-1.html
- [25] NASA Technical Reports Server, “Mission radiation modeling: Gto, leo, and beo environments,” NASA, Tech. Rep. NTRS 20220011775, 2022, relevant to GTO TID exposure analysis. [Online]. Available: <https://ntrs.nasa.gov/citations/20220011775>
- [26] AMD, “Radiation defense across all orbits and beyond,” AMD Solutions Page, 2025. [Online]. Available: <https://www.amd.com/en/products/embedded-and-semi-custom-solutions/radiation-defense>
- [27] D. Karsenty and J. Mather, “The cots-capsule csee mitigation scheme: A novel approach to shielding commercial-off-the-shelf electronics in space,” *arXiv preprint arXiv:2502.04504*, February 2025. [Online]. Available: <https://arxiv.org/abs/2502.04504>
- [28] S. Barmada, M. Zannoun, and O. Koudelka, “Ai at the edge on radiation-hardened hardware: Leveraging the high performance data processor,” *arXiv preprint arXiv:2504.03680*, April 2025. [Online]. Available: <https://arxiv.org/abs/2504.03680>
- [29] J. Ku, “Loop heat pipe overview,” *SAE Technical Paper 1999-01-2007*, 1999.
- [30] S. Semenov, “Introduction to loop heat pipes,” NASA Technical Course - TFAWS 2023, 2023. [Online]. Available: <https://tfaws.nasa.gov/TFAWS23/Proceedings/TFAWS-23-024-Semenov.pdf>
- [31] Calyos TM, “Loop heat pipes: How they work and applications,” Calyos Technical Page, 2024. [Online]. Available: <https://www.calyos-tm.com/loop-heat-pipes>
- [32] E. Kok, J. Rauer, P. Sachdeva, and P. Pikul, “Scaling bigger, faster, cheaper data centers with smarter designs,” *McKinsey & Company*, September 2023. [Online]. Available: <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/scaling-bigger-faster-cheaper-data-centers>
- [33] NASA Office of Technology, Policy, and Strategy, “Space-based solar power report,” NASA Technical Report, 2024. [Online]. Available: <https://www.nasa.gov/wp-content/uploads/2024/01/otps-sbsp-report-final-tagged-approved-1-8-24.pdf>
- [34] Neuraspace, “Responsible rules in orbit: Federated governance for a crowded and congested space,” *Neuraspace Blog*, January 2025. [Online]. Available: <https://www.neuraspace.com/blog/responsible-rules-orbit-federated-governance>
- [35] S. Johnson, “Mitigation of orbital debris in low earth orbit with a focus on mega-constellations,” Master’s Thesis, University of Illinois, 2022. [Online]. Available: <https://www.ideals.illinois.edu/items/124567>
- [36] European Space Agency (ESA), “Managing mega-constellations and space debris mitigation,” ESA Clean Space Initiative, 2025. [Online]. Available: https://www.esa.int/Space_Safety/Clean_Space/Managing_mega_constellations
- [37] V. Strelets, “Sustainable space: Satellites need harmonized spectrum and more,” *ITU Hub*, February 2025. [Online]. Available: <https://www.itu.int/hub/2025/02/sustainable-space-satellites-spectrum/>

- [38] S. Hobe, “Orbital law: Governance of resources in space,” *Oxford Research Encyclopedia of Planetary Science*, 2024.
- [39] G. Rotola and A. Williams, “Regulatory context of conflicting uses of outer space: Astronomy and satellite constellations,” *Air and Space Law*, vol. 46, no. 4, pp. 465–490, 2021.