https://doi.org/10.70844/ijas.2025.2.38

Innovative Journal of Applied Science

Research Article ISSN: 3049-3072

IJAS-25-112

The Pluralistic Future of AI: A Comprehensive Analysis of Decentralized Large Language Models

Nursan Omarov*, **Alimzahn Tokushev**

Independent Researcher

Corresponding Author: Nursan Omarov, Independent Researcher, E-mail: nursan2007@gmail.com

Received date: 09 September, 2025, **Accepted date:** 16 September 2025, 2025, **Published date:** 23 September, 2025

Citation: Omarov N, Tokushev A (2025) The Pluralistic Future of AI: A Comprehensive Analysis of Decentralized Large Language Models. Innov J Appl Sci 2(5): 38.

Abstract

The Centralized Status QuoThe current landscape of Artificial Intelligence (AI), particularly concerning Large Language Models (LLMs), is dominated by a centralized paradigm. Industry leaders such as OpenAI, Anthropic and Google DeepMind deploy and manage their models on massive computational infrastructure within expansive data centers. This architecture, where a single, colossal model serves millions of users and cloud-based APIs, has enabled unprecedented scalability, high performance and the capacity for continuous updates. However, this model is not without significant drawbacks. It requires substantial financial investment in infrastructure, is entirely reliant on constant internet connectivity and introduces considerable privacy and data control concerns, as sensitive information must be processed in the cloud. The user also frequently pays for compute capacity they do not fully utilize, leading to an inefficient cost model based on subscriptions or per-token usage.

Keywords: Artificial Intelligence (AI), Large Language Model (LLM), Centralized LLM: Cloud-based large language model, Decentralized LLM: Locally deployed large language model, Data centers, Model, API

Introduction

The centralized status quo

The current landscape of Artificial Intelligence (AI), particularly concerning Large Language Models (LLMs), is dominated by a centralized paradigm. Industry leaders such as OpenAI, Anthropic and Google DeepMind deploy and manage their models on massive computational infrastructure within expansive data centers.1 This architecture, where a single, colossal model serves millions of users *and* cloud-based APIs, has enabled unprecedented scalability, high performance and the capacity for continuous updates. However, this model is not without significant drawbacks. It requires substantial financial investment in infrastructure, is entirely reliant on constant internet connectivity and introduces considerable privacy and data control concerns, as sensitive information must be processed in the cloud. The user also frequently pays for compute capacity they do not fully utilize, leading to an inefficient cost model based on subscriptions or per-token usage [1].

The decentralized alternative and the pluralistic thesis

An alternative approach is emerging: decentralized LLMs, where smaller, optimized models are deployed and executed locally on end-user devices, such as PCs, mobiles and edge hardware. This paradigm represents a fundamental inversion of the centralized model, reminiscent of the 20th-century transition from centralized mainframe computing to personal computers. The core argument of this report is that the future of AI is not a zero-sum competition between these two models but a "pluralistic" one, where both centralized and decentralized architectures coexist and serve distinct, complementary needs. Centralized systems may continue to dominate for large-scale, complex reasoning, while decentralized systems will enable a new era of personalized, private and resilient intelligence for individuals and organizations [1].

Scope and structure OFI the report

This report will first provide a foundational comparative analysis of the centralized and decentralized LLM paradigms. Following this, it will delve into the technical enablers that make on-device AI a reality, exploring model compression, optimization and distributed training frameworks. The analysis will then extend to the economic, environmental and ethical implications of this shik. Finally, the report will present a series of domain-specific applications and case studies to illustrate the real-world impact of decentralized AI and conclude with an outlook on the future research trajectory.

Foundational Concepts: Centralized *vs.* Decentralized LLM Architectures

The centralized paradigm and its bottlenecks

The architecture of centralized LLMs is defined by its scale and unified deployment. With parameter counts reaching into the trillions, these models are housed in cloud servers and are capable of complex reasoning across diverse domains. Their advantages are clear: they offer high performance and generality, benefit from continuous updates and improvements and can handle a wide array of tasks [1].

However, this model creates a performance-to-accessibility boWleneck. The immense infrastructure investment required restricts their deployment to a handful of major corporations, creating a centralized control that can stifle innovation and limit user agency. Furthermore, their dependence on continuous internet connectivity makes them unsuitable for applications in remote areas, disaster recovery scenarios, or any environment with limited bandwidth [1]. The core trade-off here is that centralized systems offer unparalleled scale and general capabilities but are constrained by their reliance on vast, expensive and internet-dependent infrastructure.

The decentralized alternative: On-device LLMs

In contrast, the decentralized paradigm deploys smaller, optimized models directly onto end-user devices [1]. The fundamental benefits of this architecture are multifold. It offers significant cost-efficiency, as users incur a one-time device-level cost rather than recurring

subscription or per-token fees. This approach also provides robust privacy, as sensitive data remains on the device and is not processed in an external cloud, mitigating the risk of data leaks or hacks [2]. The offline functionality of these models is critical for resilience and autonomy in environments with limited or no connectivity. Finally, on-device models can be fine-tuned to narrow, user-specific tasks, enabling a high degree of personalization [1]. The fundamental trade-off of this paradigm is a sacrifice of raw, general-purpose performance for ubiquity, privacy and cost-efficiency.

Comparative analysis OFI paradigms

The core dynamic driving the evolution of LLMs is the nuanced trade-off between the scale of centralized models and the accessibility of decentralized ones. The following table expands on the initial user document to provide a comprehensive comparison across several key dimensions, illustrating how these two paradigms cater to different use cases and constraints (Table 1).

Feature	Centralized LLMs	Decentralized LLMs
Scale	Trillions of parameters	Millions-billions (optimized)
Deployment	Cloud servers in data centers	Local devices (PCs, mobiles, edge)
Connectivity	Requires continuous internet	Fully offline possible
Cost model	Subscription/per-token usage	One-time, device-level cost
Privacy	Data processed in the cloud	Data remains local
Personalization	Limited (general-purpose)	High (task-specific tuning)
Governance	Centralized (corporate control)	Distributed (user/community control)

Table 1: Comparative analysis OFI paradigms.

The Technical Enablers OFI on-Device AI

The ability to run large language models on resource-constrained devices is not a simple feat but the result of significant advancements in model compression and distributed inference. These technical innovations are the foundation upon which the decentralized AI revolution is built.

Model compression and optimization

The primary challenge for on-device deployment is the sheer size and computational demand of LLMs. Model compression techniques address this by reducing memory footprint and computational requirements while minimizing performance degradation.

Quantization

Quantization is a key technique that reduces the numerical precision of a model's weights and activations, typically from 32-bit floating-point to lower bit-width representations like 8-bit integers [3]. A straighVorward approach is:

- Post-Training Quantization (PTQ), which applies quantization aker the model has been trained. This is simple to implement but can lead to performance drops for complex models [3]. A more advanced method is:
- Quantization-Aware Training (QAT), which incorporates the effects of quantization during the training process itself, allowing the model to adapt to reduced precision and maintain higher accuracy [3]. The evolution from simple PTQ to more sophisticated techniques such as:
- Outlier-Aware Quantization and Mixed-Precision Quantization demonstrates that the technical challenge is not merely reducing size but doing so without critically compromising performance [3]. This is an active area of research aimed at bridging the gap between massive models and limited hardware.

Pruning

Pruning is the process of eliminating redundant or unnecessary parameters from a neural network, thereby reducing its size and computational requirements [3]. Research has shown that up to half of a model's parameters can be removed with almost no impact on performance, highlighting the common issue of over-parameterization in large models [4]. By identifying and trimming these excess components, pruning makes models more efficient for on-device deployment.

Knowledge distillation

Knowledge distillation is a powerful technique where a large, high-performing "teacher" model transfers its knowledge to a smaller, more efficient "student" model. The student model is trained to mimic the outputs of the teacher, effectively learning the same generalized knowledge but with a significantly smaller parameter count. A prominent example is Distil BERT, which achieves a 50% reduction in model size while retaining near-equivalent performance on most tasks [4]. This process represents a direct, symbiotic link between the centralized and decentralized paradigms. The expensive, massive-scale training of a centralized model (the teacher) is leveraged to produce a lightweight, efficient on-device model (the student), perfectly illustrating how one paradigm can feed and enhance the other in a pluralistic ecosystem.

Distributed inference architectures

For models that are still too large to fit on a single device, even aker compression, distributed inference offers a solution. Tensor parallelism partitions the model's neural network tensors, such as weight matrices, across multiple edge devices for collaborative inference. While this enables the deployment of larger models, it introduces the significant challenge of communication overhead due to frequent all-reduce operations needed to aggregate intermediate outputs across devices [5].

The Distributed Paradigm: Training and Synchronization

The principles of decentralization extend beyond model deployment to the very process of training and model synchronization. This distributed approach addresses core challenges related to privacy, resource constraints and development boWlenecks.

Federated Learning (FL) FIOR privacy-preserving training

Federated Learning (FL) is a decentralized AI strategy that allows a global model to be trained across a multitude of devices without requiring the raw data to be centralized or shared. The framework operates with a central server, or aggregator, that sends out the latest model to participating devices, or workers. Each worker then uses its own local data to train and update the model. Crucially, only the model updates—not the sensitive data—are sent back to the aggregator, which then combines these updates to refine the global model. This approach harnesses the collective computational power of distributed devices, reducing the burden on a central server and minimizing data transmission costs [6].

While FL is a significant step toward privacy-preserving training, it is not a complete solution. The framework introduces new security challenges, as adversaries may exploit the shared gradients during training to infer sensitive information about the underlying data. This necessitates the use of advanced cryptographic techniques like Secure Multi-Party Computation (SMPC) and homomorphic encryption to safeguard against such threats [6]. This demonstrates that FL, while solving one problem, introduces its own set of complex research challenges that must be addressed for it to be an andble long-term solution.

Advanced distributed architectures

Decentralized mixture OFI experts (MoE)

A Mixture of Experts (MoE) is an architecture where a "router" network directs input tokens to a sparse subset of specialized "expert" sub-networks [7]. The centralized paradigm is oken constrained by the need for a massive, high-bandwidth network fabric to synchronize gradients across thousands of GPUs during training. MoE offers an elegant solution by providing an orthogonal form of parallelism that can be applied in a decentralized manner. Instead of synchronizing gradients, the training burden is partitioned across independent expert models, each trained on its own "compute island" with no cross-communication. This distributed approach to development allows for the use of scaWered, heterogeneous hardware, democratizing the creation of massive models and alle andting the systems constraints that limit centralized training runs [8].

Federated MoE Frameworks

The integration of MoE with Federated Learning creates a sophisticated, hybrid system. Frameworks like Federated Mixture of Experts (FedMix) and FedMoE-DA allow for the training of an ensemble of specialized models within an FL setup. This architecture leverages the diversity of local client data to train specialized experts, enhancing both robustness and personalizability while maintaining privacy [9]. The combination of FL, which provides a mechanism for privacy-preserving generalization and MoE, which provides a mechanism for privacy-preserving specialization, creates a powerful system that can collectively learn from diverse data sources while maintaining individual data control and fostering domain-specific expertise [10].

Economic and Environmental Implications

Economic democratization OFI AI

The shik toward decentralized LLMs has profound economic implications. By eliminating recurring API costs, it significantly lowers the barrier to entry, allowing individuals and small businesses to leverage advanced AI without relying on corporate APIs [1]. This shik from a B2C (Business-to-Consumer) API-as-a-service model to a P2P (Peer-to-Peer) marketplace model is creating new ecosystems. PlaVorms like SingularityNET and BiWensor are emerging as decentralized marketplaces where AI models and datasets can be bought, sold and collaboratively developed. Autonomous AI agents are already transforming the Web3 economy by optimizing trades, managing liquidity pools and executing financial operations without human oversight [11].

However, the decentralized AI economy is still in its nascent stages and is subject to "considerable hype and misinformation" [12]. The presence of fraudulent projects and scams that lead to "rug pulls" is a significant concern [11]. This highlights the need for a balanced perspective that acknowledges the immense potential of this new economic model while also recognizing its speculative and risky nature.

The AI energy challenge

The widespread adoption of AI is not sustainable under the centralized paradigm due to its immense energy footprint. Data centers, primarily powered by energy-intensive GPUs, already consume more electricity than entire nations and are projected to double their energy use to 500 TWh by 2027. A single query on a centralized model can consume approximately 2.9 Wh of electricity, roughly 10 times more than a standard Google search [13]. The cumulative energy consumption from continuous inference across millions of users far exceeds the energy used in a single training run [14].

On-device AI presents a crucial solution to this environmental crisis. By processing data locally, it eliminates the need for energy-intensive data transmission to and from distant data centers. The use of specialized, energy-efficient chips for local processing results in a dramatic 100- to 1,000-fold reduction in energy consumption per task compared to cloud-based AI [13]. The technical enablers of model compression discussed earlier are directly linked to this environmental benefit, making the push for on-device AI not just a technological refinement but a strategic imperative for the industry's sustainability.

Ethical and Societal Considerations

Privacy and data confidentiality

The most prominent ethical benefit of decentralized LLMs is their ability to preserve privacy. When a model runs locally, sensitive professional or personal data never leaves the device, ensuring confidentiality [2]. This stands in stark contrast to centralized models, where corporate policies regarding data retention and privacy can change overnight and oken do not guarantee the confidentiality of data submiWed through chatbot interfaces, even for paid subscription plans [15].

The challenge OFI bias and validation

While local LLMs offer a privacy advantage, they are not immune to ethical challenges. They inherit biases—including gender, racial and socio-economic prejudices—from the foundational datasets on which they were trained. These biases can be inadvertently perpetuated and amplified in the model's outputs [16].

Furthermore, the power of local fine-tuning, while a tool for personalization, also carries the risk of introducing or reinforcing new biases based on the user's specific data [17]. This raises a critical question about accountability: With a centralized model, the responsibility for a biased output can be traced to a single corporate entity. With a decentralized, locally fine-tuned model, accountability becomes distributed and ambiguous. The report suggests that a critical and unresolved ethical and legal challenge is determining who is responsible for a harmful output—the foundational model developer, the fine-tuning plaVorm, or the end-user. This underscores the need for "trust and validation" in locally fine-tuned models, especially for sensitive applications like healthcare or law [1].

Societal polarization and echo chambers

Centralized social media plaVorms and their AI-driven recommendation algorithms, motivated by a profit imperative to maximize user engagement, are known to create "filter bubbles" and "echo chambers" [18]. These systems reinforce users' existing beliefs and can be instrumental in spreading misinformation and even radicalization [19].

The rise of personalized, on-device AI introduces a new dimension to this problem. A personal AI could theoretically act as an agent of change, proactively exposing a user to diverse viewpoints to counter the filter bubble effect [20]. However, it could also be fine-tuned by the user to align perfectly with their existing biases, creating a more potent, self-directed "personal echo chamber" that is far more difficult to govern or mitigate [18]. The problem shiks from being an external, corporate-controlled issue to a personal, user-controlled one, which is a far more complex and nuanced ethical dilemma.

The impact on human cognition

A broader societal risk of pervasive AI is the potential for cognitive offloading, where humans outsource complex problem-solving and critical thinking to AI systems [21]. Research has already revealed a "significant negative correlation between frequent AI tool usage and critical thinking abilities," suggesting that an over-reliance on these systems may come at a cognitive cost [22]. On-device AI, being always available and highly personalized, could exacerbate this trend, raising a fundamental question about the trade-off between efficiency and intellectual autonomy in a future where every person carries their own intelligent system.

Domain-Specific Applications and Case Studies

The following table provides a summary of real-world applications and case studies that highlight the practical benefits and challenges of decentralized AI across various domains (Table 2).

Domain	**Specific application**	**Benefit**	**Key technology**	**Example/case study**
Healthcare	Offline diagnostic assistants for rural areas	Resilience, privacy, cost-efficiency	On-device LLMs, Federated Learning (FL)	OfflineMedics [1], MONAI, FedMRG framework [23]
LegalTech	Local document validation and draking	Privacy, cost-efficiency, speed	On-device LLMs	LexiHK [24]
Education	Personalized offline tutors and assistants	Resilience, personalization	On-device LLMs, Retrieval-Aug mented Generation (RAG)	Khanmigo, Edukapi [25]
Aerospace	Offline AI for problem-solving	Resilience, autonomy	On-device LLMs	Astronauts relying on offline AI [1]

Table 2: Domain-specific applications and case studies.

Healthcare

Decentralized AI is poised to revolutionize healthcare. On-device LLMs can act as diagnostic assistants in rural or remote areas with limited or non-existent internet connectivity, providing crucial support in emergency situations [1]. AI models can interpret medical imaging, detect bone fractures and triage patients with greater speed and accuracy than humans in many cases [26].

A particularly compelling application is the use of Federated Learning in medical research. This framework enables collaborative model training across multiple hospitals without the need to share sensitive patient data, which is oken restricted due to privacy concerns [27]. Projects like MONAI (Medical Open Network for AI) and the FedMRG framework are demonstrating how this technology can build more robust models from diverse datasets, particularly for rare diseases, while maintaining data confidentiality [23].

LegalTech

For small law firms, where time and resources are limited, local LLMs offer significant advantages. These models can automate repetitive tasks such as document summarization, classification and initial draking of legal documents, freeing up lawyers to focus on high-value, critical thinking tasks [28]. Local deployment ensures that sensitive client information never leaves the firm's devices.

A key case study demonstrating this trend is the development of LexiHK, a fine-tuned local LLM for legal document assistance developed by the Hong Kong Department of Justice [24]. This demonstrates a policy-driven move toward the adoption of local, domain-specific models to enhance efficiency and security within the legal sector.

Education

On-device LLMs can serve as personalized, offline tutors that adapt to each student's unique learning style and pace [1]. These tools can provide real-time feedback, assist with homework and streamline administrative tasks for teachers, such as creating lesson plans and rubrics [29].

While the theoretical promise of a fully offline, private AI tutor is strong, an examination of real-world applications reveals a more nuanced reality. For example, while Khanmigo offers engaging, on-topic tutoring other plaVorms like [25].

Flexi requires an internet connection and applications like edukapi may still collect user data despite the promise of on-device functionality [30]. This illustrates a crucial point: the clear theoretical distinction between centralized and decentralized models is oken blurred in practice, with many commercial products operating as hybrids that leverage both local and cloud-based components to deliver their services [31-44].

Conclusion and Future Directions

The report concludes that the future of AI is not a simple choice between centralized or decentralized systems but a "pluralistic" coexistence where each paradigm addresses unique needs and constraints. Centralized models will continue to be the engine for large-scale, general-purpose intelligence, while decentralized models will democratize access, ensure privacy and enable resilience in a new era of on-device, personalized intelligence.

However, several challenges remain. The performance gap between on-device and centralized models, while shrinking due to advancements in compression and optimization, has not yet been fully closed. The technical complexity of distributed training, particularly in federated learning and decentralized MoE architectures, introduces new security and synchronization challenges that require ongoing research. On a societal level, the shik in ethical responsibility from a centralized corporate entity to a distributed network of users creates new legal and ethical dilemmas that are yet to be resolved. Finally, the broader societal impact on human cognition and the potential for a new form of personalized, self-directed echo chamber must be carefully navigated.

Despite these challenges, the trajectory is clear. The report reinforces the historical analogy to the personal computer revolution, where centralized mainframes gave way to a distributed computing model. A similar shik is emerging in AI, promising a future where every person can carry their own intelligent system, privately and cost-effectively, without dependence on external servers. This vision offers a more sustainable, equitable and autonomous technological future.

Conflict of interest

The author declares no conflict of interest.

References

1. Untitled.pdf
2. Offline AI made easy: How to run large language models locally. 2025.
3. Mulki R (2025) LLM optimization: Quantization, pruning and distillation. Medium.
4. Pachipulusu SC (2025) Quantization, distillation and pruning. Medium.
5. Zhang K, He H, Song S, Zhang J, Letaief KB (2025) Distributed on-device LLM inference with over-the-air computation. arXiv. [Crossref] [GoogleScholar]
6. Cheng Y, Zhang W, Zhang Z, Zhang C, Wang S, et al. (2024) Towards federated large language models: Motivations, methods and future directions. IEEE Communications Surveys & Tutorials. [Crossref] [GoogleScholar]
7. Dhanush K (2025) Mixture-of-Experts (MoE) Models in AI. Medium.
8. McAllister D, Tancik M, Song J, Kanazawa A (2025) Decentralized diffusion models. InProceedings of the Computer Vision and Pattern Recognition Conference 23323-23333. [Crossref] [GoogleScholar]
9. Reisser M, Louizos C, Gavves E, Welling M (2021) Federated mixture of experts. Semantic Scholar. [GoogleScholar]
10. Zec EL, Martinsson J, Mogren O, Sütfeld LR, Gillblad D (2020) Federated learning using mixture of experts. OpenReview. [GoogleScholar]
11. How will decentralized AI affect big tech?. (2025) Built In.
12. Sharma R (2025) Understanding Decentralized Finance (DeFi): Basics and functionality. Investopedia.
13. How on-device AI could help us to cut AI's energy demand. (2025) World.
14. Energy efficiency in AI models: Strategies for a sustainable future. (2025) Skymod.
15. Steel A (2025) Local LLMs: Ethical, secure and sustainable AI. La Linterna del.
16. Ethical implications and challenges of using language models. (2025) GeeksforGeeks.
17. Lee JK, Chung TM (2024) Detecting bias in large language models: Fine-tuned KcBERT. In International Conference on Pattern Recognition and Artificial Intelligence. Singapore: Springer Nature Singapore. [Crossref] [GoogleScholar]
18. Human Rights and Business (2025) Echo chambers and recommendation algorithms: Who decides what we see online?.
19. Algorithmic radicalization. (2025) Wikipedia.
20. Trapped in a social media echo chamber? A new study reveals how AI can offer an escape.
21. Fernández M, Keferstein M, Thomson B (2025) AI and society: Implications for global equality and quality of life.
22. Gerlich M (2025) AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies 15(1): 6. [Crossref] [GoogleScholar]
23. Che H, Jin H, Gu Z, Lin Y, Jin C, et al. (2025) LLM-driven medical report generation *and* communication-efficient heterogeneous federated learning. IEEE Transactions on Medical Imaging 1-1. [Crossref] [GoogleScholar]
24. LCQ5: Application of legal technology and artificial intelligence.
25. Meet khanmigo: Khan academy's AI-powered teaching assistant & tutor.
26. 7 ways AI is transforming healthcare. (2025) The World Economic Forum.
27. Decentralized AI: What Are the Advantages for The Healthcare Industry?. Alcimed.
28. Gavel (2025) Small law firm AI guide: Using LLMs in 2025.
29. Shapel M (2025) Top 10 AI-powered learning experience platforms in 2025. SaM Solutions.
30. Edukapi (2025) Your AI tutor 24/7 - apps on google play.
31. Dilmegani C (2025) Federated learning: 5 use cases & real-life examples. Research AIMultiple.
32. Prebys J (2025) What is decentralized AI? A beginner's guide to blockchain-powered intelligence. Polkadot.
33. Mixture of experts (2025) Wikipedia.

34. Parallelism and distributed training for maximizing AI efficiency. (2025) Exxact Blog.
35. AI+Blockchain: The most promising projects to watch in 2025. (2025) Walbi Blog.
36. Large language model. (2025) Wikipedia.
37. We'd like to use additional cookies to understand how you use the site and improve our services. (2025) UK Parliament Committees.
38. AI doc (2025) Clinical AI solutions for healthcare providers.
39. Rauniyar A, Hagos DH, Jha D, Håkegård JE, Bagci U, et al. (2023) Federated learning for medical applications: A taxonomy, current trends, challenges and future research directions. IEEE Internet of Things Journal 11(5): 7374-7398.
40. MONAI (2025) Medical open network for AI.
41. Personalized learning with AI: Transforming education. HP® Tech Takes.
42. LLM in education-the secret to smarter and personalized learning. (2025) Matellio Inc.
43. SchoolAI (2025) Reimagining Student Success.
44. Flexi - A free science and math AI tutor for every student - CK-12.