Edge AI using Ultra Low Bit LLMs

OneBit AI

January 16, 2026

Anirudh Malik, M Sparsh Mehra, Vishnuprasad Pradeepkumar Reshma
Apoorv Pundir, Harsh Tyagi

A new approach to low-bit AI models and edge-native deployment for global & private environments.

Abstract

Recent advances in large language models (LLMs) have been driven primarily by scale, high numerical precision, and centralised cloud-based infrastructure. While this paradigm has enabled impressive benchmark performance, it introduces structural constraints related to cost, energy consumption, latency, privacy, and accessibility. These constraints limit the deployability of AI systems in bandwidth-constrained, energy-limited, or privacy-sensitive environments.

In this paper, we present **OneBit AI**, an offline-first, edge-native intelligence system based on ultra-low-bit model execution. By operating in a near-binary precision regime, OneBit AI significantly reduces memory footprint, energy consumption, and inference latency while maintaining useful task-level performance. We demonstrate that meaningful language modelling and reasoning capabilities can be achieved without reliance on cloud GPUs or continuous internet connectivity. Through system-level co-optimisation of model architecture, inference runtime, and deployment hardware, OneBit AI enables deterministic, low-latency inference on commodity CPUs and ARM-based devices. Empirical comparisons against compact FP16 and INT8 models show order-of-magnitude improvements in energy efficiency, substantial reductions in memory bandwidth requirements, and faster time-per-output-token (TPOT) in edge settings.

These results suggest that high-precision, cloud-centric AI is not a prerequisite for practical intelligence. Instead, ultra-low-bit, offline-first architectures represent a viable and scalable path toward democratising AI access across global, cost-sensitive environments.

The Problem: AI Is Becoming Structurally Centralised

Compute Centralisation

Over the past decade, artificial intelligence has shifted from a distributed computing paradigm to a highly centralised one. Modern AI systems are increasingly designed around large-scale infrastructure, relying on massive GPU clusters, continuous internet connectivity, and centralised cloud APIs for both training and inference.

This architectural shift has enabled rapid advances in model capability, but it has also introduced structural dependencies that were not present in earlier generations of software. AI workloads are now tightly coupled to a small number of cloud providers and hardware vendors, making access to intelligence contingent on network availability, pricing models, and external infrastructure.

As a result, three systemic constraints emerge. *First*, cost scales directly with usage, as inference and API calls accumulate variable expenses rather than behaving like fixed infrastructure costs. *Second*, data must leave the local environment to be processed, introducing privacy, compliance, and security concerns across regulated industries. *Third*, access to AI becomes uneven across geographies and sectors, favouring regions with robust connectivity, capital availability, and electrical capacity while excluding environments where such infrastructure is impractical or unavailable.

These constraints are not accidental side effects; they are a direct consequence of an AI ecosystem optimised for centralised compute rather than for deployment diversity.

Energy as a Structural Constraint

Beyond monetary cost, modern large language models impose significant energy demands. Training state-of-the-art models routinely consumes hundreds of megawatt-hours per run, while large-scale inference deployments require sustained power draw comparable to small industrial facilities.

As AI adoption increases, aggregate energy consumption scales with both model size and usage intensity—a trajectory effectively flattened by the efficiency gains demonstrated in Figure 1.

This energy profile introduces a new class of deployment limitations. High-power computing concentrates AI infrastructure in regions with access to inexpensive electricity, advanced cooling, and dense data centre capacity. In many parts of the world, electrical availability rather than computing demand becomes the binding constraint for AI deployment. As inference workloads expand globally, the energy requirements of current AI architectures risk becoming a bottleneck to both scalability and sustainability.

If left unaddressed, this trajectory reinforces centralisation, increases environmental load, and restricts AI access to a narrow subset of locations capable of supporting its energy footprint [1] [2].

Where Cloud AI Breaks Down

While cloud-first AI systems perform well in well-connected, well-capitalised environments, they struggle in a wide range of real-world conditions. Budget-constrained organisations often find usage-based pricing unpredictable and difficult to sustain over time, even when the economic upside of AI adoption is clear. Bandwidth-limited regions face degraded performance or complete inaccessibility when connectivity is unreliable. Latency-sensitive workflows, such as real-time decision-making or human-in-the-loop systems, suffer from round-trip delays inherent to remote inference.

In privacy-critical domains, including education, research, healthcare, and regulated enterprise environments, transmitting sensitive data to external servers is either undesirable or prohibited [3]. Similarly, offline or air-gapped systems are common in industrial, research, and secure deployments and cannot rely on cloud APIs by design.

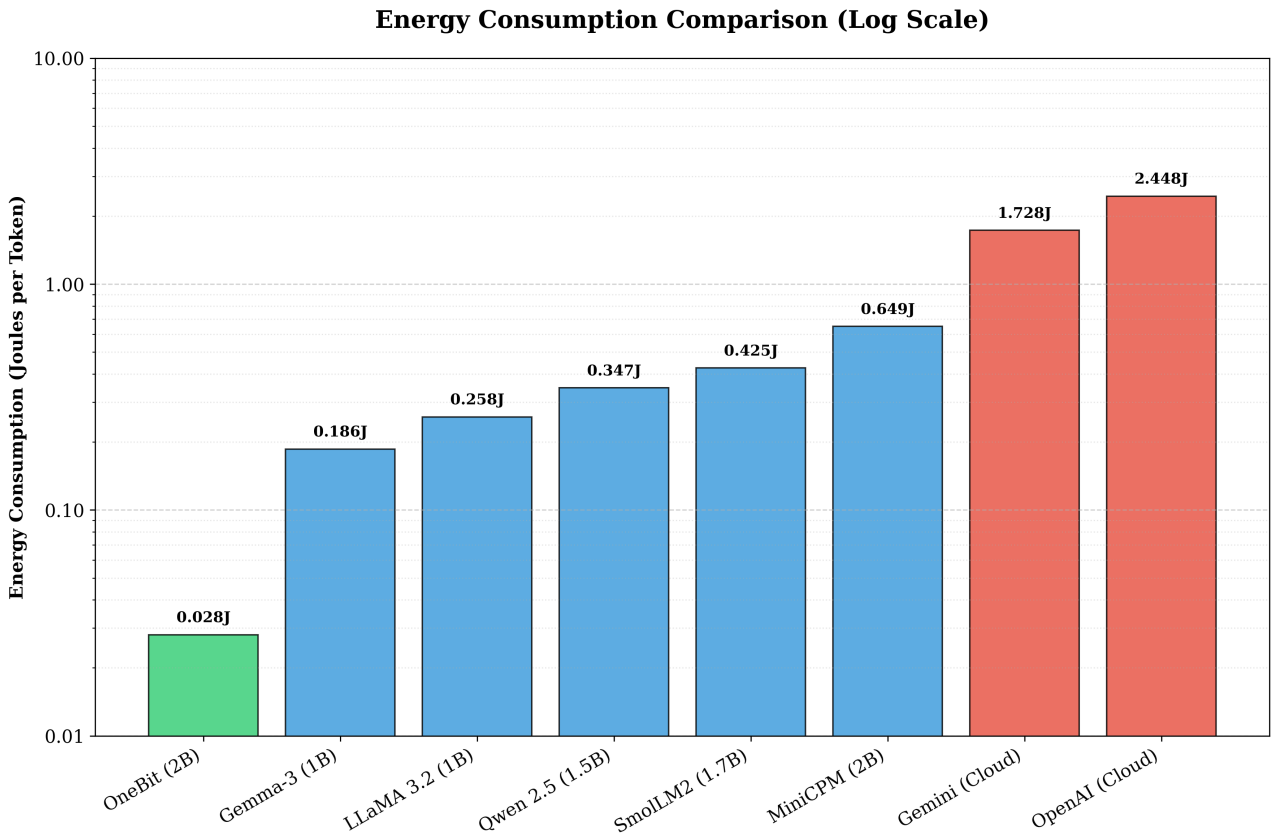

Figure 1: **Energy Consumption Benchmark (Log Scale).** A comparative analysis of energy expenditure per token in Joules. OneBit (0.028J) demonstrates an order-of-magnitude efficiency advantage over standard compact models and a >80x reduction in power requirements compared to cloud-based baselines.

These limitations do not represent niche edge cases. They describe a substantial portion of global compute environments, encompassing billions of users, devices, and workflows that are structurally underserved by cloud-only AI architectures.

Insight: Intelligence Does Not Require High Precision Everywhere

Precision Is a Design Choice, Not a Law

The dominant assumption in modern AI development is that higher numerical precision and specialised hardware are prerequisites for useful intelligence. Most large-scale models are designed and evaluated using FP16 or INT8 precision, with GPU acceleration treated as a baseline requirement and cloud deployment assumed as the default execution environment.

While this approach maximises performance on standardised benchmarks, it conflates peak capability with practical utility. In many real-world tasks—such as assistance, summarisation, classification, and interactive reasoning—full floating-point precision is not strictly necessary to produce useful outcomes.

Precision, therefore, is not a law of intelligence but a design choice. Treating it as such enables deliberate trade-offs between marginal accuracy gains and substantial improvements in cost, energy efficiency, and deployability.

Ultra-Low-Bit Models as an Efficiency Lever

Recent research and system-level experimentation indicate that meaningful reasoning behaviour can persist even as numerical precision is significantly reduced [4]. At the same time, empirical analysis shows that inference cost is often dominated not by arithmetic operations, but by memory bandwidth, data movement, and energy consumption.

By operating in an ultra-low-bit regime, these bottlenecks are reduced. Model weights become smaller, memory access becomes cheaper, and execution can shift from specialized accelerators to general-purpose CPUs and ARM-based processors. At sufficiently low precision, CPU-first inference becomes viable, removing the hard dependency on GPUs and high-power compute infrastructure [5].

This shift enables a different design objective. Instead of optimising solely for peak benchmark performance, AI systems can be optimised for efficiency, locality, and operational resilience—delivering intelligence within the physical, economic, and energy constraints of real-world environments.

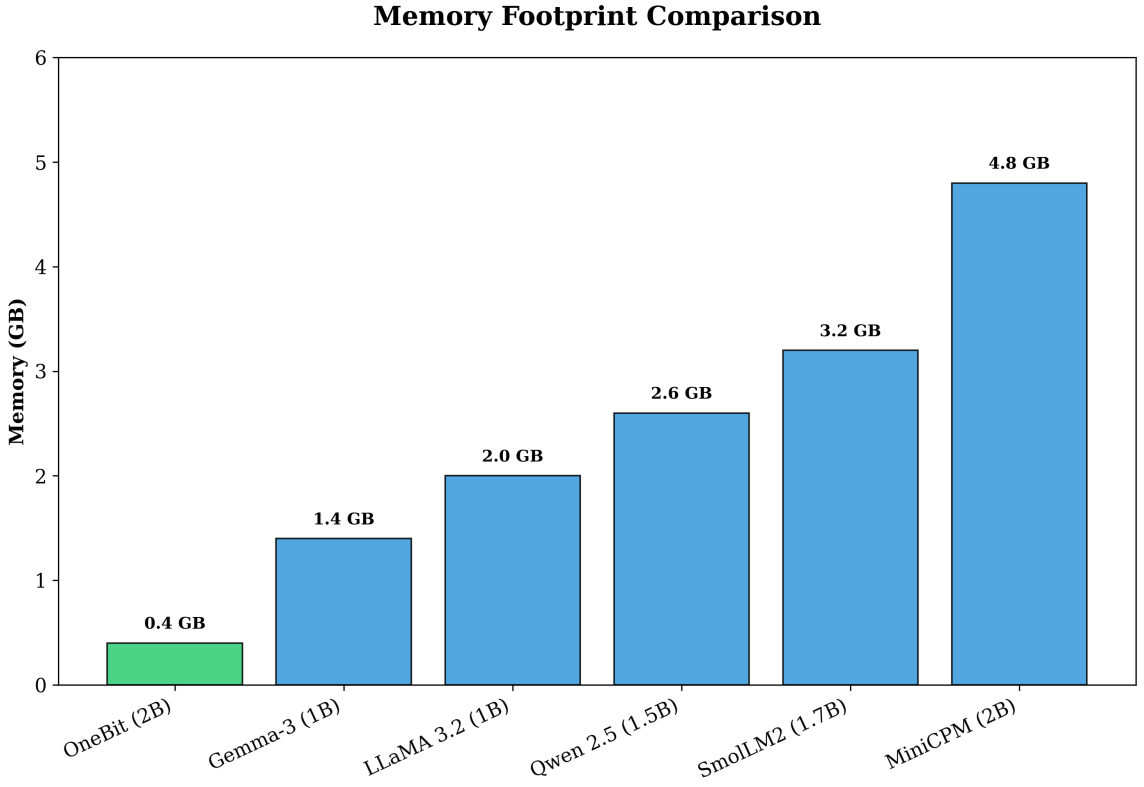

Figure 2: **Edge Model Memory Footprint Comparison.** A comparison of RAM requirements for inference. The OneBit (2B) model requires only 0.4 GB, fitting comfortably within the constraints of standard IoT and mobile devices, whereas comparable models demand significantly higher memory overhead.

This architectural efficiency is quantified in Figure 2, where the OneBit architecture requires

a fraction of the RAM demanded by comparable industry models, ensuring viability even on highly constrained hardware.

OneBit AI: System Overview

OneBit AI represents a fundamental departure from the 'Cloud-First' AI paradigm. Rather than treating AI as a remote service, we treat it as a local utility in which the system is designed as a vertically integrated stack where the model architecture, the inference logic, and the physical hardware are co-optimised.

Core Principles

To meet the rigorous demands, the OneBit ecosystem is governed by five pillars

- Offline-first
  Ensuring zero-latency and operational continuity in 'Denied, Disconnected, Intermittent, and Limited' environments.
- CPU-compatible
  By moving away from floating-point math to integer-based logic, we unlock high-speed intelligence by bypassing the 'GPU bottleneck'. This enables OneBit AI to deliver real-time performance on CPUs, turning standard processors into powerful AI engines without the energy or cost overhead.
- Edge-native
  The software footprint is minimised to run on the metal of ARM-based SBCs (Single Board Computers) without the overhead of heavy virtualisation layers.
- Privacy-preserving
  OneBit's localised processing model ensures that sensitive information never leaves the host hardware, providing a definitive solution to metadata leakage and securing the data supply chain against external eavesdropping.
- Cost-efficient by design
  We leverage existing hardware cycles, removing the need for recurring API costs or specialised GPU cooling infrastructure.

Mathematical Formalism: The Move to Integer Addition

The core innovation of the OneBit architecture is the replacement of standard `nn.Linear` layers with `BitLinear` layers. This transition is achieved through a specific quantization strategy that constrains weights to a ternary set $\mathcal{W} \in \{-1, 0, 1\}$, effectively changing the fundamental arithmetic of the network [6].

AbsMean Weight Quantization

Unlike standard post-training quantisation, OneBit models are trained using *Absolute Mean Quantization*. For a given weight matrix $W \in R^{n \times m}$, we first compute a scaling factor γ based on the average absolute magnitude of the weights:

$$\gamma = \frac{1}{nm} \sum_{i,j} |W_{ij}| \tag{1}$$

The weights are then quantised into ternary values by scaling, rounding, and clipping:

$$W_{quant} = \text{Clip}\left(\text{Round}\left(\frac{W}{\gamma + \epsilon}\right), -1, 1\right) \tag{2}$$

This forces every element in W_{quant} to strictly adopt a value from $\{-1, 0, 1\}$.

The BitLinear Operation

In a standard Transformer, the output Y is computed via matrix multiplication $Y = W \cdot X$, which requires expensive floating-point multiplication-accumulate (MAC) operations.

In the OneBit architecture, inputs X are quantised to 8-bit integers (X_{quant}) using AbsMax scaling (β). The linear operation is then reformulated to bypass floating-point multiplication entirely:

$$Y = \underbrace{(W_{quant} \otimes X_{quant})}_{\text{Integer Addition Only}} \times \frac{\beta\gamma}{Q_b} \tag{3}$$

Where:

- $W_{quant} \in \{-1, 0, 1\}$ transforms the tensor product $\otimes$ into simple addition and subtraction.
- β and γ are scalar values applied *after* the heavy computation, preserving precision without slowing down the matrix kernel.

This formulation proves that the computational complexity shifts from $O(N \cdot \text{FP}_{\text{mul}})$ to $O(N \cdot \text{INT}_{\text{add}})$, resulting in the energy efficiency gains illustrated in Figure 1.

OneBit AI: System Overview

OneBit AI represents a departure from the prevailing cloud-first paradigm in artificial intelligence. Instead of treating intelligence as a remote service accessed through centralised APIs, OneBit AI approaches AI as a local utility. The system is designed as a vertically integrated stack in which model architecture, inference logic, and deployment hardware are co-optimised for efficient, on-device execution.

Rather than optimising for peak benchmark performance in controlled environments, OneBit AI prioritises deployability, predictability, and efficiency across real-world constraints. This design philosophy enables intelligence to operate independently of continuous network connectivity or specialised accelerators.

Core Principles

The OneBit ecosystem is governed by five core principles that guide system design and deployment:

- **Offline-first**
  The system is designed to operate without reliance on continuous internet connectivity, ensuring consistent performance in disconnected, intermittent, or bandwidth-limited environments.
- **CPU-compatible**
  By shifting away from **floating-point-heavy** execution toward integer-dominant computation, OneBit AI enables efficient inference on general-purpose CPUs. This approach removes the dependency on GPUs while reducing energy consumption and operational cost.
- **Edge-native**
  The software stack is engineered to run directly on x86 or ARM based single-board computers and commodity processors without the overhead of large virtualisation layers or containerised cloud runtimes.
- **Privacy-preserving**
  All inference is performed locally. Sensitive data remains within the execution environment, minimizing exposure to external networks and reducing risks associated with data transfer and metadata leakage.
- **Cost-efficient by design**
  OneBit AI leverages existing compute resources and avoids recurring API usage fees, specialised cooling infrastructure, and cloud compute overhead.

The OneBit Stack

System efficiency in OneBit AI emerges from coordinated optimisation across four layers:

1. **Low-bit model architecture**
   Models are trained to operate in an ultra-low-bit regime suitable for execution on commodity CPUs and ARM systems. Reduced numerical precision allows replacement of expensive floating-point operations with integer arithmetic, enabling model weights to reside in cache or system memory and minimising memory bandwidth bottlenecks.
2. **Inference engine**
   A custom runtime optimised for memory efficiency and predictable latency. The engine minimises data movement through a zero-copy execution model and deterministic scheduling, ensuring stable performance independent of network conditions.
3. **Deployment layer**
   A hardware-abstraction layer that enables consistent execution across desktops, ARM SBCs, mobile devices, and plug-and-play edge hardware. Whether deployed on a portable AIR unit, an embedded PRO board, or a clustered MAX configuration, the system utilises available SIMD instructions to maximise CPU throughput.
4. **Application layer**
   User-facing interfaces delivered as offline mobile applications, local web interfaces, or em-

bedded integrations. These interfaces support tasks such as document analysis, translation, and real-time inference without requiring external connectivity.

Architecture Philosophy (Without Revealing IP)

The architectural philosophy of OneBit AI is rooted in the belief that intelligence should adapt to deployment constraints rather than dictate them.

Why We Avoid Cloud Dependence

Cloud-based AI systems are optimised for centralised throughput, multi-tenant efficiency, and vendor-controlled execution. While effective in certain contexts, this architecture introduces structural limitations for environments that require predictability, locality, and autonomy.

First, dependence on continuous connectivity introduces a single point of failure. Any network disruption renders cloud-dependent intelligence unavailable. Second, cloud inference introduces non-deterministic latency due to shared infrastructure and network variability. Third, transmitting data to external servers increases exposure to interception, traffic analysis, and compliance risk.

OneBit AI avoids these trade-offs by executing inference locally. All compute resources are dedicated to the end user and sensitive data never leaves the execution environment. Crucially, this approach ensures that latency remains stable and highly competitive, as demonstrated by the benchmark data in Figure 3.

CPU-Only as a Strategic Advantage

Designing for efficient CPU-only execution provides a set of structural advantages that extend beyond cost reduction. GPUs are optimised for throughput under highly parallel workloads, but they introduce dependencies on specialised hardware supply chains, high power envelopes, and centralised deployment models. By contrast, CPUs are ubiquitous, well-understood, and deeply integrated into existing infrastructure across consumer, enterprise, and industrial environments.

Eliminating reliance on GPUs removes a major source of supply volatility and capital expenditure. CPU-based systems can be provisioned, replaced, and scaled using commodity hardware already present in most organisations. This dramatically expands the addressable deployment surface, enabling AI inference on desktops, laptops, servers, single-board computers, and embedded systems without modification to underlying infrastructure.

CPU-first execution also aligns naturally with regulated and privacy-sensitive environments. Many such deployments explicitly restrict the use of external accelerators, cloud-based inference, or opaque runtime dependencies. A CPU-only model simplifies compliance, auditability, and long-term maintainability.

Finally, CPU-centric architectures offer margin stability over time. As general-purpose processors continue to improve incrementally and energy efficiency increases, inference cost trends flatten rather than scale unpredictably with usage. This contrasts with cloud-based GPU pricing models, where marginal cost remains tied to vendor-controlled compute markets.

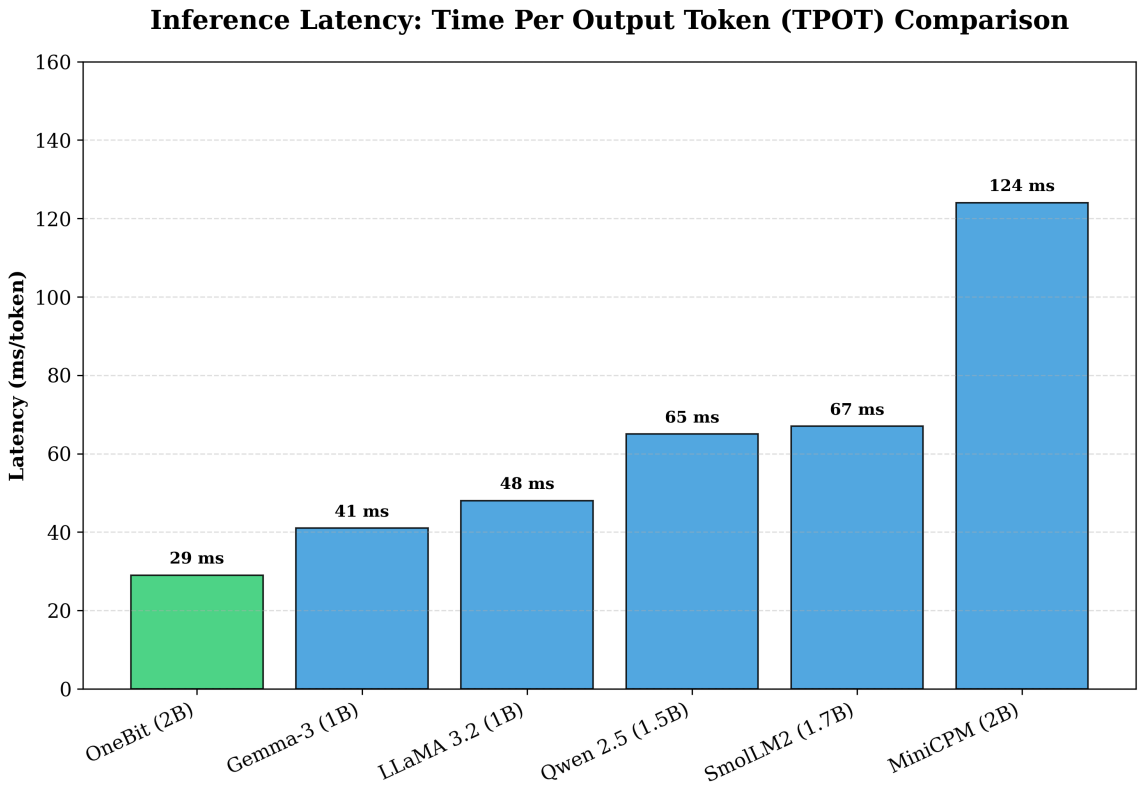

Figure 3: **Inference Latency Benchmark.** A comparison of Time Per Output Token (TPOT). With a latency of just 29ms, the OneBit architecture eliminates network variability, delivering real-time responsiveness that outperforms standard baselines by a factor of 2x to 4x.

Performance Snapshot

Performance evaluation in OneBit AI prioritises deployability and efficiency rather than peak benchmark scores. Without disclosing proprietary implementation details, current system measurements demonstrate that useful, real-time inference is achievable on modern consumer CPUs and ARM-based platforms.

Benchmarks indicate stable token generation rates suitable for interactive workloads, consistent latency across execution environments, and predictable performance under offline conditions. Importantly, system behaviour remains deterministic regardless of network availability or external load, a property that is difficult to guarantee in shared cloud environments.

Performance is assessed along three primary dimensions: energy consumption per inference, cost per inference over time, and responsiveness under constrained conditions. This framing reflects real-world deployment priorities more accurately than aggregate throughput metrics alone.

Products and Deployment Paths

OneBit AI is delivered through both software and hardware pathways, designed to accommodate a wide range of deployment constraints without altering the core execution model.

Software

The software stack is delivered as offline-first applications capable of running entirely on local hardware. Models are downloadable and executable without mandatory cloud connectivity, enabling persistent operation in disconnected or privacy-sensitive environments.

Applications are exposed through lightweight local interfaces, including native mobile applications and local web-based user interfaces. This approach allows users to interact with models using familiar tools while maintaining full control over execution and data locality.

Hardware

For environments requiring dedicated or embedded compute, OneBit AI supports a family of purpose-built hardware configurations:

- **AIR**: A portable, single-device edge compute unit designed for rapid deployment and personal or field use.
- **PRO**: An embedded or PCB-level configuration intended for integration into existing systems or products.
- **MAX**: A scalable configuration composed of multiple PRO units networked together to form a local inference cluster.

All hardware configurations are designed to execute models locally and independently. They are not thin clients and do not rely on remote inference services for core functionality.

Initial Go-To-Market Strategy

The initial go-to-market strategy focuses on environments that value autonomy, experimentation, and cost predictability over raw scale.

Early Pilots

Early deployments target educational institutions, research organisations, and developer communities. These environments provide a combination of real-world usage, technical feedback, and tolerance for early-stage systems.

Pilot deployments are structured to validate performance under diverse conditions, refine usability, and identify task categories where offline-first intelligence provides immediate value.

Expansion Path

Following pilot validation, deployment expands toward enterprise and infrastructure-adjacent environments where cost, latency, or privacy constraints make cloud-first AI impractical. Regulated sectors and bandwidth-limited deployments represent natural extensions of the initial market, as they share similar structural requirements around data locality and operational independence.

Moat: Why This Is Hard to Copy

The defensibility of OneBit AI does not arise from a single algorithmic breakthrough or isolated optimisation. Instead, it emerges from system-level integration across multiple layers of the AI stack. Each layer reinforces the others, creating a compound advantage that is difficult to replicate without rebuilding the system end-to-end.

First, OneBit AI is designed through the co-optimisation of model architecture, inference engine, and deployment environment. Decisions about numerical precision, memory layout, execution order, and runtime behaviour are made jointly rather than in isolation. This contrasts with cloud-native AI systems, where models, runtimes, and deployment infrastructure are developed independently and integrated post hoc.

Second, the system is optimised for CPU-first execution rather than GPU acceleration. This requires a fundamentally different approach to performance engineering, including careful control of memory bandwidth, cache locality, and instruction-level efficiency. Such optimisations do not transfer directly from GPU-centric stacks and require specialised expertise in low-level systems engineering.

Third, OneBit AI is built around an offline-first application architecture. Local execution, predictable latency, and independence from external services are treated as primary design constraints rather than optional features. Retrofitting offline capability into cloud-dependent systems introduces architectural friction that often negates performance and cost benefits.

Fourth, the tight coupling between software and purpose-built hardware introduces an additional layer of defensibility. While each component can operate independently, the full efficiency gains emerge only when models, runtimes, and hardware are deployed together. This coupling creates learning effects in manufacturing, thermal design, and deployment that compound over time.

Finally, operating in an ultra-low-bit regime imposes a steep learning curve. Accumulated experience in training stability, error propagation, and system behaviour at low precision constitutes tacit knowledge that is difficult to acquire through replication alone.

For these reasons, OneBit AI is not a feature that can be trivially added to an existing cloud-native stack. It represents a distinct architectural path optimised for efficiency, locality, and deployment diversity.

Open Research and Responsibility

OneBit AI recognises the importance of open research and responsible development in advancing the broader AI ecosystem. Where appropriate, we support research collaboration, transparent evaluation methodologies, and engagement with the academic and developer communities.

At the same time, not all system-level optimisations are disclosed. Certain implementation details remain proprietary to ensure long-term sustainability, operational security, and the ability to continue investing in research and infrastructure. This balance reflects a pragmatic approach: contributing to shared understanding while protecting the integrity of production systems.

Deployment decisions are guided by the principles of safety, privacy, and operational accountability. By enabling local execution and minimising data movement, OneBit AI reduces exposure

to centralised failure modes and external misuse while giving operators greater control over how and where intelligence is applied.

Roadmap (12 Months)

The near-term roadmap for OneBit AI focuses on the transition from validated system prototypes to production-ready deployments.

Key milestones include proprietary model training optimised for ultra-low-bit execution, expansion of pilot deployments across educational and research environments, and continued refinement of edge hardware platforms. In parallel, the software stack will be extended to support broader application workflows and improved developer tooling.

The roadmap also includes the rollout of offline-capable mobile applications and preparation for enterprise-grade deployments, including reliability, monitoring, and lifecycle management features.

Together, these efforts aim to establish OneBit AI as a practical, scalable foundation for efficient, local intelligence across a wide range of deployment environments.

Conclusion: The OneBit Leap

The history of computing is a pendulum swing between centralisation and distribution. We are now standing at the apex of the centralisation era, looking toward a distributed future.

Current industry efforts to address this shift rely on "distillation",taking massive, inefficient cloud models and attempting to shrink them. This approach is fundamentally limited; it carries the architectural debt of the cloud down to the device.

OneBit AI has taken a different path We did not shrink a cloud model; we rebuilt the arithmetic of intelligence from the ground up for the edge. By moving from floating-point matrix multiplication to integer-based addition, we have achieved a generational leap in efficiency that competitors cannot replicate simply by "optimising" existing architectures.

This mathematical breakthrough allows us to outgrow the market constraints. While others wait for more powerful chips to run their heavy models, we unlock supercomputer-class intelligence on the chips that exist today.

OneBit AI is not just building a smaller model. We are building the operating system for the next era of ubiquitous, private, and energy-efficient intelligence.

AI will not be limited to the cloud. The future belongs to the edge.

References

[1] Kryś J, Sharma Y and Egan J 2025 Distributed and decentralised training: Technical governance challenges in a shifting ai landscape (*Preprint* 2507.07765) URL https://arxiv.org/abs/2507.07765

[2] Wu C J, Raghavendra R, Gupta U, Acun B, Ardalani N, Maeng K, Chang G, Aga F, Huang J, Bai C, Gschwind M, Gupta A, Ott M, Melnikov A, Candido S, Brooks D, Chauhan G, Lee B, Lee H H, Akyildiz B, Balandat M, Spisak J, Jain R, Rabbat M and Hazelwood K 2022 Sustainable ai: Environmental implications, challenges and opportunities *Proceedings of Machine Learning and Systems* vol 4 ed Marculescu D, Chi Y and Wu C pp 795–813 URL https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf

[3] Elliott D and Soifer E 2022 *Frontiers in Artificial Intelligence* **Volume 5 - 2022** ISSN 2624-8212 URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.826737

[4] Hao Z, Guo J, Shen L, Luo Y, Hu H, Wang G, Yu D, Wen Y and Tao D 2025 Low-precision training of large language models: Methods, challenges, and opportunities (*Preprint* 2505.01043) URL https://arxiv.org/abs/2505.01043

[5] Shen H, Chang H, Dong B, Luo Y and Meng H 2025 *Efficient LLM Inference on CPUs* (Cham: Springer Nature Switzerland) pp 33–46 ISBN 978-3-031-85747-8 URL https://doi.org/10.1007/978-3-031-85747-8_3

[6] Ma S, Wang H, Ma L, Wang L, Wang W, Huang S, Dong L, Wang R, Xue J and Wei F 2024 The era of 1-bit llms: All large language models are in 1.58 bits (*Preprint* 2402.17764) URL https://arxiv.org/abs/2402.17764