

# LLM Infrastructure Forecast

## 1. What is a TPU?

A **TPU (Tensor Processing Unit)** is a custom-designed AI accelerator chip developed by Google specifically for machine learning workloads.

### Key Characteristics

- **Purpose-built for ML:** Optimized for matrix operations and tensor computations common in neural networks
- **High throughput:** Excels at large-scale, low-precision (8-bit) matrix multiplications
- **Architecture:** Uses a systolic array design that efficiently moves data through processing elements
- **Power efficient:** Delivers more ML performance per watt compared to general-purpose GPUs/CPUs

### TPU vs GPU

Aspect	TPU	GPU
Specialization	ML-only	General-purpose parallel computing
Best for	Large transformers, inference at scale	Varied workloads, smaller models
Flexibility	Limited	High

## 2. Will ASICs Dominate if LLMs Become Mainstream?

### Arguments For ASIC Dominance

- **Efficiency:** ASICs can be 10-100x more power-efficient than GPUs for specific workloads
- **Cost at scale:** Once designed, per-unit costs drop significantly in high volume
- **Inference dominance:** If LLMs become ubiquitous, inference will be the bulk of compute—ideal for ASICs
- **Edge deployment:** Running models on phones/devices almost certainly requires custom silicon

### Arguments Against (GPU Resilience)

- **Rapid model evolution:** LLM architectures are still changing fast. ASICs take 2-3 years to design—risky if architectures shift
- **NVIDIA's moat:** CUDA ecosystem, software stack, and developer familiarity are deeply entrenched
- **Flexibility:** GPUs can run any model; ASICs may become obsolete if paradigms change
- **Hybrid approaches:** NVIDIA is adding specialized tensor cores—blurring the line

### Current Trajectory

- Hyperscalers (Google, Amazon, Microsoft) are building custom chips (TPU, Trainium, Maia)

- Startups (Groq, Cerebras, SambaNova) are betting on specialized architectures
- NVIDIA still dominates (~80%+ of AI training market)

## Likely Outcome

A mixed ecosystem—ASICs for inference at scale and edge, GPUs for training and flexibility. If architectures stabilize, ASICs gain ground. If innovation continues rapidly, GPUs remain essential.

## 3. Enterprise LLM Deployment: Hybrid Model

The likely future is a **hybrid model** where companies run private small LLMs for routine tasks and use public cloud LLMs for heavy compute. This mirrors how companies handle compute generally (on-prem + cloud).

### Why Private Small LLMs Make Sense

- **Data privacy:** Sensitive data never leaves the network
- **Latency:** Local inference is faster for real-time applications
- **Cost predictability:** Fixed infrastructure vs. per-token API costs
- **Customization:** Fine-tuned on proprietary data, jargon, workflows
- **Compliance:** Easier to meet regulatory requirements (GDPR, HIPAA, etc.)

### Why Public LLMs for Heavy Compute

- **Frontier capabilities:** Largest models require massive infrastructure
- **Occasional use:** Doesn't justify owning the hardware
- **Rapid improvement:** API access means instant upgrades
- **Burst capacity:** Handle spikes without over-provisioning

### Emerging Deployment Patterns

Use Case	Likely Solution
Internal chatbots, code assist	Private small LLM (7B-70B)
Document search/RAG	Private, fine-tuned
Complex reasoning, research	Public frontier API
Customer-facing products	Hybrid or public
Edge/embedded	Tiny private models (<3B)

### The Analogy

It's like databases—companies run private databases for core operations but use cloud services for analytics, burst workloads, or specialized capabilities.

## 4. The Bitcoin ASIC Analogy: Will History Repeat?

Bitcoin mining evolved from CPUs → GPUs → FPGAs → ASICs, with ASICs now dominating completely. Will LLM inference follow the same path?

### Why Bitcoin ASICs Dominated Completely

- **Single, fixed algorithm:** SHA-256 never changes
- **Pure economics:** Only metric is hashes per watt per dollar
- **No flexibility needed:** The workload is 100% predictable forever
- **Winner-take-all:** Efficiency directly equals profit

### Why LLM ASICs Won't Dominate as Completely

Factor	Bitcoin	LLMs
Algorithm stability	Fixed forever	Evolving (attention → MoE → SSM?)
Workload variety	One operation	Many (models, quantizations, batch sizes)
Market maturity	15+ years	~3 years
Upgrade cycle	Rare algorithm changes	New architectures yearly

### Where LLM ASICs Will Likely Dominate

- **Edge devices** (phones, cars, IoT): ASICs will dominate—battery life is critical
- **High-volume inference:** Running the same 7B model billions of times justifies custom silicon
- **Commoditized models:** Once a model becomes stable (like Llama-class), ASICs become viable

### Likely Pattern by Use Case

Use Case	Dominant Hardware
Training	GPUs (too dynamic)
Large inference (cloud)	Mix of GPUs + specialized accelerators
Small inference (edge)	ASICs (similar to Bitcoin)

### The Key Variable

Architecture stability determines ASIC viability. If transformers remain the standard for 5+ years, ASICs will take over inference. If major shifts occur (like Mamba/SSMs gaining traction), GPU flexibility remains valuable.

## 5. The Fragmented AGI Future: Data Sovereignty Forces Decentralization

The current centralized API model (everyone sends data to OpenAI/Anthropic/Google) is unlikely to survive the path to AGI. Data security requirements in a capitalist model will force fragmentation.

## Why Centralized APIs Won't Scale to AGI

- **Data is the moat:** Corporations won't send proprietary data to potential competitors
- **Regulatory pressure:** GDPR, HIPAA, national security laws prohibit cross-border data flows
- **Competitive risk:** Training data leakage could destroy competitive advantage
- **National security:** Governments won't route sensitive queries through foreign systems

## The Emerging Tiered Model

Tier	Users	Model Type	Data Policy
Tier 1: Public	Education, researchers, public	Open source (Llama, Mistral)	Public data only
Tier 2: Enterprise	Corporations	Private fine-tuned, on-prem	Data stays internal
Tier 3: Regulated	Healthcare, finance, legal	Certified & audited	Compliance-first
Tier 4: Sovereign	Governments, defense	Air-gapped, national	Complete isolation

## Market Projection

The centralized API model (currently ~85% of AI compute market) will decline to ~10% by 2032 as enterprise moves compute on-premises, governments mandate sovereign AI capabilities, and open models become capable enough for public use.

## The Google Analogy

Just as Google Search is 'free' for public use while enterprises pay for private search appliances and governments build classified systems, AGI will fragment into:

- **Public AGI:** Ad-supported or government-subsidized for education/general use
- **Enterprise AGI:** Licensed, on-prem, fine-tuned on proprietary data
- **Sovereign AGI:** National AI capabilities, completely isolated

## Implications

- **No single AGI monopoly:** Unlike search (Google dominance), AGI will be fragmented by design
- **NVIDIA benefits:** Sells hardware to all tiers, not dependent on any single provider
- **Open source critical:** Public tier depends on open models (Llama successors)
- **Talent fragmentation:** AI researchers spread across government, enterprise, public sectors

## 6. The Power Bottleneck: Does China Win the 7-Year Race?

If power becomes the primary constraint on AI scaling, geopolitical dynamics shift dramatically. China's infrastructure advantages could prove decisive.

## The Power Problem

- **Current AI data center:** 50-100 MW typical
- **Next-gen training clusters:** 500 MW - 1 GW required
- **GPT-5 class training:** Estimated 100+ MW sustained for months
- **AGI-scale compute:** Potentially 5-10 GW dedicated facilities

## China's Structural Advantages

Factor	China	US/West
Permitting speed	Months	5-10 years
State coordination	Central planning	Fragmented jurisdictions
Grid buildout	Rapid expansion	Aging infrastructure
Nuclear expansion	150+ reactors planned	Regulatory paralysis

## 7-Year Scenario (2026-2033)

- **2026-2027:** US leads on architecture; power constraints emerge; China builds power plants
- **2028-2029:** US hits grid limits; China's new plants come online; compute parity approaches
- **2030-2033:** China achieves raw compute advantage; US forced into efficiency focus

## The Critical Question

**If scaling laws hold** (more compute = better AI): China wins through brute force power advantage

**If algorithmic breakthroughs dominate:** US/West wins through talent and research ecosystem

## Likely Outcome

A bifurcated AI world by 2033: Chinese AI sphere (raw power, state-controlled, closed) vs Western AI sphere (efficiency-focused, distributed, allied nations pooling resources). Neither achieves global AGI monopoly.

## 7. What If AGI Doesn't Scale? The Moore's Law Parallel

The assumption that 'more compute = smarter AI' may break down, just as Moore's Law eventually hit physical limits.

## The Moore's Law Template

- **1970-2010:** Exponential scaling held (transistors doubled every 2 years)
- **2010-2025:** Dennard scaling ended; gains slowed to ~3 year doubling
- **2025+:** Physical limits (atomic scale) cause further slowdown

## Four Scenarios for 2026-2036

Scenario	Assumption	2036 Outcome
Optimistic	Scaling continues	AGI achieved
Moderate	Moore's Law pattern	~3x current, no AGI
Pessimistic	Hard ceiling	~1.5x current, plateau
Plateau	Brief gains then stagnation	Near-current, no AGI

## The Binding Constraints

Capability = Minimum(Compute, Data, Algorithms, Energy). Progress stops when ANY constraint binds:

- **Training data exhaustion:** Internet-scale text already consumed
- **Compute limits:** Power constraints, chip fab limits, prohibitive costs
- **Algorithmic ceiling:** Transformer may be near-optimal, no successor paradigm
- **Energy wall:** Training runs consuming city-scale power

## The Uncomfortable Question

Current AI progress may be a **one-time windfall** from: (1) Transformer architecture, (2) Scale discovery, (3) Internet-scale training data. If no new paradigm emerges, we may be witnessing the **peak of this approach**, not the beginning of exponential takeoff.

## 8. The Data Wall: How Can OpenAI Continue Scaling?

The fundamental problem: scaling requires exponentially more data, but high-quality training data is finite and already exhausted.

## The Numbers Don't Work

Model	Training Tokens	Status
GPT-2 (2019)	~10 billion	Abundant data
GPT-3 (2020)	~300 billion	Plenty remaining
GPT-4 (2023)	~13 trillion	Used most of internet
GPT-5 (2025?)	~50+ trillion	Doesn't exist

Available high-quality internet text: ~10-15 trillion tokens. Annual new content: ~1-2 trillion.

## OpenAI's Attempted Solutions

- **Synthetic data:** AI generates training data → model collapse risk, quality degrades
- **Licensed deals:** Reddit (\$60M/yr), publishers → expensive, finite, legally contested
- **Multimodal:** Video/audio → different modality, doesn't help text reasoning
- **User data:** ChatGPT conversations → privacy laws, consent issues

- **RLHF quality:** Better curation → doesn't add new knowledge

## The Uncomfortable Reality

**OpenAI cannot continue the scaling approach** that made GPT-3→GPT-4 successful. Options: admit diminishing returns, pivot to efficiency, hope for algorithmic breakthroughs, or gamble on synthetic data.

## What This Means for AGI

- **AGI via scaling is impossible:** Not enough data exists
- **Algorithmic breakthroughs required:** Fundamentally new approaches needed
- **China's compute advantage irrelevant:** Can't train what doesn't exist
- **Open source catches up:** Diminishing returns level the field

The scaling era (2019-2024) may be over. What comes next is uncertain.

## 9. AI Sector Evolution: Cloud vs Local (2026-2029)

Data sovereignty and security concerns will drive a major shift from cloud-based AI to local/on-prem deployments.

### Year-by-Year Projection

Year	Public Cloud	Govt Cloud	Corp Cloud	Trend
2026	75%	40%	65%	Cloud dominant
2027	60%	30%	50%	Transition begins
2028	45%	15%	35%	Local majority
2029	35%	10%	25%	Local dominant

### Country-Level Patterns (2029)

- **Developed democracies** (US, EU, Japan): 25% cloud, 55% local, 20% hybrid
- **Authoritarian states** (China, Russia): 5% foreign cloud, 70% local, 25% state cloud
- **Developing nations** (India, Brazil, Africa): 45% cloud-dependent, 20% local

### Key Drivers

- **Regulatory mandates:** GDPR, China data laws, US federal requirements
- **Security incidents:** Each breach accelerates local adoption
- **Open source maturity:** Llama, Mistral make local deployment viable
- **Cost crossover:** On-prem becomes cheaper at scale

## 10. Summary & Conclusion: AI in 5 Years and AGI

Synthesizing all findings into a coherent forecast for 2026-2031.

### 5-Year Trajectory

Year	Capability	Deployment	AGI Prob
2026	GPT-4 class	Cloud dominant	0%
2028	Scaling wall	Hybrid standard	5%
2031	Uncertain	Regional blocs	15%

### AGI Probability: ~12% by 2031 (weighted average)

- **Optimistic scenario:** 35% (scaling works, data problem solved)
- **Moderate scenario:** 15% (gradual algorithmic progress)
- **Pessimistic scenario:** 5% (data wall binding, no breakthroughs)
- **Plateau scenario:** 2% (stagnation, fundamental limits)

### Key Bottlenecks Preventing AGI

- **Training data exhaustion** (90/100): Primary blocker—internet consumed
- **Algorithmic ceiling** (75/100): Transformers may be near-optimal
- **Unknown unknowns** (80/100): Fundamental limits undiscovered

### The Bottom Line

**AGI by 2031 is UNLIKELY (~12%).** AI will be more capable (2-3x, not 100x) but incrementally. AI will be more fragmented—regional, sectoral, organizational silos replacing centralized APIs.

Winners: Hardware makers, open source, enterprises with data moats, countries with energy infrastructure.

Losers: Pure API business models, scaling assumptions, imminent AGI predictions.

The 2019-2024 period may be remembered as a **one-time windfall** from transformers + scaling + internet data. What comes next is uncertain—but likely **evolution, not revolution**.

*Generated: February 2026*

*See PNG files for all visualizations*