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

Generated: February 2026

See agi_future_tiers.png and agi_market_projection.png for visualizations