COHESIONX

HARMONY IN EXPERTISE

State of AI

2025

Executive Summary

Artificial intelligence is moving from a research-led boom to an infrastructure-led buildout.

Operators, chipmakers, and policymakers converge on three realities:

(1) large-scale clusters are being designed and financed like “AI factories,” not classic data centres.

(2) the economics of inference and training are undergoing radical cost deflation; and

(3) Political acceptance of AI now depends on energy pragmatism and clear, light-touch safety reporting rather than abstract, centralised control.

The result is an industrial cycle measured in gigawatts, not merely GPUs or model releases.

The “factory” framing matters. At scale, the input bill of materials is chips + data + algorithms + electricity; the output is tokens - manufactured cognition that feeds products, research, and operations. Analysts cite a $20T impact by 2030 and an empirical rule of thumb that each dollar of enterprise AI spend returns ~$4.60 in value, which explains why hyperscalers and new entrants are committing multi-year capex in parallel with power development.

At the model layer, the price shock is unmistakable: DeepSeek 3.2 EXP lists at roughly $0.28/M tokens in and $0.42/M out (vs. ~$3.15 for premium incumbents such as OpenAI) and introduces DeepSeek Sparse Attention (DSA), which can cut cost by ~50%. Operators like 8090 describe genuine switching frictions - toolchains, safety layers, and guardrails are tuned per model - which slows hot-swaps even when price/perf is compelling, but the long-term direction is clear: multi-model routing and persistent token-cost deflation.

Competition is not just between companies but between architectures and geographies.

Closed-source U.S. frontier models dominate the proprietary tier; the strongest open-source releases of late have come from China (DeepSeek, Qwen, Gimmie). In parallel, a decentralisation push (Bittensor/TAO; Apple’s on-device roadmap) points to a hybrid future of hyperscale training plus local inference for privacy, latency, and cost.

Infrastructure and power are the bottlenecks. Case studies describe gigawatt-class builds (e.g., >1.2 GW, ~400k GPUs) using modular “Lego-block” construction, with 4,000 workers onsite and ~$15B raised per site. On the grid side, communities push back when they fear bill spikes; proposed near-term off-ramps include cross-subsidies

and mass deployment of home batteries near clusters, plus “peak shaving” (~40 hours/year) to free roughly ~80 GW of headroom.

Meanwhile, regulation is sprinting ahead of consensus: all 50 U.S. states have introduced AI bills (1,000+ in flight; 118 enacted), with California’s newer SB-53 style approach aiming transparency for frontier models (terms like “catastrophic harms” and “model autonomy” remain vague). The patchwork is a compliance trap for startups - worse than a single EU-style regime. Yet, none of it changes the central arc: AI factories will scale, GPUs will compound through software, physical-AI supply chains (e.g., magnets) will be rebuilt in the USA, and the deployment frontier will move closer to users and devices.

1. From Data Centres to AI Factories

Analysts reframe modern compute footprints as AI factories - continuous-production plants for intelligence. The difference from classic software is not semantic: operating performance is bounded by energy, thermal envelopes, and cluster-scale networking, not just developer velocity. In this frame, the goal is to align year-ahead GPU roadmaps, space, and power so that each generation’s perf/$ and perf/W leap turns directly into lower unit economics for customers and more tokens for applications.

1.1. Input Bill of Materials:

- High-performance accelerators (GPUs/TPUs)
- High-bandwidth memory
- Low-latency networking fabric
- Curated training data
- Reliable electricity supply

Output:

- Tokens (text, images, video, embeddings)
- Generated decisions
- Downstream product capabilities

1.2. Scale is accelerating

Northern Virginia - the world’s largest DC region - stood at ~4.5 GW in 2014; single facilities now target ~5 GW, implying we are building “more than a Northern Virginia” every year. Developers report ~40 GW build pipelines spanning SMRs, renewables, and gas, with modular factory-built components that snap together on site to reduce time-to-capacity.

1.3. Economic Logic

If AI manufactures cognition, the industry behaves like capital-intensive utilities: long-dated commitments, integrated power procurement, and job creation at scales more familiar to energy than to software-as-a-service. Each generation's performance-per-dollar and performance-per-watt improvements translate directly into lower unit economics and expanded application capabilities.

The factory metaphor reflects fundamental physics: power availability determines cluster size, cooling determines density, network topology constrains collective performance, and maintenance windows become as critical as software deployments.

2. The New AI Economics: Radical Cost Deflation

DeepSeek 3.2 EXP anchors the cost shock: about $0.28/M input tokens and $0.42/M output tokens - an order of magnitude (or more) cheaper than premium incumbents like Claude (~$3.15/M). It adds a sparsity mechanism (DSA) that reduces cost up to ~50% by focusing compute where attention matters most. This is not a one-off price war; it reflects architectural progress that is likely to propagate through the stack.

Teams describe the real-world friction of model switching. Tooling, evaluation harnesses, red-teaming, and fine-tuning are optimised per model; “hot-swapping” a production pipeline takes weeks or months. As a result, many builders adopt blended strategies - keeping coding co-pilots or high-stakes tasks on the best-in-class model while diverting high-volume workloads to cheaper engines. Net: deflation continues, but migration happens in waves.

Expect the buyer stack to formalise multi-model routing (policy-based selection by price/quality/latency), systematic offline evals, and shared safety tooling that travels between providers. As sparsity, quantisation, and context-length engineering improve, energy per token should fall with dollars per token - a key point several analysts make explicitly.

2.1. Price Compression Mechanisms

Multiple forces converge to reduce costs:

Architectural Innovation:

- Sparsity: DeepSeek Sparse Attention (DSA) directs computation to the most informative tokens and layers, cutting costs up to 50%
- Quantisation: Reduces arithmetic precision while maintaining accuracy
- Compiler Optimisation: Kernel fusion minimises memory movement
- Caching: Reuses intermediate results across similar prompts

Operational Efficiency:

- Larger batch processing
- Improved scheduling algorithms
- Enhanced accelerator utilisation

2.2. Market Impact

DeepSeek 3.2 EXP represents the cost shock:

- Input tokens: $0.28/M (vs. $3.15/M for premium providers)
- Output tokens: $0.42/M
- 10× reduction compared to incumbents

This is not a temporary price war but reflects architectural progress likely to propagate throughout the ecosystem.

2.3. Real-World Migration Friction

Despite compelling economics, model switching faces genuine barriers:

Technical Constraints:

- Toolchains optimised per model
- Evaluation harnesses tuned to specific behaviors
- Red-teaming and safety layers model-dependent
- Fine-tuning investments non-transferable

Migration Timeline: Production pipeline transitions require weeks to months, not days.
Practical Response: Organisations adopt blended strategies—retaining premium models for safety-critical and brand-sensitive workloads while routing high-volume tasks (extraction, translation, summarisation) to lower-cost alternatives that meet defined quality thresholds.

2.4. Emerging Buyer Stack

Multi-Model Routing:

- Policy-based model selection by quality, latency, and price targets
- Offline evaluation suites tracking regressions and drift

- Cost allocation visibility per team and feature
- Prompt management systems with versioning and rollback capabilities

Objective: Match optimal models to specific tasks and swap confidently when economics shift, rather than crowning a single "best" model.

3. Open vs. Closed Models - and the U.S. - China Split

Analysts paint a stark (if evolving) asymmetry: the U.S. dominates closed frontier models and much of the hardware and cloud stack, while China currently leads high-quality open models (DeepSeek, Qwen, Gimmie). Open release means anyone can fork weights and run them domestically; this checks big-tech power but raises national-security and supply-chain questions.

Importantly, open-source ≠ “Chinese cloud.” Forked models can be hosted on U.S. or any other infrastructure, with independent safety testing and hardening. The risk to watch is not data exfiltration so much as potential backdoors/vulnerabilities - something the competitive security community is actively probing.

Several sources note Apple’s OpenELM effort and a broader device-first posture: when you’re behind, you open up to accelerate; when you’re ahead, you tend to close. Regardless, proliferation is inevitable: “no one needs nuclear weapons; everyone needs AI.”

3.1. Current Market Structure

Closed Frontier Models (U.S.-Led):

- Dominate proprietary tier
- Lead in raw capability at the frontier
- Control cloud infrastructure

Open Source Models (China-Led):

- DeepSeek, Qwen, Gimmie lead high-quality releases
- Enable anyone to fork weights and self-host
- Reach parity for many mainstream tasks

3.2. Strategic Implications

Open Source ≠ Foreign Cloud Dependency

Organisations can host open models on domestic infrastructure with independent security testing. Key risk management practices:

- Pin exact model hashes
- Mirror weights internally
- Verify signatures from multiple sources
- Never auto-update in critical systems
- Conduct independent vulnerability assessments

Practical Risk Profile:

- Data need not leave national boundaries
- Primary concerns: hidden vulnerabilities, update provenance uncertainty
- Managed through engineering discipline, not prohibition

3.3. National Security Considerations

Physical AI supply chains (magnets, packaging, memory) require domestic capability. The strategic objective is resilience, not autarky—sufficient domestic capacity to absorb shocks, with allied depth providing redundancy.

4. Decentralisation and Local AI

AI will be hybrid: hyperscale training + local/edge inference. Crypto-incentivised networks (e.g., Bittensor/TAO) illustrate how distributed compute can harness idle capacity, while platform vendors push personal models onto phones and PCs for privacy, latency, and cost control. Analysts position this not as ideology, but as engineering sense and customer preference.

As local models absorb personal context and preference, value will shift from raw model quality to orchestration - what runs where, with what context, and under which privacy rules. This blurs today's boundaries between "cloud AI" and "device AI," letting users keep sensitive prompts and embeddings close to the metal.

4.1. The Cloud-Edge Continuum

The future is hybrid: hyperscale training and heavy inference remain centralised, while personal reasoning and retrieval increasingly occur on devices.

Drivers:

- **Privacy:** Personal context (notes, histories, calendars) stays local
- **Latency:** Eliminates round-trip delays
- **Cost**: Reduces bandwidth expenses
- **Resilience**: Services survive connectivity interruptions

Historical Pattern: Mirrors computing evolution from mainframe→PC, server→smartphone, now cloud→cloud-plus-edge continuum.

4.2. Decentralised Networks

Crypto-incentivised networks (e.g., Bittensor/TAO) demonstrate how distributed compute can harness idle capacity. Platform vendors (notably Apple with OpenELM) push personal models onto phones and PCs.

Value Shift: From raw model quality to orchestration capabilities—determining what runs where, with which context, under what privacy rules.

4.3. Enterprise Implementation

Data Classification Framework:

- Device/Private Enclave: Highly personal or regulated data
- Shared Services: Everything else

Design Requirements:

- Orchestration rules for model placement
- Data retention policies
- Audit logging requirements
- Experience optimisation without sacrificing accountability

5. Generative Video, Personas, and IP

New consumer apps (OpenAI Sora, Meta Vibes) demonstrate rapid progress in video generation and persona tools. The features analysts emphasise - opt-in use of notable figures, friend-scoped persona rights, and improving prompting/scripting -point to workflows that will feel increasingly “consumer-ready” within 12–24 months. Today is the worst they’ll ever be.

The live legal overhang concerns inputs vs. outputs. Output cloning is obviously infringing; training-set legality is more debated. One controversial design cited in your material is opt-out defaults for broader IP ingestion, which will invite more lawsuits alongside growing settlements. Expect pragmatic guardrails: licensing where feasible, transparency reports, and user-level controls.

5.1. Technical Capabilities

Consumer tools (OpenAI Sora, Meta Vibes) demonstrate rapid progress in video generation and persona creation. Quality improvements suggest consumer-ready workflows within 12–24 months.

5.2. Definitions and Scope

Personas: Configurable characters that look, sound, or behave like a person

- Notable Persons: Individuals with public recognition
- Private Individuals: Everyone else
- Friend-Scoped: Explicitly shared with limited audiences

Cloning vs. Inspiration:

- Cloning: Output reasonably passes as specific person's face/voice
- Inspiration: Captures general style without look-alike characteristics

5.3. Legal Framework

Output vs. Input Considerations:

- Outputs: Published results that closely imitate without permission trigger right-of-publicity claims
- Inputs: Training materials subject to evolving fair-use norms

5.4. Recommended Guardrails

Product Design:

1) Default persona scope to private and non-commercial
2) Require explicit consent for real-person clones
3) Embed visible labels and invisible provenance
4) Block/review sensitive contexts (elections, medical, financial)
5) Provide prominent "stop using my persona" controls

Business Practices:

- License where feasible
- Maintain transparency reports
- User-level usage controls
- Takedown mechanisms for abusive content

6. Gaming as an AI Proving Ground

Games already validate AI's value: adaptive bots keep new players engaged (a Fortnite example), and asset generation compresses content pipelines. Experiments in AI-rendered 3D "worlds without a physics engine" are underway, though seasoned engine builders doubt they can yet meet AAA scale/quality. In the interim, AI layers atop Unity/Unreal will render characters, scenes, and logic to spec.

Commercially, your sources argue AI-native titles can route around IP gatekeepers by synthesising dynamic, personalised content. Against this backdrop, the proposed $55B EA take-private is read as a bet that AI will increase the strategic value of gaming IP and distribution—and that legacy gatekeepers (e.g., consoles) can be disintermediated with time.

6.1. Current Applications

Asset Pipeline Acceleration:

- Concept art generation
- Texture creation
- Dialogue variants
- Reduced onboarding friction

Adaptive Systems:

- Non-player characters that respond to player skill
- Sustained long-term engagement
- Personalised difficulty curves

6.2. Experimental Frontiers

AI-rendered 3D worlds without traditional physics engines remain experimental. While promising, achieving AAA-grade output at scale requires further development. Near-term reality: AI layers atop Unity/Unreal rendering characters, scenes, and logic to specification.

6.3. Commercial Implications

Market Disruption:

- Dynamic, personalised content reduces expensive licensing dependencies
- New studios can compete with established publishers
- Blurred lines between game, tool, and social platform

Strategic Response: The proposed $55B EA take-private reflects expectations that AI increases strategic value of gaming IP and distribution, potentially disintermediating traditional console gatekeepers.

7. Case Study: Building a Gigawatt-Scale AI Factory

One developer describes a 1.2-GW site designed as a single coherent cluster with ~400,000 NVIDIA GPUs, built from modular components assembled like Lego blocks. Roughly 4,000 workers are on site daily; about $15B has been raised. This is not a one-off vanity project: the company reports a ~40-GW development pipeline across SMRs, renewables, and gas.

The lesson for operators: schedule power, chips, construction, and networking as a single critical path; pre-fabricate as much as possible; and assume multi-gigawatt demand becomes normal in the second half of the decade.

7.1. Physical Architecture

Modern gigawatt-class sites resemble industrial campuses:

- Prefabricated modules with integrated power distribution

- Pre-assembled cooling loops and rack systems
- Crane-installed components connected to high-capacity spine
- Low-latency fabrics connecting tens of thousands of accelerators

7.2. Critical Path Components

Long-Lead Items:

- Land acquisition and permits
- Grid interconnect studies
- Electrical equipment (transformers, switchgear)
- Fiber routes
- Substation capacity upgrades

Labor Profile:

- Peak during module installation
- Secondary peak during network/storage fit-out
- 24/7 monitoring and maintenance post-launch
- Strict safety protocols and maintenance schedules

7.3. Operational Model

Sites operate as continuous-process plants with:

- Codified templates reused across subsequent builds
- Airflow and liquid cooling optimised for density
- Heat recovery designs minimising operating costs
- Predictable maintenance windows

8. Power is the Constraint: Rates, PR, and Practical Off-Ramps

Your sources warn that the next five years are “baked” on supply, risking electricity rate spikes that could turn public opinion against AI. A notable case involved residents near Indianapolis opposing a $1B data centre over price-inflation fears. In Virginia, an estimated ~40% of power already flows to data centres.

Near-term mitigations include cross-subsidies (hyperscalers accept higher rates so household bills don’t rise) and home-battery programs around clusters. Grid operators can also shed ~40 peak hours/year to backup generation—unlocking ~80 GW of adequate capacity—while gas turbines clear their backlogs and nuclear projects advance.

8.1. Community Opposition Dynamics

Electricity availability shapes public sentiment. Residents resist when fearing higher bills or crowding out local needs.

Case Example: Indianapolis residents opposed $1B data centre over price inflation concerns. Virginia already allocates ~40% of power to data centres.

8.2. Mitigation Strategies

Tariff Innovation:

- Cross-subsidies (hyperscalers accept higher rates)
- Shield household bills from industrial demand
- Published consumption data with community audits

Technical Solutions:

- Demand-response agreements (~40 peak hours/year)
- Behind-the-meter generation and storage
- Peak shaving unlocking ~80 GW adequate capacity
- Home battery programs near clusters

Community Benefits:

- Local hiring commitments

- Procurement targets favouring regional suppliers
- Construction traffic management
- Noise abatement and landscaping

8.3. Long-Term Supply

Five-Year Outlook: Supply is "baked," creating rate spike risk

Development Pipeline:

- Medium term: New gas turbines, uprated transmission
- Long term: Advanced nuclear (SMRs), large-scale renewables
- Strategy: Sequence options showing community benefits while maintaining grid reliability

9. Chips and Fabs: Arizona's Yields and the Global Reality

Lisa Su reports that TSMC Arizona's first advanced nodes are yielding equivalently to Taiwan with a low double-digit cost premium, not the feared +50%. The message: onshoring leading-edge manufacturing is challenging but tractable, and assurance of supply justifies modest cost deltas in a world where "everyone wants a GPU."

The ecosystem remains global by design—ASML, Taiwan, Korea, Japan, and U.S. fabs as a concert. The right goal is not autarky but sufficient capability to cover national and commercial requirements, with allies providing depth and redundancy.

9.1. Domestic Capability Building

TSMC Arizona yields match Taiwan benchmarks with low double-digit cost premiums (not the feared 50%+ increase). Message: Onshoring leading-edge manufacturing is tractable, and supply assurance justifies modest cost deltas.

9.2. Ecosystem Realities

Necessarily International:

- Lithography equipment (ASML, Netherlands)
- Specialty chemicals (Japan, Europe)
- Substrates and packaging (Taiwan, Korea)

- Equipment maintenance (distributed globally)
- Strategic Objective: Resilience, not autarky—sufficient domestic capability to absorb shocks, with allied capacity providing depth.

9.3. Pinch Points

Critical Focus Areas:

- Advanced packaging
- High-bandwidth memory (HBM)
- Equipment supply chains
- Process compatibility across fabs

Investment Priorities: Multi-sourcing strategies and compatible process nodes improve shock absorption.

10. GPUs, Residual Value, and Software-Driven Uplift

NVIDIA now shares roadmaps ~12 months ahead, giving customers time to plan power, space, and capex and to time transitions. Allocation is straightforward: “place a PO.” Despite rapid cycles, residual values have held up (Hopper at ~75–80% after year one) because CUDA programmability and community optimisations deliver substantial software uplift (~4× on Hopper over the shipping window, per your material).

Strategically, the vendor objective is relentless improvement in perf/$ and perf/W so applications can “think longer” (generate more tokens) while overall TCO declines. That is the core economic engine behind the factory metaphor.

10.1. Market Dynamics

Roadmap Transparency: NVIDIA shares 12-month forward roadmaps, enabling facility planning alignment with accelerator refresh cycles.

Allocation Mechanism: "Place a purchase order"—straightforward for qualified buyers.

Residual Values: Hopper maintains 75–80% value after year one, supported by software improvements.

10.2. Software-Driven Performance Gains

Previous-generation hardware retains and gains value through:

- Compiler optimisation passes
- Kernel fusion improvements
- Smarter scheduling algorithms
- Enhanced serving layers

Example: 4× throughput improvement on Hopper over shipping window through software alone.

10.3. Strategic Implications

Vendor Objective: Relentless improvement in performance-per-dollar and performance-per-watt enables applications to "think longer" (generate more tokens) while total cost of ownership declines.

Buyer Strategy: Purchasing ahead of immediate need remains rational—well-tuned clusters deliver more next quarter than today.

11. Physical AI: Magnets, Materials, and Security

Rare-earth magnets are the feedstock to physical AI—robots, drones, and any electrified motion. MP Materials details a U.S. mine-to-magnet rebuild: a Texas magnetics factory (with GM as foundational customer) plus a novel $400M Department of Defence partnership that sets a price floor against below-cost imports and includes 50/50 profit sharing above thresholds.

The policy logic is simple: you cannot fund advanced drones and then buy magnets from strategic rivals. Public-private structures that share upside and risk can crowd-in private capital to rebuild critical inputs for AI hardware.

11.1. Critical Materials

Rare-earth permanent magnets underpin physical AI:

- Robotics actuators
- Drone propulsion
- Autonomous vehicle motors
- Smart factory equipment

11.2. Domestic Rebuild Example

MP Materials U.S. Initiative:

- Texas magnetics factory with GM as anchor customer
- $400M Department of Defence partnership
- Price floor protection against below-cost imports
- 50/50 profit sharing above thresholds

Policy Logic: Cannot fund advanced defence systems while sourcing critical components from strategic rivals.

11.3. Broader Ecosystem Requirements

Beyond Magnets:

- High-performance sensors
- Precision gears and bearings
- Power electronics
- High-reliability connectors

Reality: Physical AI is an ecosystem—reliable motion depends on integrity of every component. As software capability improves, hardware constraints increasingly set pace.

12. Talent and Jobs: Building the Workforce

The industrialisation of AI is labour-intensive. One Texas site reports ~4,000 workers onsite daily across trades, with labour imported from across the country. Median wages in some upstream roles approach $100k, and companies emphasise training pipelines and owner-operator cultures to recruit and retain talent.

Your sources also flag pipeline gaps: the U.S. graduates only ~200 mining engineers/year, far below China. Bridging these gaps—from mining to fab techs to data-centre electricians—will define the pace at which AI factories and physical-AI supply chains can scale

12.1. Labor Intensity

AI industrialisation creates jobs across skill levels:

Construction Phase:

- Electricians
- Pipefitters
- Crane operators
- HVAC specialists

Operations Phase:

- Network engineers
- Reliability specialists
- Security professionals
- Facility operators

Compensation: Median wages for many roles approach $100K, with future-facing, portable skills across regions.

12.2. Training Capacity Gaps

Current Constraints:

- U.S. graduates ~200 mining engineers annually (vs. China's thousands)
- Insufficient technical college capacity
- Slow apprenticeship program scaling

12.3. Acceleration Strategies

Effective Approaches:

- Targeted apprenticeships with paid training
- Retraining programs for adjacent trades
- Portable credentials recognising real-world proficiency
- Clear safety standards and career ladders
- Local partnerships with community colleges
- Recruitment from under-represented groups

13. Enterprise Adoption: Who Moves First (and Why)

Operators divide adopters into two clusters: (1) owner-operated firms and forward-leaning public CEOs who fear disruption (these groups mandate AI programs and accept real re-engineering cost), and (2) traditional PE portfolios, where misaligned incentives and B/C-grade management stall execution. Even with case studies and white papers, selling AI transformation into mainstream PE is described as unusually hard.

Tactically, the winners will run multi-model stacks, build durable evaluation/safety pipelines, and pursue process redesign rather than bolting AI onto legacy workflows.

13.1. Early Adopter Profiles

High-Velocity Segments:

1. Owner-Operated Companies: Internalise long-term benefits, accept near-term disruption
2. Forward-Leaning Public CEOs: Fear competitive disruption, mandate AI programs

Slow-Adopter Segments: Portfolio companies optimised for short-term financial metrics struggle with:

- Change programs competing with immediate targets
- Lower risk appetite
- Misaligned incentive structures

13.2. Overcoming Barriers

Requirements for Success:

- Clear executive sponsorship
- Incentive alignment across functions
- Roadmaps sequencing value in 90-day increments
- Visible early results

13.3. Technical Implementation

Standardisation Layer:

- Unified evaluation and safety frameworks
- Shared prompt libraries
- Central integration providing:
  - Observability
  - Cost controls
  - Access governance

Organisational Model:

- Bottom-up experimentation encouraged
- Top-down productisation of successful patterns
- Retirement of redundant tools (avoid point-solution fragmentation)

14. Regulatory Landscape and Compliance

14.1. Current State

U.S. Fragmentation:

- All 50 states introduced AI legislation
- 1,000+ bills in flight
- 118 enacted
- California SB-53 style: transparency requirements for frontier models

Terms Requiring Clarification:

- "Catastrophic harms"
- "Model autonomy"
- Reporting thresholds

14.2. Compliance Burden

State-level patchwork creates "death by paperwork"—potentially worse than single EU-style regime for startups.

Impact:

- Resource drain on smaller teams
- Inconsistent requirements
- Unclear enforcement mechanisms
- Innovation friction

14.3. Recommended Approach

Internal Process:

1. Establish lightweight safety and transparency framework once

2. Reuse everywhere with jurisdiction-specific annexes
3. Maintain clear records:
    - Data sources
    - Model evaluations
    - Choices and rationale
    - Known limitations
4. Enable rapid auditor response without scrambles

15. What the Next Five Years Could Look Like

Base case: AI factories scale; data-centre share of U.S. power rises from ~2.5% toward ~10%; GPU cycles remain valuable thanks to software uplift and CUDA; leading-edge U.S. manufacturing proves viable with low double-digit cost premiums. Open and closed models coexist, with enterprises adopting blended routing.

Upside: Architectural breakthroughs (sparsity, context, reasoning) push another order-of-magnitude cost drop and enable more local inference, easing grid pressure; new public-private templates (e.g., magnets) accelerate physical-AI supply chains; regulatory templates converge.

Downside: Energy politics sour as bills rise; state-level compliance balkanises; gas-turbine and transmission backlogs persist; and a few high-profile IP cases chill creator adoption. Peak-shaving and cross-subsidies fail to materialise at scale.

15.1. Base Case

Infrastructure:

- AI factories scale out across geographies
- Data center electricity: 2.5% → 10% of U.S. total
- Software extends useful life of deployed accelerators

Models:

- Open and closed coexist
- Businesses choose deliberately per task
- Multi-model routing becomes standard

Edge Capabilities:

- Expanded local inference

- More responsive, private personal assistants

Supply Chains:

- Increased resilience as packaging, magnetics, memory investments mature

15.2. Upside Scenario

Technical Breakthroughs:

- Sparsity, context handling, reasoning improvements
- Another order-of-magnitude cost reduction
- Routine local inference for most consumer interactions

Infrastructure:

- Energy costs moderated by demand response and storage
- Faster grid upgrade cycles

Policy:

- Practical, predictable regulatory templates
- Reduced burden on smaller firms
- Balanced innovation and safety

15.3. Downside Scenario

Political:

- Energy prices politicized
- Regulatory fragmentation grows
- Public trust erosion from misuse incidents

Infrastructure:

- Transmission upgrades lag demand
- Site development delays
- Skill shortages lengthen build times

Market:

- Reduced investment velocity
- Competitive disadvantages vs. less-regulated jurisdictions

16. Risks to Manage—and How

Energy & PR risk. If households perceive AI as the cause of higher bills, public license erodes. Mitigations from your sources: cross-subsidies, local home-battery programs, targeted peak-shaving (~40 hours ⇒ ~80 GW), and clearer rate design near clusters. Execution risk is non-trivial—these require utility partnerships and local trust.

Compliance Balkanisation. Fifty sets of rules impose “death by paperwork.” Response: pre-baked safety assessments and transparency reports portable across states; lean governance that meets SB-53-style expectations without over-engineering.

Supply-chain fragility. Even with strong U.S. capability, the system is global. Maintain redundancy (fabs, magnetics, grid equipment), cultivate allied depth, and support public-private models that neutralise mercantilist tactics

16.1. Energy and Public Trust

Risk: Without credible plans to shield households and deliver local benefits, opposition will intensify.

Mitigations:

- Publish verified consumption data
- Fund storage programs near affected communities
- Peak-shaving contracts reducing grid stress
- Visible local hiring and procurement
- Early, transparent community engagement

16.2. Compliance Balkanisation

Risk: Overlapping rules slow smaller teams disproportionately.

Mitigations:

- Establish lightweight safety/transparency process once
- Reuse with jurisdiction-specific modifications

- Maintain comprehensive documentation
- Prepare for audit requests proactively

16.3. Supply Chain Fragility

Risk: Global dependencies create vulnerability to disruptions.

Mitigations:

- Diversify suppliers for key components
- Validate alternates early
- Maintain buffer stocks where feasible
- Participate in public-private programs for critical materials
- Design for component substitution tolerance

16.4. Model Security and Provenance

Risk: Hidden vulnerabilities in open models, especially from strategic competitors.
Mitigations:

- Pin exact model hashes
- Mirror internally
- Verify signatures from multiple independent sources
- Never auto-update in production
- Conduct independent security assessments
- Maintain fallback options

17. Operator Playbook: 12–18 Month Actions

1) Co-plan power + compute. Book GPUs against the vendor roadmap and align power/space/capex on the same Gantt chart; treat factory ramp-up as the product.

2) Build a multi-model layer. Route by task, quality, and price; expect migrations in waves; invest in evals and safety tooling that travels across models.

3) Edge where it helps. Push inference local for privacy/latency/cost; reserve cloud for heavy lifts; experiment with decentralised networks where sensible.

4) Earn community license. Pair new sites with peak-shaving commitments and household battery programs; discuss rate impacts early, consider cross-subsidy constructs.

5) Shore up the physical stack. Track magnets/materials programs and allied fab capacity; prefer public-private structures that secure inputs and share upside.

17.1. Co-Plan Power and Compute

Objective: Synchronise infrastructure development as single critical path.

Actions:

- Map substation capacity and transformer lead times
- Align cooling envelopes with GPU roadmap
- Coordinate fibre routes with chip orders
- Build contingency into procurement and commissioning
- Treat schedule like construction project (single long-lead slip moves everything)

17.2. Build Multi-Model Layer

Objective: Optimize cost-quality-latency trade-offs per workload.

Actions:

- Define quality bars per use case
- Route to cheapest model meeting each bar
- Maintain champion-challenger framework
- Store prompt variants and evaluation results
- Track improvements with attribution
- Enable confident component swapping

17.3. Edge Where It Helps

Objective: Optimise privacy, latency, and cost through selective local processing.

Actions:

- Offload privacy-sensitive interactions to devices
- Deploy on-device personal context stores

- Clear boundaries between private context and shared corpora
- Reserve cloud for computation-heavy tasks
- Test latency improvements and user satisfaction

17.4. Earn Community License

Objective: Build durable local support for facilities.

Actions:

- Engage councils and residents early
- Publish impacts in accessible language
- Tie community benefits to measurable milestones
- Offer apprenticeships and training programs
- Set local procurement targets
- Demonstrate visible upside in host regions

17.5. Shore Up Physical Stack

Objective: Reduce supply chain exposure.

Actions:

- Track magnets, packaging, memory, sensors, power electronics
- Sign longer-term agreements with volume commitments
- Build testing capabilities for rapid qualification
- Participate in public-private supply programs
- Design for substitution flexibility

18. Technology and Architecture Trends

18.1. Sparsity and Efficiency

Current State: DeepSeek Sparse Attention demonstrates 50% cost reduction potential.

Trajectory: Expect broader adoption of:

- Dynamic computation allocation
- Layer-wise sparsity
- Token-level attention optimisation
- Mixture-of-experts architectures

18.2. Context and Memory

Evolving Capabilities:

- Extended context windows
- Hierarchical memory systems
- Personal context management
- Privacy-preserving retrieval

18.3. Reasoning and Planning

Research Directions:

- Multi-step reasoning chains
- Self-correction mechanisms
- Uncertainty quantification
- Explanation generation

18.4. Multimodal Integration

Current Progress:

- Text-image-video synthesis
- Cross-modal retrieval
- Unified representation spaces

Future State: Seamless integration across modalities with consistent quality and coherent outputs.

19. Strategic Recommendations by Stakeholder

19.1. For Hyperscalers and Infrastructure Providers

Priorities:

1. Secure long-term power agreements before site selection
2. Engage communities 2–3 years before groundbreaking
3. Standardise modular designs for rapid replication
4. Build cross-subsidies into rate structures
5. Invest in workforce training partnerships

19.2. For Enterprise Technology Leaders

Priorities:

1. Implement multi-model routing architecture
2. Establish cost allocation and observability
3. Build reusable safety and evaluation frameworks
4. Experiment with edge deployment for appropriate workloads
5. Document model choices and maintain audit trail

19.3. For Application Developers

Priorities:

1. Design for model-agnostic interfaces
2. Build quality benchmarks and regression testing
3. Implement progressive enhancement (fallback models)
4. Monitor unit economics continuously
5. Plan migration paths between providers

19.4. For Policymakers

Priorities:

1. Harmonise transparency requirements across jurisdictions
2. Support public-private partnerships for critical materials

3. Facilitate grid modernisation and demand response
4. Fund workforce development at scale
5. Create predictable, light-touch regulatory frameworks

19.5. For Investors

Priorities:

1. Evaluate power agreements as fundamental due diligence
2. Assess supply chain exposure and mitigation strategies
3. Understand total cost of ownership beyond chip acquisition
4. Consider regulatory compliance burden
5. Evaluate management experience with infrastructure-scale projects

20. Conclusion: From Demos to Infrastructure

20.1. The Threshold Moment

AI has crossed from viral demonstrations to manufactured intelligence at national-infrastructure scale. This transition demands thinking that integrates:

- Multi-gigawatt power planning
- Sub-$1/M token economics
- Global supply chain resilience
- Community engagement and local benefit sharing

20.2. The Winners' Profile

Organisations that prosper will:

- Synchronise GPU roadmaps with electricity availability
- Convert falling token costs into superior products
- Earn durable community licenses through visible benefit sharing
- Treat energy as first-class design constraint
- Build allied supply chain resilience from fabs to magnets

20.3. The Path Forward

Success is not automatic. It requires:

Regulatory Evolution:

- Templates preserving dynamism
- Balanced innovation and safety
- Reduced compliance fragmentation

Engineering Discipline:

- Energy-first design
- Multi-model optimisation
- Privacy-respecting architectures

Supply Chain Development:

- Domestic capability for shocks
- Allied depth for redundancy
- Public-private risk sharing

20.4. The Opportunity

The components are in motion. Costs are falling, tools are maturing, and supply chains are being rebuilt with intention. With patient execution and respect for host communities, AI can expand access to knowledge and services while creating quality employment across skilled trades.

The task now: Turn plans into projects, and projects into operating assets that serve both users and the places they live.

Report Prepared: October 2025
Next Update: Q2 2026

This report synthesises multiple primary sources and represents analysis current as of publication date. Market conditions, regulatory frameworks, and technical capabilities continue to evolve rapidly.

Conclusion: The State of AI—and Its Future

Your material depicts AI crossing a threshold: from viral demos to manufactured intelligence at national-infrastructure scale. The winners will be builders who can synchronise GPU roadmaps with gigawatt power, convert falling token costs into superior products, and earn a durable community license by sharing the benefits locally.

The path is not automatic. It requires regulatory templates that preserve dynamism, engineering that treats energy as a first-class design constraint, and supply chains—from fabs to magnets—that reflect allied resilience. The good news embedded throughout your sources is that each of these pieces is already in motion.

www.cohesionx.co.za