



Sovereign AI Framework for Developing Nations





Sovereign AI Framework for Developing Nations

Jithin VG, Aharsh MS
{jithinvg, aharsh}@bud.studio

The global AI landscape shows a significant gap in infrastructure between developed and developing countries. For instance, the United States has about 21 times more data center capacity than India. This research shows that software-based optimization strategies, architectural innovations, and alternative deployment models can greatly reduce reliance on large infrastructure. By analyzing current capacity data, emerging optimization techniques, and successful examples like DeepSeek’s cost-effective training methods, this paper demonstrates that developing countries can achieve competitive AI capabilities through strategic software innovations—such as model architecture improvements, federated inference systems, and resource-aware deployment strategies—reducing reliance on massive infrastructure investments and helping to close the 21x infrastructure gap, thereby enabling fuller participation in the global AI ecosystem.

Table of contents

1.	Executive Summary.....	05
2.	Introduction: Towards AI Sovereignty.....	06
3.	The Global Landscape of AI Initiatives and Investments.....	09
3.1.	Public-Private Partnerships.....	14
4.	The Global Race for AI Infrastructure and Energy Sovereignty.....	17
4.1.	Global Data Center Capacity and Expansion.....	17
4.2.	Compute as a Strategic Asset.....	18
4.3.	Power Capacity and Energy Consumption.....	19
5.	National Contributions to Models, Research, and Industry.....	22
6.	Weaponization of AI and Geopolitical Leverage.....	27
6.1.	GPU Export Bans and Their Impact.....	27
6.2.	AI as a Potential Leverage in National Security.....	28
7.	Sovereign AI as National Strategy.....	31
8.	Challenges to Adopting Generative AI in Developing Countries.....	35
8.1.	Digital Divide and Connectivity Gaps.....	35
8.2.	Limited Computing Power and Data Center Capacity.....	37
8.3.	Unreliable Energy Supply and Environmental Concerns.....	38
8.4.	Investment Barriers: The Capital Chasm.....	41
8.5.	High Costs of GenAI Development and Deployment.....	43
8.6.	Reliance on External Capital and Market Dynamics.....	44
8.7.	Impact of International Regulations and Export Controls.....	46
8.8.	Regionalization and Cultural Relevance: The Contextual Imperative.....	47
8.9.	Language Diversity and Low-Resource Languages.....	47
8.10.	Cultural Bias in AI Models and Datasets.....	48
8.11.	Need for Context-Specific Solutions.....	49
9.	Lessons from Previous Technological Revolutions.....	50
10.	Training a GenAI model: Different strategies & Efficient alternatives.....	54
10.1.	Pretraining Techniques.....	55
10.2.	Post-Pretraining Alignment and Fine-Tuning Techniques.....	56
10.3.	Parameter-Efficient Fine-Tuning (PEFT).....	58

11.	Alternatives to Expensive Pretraining.....	61
11.1.	Leveraging Open Models as Base Initializations.....	62
11.2.	Regionalization via Adapters.....	62
11.3.	Model Selection, Routing, and Merging.....	63
11.4.	Continuous Pretraining on Local Data.....	65
11.5.	Examples of Successful Low-Cost Adaptations.....	66
12.	Infrastructure and Optimization.....	67
12.1.	Hardware Constraints: Making Do with What You Have.....	67
12.2.	Efficient Model Architectures and Quantization.....	69
12.3.	Distributed and Hybrid Computing Approaches.....	71
12.4.	Memory and Compute-Efficient Optimizer.....	72
12.5.	Geopolitical and Infrastructure Risk Mitigation.....	74
12.6.	National AI Model Repositories and Ecosystem Building.....	76
12.7.	Hyperscaler and OEM Dependency Risk mitigation	77
12.8.	Resource-Aware Model Architectures.....	80
12.9.	Using Adapters to Prevent Catastrophic Forgetting.....	82
12.10.	Architectures for Low-Data Alignment and Multilinguality.....	83
13.	Efficient Inference for GenAI in Resource-Constrained Environment.....	86
13.1.	Efficient Model Architectures.....	86
13.2.	Split Inferencing (Edge-Cloud Collaboration).....	88
13.3.	Collaborative Decoding Strategies.....	89
13.4.	Distributed and Federated Inferencing.....	91
13.5.	Decentralized LLMs and Prompt Routing.....	93
13.6.	Strategic Use of SLMs Over LLMs.....	95
13.7.	SLM Test-Time and Inference-Time Scaling.....	97
14.	Recommendations.....	100
14.1.	Practical Roadmaps for Different Resource Profiles.....	100

Executive Summary

The global AI landscape shows a significant gap in infrastructure between developed and developing countries. For instance, the United States has about 21 times more data center capacity than India. This research shows that software-based optimization strategies, architectural innovations, and alternative deployment models can greatly reduce reliance on large infrastructure. By analyzing current capacity data, emerging optimization techniques, and successful examples like DeepSeek’s cost-effective training methods, this paper demonstrates that developing countries can achieve competitive AI capabilities through strategic software innovations—such as model architecture improvements, federated inference systems, and resource-aware deployment strategies—reducing reliance on massive infrastructure investments and helping to close the 21x infrastructure gap, thereby enabling fuller participation in the global AI ecosystem.

Key objectives of this Whitepaper

To deliver a policy-oriented, technically grounded roadmap that enables developing nations to achieve functional parity with AI leaders by:

1. **Benchmarking the Global Compute Divide:** Quantify the present gap in datacenter power (e.g., ≈ 21 GW in the U.S. vs. ≈ 1 GW in India), accelerator inventory, energy costs, and talent pools across representative developed and developing countries.
2. **Diagnosing True Constraints:** Distinguish bottlenecks that require capital-heavy fixes (power grids, fabs) from those solvable through software (kernel fusion, quantisation, alternative architectures).
3. **Curating High-Leverage Software Levers:** Catalogue and experimentally validate optimisations—FlashAttention-class kernels, BitNet-style extreme quantisation, Mamba/SSM architectures, DeepSeek-style low-cost training—that together can deliver $\geq 20\times$ aggregate efficiency.

4. **Formulating the “Chandrayaan Way” Framework:** Translate India’s frugal-innovation ethos into a repeatable playbook: design for CPU + edge first, leverage community LoRA/adapters, and federate inference to tap existing client hardware.
5. **Mapping a Phased Implementation Path:** Provide a five-year schedule, investment range, and KPI dashboard to track progress toward sovereignty in AI capability without trillion-dollar hardware outlays.



The ultimate aim is to demonstrate that, through coordinated software innovation, heterogeneous hardware utilisation, and risk-aware policy, resource-constrained nations can achieve parity in practical AI outcomes with far lower capital outlay than traditional “hardware-first” approaches require.

2. Introduction: Towards AI Sovereignty

Sovereign AI refers to a nation’s full control over the entire AI stack—including infrastructure (compute, storage, networking), data (collection, processing, governance), algorithms (models, frameworks, applications), and talent (researchers, engineers, operators). It embodies technological self-determination in the AI era. The strategic value of sovereign AI goes beyond technology. Nations with sovereign AI capabilities can:

1. **Preserve cultural and linguistic identity** by developing AI systems that reflect and understand local contexts.
2. **Ensure data sovereignty** by keeping citizen data within national borders.
3. **Foster economic growth** through homegrown AI innovation and reduced reliance on foreign technology.

4. **Protect national security** by securing critical AI infrastructure
5. **Define AI governance** based on national values and priorities

However, current AI development is largely dominated by a few major technology companies and powerful nations, creating significant risks for developing countries.

The Cost of Dependency

1. **Economic drain:** Relying on foreign cloud-based AI services can cost developing countries billions in foreign exchange each year
2. **Data colonialism:** When citizen data is processed abroad, it compromises national data sovereignty
3. **Cultural erasure:** AI models trained predominantly on Western data often fail to reflect local languages, values, and traditions
4. **Technological lock-in:** Dependence on proprietary AI systems stifles local innovation and limits long-term flexibility
5. **Security vulnerabilities:** Outsourcing critical AI infrastructure increases exposure to foreign interference and cybersecurity threats

Sovereign AI is not merely a technological aspiration; it is a fundamental matter of economic independence and national security.² Nations with robust sovereign AI capabilities gain significant advantages. They can promote digital self-determination, ensuring that algorithmic decision-making respects and protects citizen rights. This builds trust in AI applications deployed in sensitive sectors like healthcare, defense, education, and public safety. Furthermore, it allows nations to maintain economic leverage in global technology markets and support industrial competitiveness through continuous innovation.² The ability to control critical digital infrastructure and align AI systems with democratic values is foundational for building thriving local economic ecosystems around AI innovation, fostering self-reliance and long-term prosperity.²

The broad scope of principles underlying Sovereign AI, encompassing strategic interests, cultural values, legal frameworks, economic independence, and national security, indicates that nations are not simply seeking to acquire AI technology. Instead, the objective is to deeply integrate AI within their societal fabric and governance structures, safeguarding their unique values and ensuring long-term self-determination. This approach signifies a comprehensive national strategy that extends far beyond technical control, embedding AI within a nation's identity and resilience.

AI and the New Geopolitics

The world is currently experiencing a profound shift in international relations, moving away from multilateralism towards a more competitive landscape characterized by unilateralism and bilateralism. Within this dynamic, AI has rapidly emerged as a primary battleground for nations vying for economic and security leadership.³ The intensifying competition between major powers, particularly the United States and China, in the realm of AI is fundamentally reshaping bilateral relations and redefining the global balance of power. This compels nations worldwide to adapt to a rapidly evolving and uncertain international environment.⁴

The pursuit of Sovereign AI is thus driven by a dual imperative: defensive needs and offensive ambitions. On the defensive side, nations seek to protect their sensitive data, intellectual property, and critical infrastructure from foreign influence or disruption. This includes mitigating dependence on foreign-controlled environments and reducing exposure to supply chain disruptions or geopolitical tensions. Concurrently, there are significant offensive ambitions, such as gaining economic leverage, achieving global leadership in key AI sectors, and securing strategic advantage in emerging technologies. This duality highlights AI as a critical tool for both national resilience and the projection of geopolitical influence.

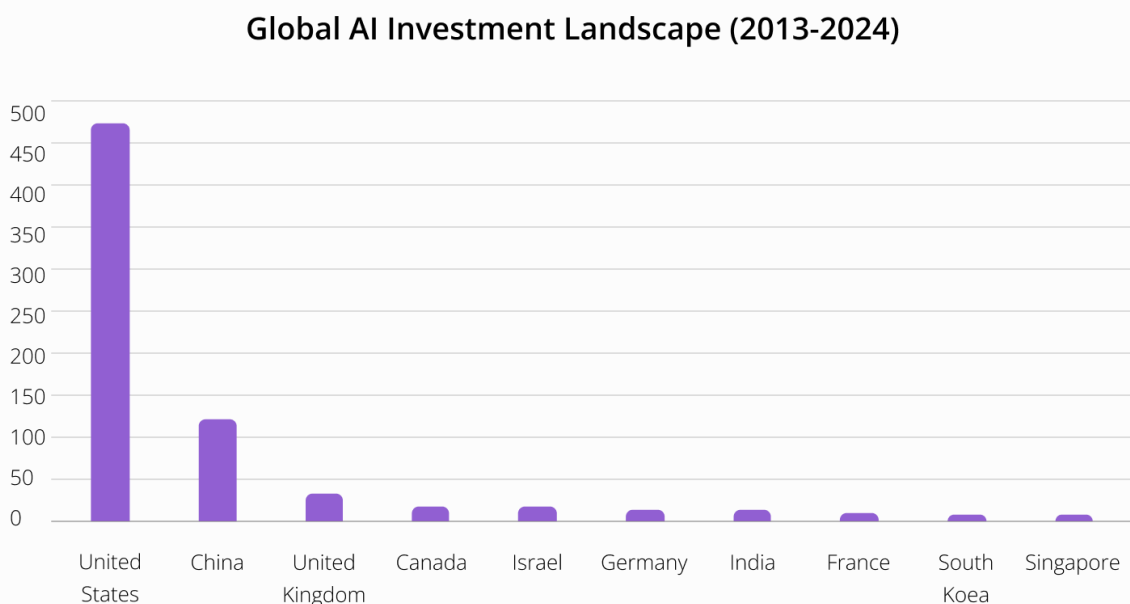


AI has become a pivotal force in the new geopolitical arena—driving both competition and national strategy. Nations are racing to harness AI not just for defense, but to assert dominance on the world stage

3. Global Landscape of AI Initiatives and Investments

The global race for AI supremacy is marked by substantial strategic investments and diverse national initiatives. Nations are committing significant resources to develop their AI capabilities, recognizing AI as a cornerstone of future economic prosperity and national security.

The approaches to fostering AI capabilities vary significantly across leading nations, reflecting their unique economic structures, political systems, and strategic priorities.



Source: Stanford HAI AI Index Report 2025

Image 1: Investments on AI by countries from 2013 to 2024

United States: The U.S. AI policy, particularly under recent administrations, emphasizes a "forward-leaning, pro-innovation, and pro-competition mindset".⁵ The federal government plays a crucial role in supporting AI research and development (R&D) in areas where private sector investment might be insufficient, focusing on national security, public infrastructure resilience, and scientific discovery.⁶ Key initiatives include the White House Office of Management & Budget's (OMB) AI Use & Procurement Requirements for federal agencies, designed to make agencies more agile

and cost-effective.⁵ The Department of Energy has also announced 16 potential federal sites for rapid AI data center construction, with operations targeted by the end of 2027 through public-private partnerships.⁵ An Executive Order on Coal-Powered AI Infrastructure has been issued to identify regions with suitable coal infrastructure to meet the escalating electricity needs of AI data centers.⁵ The U.S. aims to maintain its AI dominance through strategic R&D funding, grand challenges across federal agencies, and procurement reform that streamlines the integration of AI solutions into government operations.⁷ Private AI investment in the U.S. reached an astounding \$109.1 billion in 2024, significantly surpassing other nations; this figure is nearly 12 times higher than China's and 24 times the UK's private AI investment.⁸ Cumulatively, from 2013 to 2024, total private AI investment in the U.S. amounted to \$470.9 billion.¹⁰

China: China is driven by an ambitious strategic vision to become the global leader in AI innovation by 2030, a goal articulated in key policy frameworks such as the Next-Generation AI Development Plan (2017) and the Made in China 2025 initiative.¹¹ A state fund worth 60 billion yuan (approximately US\$8.2 billion) was launched in January by the Ministry of Industry and Information Technology (MIIT) and the Ministry of Finance, specifically for early-stage AI projects. This fund targets the entire AI supply chain, encompassing computing power, algorithms, data, and applications.¹² This initiative contributes to a broader national plan to cultivate a \$150 billion AI industry by 2030.¹² China demonstrated early commitment to AI dominance, accounting for 48% of global AI startup funding in 2017, surpassing the U.S. share of 38%.¹² From 2013 to 2024, China's total private AI investment stood at \$119.3 billion.¹⁰

European Union (EU): The EU's AI Continent Action Plan outlines an ambitious roadmap to position Europe as a global AI leader, supported by a substantial €200 billion program known as the InvestAI Initiative.¹³ A central pillar of this plan is building large-scale AI data and computing infrastructure, including the establishment of at least 13 "AI Factories" and "AI Gigafactories" across Europe. These facilities are designed to be equipped with approximately 100,000 state-of-the-art AI chips to train and develop complex AI models.¹³ The proposed "Cloud and AI Development Act" aims to triple the EU's data center capacity within the next five to seven years by addressing obstacles like suitable locations and energy resources.¹³ Other key pillars of the EU strategy include increasing access to high-quality data through a "Data Union

Strategy," fostering AI adoption in strategic sectors via an "Apply AI Strategy," and strengthening AI skills and talent through various educational and training programs.¹³

United Kingdom (UK): The UK's AI strategy, detailed in the AI Opportunities Action Plan (unveiled in January 2025), focuses on enhancing AI infrastructure, attracting investment, and integrating AI technologies into public services with the goal of positioning the UK as a global leader in the field.¹⁵ The UK is notably "doubling down on US investment," having secured significant commitments from major U.S. firms. Microsoft, for instance, committed £2.5 billion over three years to expand its next-generation AI data center infrastructure, while Amazon is investing £8 billion over five years for data center construction. Other U.S. firms like CyrusOne, ServiceNow, Cloud HQ, and CoreWeave have pledged £6.3 billion in data center infrastructure investments.¹⁶ Initiatives like "AI Growth Zones" are designed to encourage new data center construction in suitable areas, and efforts are underway to optimize access to valuable public datasets for AI innovation.¹⁵

Canada: Canada announced a substantial \$2.4 billion package in Budget 2024 to secure its AI advantage, with \$2 billion specifically allocated to building and providing access to computing capabilities and technological infrastructure for its AI researchers, startups, and scale-ups.¹⁷ The "Canadian Sovereign AI Compute Strategy" is a cornerstone of this investment, comprising up to \$700 million to mobilize private sector investment, up to \$1 billion for building public supercomputing infrastructure (including a \$705 million AI Sovereign Compute Infrastructure Program), and a \$300 million AI Compute Access Fund to support the purchase of AI compute resources by Canadian innovators and businesses.¹⁷ Canada holds the distinction of being the first country to establish a national AI strategy in 2017.¹⁷ From 2013 to 2024, Canada attracted \$15.3 billion in private AI investments.¹⁰

Japan: Japan's 2025 AI governance strategy has shifted towards a pragmatic "light-touch" approach, aiming to establish the country as "the most AI-friendly country in the world".¹⁹ This framework relies heavily on existing sector-specific laws and voluntary risk mitigation by businesses, rather than imposing sweeping AI-specific regulations.¹⁹ In February 2025, the government submitted a draft AI Bill designed to promote AI research, development, and utilization with minimal explicit

penalties for the private sector.¹⁹ Initiatives such as the Ministry of Economy, Trade and Industry's (METI) Generative AI Accelerator Challenge (GENIAC) aim to harness AI for economic growth and societal transformation.¹⁹ Japan actively supports Nvidia-led AI-computing infrastructure, collaborating with domestic cloud leaders like SoftBank and GMO Internet Group.¹⁹ Notably, SoftBank CEO Masayoshi Son and OpenAI CEO Sam Altman announced a joint venture to launch AI services in Japan, underpinned by a US\$3 billion annual licensing agreement for OpenAI technology.¹⁹ From 2013 to 2024, Japan attracted \$5.9 billion in private AI investments.¹⁰

South Korea: South Korea has articulated a clear ambition to achieve "AI G3 Status," positioning itself as a top-three AI nation by 2030.²⁰ A plan approved in April allocates over 1 trillion won (approximately \$1.3 billion) this year to secure 10,000 high-performance Graphics Processing Units (GPUs), which will be distributed to domestic firms, universities, and research institutes developing AI foundation models.²¹ Key projects include establishing a "National AI Computing Center," expanding GPU capacity to 15 times its current size, supporting the commercialization of domestically produced AI chips, and investing 193.6 billion won for Korea's "World Best LLM (large language model) Project".²⁰ The private sector is expected to invest KRW 65 trillion (approximately \$47 billion) in AI development over the next four years (2024-2027), with active government support.²⁰ From 2013 to 2024, South Korea attracted \$7.3 billion in private AI investments.¹⁰

India: India is rapidly building a robust AI computing and semiconductor infrastructure to support its burgeoning digital economy. The IndiaAI Mission, approved in 2024, allocates ₹10,300 crore (approximately \$1.2 billion) over five years to strengthen AI capabilities.²² A significant focus of this mission is the development of a high-end common computing facility equipped with 18,693 GPUs, making it one of the most extensive AI compute infrastructures globally. India also aims to develop its own indigenous GPU within the next three to five years to reduce reliance on imported technology.²² The "BharatGen" initiative, launched in 2024, stands as the world's first government-funded multimodal Large Language Model (LLM) initiative, designed to enhance public service delivery and citizen engagement.²² India also emphasizes an "AI-Ready Data Initiative" to provide anonymized datasets for startups and researchers, and leverages its Digital Public Infrastructure (DPI) model to foster innovation.²² From

2013 to 2024, India attracted \$11.1 billion in private AI investments.¹⁰

United Arab Emirates (UAE): The UAE is rapidly emerging as a global AI leader, aiming to become a top AI hub by 2031. AI is projected to contribute US\$96 billion to its economy by 2030, representing 14% of its GDP.²⁴ The "UAE Strategy for Artificial Intelligence," launched in 2017, focuses on integrating AI across key sectors such as transportation, renewable energy, education, health, and the environment.²⁵ Significant investments include Silver Lake's US\$800 million in G42 and DAMAC Properties' US\$20 billion investment in U.S. data centers.²⁴ The UAE is also expanding AI education through institutions like MBZUAI and actively integrating AI into public services, including traffic management and chatbots.²⁴ From 2013 to 2024, the UAE attracted \$3.7 billion in private AI investments.¹⁰

Saudi Arabia: Saudi Arabia's Vision 2030 drives a major shift towards AI and data-driven innovation, aiming to reduce oil dependence and foster high-tech industries. The Kingdom aspires to be among the top 15 nations in AI by 2030 and to become an exporter of AI-driven solutions.²⁶ Project Transcendence, a landmark US\$100 billion initiative, is designed to accelerate AI and advanced technology adoption.²⁶ At the LEAP 2025 technology conference in Riyadh, Saudi Arabia announced over \$14.9 billion in AI investments.²⁷ A particularly significant commitment includes a \$1.5 billion investment from U.S. AI chip startup Groq to establish the world's largest AI inferencing data center in Dammam, a collaboration with Aramco Digital.²⁷ Saudi Arabia is also partnering with Google, with a co-investment of US\$5-10 billion for AI-focused projects, including the development of Arabic-language AI models.²⁶

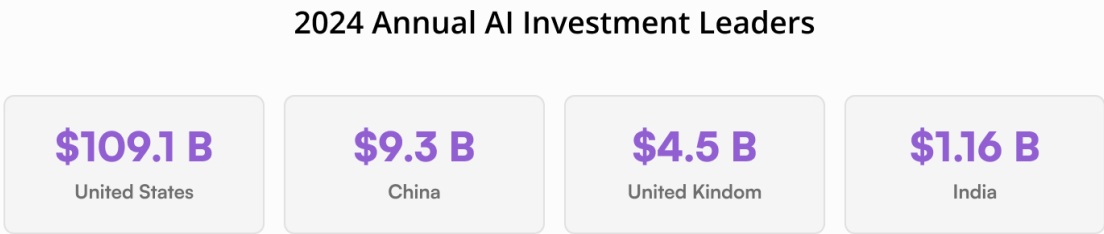


Image 2: Leaders in AI Investments for 2024

3.1 Public-Private Partnerships

Many nations are actively fostering collaboration between government, industry, and academia to accelerate AI development. The U.S. emphasizes a "collaborative approach among policymakers, industry leaders, and global partners" to ensure responsible development and continued innovation.²⁸ Canada's strategy explicitly involves mobilizing private sector investment alongside building public supercomputing infrastructure.¹⁸ Japan's government supports Nvidia-led infrastructure initiatives, and the joint venture between SoftBank and OpenAI exemplifies significant private sector collaboration.¹⁹ Similarly, the UAE and Saudi Arabia are aggressively pursuing public-private partnerships and attracting major global tech firms to invest in and contribute to their burgeoning AI ecosystems.²⁴

The comparison of these national approaches reveals that nations are pursuing AI sovereignty through fundamentally different strategic models. Some, like China and South Korea, adopt a state-led comprehensive investment model, where government funds are directed across the entire AI supply chain. In contrast, the U.S. relies more heavily on market-driven innovation, with strategic government support focusing on foundational research and procurement. Other nations, such as the UK and Japan, emphasize strategic partnerships and attracting foreign investment to bolster their capabilities. Meanwhile, countries like Canada, the EU, and India are focusing on significant domestic infrastructure build-out coupled with intellectual property protection. This divergence indicates that "Sovereign AI" is not a monolithic concept but rather a spectrum of national strategies tailored to each country's unique economic, political, and technological realities.

A common thread across these diverse strategies is the prioritization of investments in foundational elements such as compute infrastructure (GPUs, data centers) and talent development. South Korea's massive GPU acquisition, India's plans for indigenous GPU development and large compute facilities, the EU's "AI Gigafactories," Canada's multi-billion dollar compute strategy, and Saudi Arabia's commitment to the world's largest inferencing data center all underscore a consensus on the critical importance of compute power. Similarly, talent development receives significant attention in the strategies of the EU, U.S., Canada, India, UAE, and Saudi Arabia. This shared understanding suggests that controlling the means of AI production (compute) and

cultivating the necessary human capital are considered non-negotiable for achieving national AI capabilities.

The scale and nature of these national AI investments are direct reflections of intensifying geopolitical competition. Nations are actively vying for technological leadership and economic dominance. China's substantial AI fund, explicitly launched amidst rising global competition and stricter U.S. export controls, illustrates this competitive dynamic.¹² The description of AI as a "battlefield for nations competing for economic and security leadership" further emphasizes this.³ The sheer volume of investment, particularly the vast disparity in private AI investment between the U.S. and other nations, underscores the ongoing "race" for AI supremacy. This indicates that these investments are not solely aimed at fostering economic growth but are strategically designed to secure a critical advantage in a rapidly evolving global power dynamic.

Key National AI Initiatives and Investment Commitments (Selected Countries, 2024-2025)

Country	Key AI Strategy/Plan	Total Investment (if specified, with currency/timeframe)	Primary Investment Focus	Noteworthy Infrastructure Projects/Targets	Key Public-Private Partnerships/Collaborations
United States	Forward-leaning, pro-innovation, pro-competition mindset; National AI R&D Strategic Plan	\$470.9B (private, 2013-2024 sum); \$109.1B (private, 2024)	Foundational R&D, national security, public infrastructure, talent, procurement reform	16 potential DOE sites for AI data centers by late 2025; Coal-powered AI infrastructure	White House OMB, DOE, NIST, AISIC, INASI
China	Next-Generation AI Development Plan (by 2030); Made in China 2025	\$150B (AI industry by 2030); \$8.2B (state fund, early-stage AI projects); \$119.3B (private, 2013-2024 sum)	Entire AI supply chain (compute, algorithms, data, applications), indigenous models, innovation hubs	N/A (focus on domestic capabilities across stack)	Digital Silk Road
European Union	AI Continent Action Plan	€200B (InvestAI Initiative)	Large-scale AI data & computing infrastructure, data access, AI adoption, skills & talent	13 "AI Factories," "AI Gigafactories" (100,000 AI chips); Triple data center capacity (5-7 years)	GAIA-X, OVHcloud, European digital innovation hubs

United Kingdom	AI Opportunities Action Plan	£2.5B (Microsoft, 3 years); £8B (Amazon, 5 years); £6.3B (US firms, data centers)	AI infrastructure, investment attraction, public service integration, data unlocking, talent	AI Growth Zones for data center construction	Microsoft, Anthropic, CyrusOne, ServiceNow, Cloud HQ, CoreWeave, Amazon, Oracle, Salesforce, Cohere
Canada	Canadian Sovereign AI Compute Strategy; Pan-Canadian AI Strategy	\$2.4B (Budget 2024); \$2B (compute & infrastructure); \$1B (public supercomputing); \$705M (AI SCIP); \$300M (AI Compute Access Fund)	Domestic compute capacity, IP safeguarding, talent, AI safety	New AI supercomputing system (Canadian-owned & located); Secure computing facility for government/national security	AI Institutes, Digital Research Alliance of Canada
Japan	"Most AI-friendly country"; Society 5.0	\$5.9B (private, 2013-2024 sum); \$3B/year (SoftBank/OpenAI licensing)	AI innovation, economic growth, social challenges, cybersecurity	Nvidia-led AI-computing infrastructure (with SoftBank, GMO Internet Group)	SoftBank, OpenAI, METI
South Korea	National AI Strategy Policy Directions; AI G3 Status by 2030	\$1.3B (GPU purchases, 2025); KRW 65T (~\$47B) (private, 2024-2027)	GPU acquisition, domestic AI chips, LLM development, AI transformation, talent	National AI Computing Center (15x current GPU size); 10,000 high-performance GPUs	World Best LLM Project, AI semiconductor commercialization initiatives
India	IndiaAI Mission; AI for India 2030	₹10,300 crore (~\$1.2B) (5 years); \$11.1B (private, 2013-2024 sum)	Computing & semiconductor infrastructure, indigenous GPUs, data access, talent, startup support	High-end common computing facility (18,693 GPUs); BharatGen (multimodal LLM); AI-Ready Data Initiative	Sarvam AI, BCG, T-Hub MATH
United Arab Emirates	UAE Strategy for Artificial Intelligence (by 2031)	\$96B (AI contribution to economy by 2030); \$3.7B (private, 2013-2024 sum); \$800M (Silver Lake in G42); \$20B (DAMAC in US data centers)	AI infrastructure, regulations, partnerships, AI-driven sectors (finance, healthcare, energy, defense)	Expanding AI education (MBZUAI); Smart government initiatives	G42, Silver Lake, DAMAC Properties
Saudi Arabia	Vision 2030; National Strategy for Data and AI (by 2030)	\$100B (Project Transcendence); \$14.9B (AI investments at LEAP 2025); \$1.7B (AI-related companies, 2023 funding)	AI capabilities, data-driven innovation, ICT infrastructure, workforce development	World's largest AI inferencing data center (with Groq/Aramco Digital); National Data Bank	Microsoft, Oracle, Huawei, Google, PIF, Groq, Aramco Digital

4. The Global Race for AI Infrastructure and Energy Sovereignty

The rapid advancement of AI is inextricably linked to the underlying physical and energy infrastructure that supports it. This section examines the critical components of this foundation, including global data center capacity, the specialized hardware driving AI, and the escalating power demands that pose both challenges and opportunities.

4.1 Global Data Center Capacity and Expansion

Data centers serve as the physical backbone of the AI revolution, housing the vast computational resources required for training and deploying AI models. The global data center market capacity was estimated at approximately 59 gigawatts (GW) in 2023, with projections indicating a significant expansion to around 122 GW by the end of 2030.²⁹ The United States currently holds a dominant position in this landscape, hosting 51% of the world's over 1,000 hyperscale data centers.³⁰

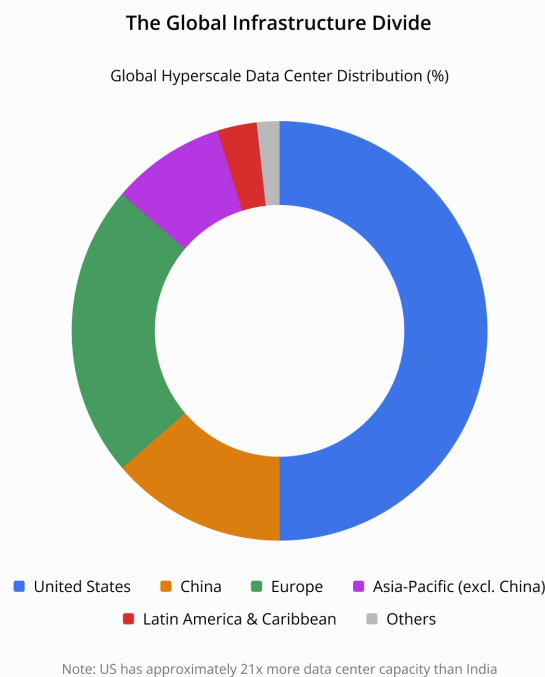


Image 3: Distribution of data centers among countries

The demand for AI-ready data center capacity is projected to rise by 33% between 2023

and 2030, reflecting the increasing adoption of AI workloads. By 2030, it is anticipated that approximately 70% of data centers will be equipped to handle advanced AI workloads.³⁰ In response to this surging demand, countries are actively pursuing expansion of their domestic capacity. The European Union, for instance, aims to triple its data center capacity within the next five to seven years through initiatives like the Cloud and AI Development Act.¹³ Similarly, the UK is launching "AI Growth Zones" to encourage the construction of new data centers.¹⁵ Saudi Arabia is making substantial investments, including plans for the "world's largest AI inferencing data center".²⁷

4.2 Compute as a Strategic Asset

Modern AI infrastructures are fundamentally reliant on specialized accelerators such as Graphics Processing Units (GPUs) or custom AI chips like Tensor Processing Units (TPUs) and Application-Specific Integrated Circuits (ASICs), rather than traditional Central Processing Units (CPUs). This preference stems from the GPUs' ability to perform many calculations in parallel, a necessity for efficiently training complex AI models.³¹ Despite this increasing demand, the underlying hardware is becoming more efficient. Hardware costs for AI computation have been declining at an estimated rate of 30% annually, while the energy efficiency of this hardware has been improving by approximately 40% each year.³³ Nations are prioritizing both the acquisition and domestic development of these critical components to secure their AI sovereignty. South Korea, for example, plans to acquire 10,000 high-performance GPUs and support the commercialization of domestically produced AI chips.²⁰ India has set an ambitious goal to develop its own indigenous GPU within three to five years, aiming to reduce reliance on imported technology.²²



AI's future hinges on advanced compute power—GPUs and custom chips are now the backbone of progress. As demand skyrockets, nations are racing to build and secure their own AI hardware to stay competitive and sovereign,

4.3 Power Capacity and Energy Consumption

The burgeoning AI sector presents a significant challenge to global power grids due to its immense energy requirements. Electricity consumption from data centers was estimated at around 415 terawatt hours (TWh) in 2024, constituting approximately 1.5% of global electricity consumption. This consumption has grown at a rate of 12% per year over the last five years.³⁴ Projections indicate that global electricity consumption for data centers will double to approximately 945 TWh by 2030, growing at about 15% per year—more than four times faster than the growth of total electricity consumption from all other sectors.³⁴

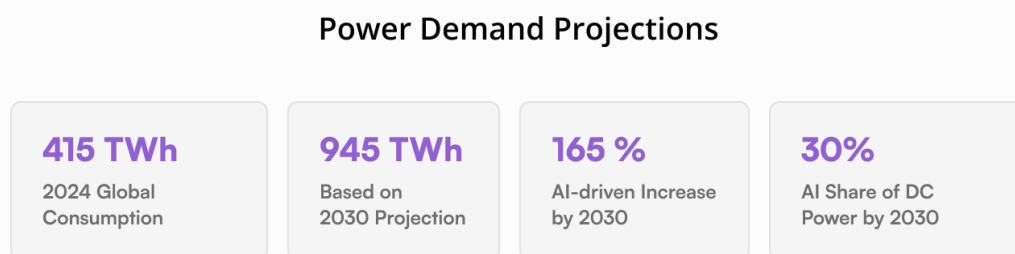


Image 4: Global Data center energy consumption

AI is a primary driver of this surge, projected to increase data center power demand by 165% by 2030 compared to 2023 levels.²⁹ AI workloads accounted for 14% of global data center power usage in 2023, a share projected to rise to 27% by 2027.²⁹ In the U.S., data centers' power demand reached 46,000 megawatts (MW) in Q3 2024, largely driven by AI and cryptocurrency mining.³⁵ Goldman Sachs Research projects that AI-driven data centers will consume an additional 200 TWh annually from 2023 to 2030, potentially accounting for 30% of all global data center power consumption by 2030.³⁵ Concerns are mounting regarding the grid's ability to keep pace with this skyrocketing demand, necessitating significant utility investment.³⁵ Some U.S. states are even exploring the use of coal-powered infrastructure for AI data centers to meet the escalating electricity needs.⁵

The escalating power demands of AI data centers elevate energy supply and grid stability into critical geopolitical and strategic assets. This means that energy sovereignty is becoming increasingly intertwined with AI sovereignty. The massive

and rapidly growing energy consumption, with projections of doubling global consumption by 2030, highlights a fundamental challenge. The explicit questioning of whether the grid can keep up and the potential for power shortages underscore that energy availability is now directly influencing national AI strategy. Consequently, nations with abundant, reliable, and affordable energy sources gain a significant competitive advantage in the AI race, making energy security a direct component of their overall AI capabilities.

The global scramble for GPUs and the push for indigenous chip development underscore the critical vulnerability inherent in AI supply chains. This situation highlights the strategic imperative for nations to control their access to advanced compute power. The emphasis on GPUs as essential for AI model training and the rapidly increasing computational demands of these models create intense pressure. The initiatives by South Korea and India to acquire or develop their own GPUs directly reflect this concern. Reliance on a few dominant manufacturers, such as Nvidia, creates a single point of failure or leverage in the global supply chain. Therefore, achieving "Infrastructure Autonomy," as defined in the core principles of Sovereign AI, becomes paramount for true national AI capabilities, pushing nations towards domestic production or highly diversified technology stacks.



The global race for AI dominance hinges not just on data and talent, but on secure, sovereign access to compute power. As nations confront the risks of over-reliance on a few GPU manufacturers, "Infrastructure Autonomy" emerges as a strategic necessity, not a luxury. At the same time, the growing energy demands of AI are forcing a critical reckoning with sustainability. These twin pressures—technological dependence and environmental strain—are driving innovation in both domestic chip development and green AI infrastructure.

The immense energy footprint of AI poses a significant environmental challenge, which could become a limiting factor for national AI ambitions. However, this challenge also serves as a potent driver for innovation in energy-efficient hardware

and sustainable data center operations. The increasing carbon emissions from AI training, with large models like Llama 3.1 405B (2024) emitting thousands of tons of carbon, and the projected rise in data center carbon emissions, indicate that unchecked growth is unsustainable. This environmental pressure, combined with the strain on power grids, will likely compel nations to invest heavily in "AI Energy Efficiency & Sustainability," transforming environmental concerns into a catalyst for advancements in green computing and potentially shaping future regulatory frameworks for AI infrastructure.

Global Data Center Power Demand Projections (2023-2030)

Year	Total Global Data Center Power Consumption (TWh)	Percentage of Global Electricity Consumption (%)	AI's Share of Data Center Power Consumption (%)	Projected Growth Rate (YoY/CAGR)	Notable Country-Specific Consumption/Growth
2023	~360 TWh (est.)	~1.3% (est.)	14%	N/A	N/A
2024	415 TWh	1.5%	N/A	12% (last 5 years)	US: 540 kWh/capita
2027	84 GW (projected)	N/A	27%	N/A	US: 46,000 MW (Q3 2024)
2030	945 TWh	~3%	30%	15% (2024-2030)	US: +130% (from 2024), ~1200 kWh/capita; China: +170% (from 2024); Europe: +70% (from 2024)

Note: TWh (Terawatt-hours) is a measure of energy consumed over time, while GW (Gigawatts) and MW (Megawatts) are measures of power capacity at a given moment. Conversions are approximate where necessary for consistency.

5. National Contributions to Models, Research, and Industry

The AI software ecosystem, encompassing model development, research output, and the emergence of national AI companies, forms the intellectual and commercial core of AI sovereignty. This section analyzes the global landscape of these contributions.

Model Contributions and Innovations by Countries

The development of notable AI models has increasingly shifted towards the industry sector, which accounted for nearly 90% of such models in 2024, a significant increase from 60% in 2023.³² This trend indicates a pronounced movement of frontier AI development from academic institutions to private enterprises. The United States continues to lead as the primary source of notable AI models, producing 40 in 2024, substantially more than China's 15 and Europe's combined total of three.⁸

Despite the U.S. lead in quantity, Chinese models have rapidly narrowed the quality gap. Performance differences on major benchmarks, such as MMLU (Massive Multitask Language Understanding) and HumanEval, shrank from double digits in 2023 to near parity in 2024.⁸ Key innovations in the AI model landscape include the rise of highly capable Small Language Models (SLMs)³³ and a dramatic reduction in AI inference costs. For instance, the cost of querying an AI model equivalent to GPT-3.5 (64.8% on MMLU) dropped over 280-fold from November 2022 to October 2024.³²

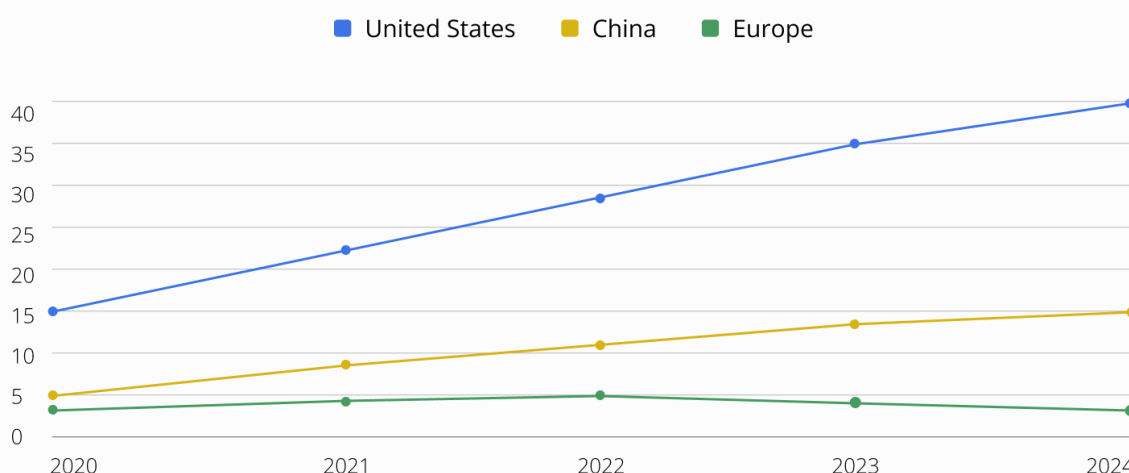
Examples of national model contributions include China's DeepSeek, which notably claims to have trained large language models (LLMs) using only a fraction of the computing power required by some top U.S.-made LLMs.¹¹



India's BharatGen stands out as the world's first government-funded multimodal LLM initiative, launched in 2024 to enhance public service delivery.²² In the open-source domain, Meta's Llama series and Mistral AI's models are recognized as key players.³³

AI Research Output and Influence

In terms of raw scientific output, China leads the global landscape in AI scientific publications, holding a 36.05% share in 2023 and demonstrating significant year-on-year growth.³⁸ India and the United States follow, with shares of 12.13% and 11.98% respectively in 2023.³⁸ While China produces the highest volume of AI research publications, the United States maintains a lead in highly influential research, specifically among the top 100 most cited publications.³²



Source: AI Index 2025 Report

Image 5: Notable AI models by region

Overall AI publication activity continues to grow robustly, nearly tripling between 2013 and 2023, and increasingly dominates computer science publications.³² Academia remains the single leading institutional producer of highly cited research.³² National AI research centers and university programs are recognized as crucial components of a sovereign AI ecosystem, fostering the intellectual capital necessary for long-term AI leadership.²

National AI Company Landscapes

The United States continues to lead significantly in private AI investment, attracting \$109.1 billion in 2024.⁸ Major U.S. AI companies include Nvidia, dominant in AI Hardware and GPUs; Google (Alphabet), a leader in AI Platforms and search integration; Amazon, with strong presence in AI in healthcare and cloud services; and Microsoft,

also active in AI in healthcare and cloud computing.⁹

China's AI industry, while receiving substantial private investment, is heavily state-backed, with a new \$8.2 billion fund focusing on the entire AI supply chain.¹² In Europe, Germany demonstrates high profitability among its AI companies, with 31% reporting profitability, and a strong focus on robotics and automation, aligning with its manufacturing base. Aleph Alpha is a notable German company specifically developing "Sovereign AI, Foundation Models".⁹ France boasts the fastest-growing AI sector in Europe, with French AI startups having attracted €11.2 billion in funding.⁹ India's Generative AI ecosystem has experienced remarkable growth, with 80% of Indian companies considering AI a core strategic priority.²² Infosys and Persistent Systems are key players in India's AI landscape, investing heavily to enhance service offerings and operational efficiencies.⁹ Globally, private investment in Generative AI showed strong momentum, attracting \$33.9 billion in 2024, an 18.7% increase from 2023.³³



The U.S. maintains a dominant lead in private AI investment, attracting \$109.1 billion in 2024, far outpacing other nations. This private-sector driven model appears highly effective in commercializing AI research into market-leading products.

The differing patterns in AI research output, where China leads in publication volume while the U.S. excels in highly influential research and notable model development, suggest distinct pathways to innovation and leadership, each with implications for long-term AI supremacy. The raw number of publications does not necessarily equate to breakthrough innovation or market-leading products. The U.S. model, largely driven by massive private investment, appears to be more effective at translating research into impactful, commercialized models. China's volume, conversely, might be part of a broader, state-backed effort to build a comprehensive knowledge base across the AI spectrum. Understanding this dynamic is crucial for discerning where true "sovereignty" in AI innovation resides.

The increasing prevalence of open-source AI models, with 65.7% of new models in 2024 being open-source, presents an interesting challenge for the concept of sovereign AI. While open-source development democratizes access to AI technologies and fosters widespread innovation, it simultaneously implies less proprietary control for individual nations over foundational models. This development suggests a shift in the focus of sovereignty from outright "ownership" of every component to a greater emphasis on "governance and customization." If foundational models are freely available, nations may not need to develop every model from scratch. Instead, their strategic focus could pivot towards fine-tuning these models, developing specific applications, and ensuring their ethical alignment within national legal frameworks. This approach could potentially reduce the "cost of sovereignty" in model development but elevates the importance of robust regulatory control and the cultivation of talent capable of leveraging and adapting open-source solutions to national needs.

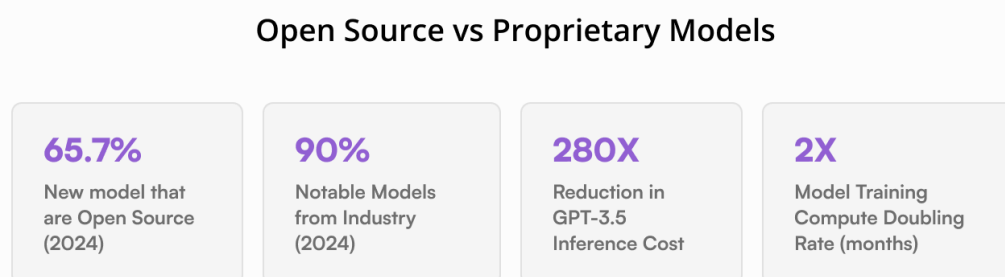



Image 6: Key trends of open-source models

Nations are actively fostering their own AI companies and ecosystems, aiming to cultivate domestic AI champions. This strategy is driven by a desire to ensure local control, maximize economic benefits, and align AI development with national values, even when it means competing with established global tech giants. The explicit focus of Germany's Aleph Alpha on "Sovereign AI, Foundation Models," India's government-funded "BharatGen" initiative, and China's emphasis on indigenous models like DeepSeek all point to a deliberate strategy to nurture national leaders in the AI space. This approach is motivated by the need to retain full control over sensitive data, comply with national regulations on data residency, and build local economic ecosystems around AI innovation, rather than relying solely on foreign providers. This highlights a strategic tension between the benefits of globalized AI development and

national aspirations for self-determination and technological independence.

AI Research Publication Share by Country (2023)

Rank	Country	Share of Global AI Scientific Publications (%)	Year-on-Year Growth (%)	5-Year CAGR (%)
1	China	36.05	+16.88	+7.92
2	India	12.13	+9.17	+9.26
3	United States	11.98	+6.23	+3.89
4	Japan	3.39	-0.74	+5.83
5	Germany	3.21	+6.28	+3.60
6	United Kingdom	2.82	+1.37	+2.47
7	Indonesia	2.27	+21.07	+27.24
8	Italy	2.23	+7.03	+3.35
9	Australia	1.69	+7.34	+3.04
10	Canada	1.66	+5.21	+2.67



The growing dominance of open-source AI models, accounting for 65.7% of new models in 2024, is redefining AI sovereignty from outright ownership to governance and customization. This shift allows nations to focus on fine-tuning open-source models, developing specific applications, and ensuring ethical alignment within national frameworks, effectively reducing the "cost of sovereignty" in model development while elevating the importance of robust regulatory control and talent cultivation.

6. Weaponization of AI and Geopolitical Leverage

The dual-use nature of AI, with its potential for both transformative civilian applications and profound military implications, presents a significant challenge in the geopolitical landscape. This section explores how AI is being weaponized, from export controls on critical hardware to its integration into national security strategies and surveillance technologies.

6.1 GPU Export Bans and Their Impact

The U.S. government has progressively tightened controls on the export of advanced semiconductor technology, devices, and tools to China, primarily to maintain U.S. leadership in this critical sector and address national security concerns.³⁹ These restrictions have targeted Nvidia's most advanced chips, including the H100, H200, and Blackwell series. As of April 2025, these controls were extended to even the less advanced H20 chip, explicitly citing the risk that such chips could be used in Chinese supercomputers for military purposes.⁴⁰ Nvidia anticipates incurring charges of up to \$5.5 billion in its fiscal Q1 due to these export restrictions on its H20 chip for China.⁴⁰ The portion of Nvidia's total revenues from China sales has already seen a significant decline, from 26% in 2022 (before restrictions) to an estimated 13% in fiscal 2025.⁴⁰

The consequences for targeted nations, particularly China, have been complex and, in some respects, counterintuitive. U.S. export controls have paradoxically accelerated China's efforts to achieve self-sufficiency in semiconductor design and production.³⁹ Operating under constrained computing environments, Chinese AI engineers are innovating in ways that prioritize efficient use of computing power. This is exemplified by DeepSeek's claim of training large language models (LLMs) with only a fraction of the computing power needed by U.S. models.³⁷ Despite the controls, circumvention efforts persist, with reports of a robust black market for restricted chips and companies finding ways to access computing resources located elsewhere.³⁷ There have been documented instances of large-scale smuggling of banned Nvidia GPUs into regions like Malaysia for re-export to China.³⁹ Furthermore, China has made notable strides in indigenous chip development, with Huawei producing its Ascend 910B chip and

Alibaba Group unveiling the C930 central processing unit based on the RISC-V architecture to counter U.S. restrictions.³⁹ Chinese scientists have also reported breakthroughs in developing carbon nanotube-based chips capable of running AI tasks.³⁹

The impact on global supply chains and U.S. competitiveness is also significant. U.S. and allied semiconductor makers have experienced substantial revenue losses due to curtailed sales in China, which directly affects their ability to fund the high levels of R&D characteristic of the industry.³⁹ These export controls risk isolating U.S. firms from the global market and inadvertently creating opportunities for Chinese competitors to fill the void.⁴¹ This "counterproductive" approach could hasten China's progress in AI chips by incentivizing domestic innovation and pushing international customers towards Chinese suppliers who can no longer rely on American technology.⁴¹

The U.S. export controls, intended to impede China's AI progress, are inadvertently accelerating China's domestic innovation and self-sufficiency in critical AI hardware and software. This dynamic potentially undermines U.S. long-term leadership and global market share. The explicit statements that "scarcity fosters innovation" and that "blocking access weakens U.S. competitiveness and accelerates China's efforts to build a self-sufficient chip industry" highlight a complex causal relationship. A policy designed for denial is, in effect, strengthening the adversary's indigenous capabilities, making true "sovereignty" through external control difficult to achieve and potentially counterproductive. The direct economic cost to U.S. firms, exemplified by Nvidia's substantial charges, further illustrates this boomerang effect.

6.2 AI as a Potential Leverage in National Security

AI has rapidly transformed into a "strategic asset" and a "battlefield for nations competing for economic and security leadership".³ The pursuit of AI supremacy is accelerating geopolitical rivalries and fundamentally redefining the global balance of power.⁴ Its incorporation into national security, economic policies, and social governance means no major power can afford to overlook or deny its significance.⁴

Military Applications of AI by Various Countries:

- **United States:** The Department of Defense is significantly scaling up AI

integration into its military operations. The potential value of all AI-related federal contracts increased by almost 1200% from 2022 to 2023, with the Pentagon accounting for the vast majority of this spending.⁴² Initiatives include Project Thunderforge, which integrates AI agents into military planning, decision-making workflows, wargaming simulations, and strategic assessments.⁴² The U.S. also operates autonomous warships like the Sea Hunter, capable of extended operations without human interaction, and utilizes AI for intelligence, surveillance, and reconnaissance through programs like Project Maven, which identifies objects and people from drone footage.⁴³

- **China:** China's military strategy, known as "intelligentized AI warfare," integrates AI across all aspects of modern war. This approach views AI weapons as enhancements to existing military systems rather than merely introducing independent weapons. This includes AI-powered psychological operations, behavioral analysis of enemies, social media manipulation, automated signals intelligence operations, and the development of real-time reactionary tactics.⁴³
- **Israel:** AI capabilities are deeply integrated into Israeli military operations, with major tech companies like Palantir, Google, Amazon, and Microsoft providing AI services. Systems such as "Gospel" and "Lavender" are used for target identification, sifting through intelligence, and pinpointing targets for drone strikes. Israel also deploys AI in ground and robotic vehicles, which have been involved in numerous engagements.⁴²
- **India:** India established its Defense Artificial Intelligence Council and Defense AI Project Agency in 2022. It has been utilizing AI in its intelligence, reconnaissance, and surveillance systems since 2021 and invests millions in AI-powered UAVs, drone swarms for offensive engagements, autonomous combat vehicles, and robotic surveillance platforms for high-altitude outposts.⁴³
- **Russia/Ukraine:** The ongoing conflict has seen significant deployment of AI-powered technology and weapons by both sides, with some observers referring to it as a "drone war." AI-powered navigation and drone swarms have notably

improved military operations, and Ukraine is using AI to intercept Russian communications and extract critical information.⁴³

- **South Korea:** South Korea has developed advanced AI-powered weaponry, including hand-held weapons and the "Super aEgis II" machine gun, which can autonomously identify, track, and engage targets up to four kilometers away, even at night. The nation is also working towards AI technology capable of independently determining whether a human is an enemy or a friend.⁴³

The extensive evidence of AI's military applications across leading global powers, ranging from autonomous weapons to intelligence gathering and strategic planning, demonstrates that AI is not merely enhancing existing military capabilities but fundamentally reshaping the nature of warfare and strategic advantage. Consequently, national control over AI development and deployment has become an existential security concern. The image below shows the AI adoption readiness and challenges of developing and developed countries.

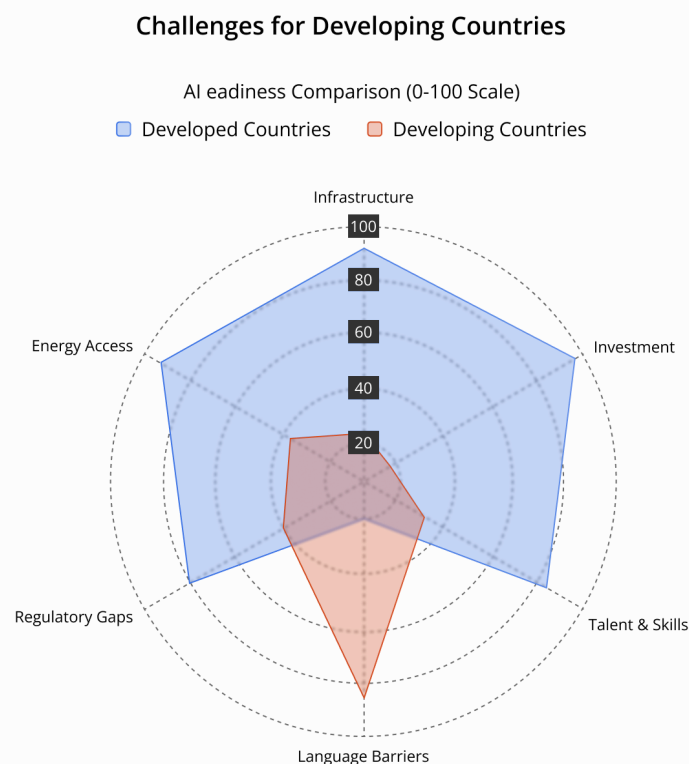


Image 7: AI adoption readiness and challenges of developing and developed countries

7. Sovereign AI as National Strategy

The preceding analysis unequivocally establishes the profound relevance of Sovereign AI as a multifaceted imperative for nations in the 21st century. Its significance is rooted in the critical need for autonomy, resilience, economic competitiveness, and ethical alignment in an increasingly complex and technologically driven global landscape.

Autonomy and Control

Sovereign AI is fundamentally about a nation's independent ability to develop, deploy, and govern AI systems without undue reliance on foreign entities.¹ This independence is paramount for protecting national data, intellectual property, and strategic interests.² For example, data sovereignty ensures that sensitive national data, such as patient information used for diagnostic models, remains within the country's legal jurisdiction, preventing its exposure to foreign control or exploitation.² Similarly, national ownership or full transparency of AI models allows nations to validate, customize, and audit these systems to align with local regulations and cultural values, thereby preventing foreign influence on critical algorithmic decision-making processes that could impact public services or national security.² In an era of increasing geopolitical competition, autonomy over AI capabilities is essential to prevent technological dependency, safeguard national security from potential backdoors or foreign leverage, and ensure that AI development serves national rather than external interests.



Sovereign AI is a critical imperative for nations, ensuring autonomy, resilience, economic competitiveness, and ethical alignment in a technologically driven world. It enables countries to develop, deploy, and govern AI systems independently, safeguarding national data and strategic interests. This approach fosters in-house capabilities, reduces vulnerability to supply chain disruptions, and drives domestic innovation and economic growth.

Resilience and Continuity

Cultivating in-house AI capabilities and diversified technology stacks is a direct strategy to reduce exposure to supply chain disruptions, export controls, and geopolitical tensions.² The impact of U.S. GPU export bans on China, leading to billions in charges for affected companies and accelerating China's indigenous chip development, serves as a stark reminder of the fragility of global supply chains and the imperative for self-reliance.³⁷ Nations like China and India are actively pursuing comprehensive investments in their AI supply chains and indigenous chip development to build resilience against such external pressures.¹² Geopolitical volatility and the weaponization of technology necessitate that nations build robust, resilient AI ecosystems capable of withstanding external shocks and ensuring the continuity of critical AI-driven services and defense capabilities.

Economic Competitiveness and Innovation

Sovereign AI plays a crucial role in fostering domestic AI ecosystems, nurturing talent pipelines, and cultivating national AI champions. This, in turn, drives economic growth and helps nations maintain leverage in global technology markets.² Significant investments in national AI computing centers, data platforms, and startup ecosystems—such as Canada's AI Compute Strategy, India's IndiaAI Mission, and Saudi Arabia's Project Transcendence—are explicitly designed to stimulate local innovation and create new high-value industries.¹⁸ AI is projected to contribute trillions to the global economy, with estimates ranging from \$7 trillion for China by 2030 to \$500 billion for India by 2025, and a potential \$15 trillion globally by 2030.¹² Nations that control their AI destiny are better positioned to capture this immense economic value, create high-quality jobs, and ensure that the benefits of AI accrue domestically, thereby avoiding a future where economic prosperity is dictated by foreign technological monopolies.⁴⁷

Ethical Alignment and Regulatory Control

A core principle of Sovereign AI is the development of AI systems that are aligned with national values, legal frameworks, and ethical norms.² This involves establishing robust regulatory bodies and frameworks to ensure responsible innovation, transparency, and fairness in AI deployment.¹⁵ Examples include the European Union's

comprehensive AI Act ² and Canada's new AI Safety Institute ¹⁷, both of which aim to address critical risks such as lack of accountability, algorithmic bias, and cybersecurity vulnerabilities.⁵ As AI becomes increasingly pervasive in decision-making processes, from public services to defense, ensuring that these systems reflect societal values and protect citizen rights is paramount. Sovereign AI empowers nations to shape AI governance according to their unique ethical and legal landscapes, preventing the imposition of foreign norms or the erosion of democratic principles through unchecked AI deployment, particularly concerning surveillance technologies.⁴⁴

The development of AI presents a paradox for globalization. While AI thrives on global collaboration, open-source contributions, and cross-border data flows, the strategic imperative of "Sovereign AI" is simultaneously driving a counter-trend towards national self-sufficiency and control. This creates an inherent tension between globalized innovation and nationalistic technological protectionism. The report highlights both extensive global collaboration, such as the SoftBank/OpenAI joint venture in Japan, and the dominance of U.S. private investment, alongside strong nationalistic drives like China's push for self-sufficiency and the EU's large-scale infrastructure build-out aimed at reducing foreign dependence. This fundamental tension will likely continue to shape the future of global AI governance, trade policies, and technological development.

The concept of Sovereign AI is not merely an aspirational goal but a direct, strategic response to the weaponization of AI and its underlying technologies as tools of geopolitical leverage. The detailed accounts of GPU export bans by the U.S. government, intended to impede China's AI progress, clearly illustrate how AI and its components are being used as instruments of power and control. Sovereign AI, with its emphasis on "Resilience and Continuity" and "Infrastructure Autonomy," directly addresses these vulnerabilities. It represents a proactive measure by nations to ensure their stability and maintain their power in a world where technological dependency can be exploited for strategic advantage.

Furthermore, nations are increasingly recognizing that a failure to achieve AI sovereignty could lead to significant economic losses and societal disadvantages, creating a sense of urgency akin to Japan's "2025 digital cliff." This term, used to describe projected scenarios where society-wide failures to adopt digital systems could

incur massive economic losses, can be generalized to the broader AI landscape. If a nation lacks control over its AI infrastructure, data, and models, it risks being left behind in the global economic transformation, becoming perpetually reliant on foreign powers for critical services, and potentially compromising its national security. This adds a critical, time-sensitive dimension to the pursuit of Sovereign AI, underscoring the imperative for rapid and decisive action.

Global Private AI Investment by Country (2013-2024/2025 Sum)

Rank	Country	Total Private Investment (in USD, Billions)	Notable % Change (YoY or since 2023)
1	United States	470.9	N/A (2024: \$109.1B)
2	China	119.3	-1.9% (since 2023)
3	United Kingdom	28.2	N/A (2024: \$4.5B)
4	Canada	15.3	N/A
5	Israel	15.0	N/A
6	Germany	11.3	N/A
7	India	11.1	N/A (2024: \$1.16B)
8	France	9.0	N/A
9	South Korea	7.3	N/A
10	Singapore	7.3	N/A



The United States dominates global private AI investment, attracting a staggering \$470.9 billion between 2013 and 2024.

8. Challenges to Generative AI Adoption at Scale in Developing Countries

Generative Artificial Intelligence (GenAI) presents an unparalleled opportunity for developing countries to accelerate economic growth and address persistent societal challenges across vital sectors such as healthcare, agriculture, and education. However, the realization of this transformative potential is significantly impeded by a complex and interconnected array of foundational barriers. These critical challenges span inadequate digital and energy infrastructure, a substantial disparity in AI investment coupled with high development costs, a severe talent and digital literacy gap exacerbated by brain drain, nascent and inconsistent regulatory frameworks, and significant issues related to the localization and cultural relevance of AI models.

These challenges are not isolated; they form a compounding vulnerability that risks widening global inequalities and entrenching technological dependency. Overcoming these impediments necessitates a holistic, multi-stakeholder approach. This includes strategic domestic investments in foundational infrastructure and human capital, targeted international cooperation to bridge resource gaps and foster knowledge transfer, and the development of adaptive, context-specific policy frameworks to ensure inclusive and sovereign AI development. Without concerted and coordinated efforts, the promise of GenAI for developing nations may remain largely unfulfilled, further entrenching existing global disparities.

The aspiration for developing countries to embrace Generative AI at scale confronts a formidable barrier in their existing infrastructure. This challenge manifests across digital connectivity, computing power, and reliable energy supply, creating a foundational bottleneck that impedes widespread AI adoption.

8.1 Digital Divide and Connectivity Gaps

The fundamental "digital divide" between developed and developing countries represents a primary impediment, exacerbating inequalities in access to AI technologies, the necessary underlying infrastructure, and essential digital literacy skills. This disparity significantly limits the equitable distribution of AI's potential

benefits across populations.¹ Global internet access remains far from universal. In 2024, a staggering 2.6 billion people—one-third of the global population—were still offline, indicating that universal connectivity is a distant prospect.⁸

The disparity in internet penetration is particularly stark: only 27% of the population in low-income countries uses the internet, compared to 93% in high-income countries. While low-income economies exhibit the highest annual growth rate in internet use (8.5% in 2024), this pace is insufficient to close the existing connectivity gap in the foreseeable future.⁸ Regionally, Africa records the lowest average internet penetration at merely 38%.⁸ Within Latin America and the Caribbean (LAC), internet access is highly uneven, with significant variations across countries and regions, leading to disparities in access to data storage, cloud computing, and AI capabilities.¹²

The impact of this digital chasm extends profoundly to education. A concerning two-thirds of the world's school-age children (1.3 billion aged 3-17) lack internet connection in their homes. This figure rises dramatically to 95% in West and Central Africa, 88% in East and Southern Africa and South Asia, and 75% in the Middle East and North Africa.⁹ This profound "digital canyon" directly impedes children's ability to connect online, compete in the modern economy, and access education, particularly critical during periods of school closures.⁹ Furthermore, a lack of reliable internet connectivity in remote areas and a shortage of digital devices in schools significantly hinder students' engagement with modern educational tools and resources.¹³ This current deficit in digital access for children is not merely a contemporary inconvenience; it fundamentally limits their ability to develop foundational digital literacy skills. As these children mature, they will enter a workforce increasingly shaped by AI, yet without the basic competencies required.



The persistent "digital divide" profoundly exacerbates global inequalities, leaving 2.6 billion people offline in 2024. This disparity in internet access, especially stark in low-income countries and regions like Africa, critically impedes equitable access to AI technologies and essential digital literacy.

8.2 Limited Computing Power and Data Center Capacity

The effective implementation and scaling of advanced AI technologies, particularly GenAI, are critically dependent on robust infrastructure, including high-speed internet, reliable electricity, and access to modern computational devices.¹⁰ Developing nations frequently possess limited technological infrastructure, which significantly impedes their capacity to deploy and scale GenAI solutions effectively.¹ AI infrastructures are heavily reliant on specialized hardware such as Graphics Processing Units (GPUs) or similar accelerators, which are indispensable for the parallel processing required to train complex AI models.¹⁴

The global landscape of AI development and deployment is characterized by a high concentration of resources. Wealthy nations and a few large tech corporations largely control the development and deployment of AI technologies, leading to significant resource concentration.¹⁰ This concentration means low-income nations often lack the necessary financial and technological resources to either compete in AI innovation or fully benefit from its advancements.¹⁰ The prohibitive costs associated with the infrastructure and computational power required for training and deploying AI models make them largely inaccessible for smaller businesses in developing countries, forcing them into reliance on external, often costly, solutions provided by dominant tech giants.¹⁰

Global investment trends underscore this challenge: global spending on AI data centers alone is projected to exceed \$1.4 trillion by 2027.¹⁴ Hyperscale data centers, which are synonymous with AI data centers due to their immense data processing capabilities, are overwhelmingly concentrated in developed nations, with the U.S. alone housing 51% of the world's over 1,000 hyperscale centers.¹⁵ In Latin America and the Caribbean (LAC), the region accounts for a mere 4.8% of global data center infrastructure, a modest share compared to the U.S. (38.5%) and G7 countries collectively (17.7%).¹² Within LAC, Brazil hosts 37.2% of the region's data centers and its only two hyperscalers, highlighting internal regional disparities.¹² Similarly, the Asia-Pacific region, despite its large population and digital fluency, is home to only 26% of the world's existing hyperscale data capacity.¹⁶

A critical consequence of this limited domestic computational capacity is that countries with insufficient infrastructure must rely on data storage and processing facilities located in other nations, thereby subjecting their data to the privacy regulations and legal frameworks of those foreign jurisdictions.¹² This directly undermines the principle of "Sovereign AI," which emphasizes independent development, deployment, and governance of AI systems aligned with national interests and legal frameworks, including infrastructure autonomy and data governance.¹⁷ This dependency is not merely a technical or economic issue; it carries significant geopolitical implications. It limits a nation's economic competitiveness, stifles local innovation, and creates a vulnerability to external political and economic pressures. This reliance effectively constitutes a form of "digital colonialism"¹⁸, where developing countries cede control over a critical future technology, hindering their ability to align AI development with their unique cultural values and strategic interests.

8.3 Unreliable Energy Supply and Environmental Concerns

Generative AI models are characterized by their increasing size, computational demands, and energy intensity.¹⁹ The training compute for notable AI models is doubling approximately every five months, indicating an exponential growth in energy requirements.¹⁹ The environmental footprint is significant: carbon emissions from AI training are steadily increasing, with models like GPT-4 (2023) emitting 5,184 tons and Llama 3.1 (2024) emitting 8,930 tons of carbon, a substantial amount compared to the average American's annual emissions.¹⁹

Projections indicate that AI-driven data centers will consume an additional 200 terawatt-hours of electricity annually from 2023 to 2030, potentially accounting for 30% of all global data center power consumption by 2030.²⁰ Overall, global electricity consumption for data centers is projected to more than double to approximately 945 TWh by 2030.²¹ The energy demands of AI data centers are immense; a typical AI-focused data center consumes as much electricity as 100,000 households, with the largest ones under construction consuming 20 times that amount.²¹ These data centers require substantial amounts of both electricity and water, placing considerable pressure on national grids and local water resources.¹² Even a relatively small data center can consume up to 25.5 million liters of water annually solely for cooling purposes.¹² In

many rural or underdeveloped areas of developing countries, frequent power shortages and unreliable electricity grids directly prevent the effective and consistent use of AI-powered tools.¹⁰

AI's Energy & Environmental Impact

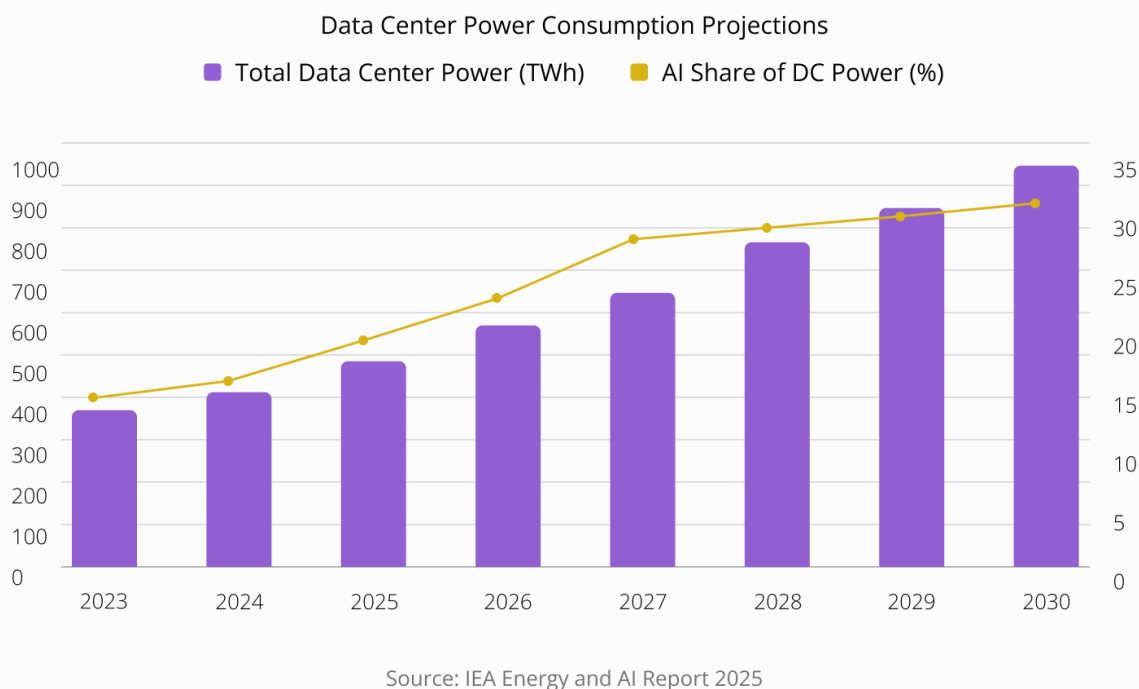


Image 8: AI's energy and environmental impact, year by year

This situation creates a critical energy-AI paradox for developing countries. While AI is presented as a powerful tool to help achieve sustainable development goals (e.g., improving agriculture, healthcare, and urban planning)³, its foundational requirement for massive energy and water resources directly strains these same critical resources and contributes to environmental concerns like increased carbon emissions.¹⁹ Many developing countries already struggle with unreliable energy supply, energy poverty, and water scarcity.¹⁰ This means that without massive, rapid investment in renewable energy infrastructure and highly energy-efficient data center technologies, scaling AI will either exacerbate their existing energy and water crises or force them to rely on carbon-intensive energy sources, thereby hindering their climate commitments. This makes "green computing" not just an environmental ideal but an economic necessity for viable AI adoption.¹⁴ The broader implication is that foreign investments in AI

infrastructure in developing countries must be inextricably linked to sustainable energy solutions. Otherwise, such investments could create long-term environmental liabilities and perpetuate a cycle of resource depletion, ultimately undermining the very notion of sustainable development.

Internet Penetration Rates and Data Center Capacity in Developing Regions

Metric	Value (2024)
Global Internet Users	5.5 billion (68% of population)
Global Offline Population	2.6 billion (32% of population)
Internet Penetration in High-Income Countries	93%
Internet Penetration in Low-Income Countries	27%
Internet Penetration in Africa	38%
School-age children unconnected at home (Global)	1.3 billion (67%)
School-age children unconnected at home (West & Central Africa)	95% (194 million)
School-age children unconnected at home (East & Southern Africa)	88% (191 million)
School-age children unconnected at home (South Asia)	88% (449 million)
School-age children unconnected at home (Middle East & North Africa)	75% (89 million)
LAC's share of global data center infrastructure	4.8%
Brazil's share of LAC data centers	37.2%
Asia-Pacific's share of global hyperscale data capacity	26%

This table provides a quantitative overview of the foundational digital infrastructure challenges facing developing regions. (Data sources: 8, 9, 12, 16) The stark contrast in

internet penetration rates between high-income and low-income countries, along with the significant number of school-age children lacking home internet access, vividly illustrates the profound "digital canyon." This lack of basic connectivity profoundly affects access to AI education and digital literacy development, creating an intergenerational barrier to AI adoption. Furthermore, the low share of global data center capacity in regions like LAC and Asia-Pacific underscores the severe limitations in local computing power, reinforcing technological dependency and raising concerns about data sovereignty. This data is crucial for understanding why other aspects of AI adoption, such as talent development and investment attraction, are so profoundly affected in these regions.

8.4 Investment Barriers: The Capital Chasm

The ambitious pursuit of Generative AI adoption in developing countries is severely constrained by significant investment barriers, creating a substantial "capital chasm" that separates them from leading AI nations.

Insufficient Public and Private Funding

A significant challenge for developing countries is the stark disparity in AI investment. The United States maintains a commanding lead, with private AI investment reaching \$109.1 billion in 2024. This figure is nearly 12 times higher than China's (\$9.3 billion) and 24 times that of the UK (\$4.5 billion).²⁴ While some emerging economies show growth, the overall scale remains limited. India, for example, attracted \$1.16 billion in private AI investments in 2024, with a cumulative total of \$11.29 billion from 2013 to 2024.²⁷ Other notable developing countries like the UAE (\$3.7 billion) and Israel (\$15 billion) have also drawn significant AI investments over the past decade, yet these amounts are still dwarfed by the investments in leading developed nations.²⁷

Globally, venture capital (VC) funding for AI companies surged in 2024, exceeding \$100 billion—an 80% increase from \$55.6 billion in 2023—with AI attracting nearly 33% of all global venture funding.²⁸ However, this investment remains heavily concentrated in high-income countries, leaving low-income nations largely excluded.¹⁰ For emerging markets and low-income countries, foundational investments in digital infrastructure, education, and data access are deemed essential. Public investment plays a particularly

critical role in areas with high social returns, such as healthcare, education, and public administration, where private markets are less likely to invest sufficiently.²⁹ In Africa, tech VC funding in 2024 saw a slight 2% decrease in total equity funding compared to 2023, reaching US\$2.2 billion. While this indicates a stabilization after a steep drop in 2023, it was largely influenced by a resurgence of a few large "megadeals" in the second half of 2024, masking underlying challenges in broader funding activity.³⁰

This situation highlights an "investment paradox" for developing countries. While these nations are recognized for their immense potential for AI to drive economic growth and solve critical societal problems³, the data clearly shows a massive disparity in private AI investment, with the overwhelming majority concentrated in the US and China. This means that despite the significant potential for high social and economic returns in developing economies—such as addressing unmet needs or tapping into less saturated markets—direct private capital inflow for AI is comparatively low. This suggests that private investors may perceive higher risks, lower immediate commercial returns, or insufficient foundational elements like robust infrastructure and skilled talent in these markets. The necessity of public investment in "high social return areas"²⁹ further indicates a market failure where private capital is hesitant to enter, necessitating government or international body intervention. This funding gap severely limits the ability of developing countries to build necessary AI infrastructure, foster local innovation, and scale AI solutions, perpetuating their reliance on foreign AI technologies and hindering their path to economic competitiveness and digital self-determination.



This surge in demand highlights a stark disparity in AI investment, with the United States attracting \$109.1 billion in private AI investment in 2024, significantly overshadowing developing countries like India (\$1.16 billion). Despite a global surge in venture capital funding for AI, it remains heavily concentrated in high-income nations, leaving low-income countries struggling to make foundational investments in digital infrastructure and education. This uneven funding landscape, coupled with the environmental burden, poses a significant hurdle for equitable AI development and access globally.

8.5 High Costs of GenAI Development and Deployment

The costs associated with training cutting-edge AI models are astronomically high, with estimates ranging from \$78 million for OpenAI's GPT-4 to \$191 million for Google's Gemini Ultra.¹¹ These prohibitive figures effectively exclude most developing countries, and even many developed ones, from participating in frontier AI research and development, contributing to an "AI oligarchy" dominated by a few global players.¹¹

Beyond initial training, the operational costs of GenAI can be unpredictable and drastically increase with usage. This is largely due to the linear cost scaling employed by most GenAI providers, meaning the cost per unit (e.g., per token) remains constant as usage volume climbs.³¹ This linear scaling makes it challenging for infrastructure providers to offer volume-based pricing, leading to unpredictable expenses for businesses integrating GenAI.³¹ Software companies integrating GenAI face a difficult choice: either pass these new, potentially volatile AI costs onto their customers or absorb them, with the latter often being unfeasible given rising usage.³¹ While the cost of *querying* (inference) certain AI models has seen dramatic reductions (e.g., a 280-fold reduction for GPT-3.5 equivalent from \$20 to \$0.07 per million tokens by October 2024 for Gemini-1.5-Flash-8B)¹⁹, this affordability primarily applies to *using* existing models, not to their initial development or large-scale customization.

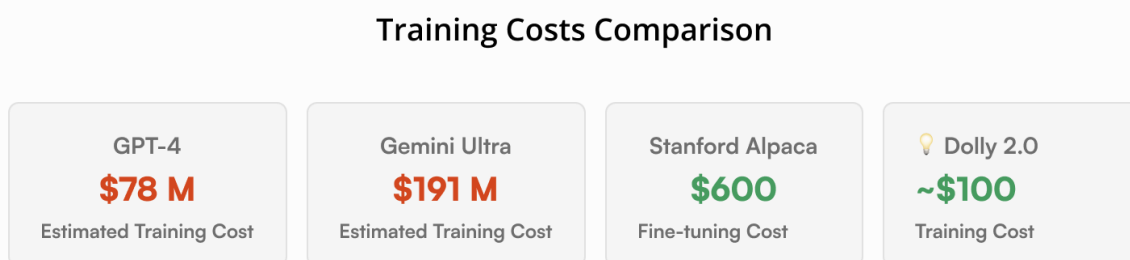


Image 9: AI's training cost comparison of different models

This situation presents a crucial distinction between "cost-efficiency" at inference and "cost-prohibitive" at training. While the application of existing AI models (inference) is becoming more accessible, the ability to develop, customize, and innovate foundational

GenAI models (training) remains prohibitively expensive. For developing countries, this means that while they might be able to afford to use pre-trained models developed elsewhere, their capacity to build their own sovereign, locally relevant, and culturally nuanced GenAI models is severely restricted. This reinforces technological dependency and limits their ability to address unique local challenges with bespoke AI solutions. Furthermore, even for inference, the unpredictable and linear scaling of usage costs³¹ poses a significant financial risk for businesses in developing countries with tighter budgets and less financial flexibility. This cost dynamic ensures that the "AI oligarchy"¹¹ of a few wealthy nations and large corporations will continue to dominate frontier AI innovation, hindering true digital self-determination for developing countries and impacting their long-term economic competitiveness and strategic autonomy.

8.6 Reliance on External Capital and Market Dynamics

While attracting foreign investment in AI is crucial for driving growth in developing countries, it frequently creates a tension with the imperative to protect local industries and national interests.¹⁸ The engagement of Western AI companies in the Global South has been characterized by some as ushering in a new era of "digital colonialism," marked by practices that are perceived as exploitative and undermining local agency and control.¹⁸ Concerns exist regarding economic "rent-seeking" by AI-dominant countries, where they might extract disproportionate value from developing nations. Furthermore, there are fears of the deliberate exclusion of AI innovations from low-income countries from major Western and Chinese markets, limiting their global participation.⁵

This reliance on external capital for AI development is not merely an economic issue but one with significant geopolitical dimensions. The research introduces concepts like "digital colonialism"¹⁸ and "global asymmetries in power"¹⁸, where dominant firms (often US-based) control critical aspects such as data sovereignty and intellectual property. This frames foreign investment not just as an economic transaction but as a strategic tool in a competitive geopolitical landscape where AI is increasingly viewed as a "battlefield" for global economic and security leadership.³³ This can lead to conflicts with a developing country's need to protect local industries, ensure data sovereignty,

and align AI development with its own strategic interests. The risk of "economic rent-seeking" and market exclusion ⁵ further highlights how foreign investment, if not carefully managed, can perpetuate rather than alleviate dependency. Consequently, developing countries must navigate a complex geopolitical minefield to secure necessary AI investment without compromising their national interests, fostering long-term technological dependency, or ceding control over their digital future.

Global Private AI Investment by Country (2013-2024 Cumulative & 2024 Annual)

Rank	Country	Total Investment (2013-2024 Cumulative, in USD Billions)	Private AI Investment (2024 Annual, in USD Billions)
1	United States	470.9	109.1
2	China	119.3	9.3
3	United Kingdom	28.2	4.5
4	Canada	15.3	N/A
5	Israel	15.0	N/A
6	Germany	11.3	N/A
7	India	11.1	1.16
8	France	9.0	N/A
9	South Korea	7.3	N/A
10	Singapore	7.3	N/A

Note: N/A indicates data not explicitly provided for the specific year/metric in the source snippets.

This table quantifies the significant financial disparity in AI investment, directly illustrating the "capital chasm" faced by developing nations. The overwhelming dominance of the United States and China in cumulative and annual private AI

investment is evident, underscoring the challenge for other nations to compete. The substantial growth in global Generative AI private investment, while positive for the sector overall, is likely concentrated in these leading economies, further highlighting the exclusion of developing countries from this high-growth segment. The African Tech VC funding figures provide a regional perspective, showing a modest scale of investment despite signs of stabilization. This data is crucial for demonstrating the magnitude of the financial challenge and reinforcing the arguments about technological dependency and the emerging "AI oligarchy."

8.7 Impact of International Regulations and Export Controls

International geopolitical dynamics significantly influence AI adoption in developing countries, particularly through export controls on critical hardware. U.S. export controls on advanced computing chips (GPUs), initially designed to impede China's AI progress, may inadvertently be accelerating it by forcing innovation in efficiency within China.⁵⁸ However, these controls have broader implications for developing nations. The new U.S. rule, effective January 2025, restricts sales of high-powered GPUs to over 100 countries, with exemptions only for 18 key allies (Tier I nations). Tier II countries, which include most of Africa, Latin America, Asia, the Middle East, and some EU members, face strict limits on GPU imports (approximately 50,000-100,000 Nvidia H100s from 2025-2027), while Tier III countries (e.g., China, Russia) are banned altogether.⁶⁰

In the long term, the limited allocation of 50,000 advanced GPUs will be insufficient for private entities in Tier II countries to develop leading AI models independently, potentially forcing them to import advanced AI models from Tier I nations.⁶¹ These restrictions also significantly hinder the development of data centers by Tier I companies within Tier II countries, as they are limited to locating no more than 25% of their computing capacity in Tier II countries overall, and no more than 7% in any single one.⁶¹

The U.S. export controls on GPUs create a multi-tiered global market for AI hardware, effectively imposing a "geopolitical chokehold" on AI development in many developing

countries. Even if these nations possess the financial capital, their access to the most advanced chips is severely restricted, forcing them to either rely on less powerful hardware, embark on the long and costly process of developing their own indigenous capabilities, or import pre-trained models from dominant players. This reinforces technological dependency and undermines efforts towards "Sovereign AI".¹⁷ The restrictions on Tier I companies building data centers in Tier II countries further limits the development of local infrastructure, directly impacting local innovation ecosystems. This policy, while primarily aimed at geopolitical rivals, has significant collateral damage for developing countries, potentially widening the AI gap and forcing them into a position of perpetual technological followership. It also complicates international collaboration and investment, as companies must navigate complex and rapidly changing export regulations.

8.8 Regionalization and Cultural Relevance: The Contextual Imperative

For Generative AI to truly benefit developing countries, it must be relevant and adaptable to their unique local contexts. However, significant challenges arise from the global AI development paradigm, which often overlooks linguistic diversity and cultural nuances, creating a barrier to widespread and equitable adoption.

Language Diversity and Low-Resource Languages

Most major Large Language Models (LLMs) currently underperform for non-English—and especially low-resource—languages. These models are often not attuned to relevant cultural contexts and are frequently inaccessible in parts of the Global South.⁶² The internet, which serves as the primary training ground for these models, is overwhelmingly Anglophone; while only 20% of the world's population speaks English at home, nearly half of the training data for major AI models is in English.⁶⁴ This inherent bias in training data leads to poor model performance for less-used languages, which in turn discourages their use and further reduces investment and interest in developing AI solutions for them.⁶⁵

Consequently, less-used languages and scripts (e.g., Ahirani, Sharda script, various non-Latin scripts) face an ever-greater risk of digital marginalization, potentially being

"wiped off the planet" as GenAI content is overwhelmingly in English.⁶⁵ Furthermore, AI models can "hallucinate" or flatten linguistic richness, struggling with regional accents and variations, and producing grammatically correct but culturally tone-deaf content.⁶⁴ This situation suggests a form of "linguistic imperialism" within AI development. The overwhelming dominance of English in AI training data and the underperformance of LLMs in low-resource and non-English languages is not merely a technical challenge; it poses a significant cultural and societal threat. If AI tools primarily cater to dominant languages, they risk accelerating the decline of minority languages and eroding cultural diversity.⁶⁵ This directly impacts education, communication, and the ability of diverse communities to engage with and benefit from AI in their own cultural contexts. The lack of culturally nuanced output⁵⁷ further alienates non-English speaking populations from AI's potential benefits. This challenge undermines the inclusive potential of AI in developing countries, risking a situation where AI's advantages are primarily accessible to English-speaking or culturally Westernized elites, thereby entrenching internal inequalities and hindering the development of AI solutions tailored to local needs and values.

Cultural Bias in AI Models and Datasets

Beyond linguistic limitations, AI models trained in one cultural context may not perform optimally in another due to fundamental differences in language, social norms, and regulatory environments.⁵⁷ A notable example is DALL-E 3, an image generation model, which, when asked to generate a picture of "breakfast," produced images of pancakes, bacon, and eggs, neglecting the diverse breakfast customs prevalent across the globe.⁵⁷ This "North American bias" ingrained in many LLMs can lead to discrimination against people from diverse cultures.⁵⁷ To mitigate this, adapting AI tools and guidelines to reflect the local context and the lived experiences of data annotators is crucial for ensuring the cultural relevance and accuracy of AI outputs.⁶⁶ However, access to representative datasets that accurately reflect African contexts and realities, for instance, remains paramount yet limited.²³

This issue points to a "contextual irrelevance" barrier to AI adoption. AI models exhibit cultural biases because they are predominantly trained on unrepresentative datasets, largely sourced from Western contexts.¹⁰ This leads to AI outputs that are not only inaccurate but also culturally insensitive or inappropriate for local use cases. For

developing countries, this means that even if they overcome infrastructure and cost barriers, the globally available AI models may not be fit-for-purpose for their unique societal, economic, and cultural challenges. This necessitates significant effort in local data collection and model adaptation, resources which developing countries often lack.²³ Failure to address this limits the practical utility and trustworthiness of AI in developing countries. If AI solutions are not culturally relevant, they will face low adoption rates, fail to address local problems effectively, and potentially reinforce existing stereotypes or discrimination. This also means that AI's transformative potential for sectors like healthcare and agriculture, which are highly context-dependent, may not be fully realized.

Need for Context-Specific Solutions

Despite these challenges, AI offers novel solutions to longstanding challenges in healthcare, agriculture, education, and infrastructure development within emerging economies.¹ Developing countries possess a unique opportunity to bypass traditional development stages through AI adoption, "leapfrogging" directly to advanced digital solutions tailored to their specific contexts.⁷ This requires not only adapting AI tools to suit the rich cultural and linguistic diversity prevalent in developing countries but also seamlessly integrating AI applications into their existing systems.¹³

This highlights a "local innovation imperative" that stands in contrast to a "global template trap." The data indicates that AI's true potential in developing countries lies in its ability to provide "context-specific solutions" and enable "leapfrogging".¹ However, the challenges of language diversity and cultural bias⁵⁷ push developing countries towards a "global template trap," where they are compelled to adopt AI models and solutions primarily designed for Western contexts. This creates a tension between the *potential* for localized innovation and the *reality* of relying on globally available, often culturally misaligned, AI. The success of national initiatives, such as the IndiaAI Mission⁶⁸ and AI for India 2030³, hinges on their ability to foster indigenous development that directly addresses unique local needs, rather than simply importing foreign solutions. Failure to foster local innovation and adapt AI to specific contexts will limit the transformative impact of AI in developing countries, leading to suboptimal outcomes and missed opportunities for sustainable development, and risks creating a new form of technological dependence.

9. *Lessons from Previous Technological Revolutions*

Artificial intelligence is consolidating at an extraordinary pace inside a small circle of rich states and hyperscale companies. If present trajectories hold, the next twenty years will see a decisive concentration of computing power, algorithmic standards and data capital in the United States, China, the European Union and a few close partners. For the rest of the world the danger is not abstract: missing the AI wave could replicate—and in several respects deepen—the structural dependence that followed earlier technological revolutions.

What history tells us

Throughout the last two centuries transformative technologies have repeatedly redrawn the map of global power. When Britain mechanised cotton spinning and laid its first railways, late-industrialising regions were forced into roles as raw-material suppliers and captive markets. During the age of electrification, countries able to wire entire national grids multiplied their productivity, while those without the capital watched finished goods pour in from abroad. Control of oil after 1945 reorganised foreign policy, enabling producers to dictate prices and conditions to import-dependent states. And over the past quarter-century, network effects on the internet have allowed a handful of U.S. platforms to become global utilities, reducing many local tech sectors to resellers.



AI is rapidly concentrating within a few rich states and hyperscale companies, risking a deepened global technological divide. This trend mirrors historical technological revolutions where first-movers established dominance, leaving latecomers structurally dependent.

The logic is always the same: high entry costs, tight intellectual-property protection and reinforcing network effects let first movers define the rules for everyone else. AI magnifies each of those ingredients. Frontier models are trained on tens of thousands

of cutting-edge GPUs, cost hundreds of millions of dollars, and improve through every new user prompt—meaning even passive participation by latecomers feeds the incumbents’ advantage.

How dependence will form

Export-controlled compute: Nine out of ten of the most powerful accelerators already ship from fabs under U.S., Japanese or Dutch jurisdiction. Licensing regimes that once applied to nuclear centrifuges are now being extended to advanced chips. Artificial scarcity locks public laboratories in Africa, South-East Asia and Latin America onto older silicon; their work is condemned to trail state-of-the-art by several model generations.

Algorithmic rent extraction: Where today businesses import petroleum, tomorrow they may import cognition: language translation, contract drafting, medical triage, logistics planning—all metered by the token and priced in hard currency. By the mid-2030s, a mid-income country of a hundred million people could spend more on AI services than it now pays for refined fuel, draining reserves and crowding out domestic investment.

Standards capture: Safety evaluations, watermarking methods and audit pipelines are already being written inside the largest AI clusters. Regulators in smaller economies will have little practical choice but to adopt those tool-chains sight-unseen or find themselves cut off from global supply chains. Nuances of local law, minority languages and indigenous knowledge simply will not appear in the reference suites that decide whether a system is “responsible” or “safe”.

Brain-drain flywheel: Where compute is scarce, research talent migrates—physically or through remote contracts—to the cloud regions that hold the chips. Universities at home become feeder programmes; public-sector AI projects wither; income gaps widen between a tiny cohort of offshore contractors and the rest of the labour market, sowing social tension.

Data enclosure: Vast troves of agricultural, health and climate data from the Global South are already flowing into open-science repositories funded by Northern foundations. Once refined into proprietary models, the insights are resold to the

original data owners on a subscription basis, echoing the way colonial botanists exported seed genomes that later returned as patented hybrids.

Energy bottlenecks: Training and serving ever-larger models could require an additional five hundred gigawatts of steady electricity worldwide by 2035. States without abundant renewables or nuclear baseload will confront a bitter choice: ration domestic power or outsource their AI ambitions to foreign clouds, often on terms bundled with LNG plants or debt-financed grid upgrades.

Kill-switch diplomacy: As customs screening, tax collection and health triage move onto proprietary AI-as-a-service platforms, a temporary suspension—intentional or accidental—can paralyse an entire government. The mere possibility of outage becomes leverage in bilateral negotiations, the digital equivalent of closing a canal or turning off a pipeline.

Cultural homogenisation: Generative systems trained mainly on English, Mandarin and major European languages will dominate search, entertainment and education. Minority languages risk falling below the digital threshold of usability; local creative industries are sidelined by algorithms that recommend content with global but not necessarily local resonance.



Dependence will form as developing nations face restricted access to advanced AI compute and talent, forcing reliance on foreign-controlled AI services and standards. This creates vulnerabilities in critical functions, drains resources, and risks cultural homogenization.

Likely macro-impacts by 2045

If nothing interrupts these forces, import bills for AI services could rival today's energy imports, current-account deficits will widen, and domestic ICT value-added could shrink to a fraction of its present share of GDP. Only a small percentage of the most-cited AI research is likely to originate outside the core countries, leaving peripheral economies dependent on external intellectual property just as they once depended on imported turbines or chemical patents.

Socio-political consequences

Fiscal stress will tempt governments back to the IMF, now negotiating not only fiscal targets but also clauses on digital governance. Policy autonomy will narrow as essential functions run on foreign clouds. Authoritarian leaders may purchase turnkey predictive-policing suites, entrenching themselves with tools designed and updated abroad. The traditional ladder of export-led industrialisation could break: as rich markets automate textile sewing and electronics assembly, the jobs that once lifted East Asia out of poverty will disappear before wages in Africa or South Asia converge. And as languages vanish from the algorithmic mainstream, cultural self-determination—and the pluralism that underpins democratic resilience—will erode.

Why catch-up will be harder than before

Speed is one reason: AI capability doubles every few months, not every few decades. Intangibility is another: weight files can be embargoed without the public drama of warships or missile tests, reducing political urgency until dependence is entrenched. Finally, every prompt a latecomer submits to a frontier model improves that model; by trying to close the gap, followers deepen it.



If current AI trends persist, developing nations by 2045 could face substantial AI service import bills, reduced domestic ICT value, and increased intellectual property dependence. Socio-political consequences include fiscal stress, diminished policy autonomy, and the loss of traditional industrialization pathways. Catching up will be exceptionally difficult due to AI's rapid advancements, the subtle nature of technology embargoes, and the fact that latecomers' usage inadvertently enhances dominant models, further widening the gap. This scenario portends a future where economic and societal progress for many hinges on external AI capabilities.

10. Training a GenAI model: different strategies & Efficient alternatives

Pretraining a Large Language Model (LLM) traditionally involves training a transformer-based model on massive unlabeled text corpora. This process uses diverse datasets (e.g. web crawl data, books, code, Wikipedia) amounting to hundreds of billions of tokens. For example, OpenAI's GPT-3 (175 billion parameters) was trained on about 300 billion tokens and required an estimated 1.3 million kWh of electricity (roughly equivalent to \$4–\$5 million USD in compute cost). Meta's LLaMA-2 (70B) needed 1.72 million GPU-hours on 2048 A100 GPUs, consuming ~0.688 GWh of energy. These figures translate to *millions of dollars* in electricity and hardware—LLaMA-2's training run emitted ~291 metric tons of CO₂. Newer models are even costlier: LLaMA-3.1 (405B) reportedly required 39.3 million GPU-hours on H100 GPUs (27.5 GWh of electricity), although improved hardware efficiency kept its carbon footprint similar to LLaMA-2's at ~240 tons CO₂.

Such resource requirements demand specialized infrastructure. **Compute Clusters:** Pretraining often runs on large clusters of accelerators (GPUs or TPUs). For instance, LLaMA-3.1 training used 32,000 NVIDIA H100 GPUs in parallel. These clusters must also be fed by high-throughput storage for the terabyte-scale datasets and supported by robust networking. **Memory and Storage:** A 175B-parameter model in FP16 requires ~350GB just to store weights, so distributed memory across many devices is essential. Datasets on disk can exceed hundreds of terabytes once processed. Checkpoints (snapshots of model weights during training) can be hundreds of GB each, incurring storage and I/O costs.

Training Time and Energy: Even with massive parallelism, pretraining can take many days or weeks. This results in significant energy consumption and carbon emissions. A study estimated training GPT-3 consumed ~1.3 GWh of electricity, comparable to powering hundreds of homes for a year. Such power usage raises environmental and cost concerns, especially for organizations without subsidized compute. Moreover, the financial burden means only a few tech giants or well-funded institutions have traditionally undertaken training models from scratch.

Challenges and Redundancies: A notable drawback is that many LLM projects repeat similar expensive steps. The same Common Crawl web texts and Wikipedia articles are used to pretrain multiple models, leading to *redundant compute spend* across the field. Each new model often “relearns” basic grammar and facts from scratch. This redundancy has prompted calls for more reusable *foundation models* and shared checkpoints to avoid duplicating efforts. Another challenge is *data overlap and quality*: Large corpora contain redundant or low-quality data that can slow training or affect performance. Techniques like data deduplication can improve efficiency, but are not yet. In sum, standard pretraining is extremely costly in terms of data, compute, time, and energy. These hurdles motivate the search for more efficient approaches to LLM development, as we explore in this whitepaper.

Overview of training methodologies

This section surveys the spectrum of training methodologies for LLMs and multi-modal LLMs, from how the base models are pretrained to various fine-tuning and adaptation techniques that minimize full retraining needs. We cover traditional pretraining, continuous pretraining, alignment via human feedback, efficient fine-tuning methods, and novel paradigms like model merging and swarm learning.

10.1 Pretraining Techniques

Standard Pretraining: This refers to the one-time training of a model from random initialization on a broad corpus. As described, it is compute-intensive and typically done only for foundational models. The entire model’s weights are learned by optimizing a self-supervised objective (predicting masked tokens or next tokens) over massive text data. Once this phase is completed, the model has general language proficiency but lacks task-specific alignment (it may produce raw, unaligned outputs). The cost and difficulty of this step incentivize doing it only once per model size; after that, one can reuse the pretrained checkpoint for multiple purposes.

Continuous Pretraining (CPretraining): Continuous pretraining means further training an already pretrained model on new data to refresh or specialize its knowledge. Instead of starting from scratch, we initialize with an open or existing model and *continue* the language modeling training on additional corpora relevant to

the target domain or timeframe. This approach updates the model's general knowledge without the full cost of training from random weights. For example, an enterprise might take an open 7B model and continuously pretrain it on internal documents or recent news to keep it up-to-date. Continuous pretraining can be much cheaper than initial pretraining. Gili Nachum estimates that continuing to pretrain a 7B model on ~1 billion tokens (e.g. 5,000 domain PDFs) would take on the order of 57 hours on 8×A100 GPUs (costing around \$2.3K on AWS)— a feasible expense for many companies. Key considerations in continuous pretraining are avoiding *catastrophic forgetting* (losing the original knowledge) and managing domain shift. Usually, a small learning rate and sometimes intermixing some of the original pretraining data can help the model retain general ability while learning new content.



Standard pretraining involves training a foundational model from scratch on a vast dataset, a compute-intensive process resulting in general language proficiency. In contrast, continuous pretraining further trains an already established model on new, specialized data to refresh or update its knowledge, offering a more cost-effective way to adapt models for specific domains or recent information.

10.2 Post-Pretraining Alignment and Fine-Tuning Techniques

Once an LLM is pretrained (either from scratch or via continuous updates), it often undergoes post-training alignment to make it more useful and safe for end-users. Several methodologies exist:

1. **Supervised Fine-Tuning (SFT):** This involves further training the model on task-specific or instruction-following data via supervised learning. For instance, ChatGPT's precursor (InstructGPT) was fine-tuned on prompt-response pairs written by humans to teach the model to follow user instructions. SFT is relatively straightforward: given example inputs and desired outputs, adjust the

model's weights to better produce the desired outputs. This phase typically uses a much smaller dataset (tens of thousands of examples) and is far less costly than full pretraining – often a few GPU-hours to days. However, full-model fine-tuning can still be memory-intensive; it updates all parameters, which for a 30B+ model requires multi-GPU setups or memory optimization (since gradients and optimizer states for all weights must be stored).

2. **Reinforcement Learning with Human Feedback (RLHF):** RLHF has been crucial to aligning LLMs with human preferences. In RLHF, the model is optimized not just to predict text, but to produce *helpful, harmless* responses as judged by humans. The typical RLHF pipeline involves three steps;

- **Step 1, SFT** – start with a supervised fine-tuned policy model (π_{SFT}) that can follow basic instructions;
- **Step 2, Reward Modeling** – train a separate reward model to score outputs by quality, using a dataset of human preference comparisons;
- **Step 3, RL fine-tuning** – further train the policy (now π_{RL}) using an algorithm like Proximal Policy Optimization (PPO) to maximize the reward model's score while constraining the policy not to drift too far from the SFT model.

While effective, RLHF is complex and resource-intensive: it requires human or high-quality proxy feedback and multiple training stages. Moreover, PPO-based RLHF introduces a “critic” (value) network which doubles memory usage and adds instability.

3. **Advanced RLHF Variants (RLOO, GRPO, DPO):** Recently, researchers have proposed lighter-weight alternatives to PPO for RLHF. *Reinforce Leave-One-Out (RLOO)* and *Group Regularized Policy Optimization (GRPO)* are two such algorithms that eliminate the need for a separate value (critic) network, reducing complexity. Instead, they estimate the advantage of a response by comparing it to other sampled responses for the same prompt (for RLOO) or normalizing within a batch (for GRPO). This removal of the value model cuts memory usage

in half and simplifies training. However, RLOO and GRPO must carefully estimate advantages to remain stable. These methods can still suffer from high variance or *reward hacking* (over-optimizing to the reward model) if not properly regularized. Another notable approach, *Direct Preference Optimization* (DPO), foregoes the RL step entirely by directly fine-tuning on the comparison data via a calibrated objective, achieving results comparable to PPO-based RLHF with less complexity. The emergence of RLOO, GRPO, DPO, and others indicates active research to reduce the burden of alignment training while preserving the benefits of RLHF.

10.3 Parameter-Efficient Fine-Tuning (PEFT)

Instead of fine-tuning *all* tens of billions of weights of an LLM for each new task (which is memory-intensive and risks overfitting or forgetting), parameter-efficient fine-tuning (PEFT) methods update only a small subset of parameters or introduce additional small modules. This drastically lowers GPU memory requirements and allows reusing a single base model for many purposes via different *adapter* modules.

1. **Adapters:** Originally developed for Transformer models in NLP, adapter layers are small bottleneck networks inserted at various points (e.g. after the feed-forward or attention sublayers) and trained for the downstream task while the original model weights remain frozen. Adapters act as task-specific “patches.” Because only the adapter weights (often <5% of total parameters) are trained, the approach is efficient in data and compute, and it preserves the original model’s knowledge, avoiding catastrophic forgetting. An added benefit is modularity: one can keep a library of adapter modules for different domains or languages and hot-swap them on the same base model. This approach was used to great effect in multi-domain models and allows *regionalization* – e.g. adding a country- or language-specific adapter to a global model.
2. **LoRA (Low-Rank Adaptation):** LoRA is a popular PEFT technique that injects trainable low-rank matrices into each layer’s weight update. In practice, LoRA adds a pair of small rank-decomposed matrices (A and B) that adjust the output of the transformer layer. Only A and B are learned (a few million parameters

even for a 7B model), while the original weight stays fixed. LoRA adds *no* inference latency because at runtime the low-rank updates can be merged into the main weight matrix. This approach has been *highly* successful in fine-tuning LLMs with minimal compute. It has been reported that LoRA fine-tuning can reduce training overhead by up to 70% compared to full-model tuning. LoRA's efficiency and simplicity (no model architecture change aside from additional weights) made it a default for many LLM fine-tuning projects in 2023.

3. **DoRA (Weight-Decomposed LoRA):** Despite LoRA's success, a gap often remains between LoRA-tuned models and fully fine-tuned models in accuracy. In 2024, NVIDIA researchers introduced DoRA, which aims to close this gap. DoRA stands for *Weight-Decomposed Low-Rank Adaptation*. It factorizes each weight matrix in the pretrained model into two components: a *magnitude* vector and a *direction* (unit vector) for each weight, then applies low-rank adaptation on the high-dimension directional component. By doing so, DoRA can adjust not just a small additive delta, but also modulate the weight magnitudes, improving capacity. Importantly, DoRA still avoids any increase in inference compute – after fine-tuning, the adapted weights can be merged back. Empirically, DoRA has outperformed LoRA on benchmarks (e.g. +3 to 4 points higher accuracy on reasoning tasks with LLaMA models). It improves training stability as well. In short, DoRA is a drop-in enhancement over LoRA that yields closer-to full fine-tuning performance with the same efficiency benefits.
4. **QLoRA:** A major recent breakthrough in PEFT is QLoRA (Quantized LoRA). QLoRA tackles the *memory bottleneck* by quantizing the *base model* to 4-bit precision during fine-tuning, while still learning LoRA adapters in 16-bit. By backpropagating through a 4-bit model, QLoRA drastically cuts RAM usage—enough to fine-tune a 65B model on a single 48GB GPU. This democratized experimenting with very large models. Notably, QLoRA preserved full 16-bit fine-tuning performance; the team demonstrated a 33B and 65B LLaMA tuned via QLoRA (nicknamed “Guanaco”) that achieved 99.3% of ChatGPT's performance on a benchmark, after just 24 hours of training on one

machine. QLoRA introduced technical innovations like a new quantization data type (NF4) for minimal accuracy loss and double-quantization to reduce memory further. The success of QLoRA means even resource-constrained teams can *iterate on large models* by fine-tuning, without needing a mega-cluster.

5. **Other PEFT Variants:** Many other techniques fall under PEFT: Prefix Tuning (prepends learnable tokens to prompts), Prompt Tuning (optimizes an embedded prompt), and Adapter fusion. There are also research efforts to combine ideas: e.g. DeFT (Data-Efficient Fine-Tuning) which selects a *core subset of data* to fine-tune on, thereby reducing the data needed by up to ~70% while maintaining performance. Another direction is reducing *activation footprint* during fine-tuning (to save memory) – for example by pruning unnecessary neurons (some work has humorously noted many transformer layers are partly redundant). The key takeaway is that PEFT methods greatly minimize the burden of adapting LLMs. Instead of retraining a model from scratch or fine-tuning billions of weights for each new task, one can train a few million (or even thousand) parameters to get excellent results. This makes LLM fine-tuning viable for enterprises and researchers without access to huge compute clusters.
6. **Model Merging and Swarm Models:** Beyond fine-tuning a single model, a creative approach is to *merge or ensemble* multiple models to combine their strengths. Model merging usually means taking two or more models (with the same architecture) and literally combining their weights (e.g. by weighted averaging) to create a new model. This can extend a model’s capabilities without training on all tasks simultaneously. For instance, one could merge a model fine-tuned for legal QA with another fine-tuned for medical QA, yielding a single model with both skills. Merging can be as simple as linear interpolation of weights, or more advanced like SVD-based merges or nonlinear interpolation (e.g. SLERP in weight space). Recent “model soups” and “Franken-models” attest to merging’s potential: some top entries on open LLM leaderboards are *merges* of other fine-tuned models. There are even *Frankenstein MoEs*, which merge models by creating a Mixture-of-Experts with different expert weights frozen.

Meanwhile, swarm models or multi-agent ensembles involve orchestrating several LLMs at runtime. In a swarm, each model (agent) might specialize (one could be good at coding, another at language translation), and a router or voting system decides which model(s) handle a given query. This concept is analogous to an expert committee – it can outperform any single model if done well. Early research on *foundation model swarms* suggests that carefully optimized cooperation among models (through learned graphs or controllers) can yield robustness and better overall performance. For resource-constrained users, swarm approaches are promising because multiple smaller models (which are easier to train or fine-tune individually) can collectively cover ground that a huge monolithic model would, spreading out the compute requirements.



Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Adapters, LoRA, DoRA, and QLoRA, significantly reduce the computational burden of adapting large language models (LLMs) by training only a small subset of parameters or introducing minimal new modules. This allows for efficient task-specific customization without retraining the entire model, making LLM fine-tuning more accessible. Complementing this, model merging and swarm models offer creative ways to combine the strengths of multiple LLMs, either by merging their weights or orchestrating them as a multi-agent ensemble, effectively extending capabilities and distributing compute requirements.

11. Alternatives to Expensive Pretraining

Given the immense expense of pretraining from scratch, researchers and practitioners have developed alternative strategies to bootstrap powerful models without incurring prohibitive costs. These techniques leverage existing open models, modular adaptations, and creative combinations to achieve strong results for specific domains, languages, or tasks.

11.1 Leveraging Open Models as Base Initializations

Perhaps the most straightforward shortcut is: *don't reinvent the wheel*. If a reasonably good model already exists, use it as the starting point. The proliferation of open-source LLMs (GPT-J, GPT-NeoX, Bloom, LLaMA, Falcon, etc.) offers many foundation models that can be downloaded and reused. Organizations with limited resources can take a model like LLaMA-2 (released by Meta) and fine-tune or extend it, rather than training a new model from scratch. This practice has enabled dozens of new derivatives at a fraction of the original training cost. For example, Stanford's Alpaca model took Meta's 7B LLaMA and fine-tuned it on 52,000 instruction-response examples generated by a larger model (text-davinci-003). The total expense was under \$600 (OpenAI API fees <\$500 for data generation + a few hundred for fine-tuning compute). Alpaca's performance turned out to be remarkably close to that of the much larger text-davinci-003 on the evaluated tasks – all achieved by repurposing an existing pretrained model and *minimal new training*. This “open model + small fine-tune” recipe has been replicated widely (e.g. Vicuna, Dolly, and many other ChatGPT-like models built on LLaMA or OPT). It demonstrates that enterprises or countries can get a high-quality model by building on an open foundation, instead of paying the full cost themselves.

Reusing open models also promotes *transparency and sovereignty*. For example, if a country is concerned about reliance on API access to a foreign proprietary model, they might take an open model and adapt it to their language. The UAE's Technology Innovation Institute followed this path to create Falcon, an Arabic and English LLM, by gathering regional data and training on top of existing architectures (they released Falcon openly so others could further leverage it). Similarly, Bloom (176B), developed via a global collaboration, was explicitly intended as a public initialization that any language community could fine-tune for their needs, rather than everyone collecting and training on the same Common Crawl again.

11.2 Regionalization via Adapters

A powerful technique for local adaptation is to use adapters or LoRA modules targeted to region-specific data. Instead of one monolithic model trying to cover all languages

and dialects (which requires enormous data), one can train small adapter modules on local languages or domains and plug them into a global model. Because adapters don't overwrite the original weights, the base model's multilingual knowledge remains intact, and the adapter augments it with regional specifics. For instance, one could have a core model pretrained on multiple languages, and then add a *Swahili adapter*, *Hindi adapter*, etc., each trained on a relatively small corpus in that language. This approach was explored in the *MAD-X* framework for cross-lingual transfer with BERT: language adapters allowed adding new languages without catastrophic forgetting of others. In the LLM context, this means a country could take an existing large model and inject local cultural and linguistic knowledge via an adapter. It is much cheaper than full pretraining in that language. Indeed, there are initiatives in Africa (e.g. Masakhane's projects or Lelapa's Inkuba LM) that fine-tune adapters for African languages on top of existing models. Such efforts drastically lower the barrier for including under-represented languages in AI. Regional adapters can also encode cultural norms or domain jargon (finance, law) prevalent in that region, which would be too niche to appear often in a general pretraining corpus.

A concrete example is China's approach to LLMs: researchers have used open English models and *continued pretraining* them on large Chinese text corpora, then fine-tuning with adapters for alignment. This two-step process (base reuse + regional pretraining) proved effective for creating competitive Chinese MLLMs at a lower cost. We also see multinational companies like Orange partnering with model providers to fine-tune LLMs for African French, Arabic dialects, etc., rather than building new models from zero. The key point is that adapters enable modular regionalization: one base model can serve many locales, each with its adapter. This avoids training and maintaining separate full models per locale, saving enormous compute.

11.3 Model Selection, Routing, and Merging ("Advantages-Based" Strategies)

Instead of a single all-encompassing model, another strategy uses *multiple specialized models* and intelligently selects or combines them for a given task/query:

1. **Specialist Models:** An enterprise might curate a suite of smaller models, each an expert in something (one for software code, one for customer support dialogue, one for legal text). Then a light-weight router algorithm (possibly a classifier or a prompt-based switch) chooses which model to apply for each user query. This way, each model is simpler and trained on a focused dataset, making their training feasible on limited resources. The router ensures the query is handled by the best model (advantage-based selection). For instance, a query containing programming terms could be routed to a code model. This approach is reminiscent of Mixture-of-Experts, but can be done at a higher level without an integrated MoE architecture. It trades a bit of complexity (managing N models) for potentially large efficiency gains (N smaller trainings instead of one huge training).
2. **Ensembling and Voting:** In cases where quality is paramount, multiple models can each generate an answer, and then an ensemble method (like majority voting or a separate evaluator model) picks the final output. This “swarm” style can improve accuracy without any single model needing to be state-of-the-art – the collective compensates via diversity. While running several models is slower, one can constrain this to critical uses. Notably, OpenAI’s early work on GPT-4 hints that they used model ensembling techniques (though details are proprietary). In the open community, projects like Swarms and others are exploring multi-agent LLM systems.
3. **Model Merging:** As discussed in Section 2, merging fine-tuned models is another alternative to training one model on all data. For example, instead of one costly run on a combined dataset, train two smaller runs on parts, then merge. A real scenario: Company A fine-tunes a base model on legal contracts, Company B on medical texts. Rather than each company training a model on the union of legal+medical (which doubles data and cost), they could share their fine-tuned weights and merge them. Techniques like weight interpolation or task vector addition allow this combination. The result approximates what a joint training would have achieved, at a fraction of compute (each party did half). There are challenges – merged models can have inconsistencies or require

some additional tuning – but research is showing promising results where merges even *outperform* original models on certain benchmarks. This hints that merging isn't just cost-saving; it might create an ensemble-like effect within one set of weights.

11.4 Continuous Pretraining on Local Data

A special case of *alternatives to full pretraining* is when you *do need* to train on a lot of text, but you start from a strong base model rather than from scratch. We touched on this in continuous pretraining: for example, an institution with a large proprietary dataset (say millions of domain-specific documents) can continue pretraining a public model on this data. This effectively blends the general knowledge of the public model with the specific knowledge in the local data. Because the model already knows how to form sentences and has broad facts, the additional training is far more sample-efficient than starting fresh. Empirical evidence supports this: continued pretraining on domain data yields sizable gains on domain tasks (e.g. a medical LLM pretrained further on medical journals will do better in that domain's QA).

Importantly, continuous pretraining can be done gradually and periodically – a form of *streaming training*. An enterprise might schedule regular updates where new data (say, the past month's documents or locally relevant web content) is used to refresh the model. This keeps the AI up-to-date with current information, addressing the “knowledge cutoff” problem of static models. Several literature case studies have shown this approach: one case found that updating a model on news articles enabled it to answer questions about current events much better than the original. Another example is the open-source project RedPajama which sought to re-pretrain a Llama model on a fresh web crawl, indicating that community-driven continued training is viable.

One has to manage *forgetting* in continuous pretraining. Techniques such as AdapterSwap train new adapters on new data and occasionally swap or merge adapters to integrate new knowledge while keeping old knowledge safe. This can be seen as a hybrid of adapters and continuous training – a promising research direction to enable *lifelong learning* in LLMs without catastrophic forgetting. The overarching message is

that pretraining need not be a one-shot, all-or-nothing endeavor: by incrementally training on local data, even nations with moderate compute can build competitive models over time.

11.5 Examples of Successful Low-Cost Adaptations

To illustrate these alternatives, consider a few concrete examples:

1. **Stanford Alpaca (2023):** Already discussed, Alpaca distilled ChatGPT capabilities into a 7B model for ~\$600 by leveraging OpenAI’s API and an open base model. This sparked a wave of similar projects because it showed what *minimal data + existing model* can achieve.
2. **Databricks’ Dolly (2023):** Dolly was a 6B model fine-tuned on a small (~15K record) instruction dataset crowdsourced by Databricks. They started from EleutherAI’s GPT-J (an open 6B model) and produced a useful chatbot for essentially the cost of a few hours on 1 GPU. Dolly’s quality wasn’t state-of-the-art, but it was *good enough* for many internal applications and proved the concept of training “ChatGPT for less than \$100”.
3. **BloombergGPT (2023):** Bloomberg built a 50B model for finance by mixing an open public dataset with a large internal financial text dataset. While they did train from scratch (since no finance-focused base existed at the time), the project highlighted that domain-specific LLMs can be obtained by *augmenting a general corpus with domain data*. If BloombergGPT were done now, one could imagine taking an existing 50B model and only doing the latter half of training on the finance data – saving time.
4. **Lelapa’s Inkuba (2024):** A project out of South Africa (Lelapa AI) trained an LLM covering several African languages (Swahili, Zulu, Hausa, etc.). Rather than gather a mammoth corpus for each language, they likely leveraged multilingual transfer: e.g. use an existing multilingual model or translate data from English. This significantly lowers the barrier for African NLP. It showcases how *smart data augmentation and transfer* can replace brute-force data collection.

These case studies underscore that through *openness, clever data generation, and modular training*, high-quality models can be developed at a fraction of the traditional

cost – making LLM technology more accessible to enterprises and regions with limited resources.

12. Infrastructure and Optimization

Infrastructure choices and software optimizations play a pivotal role in reducing the cost of training and deploying LLMs. This section discusses how to make the most of limited or heterogeneous hardware – including CPUs, non-NVIDIA accelerators, and edge devices – and highlights techniques to maximize efficiency (from quantized models to distributed systems and optimizer improvements).

12.1 Hardware Constraints: Making Do with What You Have

Not everyone has a state-of-the-art GPU cluster. Many organizations must work with CPUs or consumer-grade GPUs. Fortunately, recent advances enable LLM use on modest hardware:

1. **CPU-Only Environments:** While CPUs are far slower than GPUs for training large models, they can handle smaller models or reduced-precision inference. Optimized libraries like Bud Runtime, Intel’s oneDNN and AI accelerators (like Intel’s DL Boost on Xeon) offer some speedup for matrix ops. A notable project is LLAMA.cpp, which brought LLaMA model inference to commodity laptops by optimizing the transformer operations for CPUs and using 4-bit quantization. It showed that a 7B model can generate text on a CPU in reasonable time (a few tokens per second). For CPU training, frameworks exist to distribute a model across many CPU cores or machines (Facebook’s *Fully Sharded Data Parallel* can use CPU memory to store model shards). It’s not efficient for full pretraining, but fine-tuning a smaller model on CPUs is sometimes feasible overnight. In any case, supporting CPU inference is important for broad deployment, since servers and devices without GPUs will need to run these models.

2. **Apple Silicon (M1/M2) and Mobile Chips:** Apple's M1/M2 chips have built-in neural engines and strong GPUs that punch above their weight in ML tasks. Apple has actively optimized CoreML and PyTorch (with an MPS backend) for these chips. In fact, Apple demonstrated running a 8B parameter LLaMA-3.1 model entirely on an M1 Max at ~33 tokens/second – a real-time speed. This was achieved by converting the model to Apple's neural engine format (16-bit weights with throughput-optimized kernels) and using the unified memory efficiently. Likewise, mobile phones are now seeing LLMs: e.g. an iPhone 14 Pro can run a 7B model like Mistral 7B at a slower rate (maybe 1-2 tokens/sec) thanks to 4-bit quantization and offloading to the Neural Engine. These developments mean personal devices can host an AI assistant without any cloud service, ensuring privacy and offline capability. For enterprises, leveraging employees' smartphones for some AI tasks (on-device) might reduce cloud inference costs.
3. **Intel GPUs (Arc, Data Center Max):** NVIDIA's dominance in AI compute has meant less software support for others, but that's changing. Intel's Arc series (for consumers) and their Data Center GPU Max (Ponte Vecchio architecture) are capable hardware that often goes underutilized. Projects like Intel's *BigDL-LLM* and *IPEX (Intel Extension for PyTorch)* have started enabling LLM fine-tuning and inference on Intel GPUs and CPUs. For example, Intel published guides on fine-tuning LLaMA-2 on their GPUs using BigDL, and demonstrated LoRA fine-tuning on a GPU Max cluster. While performance and software maturity are still catching up to CUDA, these alternatives can be cost-effective (Intel's data center GPUs might be cheaper or more available in some regions than A100/H100). Embracing heterogeneous hardware by using frameworks that abstract the device (like oneAPI or DirectML) can let an organization use whatever silicon they can procure – be it AMD, Intel, or older NVIDIA cards – thus mitigating supply or embargo risks.
4. **Edge Devices and IoT:** Beyond phones, think of Raspberry Pi or microcontrollers – running an LLM here sounds fanciful, but tiny models (perhaps 100M parameters distilled from a larger model) could run on such

devices for simple tasks (like voice commands in appliances). Techniques like *algorithmic distillation* (training a smaller student model on the outputs of a large teacher) are used to produce compact models that preserve some capabilities of LLMs. An example is the 30x smaller student models from Meta’s DistilBERT and others – not an “LLM” in parameter count, but often sufficiently fluent for narrow applications. This could be critical for *edge AI* where network connectivity is limited or latency must be ultra-low.



Hardware constraints can be overcome by optimization and smart scaling-down. The landscape of LLM deployment is expanding from cloud GPU clusters to Macs, phones, and beyond, enabling more inclusive AI adoption.

12.2 Efficient Model Architectures and Quantization

The architecture of a model directly impacts its compute and memory needs. Standard Transformers are heavy, but researchers are inventing more efficient variants:

1. **BitNet – 1-bit Transformers:** An extreme but promising idea is training models with binary or 1-bit weights. Microsoft’s BitNet is a recent architecture that does exactly this. In BitNet, the linear layers are replaced with *BitLinear* layers that constrain weights to 1-bit during training (with some tricks to maintain stability). The result is a model that can have an 8x smaller memory footprint (1-bit vs 8-bit) and potentially use correspondingly less energy. In their experiments, BitNet models achieved comparable perplexity to FP16 models while significantly reducing memory and energy usage. Moreover, BitNet models followed similar scaling laws to full-precision ones, suggesting they can be scaled up without hitting a performance wall. This is a radical reduction of the *pretraining burden*: a 1-bit 100B-parameter model effectively might consume resources like a 12.5B model in 8-bit. BitNet is still cutting-edge research, but it signals a future where model sizes (in terms of effective parameters) can grow with far less cost.
2. **Low-Bit Quantization:** Even outside of specialized architectures, applying post-training quantization dramatically reduces resource needs. It’s now

common to serve models in 8-bit or 4-bit integer formats with negligible accuracy loss. For training, 8-bit optimizers (such as 8-bit Adam from the BitsandBytes library) reduce memory by half for optimizer state, and mixed precision training (FP16/BF16) is standard. The frontier is *training in 4-bit or 2-bit*. QLoRA already demonstrated fine-tuning in 4-bit effectively. Going lower often requires new techniques (since naive 2-bit training can fail due to quantization error). Research into ternary or binary networks (like XNOR nets in vision) is being adapted to LLMs. If successful, one could imagine pretraining a model in say 4-bit from the start, cutting the GPU memory and communication costs by 4x. This would directly translate to needing fewer GPUs or fitting larger models in the same budget.

3. **Efficient Attention and Sparse Models:** Large model training is also benefitting from architectural tweaks that reduce complexity. *Sparse Transformers* can cut down the $O(n^2)$ attention cost. For example, a MoE model with 16 experts (each smaller) can outperform a dense model of the same size with less compute, because for each token only 2 experts are active (making the forward pass sparse). Google's Switch Transformer showed massive models (trillion+ parameters) could be trained at the same cost as dense models 1/4 the size, thanks to MoE routing. For enterprises, using MoE or other conditional computation means you don't have to run every computation for every input – saving time. Another innovation is *Retentive Networks or RWKV*, which re-imagine the transformer with RNN-like characteristics to allow streaming and potentially lower memory usage. These are more experimental but indicate that the community is actively trying to find architectures that get more out of each FLOP.



By embracing quantization and architecture innovation, one can *significantly cut down the effective compute requirements* for training and inference. Quantization in particular is a low-hanging fruit – there is little reason today to serve an LLM in full 16-bit precision when 4-bit works. The savings multiply across the whole pipeline.

12.3 Distributed and Hybrid Computing Approaches

When local hardware is limited, why not combine resources across many machines? Distributed computing for LLMs comes in two main flavors: within an organization (cluster or cloud) and across organizations (collaborative networks).

1. **Cluster-based Distributed Training:** This is the typical approach where a training job is split over multiple GPUs/nodes. Libraries like PyTorch Lightning, DeepSpeed, and Ray Train simplify doing data-parallel or model-parallel training on a cluster. For enterprises with several smaller GPU machines, using *distributed data parallel (DDP)* allows training a larger model than any single machine's memory. Techniques like Fully Sharded Data Parallel (FSDP) partition not just data but model parameters and optimizer state across nodes, enabling training of models that wouldn't fit otherwise. While distributed training has overhead (communication costs, engineering complexity), it can harness a pool of mid-range hardware to achieve something approximating a high-end system. As an example, if you have 10 machines each with an 8GB GPU, you could train a 10x larger model by sharding it across them (with careful synchronization). This is often how academic labs tackle LLMs with limited budget – by efficiently using a handful of GPUs with software tricks.
2. **PETALS & Volunteer Computing:** A novel development is decentralized inference/training networks like PETALS. Petals uses a peer-to-peer swarm of volunteer hosts, each hosting part of an LLM's layers, to collectively serve or even fine-tune the model. It's akin to BitTorrent for LLMs: anyone can contribute some GPU memory, and in return, everyone can use the large model. Petals has successfully demonstrated running inference for 100+ billion parameter models over the internet 10× faster than offloading to disk. It achieves this through pipeline parallelism over network: a forward pass is split among hosts, and a distributed routing algorithm finds an efficient path. Fault tolerance is handled by replicating blocks and dynamically reconfiguring if a host drops. This approach is highly compelling for countries or groups that individually only have a few GPUs: together, they can form a *virtual supercomputer*. For instance, 50 volunteers with one 24GB GPU each could host a

$50 \times 24 = 1200$ GB model in theory – well beyond any single participant’s ability. Petals also supports collaborative fine-tuning, meaning a community could fine-tune a model on a shared dataset by each doing a small part of the work, instead of one entity bearing it all.

3. **Hybrid Edge-Cloud Solutions:** Another approach is splitting computation between local devices and cloud servers. For example, where an SLM runs at the client or edge level generating drafts that are verified at the client side using a reward model, and if the reward doesn't meet the accuracy of a larger model then the larger model is requested to create/verify the draft, any such edits are cached at the edge for future uses as well. This can reduce the cloud/server usage by up to 90% while ensuring the accuracy of the cloud model at the client/edge.

In all, distributed and hybrid strategies enable scaling beyond local limitations. They demand smart orchestration, but a well-designed distributed training can turn a network of ordinary machines into a formidable LLM factory. Similarly, volunteer and federated approaches can democratize access by pooling resources of many actors to achieve what none could alone.

12.4 Memory and Compute-Efficient Optimizers

Optimizers are the algorithms that update model weights during training (SGD, Adam, etc.). The choice of optimizer affects not just model convergence but also memory and compute overhead. For LLMs, AdamW has been a standard, but it requires keeping two extra momentum tensors of the same size as the model (doubling memory) and performs many math operations per step. New optimizers aim to be lighter:

1. **Adafactor:** Adafactor is a variant of Adam that reduces memory usage by not keeping a full second moment matrix for each weight; instead it factorizes the second-moment statistics into per-row and per-column vectors. This cuts memory substantially (especially for very large layers) and was used by Google to train T5 models. Adafactor in its basic form has no momentum, though later versions allow a memory-efficient momentum. It achieves nearly the same convergence as Adam on large-scale tasks with far less memory overhead. For

someone training a model at the edge of GPU memory, switching to Adafactor can make the difference between fitting or not fitting the model.

2. **Lion (EvoLved Sign Momentum):** Lion is a recently introduced optimizer that was discovered through neural optimizer search. It's essentially Adam but only uses the *sign* of the gradients with momentum for updates. The crucial aspect is that Lion only keeps momentum, dropping the second moment tracking of Adam. This means it uses roughly half the memory of Adam and also has simpler update computations (no bias-corrected variance). Empirical results showed Lion can slightly outperform Adam in quality while being more memory- and compute-efficient. It has been tested on vision and language models, and generally if one tunes learning rates appropriately, it converges similarly. The benefit for low-resource training is that you can train larger batches or models before running out of memory, and also potentially see faster step times due to fewer operations.
3. **Sophia, Apollo and Others:** There's a proliferation of optimizers (Sophia, AdaClamp, Shampoo, etc.) each with their pros/cons. *Sophia* for instance tries to approximate second-order information cheaply, which can converge in fewer steps (saving compute) albeit with some overhead per step. *Apollo* uses adaptive momentum. *SGD with momentum* is actually the most compute-efficient (least ops per step), but it's been found to yield worse final performance in LLM training – one study noted that all adaptive optimizers (Adam, Lion, Adafactor, etc.) performed similarly and much better than plain SGD on language modeling. Thus, completely dropping adaptivity to save compute is not wise. However, one insight from that study was that only certain parameters (like output layer and layer norms) truly need per-parameter learning rates; freezing learning rates for others didn't hurt much. This suggests future optimizers might strategically apply adaptive logic to a subset of parameters and use simpler updates for the rest, getting the best of both worlds.
4. **Memory-Efficient Training Techniques:** Beyond optimizers, there are training loop tricks to reduce memory: gradient checkpointing (trading recomputation

for memory), optimizer state sharding (as in ZeRO), and even offloading gradients to CPU if GPU memory is tight (slower but sometimes necessary). Researchers have also looked at *activation sparsity* – if a large portion of neurons are inactive (ReLU-like sparsity), one can skip gradient updates for them (this is in early stages for transformers though). Another approach is Online Subspace Training, where the idea is to restrict gradients to a lower-dimensional subspace at a time, reducing the number of variables being actively updated (thus reducing optimizer memory). Techniques like this effectively say: you don't need to adjust all 100% of weights simultaneously, you can update a subset, offload the rest, and swap through them – potentially reducing memory and maybe even noise.

In practice, a simple change like switching to 8-bit optimizers or Lion can give immediate memory savings. For a small team, that might let them run a 13B model fine-tune on a single 24GB GPU (which is possible with QLoRA + 8-bit optimizers), instead of needing 2-3 GPUs with the default Adam. Compute-efficient optimizers that converge faster (like reaching the same accuracy in 50% fewer steps) effectively halve the compute cost if they work as advertised, which is another huge win when every GPU-hour counts.

12.5 Geopolitical and Infrastructure Risk Mitigation

As countries and enterprises invest in AI, they face strategic decisions beyond just technical ones. This section discusses mitigating risks related to hardware supply, cloud dependency, and preserving sovereignty over AI capabilities. Key considerations include using heterogeneous hardware, maintaining national model repositories, avoiding over-reliance on hyperscalers, and weighing the benefits of owning vs. leasing infrastructure.

Heterogeneous Hardware Strategies

Relying on a single vendor or type of hardware can be risky. Supply chain disruptions, export controls, or price changes can stall an AI initiative. The prudent strategy is to design solutions that are hardware-agnostic and heterogeneous:

1. **Multi-Vendor Support:** Ensure that your LLM software stack (training code, inference servers) can run on NVIDIA, AMD, Intel, or even emerging AI chips with minimal friction. This might mean using frameworks like ONNX Runtime or PyTorch with oneAPI that can target multiple backends. By keeping flexibility, an organization can pivot if, say, NVIDIA GPUs become hard to get or expensive. We see this thinking in the Chinese AI community, where U.S. export restrictions on high-end GPUs have prompted investment in domestic accelerators (e.g. Huawei Ascend, Alibaba Hanguang). Their LLM implementations are being ported to those platforms. Western companies similarly might want at least a proof-of-concept running on AMD Instinct GPUs or Intel GPUs to avoid a monopoly-induced risk.
2. **Older Hardware and Mixed Clusters:** Not everyone can buy new GPUs at will. But perhaps one has a cluster of older V100s or even TPUs. Using them effectively is part of heterogeneous strategy. Techniques like model slicing can allocate different layers of a model to different hardware types based on their strengths. For instance, an attention layer might run fine on a CPU if it's small, while big matrix multiplies go to a GPU. If bandwidth allows, mixing CPU and GPU in a training job (pipelines, offloading) can increase utilization of all resources. While not ideal, this can be a bridge solution in resource-limited environments. The *risk* of heterogeneous setups is software complexity – however, new orchestration tools and ML compilers (like TVM, TensorRT with fallbacks) are making it easier.
3. **Federated and Collaborative Training:** From a geopolitical standpoint, if data cannot leave a country due to privacy (think EU's GDPR or a nation's data sovereignty laws), one can bring compute to the data via federated learning. For example, hospitals in different regions might each train the model on local data and only share model updates (not raw data) with a central aggregator. This mitigates risk of data exposure and can leverage distributed data sources without a single data center. The flip side is increased communication cost and complexity. Still, for certain applications (like medical LLMs across borders), this is an attractive, risk-aware approach.

12.6 National AI Model Repositories and Ecosystem Building

A key part of *AI readiness* for a country is having control and access to models and tools. Relying solely on foreign APIs or closed models can be a strategic vulnerability. Thus, we see moves towards national AI model catalogues and open ecosystems:

1. **Hugging Face as a Global Model Hub:** The rise of Hugging Face’s Transformers Hub has provided a centralized place to share models (over 100,000 models as of 2025). Many governments and institutions actively use it. However, it is a commercial entity based in the U.S./France; some nations may prefer hosting their own repository for critical models (especially if internet access is an issue or they want curation). The concept of a *national model catalogue* is to have an official repository (perhaps run by a government lab or consortium) where validated models (and possibly domain-specific ones, like for healthcare or education) are stored and made available to domestic companies. China’s ModelScope, backed by Alibaba, is an example of this approach: it’s a platform hosting many AI models including LLMs, with an eye towards the Chinese developer community and compliance with local regulations.
2. **Model Zoos and Adapter Repositories:** Beyond full models, the sharing of fine-tuning components (like LoRA adapters or prompts) is important for collaboration. Communities have created things like *AdapterHub* (for sharing adapter modules for various tasks). Encouraging a culture of sharing these “building blocks” can accelerate progress nationwide. If one university fine-tunes an adapter for a local language, they can publish it for others to plug into the base model. This avoids duplicate work. Policymakers could incentivize such sharing (e.g. require that publicly funded AI projects release their models/adapters to the national repository).
3. **Data and Tooling Support:** A model is only as good as its training data and tools. National efforts should also include curated datasets (like a national corpus including government documents, literature, etc., cleaned and prepared

for training) and investment in open-source tooling (frameworks, libraries) that locals can use without legal or cost barriers. For example, the Indian government's *Bhashini* initiative is creating translation datasets and models for Indian languages, hosted for public use. This mitigates the risk of Indian tech being stuck with only English-proficient models or paying for expensive translation APIs.

By cultivating a robust internal AI ecosystem – models, data, skills – a country ensures it can adopt AI on its own terms and continue progress even if external access is cut off or becomes too costly. It also reduces *brain drain*, as local AI talent sees that they can do cutting-edge work at home with these shared resources.

12.7 Hyperscaler and OEM Dependency Risk mitigation with hardware selection methodologies

Hyperscalers (the big cloud providers: Amazon AWS, Microsoft Azure, Google Cloud) offer attractive on-demand compute for AI, but over-reliance on them has downsides:

1. **Cost and Lock-in:** Cloud is essentially renting. For sporadic or initial experiments, it's great because you avoid capital expenditure. But for sustained workloads, the costs can dwarf owning hardware. Estimates have shown that training large models on cloud can be 2× or more the cost of on-premise over time. Cloud providers also have egress fees (getting your data/model out) and you might build your stack around their proprietary services, making it hard to switch – *lock-in*. If a hyperscaler changes their pricing or terms (or faces outages), your project could be disrupted. To mitigate this, some organizations adopt a *multi-cloud strategy* (spreading work across AWS/Azure/GCP to avoid single-provider risk) or design portable workflows (e.g. using Kubernetes or Terraform so things can be moved). However, multi-cloud can be complex and you might lose volume discounts.
2. **Geopolitical Risks:** Using foreign cloud providers may raise sovereignty issues. E.g., EU regulators worry about sensitive data being processed by U.S. companies

(hence the push for “sovereign cloud” solutions that keep data in-country). If relations sour or export rules shift, a cloud provider could theoretically restrict service to certain users or regions. This is not a purely hypothetical scenario – we’ve seen instances of software access being revoked due to sanctions. Owning at least part of the inference infrastructure (for critical systems) ensures continuity under various circumstances.

3. **Hardware OEM Dependency:** The majority of advanced AI runs on NVIDIA GPUs. This concentration poses a risk: if NVIDIA cannot deliver chips (due to supply chain or export restrictions), it bottlenecks AI progress. We already see the U.S. restricting top-tier NVIDIA GPUs to certain countries. One mitigation is stockpiling (some firms are literally buying and hoarding years’ worth of GPUs when they can). Another is nurturing alternative hardware ecosystems: AMD’s MI250/MI300 GPUs, FPGA-based accelerators, or indigenous chip development (like Europe’s EPI or India’s upcoming AI chips). Governments might give grants or form partnerships to ensure they have at least one home-grown option for AI computing – much like how some countries ensure they can build their own supercomputers independent of foreign tech.

A balanced approach is prudent: use hyperscalers for elasticity and when trying things out, but for core long-term workloads, consider investing in owned infrastructure. And avoid betting everything on one vendor’s roadmap.

Owning Infrastructure vs. Leasing (Cloud)

This is an age-old debate with new twists in the LLM era. Owning means buying servers/GPUs and running them (on-premise or in co-location), whereas leasing means using cloud or HPC centers on rental basis.

Benefits of Owning:

1. **Lower Long-term Cost:** If utilization is high (e.g. training models or serving inference 24/7), owning is generally cheaper. A Deloitte analysis found on-prem HPC could be ~50% the cost of equivalent cloud hours when hardware is well-utilized. Essentially, you pay a fixed cost up front, then depreciation, but

you're not paying the cloud's premium and profit margin. Organizations like Meta, Google always build their own for this reason at scale.

2. **Control and Customization:** When you own hardware, you can optimize it specifically for your workloads (choose specialized interconnects, tune cooling, even modify hardware). You're not beholden to cloud configurations. You also control scheduling – no surprise interruptions for maintenance or noisy neighbors on shared cloud nodes. For enterprises that need predictable performance (financial models, etc.), this control is valuable.
3. **Data Governance:** Keeping everything on-prem means data doesn't leave your facility. For highly sensitive data, this is non-negotiable. Some sectors (defense, healthcare) often mandate on-prem or approved private cloud only.
4. **Strategic Independence:** Nationally, having sovereign compute (like a state-funded supercomputing facility) ensures that academia and industry have access to AI compute even if external services become limited. Countries like France have the Jean Zay supercomputer which was used to train the Bloom model – showing how national infrastructure can produce globally relevant outcomes.

Benefits of Leasing (Cloud):

1. **No CapEx and Quick Scaling:** You avoid the huge upfront cost of buying, and can scale up and down quickly. This is great for startups or research groups who occasionally need a burst of 100 GPUs for a week but then could go idle – owning those 100 GPUs to use rarely would be wasteful, whereas cloud you just pay for that week.
2. **Maintenance and Updates:** The cloud provider manages hardware failures, upgrades to new GPU generations, etc. If a GPU dies, it's their problem to replace it. On-prem, you need an ops team to handle this. Cloud also provides easy access to a variety of hardware (TPUs, latest GPUs immediately, etc.), while if you

bought GPUs, you might be stuck with last-gen until you can afford new ones.

3. **Global Availability:** Cloud data centers are worldwide, so you can run compute close to where your users are or duplicate across regions for resiliency. An on-prem data center is usually one location (with maybe a backup site) – harder to achieve that geo-redundancy on your own unless you’re big.

Hybrid: Many larger organizations choose a hybrid model – keep a baseline on-prem capacity for steady workloads, burst to cloud for spikes. This can yield cost savings while retaining flexibility. For example, a company might train models on-prem but do hyperparameter sweeps or short-term experiments on cloud VMs to not clog their own cluster.

From a risk perspective, owning infrastructure insulates you from external shocks like sudden cloud price hikes, data transfer restrictions, or geopolitical embargoes. However, it requires capital expenditure and know-how to operate effectively. A nation might invest in a national AI Supercomputing Center (like how some have petaflop supercomputers for science) to support domestic AI needs – effectively acting as a private cloud for its citizens.



Enterprises and countries should evaluate their steady-state compute needs vs. peak needs, budget constraints, data sensitivity, and choose an infrastructure strategy that hedges risks accordingly. For mission-critical AI (defining products or national projects), owning core infrastructure is often recommended for the assurance it provides, whereas cloud can supplement for non-critical or variable demands.

12.8 Resource-Aware Model Architectures

Designing models and training schemes that inherently consider limited data and compute is crucial for inclusive AI development. In this section, we discuss approaches

for *localized model adaptation*, preventing catastrophic forgetting when expanding a model's knowledge, and architectures conducive to low-data, multilingual scenarios.

Localized Model Adaptation for Diverse Cultures & Low-Resource Languages

One size does not fit all in language models. Countries with multiple languages or distinct cultural contexts face the challenge of adapting LLMs that were often pretrained predominantly on English and other high-resource languages. Resource-aware adaptation means using the data you have efficiently:

1. **Multilingual Joint Training:** If data for each individual low-resource language is scarce, training a single model on many languages together can help. The model learns a shared representation that transfers knowledge between languages (e.g., it might learn a concept in English and align it with the word in Swahili if given some parallel data). Models like mBERT, XLM and mT5 followed this approach, allowing over a hundred languages to be handled by one model. For LLMs, the same idea applies – *mix languages in pretraining*. The catch is that model capacity gets divided among languages, so very low-resource ones might still not get enough relative share. A remedy is to up-sample the low-resource language data during training and/or add language-specific tokens that help the model identify and separate languages. The BigScience project did this with Bloom, including 46 natural languages and 13 programming languages in training, and specifically making sure “long tail” languages (like Lao, Maltese, etc.) were repeated more times so the model sees them sufficiently.
2. **Culturally Diverse Data:** Cultural knowledge is not just language; it's also about idioms, social norms, local facts. A model pretrained on Western internet might not know much about folklore or popular culture in another country. Incorporating *local texts* (newspapers, novels, social media from that country) during training or via continuous pretraining can give the model more culturally relevant knowledge. Moreover, fine-tuning on dialogue data that reflects local conversational styles can make the AI's responses feel more

natural to users in that culture.

3. **Domain Adaptation on Sparse Data:** In some cases, even domain data is limited (e.g. a specialized scientific field). Techniques like few-shot learning or data augmentation can help. For example, if you only have 1,000 legal Q&A pairs, you might prompt a larger LLM to generate more synthetic Q&A in that domain (similar to how Alpaca generated its data). Knowledge distillation from a model that has seen more (maybe a teacher model that can access a bigger corpus) into a smaller model is another tactic. Essentially, you use any external knowledge source to amplify the effect of your small dataset.

12.9 Using Adapters to Prevent Catastrophic Forgetting

We've touched on adapters for efficiency, but they also shine for incremental learning. Catastrophic forgetting is when a model fine-tuned on new data loses performance on its original capabilities or knowledge. Adapters offer a solution: isolate new knowledge in dedicated parameters.

Imagine you have a base LLM that's good at general tasks. Now you want to teach it a new skill (say, writing poetry) or update it with this year's events. If you fine-tune the whole model on the new data, it might overwrite weights that were important for other tasks (suddenly it might get worse at factual Q&A while becoming poetic). Instead, you attach a fresh adapter (or LoRA module) and train that on the new data, keeping the original weights frozen. The adapter learns the new skill; if it's small relative to the model, it won't interfere much with existing functions. At inference, the model with adapter can do the new skill, but if needed, you could also disable/unplug the adapter to get back the original behavior.

This approach has been validated in research: one study on continual learning found that using separate adapters for each new task greatly mitigated forgetting, since the core model wasn't being rewritten. Some techniques even enable conditional routing: e.g., a task token can activate the relevant adapter for the task. Google's *PALM* used a

form of this for multitask training, and there's work on learning routing to appropriate expert adapters given the input.

Additionally, *soft modularization* methods like SLIM (Soft LoRA Injection Mixture) allow a model to blend between multiple LoRA adapters and the identity (no change) dynamically. SLIM showed that by gating the influence of LoRA on a per-input basis, the model could retain general ability and only apply the specialized weights when appropriate, further reducing interference between tasks.

For multilingual models, language-specific adapters have been used to add a new language without hurting existing ones: train a new adapter on that language, keep others frozen. Facebook's LASER and similar projects successfully added support for languages this way. When generating, the model uses the adapter corresponding to the language it's outputting in.



Adapters act as “memory compartments” – each new knowledge area gets its own compartment, so it doesn't overwrite the others. This strategy is extremely useful for enterprises updating models with new data continuously (like a search engine's LLM that learns from new documents daily) or for joint models that serve many tasks.

12.10 Architectures for Low-Data Alignment and Multilinguality

Beyond training tricks, certain architectural designs inherently facilitate doing more with less data:

1. **Instruction Tuning with Mixture-of-Tasks:** Instead of needing a large dataset for every distinct instruction/task, models are often *jointly tuned on a mixture of many tasks' small datasets*. This approach, used in T0 and FLAN, aligns the model to the “format” of following instructions generally. The model then can generalize to new unseen tasks (zero-shot) surprisingly well. What this means

for low-resource settings is that you don't necessarily need *in-domain data* for every *capability* – you can leverage public multitask data to create a broadly instruction-following model, then just lightly tweak it to your specific domain. This massively reduces data requirement for alignment. OpenAI's usage of instruction tuning on GPT-3 (to get text-davinci-002) is a case in point: they likely used a wide array of instruction examples (from summarization to coding to QA) to make a single model adaptable to anything.

2. **Multi-Modal, One-Model Approaches:** For countries that have diverse data modalities but not a lot in each (e.g., some text, some speech transcripts, some images with captions in a local language), a single multi-modal model that learns from all can be beneficial. Recent M-LLMs like LLaVA and PaLI-X combine vision and language, and some also combine speech. A multi-modal model can use supervision across modalities to strengthen its language understanding. For instance, an image of a local landmark with a caption teaches the model about that landmark in a way that pure text might not. Multi-modal training can act as an augmented data source – e.g. if you don't have text about a cultural concept, maybe you have pictures and descriptions of it, which then inform the language side of the model. This is especially relevant in areas where oral or visual culture is richer than written text.
3. **Retrieval-Augmented Models:** One clever way to handle low internal knowledge is not to put everything into the model's parameters. Retrieval-Augmented Generation (RAG) architectures equip the model with a retriever that can fetch relevant documents from an external database (which could be as simple as Wikipedia or a curated document set). Then the model conditions on those retrieved facts to produce the answer. This means the model itself can be smaller and trained on less, as long as it knows *how to read retrieved context*. For local use cases, one could maintain a local knowledge base (for example, government documents, local news) and let the model search that when needed. Instead of pretraining the LLM on all that data (expensive), you just need a solid retriever (which can be built using smaller models or even keyword search) and some fine-tuning of the LLM to incorporate retrieved info.

This setup is data-efficient because you don't require the LLM to memorize everything – it can look up details on the fly. It's like giving the model a bookshelf so it doesn't have to hold every book word-for-word in its brain.

4. **Sparse Activation Models:** Mentioned earlier, mixture-of-experts (MoE) models have subsets of weights active per input. This inherently allows training each “expert” on the data pertinent to it (for example, an expert for each language, or each topic). In Google's recent Switch-C and GLaM models, they had experts that clearly specialized in certain languages or genres. Because each token only trains one expert, the effective data needed per expert is lower (it's not seeing irrelevant data). If some experts are designated for low-resource languages, they can focus on those and not be diluted by high-resource language data. The router ensures they only fire for the appropriate inputs. This targeted learning is an architectural way to cope with data imbalance.
5. **Feedback and Reward Models for Low-Data Alignment:** When human examples are scarce, training a reward model on even a tiny set of human preferences and then using it to do automated reward optimization (like through RLAIIF – Reinforcement Learning from AI Feedback) can stretch a few human data points much further. Facebook's HH- series of models (like HH-RLHF) showed that you can first fine-tune on some general instructions, then use a *handful* of human preference data to train a reward model that captures, say, “avoid toxic replies”, and then use that reward model to guide generations. Essentially the reward model amplifies the impact of limited human feedback. Direct Preference Optimization (DPO) similarly can work with small comparison datasets to align a model without a full RL loop. For a policymaker wanting an aligned model but not having the budget for millions of human annotations, focusing on training a reliable reward model on a smaller curated set, and letting it drive self-training of the LLM, is a viable strategy.

13. Efficient Inference Systems for GenAI in Resource-Constrained Environments

Developing countries often face resource constraints that make deploying massive Generative AI models challenging. Recent open-source research and frameworks have focused on efficient inference techniques to enable large-scale GenAI deployment on limited GPU/CPU infrastructure. This survey covers seven major areas of advancement: efficient model architectures, split inferencing, collaborative decoding, distributed/federated inference, decentralized multi-model approaches with prompt routing, the strategic use of Small Language Models (SLMs), and test-time scaling techniques. Each area is crucial for maximizing tokens-per-dollar, minimizing energy (Watts/TFLOPs) and memory bandwidth (MBW) use, and reducing latency in low-resource settings.

13.1 Efficient Model Architectures

State-of-the-art model architectures are being redesigned for efficiency. These approaches trade off some complexity or precision for dramatic gains in speed, memory, and energy usage:

BitNet (1-bit Transformers): BitNet is an extreme quantization architecture where weights are 1-bit and activations are low-bit, trained from scratch to preserve accuracy. Microsoft's open BitNet models (e.g. BitNet b1.58) achieve remarkable efficiency – bitnet.cpp runs a 100B-parameter BitNet on a single CPU at 5–7 tokens/s, with $2.4\times$ – $6.2\times$ speedups and ~72–82% less energy on x86 CPUs versus FP16 models. By drastically reducing memory bandwidth needs, BitNet improves tokens per dollar and watt, showing that aggressive quantization can maintain performance with far lower compute cost.

Mamba (Linear-Time SSM Architecture): Mamba replaces self-attention with a Selective State-Space Model (SSM) that runs in linear time, eliminating the quadratic scaling of transformers. A pure Mamba model enjoys $5\times$ higher throughput than a

transformer and constant memory usage (no growing KV cache). Notably, a 3B Mamba matches the accuracy of a 6B Transformer. Hybrid LLMs interleaving Mamba and transformer layers (e.g. Jamba , Samba) further combine strengths. The open 9B Bamba model (Hybrid Mamba2) demonstrates 2.5× throughput and 2× lower latency than a standard 8B transformer at inference. These hybrid LLMs alleviate memory bottlenecks (constant-size caches) while retaining accuracy, making long-context inference more feasible on limited hardware.

Liquid Neural Networks: Liquid Neural Networks (LNNs) are inspired by dynamical systems, with neurons described by continuous-time equations that adapt to input streams. They excel in low-resource scenarios by being highly efficient and robust . For example, an LNN with only 19 neurons achieved parity on a task that normally needed 100k conventional neurons – an enormous reduction in model size. LNNs can also dynamically adjust to changing data, reducing retraining needs. Critically, LNN models operate on edge devices with an order-of-magnitude lower power consumption than transformers. Early results show 10× lower power usage in some cases, with LNN-based “Liquid” foundation models delivering state-of-the-art class performance with a smaller memory footprint. This makes LNNs promising for energy-efficient AI in regions with limited power or hardware.

Multi-Head Linear Attention (MLA): Transformers with linear-time attention drastically cut complexity. MLA refers to techniques that make self-attention linear in sequence length, often by kernel feature maps or low-rank approximations. Recent implementations like FlashMLA focus on memory bandwidth efficiency during autoregressive decoding. By managing KV caches in blocks and tiling computation in GPU shared memory, FlashMLA avoids memory stalls and serves multiple requests efficiently. In practice, linear attention methods (e.g. TransMLA) have shown expressive power on par with full attention while scaling to long sequences more gracefully. This translates to lower latency per token on long inputs, ideal for low-bandwidth environments.

Multiplication-Free Attention: Removing expensive matrix multiplications further improves efficiency, especially on CPUs. NoMAD-Attention introduces an attention algorithm that replaces multiply-add operations with ultra-fast in-register table lookups on modern CPUs. This technique maintained output quality while speeding

up a 4-bit quantized LLaMA-7B by up to $2\times$ at 16k context. Likewise, ShiftAddLLM uses bitwise shifts and additions instead of multiplications in attention and feedforward layers. The result is over 80% reduction in memory and energy usage versus the original model, at comparable accuracy to aggressive 2–3 bit quantization. These multiplication-free mechanisms improve W/TFLOP efficiency by exploiting cheaper operations, which is valuable for deployment on CPUs and low-power devices common in developing regions.

Overall, these architectural innovations (quantization-aware models, state-space and liquid networks, efficient attention, etc.) dramatically improve tokens per second per watt. For instance, BitNet’s 1-bit design tackles the memory bandwidth wall, and Mamba-based hybrids eliminate context-length penalties. By evaluating them on metrics like energy per token or latency, researchers have shown $5\times$ – $10\times$ gains in throughput and large drops in energy use. Such efficient open-source architectures are enabling local inference of models previously considered infeasible outside major data centers.

13.2 Split Inferencing (Edge–Cloud Collaboration)

When devices have limited capacity, split inference techniques partition the workload between the client (edge) and server (cloud). This collaborative approach minimizes on-device computation while reducing server load and preserving privacy:

Layer/Stage Partitioning: One strategy is to run the early layers of a model on the edge device and send the intermediate activations to a server that completes the forward pass. This reduces data transmission (only features, not raw data, are sent) and balances compute. For example, SplitNN paradigms and recent systems like SplitLLM use dynamic programming to find an optimal split point given the network and device speeds. Experiments show a properly chosen split can halve server workload with negligible loss in quality, improving overall throughput of the system. The challenge is the communication overhead of sending activations, which must be offset by the computation saved.

Tiered Collaborative Decoding: Beyond a simple one-time split, some frameworks employ a tiered decoding pipeline. Jupiter (2025) is a system for multi-edge

collaboration in generative LLM inference that separates the prefill phase and autoregressive decoding phase across tiers. In Jupiter, a flexible pipeline parallelism loads the initial prompt across several edge devices (or edge+cloud) for the heavy first forward pass, then uses an outline-based pipeline (with speculative decoding) for the token-by-token generation. This tiered approach yielded up to $26\times$ lower end-to-end latency compared to single-device inference. Essentially, easy portions of the computation are distributed broadly (many small devices), and the final assembly is done in a coordinated way, maximizing throughput without overloading any one node.

Client-Side Draft, Server Verification: Another collaborative pattern is having the edge device generate a draft output using a small local model, and then the server's large model validates or corrects it. This is a form of split computation across two passes. It balances compute (most tokens proposed by the cheap edge model) and bandwidth (only a draft and minimal feedback cross the network). It also keeps the user's prompt and draft mostly local, addressing privacy. Approaches like Hybrid SLM-LLM inference implement this: an on-device SLM produces tokens, and a cloud LLM intervenes only when needed to maintain quality. The hybrid edge-cloud method by Hao et al. achieved LLM-level accuracy with only 25% of the usual LLM compute cost by using a TinyLlama SLM for the bulk of work and calling the cloud LLM sparingly. This tiered decoding ensures state-of-the-art results while drastically cutting inference cost – a boon for regions with sparse GPU resources.

Split inferencing offers privacy and efficiency: user data can be processed locally up to a point (sensitive feature extraction on-device), and only non-sensitive representations are sent to powerful servers. It also reduces latency by parallelizing work. The key is careful scheduling to avoid communication becoming a bottleneck. Recent research addresses this with compression of intermediate data and asynchronous pipelines (as seen in Jupiter's design). For developing countries with patchy connectivity and limited hardware, split inference provides a practical balance between local processing (for data locality) and remote processing (for heavy lifting), enabling GenAI services that are both fast and compliant with data privacy needs.

13.3 Collaborative Decoding Strategies

Beyond static splitting, collaborative decoding techniques allow multiple models to

work together during generation to speed up inference and reduce cost:

Speculative Decoding: This method uses a faster draft model to generate several tokens ahead, then has the large model verify them in one go. If the large model agrees on the draft tokens, they are accepted, otherwise it falls back to normal generation for that step. Google’s speculative decoding showed 2–3× speedups with no loss in output quality, since the final distribution is identical to the original LLM’s. In effect, the large model “skips” computation by trusting the small model’s speculation most of the time. This reduces latency and cost per output by requiring fewer sequential calls to the big model. Speculative decoding has been widely adopted (e.g. in Google’s production systems) because it guarantees the same results as standard decoding while reducing inference time by over 2×.

EAGLE: Extrapolation Algorithm for Greater LLM Efficiency (EAGLE) is a state-of-the-art speculative decoding framework. It extrapolates internal features of the LLM’s decoder to predict future tokens efficiently. EAGLE was found to be the fastest known decoding method (as of 2024), achieving ~ 3× faster generation than vanilla decoding and outperforming earlier methods like Lookahead and Medusa. Notably, EAGLE sped up a 13B model’s decoding by 3× while provably preserving the same output distribution. Its successor EAGLE-2 further boosts speed (4× faster than normal decoding) by using dynamic draft trees that adjust to the confidence of token predictions. These techniques show that collaboration within an LLM’s layers (by predicting high-level features) can yield big inference gains.

Medusa: Medusa takes a different approach by adding multiple decoding heads to a single LLM, allowing it to predict several tokens in parallel. Instead of a separate draft model, the LLM itself is augmented (via fine-tuning) with extra output heads that jump ahead in the sequence. During generation, Medusa’s heads propose multiple next tokens (forming a tree of possibilities) and a lightweight procedure then selects the longest valid token sequence from those candidates. This effectively parallelizes autoregression. Medusa is simpler to deploy (no second model needed) and only the new heads are trained (a parameter-efficient fine-tune). The latest results show 2.2–3.6× speedups on Vicuna and other models. For example, Medusa-1 achieved ~2× faster generation for Vicuna-7B with batch size 1. Medusa’s approach demonstrates that even single models can “collaborate with themselves” by internally branching out

multiple token predictions – democratizing fast decoding without complex system orchestration.

SLM Primary + LLM Verification: In scenarios aiming for maximum cost savings, a small language model can do most of the work and a large model only verifies or corrects outputs. This idea is seen in Coarse-to-Fine decoding or frameworks like CoSine (Collaborative Speculative Inference). CoSine uses multiple specialized SSMs (small speculative models) as drafters that generate candidate tokens, which are then verified in parallel by the LLM. By decoupling drafting and checking across distributed nodes, and routing each query to the most “expert” small model, it achieves higher acceptance rates and efficiency. CoSine improved inference throughput by 32.5% and latency by 23% over standard speculative decoding in a multi-node setup. Similarly, the Hybrid SLM-LLM edge-cloud method described earlier (TinyLlama + cloud GPT) can be viewed as collaborative decoding: the SLM generates tokens and only uncertain cases invoke the LLM. These approaches prove that letting a cheaper model drive the generation and a powerful model oversee it can maintain quality while significantly cutting inference cost – an attractive strategy when GPU hours are scarce.



Collaborative decoding methods exploit redundancy in generation: a smaller model (or extra heads) can guess the next tokens much faster, and the big model can confirm them in one pass instead of generating each token itself. The benefits are substantial – speculative decoding and its variants often double or triple decoding speed, which translates directly into lower inference latency and cost per query. This is especially beneficial in resource-constrained deployments where large models are slow or expensive to run – by pairing them with efficient helpers, one can get LLM-level output at a fraction of the time and compute .

13.4 Distributed and Federated Inferencing

Decentralizing the load across many modest devices is another avenue to enable GenAI

at scale without a single supercomputer. Distributed inference frameworks and federated setups harness aggregate compute while respecting data locality:

Petals (BitTorrent-style Inference): Petals is an open-source system that lets anyone run parts of a large model and join a peer-to-peer network for serving inference. It breaks the model into layers (transformer blocks) and distributes them across participants' machines. At inference, the input passes through a sequence of peers hosting successive layers – analogous to a model-parallel pipeline over the internet. Remarkably, Petals can serve models with up to 176B parameters (BLOOM) on consumer GPUs at about 1 token/second. This outperforms offloading to disk and approaches interactive speeds. Petals also supports fine-tuning by exposing transformer block outputs to attach LoRA adapters. The benefit is decentralization: no single node needs to hold the entire model or dataset, aligning with federated principles. Petals' collaborative approach effectively creates a volunteer-driven inference cluster that scales out model serving across many nodes. It demonstrates that developing countries could pool existing low-end GPUs to collectively host an LLM that none could run alone.

Federated / Privacy-Preserving Inference: In sensitive applications, federated inference ensures data stays on the local device. Techniques include secure model aggregation (each client runs the model on their data and only shares anonymized outputs or gradients) and split learning with encryption for intermediate transmissions. For instance, a health NLP application might run an SLM on a hospital's server on-site, then send encoded representations to a central server running the heavy LLM part – this way raw patient data never leaves the hospital. Homomorphic encryption can further secure any offloaded computation, albeit with added latency. Another angle is model encryption: distributing encrypted model weights to edge devices that decrypt and run locally (useful if model IP needs protection but data is local). While still an emerging area, frameworks like FedML are extending federated learning to LLM inference, and research shows it's feasible to update LLMs on local data and periodically sync small weight updates – keeping data local and only sharing model improvements.

CPU/GPU Decentralized Clusters: Even without special software, communities have experimented with coordinating many CPU-only machines to serve LLMs. One

example is using Ray or MPI to shard the model across cheap cloud instances or donated computers. The key bottleneck is interconnect bandwidth; projects like Petals address this with compression and scheduling to hide network latencies. Another project, EdgeShard, specifically looked at collaborative edge computing for LLMs, where portions of the model are hosted on a cluster of edge devices (Raspberry Pi's, smartphones, etc.) and the rest on a server. They found that smart partitioning and parallel execution can make such setups competitive with single-node inference, especially for large sequence inputs that can be split. These distributed setups inherently improve resilience (no single point of failure) and allow scaling out with commodity hardware, which is cost-effective in many regions.

In all, distributed and federated inference techniques emphasize joining forces – whether it's individuals pooling hardware via Petals or organizations keeping data local and only sharing learned representations. For developing regions, this means an LLM-driven service could be powered by a network of ordinary computers (even across different cities) instead of an expensive centralized GPU farm. Additionally, by keeping computation close to data, they offer compliance with data sovereignty laws and latency benefits for local users. The trade-offs are in complexity and potential communication cost, but ongoing innovations (like efficient layer splitting and on-device optimization) are rapidly closing the gap, making decentralized LLM serving a viable reality.

13.5 Decentralized LLMs and Prompt Routing

Instead of one giant model, another approach is to use multiple smaller models working together to achieve high accuracy. These decentralized LLM frameworks often incorporate a prompt routing mechanism to dispatch queries to the best model(s) for the job:

Prompt-to-Leaderboard (P2L): Prompt-to-Leaderboard is a recent method that dynamically evaluates which model among a pool is most likely to excel on a given prompt. It trains a meta-model (P2L model) that takes a prompt and outputs a leaderboard ranking of models for that specific prompt. In deployment, the system can then route the prompt to the top-ranked model (or ensemble of models). This achieves per-prompt model specialization. Impressively, a P2L-powered router deployed on the

LMarena benchmark outperformed every individual model, ranking #1 on the Chatbot Arena by selecting the optimal model for each query. By leveraging several SLMs (each of which might be fine-tuned in different domains or have different strengths), P2L ensures that, for example, a coding question goes to a code-specialized 6B model, while a creative writing task goes to a dialogue-optimized 7B model. The net effect is an ensemble system that is both high-accuracy and compute-efficient: each prompt only runs on a small model (not a monolithic 70B), but the aggregate performance rivals very large models because of this targeted selection.

Mixture-of-Experts & Model Routing: The concept of routing prompts to different experts has antecedents in Mixture-of-Experts (MoE) models, but those typically route tokens within a single large model. Here we focus on external routing: having multiple independent models (which could all run on CPUs or small GPUs) and a coordinator deciding how to split the task. Some systems use domain classifiers or embeddings to choose an appropriate model. For instance, an architecture might include a language detector that routes the input to an SLM fluent in the detected language, or a task classifier that decides if the query is about finance, medicine, etc., and hands off to a model tuned in that field. This avoids loading an unnecessarily large general model. Research shows that ensembles of specialized smaller models can exceed a general LLM’s performance on a broad evaluation. DeepMind’s Chinchilla ensemble (12B × 4 models) matched GPT-3.5 on certain tasks with lower total parameters. The key is that each model contributes where it’s strongest. Routing strategies include hard routing (one model handles the query) or soft voting (all relevant models generate an answer and a final module picks or merges them).

Collaborative Generation (Token-Level Fusion): A more tightly integrated approach allows multiple models to interleave their generation at the token level. Recent work from MIT on CoLLM trains a base model to call upon other models mid-sentence. For example, a 7B model might generate an answer outline but whenever a factual detail is needed, it defers to a 34B knowledge model to fill in those tokens. In experiments, this token-level collaboration yielded joint systems that beat any individual model on cross-domain tasks. One learned pattern was template filling – the smaller model would lay out the answer format (leveraging its strength in following instructions), and insert calls to a larger model for complex subtasks like math or factual recall. This is

analogous to a team where a junior member does most of the writing but asks a senior expert for specific pieces. Such coordinated multi-LLM systems are complex but show the potential of micro-models on CPUs collaborating to emulate a much larger model. In practice, implementing this might involve a supervising program that manages multiple model instances and merges their outputs (ensuring consistency in the final text).

The benefit of decentralized multi-model approaches is flexibility and cost-effectiveness. Each small model can be deployed on affordable hardware and even fine-tuned independently on locally relevant data (e.g. a model tuned for Swahili language, another for legal text). A routing system or collaborative scheme then unites them to deliver strong overall performance. This also aligns with modularity: new specialist models can be added to the pool to improve certain capabilities without retraining a giant model from scratch. For regions with limited resources, it may be easier to obtain or train several 1–7B models (possibly on different machines or by different groups) than a single 70B model. Through prompt routing and ensemble techniques, these communities can still achieve high accuracy AI services collectively, exemplifying the principle that “many small models can work together to solve big problems.”

13.6 Strategic Use of SLMs Over LLMs

In many common use cases, deploying a Small Language Model (SLM) is far more practical and can be surprisingly effective relative to a Large Language Model. SLMs (ranging from say 100M to 7B parameters) offer drastically lower resource requirements, and with fine-tuning they often match or surpass larger models on specialized tasks. This is critical for developing countries where hardware is limited and inference costs must be minimal.

Efficacy of SLMs on Common Tasks: It has been observed that a smaller model fine-tuned on a domain can outperform a much larger general model on that domain. For example, a 770M-parameter DistilBERT fine-tuned for sentiment analysis can beat a 13B generic model in accuracy for that task, while being faster and lighter. Similarly, flan-tuned T5 or LLaMA models (~3B–7B) achieve very strong results on translation, summarization, and Q&A when appropriately fine-tuned. One analysis noted that a

fine-tuned small model “can often outperform larger (more expensive) models” when the task is well-defined and data is available for tuning. In real deployments like chatbots for government services, an SLM of a few billion parameters fine-tuned on the target language and domain can provide excellent answers with a tiny fraction of the compute needed for GPT-class models. Crucially, these models can run on a single modest GPU or even CPU (with quantization), meaning areas with only basic computing infrastructure can still host AI services locally.

Open-Source SLM Examples: There is a rich ecosystem of open SLMs that are suitable for low-resource settings. Meta’s LLaMA-2 7B and its derivatives (e.g. Alpaca, Vicuna-7B) are often used as a starting point; when quantized to 4-bit integers, they can execute on consumer laptops. TinyLlama (1.1B) is another project explicitly focusing on pretraining a small model from scratch on extensive data, yielding surprisingly competent language understanding in a 1B-scale model. For multilingual needs, models like Bloom 3B or XLM-R (0.5B) are available and cover many languages of Africa, Asia, etc. These can be fine-tuned on local languages or dialects with low computational cost. We also see specialized SLMs like MedAlpaca (medical dialogue, ~7B) or LegalBERT (110M) which, in their domain, significantly outperform baseline LLMs that haven’t seen that domain data. Such models underscore that bigger is not always better if the model is well-matched to the use case.

Fine-Tuning and Adaptation on Minimal Hardware: Techniques like Low-Rank Adaptation (LoRA) and other PEFT methods have made it feasible to fine-tune SLMs on as little as a single GPU (even a 12GB card) or aggregated low-cost instances. LoRA adapts only a small fraction of parameters, meaning one can, for instance, adapt a 7B model to a new task by training just 30 million parameters, greatly reducing memory and time. There have been demonstrations of fine-tuning a 6B model on a Raspberry Pi cluster and achieving good performance. Moreover, quantization-aware training and distillation can further compress models – e.g. taking a 2B model and creating a distilled 300M version that runs on mobile. These approaches allow local researchers and developers to customize models for their community’s needs (e.g. a Swahili summarization model) without requiring cloud-scale compute. The end result is an SLM that is tailored, efficient, and private.

Choosing SLMs over LLMs is often a trade-off between general ability and practical

deployability. In a constrained environment, the latter usually wins out – a slightly less sophisticated answer from a model that can actually be deployed is better than an ideal answer from an unusable model. Encouragingly, with good data and fine-tuning, the gap in quality can be very small. Users have found that for many routine applications (customer support bots, elementary education tutors, information lookup, etc.), a well-tuned 7B model provides an experience on par with a 70B model, at perhaps 1/10th the runtime cost. Additionally, running entirely on local servers or devices avoids reliance on expensive internet connectivity and cloud APIs, which aligns with the self-reliance goals of many communities.



Small models are the workhorses of resource-constrained AI, and strategic use of SLMs – boosted by fine-tuning and quantization – enables widespread GenAI adoption even outside tech-rich regions.

13.7 SLM Test-Time and Inference-Time Scaling

To further boost SLM performance to rival larger models, researchers are exploiting test-time (inference-time) computation strategies – essentially making the model “think more” without changing its weights. These methods give small models an edge by using extra computation during inference selectively:

Test-Time Ensembling and Self-Consistency: One straightforward approach is to use ensembles or multiple decoding passes at inference and combine the results. For instance, Self-Consistency decoding generates multiple independent answers (via sampling different reasoning paths) and then uses majority voting or a scoring function to pick the best answer. This has been especially effective for math and reasoning problems: a 2.7B model, when asked to produce 10 different reasoning chains and then taking the most agreed-upon answer, can significantly improve accuracy – sometimes matching a 6B model that doesn’t use self-consistency. This is leveraging the idea that while a single pass of an SLM might make an error, the ensemble of its own outputs reduces variance. In general, test-time ensembling (running an SLM N times)

times) scales computation linearly with N , but if N is small (e.g. 5 or 10) it may still be much cheaper than running a $10\times$ larger model once, while yielding similar results. Researchers have noted diminishing returns after a point, but carefully chosen self-consistency can yield large gains in reasoning correctness without any model retraining.

Inference-Time Iterative Refinement: Another powerful idea is allowing the model to iteratively improve its output. Techniques like scratchpad or tree-of-thought prompting let an SLM generate a draft solution, then verify or critique it, and generate a refined solution, possibly repeating this loop. Essentially, the model spends more FLOPs per query by checking its work. A recent paradigm called ReST (Recurrently Self-Improve at Test-time) gives the model a chance to update its answer using a verifier model as feedback. It was found that allocating a fixed extra compute budget to an SLM in this way can outperform a much larger model: in a FLOPs-matched comparison, a smaller base model with test-time compute outperformed a model $14\times$ larger on certain prompts. This underscores that “thinking longer” can substitute for “thinking bigger.” For example, an SLM might use 4 passes to refine an answer (using as much total compute as a big model would in 1 pass) and end up more accurate than the big model. This is highly relevant to low-resource settings: rather than scaling up model size (which increases memory and hardware requirements), one can keep a smaller model but run it a bit longer on available hardware to boost quality when needed.

Auto-Selectable Adapters (Dynamic Routing at Inference): Building on the adapter-based fine-tuning in Section 6, researchers have created systems where multiple adapters (each tuned for a different domain or task) are loaded into the model, and at inference an automated mechanism selects which adapter(s) to apply. One such framework is the Adapters Selector (AS). It trains a lightweight “middleman” network that looks at the input and decides, for example, “This looks like a legal question, activate the Legal adapter”. The selected LoRA adapter is then merged for that forward pass. This yields an on-the-fly multi-skill model without the overhead of a giant model that knows everything. In one experiment, a 770M base model with a suite of domain adapters and an adapter-selector outperformed a 7B model on a mix of domain tasks, using each adapter only when appropriate. Another work, MeteoRA, explicitly aims for

autonomous LoRA selection during inference. These methods essentially scale capacity at inference by having specialized weights ready and choosing the right subset – analogous to having an expert on call. The result is a flexible system that can handle many tasks with SLM-level efficiency, because only a small portion of extra weights (the relevant adapter) is active per query. This is ideal when an edge server has to deal with various use cases: it can host, say, 10 small adapters (which is memory-feasible) covering different languages or functions, and activate them as needed to achieve quality comparable to a larger unified model.

“Think More” Paradigm: A philosophical shift in AI scaling, often termed test-time compute scaling, is emerging: instead of investing all resources in training ever larger models (parameter scaling), invest in mechanisms that allow models to use more computation per query intelligently. For small models, this is a great equalizer. As one article put it, “Train less, think more” – meaning even if you can’t train a gigantic model, you can get a lot of mileage by making a modest model cognitively robust through inference-time strategies (self-reflection, planning steps, verifier loops, etc.). This area is rapidly developing. For instance, self-verification techniques have an SLM generate not just an answer but also a confidence or rationale, which is used to decide if it should try again or defer to another model. All these fall under giving the model additional opportunities at inference to get it right, thereby amplifying its effective performance.

In practice, many of these test-time scaling techniques can be combined with the earlier strategies. One can imagine a small, efficient model (BitNet or Mamba) that also employs speculative decoding, uses self-consistency voting, and has adapters for domain specialization – layering multiple efficiency and quality boosts. The exciting implication is that resource-constrained environments can achieve top-tier AI performance by smart utilization of compute, rather than brute-force scaling. By optimally using every FLOP at inference, whether through ensembling or iterative thinking, even a device with a single GPU or a few CPU cores can produce results on par with far larger systems. This democratizes access to powerful GenAI: instead of needing the latest 100B-parameter model, one can deploy a lean open-source model and rely on these inference hacks to punch above its weight.

Across model design, distributed systems, and clever decoding, the open-source

community has made remarkable progress in making GenAI inference efficient and accessible. Efficient architectures (quantized, hybrid, or recurrence-based) slash the per-token cost. Split and collaborative inference leverage every device and model in concert, rather than depending on a single heavy model. Decentralized and federated schemes enable scaling out with community resources while respecting data boundaries. And importantly, the emphasis on SLMs – training or adapting smaller models – acknowledges the reality that “small is beautiful” when resources are scarce. By further augmenting SLMs with test-time computation tricks and modular adapters, their performance can approach that of unwieldy LLMs at a fraction of the infrastructure demands.

For researchers, policymakers, and developers in Asia, Africa, Latin America, and elsewhere, these developments are empowering. They mean that state-of-the-art language technology no longer strictly requires state-of-the-art hardware – ingenious algorithms and collaboration can overcome the resource gap. As open benchmarks and leaderboards begin to account for efficiency (e.g. “tokens per dollar” competitions), we see a shift in AI towards frugality and inclusion. The survey of techniques here provides a toolkit for building large-scale AI services in low-resource settings: from choosing the right model architecture and size, to distributing inference across clients and servers, to using multiple small models and extra compute to boost accuracy. The overarching trend is clear: efficient GenAI inference is becoming practical everywhere, through open-source innovation that makes “doing more with less” a reality in the realm of AI.

14. Recommendations

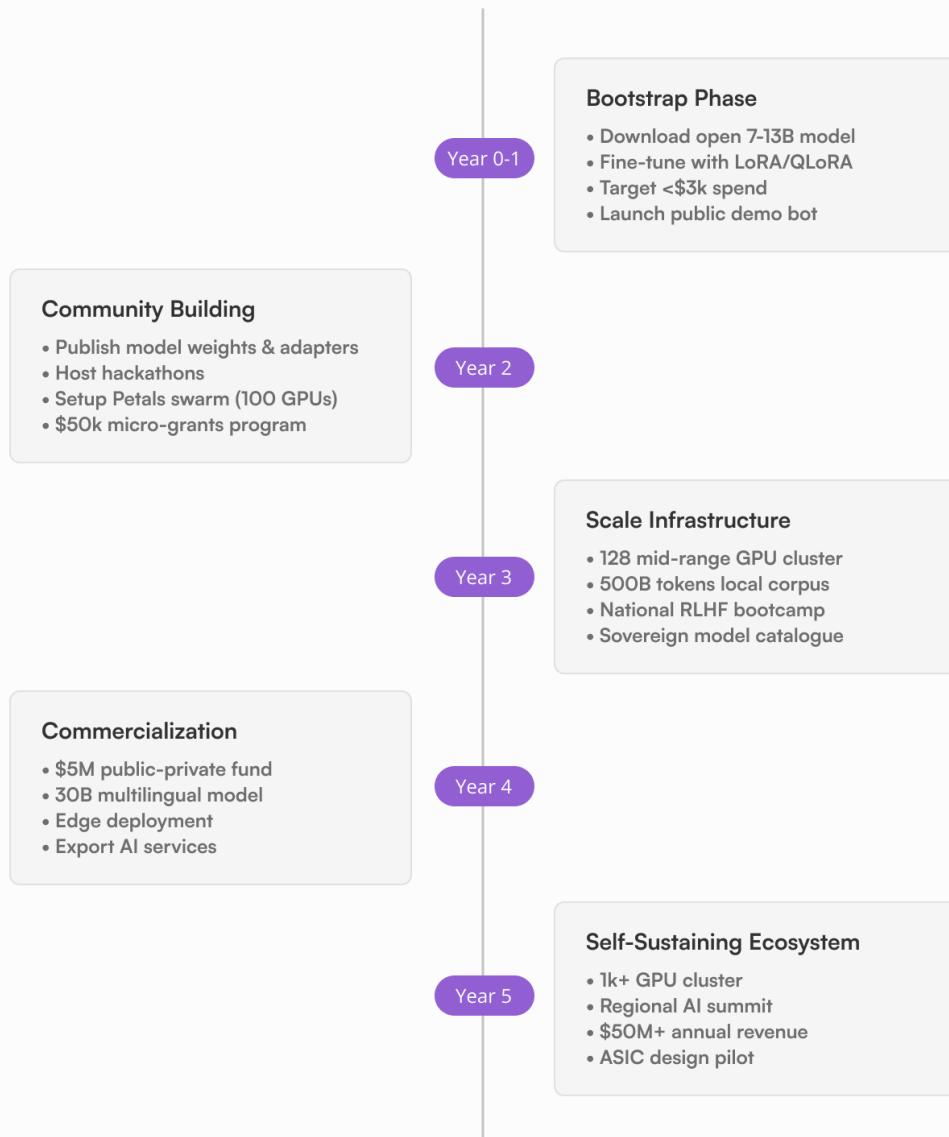
Bringing all these threads together, we provide concrete recommendations for various stakeholders – from national policymakers to enterprise AI leaders – on how to minimize pretraining burdens and enable scalable AI adoption under constrained resources.

14.1 Practical Roadmaps for Different Resource Profiles

For Nations / Large Consortia: If you have moderate infrastructure (e.g. a national supercomputer with a few hundred GPUs), consider a stepwise approach:

1. Start with an open base model that aligns with your size constraints (for instance, a 7B or 13B model if you cannot train 100B). Leverage a model like LLaMA, Qwen, or Deepseek as the foundation – this saves you an order of magnitude in cost.
2. Identify critical domains and languages and gather data (even if small) for those. Prioritize continuous pretraining or fine-tuning on those domains to specialize the model. For example, a country might focus on government documents, local news, and any public bilingual text to improve local language capability.
3. Use parameter-efficient fine-tuning extensively. Rather than one monolithic model trying to do everything, train adapters for specific tasks (e.g. an adapter for healthcare inquiries, one for tourism information). Maintain a repository of these adapters (and their training data provenance) for reuse.
4. Validate via benchmark tests relevant to your use cases (like a local language QA test, or a business process automation test). Iteratively improve by adding data or adjusting training based on where the model falls short.
5. Encourage domestic expertise development: involve local universities in creating and reviewing training datasets and in building human feedback loops to align the model with cultural norms (this could be via a crowd-sourced effort to label or rank model outputs in the local language).

Strategies for Resource-Constrained Environments



Start with the state of the art open base model, build post training and data processing recipes to maximise optimal results and value. Build up the initial momentum through opensource system with the least possible costs to build up a positive feedback loop to be able to exponentially grow value generation, investments, knowledge generation and adoption from local, foreign investors/enterprises/startups in the regional AI initiatives.

Year	North-Star Objective	Core Actions (Tech + Policy + Community)	Expected Outputs & KPIs	Feedback-Loop Levers
0 – 1 “Seed”	Bootstrap one working LLM that speaks local languages & proves ROI	<ul style="list-style-type: none"> Download an open 7 B–13 B base model (e.g. LLaMA-2). Fine-tune with LoRA / QLoRA on a single 48 GB GPU or Apple-Silicon lab server; target < \$3 k spend. (arxiv.org, arxiv.org) Crowd-collect ~20 k high-quality prompts in 2–3 official languages; use Alpaca-style synthetic expansion to 50 k prompts. (medium.com) Launch a public demo bot; require government and two local banks to pilot use-cases. 	<ul style="list-style-type: none"> Local-language chatbot with $\geq 70\%$ helpfulness rating. 5 pilot integrations (justice help-desk, ag-advice SMS, etc.). 20 volunteer GPUs registered for future use. 	<ul style="list-style-type: none"> Pilot savings/revenues fund next GPU purchases; demo success attracts press → talent signup.
2 “Sprout”	Turn pilots into shared assets & seed community	<ul style="list-style-type: none"> Publish model weights + LoRA adapters in a national hub (Git-based). Join Masakhane-style research collaboratives; host 4 hackathons. (arxiv.org) Spin up a Petals swarm; target 100 volunteer GPUs (universities, telcos). (research.yandex.com) Government micro-grants: \$50 k total for 10 student 	<ul style="list-style-type: none"> 30 LoRA adapters online. 100 M inference tokens/month served via swarm. First 3 GenAI startups incorporated. 	<ul style="list-style-type: none"> Open-source adapters → faster prototypes → more demos → more volunteers donating GPUs/data.

Year	North-Star Objective	Core Actions (Tech + Policy + Community)	Expected Outputs & KPIs	Feedback-Loop Levers
		teams building domain adapters (health, fintech, tourism).		
3 “Scale”	Establish local compute & talent flywheel	<ul style="list-style-type: none"> Procure a modest on-prem cluster: 128 mid-range GPUs + 1 PB NVMe. Require cluster time credits be repaid as open artifacts (data cleaned, adapters, eval sets). Launch national RLHF boot-camp (200 annotators part-time; reward-model built for safety & cultural norms). Create a sovereign model catalogue (mirrors Hugging Face; offline access). 	<ul style="list-style-type: none"> 500 B tokens of local corpora cleaned. 10 reward models (safety, tone, factuality). 20 % of university CS grads complete GenAI electives. 	Credits-for-compute trades outputs for wider community use ⇒ compound reuse of data & code.
4 “Harvest”	Commercialize & reinvest	<ul style="list-style-type: none"> Public–private GenAI fund (seed \$5 M) offers matched investments to startups using catalogue assets. Release first 30 B multilingual model continuously pretrained on local data. Push retrieval-augmented micro-services to edge (phones, telco POPs) for low-latency citizen apps. Export AI translation API to neighboring states—begin foreign 	<ul style="list-style-type: none"> 25 funded startups; 1,000 local jobs. 10× growth in inference calls; 40 % served on edge CPUs. Export revenue covers ≥25 % of national GPU OPEX. 	Startup exits & export fees finance cluster expansion; success stories pull diaspora talent home. (wired.com)

Year	North-Star Objective	Core Actions (Tech + Policy + Community)	Expected Outputs & KPIs	Feedback-Loop Levers
		revenue.		
5 “Flywheel”	Self-sustaining, regionally competitive ecosystem	<ul style="list-style-type: none"> • Upgrade cluster to ≥ 1 k GPUs (mix of local OEM, refurbished, heterogeneous).• Host regional GenAI summit; open cross-border Petals mesh (≥ 1 k GPUs).• Standardize open-adapter marketplace with revenue-sharing smart contracts (on-chain accounting).• Incubate ASIC design pilot for 4-bit inference accelerator with university fab partner. 	<ul style="list-style-type: none"> • Nation’s models rank in top-5 on African LLM leaderboard. • Annual AI export revenue $\geq \\$50$ M. • Net positive GPU trade balance (buy partly funded by royalties). 	Marketplace royalties + export cash feed R&D; public success story attracts further volunteers and foreign investment → reinforcing growth loop.

Key Risk-Mitigation Tactics

1. Multi-vendor hardware stack avoids single-supplier shocks; software built on device-agnostic runtimes.
2. National model catalogue mirrors global hubs to guard against external takedowns.
3. Continuous adapter updates prevent catastrophic forgetting while absorbing new data without full retrains.
4. RLHF reward models encode local norms, reducing reliance on foreign moderation tooling.

Sources: This survey is based on a range of recent papers, repositories, and technical reports. Key references include Microsoft’s BitNet work on 1-bit LLMs, the Mamba state-space model paper and Hybrid Mamba evaluations, MIT’s research on Liquid Neural Networks, the FlashMLA inference optimization note, and the NoMAD and ShiftAdd multiplication-free attention papers. For split and collaborative inference, we cited the SplitLLM report, Jupiter system results, and hybrid edge-cloud experiments. Collaborative decoding techniques referenced Google’s speculative decoding blog, as well as the EAGLE and Medusa projects. Distributed inference insights came from the Petals paper and others, while prompt routing and multi-model collaboration were illustrated by the P2L paper and Shen et al.’s collaborative decoding work. Finally, for SLM usage and test-time scaling, we drew on surveys and analyses that highlight the competitive performance of fine-tuned small models and the power of inference-time computation scaling. These and other references provide further technical details and benchmark results supporting the statements in this survey.

Works cited

1. [www.zadara.com,
https://www.zadara.com/glossary/sovereign-ai/#:~:text=Sovereign%20AI%20refers%20to%20the,cultural%20values%2C%20and%20legal%20frameworks](https://www.zadara.com/glossary/sovereign-ai/#:~:text=Sovereign%20AI%20refers%20to%20the,cultural%20values%2C%20and%20legal%20frameworks)
2. Sovereign AI - Zadara,
<https://www.zadara.com/glossary/sovereign-ai/>
3. AI Sovereignty: National Economic Competitiveness and Security - IDC Europe Blog,
<https://blog-idceurope.com/ai-sovereignty-national-economic-competitiveness-and-security/>
4. AI Rivalries: Redefining Global Power Dynamics - TRENDS Research & Advisory,
<https://trendsresearch.org/insight/ai-rivalries-redefining-global-power-dynamics/>
5. April 2025 AI Developments Under the Trump Administration | Inside Government Contracts,
<https://www.insidegovernmentcontracts.com/2025/05/april-2025-ai-developments-under-the-trump-administration/>
6. Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan - Federal Register,
<https://www.federalregister.gov/documents/2025/04/29/2025-07332/request-for-information-on-the-development-of-a-2025-national-artificial-intelligence-ai-research>
7. A Policy Blueprint for US Investment in AI Talent and Infrastructure | Andreessen Horowitz,
<https://a16z.com/a-policy-blueprint-for-us-investment-in-ai-talent-and-infrastructure/>
8. AI Index 2025: State of AI in 10 Charts | Stanford HAI,

- <https://hai.stanford.edu/news/ai-index-2025-state-of-ai-in-10-charts>
9. AI Companies 2025: Who's Winning the Global Race? – AllAboutAI.com,
<https://www.allaboutai.com/resources/ai-statistics/companies/>
 10. Top 10 countries by total AI investment (2025): Where does India rank globally?,
<https://indianexpress.com/article/trending/top-10-listing/top-10-countries-by-total-ai-investment-2025-where-does-india-rank-globally-9962276/>
 11. China's AI Strategy: A Case Study in Innovation and Global Ambition,
<https://trendsresearch.org/insight/chinas-ai-strategy-a-case-study-in-innovation-and-global-ambition/>
 12. China unveils \$8.2b AI fund to strengthen AI industry – Tech in Asia,
<https://www.techinasia.com/news/china-unveils-8-2b-ai-fund-to-strengthen-ai-industry>
 13. EU Invests €200 Billion Towards Becoming a Global Leader in AI | Perkins Coie,
<https://perkinscoie.com/insights/update/eu-invests-eu200-billion-towards-becoming-global-leader-ai>
 14. EU's AI Continent Action Plan: a turning point for digital sovereignty – Telefónica,
<https://www.telefonica.com/en/communication-room/blog/eus-ai-continent-action-plan-turning-point-digital-sovereignty/>
 15. UK's "AI Opportunities Action Plan" – RPC,
<https://www.rpclegal.com/snapshots/technology-digital/spring-2025/uks-ai-opportunities-action-plan/>
 16. The UK's AI Strategy Risks Entrenching the Power of Big Tech | TechPolicy.Press,
<https://www.techpolicy.press/the-uks-ai-strategy-risks-entrenching-the-power-of-big-tech/>
 17. Securing Canada's AI advantage,
<https://www.pm.gc.ca/en/news/news-releases/2024/04/07/securing-canadas-ai>
 18. Canadian Sovereign AI Compute Strategy,
<https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy>
 19. Less regulation, more innovation in Japan's AI governance | East Asia Forum,
<https://eastasiaforum.org/2025/05/21/less-regulation-more-innovation-in-japans-ai-governance/>
 20. National AI Strategy Policy Directions – Press Releases – 과학기술정보통신부 >,
<https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=1040&searchOpt=ALL&searchTxt=>
 21. South Korea bets \$1.3B on AI hardware, critics say talent overlooked,
<https://www.chosun.com/english/industry-en/2025/04/22/6R32I3L5RNHMIQLA7U3MYE3DE/>
 22. India's AI Revolution – PIB,
<https://www.pib.gov.in/PressReleasePage.aspx?PRID=2108810>
 23. Why AI for India 2030 is a blueprint for inclusive growth | World Economic Forum,
<https://www.weforum.org/stories/2025/01/ai-for-india-2030-blueprint-inclusive-growth-global-leadership/>
 24. AI Investment and Business Opportunities in the UAE: Growth, Regulations, Key Sectors,
<https://neweconomyexpert/publications/258009/>
 25. UAE shapes future with pioneering digital infrastructure, AI innovation – ET CIO,
<https://cio.economictimes.indiatimes.com/news/artificial-intelligence/uae-shapes-future-with-pioneering-digital-infrastructure-ai-innovation/120682775>
 26. Saudi Arabia's AI & Tech Ambitions: Investment Opportunities – Middle East Briefing,

- <https://www.middleeastbriefing.com/news/saudi-arabias-ai-and-tech-innovation-drive-opportunities-for-global-investors/>
27. Realigning US-Saudi relations for the AI era | Middle East Institute,
<https://www.mei.edu/publications/realigning-us-saudi-relations-ai-era>
 28. A Strategic Vision for US AI Leadership: Supporting Security, Innovation, Democracy and Global Prosperity | Wilson Center,
<https://www.wilsoncenter.org/article/strategic-vision-us-ai-leadership-supporting-security-innovation-democracy-and-global>
 29. AI to drive 165% increase in data center power demand by 2030 | Goldman Sachs,
<https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030>
 30. AI Data Center Statistics 2025-2024 - AI Stratagems,
<https://aistratagems.com/ai-data-center-statistics/>
 31. AI Infrastructure Investment: The Ultimate Guide for Investors | SmartDev,
<https://smartdev.com/the-rise-of-ai-infrastructure-investment/>
 32. Research and Development | The 2025 AI Index Report | Stanford HAI,
<https://hai.stanford.edu/ai-index/2025-ai-index-report/research-and-development>
 33. The State of Artificial Intelligence in 2025 - Baytech Consulting,
<https://www.baytechconsulting.com/blog/the-state-of-artificial-intelligence-in-2025>
 34. Energy demand from AI - IEA,
<https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>
 35. U.S. Data Centers' Power Demand Surges to 46,000 MW: What's Driving the Growth?,
<https://carboncredits.com/u-s-data-centers-power-demand-surges-to-46000-mw-what-s-driving-the-growth/>
 36. Artificial Intelligence Index Report 2025 - AWS,
https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf
 37. DeepSeek shows the limits of US export controls on AI chips - Brookings Institution,
<https://www.brookings.edu/articles/deepseek-shows-the-limits-of-us-export-controls-on-ai-chips/>
 38. Global Total Number of Scientific Publications in Artificial Intelligence Share by Country (Units (Publications)) - Report Linker,
<https://www.reportlinker.com/dataset/c7a7f8eae968fd302788b2e529a126109612efb>
 39. The Limits of Chip Export Controls in Meeting the China Challenge - CSIS,
<https://www.csis.org/analysis/limits-chip-export-controls-meeting-china-challenge>
 40. Nvidia Makes Mess Afterhours, Discloses \$5.5 Billion in Charges due to US Export Restrictions on its H20 Chip for China | Wolf Street,
<https://wolfstreet.com/2025/04/15/nvidia-makes-mess-afterhours-discloses-5-5-billion-in-charges-due-to-us-export-restrictions-on-its-h20-chip-for-china/>
 41. Overly Stringent Export Controls Chip Away at American AI Leadership | ITIF,
<https://itif.org/publications/2025/05/05/export-controls-chip-away-us-ai-leadership/>
 42. 8 countries that are scaling up AI in their military - Yahoo,
<https://www.yahoo.com/news/8-countries-scaling-ai-military-134200501.html>
 43. Which Countries Are Experimenting With AI-Powered Weapons? - 24/7 Wall St.,
<https://247wallst.com/military/2025/04/16/which-countries-are-experimenting-with-ai-powered-weapons/>
 44. The Authoritarian Risks of AI Surveillance - Lawfare,
<https://www.lawfaremedia.org/article/the-authoritarian-risks-of-ai-surveillance>
 45. The peace of Japan and the AI - Japan Up Close,

- https://japanupclose.web-japan.org/policy/p20250228_1.html
46. How AI can enable public surveillance – Brookings Institution,
<https://www.brookings.edu/articles/how-ai-can-enable-public-surveillance/>
 47. Economists blame Trump tariffs, AI explosion for threatening global economy – The Register,
https://www.theregister.com/2025/05/29/economists_blame_trump_tariffs_ai/
 48. Chief Economists Warn Global Growth Under Strain from Trade Policy Shocks and AI Disruption – The World Economic Forum,
<https://www.weforum.org/press/2025/05/chief-economists-warn-global-growth-under-s-train-from-trade-policy-shocks-and-ai-disruption/>
 49. Role of AI in Developing Countries – International Journal of Scientific Research and Engineering Trends,
https://ijsret.com/wp-content/uploads/2024/09/IJSRET_V10_issue5_375.pdf
 50. Maximizing AI Potential for Economic Growth in Developing ...,
<https://moderndiplomacy.eu/2023/11/20/maximizing-ai-potential-for-economic-growth-in-developing-countries-balancing-innovation-and-protection/>
 51. Why AI for India 2030 is a blueprint for inclusive growth | World Economic Forum,
<https://www.weforum.org/stories/2025/01/ai-for-india-2030-blueprint-inclusive-growth-global-leadership/>
 52. National AI Strategy Policy Directions – Press Releases – 과학기술정보통신부 >,
<https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=1040&searchOpt=ALL&searchTxt=>
 53. (PDF) Artificial intelligence for low income countries – ResearchGate,
https://www.researchgate.net/publication/385248805_Artificial_intelligence_for_low_income_countries
 54. How will AI impact jobs in emerging & developing economies? – VoxDev,
<https://voxdev.org/topic/labour-markets/how-will-ai-impact-jobs-emerging-and-developing-economies>
 55. AI For Industrial Transformation In Developing Countries,
<https://aifod.org/ai-for-industrial-transformation-in-developing-countries/>
 56. Facts and Figures 2024 – Internet use – ITU,
<https://www.itu.int/itu-d/reports/statistics/2024/11/10/ff24-internet-use/>
 57. Two thirds of the world's school-age children have no internet access at home, new UNICEF-ITU report says,
<https://www.unicef.org/eap/press-releases/two-thirds-worlds-school-age-children-have-no-internet-access-home-new-unicef-itu>
 58. AI and the Digital Divide,
<https://www.unaligned.io/p/ai-and-the-digital-divide>
 59. The Hidden Multiplier: Unraveling the True Cost of the Global AI ...,
<https://technologyandsociety.org/the-hidden-multiplier-unraveling-the-true-cost-of-the-global-ai-skills-gap/>
 60. Data in the Clouds, Centers on the Ground: The Role of Data ...,
<https://www.undp.org/latin-america/blog/data-clouds-centers-ground-role-data-centers-lacs-digital-future>
 61. AI and Education: Challenges in Developing Countries – Prime ...,
<https://primeproductionltd.com/ai-in-developing-countries/>
 62. AI Infrastructure Investment: The Ultimate Guide for Investors | SmartDev,
<https://smartdev.com/the-rise-of-ai-infrastructure-investment/>

63. AI Data Center Statistics 2025-2024 – AI Stratagems,
<https://aistratagems.com/ai-data-center-statistics/>
64. Demand for Data Centers Surges in Asia Amid Global AI Boom ...,
<https://www.gtlaw.com/en/insights/2025/4/demand-for-data-centers-surges-in-asia-amid-global-ai-boom>
65. Sovereign AI – Zadara,
<https://www.zadara.com/glossary/sovereign-ai/>
66. How Leaders in the Global South Can Devise AI Regulation That Enables Innovation,
<https://institute.global/insights/tech-and-digitalisation/how-leaders-in-the-global-south-can-devise-ai-regulation-that-enables-innovation>
67. Research and Development | The 2025 AI Index Report | Stanford HAI,
<https://hai.stanford.edu/ai-index/2025-ai-index-report/research-and-development>
68. U.S. Data Centers' Power Demand Surges to 46,000 MW: What's Driving the Growth?,
<https://carboncredits.com/u-s-data-centers-power-demand-surges-to-46000-mw-what-s-driving-the-growth/>
69. Executive summary – Energy and AI – Analysis – IEA,
<https://www.iea.org/reports/energy-and-ai/executive-summary>
70. AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works – News – IEA,
<https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works>
71. African entrepreneurs are redefining how AI can drive sustainable development,
<https://www.undp.org/digital/blog/african-entrepreneurs-are-redefining-how-ai-can-drive-sustainable-development>
72. AI Index 2025: State of AI in 10 Charts | Stanford HAI,
<https://hai.stanford.edu/news/ai-index-2025-state-of-ai-in-10-charts>
73. AI Companies 2025: Who's Winning the Global Race? – AllAboutAI.com,
<https://www.allaboutai.com/resources/ai-statistics/companies/>
74. Artificial Intelligence Index Report 2025 – AWS,
https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf
75. Top 10 countries by total AI investment (2025): Where does India rank globally?,
<https://indianexpress.com/article/trending/top-10-listing/top-10-countries-by-total-ai-investment-2025-where-does-india-rank-globally-9962276/>
76. The State of the Funding Market for AI Companies: A 2024 – 2025 ...,
<https://www.mintz.com/insights-center/viewpoints/2166/2025-03-10-state-funding-market-ai-companies-2024-2025-outlook>
77. The global impact of AI: Mind the gap | CEPR,
<https://cepr.org/voxeu/columns/global-impact-ai-mind-gap>
78. 2024 Africa Tech Venture Capital – Partech,
<https://partechpartners.com/africa-reports/2024-africa-tech-venture-capital-report>
79. Unpacking GenAI Costs: New Challenges And Opportunities In ...,
<https://www.forbes.com/councils/forbestechcouncil/2025/05/27/unpacking-genai-costs-new-challenges-and-opportunities-in-software-monetization/>
80. The State of Artificial Intelligence in 2025 – Baytech Consulting,
<https://www.baytechconsulting.com/blog/the-state-of-artificial-intelligence-in-2025>
81. AI Sovereignty: National Economic Competitiveness and Security – IDC Europe Blog,
<https://blog-idceurope.com/ai-sovereignty-national-economic-competitiveness-and-security/>

82. AI Rivalries: Redefining Global Power Dynamics – TRENDS Research & Advisory,
<https://trendsresearch.org/insight/ai-rivalries-redefining-global-power-dynamics/>
83. Report: US Is Starting to Suffer AI Talent Brain Drain – Tech.co,
<https://tech.co/news/report-us-ai-brain-drain>
84. Exploring the landscape of essential health data science skills and ...,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC11985845/>
85. (PDF) Exploring the landscape of essential health data science skills and research challenges: a survey of stakeholders in Africa, Asia, and Latin America and the Caribbean – ResearchGate,
https://www.researchgate.net/publication/390275338_Exploring_the_landscape_of_essential_health_data_science_skills_and_research_challenges_a_survey_of_stakeholders_in_Africa_Asia_and_Latin_America_and_the_Caribbean
86. apmg-international.com,
<https://apmg-international.com/article/what-digital-skills-gap#:~:text=The%20digital%20skills%20crisis%20is,economies%2C%20societies%2C%20and%20workforces.>
87. AI goes to school: The global AI education race, opportunities and perils – Development Aid,
<https://www.developmentaid.org/news-stream/post/194647/ai-transforming-education>
88. New ITU coalition to bridge AI skills gap for developing countries ...,
<https://developingtelecoms.com/telecom-business/humanitarian-communications/17877-new-itu-coalition-to-bridge-ai-skills-gap-for-developing-countries.html>
89. 60+ Stats On AI Replacing Jobs (2025) – Exploding Topics,
<https://explodingtopics.com/blog/ai-replacing-jobs>
90. AI's impact on jobs may be smaller in developing countries – World Bank Blogs,
<https://blogs.worldbank.org/en/investinpeople/AI-impact-on-jobs-may-be-smaller-in-developing-countries>
91. oxfordinsights.com,
<https://oxfordinsights.com/wp-content/uploads/2024/12/2024-Government-AI-Readiness-Index-2.pdf>
92. Arm AI Readiness Index,
<https://www.arm.com/resources/report/ai-readiness>
93. UAE shapes future with pioneering digital infrastructure, AI innovation – ET CIO,
<https://cio.economictimes.indiatimes.com/news/artificial-intelligence/uae-shapes-future-with-pioneering-digital-infrastructure-ai-innovation/120682775>
94. Data protection and privacy laws now in effect in 144 countries – IAPP,
<https://iapp.org/news/a/data-protection-and-privacy-laws-now-in-effect-in-144-countries>
95. The AI governance balancing act: Navigating opportunities and risks – World Bank Blogs,
<https://blogs.worldbank.org/en/digital-development/the-ai-governance-balancing-act-navigating-opportunities-and-risks>
96. AI Laws: Global Insights & Africa's Regulatory Future | Data Privacy & Innovation,
<https://techcultureafrica.com/ai-laws-africa>
97. Addressing AI Bias and Fairness: Challenges, Implications,
<https://smartdev.com/addressing-ai-bias-and-fairness-challenges-implications-and-strategies-for-ethical-ai/>
98. AI Governance in the Global South: Turning Risk into Relevance – Modern Diplomacy,
<https://moderndiplomacy.eu/2025/05/09/ai-governance-in-the-global-south-turning-risk-into-relevance/>

99. How AI can enable public surveillance – Brookings Institution,
<https://www.brookings.edu/articles/how-ai-can-enable-public-surveillance/>
100. Navigating the Intersection of AI, Surveillance, and Privacy: A Global Perspective – Sustainable Development Goals,
https://sdgs.un.org/sites/default/files/2024-05/Francis_Navigating%20the%20Intersection%20of%20AI%2C%20Surveillance%2C%20and%20Privacy.pdf
101. The Global Expansion of AI Surveillance | Carnegie Endowment for International Peace,
<https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance>
102. Economists blame Trump tariffs, AI explosion for threatening global economy – The Register,
https://www.theregister.com/2025/05/29/economists_blame_trump_tariffs_ai/
103. Chief Economists Warn Global Growth Under Strain from Trade Policy Shocks and AI Disruption – The World Economic Forum,
<https://www.weforum.org/press/2025/05/chief-economists-warn-global-growth-under-strain-from-trade-policy-shocks-and-ai-disruption/>
104. Navigating algorithm bias in AI: ensuring fairness and trust in Africa – ResearchGate,
https://www.researchgate.net/publication/385737925_Navigating_algorithm_bias_in_AI_ensuring_fairness_and_trust_in_Africa
105. Achieving the Potential of AI Across Cultures – Horasis,
<https://horasis.org/achieving-the-potential-of-ai-across-cultures/>
106. DeepSeek shows the limits of US export controls on AI chips – Brookings Institution,
<https://www.brookings.edu/articles/deepseek-shows-the-limits-of-us-export-controls-on-ai-chips/>
107. Overly Stringent Export Controls Chip Away at American AI Leadership | ITIF,
<https://itif.org/publications/2025/05/05/export-controls-chip-away-us-ai-leadership/>
108. Why export limits on GPUs threaten global AI and data center markets – CO/AI,
<https://getcoai.com/news/why-export-limits-on-gpus-threaten-global-ai-and-data-center-markets/>
109. US Imposes Export Restrictions on Advanced Integrated ... – PISM,
<https://pism.pl/publications/us-imposes-export-restrictions-on-advanced-integrated-circuits>
110. Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts | Stanford HAI,
<https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts>
111. Studies explore challenges of AI for low-resource languages – Tech Brew,
<https://www.emergingtechbrew.com/stories/2025/05/05/ai-low-resource-languages>
112. Mind your language: The battle for linguistic diversity in AI – UN News,
<https://news.un.org/en/story/2025/03/1161406>
113. GenAI and the Global South: Language Preservation and Perception Management – The Geostrata,
<https://www.thegeostrata.com/post/genai-and-the-global-south-language-preservation-and-perception-management>
114. Moving toward truly responsible AI development in the global AI market,
<https://www.brookings.edu/articles/moving-toward-truly-responsible-ai-development-in-the-global-ai-market/>
115. Use of AI Technology to Support Data Collection for Project Preparation and

- Implementation: A 'Learning-by-doing' Process,
https://gps.worldbank.org/sites/gps/files/knowledge_products/2021/Use%20of%20AI%20technology%20to%20support%20data%20collection.pdf
116. India's AI Revolution – PIB,
<https://www.pib.gov.in/PressReleasePage.aspx?PRID=2108810>
117. AI Governance Frameworks: Guide to Ethical AI Implementation – Consilien,
<https://consilien.com/news/ai-governance-frameworks-guide-to-ethical-ai-implementation>
118. Securing Canada's AI advantage,
<https://www.pm.gc.ca/en/news/news-releases/2024/04/07/securing-canadas-ai>
119. Canadian Sovereign AI Compute Strategy,
<https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy>
120. A Strategic Vision for US AI Leadership: Supporting Security, Innovation, Democracy and Global Prosperity | Wilson Center,
<https://www.wilsoncenter.org/article/strategic-vision-us-ai-leadership-supporting-security-innovation-democracy-and-global>
121. A Policy Blueprint for US Investment in AI Talent and Infrastructure | Andreessen Horowitz,
<https://a16z.com/a-policy-blueprint-for-us-investment-in-ai-talent-and-infrastructure/>
122. AI Investment and Business Opportunities in the UAE: Growth, Regulations, Key Sectors,
<https://neweconomy.expert/publications/258009/>
123. Data, AI, and Emerging Technologies – World Bank,
<https://www.worldbank.org/en/topic/digital/brief/emerging-technologies>
124. Global Trends in AI Governance – World Bank Documents and Reports,
<https://documents1.worldbank.org/curated/en/099120224205026271/pdf/P1786161ad76ca0a1ba3b1558ca4ff88ba.pdf>
125. The Global Impact of AI: Mind the Gap, WP/25/76, April 2025,
<https://www.imf.org/-/media/Files/Publications/WP/2025/English/wpia2025076-print-pdf.ashx>