Lablup Inc.

The AI Landscape of Early 2025:

Reasoning, Sovereignty, And the Rise of Agentic AI

Jeongkyu Shin, Lablup Inc.

Table of Contents

Lablup Inc.

The AI Landscape of Early 2025: Reasoning, Sovereignty, And the Rise of Agentic AI

Executive Summary

As we reach mid-2025, the artificial intelligence (AI) ecosystem is undergoing a fundamental paradigm shift amid intensifying geopolitical competition. This analysis examines the major transformations from the second half of 2024 through July 2025, aiming to provide guidance for navigating this rapidly evolving landscape.

The most significant change in the first half of this year has been the shift in how AI models improve their performance—from the capital-intensive 'Train-Time Compute' approach to 'Test-Time Compute,' which invests more computation during inference to induce deeper reasoning. This transformation has catalyzed the emergence of 'Reasoning Models' such as OpenAI's o1 and China's DeepSeek R1, creating economic ripple effects that lower the barriers to entry for frontier model development. This became the technical backdrop for the 'DeepSeek Shock' of January 2025, when DeepSeek's claims of achieving frontier performance at remarkably low cost sent shockwaves through global markets. This event elevated the U.S.-China technology competition to a new dimension, demonstrating that Chinese AI capabilities had reached world-class levels in software and algorithmic efficiency despite hardware sanctions.

This competition is spreading into an infrastructure war. Data centers are transitioning to liquid cooling systems to handle power densities exceeding 250kW per rack, with nuclear energy being mobilized to secure stable power supplies. This signifies the emergence of energy sovereignty as a new geopolitical variable directly linked to AI sovereignty. In the AI accelerator market, NVIDIA is consolidating its dominance through integrated 'AI Factory' platforms with its Blackwell Ultra and Rubin roadmap, while AMD challenges with cost, availability, and openness as its weapons through the MI400 series and open ROCm software stack.

From a geopolitical perspective, 'Sovereign AI' has emerged as a core national security agenda. Major nations including the United States, United Kingdom, France, and Japan have announced massive investment plans to control their own AI models, infrastructure, and data, while South Korea aims to leap into the global AI top three with a 100 trillion won investment plan.

Meanwhile, the first half of 2025 saw market differentiation with the release of frontier models specialized in specific domains, including Claude 4, Gemini 2.5 Pro, and Llama 4. Competition for leadership is intensifying particularly in high-value areas such as coding, scientific reasoning, and multimodal capabilities. In the coding AI sector, autonomous 'AI teammates' like GitHub Copilot Agent and Devin are fundamentally changing developers' roles, though a 'productivity paradox' has been observed where actual skilled developers' productivity decreases despite impressive benchmark performance, presenting new challenges.

In conclusion, the 2025 AI ecosystem stands at an inflection point where technological innovation is rapidly reshaping economic and geopolitical landscapes. The universalization of reasoning models, the rise of sovereign AI, the intensification of infrastructure wars, and the proliferation of agentic AI will serve as both threats and opportunities for all enterprises and nations.

Chapter 1

The New Frontier: Reasoning Models and the Test-Time Compute Revolution

1.1. Paradigm Shift: From Train-Time to Test-Time Compute

Until mid-2024, performance improvements in large language models (LLMs) primarily came through what's known as 'Train-Time Compute'—scaling three key elements: model size, dataset size, and training computation.[1] While highly effective, this approach became increasingly untenable as pre-training model scales grew exponentially, with training costs reaching billions of dollars.

This cost barrier triggered a fundamental paradigm shift in AI development. Research labs like OpenAI and DeepMind confirmed that the scaling laws previously applied to training also apply to the inference stage.[1] This means models can "think" more deeply and solve complex problems through 'Test-Time Compute'—investing more computational resources during the inference process when models answer questions. This discovery prompted AI developers to reconsider their resource allocation strategies. Instead of focusing solely on costly pre-training, developers began investing more in inference optimization.[1]

This shift led to the emergence of a new class of models called 'Reasoning Models (RLMs)' or 'Rumination Models.' These models are explicitly trained or prompted to go through step-by-step reasoning processes like 'Chain of Thought' before providing final answers.[2] Validated in benchmarks across fields like mathematics, physics, and chemistry, this principle has brought AI's reasoning capabilities significantly closer to human expert cognition.[1]

1.2. Pioneers of Reasoning: OpenAI's o1 and DeepSeek's R1

The first major model to successfully demonstrate test-time compute's potential was OpenAI's o1, released in September 2024.[2] This model was trained through reinforcement learning (RL) to "think productively," resulting in dramatic performance improvements on complex reasoning benchmarks like math problems and competitive programming. For example, while GPT-4o achieved only 9.3% accuracy on International Mathematical Olympiad qualifying problems, o1 recorded an impressive 74.4%.[4]

However, this enhanced reasoning capability came at a significant cost. The o1 model was nearly 6 times more expensive and 30 times slower than GPT-4o.[4] This established a new trilemma in AI performance evaluation: speed versus cost versus reasoning depth.

In this context, DeepSeek's R1 model, released in January 2025, sent shockwaves through the market.[5] R1 demonstrated reasoning capabilities on par with o1 while offering service at much lower costs through reinforcement learning techniques like GRPO (Group Relative Policy Optimization) and efficiency technologies such as multi-head latent attention.[2] This model's emergence became the key catalyst for the 'DeepSeek Shock' discussed later.

1.3. Frontier Convergence and Narrowing Gaps

Another characteristic of the AI model market in the first half of 2025 was the convergence of top-tier model performance. Based on the Chatbot Arena leaderboard, the Elo score difference between first and tenth place models dropped dramatically from 11.9% in 2023 to 5.4% in early 2025.[4] This suggests frontier competition has become extremely fierce as more developers can produce high-quality models.

The performance gap between open-source and closed-source models has also nearly disappeared. The performance difference in Chatbot Arena, which reached 8.04% in early 2024, narrowed to 1.70% by February 2025.[4] This trend provides powerful AI technology to a broader range of developers, serving as an important force challenging the dominance of large, closed research labs.

1.4. Future Trajectories: Next-Generation Reasoning Architectures

The test-time compute paradigm is shaping the evolution of future AI architectures.

Hybrid Strategies: The future will be dominated by hybrid strategies combining train-time and test-time compute to achieve optimal performance under resource constraints.[1]

Adaptive Inference and API Evolution: Not all questions require deep thinking. Future APIs may introduce 'inference budget' settings where users can adjust reasoning depth and pay accordingly. This could lead to dual-tier AI systems where simple questions are handled by fast, cheap general LLMs while complex tasks are processed by slow, expensive RLMs.[1]

lablup

Agent Search and Advanced RAG: Reasoning models are being integrated with multi-hop retrieval and agent search frameworks like Search-o1 and DeepRAG. This evolution goes beyond simply finding information to dynamically exploring information, refining queries, and generating more sophisticated answers.[2]

Test-Time Training: One of the most cutting-edge research areas involves models fine-tuning parameters in real-time during inference. This breaks down the boundary between training and inference, potentially offering revolutionary flexibility where models dynamically adapt to given tasks even after deployment.[1]

These changes represent economic disruption that fundamentally alters the economics of AI development beyond mere technical evolution. Test-time compute directly attacks the 'capital for training' moat of established leaders like OpenAI. By enabling frontier-level reasoning capabilities without massive initial training costs, it lowers the threshold for high-performance AI development. DeepSeek R1's achievement of o1-level performance with far lower operational costs and training budgets provides clear evidence.[5] This means innovation in reasoning algorithms and efficiency (e.g., DeepSeek's GRPO) has become as important as the ability to secure large-scale training funding, creating direct conditions for powerful competitors to emerge from new regions like China or Europe.

Chapter 2

China's Rise: DeepSeek Shock and China's AI Trajectory

2.1. The January 2025 Tremor: DeepSeek's Market Entry

In January 2025, the global AI market was rocked when DeepSeek, a previously unknown Chinese company, released its reasoning model R1 and foundation model V3. Notably, R1, distributed for free, topped the download charts on the U.S. Apple App Store.[5] DeepSeek claimed its V3 model matched GPT-4's performance and its R1 model equaled OpenAI's o1. What stunned the market was the cost: they claimed the final training run for the V3 model used only $5.6 million and 2,000 Nvidia chips, a stark contrast to the billions of dollars and over 10,000 chips reportedly required for competing American models.[5] This event triggered NVIDIA's largest single-day stock drop and prompted the U.S. government to announce the $500 billion 'Stargate Project' AI investment plan in response.[5]

2.2. Nuanced Analysis: Dissecting the 'DeepSeek Shock'

Deep analysis of the 'DeepSeek Shock' reveals the event carries complex implications beyond simple cost reduction.

The Cost Trap: The $5.6 million figure was an 'apples and oranges' comparison difficult to directly compare with costs announced by American companies. This amount likely represented only the 'final training run' cost, excluding R&D, labor, hardware purchases, and preliminary experiments.[6]

lablup Inc.

Alignment with Industry Trends: AI training costs were already decreasing by half approximately every 8 months due to algorithmic and hardware advances. Given this pace of progress, the final training cost for a GPT-4 level model could be estimated at around $3 million by early 2025. Therefore, DeepSeek's efficiency gains were more an extension of existing industry trends like Moore's Law rather than a revolutionary breakthrough.[6]

The True Disruption DeepSeek's real significance lay not in cost but in accessibility. By releasing powerful open-source models under MIT license, they changed the landscape of model availability and posed a new kind of challenge to the West.[2]

2.3. China's Broader AI Offensive: Zhipu AI and Baidu

DeepSeek's emergence was not an isolated event. It was a signal of the maturation of China's entire AI ecosystem.

Zhipu AI (Z.ai): In April 2025, Zhipu AI released open-source models GLM-4 (foundation) and GLM-Z1 (reasoning) in 32B and 9B parameter sizes.[7] Notably, GLM-Z1-32B model demonstrated reasoning performance competitive with the larger 671B parameter DeepSeek-R1 or GPT-4o, while boasting exceptional inference speed and efficiency at 200 tokens per second.[7] They also showcased technical prowess with 'Rumination', capable of autonomous search and tool use.[8]

Baidu: In March 2025, Baidu released Ernie 4.5 and reasoning model Ernie X1, offering them for free and claiming to surpass GPT-4.5 on some benchmarks.[10] Baidu pursues a full-stack strategy integrating Ernie into its search, cloud (Qianfan), and autonomous driving (Apollo) platforms.[11]

In June 2025, they announced plans to open-source their next-generation model Ernie 5, further fueling competition.[12]

As a result of these efforts, by the end of 2024, the performance gap between leading U.S. and Chinese models on major benchmarks like MMLU, MATH, and HumanEval had dramatically narrowed from double-digit percentage points to low single digits.[4]

Table 2.1: China's AI Offensive - Frontier Model Comparison (H1 2025)

Model	Developer	Release Date	Parameter Size (Reported)	Key Innovation	Performance Claims	Cost/Efficiency Claims
R1	DeepSeek	Jan 2025	671B	GRPO, test-time compute	On par with o1	Much cheaper than o1
V3	DeepSeek	Jan 2025	Undisclosed	High-efficiency training	On par with GPT-4	$5.6M training cost
GLM-Z1-32B	Zhipu AI	Apr 2025	32B	Fast inference, rumination	Competitive with DeepSeek-R1	1/30 cost of R1
Ernie X1	Baidu	Mar 2025	Undisclosed	Reasoning, ecosystem integration	On par with DeepSeek R1	Half price of R1
o1	OpenAI	Sep 2024	Undisclosed	Test-time compute	Dominates GPT-4o	High cost, low speed
GPT-4o	OpenAI	May 2024	Undisclosed	Multimodal	SOTA (at time)	Baseline model

2.4. Strategic and Geopolitical Implications

The 'DeepSeek Shock' created deep ripples across investment, policy, and security domains.

Shift in Investor Perspective: Investors began reassessing Chinese technology companies like Alibaba and Baidu, paying attention not only to 'enablers' like NVIDIA in the AI value chain but also to 'providers' and 'beneficiaries.'[5]

China's Policy Crossroads: DeepSeek's success restored China's technological confidence after its 'catch-up' period (late 2022-early 2025). However, this confidence collides with a sluggish domestic economy, creating policy dilemmas between promoting growth and strengthening ideological control.[14]

lablup Inc.

Open-Source Security Dilemma: The emergence of high-performance Chinese open-source models presents significant security challenges to the Western world. If these models are widely adopted globally, they could influence global AI infrastructure through hidden censorship biases or vulnerabilities.[6]

The 'DeepSeek Shock' was not merely a single company's cost innovation but an indicator of the maturation of China's entire AI ecosystem. It demonstrated that Chinese companies have mastered not only model scaling but also algorithmic efficiency and strategic open-source utilization despite U.S. chip sanctions. The revelation that America's core containment strategy of restricting hardware access was incomplete, actually spurring Chinese software and algorithmic innovation. The new geopolitical battlefield is shifting from who creates the best models to who controls the foundational open-source stack. If Chinese models become the default foundation for startups and researchers worldwide based on performance and accessibility, China will secure immense soft power and embed its standards and values in the AI technology substrate for the next decade.

Chapter 3

The AI Infrastructure War: Power, Cooling, and Geopolitics

3.1. The Data Center Revolution: Redesigning for AI

AI workloads are fundamentally transforming data center design from the ground up. Traditional data centers cannot handle the extreme computational density that AI demands.

Ultra-High Density: AI workloads have pushed power density per rack beyond 250kW, far exceeding what traditional air-cooling systems can handle.[15]

Transition to Liquid Cooling: To manage this tremendous heat, operators are rapidly adopting liquid cooling and immersion cooling technologies. These are no longer niche solutions but mainstream technical requirements for high-density AI clusters.[15]

Edge Computing Proliferation: Demand for real-time AI decision-making is moving computational power from centralized mega-facilities to the edge. Micro data centers are being built in urban centers, manufacturing sites, and next to renewable energy facilities to reduce latency.[15]

3.2. The Energy Procurement War: Insatiable Power Demands

AI's explosive growth is placing unprecedented strain on power infrastructure.

Exponential Demand Growth: Global data center power demand is projected to more than double by 2030, exceeding Japan's entire power consumption.[16] This is creating enormous burden on national power grids.

The Nuclear Option: To secure stable, low-carbon power, tech giants are turning to nuclear energy. Amazon is building a data center next to a nuclear power plant in Pennsylvania, while Microsoft and Meta have signed nuclear power purchase agreements.[16] Small Modular Reactors (SMRs) are emerging as key elements of long-term energy strategies.[15]

Sustainability Mandates and Regulatory Pressure: Concerns about grid and resource depletion are increasing pushback from governments and communities. Europe is introducing stricter environmental regulations, and some U.S. regions have temporarily halted new data center development.[15] Sustainability and transparent reporting are now becoming 'licenses to grow.'

3.3. From Backroom to Battlefield: The Geopolitics of Infrastructure

Data centers are no longer mere technical facilities but core assets in geopolitical competition.

Concentration of Power: Approximately 51% of global data centers are concentrated in the United States. This demonstrates America's digital dominance while revealing how dependent other nations are on U.S.-based cloud infrastructure.[17]

The 'Digital Cold War': U.S.-China strategic competition has intensified as both view AI leadership as a core element of national power. This has led to rising trade barriers and technology decoupling, even spawning the term 'digital iron curtain.'[17]

Hardware Bottleneck Strategy: Through mid-2025, the U.S. intensified its campaign to block China's access to cutting-edge AI hardware. By banning even specially designed chips compliant with previous export regulations, it sent a clear message of maintaining hardware superiority at all costs.[17]

The AI infrastructure competition is creating a new global power hierarchy where energy sovereignty becomes a direct prerequisite for digital and AI sovereignty. Nations with stable, abundant, low-carbon energy (e.g., France's nuclear power grid[18]) gain significant strategic advantages in attracting AI investment and building domestic capabilities. As AI training and inference consume enormous energy[16] and national grids experience bottlenecks,[15] data center operators are seeking large-scale direct power sources like nuclear energy.[16] Therefore, nations that can combine and provide energy with favorable regulations, like France creating AI Growth Zones based on nuclear power and SMRs,[18] will become key hubs for the physical infrastructure underpinning the AI economy. The global AI power map will increasingly be drawn not just by the location of tech companies but by energy policy and infrastructure.

Chapter 4

The Silicon Engine: Accelerators, NPUs, and the Journey Toward Independence

4.1. The Data Center Revolution: Redesigning for AI

The AI hardware market continues its explosive growth. Expected to grow from $66.8 billion in 2025 to $296.3 billion by 2034, with a CAGR of 18%.[20] GPUs still maintain a dominant position in this market, accounting for approximately 39% share as of 2024.[20]

Lablup Inc.

However, AI accelerator performance is increasingly constrained by memory bandwidth. This has caused demand for High Bandwidth Memory (HBM) to skyrocket, accounting for 47% of the AI memory market in 2024. Companies like Micron are accelerating technology development to address this bottleneck, releasing next-generation HBM4 chips in July 2025.[20]

4.2. The NPU Expansion: AI Heading to the Edge

Beyond data centers, AI is rapidly spreading into personal devices, led by Neural Processing Units.

Market Growth: The embedded AI NPU market is estimated at $15 billion in 2025, with an expected annual growth rate of 25%.[21] (Other reports estimate the broader NPU market will reach $2.4 billion by 2031, showing variance based on market definition[22]).

Key Drivers: The main growth drivers are low-latency, on-device processing capabilities required by edge computing, Internet of Things (IoT), smartphones, and automotive (ADAS, autonomous driving).[21]

Industry Applications: Mobile devices consume over 200 million NPUs annually, while the automotive industry consumes over 100 million.[21] Intel's Core Ultra and Qualcomm's Snapdragon chips already include dedicated NPUs, and BMW plans to integrate DeepSeek's AI solutions (NPU-based) in vehicles produced in China by the end of 2025.[22]

4.3. The End-to-End Platform War: NVIDIA vs AMD

The frontline of the AI chip war has evolved beyond individual chip performance competition to platform competition encompassing entire ecosystems.

NVIDIA's Dominance Strategy

Roadmap: NVIDIA maintains a rapid one-year cycle roadmap. Blackwell Ultra GPU, an enhanced version of Blackwell, is scheduled for release in the second half of 2025, offering 50% more HBM3e memory (288GB) and FP4 performance.[23] Next-generation architecture Rubin is scheduled for 2026, with Rubin Ultra in 2027.[24]

Full-Stack Ecosystem: NVIDIA's powerful moat combines hardware (DGX/HGX systems like GB300 NVL72), proprietary high-speed networking (NVLink, CX-8 NICs), and the dominant CUDA software ecosystem into an end-to-end 'AI Factory' platform.[26] They sell fully integrated rack-scale solutions, reducing customer deployment time and complexity.[23]

AMD's Strategic Challenge

Counter-Positioning: AMD targets key customer pain points against NVIDIA's near-monopoly position: cost, availability, and openness.[26]

Hardware: The MI400 series GPUs are positioned as cost-effective alternatives, aiming to capture 15-20% of the AI GPU market by the end of 2025.[29]

Lablup Inc.

Software and Ecosystem: AMD is investing heavily in the open-source ROCm software stack (version 7 in 2025) to break CUDA dependency, aiming for immediate support of major frameworks.[30] Like NVIDIA, AMD now offers fully integrated rack-scale solutions combining CPU, GPU, and networking.[26]

Table 4.1: AI Accelerator Competition: NVIDIA vs AMD (H1 2025)

Category	NVIDIA (GB300 NVL72 w/ Blackwell Ultra)	AMD (MI400-based Rack Systems)
GPU Architecture	Blackwell Ultra	CDNA 4 (MI400 Series)
Memory (Capacity/Type)	288GB HBM3e per GPU	192GB HBM3 per GPU
Key Performance Metric	1.1 ExaFLOPS at FP4 (rack-level)	1.5 TFLOPS (chip-level, FP16 est.)
Interconnect Technology	NVLink 5.0, CX-8 InfiniBand	Infinity Fabric
Software Stack	CUDA, cuDNN, TensorRT	ROCm 7, PyTorch/TF support
System-Level Offering	Fully integrated DGX/HGX rack-scale systems	Fully integrated rack-scale systems
Market Approach	High performance, full-stack, proprietary ecosystem (AI Factory)	Price competitive, available, open ecosystem

4.4. The Hyperscaler Gambit: The Rocky Road to Silicon Independence

Hyperscalers like Google (TPU), Amazon (Trainium/Inferentia), and Microsoft (Maia) are investing heavily in developing their own AI accelerators to reduce dependence on NVIDIA and optimize hardware for their specific workloads.[20] Amazon is investing $100 billion in its 'Project Rainier' data center centered around its Trainium 2 chips.[20]

However, designing custom AI chips is an extremely complex and challenging task. In June 2025, Microsoft had to delay production of its 'Maia' AI chip by six months, a representative case showing significant development and scalability issues.[20]

The AI hardware market is dividing into two distinct strategic battlegrounds. The first is the 'Merchant Market', where NVIDIA and AMD compete to supply end-to-end rack-scale systems to enterprises and smaller cloud providers. The second is the 'Captive Market', a costly and difficult competition where hyperscalers build their own vertically integrated infrastructure. NVIDIA's true long-term competitor may not be AMD but the success of its largest customers' own silicon development. NVIDIA's strategy of shortening its roadmap to one-year cycles and deepening software and networking moats is a direct defense against this captive market threat.

Chapter 5

The Rise of Sovereign AI: Age of Competition, National Strategies

5.1. The New Global Agenda: From Cooperation to Competition

In 2025, the center of gravity in global AI policy discourse has dramatically shifted from abstract discussions about existential risks to a competitive 'Global AI Race' where nations prioritize domestic technology leadership, secure data centers, and national adoption mandates.[31] This stems from fears of depending on foreign technology for critical infrastructure.[17] This trend has crystallized into the concept of 'Sovereign AI'—policies ensuring nations control their own AI models, infrastructure, and data to align with national interests.[19]

Lablup Inc.

5.2. The New Global Agenda: From Cooperation to Competition

Major nations are jumping into the sovereign AI competition with their own visions and strategies.

United Kingdom

Vision: Transform the UK from an 'AI adopter' to an 'AI maker.'[19]

Initiatives: Announced the 'Compute Roadmap' in July 2025 and 'AI Opportunities Action Plan' in January. Key components include 20x expansion of AI Research Resources (AIRR), establishment of a £500 million Sovereign AI Unit, and creation of AI Growth Zones to attract private data center investment.[19]

United States

Vision: "Remove barriers to American AI leadership."[31]

Initiatives: In early 2025, announced the 'AI Action Plan' to expedite data center construction permits.[16] In response to the 'DeepSeek Shock,' President Trump announced the $500 billion 'Stargate Project' AI investment.[5]

France

Vision: Become a global 'AI powerhouse' and build a 'third way' alliance countering the U.S. and China.[34]

Initiatives: Phase 3 national strategy starting in 2025 includes €2.5 billion investment and €109 billion international funding initiative.[18] Key focus is supporting national champion Mistral AI and building sovereign computing infrastructure including a 500,000 GPU supercomputer partnership powered by nuclear energy.[18]

Japan

Vision: Make Japan "the world's most AI-friendly nation" through a 'light-touch' regulatory approach.[37]

Initiatives: Enacted the 'AI Promotion Act' in May 2025. Instead of strict rules or penalties, the law focuses on establishing a National AI Strategy Headquarters, promoting voluntary cooperation, and providing public support (funding, infrastructure access) to reverse sluggish private AI investment.[38] The FY2025 budget allocated approximately ¥196.9 billion for AI-related activities.[41]

Table 5.1: Sovereign AI Strategies - Global Comparison (2025)

Country	Key Policy/Initiative	Vision	Total Investment (Announced)	Regulatory Approach	Strategic Model
United States	AI Action Plan, Stargate Project	Maintain AI leadership	$500B	Innovation promotion, infrastructure acceleration	Fortress
United Kingdom	Compute Roadmap, AI Growth Zones	AI Maker nation	£1B+	Pro-business, deregulation	Lighthouse
France	National AI Strategy Phase 3, Mistral support	AI powerhouse, third way	€109B (incl. int'l funding)	Sovereignty enhancement, ethical framework	Alliance
Japan	AI Promotion Act	Most AI-friendly nation	¥196.9B (FY2025)	Light-touch, voluntary cooperation	Lighthouse
South Korea	National AI Strategy	Global AI Top 3	₩100T (5-year)	ecosystem building	Hybrid

Lablup Inc.

5.3. Deep Dive: South Korea's National AI Strategy

South Korea is pursuing an ambitious national strategy to enter the leading group in global AI competition.

Vision: The goal is to elevate South Korea to one of the world's top three AI economic powers.[42]

Investment Scale: The strategy's core is a massive investment plan totaling 100 trillion won (approximately $72 billion) over five years.

Funding Sources: This plan is based on public-private partnerships, mobilizing 30 trillion won from the national budget, 5 trillion won from local governments, and 65 trillion won from the private sector.[42]

Objectives: Building on previous foundations, this plan aims to create a stable and organized national ecosystem for AI development with much greater scale and consistency.[42] The 2025 implementation plan for national strategic technologies including AI allocated 6.8 trillion won for that year.[43]

Sovereign AI strategies present a fundamental philosophical divergence. The U.S. and China's 'Fortress' model focuses on massive state-led investment and technological dominance / containment. France's 'Alliance' model seeks to create a counterbalance through coalitions with like-minded nations. Japan and the UK's 'Lighthouse' model focuses on creating attractive, business-friendly environments to draw global talent and private investment.

Each strategy faces the 'Sovereign Trilemma'—the inability to simultaneously achieve:
(1) complete technological autonomy
(2) access to global innovation and talent
(3) maximum economic growth.

The 'Fortress' model may sacrifice (2) and (3) for (1), while the 'Lighthouse' model is open to (2) for (3) while sacrificing (1). South Korea's massive public-private investment model can be interpreted as an attempt to aggressively pursue all three simultaneously.

Chapter 6

Innovators: Startup Approaches and Big Tech Power Dynamics

6.1. The Vertically Integrated Challenger: Mistral AI's Market Strategy

In July 2025, France's Mistral AI announced a significant strategic shift beyond being a simple model developer to becoming a 'vertically integrated AI cloud provider.'[44] This ambitious plan is backed by up to $1 billion in equity investment from investors including Abu Dhabi's MGX, and hundreds of millions of euros in debt financing to build the sovereign AI cloud service 'Mistral Compute' in France.[44] This extends the partnership to build Europe's largest data center campus with NVIDIA and MGX.[36]

This move positions Mistral not as a mere startup but as a key pillar of France and Europe's AI sovereignty strategy. French President Emmanuel Macron praised this as a "game changer" for technological independence.[44] Mistral offers enterprise, agent-enabled, privacy-first solutions, emphasizing flexibility to deploy anywhere—on-premises, cloud, or edge.[45]

6.2. The Infrastructure Specialist: Lablup's AI Foundation Technology Market

Seoul-based Lablup, a Series A startup founded in 2015, has raised $9.64 million to date.[46] Rather than developing large models directly, Lablup focuses on solving critical problems in AI infrastructure.

Core Product: Lablup's flagship product is 'Backend.AI,' an open-source platform that manages and scales AI workloads across multiple GPU clusters and clouds.[48]

Key Innovation: Lablup's core technical competitiveness lies in 'container-level GPU virtualization' and a vertically integrated platform from OS to end-user interface. Lablup's platform uses AI technology to virtualize GPU resources and efficiently manage network and storage usage, enabling multiple container-based sessions to efficiently share GPUs and allowing numerous users and services to train AI or serve models with minimal contention. This is a critical technology that reduces the waste of expensive AI hardware resources, a chronic problem in MLOps (machine learning operations), and is patent-protected in the United States, Japan, and South Korea.[47]

6.3. The Giants' Stage: Big Tech Competition Landscape

Big Tech AI competition in 2025 is unfolding with each company maximizing its strengths.

Microsoft: Focuses on the enterprise market centered around the 'Copilot' brand. The strategy integrates AI assistants into all aspects of work and life across Office, Windows, GitHub, based on Azure OpenAI Service and OpenAI models like o3.[49] At Build 2025, they introduced Copilot Studio and Azure AI Foundry enabling enterprises to build their own agents.[49]

Google: Under an 'AI-first' vision, integrates its powerful in-house Gemini models across its vast consumer ecosystem including Search and Android, as well as enterprise services like Workspace and Google Cloud Vertex AI.[49] Google pursues an open ecosystem, leading protocols like A2A (Agent2Agent).[49]

Meta: Continues to lead the open-source path with the release of Llama 4 in April 2025. This positions them as a key alternative to closed models from OpenAI or Anthropic.[51]

Anthropic: Focuses on safety and enterprise readiness. The Claude 4 family released in May 2025 leads in coding and complex reasoning, deployed through major cloud platforms like AWS Bedrock and Google Cloud Vertex AI for enterprises prioritizing reliability and controllability.[51]

6.4. The Uneasy Truce: The Paradox of Safety Cooperation

Despite fierce competition, July 2025 saw an unusual move when researchers from competing labs including OpenAI, Anthropic, Meta, and Google jointly published a paper on the importance of 'Chain of Thought Monitorability' for AI safety.[3]

They warned that while observing models' step-by-step reasoning processes (CoT) is an important tool for detecting deception or harmful intent, paradoxically, the very process of optimizing model performance could lead models to hide these 'bad thoughts' or shorten reasoning processes, reducing transparency and making control more difficult.[3] This suggests a direct tension between improving AI capabilities and ensuring safety.

The AI startup landscape has entered a phase of maturation and specialization. The 'build a better LLM' race has become unaffordable for most startups. Successful new entrants are choosing to either (A) become vertically integrated sovereign champions with massive state and private support (Mistral), or (B) become hyper-specialized tool providers solving critical, painful niche problems in the MLOps stack (Lablup). Mistral builds an entire stack combining models with cloud services instead of just selling models, securing massive funding by aligning with EU sovereignty as a national strategic goal.[36] Meanwhile, Lablup solves the inefficient GPU usage problem caused by the model boom instead of making models. Their fractional GPU technology provides direct value to every company struggling with high AI infrastructure costs.[47] The most promising strategies for AI startups in 2025 are to 'go big' by becoming national champions or 'go deep' by solving technical problems created by giants. The middle ground is disappearing.

Lablup Inc

Chapter 7

The H1 2025 Frontier Model Landscape

7.1. A New Generation of Intelligence: Major Model Releases

The first half of 2025 saw successive releases of frontier models rapidly pushing the technological frontline. xAI (February), Meta (April), Anthropic (May), Google (June), and xAI again (July) announced major model families, forming an intensely competitive landscape.[51]

Major Models

Grok 3 (xAI, February 2025) & Grok 4 (xAI, July 2025): Specialized in reasoning and knowledge utilizing real-time data from the X platform, offering features like 'Think Mode' for deep reasoning.[51]

Llama 4 (Meta, April 2025): An important evolution in open-source models, offering native multimodal capabilities and mixture-of-experts (MoE) architecture in various sizes (Scout, Maverick, Behemoth).[51]

Claude 4 (Anthropic, May 2025): A model family consisting of Opus 4 and Sonnet 4, establishing itself as an industry leader in coding and complex software engineering tasks.[51]

Gemini 2.5 Pro (Google, June 2025): A multimodal powerhouse with a massive 1 million token context window and advanced reasoning capabilities through 'Deep Think' mode.[50]

OpenAI Lineup: While not releasing a completely new model named 'GPT-5,' maintained high competitiveness by continuously advancing the GPT-4 family through GPT-4o and reasoning-specialized o1/o3 models.[4]

Lablup Inc.

7.2. Performance and Specialization: Winners by Domain

The frontier model market has shifted beyond general performance competition to specialization competition for dominating specific high-value domains.

Coding and Software Engineering: Claude 4 emerged as the leader in this field, achieving a record 72.7% on the challenging SWE-bench Verified benchmark.[51] Gemini 2.5 Pro also showed strong performance, leading WebDev Arena and scoring high on code editing benchmarks.[52]

Mathematics and Scientific Reasoning: Grok 3/4 and Gemini 2.5 Pro stood out. Grok 3 scored 93.3% on AIME 2025, Gemini 2.5 Pro recorded impressive scores of 86.7% and 84% on USAMO 2025 benchmarks.[51] Claude 4 Opus also showed high performance with 90% on AIME 2025.[51]

Multimodal (Image, Audio, Video): Gemini 2.5 Pro was rated the most versatile model with native support for text, image, audio, and video inputs, leading benchmarks like VideoMME with 84.8%.[51] GPT-4o was praised as the most natural and flexible in voice interactions thanks to end-to-end multimodal training.[52]

Long Context: Gemini 2.5 Pro solidified its leadership in this field with a 1 million token context window, capable of processing an entire book or large codebase at once.[52] Grok 3 and Llama 4 Scout also offer massive context windows of 1 million and 10 million tokens respectively.[51]

Table 7.1: Frontier Foundation Model Benchmark Comparison (July 2025)

Model	Developer	Release Date (H1 2025)	Key Specialization	SWE-bench Verified (%)	AIME 2025 (%)	GPQA (%)	Multimodal (VideoMME %)	Context Window (Tokens)
Claude Opus 4	Anthropic	May	Coding, Enterprise	72.5	90.0	83-84	-	200K
Grok 4	xAI	July	Real-time knowledge, Reasoning	79.4 (LiveCodeBench)	93.3	84.6	-	1M
Gemini 2.5 Pro	Google	June	Multimodal, Long context	63.8	86.7	84.0	84.8	1M
Llama 4 Maverick	Meta	April	Open-source, Multimodal	Competitive with GPT-4o	-	-	Excellent	1M
GPT-4o / o3	OpenAI	-	General purpose, Natural interaction	54.6 (GPT-4.1)	90.5 (MGSM)	-	Excellent	128K

7.3. Open-Source vs Closed-Source Status

As mentioned in Chapter 1, by early 2025, the performance gap between top open-source models like Llama 4 and closed-source models like GPT-4o or Claude 4 had nearly disappeared on many benchmarks.[4] As Meta continues to lead the open-source camp with Llama 4 and Mistral provides powerful open models, closed-source providers like OpenAI and Anthropic face pressure to differentiate through specialized features, enterprise-level security, and ease of use through platforms.

The 2025 frontier model market is de-commodifying through specialization. While 2023-2024 competition was about ranking first on general leaderboards like MMLU, 2025 competition has shifted to dominating specific high-value vertical markets like coding (Anthropic), scientific reasoning (xAI/Google), and multimodal (Google). This is leading sophisticated enterprises to adopt 'multi-model strategies', no longer relying on single providers but using Claude for development teams, Gemini for marketing teams, and Grok for R&D teams. This increases demand for orchestration layers that can route tasks to the most suitable specialized models.

Lablup Inc.

Chapter 8

The Proliferation of Coding AI: From Assistant to Autonomous Teammate

8.1. A New Generation of Intelligence: Major Model Releases

The paradigm of AI coding tools has dramatically shifted beyond simple assistance to autonomous execution. GitHub Copilot, first released in 2021, was a 'smart assistant' providing code auto-completion, with a 'human-in-the-loop' approach where developers controlled every step.[53]

GitHub Copilot Agent, unveiled at Microsoft Build 2025, introduced the new concept of an 'AI Teammate.' When developers assign a GitHub issue, the agent autonomously plans, codes, and tests, then submits a pull request for human review. This represents a shift to 'human-on-the-loop' where developers serve as final reviewers and decision-makers, part of Microsoft's broader 'Agentic DevOps' strategy.[53]

Meanwhile, Devin from startup Cognition AI, introduced in early 2024, branded itself as the first 'AI Software Engineer,' demonstrating abilities to handle complex end-to-end tasks, learn unfamiliar technologies, and even take on real work from platforms like Upwork.[58]

8.2. The Productivity Paradox: Benchmarks vs Reality

Agent models have achieved remarkable success on benchmarks. While 2023 models solved about 5% of SWE-Bench tasks, top models like Anthropic's Claude 4 in mid-2025 solved over 70%.[4] Devin also showed a 13.86% solve rate on SWE-bench, significantly exceeding the previous record of 1.96%.[58]

However, reality contrasts sharply with these benchmark performances. A groundbreaking randomized controlled trial (RCT) by METR published in July 2025 found that skilled open-source developers using latest AI tools like Claude 3.7 actually took 19% longer to complete real work tasks.[61]

The most surprising finding was the disconnect between developer perception and reality. Developers believed AI improved their speed by 20%, when in fact their speed had decreased.[61] Researchers explain this paradox by noting that benchmarks only evaluate self-contained tasks, underestimating real-world complexity. Mature real codebases require implicit context and high quality standards including documentation, style, and test coverage, which AI struggled to follow, requiring developers to spend significant time reviewing and fixing AI-generated code.[61]

8.3. Enterprise Adoption and Strategic Implications

Despite mixed productivity results, enterprise adoption of AI agents is accelerating. According to a Cloudera survey, 96% of IT leaders plan to increase AI agent usage within the next year, with coding assistants at 62% being one of the most popular use cases.[64]

Financial giants like Goldman Sachs have already piloted Devin, with plans to deploy hundreds to thousands of instances.[65]

The generative AI coding assistant market is expected to grow from approximately $26 million in 2024 to about $98 million by 2030, while the broader 'AI code tools' market including platforms and services already reaches billions of dollars.[60]

However, successful adoption requires more than purchasing licenses. Companies must establish clear governance policies, strengthen code review processes specifically for AI-generated code, manage new security risks, and provide comprehensive developer training in new skills like advanced prompt engineering.[60]

The current generation of 'AI Software Engineers' is not replacing developers but changing the nature of development work and creating new management and operational challenges. The core value lies not in simple speed improvements but in reducing cognitive load for specific tasks, requiring fundamental changes in team structures and workflows. ROI for companies like Goldman Sachs will come not from firing developers but from successfully restructuring engineering organizations into human-agent teams where humans serve as architects and quality controllers for teams of AI 'junior developers.' This is a significant execution challenge requiring substantial investment in new processes, governance, and training.

Lablup Inc.

References

[1] C. C. Chieh, "Understanding Reasoning Models & Test-Time Compute: Insights from DeepSeek R1," Medium, 2024. [Online]. Available: https://medium.com/@cch.chichieh/understanding-reasoning-models-test-time-compute-insights-from-deepseek-r1-d30783070827

[2] Zeta Alpha, "Trends in AI February 2025: Reasoning Models," 2025. [Online]. Available: https://www.zeta-alpha.com/post/trends-in-ai-february-2025-reasoning-models

[3] R. Rajkumar, "Researchers from OpenAI, Anthropic, Meta, and Google issue joint AI safety warning - here's why," ZDNET, 2025. [Online]. Available: https://www.zdnet.com/article/researchers-from-openai-anthropic-meta-and-google-issue-joint-ai-safety-warning-heres-why/

[4] Stanford Institute for Human-Centered Artificial Intelligence, "The AI Index 2025 Annual Report: Technical Performance," 2025. [Online]. Available: https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance

[5] M&G Investments, "China's DeepSeek shakes up the global AI race," 2025. [Online]. Available: https://www.mandg.com/investments/private-investor/en-gb/in-the-spotlight/m-g-insights/2025/06/chinas-deepseek-shakes-up-the-global-ai-race

[6] RAND Corporation, "What DeepSeek Really Changes About AI Competition," 2025. [Online]. Available: https://www.rand.org/pubs/commentary/2025/02/what-deepseek-really-changes-about-ai-competition.html

[7] Z.ai, "Z.ai Unveils New GLM Open-Source Models with World-Class Reasoning Performance," PR Newswire, 2025. [Online]. Available: https://www.prnewswire.com/news-releases/zai-unveils-new-glm-open-source-models-with-world-class-reasoning-performance-302429306.html

[8] Synced, "Zhipu.AI's Open-Source Power Play: Blazing-Fast GLM Models & Global Expansion Ahead of Potential IPO," 2025. [Online]. Available: https://syncedreview.com/2025/04/16/zhipu-ais-open-source-power-play-blazing-fast-glm-models-global-expansion-ahead-of-potential-ipo/

[9] AIbase, "Zhipu Releases New Generation Open-Source GLM Model: 32B Parameters, Rivaling DeepSeek R1 with 8x Faster Speed," 2025. [Online]. Available: https://news.aibase.com/en/news/17156

[10] Wikipedia, "Ernie Bot," 2025. [Online]. Available: https://en.wikipedia.org/wiki/Ernie_Bot

[11] Zacks Investment Research, "Baidu's AI Push Gains Momentum: Is ERNIE Enough to Power Ambitions?," Nasdaq, 2025. [Online]. Available: https://www.nasdaq.com/articles/baidus-ai-push-gains-momentum-ernie-enough-power-ambitions

[12] A. Garg, "After DeepSeek, China Baidu to open source its Ernie AI chatbot," India Today, 2025. [Online]. Available: https://www.indiatoday.in/technology/news/story/after-deepseek-china-baidu-to-open-source-its-ernie-ai-chatbot-2748296-2025-06-30

[13] Digital Watch Observatory, "Baidu to launch Ernie 5 AI in 2025," 2025. [Online]. Available: https://dig.watch/updates/baidu-to-launch-ernie-5-ai-in-2025

[14] Carnegie Endowment for International Peace, "China's AI Policy in the DeepSeek Era," 2025. [Online]. Available: https://carnegieendowment.org/research/2025/07/chinas-ai-policy-in-the-deepseek-era?lang=en

[15] JLL, "The AI-driven data center revolution: Why 2025 is a defining year," 2025. [Online]. Available: https://spark.jllt.com/resources/blog/the-ai-driven-data-center-revolution-why-2025-is-a-defining-year

[16] A. St. John, "Trump's plan to boost AI and build data centers could strain the nation's power grid," Associated Press, 2025. [Online]. Available: https://apnews.com/article/trump-artificial-intelligence-energy-data-centers-f216660b80f992ae303b348dac0b2f87

[17] World Economic Forum, "AI geopolitics, data centres and the new technological rivalry," 2025. [Online]. Available: https://www.weforum.org/stories/2025/07/ai-geopolitics-data-centres-technological-rivalry/

Lablup Inc.

[18] Chambers and Partners, "Artificial Intelligence 2025: France Trends and Developments," 2025. [Online]. Available: https://practiceguides.chambers.com/practice-guides/artificial-intelligence-2025/france/trends-and-developments

[19] Department for Science, Innovation and Technology, "Engines of AI primed to accelerate new breakthroughs, economic growth, and transform the UK into an AI maker," 2025. [Online]. Available: https://www.gov.uk/government/news/engines-of-ai-primed-to-accelerate-new-breakthroughs-economic-growth-and-transform-the-uk-into-an-ai-maker

[20] Global Market Insights, "AI Hardware Market Size, Share & Trends Analysis Report," 2025. [Online]. Available: https://www.gminsights.com/industry-analysis/ai-hardware-market

[21] Market Report Analytics, "Embedded AI Neural Processing Unit (NPU) Market," 2025. [Online]. Available: https://www.marketreportanalytics.com/reports/embedded-ai-npu-382260

[22] IndustryARC, "Neural Processing Units (NPUs) Market," 2025. [Online]. Available: https://www.industryarc.com/PressRelease/4540/Neural-Processing-Units-(NPUs)-Market

[23] CRN, "10 Big Nvidia GTC 2025 Announcements: Blackwell Ultra, Rubin Ultra, DGX Spark And More," 2025. [Online]. Available: https://www.crn.com/news/ai/2025/10-big-nvidia-gtc-2025-announcements-blackwell-ultra-rubin-ultra-dgx-spark-and-more

[24] N. Owens, "Nvidia details AI roadmap with new chips, robots and more," Manufacturing Dive, 2025. [Online]. Available: https://www.manufacturingdive.com/news/nvidia-details-ai-roadmap-Rubin-Blackwell-Ultra-chips-robots/742902/

[25] eWeek, "NVIDIA Shows More AI Infrastructure at GTC 2025," 2025. [Online]. Available: https://www.eweek.com/news/nvidia-gtc-2025-ai-infrastructure-blackwell-ultra-vera-rubin-ai/

[26] Datacenters.com, "AMD's Full-Rack AI Systems: Challenging Nvidia's Dominance in 2025," 2025. [Online]. Available: https://www.datacenters.com/news/amd-s-full-rack-ai-systems-challenging-nvidia-s-dominance-in-2025

[27] SemiAnalysis, "Nvidia GTC 2025: Built For Reasoning - Vera Rubin, Kyber, CPO, Dynamo Inference, Jensen Math & Feynman," 2025. [Online]. Available: https://semianalysis.com/2025/03/19/nvidia-gtc-2025-built-for-reasoning-vera-rubin-kyber-cpo-dynamo-inference-jensen-math-feynman/

[28] A. Greengart, "GTC 2025: Nvidia's AI Lead is Secure," Techsponential, 2025. [Online]. Available: https://www.techsponential.com/reports/gtc2025

[29] AInvest, "Unlocking AI Potential: AMD & NVIDIA's Valuation Edge in the 2025 Tech Landscape," 2025. [Online]. Available: https://www.ainvest.com/news/unlocking-ai-potential-amd-nvidia-valuation-edge-2025-tech-landscape-2506/

[30] YouTube User, "AMD's AI Platform Strategy 2025," 2025. [Online]. Available: https://www.youtube.com/watch?v=4iZPPypUa9w

[31] Atlantic Council, "Navigating the new reality of international AI policy," 2025. [Online]. Available: https://www.atlanticcouncil.org/blogs/geotech-cues/navigating-the-new-reality-of-international-ai-policy/

[32] Beaumont Capital Markets, "France's AI Strategy: Sovereignty, Ethics, & Industry Growth," 2025. [Online]. Available: https://beaumont-capitalmarkets.co.uk/frances-ai-strategy-sovereignty-ethics-industry-growth/

[33] Technology Law Dispatch, "The UK government's AI growth agenda in 2025," 2025. [Online]. Available: https://www.technologylawdispatch.com/2025/04/artificial-intelligence/the-uk-governments-ai-growth-agenda-in-2025/

[34] P. Chavez, "France Pursues an AI 'Third Way'," CEPA, 2025. [Online]. Available: https://cepa.org/article/france-pursues-an-ai-third-way/

[35] Présidence de la République, "Make France an AI Powerhouse," 2025. [Online]. Available: https://www.elysee.fr/admin/upload/default/0001/17/d9c1462e7337d353f918aac7d654b896b77c5349.pdf

[36] Freshfields Transactions, "Inside Infrastructure: Focus on France," 2025. [Online]. Available: https://transactions.freshfields.com/post/102kuzs/inside-infrastructure-focus-on-france

[37] East Asia Forum, "Less regulation, more innovation in Japan's AI governance," 2025. [Online]. Available: https://eastasiaforum.org/2025/05/21/less-regulation-more-innovation-in-japans-ai-governance/

Lablup Inc.

[38] Future of Privacy Forum, "Understanding Japan's AI Promotion Act: An Innovation-First Blueprint for AI Regulation," 2025. [Online]. Available: https://fpf.org/blog/understanding-japans-ai-promotion-act-an-innovation-first-blueprint-for-ai-regulation/

[39] DDG, "Japan's 2025 AI Promotion Act: Structuring Innovation Through Soft Regulation," 2025. [Online]. Available: https://www.ddg.fr/actualite/japans-2025-ai-promotion-act-structuring-innovation-through-soft-regulation

[40] Baker McKenzie, "Japan's AI Bill Advances Toward Enactment," 2025. [Online]. Available: https://connectontech.bakermckenzie.com/japans-ai-bill-advances-toward-enactment/

[41] Chambers and Partners, "Artificial Intelligence 2025: Japan," 2025. [Online]. Available: https://practiceguides.chambers.com/practice-guides/artificial-intelligence-2025/japan

[42] Jipyong LLC, "South Korea's National AI Strategy," 2025. [Online]. Available: https://www.jipyong.com/en/board/newsletters_main.php?seq=2118

[43] Ministry of Science and ICT, Republic of Korea, "2025년 국가전략기술 육성 시행계획," 2025. [Online]. Available: https://nsp.nanet.go.kr/plan/subject/detail.do?nationalPlanControlNo=PLAN0000051325

[44] PYMNTS, "Report: Mistral Aims to Raise Financing for AI Cloud Service," 2025. [Online]. Available: https://www.pymnts.com/artificial-intelligence-2/2025/report-mistral-aims-to-raise-financing-for-ai-cloud-service

[45] Mistral AI, "Mistral AI Homepage," 2025. [Online]. Available: https://mistral.ai/

[46] Tracxn, "Lablup Company Profile," 2025. [Online]. Available: https://tracxn.com/d/companies/lablup/__X6TyRocL56l9E4tLliF0KgF5hx2qCo7yiINM49QleD4

[47] PitchBook, "Lablup Company Profile," 2025. [Online]. Available: https://pitchbook.com/profiles/company/181762-93

[48] CommunicAsia, "Lablup Inc. - Exhibitor 2025," 2025. [Online]. Available: https://asiatechxsg.com/communicasia/sponsors/lablup-inc-exhibitor-2025/

[49] R. G. Infante, "Generative AI Showdown 2025: Microsoft vs. Google vs. Amazon," Medium, 2025. [Online]. Available: https://medium.com/@roberto.g.infante/generative-ai-showdown-2025-microsoft-vs-google-vs-amazon-6060841f291c

[50] M. Frąckiewicz, "Google Gemini's 2025 Takeover: How Google's AI Powerhouse Stacks Up," TS2 Tech, 2025. [Online]. Available: https://ts2.tech/en/google-geminis-2025-takeover-how-googles-ai-powerhouse-stacks-up-against-gpt-4-claude-more/

[51] Collabnix, "Comparing Top AI Models in 2025: Claude, Grok, GPT, Llama, Gemini, and DeepSeek," 2025. [Online]. Available: https://collabnix.com/comparing-top-ai-models-in-2025-claude-grok-gpt-llama-gemini-and-deepseek-the-ultimate-guide/

[52] FelloAI, "We Tested Grok 4, Claude, Gemini, GPT-4o: Which AI Should You Use In July 2025?," 2025. [Online]. Available: https://felloai.com/2025/07/we-tested-grok-4-claude-gemini-gpt-4o-which-ai-should-you-use-in-july-2025/

[53] SCB Tech X, "GitHub Copilot 2025: Not Just an 'Assistant,' but the 'AI Teammate' Revolutionizing the Industry," 2025. [Online]. Available: https://scbtechx.io/blogs/technology/github-copilot-2025-not-just-an-assistant-but-an-ai-teammate-revolutionizing-devops/

[54] N-School, "Evolution of GitHub Copilot to Copilot X," 2025. [Online]. Available: https://www.n-school.com/evolution-of-github-copilot-to-copilot-x/

[55] GitHub, "GitHub Introduces Coding Agent For GitHub Copilot," 2025. [Online]. Available: https://github.com/newsroom/press-releases/coding-agent-for-github-copilot

[56] Thurrott.com, "Build 2025: Big Updates for GitHub Copilot, Open Source Implementation in Visual Studio Code," 2025. [Online]. Available: https://www.thurrott.com/a-i/github-copilot/321127/build-2025-big-updates-for-github-copilot-open-source-implementation-in-visual-studio-code

[57] Microsoft Azure, "Agentic DevOps: Evolving software development with GitHub Copilot and Microsoft Azure," 2025. [Online]. Available: https://azure.microsoft.com/en-us/blog/agentic-devops-evolving-software-development-with-github-copilot-and-microsoft-azure/

Lablup Inc.

[58] The Cognition Team, "Introducing Devin, the first AI software engineer," Cognition, 2024. [Online]. Available: https://cognition.ai/blog/introducing-devin

[59] Aegissofttech, "Unveiling the Danger: Could Devin AI Spell the End for Software Developers?," 2025. [Online]. Available: https://www.aegissofttech.com/insights/devin-ai/

[60] P. Ajith, "AI Code Assistants – Comprehensive Guide for Enterprise Adoption," 2025. [Online]. Available: https://ajithp.com/2025/06/23/ai-code-assistants-enterprise-adoption-guide/

[61] J. Becker, N. Rush, E. Barnes, and D. Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," METR, 2025. [Online]. Available: https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

[62] N. Kobie, "Think AI coding tools are speeding up work? Think again – they're actually slowing developers down," ITPro, 2025. [Online]. Available: https://www.itpro.com/software/development/think-ai-coding-tools-are-speeding-up-work-think-again-theyre-actually-slowing-developers-down

[63] arXiv, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," 2025. [Online]. Available: https://arxiv.org/abs/2507.09089

[64] S. Christian, "2025 Is the Year to Adopt AI Agents: Cloudera Survey," CDO Magazine, 2025. [Online]. Available: https://www.cdomagazine.tech/branded-content/2025-is-the-year-to-adopt-ai-agents-cloudera-survey

[65] The HR Digest, "Goldman Sachs Just Hired AI Engineer Devin to Kickstart Its Hybrid Workforce," 2025. [Online]. Available: https://www.thehrdigest.com/goldman-sachs-just-hired-ai-engineer-devin-to-kickstart-its-hybrid-workforce/

[66] ResearchAndMarkets.com, "Generative Artificial Intelligence Coding Assistants Strategic Research Report 2025," Business Wire, 2025. [Online]. Available: https://www.businesswire.com/news/home/20250319490646/en/Generative-Artificial-Intelligence-Coding-Assistants-Strategic-Research-Report-2025-Market-to-Reach-%2497.9-Billion-by-2030-at-a-CAGR-of-24.8-Driven-by-Growing-Adoption-of-Low--and-No-Code-Platforms---ResearchAndMarkets.com

[67] GetDX, "AI code generation: Best practices for enterprise adoption in 2025," 2025. [Online]. Available: https://getdx.com/blog/ai-code-enterprise-adoption/

Lablup Inc.

Disclaimer

The content of this article is provided for informational purposes only and represents the personal views of the author in their individual capacity. It does not represent or reflect the views, policies, or official position of the company, nor does it constitute any form of professional, business, or investment advice. While reasonable efforts have been made to reference publicly available sources, the information contained herein may include opinions, interpretations, or data that have not been independently verified. The company makes no representations or warranties, express or implied, as to the accuracy, completeness, or reliability of the information provided. Any reliance placed upon this content is strictly at the reader's own risk, and the company disclaims all liability for any direct, indirect, or consequential loss or damage arising from its use.