THE FTI DELTA PERSPECTIVE

October 2024

Saudi Arabia's Bold AI Vision

A leap towards AI independence

FTI DELTA

Saudi Arabia's Bold AI Vision

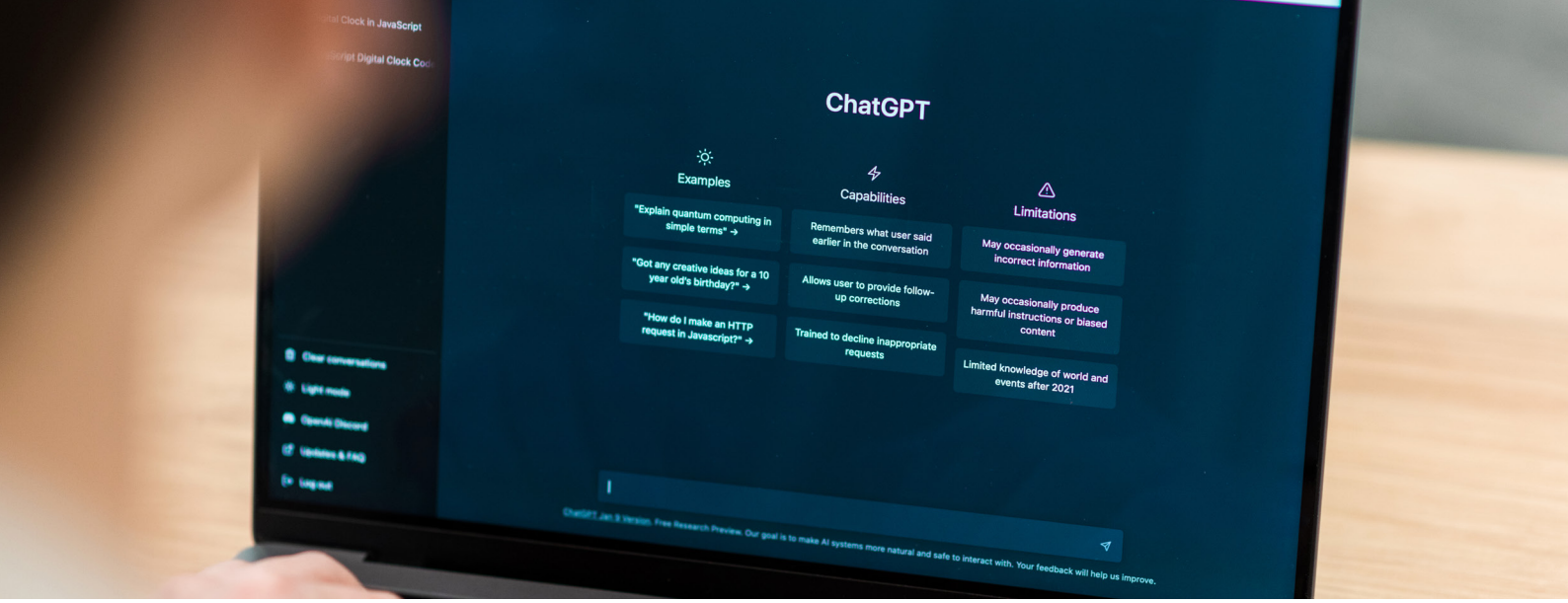
A leap towards AI independence

Authors:
Ahmed Abdulla Associate Partner
Javier Álvarez Senior Managing Director

Saudi Arabia is rapidly transforming its economy, with artificial intelligence (AI) as a key driver of its diversification strategy. The Kingdom is positioning itself as a global leader in AI by focusing on developing cutting-edge technologies, building a skilled workforce and fostering a collaborative innovation ecosystem.

To support these ambitions, Saudi Arabia is making significant investments in AI and related infrastructure, including a $40 billion tech fund and targeted investments in AI companies and startups.

However, achieving AI independence remains a significant challenge for the Kingdom. This is due to a combination of factors, including the rapidly evolving AI chip and technology landscapes. Additionally, the complexities of global supply-chain constraints, regulations and export restrictions pose further hurdles. To maintain its leading position in the AI landscape, Saudi Arabia must proactively address these challenges.



Saudi Arabia's economic landscape is undergoing a seismic shift, with AI set to be a key catalyst in its ambitious diversification plans. The Kingdom's AI ambitions reflect a comprehensive strategy to position the country at the forefront of technological innovation in several key ways. The Kingdom aims to establish itself as the world's premier hub for AI innovation, developing advanced technologies to tackle global challenges. For example, Saudi Arabia has an ambition to make its locally developed Large Language Model (LLM), i.e., the Arabic Large Language Model (ALLaM), to be the best large Arabic language model in the world. [1] Simultaneously, it seeks to transform its workforce by nurturing a steady stream of AI expertise across various industries. To foster widespread AI adoption, the country is cultivating a cutting-edge and collaborative AI ecosystem. Finally, Saudi Arabia plans to implement foundational legislation to attract AI businesses and talent, further solidifying its position as a global AI leader. [2]

These ambitions are driving substantial transformative investments, from sector-wide initiatives down to targeted investments in specialist companies. For example, the government announced plans to launch a $40 billion tech fund focusing on AI, with an emphasis on cloud computing and digital infrastructure. [3] At a more targeted level, Saudi's Public Investment Fund (PIF)-backed SCAI (Saudi Company for Artificial Intelligence) announced plans to invest $776 million in a joint venture with SenseTime to develop AI across the Kingdom. [4] Additionally, the government announced plans to invest $1 billion in GAIA, its GenAI startup accelerator program. [5] Furthermore, PIF-backed Alat announced plans to invest $200 million alongside Dahua Technology for AI-powered, vision-centric solutions in intelligent cities. [6] Finally, Kingdom Holding participated in a $6 billion funding round for the GenAI company xAI. [7]

ChatGPT took the world by surprise, but we are yet to see a wide adoption of Gen AI applications at the enterprise and government level

The adoption of Gen AI is impacted by several factors, including global politics, funding, regulations and the availability of technology enablers (e.g., talent, data, open-source models, compute capacity). Gen AI holds the promise of revolutionary benefits for enterprises, and its future adoption by businesses and governments is poised to trigger an unprecedented surge in computational demand. However, the technology is in its nascency, and we are yet to witness its widespread implementation across enterprise and government sectors. [8, 9]

The future of Gen AI is set to bring about significant advancements in functionality. First, agentic workflows–business workflows that use AI agents in parts of the steps–are expected to enable chaining of agents to deliver an outcome. [10] Second, multimodal systems will enable the processing of diverse types of input, including combinations of audio, text and video, enabling more versatile and context-aware generative capabilities. [11] Third, efforts are underway to expand the scope of AI models, aiming to provide expert advice across broader areas of human knowledge.

The aforementioned evolution coupled with increased adoption will substantially raise computing demand. [12] In order to support these advanced systems across various industries and governmental functions, a nationwide AI independence strategy will be needed.

AI independence is key to maximizing the benefits of Gen AI for Saudi Arabia

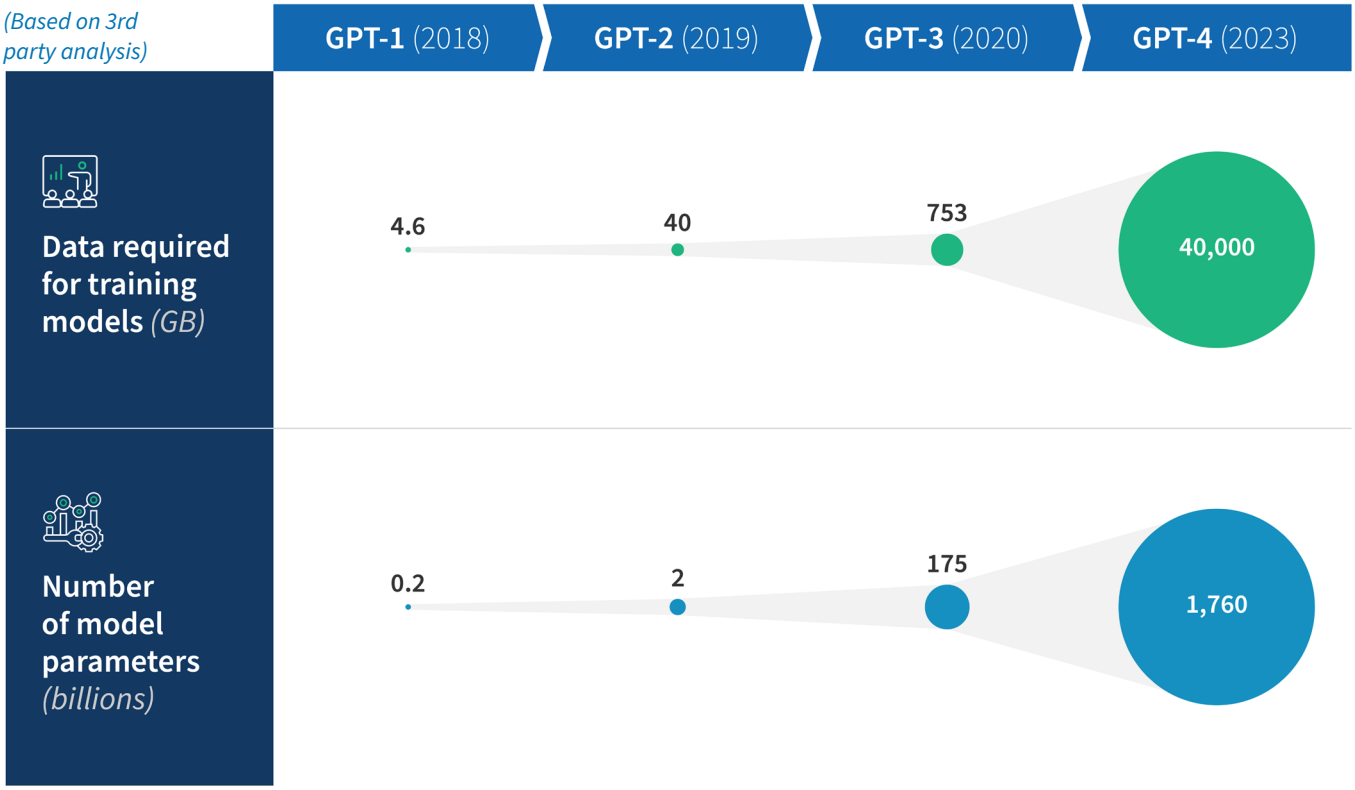
Consumers are readily adopting Gen AI, as they are accustomed to providing their data for access to basic services from the likes of Facebook to Google. [13] On the other hand, for governments and enterprises it is key that data, models, chips and Data Centers (DCs) are localized to secure benefits from Gen AI models and AI independence.

There are several key elements needed to secure AI independence: (1) control over quality local data to ensure differentiation, given the availability of open-source models; (2) model fine-tuning using local data is essential for enterprises and governments to fully leverage AI capabilities, as generic models often fall short in addressing specific use cases; (3) a secure supply of advanced AI chips will ensure the steady stream of computational capacity necessary to develop advanced models and maintain AI independence; (4) the localization of next-generation DCs with new layouts and advanced cooling technology will be key to meeting the increasing computational requirements.

AI infrastructure in Saudi Arabia and the implications for Gen AI

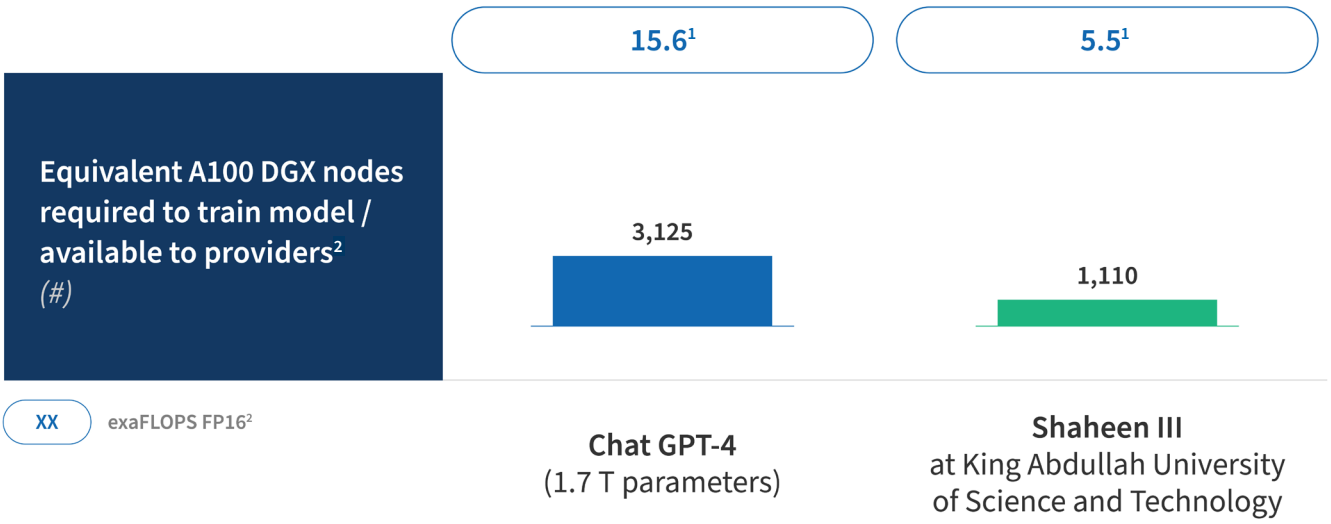
A shortage of AI chips has the potential to bottleneck Gen AI development in the Kingdom. A real-world example can be seen in the latest Gen AI models. The number of parameters contained in leading Gen AI models continues to grow; e.g., GPT-4 grew ~10x versus its previous release (1,760 billion vs. 175 billion in ChatGPT-3). [14, 15]

Training models of this scale requires a significant amount of computational power, and it is rumored that GPT-4 was trained on 25,000 A100 chips (equivalent to 3,125 A100 DGX nodes). [16] These computational requirements pose a challenge for current computing infrastructure in the Kingdom; King Abdullah University of Science and Technology's (KAUST) Shaheen III supercomputer, for example, has the equivalent of 1,110 A100 DGX nodes. [17, 18][1] This gap is likely to widen with the advent of GPT-5 and the next generation of Gen AI models. The Kingdom is actively addressing this challenge and is optimistic about gaining access to additional high-performance AI chips from the United States in the coming year. [19]



Sources: What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher (Thompson, 2022 [dataset updated 2024]); The Memo - GPT-4 (Alan D. Thompson, 2024); FTI Delta analysis

[1] Equivalent to 36% of GPT-4's training nodes.



Notes: [1] Floating points operations per second at half-precision floating-point format. Chat GPT-4 15.6 exaFLOPS based on 25,000 A100 chips producing 624 teraFLOPS FP16 each. Shaheen III 5.5 exaFLOPS FP16 is based on 2,800 Nvidia GH chips producing 1,979 teraFLOPS FP16 each [17]. Other sources have indicated Shaheen III has 3,000 Nvidia GH chips [18], and we have selected the lower bound for prudence.
[2] Chat GPT-4 assumes 25,000 A100 chips to train the model and assumes 8 A100 chips per A100 DGX node. Shaheen III assumes the equivalent number of A100 chips to produce 5.5 exaFLOPS FP16 and assumes 8 A100 chips per A100 DGX node. Each A100 chip is assumed to produce 624 teraFLOPS FP16.

Sources: Top 500, Nvidia, "The Secrets of GPT-4 Leaked?" (Martin Treiber, 2023); "Saudi University KAUST taps HPE Cray for 100 petaflops Shaheen III supercomputer" (Dan Swinhoe, 2022); FTI Delta analysis. [2,3]

The advancement in chips is also changing DC requirements towards high-power density coupled with advanced cooling technology. This shift may present a challenge to the Kingdom, as many of the new DC announcements are focused on hyperscalers and may not be able to directly address the need for high-power density DCs.

As Gen AI is in its nascency, the solution for AI independence is rapidly evolving

Leading AI chip providers cannot provide a silver bullet for computing needs. Varying levels of export restrictions have been applied to advanced chips by the U.S. government, limiting their supply to only certain parts of the world. [20] The significant demand for advanced AI chips, coupled with supply chain constraints, is resulting in long lead times and delaying solution development. [21] To address the supply chain challenge, KSA is planning to localize 50 semiconductor design companies by 2030. [22]

There are multiple startups offering AI-specific chips to compete with the market leaders, and it is not clear what the best long-term solutions are. [23] Notable players include Graphcore, which offers AI DC-ready compute pods [24] and has recently become a wholly owned subsidiary of SoftBank Group [25]; Groq and its proprietary Groq Language Processing Unit (LPU) chipset offer exceptionally fast AI inference [26, 27]; Samba Nova's proprietary chips offer enterprise-grade AI as well as access to other open-source models [28, 29]; and finally, Cerebras's Wafer-Scale solution provides cluster-scale performance on a single chip. [30] The Kingdom is collaborating with these leading AI chip players, e.g., SambaNova Systems [31], Cerebras [32] and Groq [33], to ensure access to their latest technologies.

The landscape of open-source Gen AI models continues to evolve: for example, Meta is leading the charge to develop open-source ecosystems [34], with multiple models and frameworks now available. [35]

The AI regulatory landscape needs to evolve as fast as the technology to ensure there is a concrete regulatory framework enabling AI. Saudi Arabia has taken the first step to develop AI principles and now can move towards developing clear policies, governance mechanisms, standards and controls for AI. [36]

The continued evolution of AI chipsets is driving the need for a new generation of DCs capable of hosting

high-power density equipment. These DCs will need to support rapid and scalable AI deployments aligned to the latest model developments. These complex technical requirements exist alongside the need to minimize the social and environmental impact of the next-generation facilities. [37]

Ensuring AI independence includes exploring alternative chipset architectures, leveraging open-source designs and developing local supply chains. These strategies collectively enhance technological capabilities, foster innovation and reduce external dependencies. By implementing this comprehensive approach, the Kingdom can strengthen its position in the rapidly evolving AI landscape while promoting self-reliance and security.

References

[1] Asharq Al-Awsat, "SDAIA Concludes GAIN Summit in Riyadh with Local, Int'l Agreements," 13 September 2024. [Online]. Available: https://english.aawsat.com/business/5060516-sdaia-concludes-gain-summit-riyadh-local-int%E2%80%99l-agreements

[2] SDAIA, "National Strategy for Data & AI," 1 January 2020. [Online]. Available: https://ai.sa/

[3] M. Farrell and R. Copeland, "Saudi Arabia Plans $40 Billion Push Into Artificial Intelligence," *The New York Times*, 19 March 2024. [Online]. Available: https://www.nytimes.com/2024/03/19/business/saudi-arabia-investment-artificial-intelligence.html

[4] "Saudi PIF's SCAI to invest $776m to boost AI in Kingdom | 1," *Arab News*, 13 September 2022. [Online]. Available: https://www.arabnews.com/node/2161476/business-economy

[5] C. Malin, "Saudi backs GAIA with $1 billion," 04 March 2024. [Online]. Available: https://www.middleeastainews.com/p/saudi-backs-gaia-with-1-billion

[6] Alat, "Alat and Dahua partner to establish global AIoT vision-centric products and solutions business in the Kingdom of Saudi Arabia," 20 February 2024. [Online]. Available: https://alat.com/en/newsroom/strategic-partnership-between-alat-and-dahua/

[7] xAI, "Series B Funding Round," 26 May 2024. [Online]. Available: https://x.ai/blog/series-b

[8] AI Infrastructure Alliance, "Enterprise Generative AI Adoption," 1 August 2023. [Online]. Available: https://ai-infrastructure.org/enterprise-generative-ai-adoption-report-aug-2023/

[9] ABI Research, "Whitepaper | Generative AI in the Enterprise Sector," 1 August 2023. [Online]. Available: https://go.abiresearch.com/lp-generative-ai-in-the-enterprise-sector

[10] A. Ng, "Agentic Design Patterns Part 2: Reflection," The Batch, 27 March 2024. [Online]. Available: https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-2-reflection/

[11] Meta, "Multimodal generative AI systems," 12 December 2023. [Online]. Available: https://ai.meta.com/tools/system-cards/multimodal-generative-ai-systems/

[12] GPUnet, "Generative AI fuels comute demand: Urgency for sustainable solutions," 6 May 2024. [Online]. Available: https://medium.com/@GPUnet/generative-ai-fuels-compute-demand-urgency-for-sustainable-solutions-49b9b3464c80

[13] Adobe, "The Age of Generative AI: Over half of Americans have used generative AI and most believe it will help them be more creative," 22 April 2024. [Online]. Available: https://blog.adobe.com/en/publish/2024/04/22/age-generative-ai-over-half-americans-have-used-generative-ai-most-believe-will-help-them-be-more-creative

[14] A. D. Thompson, "What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher," LifeArchitect.ai, 2022. [Online]. Available: https://lifearchitect.ai/whats-in-my-ai/

[15] A. D. Thompson, "GPT-4 - Alan D. Thompson - Life Architect," 1 March 2023. [Online]. Available: https://lifearchitect.ai/gpt-4/

[16] M. Treiber, "The Secrets of GPT-4 Leaked?," 17 July 2023. [Online]. Available: https://www.ikangai.com/the-secrets-of-gpt-4-leaked/

[17] D. Swinhoe, "Saudi University KAUST taps HPE Cray for 100 petaflops Shaheen III supercomputer," 26 September 2022. [Online]. Available: https://www.datacenterdynamics.com/en/news/saudi-university-kaust-taps-hpe-cray-for-100-petaflops-shaheen-iii-supercomputer/

[18] M. Murgia, A. England, Q. Liu and E. Olcott, "Saudi Arabia and UAE race to buy Nvidia chips to power AI ambitions," 14 August 2023. [Online]. Available: https://www.ft.com/content/c93d2a76-16f3-4585-af61-86667c5090ba

[19] NBC New York, "Saudi Arabia expects to get access to Nvidia's high performance chips 'within the next year'". [Online]. Available: https://www.nbcnewyork.com/news/business/money-report/saudi-arabia-expects-to-get-access-to-nvidias-high-performance-chips-within-the-next-year/5792133

[20] B. Nguyen, "US slows Nvidia and Intel AI chip exports to the Middle East," 31 May 2024. [Online]. Available: https://qz.com/us-slowing-nvidia-intel-amd-ai-chip-exports-middle-east-1851512692

[21] L. Mearian, "AI chip shortages continue, but there may be an end in sight," 7 May 2024. [Online]. Available: https://www.computerworld.com/article/2098937/ai-chip-shortages-continue-but-there-may-be-an-end-in-sight.html

[22] Arab News, "Saudi Arabia launches 'National Semiconductor Hub' to drive industry localization," 5 June 2024. [Online]. Available: https://www.arabnews.com/node/2524301/business-economy

[23] K. Freund, "AI Chip Vendors: A Look At Who's Who In The Zoo In 2024," 18 February 2024. [Online]. Available: https://www.forbes.com/sites/karlfreund/2024/02/13/ai-chip-vendors-a-look-at-whos-who-in-the-zoo-in-2024/

[24] Graphcore, "IPU Products," Graphcore, 2024. [Online]. Available: https://www.graphcore.ai/products

[25] Graphcore, "Graphcore joins SoftBank Group to build next generation of AI compute," 11 July 2024. [Online]. Available: https://www.graphcore.ai/posts/graphcore-joins-softbank-group-to-build-next-generation-of-ai-compute

[26] Groq, "Why Groq," [Online]. Available: https://wow.groq.com/why-groq/

[27] Artificial Analysis, "Groq - Quality, Performance & Price Analysis | Artifical Analysis," 2024. [Online]. Available: https://artificialanalysis.ai/providers/groq

[28] SambaNova, "Playground," 2024. [Online]. Available: https://sambaverse.sambanova.ai/playground

[29] SambaNova, "SN40L RDU AI Chip: Powering SambaNova Suite | SambaNova," [Online]. Available: https://sambanova.ai/technology/sn40l-rdu-ai-chip

[30] Cerebras, "The future of AI is Wafer-Scale," [Online]. Available: https://www.cerebras.net/product-chip/

[31] Arab News, "GAIN Summit: stc launches AI lab, inks partnerships," 16 September 2024. [Online]. Available: https://www.arabnews.com/node/2571659/corporate-news

[32] The Daily News, "Cerebras Signs MoU to Help Accelerate the Deployment of AI," 11 September 2024. [Online]. Available: https://www.galvnews.com/cerebras-signs-mou-to-help-accelerate-the-deployment-of-ai/article_8fa6bece-dcd4-5eb8-a239-75eae25a6e36.html

[33] Saudi Gazette, "Aramco Digital, Groq Partner to build world's largest AI Inferencing Data Center," 12 September 2024. [Online]. Available: https://www.saudigazette.com.sa/article/645478/SAUDI-ARABIA/Aramco-Digital-Groq-Partner-to-build-worlds-largest-AI-Inferencing-Data-Center

[34] M. Zuckerberg, "Open Source AI Is the Path Forward | Meta," 23 July 2024. [Online]. Available: https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/

[35] Hugging Face, 5 August 2024. [Online]. Available: https://huggingface.co/models

[36] White & Case, "AI Watch: Global regulatory tracker - Saudi Arabia | White & Case LLP," 20 June 2024. [Online]. Available: https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-saudi-arabia

[37] D. Watkins, "The AI-ready data center - DCD," 26 February 2024. [Online]. Available: https://www.datacenterdynamics.com/en/opinions/the-ai-ready-data-center/

FTI Delta is a global industry-specialized strategy consulting practice delivering end-to-end transformation. Our unrivaled team of experts offers a wide range of services that create value throughout the entire strategy-to-execution journey, serving top-tier corporations, private investors, mid-market companies and government authorities. FTI Delta is part of FTI Consulting (NYSE: FCN), a leading global advisory firm. For more information, please visit **ftidelta.com** and follow us on LinkedIn @FTI-Delta

FTI Consulting is an independent global business advisory firm dedicated to helping organizations manage change, mitigate risk and resolve disputes: financial, legal, operational, political and regulatory, reputational and transactional. FTI Consulting professionals, located in all major business centres throughout the world, work closely with clients to anticipate, illuminate and overcome complex business challenges and opportunities