



TMT Predictions 2025: Bridging the gaps

Deloitte predicts 2025 will be a “gap year” for gen AI and the TMT sector, marked by eight critical gaps that need to be bridged for today’s potential to be realized

ARTICLE • 16 MINUTE READ

As we look ahead to 2025 and beyond, it's clear that TMT is on the verge of a significant leap forward, largely powered by rapid generative AI adoption. But to get there, the industry will need to close gaps, including: balancing gen AI infrastructure investments with monetization, addressing gender disparities in gen AI usage, managing the energy consumption of gen AI data centers, tackling trust concerns surrounding deepfake content, discovering how best to use gen AI in media and gaming, and harnessing the power of gen AI agents to manage and act in real time. Further gaps exist in streaming video and cloud spending. Plus, there are non-gap predictions, around new smartphones and PCs with gen AI chips on them, new stadiums and other sports infrastructure leveling up the fan experience, and telco consolidation, specifically of wireless players. Overcoming these hurdles will be important to help businesses and industries thrive.

This blank space is caused due to formatting limitations. Text continues on next page.

Closing the gap for a brighter future

We are at a pivotal moment in the history of human invention. Future generations will certainly look back on the choices we make today. Deloitte's prediction that 2025 will be a "gap year" for generative AI underscores this significant inflection point. These gaps—spanning infrastructure investment, gender disparities, energy consumption, trust deficits, and the capabilities of gen AI agents—are not just challenges for the industry but societal imperatives. How we collectively address these gaps will define the legacy we create.

Beyond gen AI, advancements in cloud computing and telecommunications are expected to bring unprecedented efficiencies, new business models, and augmented consumer experiences. Investments in sports infrastructure and the increasing prominence of women's sports can act as catalysts for economic and social development. These trends reinforce the industry's role in fostering innovation that enhances businesses, consumers, and broader communities.

The 2025 TMT Predictions represent opportunities to create lasting impact. By navigating the path forward with trust, inclusivity, and sustainability at the forefront, industry advancements can benefit not only the current generation, but all those who follow. Together, we can rise to the occasion and bridge the gap to a brighter tomorrow.

-Lara Abrash, chair, Deloitte US

Eight gaps that mark 2025 as a "gap year" for TMT

1. **The gen AI infrastructure and monetization gap.** As we predicted last year, companies are spending tens of billions of dollars on chips and further hundreds of billions to build gen AI data centers for training and inference of gen AI models. While some companies offering gen AI enterprise software are seeing incremental revenues, the investment is 10 times (or more) higher than the return, at least for now. Those spending the most might suggest that the risk of underinvesting in gen AI is higher than the risk of overinvesting. But the gap persists and seems to be widening.
2. **The gen AI data center electricity and sustainability gap.** Proposed gen AI data centers require unprecedented amounts of power, preferably low carbon, which is creating a gap between their needs and the capacities of electrical grids, and companies' sustainability targets. Much is being done to close it by hyperscalers, chip companies, and utilities around the world, but the gap is expected to remain in 2025.
3. **The gen AI gender gap.** Women are less likely than men to use gen AI tools for both work and play. Some of this is due to lack of trust, but women's usage of gen AI is expected to catch up to men's usage ... within the year in some markets.
4. **The gen AI deepfake trust gap.** The proliferation of deepfake gen AI content (images, video, and audio) is making it harder for consumers, as a society, to trust their own eyes and ears. That gap needs to be bridged by the gen AI ecosystem comprehensively and immutably labeling gen AI content, as well as reliably and accurately detecting fake images in real time. The marginal cost of creating convincing deep fakes is falling, and the cost of detection needs to fall at an equivalent pace to help close the gap.
5. **The studio gen AI usage gap:** Many expect large studios to be using gen AI for content production, and some are, but there is a gap between those expectations and reality. Many are cautious about challenges

with intellectual property inherent to generative content, but they are keen to gain enterprise capabilities that can reduce time, lower costs, and expand their reach.

- 6. **The autonomous gen AI agent gap.** The prospect of autonomous bots that can consistently and reliably complete discrete tasks and orchestrate entire workflows is tantalizing. Agentic AI pilots are launching in 2024—will they reach widespread adoption in 2025?
- 7. **The streaming video gap.** Many media and entertainment companies assumed consumers would “buy and hold” multiple subscriptions. Instead, customers are looking to cut costs by bundling their favorite subscriptions and dropping others. We now see the number of services per household not merely stagnating but shrinking, and streamers increasingly relying on bundling to help fill the growth gap and using other parties to aggregate and distribute their content.
- 8. **The cloud spending gap.** One of the original selling points of using the cloud was the claim that it was cheaper, but in reality, spending is often decentralized and poorly controlled. Some buyers are leaning into FinOps to bridge the gap between promised cost savings and current spending to manage their cloud spending and potentially save billions.

New this year

This year, we’re introducing two new sections containing 11 additional mini predictions between them. Our *Updates* segment revisits seven topics from previous TMT Predictions reports to ask: “How’d we do?” while also exploring the latest predictions on those subjects. (Spoiler: We did really well!) Next, in our *Rising trends* section, we unveil four cutting-edge topics in TMT. While these emergent themes may not have made it into mainstream forecasts just yet, we believe they could be the hidden gems of tomorrow's industry conversations.

2025 topics

Here's a quick look at our main topics, plus those from our two new sections:

Women and generative AI: The adoption gap is closing fast, but a trust gap persists

For women to reap the full rewards of gen AI, tech companies should work to increase trust, reduce bias, and strive for more representative workforces

Deloitte predicts that women’s use of generative AI will equal or exceed that of men in the United States by the end of 2025. Although their use of gen AI was half that of men’s in 2023, their pace of adoption suggests they’ll reach parity in the United States within the next year and in many European countries within one to two years. Despite accelerating adoption, women express less trust that gen AI providers will keep their data secure—which may inhibit their full engagement and their gen AI spending. To help overcome this, tech companies should enhance data security and data management practices, mitigate AI bias, and improve women’s representation in their AI ranks.

As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions

The tech industry should optimize infrastructure, rethink chip design, and collaborate with electricity providers to help secure a more sustainable future for data centers

AI-driven data center power consumption will continue to surge, with Deloitte predicting data centers will only make up about 2% of global electricity consumption by 2025. This growth is expected to emerge from high-density data center infrastructure to support massive computing power and cooling needs. But regulatory, infrastructure, and cost issues are posing challenges for electricity generation and the grid to keep pace with data centers' unprecedented demand for 24/7 reliable energy. Technology and electric power industries can jointly address these issues by increasing the use of carbon-free sources, improving energy efficiency in gen AI chips and algorithms, and re-balancing compute-intensive AI workloads.

Ambitious stadium projects aim to bridge public-private investment goals

Sports owners transform stadiums into destinations, driving socioeconomic growth, community engagement, and revenue diversification

The sports industry has repeatedly demonstrated its ability to act as a catalyst for economic and social development, with stadiums largely at the epicentres of their communities. Investment in sports infrastructure is seeing an upward trend, as these developments often instigate wider returns to both the public and private sectors. With growth as a common goal between the public and private sectors, governments and communities can work with sports investors to provide supplementary investments in infrastructure and supercharge the socioeconomic impact of sports, enhance fan engagement, and diversify revenue streams for the organization. In 2025, we expect to see new stadium developments continue at pace, with almost half of these new projects expected to take place across North America and Europe.

Autonomous generative AI agents: Under development

Autonomous gen AI agents—agentic AI—could increase the productivity of knowledge workers and make workflows of all kinds more efficient. But the “autonomous” part may take time for wide adoption.

Deloitte predicts that in 2025, 25% of companies that use gen AI will launch agentic AI pilots or proofs of concept, and this figure will grow to 50% in 2027. Agentic AI has the potential to complete complex tasks autonomously, improving the productivity and efficiency of knowledge workers. Some of today's most promising applications include software development, customer support, cybersecurity, and regulatory compliance. The pace of improvement for agents is accelerating, but like most new technologies, widespread use will take time. That said, some agentic AI applications, in some industries, and for some use cases, may see actual adoption into existing workflows in 2025.

Deepfake disruption: A cybersecurity-scale challenge and its far-reaching consequences

As the effort to detect and combat fake content escalates, the costs of maintaining a credible internet may fall on consumers, creators, and advertisers alike

As AI generates increasing amounts of online images and video, questions around content authenticity and the potential harms of fake content grow more urgent. Online platforms, tech companies, and media players are taking two complementary approaches: using technology (often AI) to detect and flag fakes and using cryptographic metadata to assure provenance of authentic media assets. Deloitte predicts that this market will follow a similar pattern as that of cybersecurity, with bad actors finding ways to thwart detection tools and industries collaborating to confirm the credibility of at least some online content.

Cloud gets lean: ‘FinOps’ makes every dollar work harder

Enterprise cloud spend is growing, and using FinOps strategies can make each dollar work harder. Companies can save money, boost value, and build cross-functional cohesion.

Global cloud spend is set to top US\$825 billion in 2025,¹ but ask an organization’s leaders what they spend, and it might be difficult for them to answer. However, in 2025 Deloitte predicts more companies than ever will turn to “FinOps”, a set of tools and strategies to measure and optimize cloud spend, to save an estimated US\$21 billion. Companies can start simple, acting to reduce cloud waste, take advantage of discounts, and proactively right-sizing compute, network, and storage. But advanced companies could also drive cultural change, such as making business units financially accountable for their portion of the cloud bill. The goal is a “cloud unit economics” model—linking each dollar of spend to the business value it generates, so that companies can make more effective decisions about IT.

On-device generative AI could make smartphones more exciting —if they can deliver on the promise

With specialized chips and extensive mobile OS integration, smartphones could become smart—even intelligent. Will users embrace the new approach?

Deloitte predicts that in 2025, global smartphone shipments will see a modest lift to around 7%, up from about 5% annual growth in 2024. Some of this lift will be due to the device upgrade cycle, which has been down the past two years, and some will be from early adopters seeking new generative AI capabilities. Smartphones with on-device generative AI capabilities will test the value of features like intelligent assistants and conversational interfaces; the capabilities of small models running on-device; and the business models seeking economic value from the capital intensity of the generative AI buildout. There is excitement about generative AI, but can the technology deliver on its promises, and will users embrace a new way of interacting with the most widely used consumer device?

Large studios will likely take their time adopting generative AI for content creation. Social media isn’t hesitating.

Hollywood (and others) may be cautious about using gen AI for content creation, but they will likely be quicker to adopt it for operations and distribution

In 2025, Deloitte predicts that the biggest TV and film studios—especially those in the United States and European Union—will be cautious in adopting generative AI into their creative workflows, with less than 3% of their production budgets going to these tools. But we also predict that operational spending will expand by 10% to integrate generative AI enabled tools for more bread-and-butter functions like contract and talent management, permitting and planning, marketing and advertising, and localization and dubbing of content that can expand their reach into diverse global markets. This approach can help studios slow the potential disruptions that gen AI can pose to talent and content, while more quickly adopting gen AI tools that can help reduce costs and accelerate performance across their businesses.

Reevaluating direct-to-consumer: The shift toward video aggregators

Video content creators may need more distributors to reach their total addressable market

Consumers are expected to continue to stack streaming video services and have more than one standalone subscription at a time in 2025, but Deloitte predicts that the stack is about to get shorter. According to Deloitte surveys, after reaching more than four services per consumer in the United States and over two in most European markets in 2023 and 2024, we appear to have passed the peak, and the number of standalone services will slowly decline in most markets. Instead of consumers directly subscribing to each content provider's service, there will likely be increased aggregation, where intermediaries—ranging from telcos to grocery stores to tech platforms to streamers themselves—combine multiple content sources in a single package. This trend is likely a win for many of the players, keeping costs under control and creating a stable and sustainable streaming ecosystem for 2025 and beyond.

Wireless telecom consolidation speeds up ... where regulators allow

In many markets, smaller wireless telecoms see slow growth, low profits, and have debt to repay. M&A, specifically combining assets or even entire consumer-facing companies, may help where it gets approved by regulators.

In some markets, especially Europe and Asia, there is an increasing perception that wireless markets are too fragmented, subscale and unsustainable, with the smaller third and fourth place operators not able to invest in networks over the long term. Although these markets have historically kept the number of operators high, recent conversations have seen opportunity to allow or even encourage consolidation. Deloitte predicts that although it is expected to be a slow process, and regulators will have their conditions, there will be an increased pace of consolidation, beginning in 2025 and continuing, creating a more viable and sustainable wireless ecosystem, especially in smaller markets.

Updates

This year we're looking at seven previous predictions to see how we did, and what the latest developments are:

Generative AI comes to the enterprise edge: 'On-prem AI' is alive and well

Owning their own servers gives companies a more private, secure, flexible, and possibly cheaper IT environment for AI

Deloitte predicts that although gen AI via cloud will continue to be the dominant option in 2025, about half of the enterprises worldwide will add AI data center infrastructure on premises, primarily to protect their IP and sensitive data, comply with data sovereignty or other regulations, and save costs. Deloitte's 2024 *State of Generative AI in the Enterprise* Q2 survey noted 80% of companies with "very high" AI expertise reported spending more on AI in the cloud ... but 61% are investing more in their own hardware. Enterprise gen AI will likely be a hybrid approach, with enterprises doing some in the cloud, and some on premise.

(Re)defining the investment case for women's sports

Rising women's sports revenue fuels investor interest and valuation records

The increasing professionalization and commercialization of women's sports around the world is garnering the attention of fans, sponsors and—critically—investors. In 2024, we predicted the women's elite sports market would generate over \$1 billion in revenue. In North America, clubs are recognizing record valuations, from Angel City FC in the National Women's Soccer League at a valuation of \$250 million² to Las Vegas Aces in the Women's National Basketball Association at \$140 million.³ Elsewhere, organizations are creating innovative structures to channel investment into their women's teams and emphasizing strategic growth, independent leadership, and commercial opportunities. In 2025, we expect to see an expanding group of investors—including institutional investors, private equity and high net worth individuals—take more note.

Fixed wireless access: Contrary to popular opinion, adoption may continue to grow

With US FWA net adds likely being slightly lower than last year, and some markets not expected to take off until 2026 ... there may be pockets of unrealized or potential growth out there, both in the US and globally

Fixed wireless access (FWA)—when consumers and enterprises get their home broadband over a fixed cellular device (mainly 5G) rather than via wires—has been *the* 5G growth story over the last few years in the United States, with well more than 10 million homes expected to be connected by the end of 2024. However, growth is slowing, with the first quarter of 2024 net additions lower than Q1 2023, and a potential slowdown anticipated in 2025. Despite this, Deloitte predicts global FWA net additions will rise by 20% annually in 2025 and 2026 (in line with our 2022 Prediction on FWA), driven by growth in new enterprise FWA and markets that are not as large as the United States or India individually but still add up to millions in new subscriptions annually.

5G standalone appears to be at a standstill: Will 6G run late?

Telecoms reassess investments in 5G standalone and delay 6G progress amid ROI concerns

The deployment of 5G standalone networks is progressing more slowly than expected. Telecom companies may be hesitant to invest heavily in this next generation of 5G in part due to underwhelming returns on their existing 5G investments, making the rollout of 6G seem further away than ever. In 2022, Deloitte Global predicted that the number of telcos investing in 5G SA networks would double from more than 100 operators in 2022 to at least 200 by the end of 2023, but that has not happened: as of March 2024, only 49 operators (out of 585 who have launched 5G globally) have deployed, launched, or soft-launched 5G SA networks.⁴ In 2025, Deloitte predicts that fewer than 20 additional networks will be upgraded to standalone, keeping 5G SA at around 12% of all 5G deployments.

Open RAN mobile networks and vendor choice: Single vendor now, multivendor when?

Open RAN's journey toward a diverse, multivendor ecosystem is marked by slow growth and complex challenges

Open Radio Access Network (Open RAN) aims to democratize networks by providing mobile network operators (MNOs) who build RANs with greater choice and more flexibility. In 2021, Deloitte predicted that global active Open RAN deployment would double from 35 to 70. We were too optimistic: as of March 2024, the ongoing Open RAN deployments and trials stand at 45, with only two networks globally being multivendor Open RAN.⁵ The transition towards a diverse, multivendor ecosystem is proving slower and more complex than initially anticipated, and realizing true multivendor Open RAN may take a while. Deloitte predicts that no additional multivendor Open RAN networks will be deployed or announced in 2025.

Despite quantum's slow start, don't be slow to start your defense against it

Quantum drug discovery and financial modeling are likely several years away, but the time needed to upgrade cyber defenses for the quantum age likely necessitates prompt action

As Deloitte predicted in past reports, quantum computers remain works in progress, with few real-world use cases where they offer a computing advantage, at least for now. But the threat of “harvest now, decrypt later” attacks, where threat actors gather encrypted data, store it for years, and then unlock it with cryptographically relevant future quantum computers at some point has reached a tipping point. Deloitte predicts that the number of companies, and the dollars spent, working on implementing post quantum cryptography standards will quadruple in 2025 compared with 2023. Post-quantum cryptographic solutions are expected to span the gamut in 2025, from enterprise and hyperscalers to consumer smartphones and messaging services.

RISC-V: Closing the geopolitical gen AI loophole

An open-source alternative to proprietary chip design is gaining popularity among CPU designers. Its potential role in gen AI for markets under export restriction is a new twist for a new technology.

As predicted in Deloitte's 2022 *Global TMT Predictions* report, open-source RISC-V (pronounced risk five) CPU chips revenues are expected to be close to US\$1 billion by 2024. And RISC-V based SoC shipments could be close to two billion units. But in a new development, there are discussions in the United States about restricting RISC-V chips exports, as it involves potential national security risks, given how China is “making significant investments in RISC-V chip design architecture” to undermine US export controls and “leapfrog” US technological leadership in chip design. Gen AI servers need both GPUs (already under various export restrictions) and CPUs to control data flows. Closed-source CPU architectures are already under US export restrictions, and therefore, restricting RISC-V based designs may close an apparent alternative.

Rising trends

Keep your eye on these newly emerging trends. We predict they could soon become the center of attention, transforming the conversation and shaping the future of the industry:

Generative AI and cyber: Big risks, but big opportunities too

Recognizing generative AI's potential for enabling both threats and cyber solutions, cybersecurity professionals are exploring ways to harness its power to counter emerging risks and help fortify the technology environment

From the 2024 Deloitte-NASCIO Cybersecurity Study, nearly three quarters of security experts surveyed said the cyber threat from AI was high. Gen AI-based cyberattacks look like they will have doubled or tripled in 2024, and Deloitte predicts they will grow again in 2025, used by threat actors writing malicious phishing emails, deepfakes, or software code for malware attacks. Tech companies who make gen AI tools will likely develop guardrails to prevent malicious use in 2025. While gen AI tools can be used by threat actors for malicious purposes, the same tools can also be used by defenders to help improve security processes, monitoring, and risk management.

Silicon building blocks: Chiplets could move Moore’s Law forward

Chiplets promise to deliver more flexible, scalable, and efficient systems for AI and high-performance computing environments, at higher yields

Chiplets—a heterogeneous technological architecture to develop and package semiconductors—enable high-speed data transfers, reduce latency, and help optimize PPA (power, performance, and area). Deloitte predicts worldwide advanced packaging revenue based on chiplets will more than double from an estimated US\$7 billion in 2021 to reach US\$16 billion in 2025. Chiplets are already used and explored in some of the fast-growing markets such as AI accelerators (especially generative AI), high performance computing, and telecommunications applications. They’re enabling the semiconductor industry to continue increasing performance and yield.

B/OSS: Telcos modernize their business and operational support systems software

Telcos’ back-end business and operations software market is growing slowly but modernizing it—by adopting SaaS and microservices architecture, moving to the cloud and more—is a hot spot of growth for software vendors and an opportunity for telcos to do more with 5G, fiber, and AI

Historically, telcos maintained separate IT systems: business support systems (BSS) for customer orders, customer relationship management (CRM), and billing, and operational support systems (OSS) for order management, network inventory, and operations, often custom-built and hardware defined. These systems were typically on-premises and composed of vertical, siloed solutions. However, by 2025, many telcos are expected to modernize and integrate these systems, driven by evolving customer expectations and new digital revenue streams. Deloitte predicts that the global B/OSS market will reach \$70 billion by 2025, growing at 5% annually. Cloud-based solutions and software-as-a-service offerings are expected to grow significantly faster, at 22% and 18% annually. Most of the growth in the next few years is expected to come from the Americas, Middle East, North Africa, and emerging Asia-Pacific regions.

Silicon photonics: Gen AI communicates at lightspeed

Propelled by the demanding requirements of gen AI, optical devices on silicon are stepping out of research labs and into the limelight of data centers

Deloitte predicts that sales of silicon photonics chips used as optical transceivers will grow at a compounded annual growth rate of 25% from 2023 to 2025 to reach US\$1.25 billion in 2025. These chips allow gen AI data centers to communicate at lightspeed, use components that are smaller and cheaper, consume less energy, and produce less heat than the traditional alternatives. In 2025, the main driver of silicon photonics adoption is expected to be in data center applications, specifically for those running gen AI training and inference—especially where data needs to travel anywhere from 10 cm to 10 meters between chips, trays, and racks.

By	Ariane Bucaille France	Kevin Westcott United States
	Gillian Crossan United States	Lara Abrash United States

Endnotes

- 1. Gartner, “*Gartner forecasts worldwide public cloud end-user spending to surpass \$675 billion in 2024*,” press release, May 20, 2024.
- 2. Angle City FC, “*Willow Bay and Bob Iger to become Angel City's new controlling owners*,” July 17, 2024.
- 3. Josh Sim, “*Las Vegas Aces valued at US\$140m as average WNBA team hits US\$96m*,” *SportsPro*, June 18, 2024.
- 4. GSA, “*5G - GSA Market Snapshot March-2024*,” March 4, 2024.
- 5. TeckNexus, “*Current State of Open RAN – Countries & Operators deploying & trialing Open RAN*,” March 10, 2024.

Acknowledgements

We wish to thank **Duncan Stewart**, **Jeff Loucks**, and **Paul Lee**, plus the entire team, for their work on the Predictions report.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

Women and generative AI: The adoption gap is closing fast, but a trust gap persists

For women to reap the full rewards of gen AI, tech companies should work to increase trust, reduce bias, and strive for more representative workforces

ARTICLE • 9 MINUTE READ

Deloitte predicts that the experimentation with and use of generative AI by women will equal or exceed that of men in the United States by the end of 2025.¹ Although women’s use of gen AI was half that of men’s in 2023, their pace of adoption suggests they’re likely to reach parity within the next year.² While this parity prediction is for the United States, the gen AI gender gap is a global phenomenon: In European countries, where the use of gen AI has been surveyed, our analysis not only identified significant gender adoption differences but also revealed that women are making up ground rapidly.³ These countries will likely close the adoption gender gap within the next two years—and the global challenges and opportunities for adoption will likely mirror the US findings.

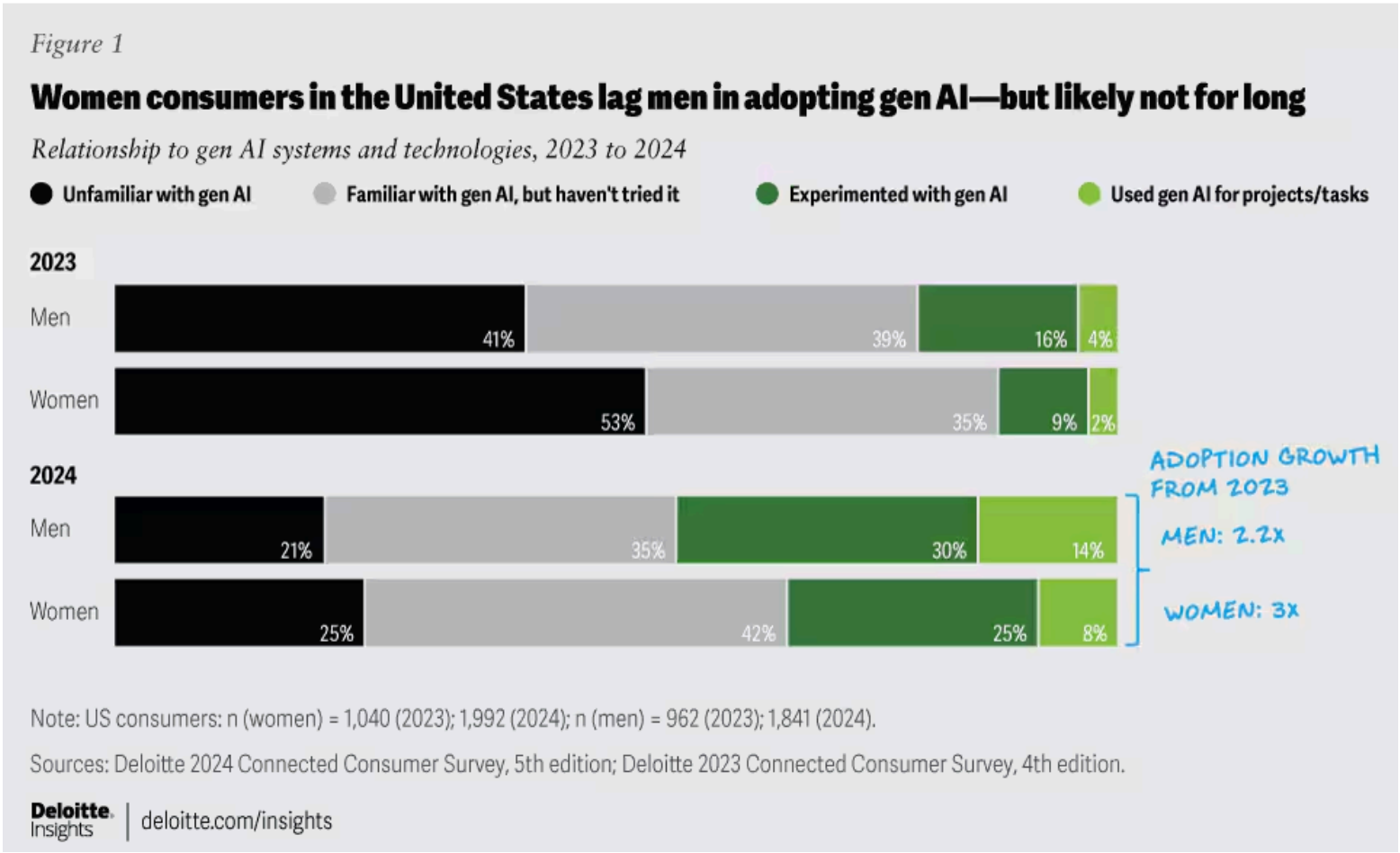
Despite accelerating their gen AI adoption, women express less trust than men that gen AI providers will keep their data secure.⁴ This “technology trust gap” could inhibit women’s regular use of the technology and full participation in new gen AI applications, as well as slow down their future purchasing of gen AI products and services. To help overcome this trust gap, tech companies should enhance their data security, implement clearer data management practices, and provide greater data control.

AI model bias can also have a negative impact on trust.⁵ Women constitute less than one-third of the AI workforce,⁶ and most AI workers feel that AI will produce biased results as long as their field continues to be male dominated.⁷ Increasing women’s presence in the field can help reduce gender bias in AI, as well as give women a greater role in steering the future of the technology.

The gen AI adoption gap is closing rapidly

Recent Deloitte research has highlighted a gender gap in generative AI adoption across various geographies. For the past two years, the Deloitte Connected Consumer Survey has investigated the adoption of gen AI by US consumers as part of its research into digital life.⁸ Our analysis revealed that women in the United States have been lagging in taking up this emerging technology (figure 1): In 2023, women’s adoption of gen AI was roughly half that of men (11% of women reported experimenting with gen AI or using it for projects and tasks beyond experimentation, vs. 20% of men). In 2024, the same survey revealed that gen AI adoption overall had more than doubled, but the gender gap remained: Thirty-three percent of women surveyed reported using or experimenting with gen AI, vs. 44% of men.

The gen AI gender gap has been noted in other geographies too: Deloitte UK’s 2024 Digital Consumer Trends survey of UK consumers reported that 28% of women were using gen AI, vs. 43% of men.⁹ Analysis of this study, as well as Deloitte UK’s European study on gen AI and trust, revealed double-digit differences between women’s and men’s adoption of gen AI in 12 additional European countries.¹⁰



In the United States, women are rapidly closing the adoption gap. In the past year, the proportion of US women surveyed who have adopted gen AI *tripled*—outpacing the 2.2x rate of growth for men.¹¹ Analysis of current adoption levels and these rates of growth allows Deloitte to predict that the proportion of women experimenting with and using gen AI for projects and tasks will match or surpass that of men in the United States by the end of 2025.¹²

Full engagement may be harder to achieve

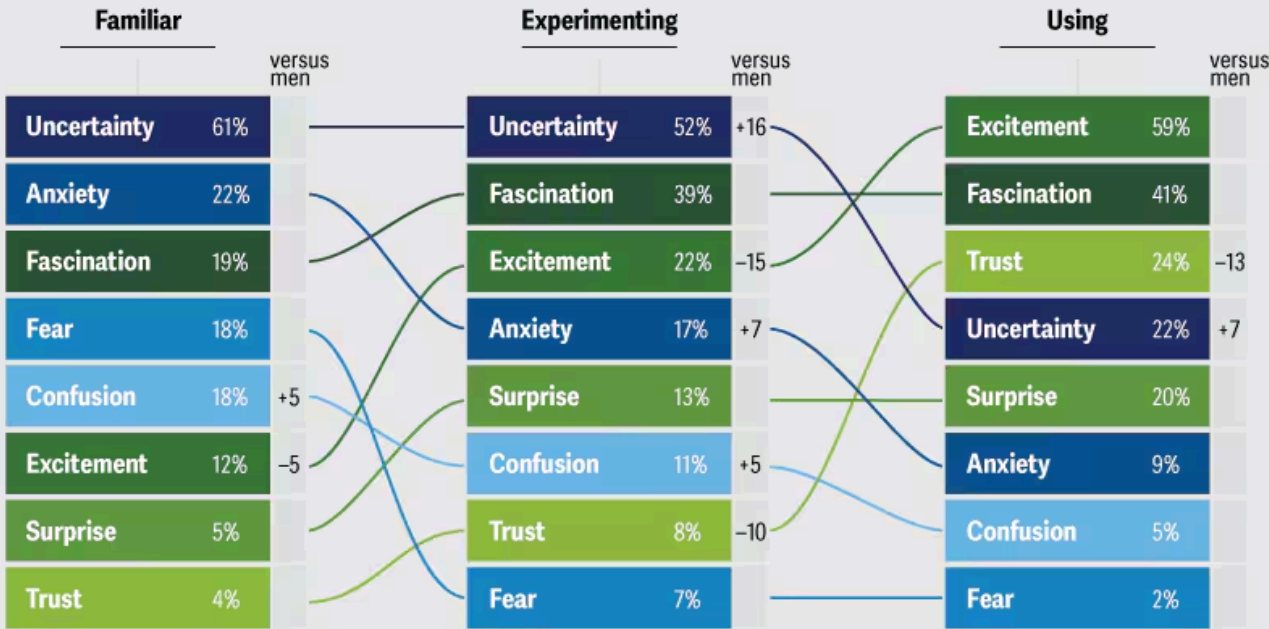
While the trend is encouraging, reaching adoption parity won’t automatically ensure that women will incorporate gen AI into their everyday workflows. Indeed, among gen AI users surveyed in Deloitte’s 2024 Connected Consumer Survey, 34% of women say they use the technology at least once a day, vs. 43% of men.¹³ And among gen AI users who reported using it for professional tasks, 41% of women currently feel that gen AI substantially boosts their productivity, vs. 61% of men.¹⁴ Tech companies and other organizations looking to benefit from using gen AI should heed these differences and take active steps to improve women’s engagement.

The contrasts between genders may stem partly from a striking difference in perspective on trust.¹⁵ As women progress from familiarity with gen AI into experimentation and use, negative emotions of uncertainty, anxiety, fear, and confusion diminish, while positive feelings of fascination, excitement, surprise, and trust grow (figure 2).¹⁶ However, at both the experimentation and project and task use levels, women’s feelings of trust toward the technology are significantly lower than men’s, and their feelings of uncertainty remain higher. Indeed, only 18% of women surveyed who are experimenting with or using generative AI indicated having “high” or “very high” trust that the providers of the gen AI capabilities they use will keep their data secure—whereas, for male adopters, that number has reached 31%.¹⁷

Figure 2

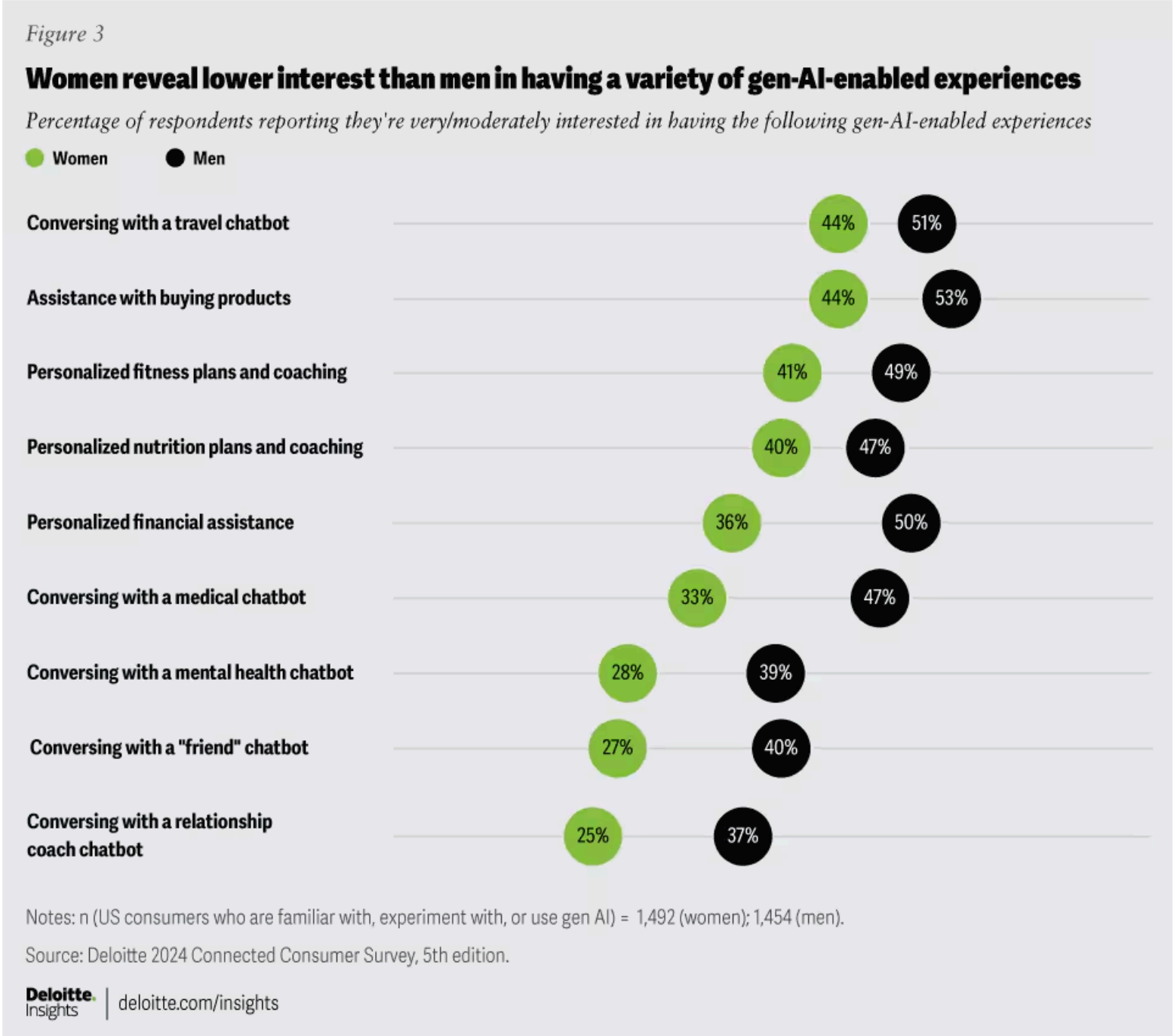
As women gain more experience using gen AI, negative emotions diminish and positive emotions, including trust, grow—but the feeling of trust still lags behind men

Percentage of women selecting each as a top-two emotion they feel toward gen AI—whether they're familiar with it, experimenting with it, or using it for projects/tasks



The trust gap is not unique to gen AI, but extends to broader tech services and interactions: While 54% of women surveyed in Deloitte’s 2024 Connected Consumer Survey agree that the benefits they get from online services outweigh their data privacy concerns (an improvement from 46% in 2023), more men agree (62%).¹⁸ Last year, we reported that women are more wary than men about how their personal data is used and protected and that it was affecting their willingness to share data, particularly when it comes to sensitive health and fitness metrics.¹⁹ Women may perceive the potential consequences of a security breach or data misuse as more significant.²⁰

The growing popularity of generative AI may exacerbate these longstanding issues around data privacy and tech.²¹ As users interact with gen AI, the systems may feed users’ data back into their AI models—and experts say it’s not necessarily clear or easy to opt out of having one’s data used for AI training.²² As consumers start to converse with gen AI for advice on sensitive, personal topics, the data privacy and security stakes grow. Indeed, the trust gap around data privacy and security may underpin the differences we’re seeing between women’s and men’s levels of interest in having a variety of gen AI experiences in the future (figure 3).²³ Surveyed women are somewhat less interested than men in interacting with gen AI on less-sensitive topics (such as travel, shopping, fitness, and nutrition), but they are *substantially* less interested than men in engaging with gen AI around more sensitive topics (such as personal finances and relationships, and medical or mental health issues).



The trust gap may also contribute to less excitement among women to purchase new gen AI technologies. Tech companies are beginning to sell laptops, tablets, and smartphones with embedded AI chips designed to improve functionality (for example, summarizing information in real time, generating photos and videos, and instantly translating foreign languages).²⁴ When Deloitte’s 2024 Connected Consumer Survey asked whether new AI functionality will have any effect on their plans to upgrade devices, fewer women said they’re likely to upgrade their devices sooner compared to men.²⁵ For example, while 43% of men with smartphones said embedded AI would make them very or somewhat likely to upgrade their phone sooner than planned, only 32% of women said the same (conversely, 58% of women said it would have no effect on their upgrade plans, vs. 50% of men). And when it comes to laptops, 41% of men said on-device AI would make them very or somewhat likely to upgrade those devices sooner, vs. 28% of women. With women controlling or influencing an estimated 85% of consumer spending, their lower enthusiasm for upgrading to devices with AI could pose an issue for tech providers.²⁶

The trust gap is not the only factor holding women back from maximizing their use of gen AI. Women gen AI users surveyed are less likely to feel that their company actively encourages their use of the technology at work (61% of women users feel this way, vs. 83% of men).²⁷ And while 49% of women gen AI users say that their company invests in training employees on how to use generative AI, that falls short of the 79% of men reporting the same. Whether these numbers reflect differences in perception or actual experiences with access to training programs and encouragement in the workplace, companies should pay heed and work to close the gaps.

Women in tech are forging ahead with gen AI—but better representation is needed

In the tech industry, there is a different story about gen AI adoption entirely—and women working in tech may hold clues for fostering greater gen AI engagement by women overall in the future. Not surprisingly, the industry creating AI products and services has higher levels of gen AI adoption among its employees: In Deloitte’s 2024 Connected Consumer Survey, 70% of women and 78% of men working in the tech industry reported experimenting with gen AI or using it for projects or tasks—far outpacing nontech women (32%) and men (40%).²⁸ What may be more surprising is that women working in the tech industry appear to be moving beyond gen AI experimentation and into using it for projects and tasks faster than their male counterparts (44% vs. 33%). And both groups are anticipating greater benefits: About 7 in 10 women and men working in tech expect their use of gen AI to “substantially boost” their productivity at work a year from now.²⁹

What’s more, there’s no notable trust gap between tech women and men. Both groups have greater trust in generative AI than adopters overall: More than 40% of tech women and men using or experimenting with gen AI reported having “high” or “very high” trust that gen AI providers will keep their data secure.³⁰ In both groups, 75% of those surveyed agree that the benefits they get from online services outweigh their privacy concerns—vs. just 54% of women and 60% of men working outside tech.³¹ It’s likely that women in the tech industry have a better understanding of how gen AI works than nontech workers, and that their heavier professional use of gen AI has increased their comfort level and shown them how they can benefit from the technology. Moreover, most tech women who use gen AI reported that their companies encourage its use (84%) and provide training (72%)—in contrast, among women using gen AI in other industries, far fewer reported that their companies encourage its use (55%) or provide training (45%).³²

Despite the greater adoption of AI by women in the tech industry, there’s a relative lack of women working in AI roles. Women only make up about 30% of the AI-related workforce, which is comparable to their representation in STEM fields overall.³³ This underrepresentation of women in AI could have serious implications for the development and deployment of AI systems across various domains and sectors.

One of the major challenges posed by the relative lack of women in the AI workforce is the risk of perpetuating gender bias against women in AI applications.³⁴ As many as 44% of AI systems across industries exhibit gender bias, which can negatively affect outputs from AI systems in ways that continue to marginalize and underrepresent women.³⁵ For instance, gender bias in AI can lead to bias in hiring practices,³⁶ lower quality health care,³⁷ and reduced access to financial services for women.³⁸ And Deloitte research has shown that bias in AI models can erode employee and customer trust.³⁹ Bringing more women into AI jobs can be crucial for achieving gender equality and ensuring that AI benefits society.⁴⁰

Bottom line

There are several reasons why tech companies should work toward increasing women’s engagement with gen AI. First, with women controlling or influencing most consumer purchasing, failing to get women on board with frequent gen AI use could increase the risk that AI products and services won’t achieve their expected potential. Second, if women don’t engage with gen AI tools as fully as male employees, companies could risk not achieving the productivity gains they might expect to see after investing in gen AI. And, because gen AI depends upon collecting and building upon interaction data, the underrepresentation of women’s interactions could exacerbate biases in AI models.⁴¹ Finally, if women don’t participate in emerging gen AI use cases as fully as they could, that may keep them from maximizing future tech benefits (for example, advantages of chatbot interventions in medical or mental health) and deepen existing inequities.⁴²

To help bolster women’s trust in gen AI, tech companies should work to address the potential risks associated with the technology. Deloitte’s 2024 Connected Consumer Survey found that earning trust may depend at least partially on improving the transparency of tech companies’ data privacy and security policies, as well as making it easier for consumers to control their personal data.⁴³ Tech companies should consider prioritizing robust data security measures and communicating their data-handling practices more effectively. Making it simpler for consumers to understand what data gets collected and how it’s used, along with providing easier ways to control that use (such as prompting users at appropriate points to make informed choices about the use of their data) may not only build trust but could also confer a competitive advantage. But it’s not just tech companies that should pay heed to potential gen AI risks: Eighty-four percent of survey respondents believe that governments should do more to regulate the way companies collect and use consumer data.⁴⁴

Across industries, companies that want to achieve full use of gen AI by men and women workers should take care to encourage the use of gen AI capabilities. Beyond various popular professional use cases—document editing, web searches, summarizing materials, and research assistance—companies can embrace industry-specific ways to use generative AI.⁴⁵ Maximizing the use of gen AI by employees may require establishing training programs.

Striving for full consumer engagement in generative AI is a commendable objective, but it may be more difficult to achieve without equitable representation among the people who develop generative AI technologies. To increase the diversity and inclusion of women in AI roles, companies should consider focusing on creating workplaces that meet the needs of those they employ. For example, a study of women in AI noted that work/life balance is the most important factor for their job satisfaction, which includes elements such as having a flexible working schedule or being able to work remotely.⁴⁶ Women also reported looking for jobs with women in leadership, transparency around pay and promotions, and zero-tolerance policies for harassment and abuse.⁴⁷ Attracting more women to the field may also involve providing more education and training opportunities for women to learn AI skills and competencies. It could also be beneficial to create more mentorship and networking programs that allow women in AI to share their experiences and support one another, and to provide funding for more women to participate in AI research and innovation projects. As women’s role in developing gen AI grows, it’s likely that there will be applications and systems that engage *all* women more.

By	Susanne Hupfer United States	Bree Matheson United States
	Gillian Crossan United States	Ariane Bucaille France
	Jeff Loucks United States	

Endnotes

1. To understand consumer attitudes toward digital life, the Deloitte Center for Technology, Media & Telecommunications surveyed 3,857 US consumers in the second quarter of 2024 and 2,018 US consumers in the second quarter of 2023. These 2024 and 2023 Connected Consumer Surveys collected data on consumers' reported adoption of generative AI, including experimentation and use for projects and tasks (beyond experimentation). By analyzing longitudinal adoption data and calculating the rate of change in adoption from 2023 to 2024 for men and women, we are able to project that women will close the adoption gap by the end of 2025; see: Jana Arbanas et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey, 5th edition*, Deloitte, December 3, 2024; Jana Arbanas, Paul H. Silvergate, Susanne Hupfer, Jeff Loucks, Prashant Raman, and Michael Steinhart, "[Balancing act: Seeking just the right amount of digital for a happy, healthy connected life](#)," *Deloitte Insights*, Sept. 5, 2023.
2. Ibid.
3. Our analysis was conducted from August to October 2024, based on data from Deloitte UK's 2023 and 2024 Digital Consumer Trends surveys, as well as a 2024 Deloitte UK survey of European consumers on the topic of generative AI; see: Paul Lee and Ben Stanton, "[Generative AI: 7 million workers and counting](#)," Deloitte, June 25, 2024; Jonas Malmund, Frederik Behnk, and Joachim Gullaksen, "[Generative AI is all the rage](#)," Deloitte, 2023; Roxana Corduneanu, Stacey Winters, Jan Michalski, Richard Horton, and Ram Krishna Sahu, "[Europeans are optimistic about generative AI but there is more to do to close the trust gap](#)," *Deloitte Insights*, Oct. 10, 2024.
4. Analysis based on Deloitte's 2024 Connected Consumer Survey; see: Arbanas et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
5. Don Fancher, Beena Ammanath, Jonathan Holdowsky, and Natasha Buckley, "[AI model bias can damage trust more than you may know. But it doesn't have to](#)," *Deloitte Insights*, Dec. 8, 2021.
6. World Economic Forum, "[Global gender gap report 2023](#)," June 2023.
7. Deloitte AI Institute, "[Women in AI](#)," accessed November 2024.
8. Jana Arbanas et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey, 5th edition*, Deloitte, publishing December 3, 2024; Arbanas, Silvergate, Hupfer, Loucks, Raman, and Steinhart, "[Balancing act](#)."
9. Deloitte, "[Generative AI: 7 million workers and counting](#)," accessed November 2024.

10. The Digital Consumer Trends study conducted in various countries in 2024 revealed gen AI adoption gaps of 17 points in Denmark; 12 points in Sweden, Italy, and the Netherlands; 11 points in Belgium; and 10 points in Norway. Additional analysis of a Deloitte European gen AI study revealed gen AI adoption gaps ranging from 10 to 15 points in 11 European countries studied (Belgium, France, Germany, Ireland, Italy, the Netherlands, Poland, Spain, Sweden, Switzerland, and the United Kingdom); see: Deloitte, “[Generative AI](#)”; Deloitte, “[Generative AI is all the rage](#),” accessed November 2024; Corduneanu, Winters, Michalski, Horton, and Sahu, “[Europeans are optimistic about generative AI but there is more to do to close the trust gap](#).”
11. Analysis based on 2024 and 2023 Deloitte Connected Consumer Surveys; see: Arbanas et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*; Arbanas, Silverglate, Hupfer, Loucks, Raman, and Steinhart, “[Balancing act](#).” Deloitte, “[Generative AI](#).”
12. Ibid.
13. Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
14. Ibid.
15. Ibid.
16. Ibid.
17. Ibid.
18. Ibid.
19. For example, only 43% of women we surveyed in the Deloitte 2023 Connected Consumer Survey who owned smart watches or fitness trackers said that they share the data collected by those devices with their health care provider, vs. 57% of men; see: Susanne Hupfer, Jennifer Radin, Paul H. Silverglate, and Michael Steinhart, “[Tech companies have a trust gap to overcome—especially with women](#),” Deloitte Insights, Nov. 8, 2023.
20. These fears may be warranted. Consider that most health apps—along with the data they gather and transmit—are not covered by the Health Insurance Portability and Accountability Act, which means the data may be shared or sold to third parties; see: Steve Alder, “[Majority of Americans mistakenly believe health app data is covered by HIPAA](#),” The HIPAA Journal, July 26, 2023.
21. Ina Fried, “[Generative AI’s privacy problem](#),” Axios, March 14, 2024; Federal Trade Commission, “[AI companies: Uphold your privacy and confidentiality commitments](#),” Jan. 9, 2024.
22. Ibid; Matt Burgess and Reece Rogers, “[How to stop your data from being used to train AI](#),” Wired, April 10, 2024.
23. Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.

24. Baris Sarer, Ricky Franks, Cheryl Ho, and Jake McCarty, “[AI and the evolving consumer device ecosystem](#),” *The Wall Street Journal*, April 24, 2014; Sam Reynolds, “[AI-enabled PCs will drive PC sales growth in 2024, say research firms](#),” *Computer World*, Jan. 11, 2024; Clare Conley, “[Generative AI in 2024: The 6 most important consumer tech trends](#),” *Qualcomm*, Dec. 14, 2023.
25. Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
26. Monique Woodard, “[Unlocking the trillion-dollar female economy](#),” *TechCrunch*, May 21, 2023.
27. Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
28. Ibid.
29. Across industries, 51% of women workers using gen AI anticipate it would substantially boost their productivity at work a year from now, vs. 64% of men; see: Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
30. Tech women and men are statistically tied: Forty-two percent of tech women who use or experiment with gen AI have “high” or “very high” trust that gen AI providers will keep their data secure, and another 40% report moderate trust, while 47% of tech men report “high” or “very high” trust and another 30% report moderate trust; see: Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
31. Ibid.
32. Greater proportions of men in the tech industry who use gen AI report that their employers encourage its use (93%) and provide training (91%). While there’s still a gender gap in these views among workers in the tech industry, the gap is significantly smaller than among men and women working in other industries; see: Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey*.
33. World Economic Forum, “[Global gender gap report 2023](#).”
34. Deloitte, “[Generative AI](#).”
35. Genevieve Smith and Ishita Rustagi, “[When good algorithms go sexist: Why and how to advance AI gender equity](#),” *Stanford Social Innovation Review*, March 31, 2021.
36. Charlotte Lytton, “[AI hiring tools may be filtering out the best job applicants](#),” *BBC*, Feb. 16, 2024.
37. Carmen Niethammer, “[AI bias could put women’s lives at risk - A challenge for regulators](#),” *Forbes*, March 2, 2020.
38. Ryan Browne and MacKenzie Sigalos, “[A.I. has a discrimination problem. In banking, the consequences can be severe](#),” *CNBC*, June 23, 2023.

39. Fancher, Ammanath, Holdowsky, and Buckley, “*AI model bias can damage trust more than you may know. But it doesn’t have to.*”

40. World Economic Forum, “*Global gender gap report 2023.*”

41. Smith and Rustagi, “*When good algorithms go sexist.*”

42. Hyun-Kyoung Kim, “*The effects of artificial intelligence chatbots on women’s health: A systematic review and meta-analysis,*” Healthcare, Feb. 23, 2024; Sheryl Jacobson and Jen Radin, “*Can FemTech help bridge a gender-equity gap in health care?*” Deloitte, Oct. 5, 2023; Karen Taylor, “*Why investing in FemTech will guarantee a healthier future for all women,*” Deloitte UK, June 23, 2023.

43. Arbanas, et al., *Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.*

44. Ibid.

45. Deloitte AI Institute, “*The generative AI dossier: A selection of high-impact use cases across six major industries,*” April 3, 2023.

46. Women in AI, “*WAI at work: Shaping the future of work for women in AI,*” 2022.

47. Ibid.

Acknowledgements

Authors would like to thank **Duncan Stewart, Paul Lee, Ben Stanton, Vipul Mehta, Roxana Corduneanu, Michael Steinhart, Michelle Dollinger, Jeff Stoudt, Catherine King, Elizabeth Fisher, Andy Bayiates, Prodyut Borah, Molly Piersol**, Deloitte Insights team.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions

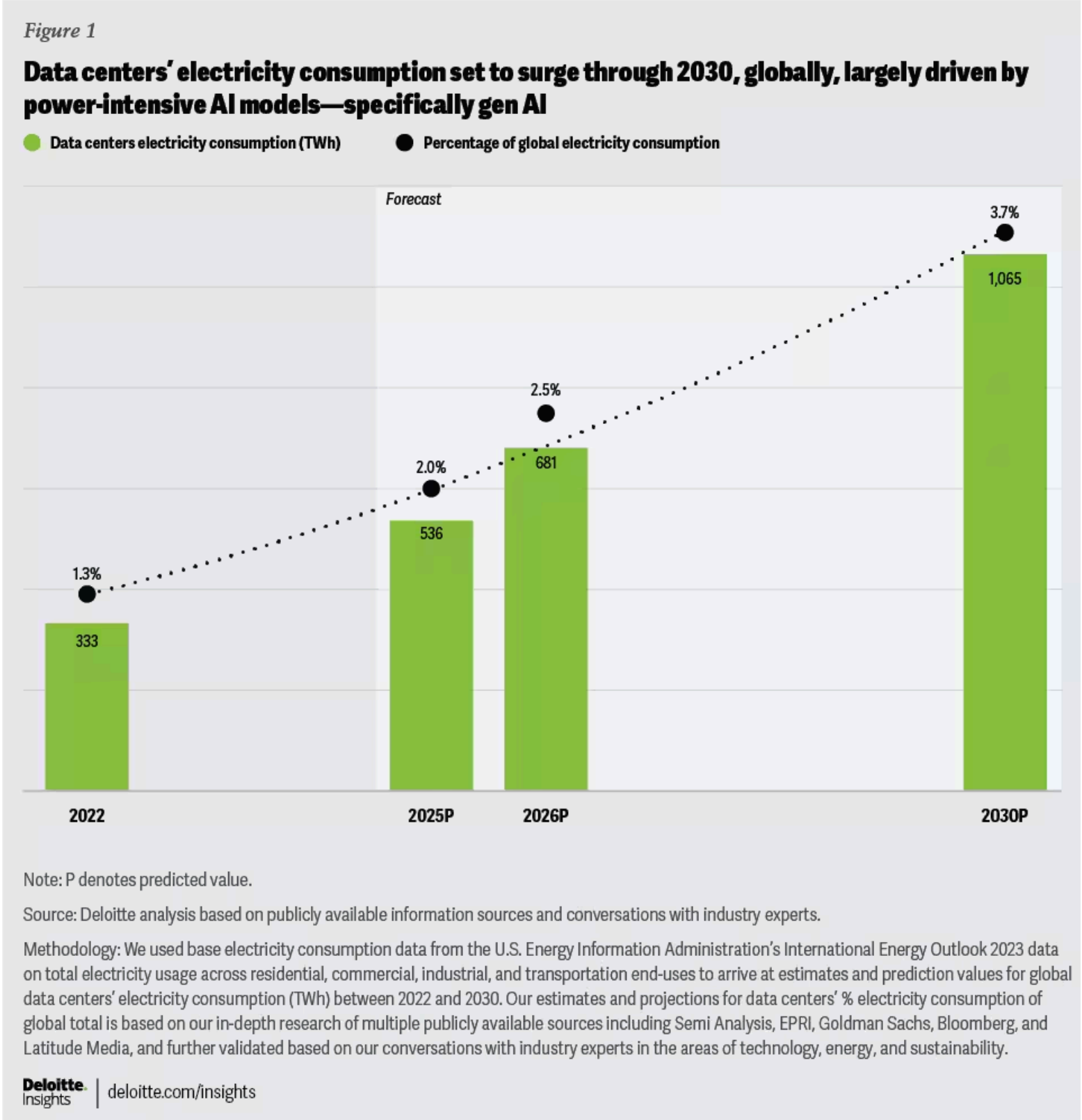
The tech industry should optimize infrastructure, rethink chip design, and collaborate with electricity providers to help secure a more sustainable future for data centers

ARTICLE • 14 MINUTE READ

AI-driven data center power consumption will continue to surge, but data centers are not—in fact—that big a part of global energy demand. Deloitte predicts data centers will only make up about 2% of global electricity consumption, or 536 terawatt-hours (TWh), in 2025. But as power-intensive generative AI (gen AI) training and inference continues to grow faster than other uses and applications, global data center electricity consumption could roughly double to 1,065 TWh by 2030 (figure 1).¹ To power those data centers and reduce the environmental impact, many companies are looking to use a combination of innovative and energy-efficient data center technologies and more carbon-free energy sources.

Nonetheless, it's an uphill task for power generation and grid infrastructure to keep pace with a surge in electricity demand from AI data centers. Electricity demand was already growing fast due to electrification—the switch from fossil-fueled to electric-powered equipment and systems in the transport, building, and industrial segments—and other factors. But gen AI is an additional, and perhaps, an unanticipated source of demand. Moreover, data centers often have special requirements as they need 24/7 power supply with high levels of redundancy and reliability, and they're working to have it be carbon-free.

Estimating global data centers' electricity consumption in 2030 and beyond is challenging, as there are many variables to consider. Our assessment suggests that continuous improvements in AI and data center processing efficiency could yield an energy consumption level of approximately 1,000 TWh by 2030. However, if those anticipated improvements do not materialize in the coming years, the energy consumption associated with data centers could likely rise above 1,300 TWh, directly impacting electricity providers and challenging climate-neutrality ambitions.² Consequently, driving forward innovations in AI and optimizing data center efficiency over the next decade will be pivotal in shaping a sustainable energy landscape.



Some parts of the world are already facing issues in generating power and managing grid capacity in the face of growing electricity demand from AI data centers.³ Critical power to support data centers' most important components—including graphics processing unit (GPU) and central processing unit (CPU) servers, storage systems, cooling, and networking switches—is expected to nearly double between 2023 and 2026 to reach 96 gigawatts (GW) globally by 2026; and AI operations alone could potentially consume over 40% of that power.⁴ Worldwide, AI data centers' annual power consumption is expected to reach 90 terawatt-hours by 2026 (or roughly one-seventh of the predicted 681 TWh of all data centers globally), roughly a tenfold increase from 2022 levels.⁵ As such, gen AI investments are fueling demand for so much electricity that in the first quarter of 2024, global net additional power demand from AI data centers was roughly 2 GW, an increase of 25% from the fourth quarter of 2023 and more than three times the level from the first quarter of 2023.⁶ Meeting data center power demand can be challenging because data center facilities are often geographically concentrated (especially in the United States) and their need for 24/7 power can burden existing power infrastructure.⁷

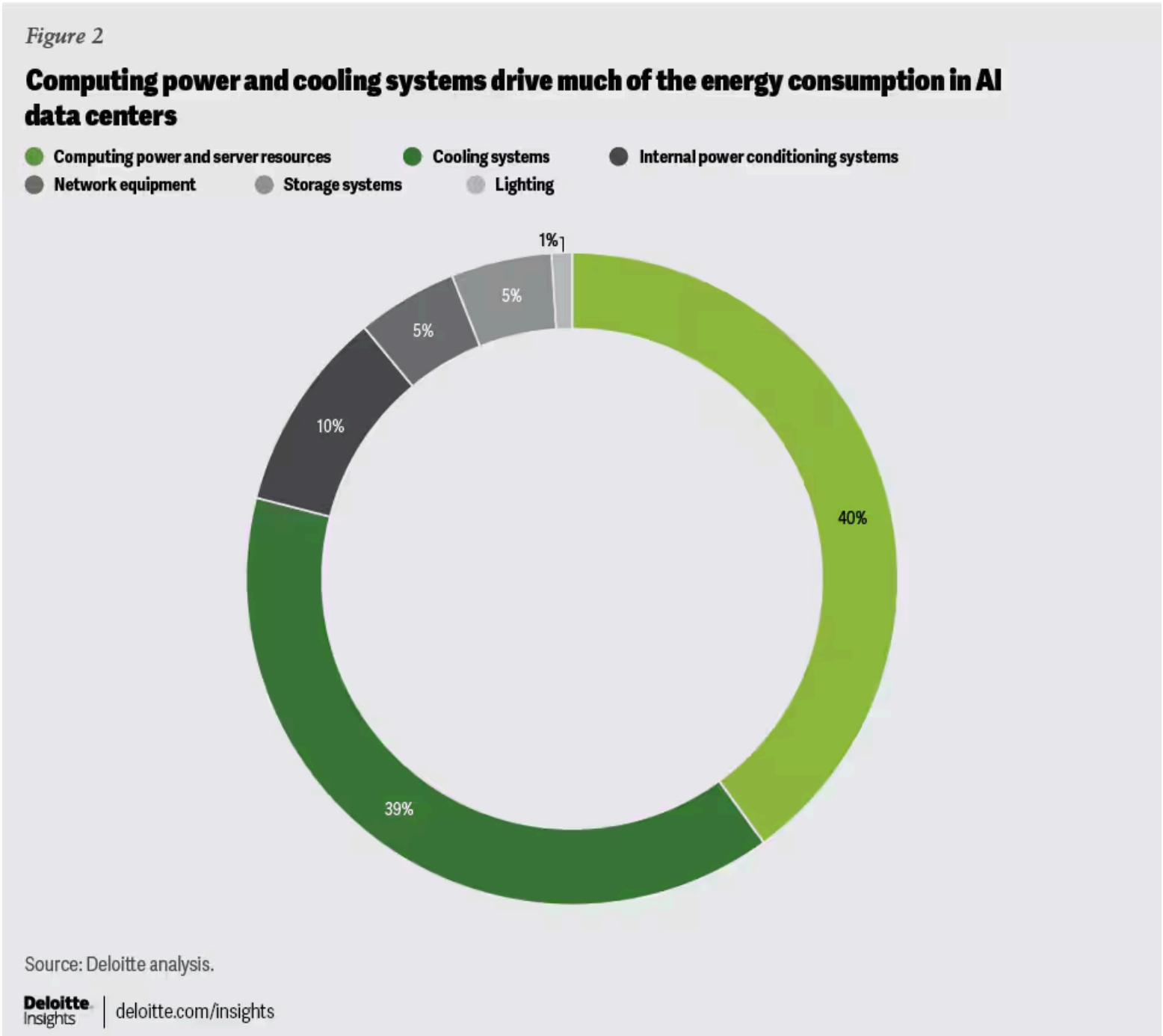
Deloitte predicts that both the technology and electric power industries can and will jointly address these challenges and contain the energy impact of AI—more specifically, gen AI. Already, many big tech and cloud providers are investing in carbon-free energy sources and pushing for net-zero targets,⁸ demonstrating their commitment to sustainability.

Hyperscalers plan massive expansion of gen AI data centers to help support growing customer demand

The surge in electricity demand is largely due to hyperscalers’ plans to build out data center capacity, globally.⁹ As AI demand—specifically gen AI—is expected to grow, companies and countries are racing to build more data centers to meet that demand. Governments are also establishing sovereign AI capabilities to maintain tech leadership.¹⁰ The data center real estate build-out has reached record levels based on select major hyperscalers’ capital expenditure, which is trending at roughly US\$200 billion as of 2024, and estimated to exceed US\$220B by 2025.¹¹

Moreover, Deloitte’s State of Generative AI in the Enterprise survey noted that enterprises have been mostly piloting and experimenting until now.¹² But as they experiment with getting value from gen AI, respondents are seeing tangible results and so intend to quickly scale up beyond pilots and proofs of concept. If usage grows as the technology matures, hyperscalers’ and cloud providers’ capital expenditure will most likely remain high through 2025 and 2026.

Two broad areas drive most of the electricity consumption in a data center: computing power and server resources like server systems (roughly 40% data center power consumption) and cooling systems (consume 38% to 40% power). These two are the most energy-intensive components even in AI data centers, and they will continue to fuel data centers’ power consumption. Internal power conditioning systems consume another 8% to 10%, while network and communications equipment and storage systems use about 5% each, and lighting facilities usually use 1% to 2% of power (figure 2).¹³ With gen AI requiring massive amounts of power, data center providers—including the hyperscalers and data center operators—need to look at alternate energy sources, new forms of cooling, and more energy-efficient solutions when designing data centers. Several efforts are already underway.



Gen AI is contributing to increased electricity demand

Data centers' energy consumption has been surging since 2023, thanks to exploding demand for AI.¹⁴ Deploying advanced AI systems requires vast numbers of chips and processing capacity, and training complex gen AI models can require thousands of GPUs.

Hyperscalers and large-scale data center operators that are supporting gen AI and high-performance computing environments require high-density infrastructure to support computing power. Historically, data centers relied mainly on CPUs, which ran at roughly 150 watts to 200 watts per chip.¹⁵ GPUs for AI ran at 400 watts until 2022, while 2023 state-of-the-art GPUs for gen AI run at 700 watts, and 2024 next-generation chips are expected to run at 1,200 watts.¹⁶ These chips (about eight of them) sit on blades placed inside of racks (10 blades per rack) in data centers, and are using more power and producing more heat per square meter of footprint than traditional data center designs from only a few years ago.¹⁷ As of early 2024, data centers typically supported rack power requirements of 20 kW or higher. But the average power density is anticipated to increase from 36 kW per server rack in 2023 to 50 kW per rack by 2027.¹⁸

Total AI computing capacity, measured in floating-point operations per second (FLOPS), has also been increasing exponentially since the advent of gen AI. It's grown 50% to 60% quarter over quarter globally since the first quarter of 2023 and will likely grow at that pace through the first quarter of 2025.¹⁹ But data centers don't only measure capacity in FLOPS, they also measure megawatt hours (MWh) and TWh.

Gen AI's multibillion parameter LLMs and the multibillion watts they consume

Gen AI large language models (LLMs) are becoming more sophisticated, incorporating more parameters (variables that enable AI learning and prediction) over time. From the 100 to 200 billion parameter models that were released initially during 2021 to 2022, current advanced LLMs (as of mid-2024) have scaled up to nearly two trillion parameters, which can interpret and decode complex images.²⁰ And there's competition to release LLMs with 10 trillion parameters. More parameters add to data processing and computing power needs, as the AI model must be trained and deployed. This can further accelerate demand for gen AI processors and accelerators, and in turn, electricity consumption.

Moreover, training LLMs is energy intensive. Independent research of select LLMs that were trained on more than 175 billion parameters of data sets demonstrated that they consumed anywhere between 324 MWh and 1,287 MWh of electricity for each training run ... and models are often retrained.²¹

On average, a gen AI-based prompt request consumes 10 to 100 times more electricity than a typical internet search query.²² Deloitte predicts that if only 5% of daily internet searches, globally, use gen AI-based prompt requests, it would require approximately 20,000 servers (with eight specialized GPU cores in each of the servers) that consume 6.5 kW on an average per server to fulfill the prompt requests, amounting to an average daily electricity consumption of 3.12 GWh and annual consumption of 1.14 TWh²³—which is equivalent to annual electricity consumed by approximately 108,450 US households.²⁴

Data center demand could present challenges and opportunities for power sector transition

The electric power sector was already planning for rising demand: Many in the industry predicted as much as a tripling of electricity consumption by 2050 in some countries.²⁵ But that trajectory has recently accelerated in some areas due to burgeoning data center demand. Previous forecasts in many countries have projected rising power demand due to electrification as well as increasing data center consumption and overall economic growth. But recent sharp spikes in data center demand, which could be just the tip of the iceberg, reveal the growing magnitude of the challenge.²⁶

Round-the-clock, carbon-free electricity that many tech companies seek can be hard to come by, especially in the short term.

This comes against the backdrop of a multi-decade power industry transformation, as electric companies build, upgrade, and decarbonize electric grid infrastructure, and digitalize systems and assets. In many areas, electric companies are also hardening assets against increasingly severe weather and climate events and protecting networks from rising cybersecurity threats.²⁷ Electric power grids in some countries are struggling to keep up with demand, especially for low- or zero-carbon electricity. In the United States, data centers are anticipated to represent 6% (or 260 TWh) of total electricity consumption in 2026.²⁸ The United Kingdom may witness a sixfold growth in electricity demand within a period of just 10 years, largely due to AI.²⁹ In China, data centers—including the ones that power AI—will likely make up 6% of the country's total electricity demand by 2026.³⁰ Data centers could also add to China's pollution problem, since the country's power generation is dominated by coal, which accounted for 61% of its energy use and 79% of its carbon dioxide emissions in 2021.³¹

Some countries that are facing the rising demand for electricity from data centers are responding with regulations. For instance, in Ireland, existing data centers consume a fifth of the country's total electricity consumption and this is only expected to grow further as AI-driven data centers spring up more; households are even lowering their power consumption.³² Temporarily, Ireland halted the construction of new data centers connected to the grid, but has since reversed that position.³³ Like Ireland, even the city of Amsterdam halted new data center construction to support sustainable urban development.³⁴ Singapore announced new sustainability standards for data centers that require operators to gradually increase the overall operating temperatures of their facilities to 26°C or higher. Higher operating temperatures reduce the demand for cooling and lower power consumption, but at the cost of shortening the lifespan of the chips.³⁵

The urgency and geographic concentration of data center demand—and the requirement for 24/7 carbon-free energy—can further complicate the challenge for tech companies and electricity providers, in addition to new demand from electrification, manufacturing, and other sources. The largest data center market globally is in northern Virginia,³⁶ and the local utility, Dominion Energy, expects power demand to grow by about 85% over the next 15 years, with data center demand quadrupling.³⁷ The round-the-clock, carbon-free electricity that many tech companies seek can be hard to come by, especially in the short term. Electricity providers are exploring multiple avenues to help meet demand while maintaining reliability and affordability. In addition to new renewable energy and battery storage, many electricity providers have also announced plans to build natural gas-fired power plants, which are not carbon-free.³⁸ This could potentially make it more challenging to meet utility, state, and even national decarbonization targets.³⁹

Despite being poised to consume massive amounts of clean energy, AI could also potentially help hasten the clean energy transition: Some utilities are already using AI to enable electric grids to operate more cheaply, efficiently, and reliably through improved weather and load forecasting, enhanced grid management and renewable asset performance, faster storm recovery, better wildfire risk assessment, and more.⁴⁰

Data center cooling is water-intensive

Next-generation CPUs and GPUs have higher thermal density properties than their predecessors. At the same time, some server vendors are packing more and more power-hungry chips into each rack in an endeavor to cater to the growing demand for high-performance computing and AI applications. But denser racks will demand more water, especially to cool the gen AI chips. AI data centers' freshwater demand could be as much as 1.7 trillion gallons (at the higher end) by 2027.⁴¹ A hyperscale data center that intends to manage excess heat with air-based cooling and evaporated drinking water would require over 50 million gallons of water every year (or roughly what it takes to make 14,700 smartphones).⁴² This water cannot be returned to the aquifer, reservoir, or water supply where it came from.⁴³

Air-based cooling alone uses up to 40% of a typical data center's electricity consumption. Therefore, data centers are looking at alternatives to traditional air-based cooling methods, mainly into liquid cooling, as its higher thermal transfer properties could help cool high-density server racks and enable them to reduce power usage by as much as 90% when compared with air-based methods.⁴⁴ As liquid cooling directly delivers cooling to server racks, it can support dense power racks on the order of 50 kW to 100 KW or more.⁴⁵ Moreover, it may help eliminate the need for chillers, which were traditionally used for producing the cooling water.

However, despite liquid cooling technology's promise to help save energy across the data center stack,⁴⁶ it's still in its early days and is yet to be widely adopted or integrated into AI data centers, globally.⁴⁷ Moreover, water is a finite resource, and therefore its cost and availability will likely affect future decisions about its usage.

The tech industry is moving toward more sustainable solutions and carbon-free sources

To help expedite the move toward using carbon-free sources to power AI data centers, tech industry majors continue to be aggressive in their pursuit of renewable energy by way of power purchase agreements, or long-term contracts with renewable energy providers.⁴⁸ These deals have helped bankroll renewable energy projects by enabling them to secure financing. In some cases, technology companies are working with electricity providers and innovators to help test and scale promising energy technologies, including advanced geothermal, advanced wind and solar technologies, hydropower, and even underwater data centers.

In some areas, local grid constraints and long interconnection times for new renewable and battery storage facilities are causing delays in connecting these resources to the electric grid.⁴⁹ These delays, which can be as long as five years in the United States, are often due to high demand and insufficient transmission infrastructure. As a result, tech companies are increasingly pursuing onsite, sometimes off-grid, energy solutions.⁵⁰ Additionally, they are investing in new technologies such as long-duration energy storage and small modular nuclear reactors to help address these challenges. In some cases, tech companies and utilities are planning to coordinate to bring innovative clean energy technologies to scale, which could eventually benefit other organizations and help decarbonize the electric grid faster.⁵¹ Many of these research and development programs, pilots and other clean energy investments may take years before reaping benefits, demonstrating return on investment, and becoming commercially viable.⁵² For example, small, modular nuclear reactors are still in early development stages, and may not be a near-term zero-carbon solution.⁵³

The technology sector consistently dominates US corporate renewable procurement and accounted for more than 68% of the nearly 200 deals of associated contracted capacity tracked over the 12 months prior to Feb. 28, 2024.⁵⁴ Similarly, hyperscalers and data center operators in India are increasingly using solar to power their data centers in the region.⁵⁵ Without these purchase commitments, many renewable energy projects would not be built.⁵⁶

As such, the tech industry's role in bankrolling clean energy technologies to help bring them to scale will continue to be valuable. In some cases, they're working directly with innovators and renewable energy producers, and in other cases, they're partnering with utilities.⁵⁷ Importantly, the way tech companies inject capital to help advance the clean energy transition is critical, as neither the innovators nor the power industry would typically have the level of financial resources that the tech industry possesses.

Bottom line

What should the broader tech industry, hyperscalers, data center operators, utilities, and regulators do to help meet gen AI demand sustainably? Several considerations for hyperscalers and the broader tech industry are more or less in line with what we presented in Deloitte Global's 2021 prediction on cloud migration.⁵⁸ Though demand drivers may have changed and the pace of change has accelerated, the industry is working to achieve balance between sustainability and expedited time to market, while keeping data centers rising energy demands under control and finding more sustainable ways to power AI—specifically gen AI.

Do we need to keep racing to build bigger and bigger foundational models (for example, more than a trillion parameter models), or are smaller models sufficient, while being more sustainable?

1. Make gen AI chips more energy-efficient: Already, a new generation of AI chips can perform AI training in 90 days, consuming 8.6 GWh. This would be less than one-tenth the energy that the previous generation of chips takes to do the same function on the same data.⁵⁹ Chip companies should continue to work with the broader semiconductor ecosystem to help intensify focus on improving FLOPS/watt performance, such that, future chips can train AIs several times larger than the largest AI systems currently available while using less electricity.

2. Optimize gen AI uses and shift processing to edge devices: This includes assessing whether it's energy-efficient to do training and inference in the data center, or on the edge device, and accordingly rebalance data center equipment needs. Edge not only can support applications where response times are critical, but also for those use cases where sensitive data is involved, or privacy needs are high. It also helps save network and server bandwidth, rerouting gen AI workloads to local and near-location or co-location devices, while only transmitting select AI workloads to data centers.⁶⁰

3. Implement changes in gen AI algorithms and rightsize AI workloads: Do we need to keep racing to build bigger and bigger foundational models (for example, more than a trillion parameter models), or are smaller models sufficient, while being more sustainable? Already, startups are developing on-device multimodal AI models, which do not require energy-intensive computations in the cloud.⁶¹ Customers should fine-tune and adjust the size of their AI workloads and go for targeted gen AI models (including preexisting models and training only when needed) based on real business needs, which can minimize energy use. Additionally, depending on specific needs with AI inferencing (for example, doing inference in real time and when latency is critical), CPUs can be more advantageous and efficient.⁶²

4. Form strategic partnerships to serve local and cluster-level AI data center needs: For several small to midsize organizations (including universities) that may find it hard to tap into gen AI data center capacity, those organizations should work with specialized data center operators and cloud service providers that focus on delivering high-performance computing solutions for smaller high-performance computing GPU-cluster co-locations.⁶³ A corollary: Data centers can then actively track usage and availability for potential opportunities and demand pockets to help deliver near-term co-location services.

5. Collaborate with other stakeholders and sectors to make an overall positive environmental impact: The various ecosystem players—including hyperscalers, their customers, third-party data center operators and co-location service providers, electricity providers, the local regulators and municipalities, and the real estate firms—should have ongoing conversations around what’s feasible and viable for the business, environment, and society.⁶⁴ That collaboration should encompass multiple aspects including determining potential strategic co-location needs (where a data center company rents computing and server resources to one or more companies), assessing cooling needs such as adequate temperatures in liquid cooling systems, identifying solutions to manage heat and wastewater, and figuring out recycling needs. For example, in Europe, data center operators are directing waste heat to warm swimming pools in the vicinity.⁶⁵ Electricity providers should consider working more closely with the tech industry to understand how to meet data center energy demand, while identifying ways the tech companies could potentially help fund and scale new energy technologies, which is a vital step to bring more clean energy to the grid.

The holistic efforts of hyperscalers and electricity providers to help increase the use of carbon-free sources to power data centers—including the ones being built exclusively for gen AI—may bear fruit in the longer term.

By	Karthik Ramachandran India	Duncan Stewart Canada
	Kate Hardin United States	Gillian Crossan United States

Endnotes

1. Deloitte analysis based on publicly available information sources and conversations with industry specialists. We used base electricity consumption data from the US Energy Information Administration's (EIA) International Energy Outlook 2023 data on total electricity usage across residential, commercial, industrial, and transportation end uses (a reference to US Energy Information Administration, "[Table: Delivered energy consumption by end-use sector and fuel](#)," accessed Nov. 4, 2024) to arrive at estimates and prediction values for global data centers' electricity consumption (TWh) between 2022 and 2030. Our estimates and projections for data centers' percent electricity consumption of global total are based on our research of multiple publicly available sources including SemiAnalysis, EPRI, Goldman Sachs, Bloomberg, and Latitude Media, and further validated based on our conversations with subject matter specialists in the areas of technology, energy, and sustainability. Total energy consumption by end-use sector and fuel (as noted from the aforementioned table from EIA's International Energy Outlook 2023 data), globally, is estimated and forecast at 26,787 TWh in 2025, 27,256 TWh in 2026, and 29,160 TWh in 2030— increasing from 25,585 TWh back in 2022.
2. As noted in endnote 1 above, we arrived at 2022 to 2030 data, estimates, and predictions based on a combination of in-depth secondary research of multiple publicly available sources, and validated further from our discussions with subject matter specialists. Also, see Prof. Dr. Bernhard Lorentz, Dr. Johannes Trüby, and Geoff Tuff, "[Powering artificial intelligence](#)," Deloitte Global, November 2024."
3. One-fifth of Ireland's electricity is consumed by data centers, and this is expected to grow, even as households are lowering their electricity use. To read further, see: Chris Baraniuk, "[Electricity grids creak as AI demands soar](#)," BBC, May 21, 2024.
4. Dylan Patel, Daniel Nishball, and Jeremie Eliahou Ontiveros, "[AI data center energy dilemma: Race for AI data center space](#)," SemiAnalysis, March 13, 2024.
5. Ibid.
6. Data center BMO report, Communications Infrastructure, "1Q24 data center leasing: Records are made to be broken," April 28, 2024; Moreover, due to strong demand from cloud providers and AI workloads, the data center primary market supply in the United States alone was up 26% year over year to 5.2 GW in 2023, and more are under construction. See further: CBRE, "[North America data center trends H2 2023](#)," March 6, 2024.
7. Lisa Martine Jenkins and Phoebe Skok, "[Mapping the data center power demand problem, in three charts](#)," Latitude Media, May 31, 2024.
8. Based on our analysis of multiple publicly available information and reports from what companies self-report, and further validated from third-party sources.
9. For context, hyperscalers are large cloud service providers and data centers that offer huge amounts of computing and storage resources typically at enterprise scale. See: Synergy Research Group, "[Hyperscale operators and colocation continue to drive huge changes in data center capacity trends](#)," Aug. 7, 2024.
10. Yifan Yu, "[AI's looming climate cost: Energy demand surges amid data center race](#)," Nikkei Asia, June 12, 2024.

11. Data center BMO report, Communications Infrastructure, “1Q24 data center leasing: Records are made to be broken,” April 28, 2024. Further, Deloitte analysis based on information from select tech companies’ publicly available sources such as earnings releases and Dell’Oro Group’s market research data on data center IT capital expenditure shows that if we consider the capital expenditure spending of other data center providers, including third-party operators and outsourced cloud service providers, data centers’ aggregate capital expenditure spending could be at least US\$250 billion in 2025. See: Baron Fung, “[Market research on data center IT capex](#),” Dell’oro Group, accessed Nov. 4, 2024.
12. Nitin Mittal, Costi Perricos, Brenna Sniderman, Kate Schmidt, and David Jarvis, “[Now decides next: Getting real about generative AI](#),” Deloitte’s State of Generative AI in the Enterprise quarter two report, Deloitte, April 2024.
13. Deloitte analysis based on publicly available research reports including: Wania Khan, Davide De Chiara, Ah-Lian Kor, and Marta Chinnici, “[Advanced data analytics modeling for evidence-based data center energy management](#),” *Physica A* 624, 2023; Kazi Main Uddin Ahmed, Math H. J. Bollen, and Manuel Alvarez, “[A review of data centers energy consumption and reliability modeling](#),” in *IEEE Access* 9, 2021: pp. 152536–152563.
14. Tom Dotan and Asa Fitch, “[Why the AI industry’s thirst for new data centers can’t be satisfied](#),” *The Wall Street Journal*, April 24, 2024.
15. Noam Broussard, “[Examining the impact of chip power reduction on data center economics](#),” *Semiconductor Engineering*, March 12, 2024.
16. Based on our analysis of multiple publicly available sources including: Michael Studer, “[The energy challenge of powering AI chips](#),” *Robeco*, Nov. 6, 2023; Agam Shah, “[Generative AI to account for 1.5% of world’s power consumption by 2029](#),” *HPCwire*, July 8, 2024.
17. From our study and analysis of select gen AI data center chip solutions offered by major AI chip vendors, further corroborated with publicly available third-party sources including: Beth Kindig, “[AI power consumption: Rapidly becoming mission-critical](#),” *Forbes*, June 20, 2024.
18. Jones Lang LaSalle, “[Data centers 2024 global outlook](#),” Jan. 31, 2024; Doug Eadline, “[The gen AI data center squeeze is here](#),” *HPCwire*, Feb. 1, 2024; Per IDC, besides graphics processing unit, servers, data centers also need to grapple with a corresponding growth in storage capacity, which is likely to double between 2023 and 2027 to reach 21 zettabytes in 2027. See: John Rydning, “[Worldwide Global StorageSphere forecast, 2023 to 2027: Despite decreased petabyte demand near term, the installed base of storage capacity continues to grow long term](#),” IDC Corporate, May 2023.
19. Patel, Nishball, and Eliahou Ontiveros, “[AI data center energy dilemma](#).”
20. Sean Michael Kerner, “[What are large language models?](#)” *TechTarget*, May 2024; Yu, “[AI’s looming climate cost](#).”
21. Alex de Vries, “[The growing energy footprint of artificial intelligence](#),” *Joule* 7, no. 10 (2023): pp. 2191–2194.

22. Eren Çam, Zoe Hungerford, Niklas Schoch, Francys Pinto Miranda, and Carlos David Yáñez de León, “[Electricity 2024: Analysis and forecast to 2026 report](#),” International Energy Agency, accessed Nov. 4, 2024.
23. Deloitte analysis based on publicly available reports and sources including: de Vries “[The growing energy footprint of artificial intelligence](#),” pp. 2191–2194.
24. Deloitte analysis based on data related to energy use and electricity consumption in homes in the United States. See: US Energy Information Administration, “[Use of energy explained](#),” accessed Dec. 18, 2023.
25. Darren Sweeney, “[Utility execs prepare for ‘tripling’ of electricity demand by 2050](#),” S&P Global, April 19, 2023.
26. Robert Walton, “[US electricity load growth forecast jumps 81% led by data centers](#),” Utility Dive, Dec. 13, 2023.
27. Aaron Larson, “[How utilities are planning for extreme weather events and mitigating risks](#),” POWER, March 13, 2024.
28. Çam, Hungerford, Schoch, Miranda, and de León, “[Electricity 2024](#).”
29. Baraniuk, “[Electricity grids creak as AI demands soar](#).”
30. Yu, “[AI’s looming climate cost](#).”
31. Data on China’s energy use and CO2 emissions sourced from International Energy Agency, accessed September 25, 2024. See: International Energy Agency, “[China’s energy use](#),” accessed Nov. 4, 2024; International Energy Agency, “[China’s CO2 emissions](#),” accessed Nov. 4, 2024.
32. Baraniuk, “[Electricity grids creak as AI demands soar](#).”
33. Paul O’Donoghue, “[Build it and they will hum: What next for Ireland and data centers?](#)” *The Journal*, Sept. 2, 2024.
34. Hosting Journalist, “[City of Amsterdam puts halt to new data center construction](#),” Dec. 21, 2023.
35. With every 1C increase, operators could save 2% to 5% on the energy they use for cooling equipment. To read further, see: Inno Flores, “[Singapore unveils green data center road map amid AI boom that strains energy resources](#),” Tech Times, May 30, 2024.
36. Julie R. Peasley, “[Ranked: Top 50 data center markets by power consumption](#),” Visual Capitalist, Jan. 10, 2024.
37. Whitney Pipkin, “[Energy demands for Northern Virginia data centers almost too big to compute](#),” Bay Journal, June 18, 2024.
38. Zach Bright, “[Southeast utilities have a ‘very big ask’: More gas](#),” E&E News, Jan. 22, 2024.

39. Ibid.
40. Robert Walton, “*AI is enhancing electric grids, but surging energy use and security risks are key concerns*,” Utility Dive, Oct. 23, 2023.
41. Karen Hao, “*AI is taking water from the desert*,” *The Atlantic*, March 1, 2024.
42. Deloitte analysis based on publicly available information sources including: Jennifer Billock, “*Photos: How much water it takes to create 30 common items*,” North Shore News, Jan. 19, 2023.
43. Hao, “*AI is taking water from the desert*”; One case in point is China—where its data centers’ annual water consumption is expected to increase from around 1.3 billion cubic meter as of 2023 to over 3 billion cubic meter by 2030. To read further, see: Yu, “*AI’s looming climate cost*.”
44. Eadline, “*The gen AI data center squeeze is here*.”
45. Diana Goovaerts, “*Data center operators want to run chips at higher temps. Here’s why*,” Fierce Network, June 11, 2024.
46. Scott Wilson, “*Is immersion cooling the answer to sustainable data centers?*” Ramboll, Dec. 13, 2023.
47. David Eisenband, “*100+ kW per rack in data centers: The evolution and revolution of power density*,” Ramboll, March 13, 2024; Direct-to-chip cooling (also known as cold plate liquid cooling or direct liquid cooling) cools down servers by distributing heat directly to server components, while, immersion cooling involves submerging servers and components in a liquid dielectric coolant that also helps prevent electric discharge.
48. Based on Deloitte’s analysis of developments and announcements from select major cloud hyperscalers and tech companies— and information gathered from publicly available sources (time period: 2023 and first three quarters of 2024).
49. Joseph Rand, Nick Manderlink, Will Gorman, Ryan Wiser, Joachim Seel, Julie Mulvaney Kemp, Seongeun Jeong, and Fritz Kahrl, “*Queued up: 2024 edition*,” Lawrence Berkeley National Laboratory, April 2024.
50. Based on Deloitte’s analysis of developments and announcements from select major cloud hyperscalers and tech companies— and information gathered from publicly available information sources between the first quarter of 2023 and the third quarter of 2024.
51. Julian Spector, “*Duke Energy wants to help Big Tech buy the 24/7 clean energy it needs*,” Canary Media, June 11, 2024.
52. For example, it’s not easy to submerge and drop a 1,300-ton data center unit underwater, especially since it demands special equipment to withstand pressure and corrosion caused by seawater. Moreover, there are concerns related to its impact on marine life.

53. David Schlissel and Dennis Wamsted, “*Small modular reactors: Still too expensive, too slow, and too risky*,” Institute for Energy Economics and Financial Analysis, May 2024.
54. Deloitte's analysis of data and information gathered from multiple reports from S&P Global Market Intelligence, published during March and August 2024.
55. Manish Kumar, “*India’s data center boom opens up a fresh segment for green developers*,” Saur Energy International, July 1, 2024.
56. Naureen S. Malik and Bloomberg, “*With AI forcing data centers to consume more energy, software that hunts for clean electricity across the globe gains currency*,” *Fortune*, Feb. 25, 2024.
57. Based on Deloitte’s analysis of developments and announcements from select major cloud hyperscalers, tech companies, and power and utility players—on publicly available information sources during 2023 and the first three quarters of 2024.
58. Duncan Stewart, Nobuo Okubo, Patrick Jehu, and Michael Liu, “*The cloud migration forecast: Cloudy with a chance of clouds*,” *Deloitte Insights*, Dec. 7, 2020.
59. Wylie Wong, “*Nvidia launched next-generation Blackwell GPUs amid AI ‘arms race’*,” Data Center Knowledge, March 19, 2024; For instance, Nvidia notes that it can train a very large AI model using 2,000 Grace Blackwell chips in 90 days, consuming 4 MW power. In comparison, it would take as much as 8,000 of the previous generation chips to do the same work within the same time, consuming 15 MW power.
60. To read further, see section “Generative AI comes to the enterprise edge: ‘On prem AI’ is alive and well” in our 2025 TMT Predictions chapter on “*Updates*”; Additionally, see: Sabuzima Nayak, Ripon Patgiri, Lilapati Waikhom, and Arif Ahmed, “*A review on edge analytics: Issues, challenges, opportunities, promises, future directions, and applications*,” *Digital Communications and Networks* 10, no. 3 (2024): pp. 783–804.
61. Yu, “*AI’s looming climate cost*.”
62. Luke Cavanagh, “*GPUs vs. CPUs in the context of AI and web hosting platforms*,” Liquid Web, Aug. 20, 2024.
63. Eadline, “*The gen AI data center squeeze is here*.”
64. Goovaerts, “*Data center operators want to run chips at higher temps. Here’s why*.”
65. Baraniuk, “*Electricity grids creak as AI demands soar*.”

Acknowledgements

The authors would like to thank **Dilip Krishna, Marlene Motyka, Jim Thomson, Adrienne Himmelberger, Thomas Schlaak, Freedom-Kai Phillips, Johannes Truby, Clement Cabot, Negina Rood, Ankit Dhameja, Suzanna Sanborn, and Akash Chatterji** for their contributions to this article.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

Ambitious stadium projects aim to bridge public-private investment goals

Sports owners transform stadiums into destinations, driving socioeconomic growth, community engagement, and revenue diversification

ARTICLE • 10 MINUTE READ

Investment in sports infrastructure such as stadiums, playing grounds, and training facilities, is seeing an upward trend, as these developments often instigate broad social and economic returns to both the public and private sectors. This trend has seen a renewed focus in recent years, as sports teams in regions such as North America, Europe, and Asia Pacific invest heavily in developing infrastructure. With growth as a common goal, governments and communities can work with sports investors to provide supplementary infrastructure like transportation links, and community resources can help supercharge the socioeconomic impact of sports. In the next year, multiple infrastructure projects are expected to take place that could increase economic returns for communities, as well as sports becoming further embedded in the heart of culture and society.

In 2025, Deloitte predicts that over 300 global sports stadiums will have begun renovations or new builds. Almost 50% of these new stadium infrastructure projects are expected to take place across North America and Europe, according to Deloitte analysis of sports infrastructure developments taking place. The growth of stadium investment across Europe, primarily focused on soccer stadiums, attempts to attract a new wave of fans and could provide revenue diversification opportunities for the organizations setting out on these programs. In doing so, stadium developments can help reach targets for maximizing return on investment for private investors as well as accelerating socioeconomic objectives for the public sector. Stadium investment is also likely to increase across multiple global regions, as sports-led regeneration projects take place and fans increasingly demand innovative touchpoints both inside and outside the stadium.

A community-centric approach to venue development

Sports organizations can bring communities together, improve feelings of civic pride and cohesion, and further diversify the cultural offerings of a city. Delivering a sports-led regeneration program using sports venues as anchors often requires collaboration with governments and other key stakeholders to help deliver strategic initiatives; engage with the community; and develop a sustainable, thriving development where people want to live and visit.

The development of sports stadiums may no longer be able to hold focus on a single club's interests. In developing new and enhanced stadiums, community benefits should infiltrate various aspects of decision-making.

In April 2024, Knighthed Capital Management, owners of now English Football League One club Birmingham City, announced its plans to create a Sports Quarter anchored around a new world-class stadium.¹ Chairman Tom Wagner outlined his vision for the ambitious project to include the stadium, men’s and women’s training grounds, and academy teams all in one location within walking distance of the city center.² Furthermore, he’s alluded to communications with hotels and other commercial entities that are interested in being based in the site and part of the regeneration of east Birmingham.³ Through this, the “Blues” will be fully embedded as part of the city and act as a “beacon for excellence” that is recognized worldwide, according to Wagner.⁴ The Sports Quarter project is anticipated to cost between £2 billion and £3 billion, with hopes to spur long-term socioeconomic impact for the West Midlands community.⁵ The Knighthed ownership group is engaged with government and public sectors on a range of strategic priorities.⁶

In Major League Baseball, the Tampa Bay Rays signed a deal in July 2024 with the City of St. Petersburg, Florida, to build a new ballpark.⁷ The development group for the new stadium pledged to create 1,250 units of affordable housing, 30,000 construction jobs, and 7,000 permanent jobs, with some reserved for local and historically disadvantaged residents.⁸ The project leaders have reiterated their commitment to bridging the generational wealth gap in the surrounding communities with this new venue, considering the project a “failure” if they do not reach that objective.⁹

Enhancing fan engagement by meeting generational preferences

Fans of different generations are consuming sports in different ways. According to Deloitte UK’s [The Future of Sport 2024](#), 84% of global sports leaders surveyed said they expected different consumption preferences to be one of the most impactful next-generation trends over the next five years. Sports organizations should balance embracing the core traditions of a game-day experience with Generation Z and Generation Alpha’s heightened expectations of entertainment.¹⁰

As a first step in developing the stadium experience, organizations should consider the basics: comfort and safety, view, quality on-field product, and an exciting atmosphere. These attributes are important for many fans and should be mastered before looking at any advanced plans.

After establishing these foundations, some organizations may look to differentiate their experiences by providing end-to-end entertainment options for fans before, during, and after the game. Not only can this help enable greater spending of both time and money for fans at these stadiums, but it can also foster a greater sense of community around the sports organizations. By integrating community culture into the fabric of the stadium, the game-day experience can be uniquely localized.

The Toronto Blue Jays embarked on new renovations on their stadium at the Rogers Centre, with two phases of work aimed at elevating the fan experience. Phase one, completed in 2023, unveiled five distinct outfield “neighborhoods” within the ballpark stands, encompassing local cuisine and entertainment, representing differentiated fan experiences, and providing social spaces in each.¹¹ The renovations also included upgraded digital technology such as “Tap N Go,” a new automated market service for food and beverage, as well as “Walk Thru Bru,” a self-serve beverage-concession stand to speed up service times.¹² Phase two of the Rogers Centre renovation includes fan-centric adjustments such as angling seats toward home plate for improved sightlines.¹³

The smart stadium

The next generation tends to consume sports in a digital-first manner, and their game-day experiences are often no different.¹⁴ Some sports organizations are designing “smart stadium districts” that integrate advanced technology to personalize the fan experience.¹⁵ The global smart stadium market is growing, with a 2024 market size of over US\$8 billion, which is expected to reach more than US\$38 billion by 2033.¹⁶

The requirements for fan engagement are changing, as Gen Z and Gen Alpha prefer shorter, more dynamic content.¹⁷ Experiences are becoming a differentiator for organizations, as fans increasingly want value and will pay for greater experiences over material objects.¹⁸ New builds are incorporating elements of gaming, merchandising operations, and designs that consider “second-screen syndrome,” in which the majority of fans tend to look at secondary screens while watching sports. According to Deloitte research, 77% of sports fans surveyed say they’ve done at least one additional activity related to a game while watching a sporting event, including looking up stats, using social media, or betting.¹⁹ New stadiums are using integrated technology to broadcast these elements, keeping a greater portion of fans’ focus within the stadium.²⁰

The National Basketball Association’s Los Angeles Clippers unveiled their new arena, the Intuit Dome, built to differentiate the fan experience at the forefront. A key highlight of the new arena is the custom-built “Halo Board,” which optimizes sight lines from all seats and prioritizes the viewing experience of upper-bowl seating. The double-sided video board hanging above center court contains a game feed, “coaches corner” with in-depth statistics, instant replays, Steve Cam (keeping tabs on Clippers owner Steve Ballmer), player profile features including photos and other personal promotions such as player foundation information, and more. To enhance the fan experience, the Halo Board will also reinvent the coveted T-shirt toss with T-shirt cannons attached, enabling fans in the upper tiers to be able to receive merchandise as well. The Intuit Dome plans to reach levels of engagement not seen before in sports, as fans will be rewarded for their cheering and provided gaming consoles at each seat for use in game-day entertainment activities, further gamifying fandom.²¹

The next generation of fans will likely expect personalized, seamless, and on-demand experiences. Stadium districts can incentivize fans to stay longer and enjoy different offerings including food, music and culture, as well as social spaces for different types of fans. Sports organizations are building these districts to give local and visiting fans a premium experience and generating new touchpoints for them while they are in the area.²² New technologies are streamlining purchasing by making it quicker to purchase, such as click and collect merchandise or food and beverage, as well as ticket software that can personalize messaging for each fan.²³ Additionally, fans are entering the stadium in new ways, often contributing to a more convenient fan experience. Mercedes-Benz Stadium partnered with Delta to create Fly-Through Lanes, which uses facial recognition to gain quick entry for fans visiting the stadium.²⁴

Embracing infrastructure and tech to help diversify revenue streams

Sports owners are also using enhanced infrastructure and digital technology to further diversify their revenue generation.

The success of sports organizations is evolving as, historically, global sports organizations typically relied heavily on broadcast revenue to supplement spend on wages and other costs.²⁵ Generally, an increase in central broadcast revenue would lead to an increase in salary caps for North American sports leagues and spending allowed under certain cost-control regulations across European sports.²⁶

Organizations seem to recognize that an overreliance on one revenue stream can place a ceiling on earning potential and expose them to market shocks, such as the COVID-19 pandemic. As elite sports clubs hold cultural and commercial capital, some organizations are leveraging their commercial assets to help bolster revenue.²⁷ Whereas ticketing and broadcast revenue are generally capped at capacity or out of scope for individual clubs to negotiate, commercial revenue can be an underutilized lever for organizations to grow within their control.²⁸ The evolution of stadiums into broader entertainment districts can help sports organizations increase their commercial footprint by expanding their offerings. For example, the Premier League’s Tottenham Hotspur has leveraged its new stadium to increase commercial revenue from £72 million in the 2016 to 2017 season to £227 million in the 2022 to 2023 season, supported by non-soccer events such as hosting National Football League games and concerts.²⁹

Recent renovations of European soccer giant Real Madrid's Santiago Bernabéu stadium in 2023 saw the club recognize record revenue in the 2022 to 2023 season, with every business line experiencing growth (outside of broadcasting rights, which are in the middle of a contract with broadcasters).³⁰ In July 2024, the club announced it had generated over €1 billion in revenue, the highest figure ever generated by a soccer club.³¹ The club surpassed the billion-euro mark by initiating new business ventures during the latter part of the financial year, primarily through hosting major events and introducing premium VIP experiences.³² For example, Real Madrid brought in €18 million from Colombian pop singer Karol G's four concerts held at the stadium.³³ The club is expected to finish renovations of the stadium in the upcoming season and is looking to add non-soccer-related income in future years.³⁴

Emerging trends in sports infrastructure

From responding to the growth in women's sports to driving greater sustainability, sports organizations are working to embrace new fan experiences and priorities. This evolution in sports infrastructure aims to create a more inclusive, innovative, and responsible future for the industry.

Infrastructure for women's sports

As women's sports continue to garner attention and rising valuations, organizations are beginning to focus more on women's sport-specific infrastructure. In the National Women's Soccer League, the Kansas City Current christened their US\$117 million riverfront stadium, CPKC Stadium, recognized as the first stadium built specifically for a women's professional sports team.³⁵ Further increasing the socioeconomic impact of the stadium, plans have been approved for a US\$650 million mixed-use development, including apartments, hotels, restaurants, and retailers, to be built alongside the soccer stadium.³⁶ The stadium is expected to generate almost US\$50 million in annual economic impact to Kansas City.³⁷

Across the Women's National Basketball Association, new training facilities are contributing to an improvement in the on-court product and, in turn, the overall valuation of the franchises. The Las Vegas Aces opened their facility in 2023, with the Seattle Storm, Phoenix Mercury, and Chicago Sky all following suit.³⁸

In England, Women's Super League club Brighton & Hove Albion is in the planning phase to build a stadium purpose-built for its women's team.³⁹ Manchester City Women has also had plans approved for a purpose-built training facility at the club's City Football Academy site.⁴⁰

Premium and personalized hospitality offerings

Hospitality may be moving away from traditional corporate offerings and is now often being used as a tool to create more accessible, differentiated experiences for all demographics. Sports organizations are working to partner with higher-end brands in these spaces to offer more premium hospitality services, from celebrity chefs to take-out goodie bags.⁴¹ Formula One provides an example of premium hospitality offerings in its Paddock clubs, engaging with social influencers, celebrities, and brand partners.⁴² As sports organizations renovate and build new stadiums, hospitality spaces can also be used more flexibly, allowing for hosting of different types of events.

Sustainability

New sports infrastructure projects are increasingly incorporating sustainability principles into planning. In the case of sports-led regeneration projects, having a focus on sustainability can help unlock public funding elements by demonstrating positive environmental and social practices.⁴³ While there are community benefits to sustainability, incorporating these elements into infrastructure can also help reduce negative impacts and can become benefits over time, including reduced energy bills. Simultaneously this can improve brand affinity, which can lead to more partnership opportunities and fan engagement. The sports industry is heavily intertwined with climate change as both a contributor and an affected party. Two of the largest contributors to global carbon emissions are large construction projects and transportation,⁴⁴ both of which new stadium developments will contribute to. However, sports are also feeling the consequences of climate change as heatwaves and other extreme weather conditions may negatively affect competitions, host locations, and athlete welfare.⁴⁵

As sports organizations consider these real estate development projects, careful thought around sustainability practices and strategy should be integrated into planning conversations.

Bottom line

Sports organizations across the world are looking to infrastructure development to help increase capacity and enhance the lifetime value of their fans. For private investors and owners, stadium districts can provide an opportunity to diversify revenue, capitalize on stadium usage year-round instead of solely on game days, and contribute to enterprise value. Digital touchpoints can also provide the organization with enhanced fan data, to be able to better personalize and target products.

For public investors and governments, contributing to the development of sports infrastructure projects can contribute to broader community benefit. Sports organizations can work to foster a sense of community, improve health and well-being outcomes, and attract tourist foot traffic.

Elite sports⁴⁶ have emerged as a powerful catalyst for economic and social growth, which can align public and private investment agendas. In the near future, organizations could use their stadiums to help breach the boundaries of sports, entering into wider entertainment and digital offerings.

By	Jennifer Haskel United Kingdom	Pete Giorgio United States
	Alice John United Kingdom	Kevin Westcott United States

Endnotes

1. Alex Dicken, “[Tom Wagner reveals timeline for new Birmingham City stadium as Knighthead pledge billions](#),” Birmingham Live, April 9, 2024.
2. Ibid.
3. Ibid.
4. Alex Dicken, “[Another reason for Tom Wagner’s Birmingham City takeover has now become clear](#),” Birmingham Live, Sept. 26, 2023.
5. Dicken, “[Tom Wagner reveals timeline for new Birmingham City stadium as Knighthead pledge billions](#).”
6. Dicken, “[Another reason for Tom Wagner’s Birmingham City takeover has now become clear](#).”
7. Hines, “[Hines and Tampa Bay Rays gain approval of new ballpark, historic gas plant district development](#),” press release, July 31, 2024.
8. FOX 13 News Staff, “[Tampa Bay Rays, city of St. Pete sign deal to build new ballpark, keeping team in town for 30 years](#),” FOX 13 News, July 31, 2024.
9. Ibid.
10. Jamie Pugh and Zoe Burton, The Future of Sport 2024, Deloitte, Sept. 2, 2024.
11. Major League Baseball, “[Blue Jays showcase all-new 100 level seating bowl at Rogers Centre, as part of multi-year renovations](#),” April 4, 2024.
12. Populous, “[Blue Jays unveil completed outfield district of Rogers Centre renovations, designed by Populous](#),” April 6, 2023.
13. Toronto Blue Jays, “[100 level renovation](#),” accessed Nov. 5, 2024.
14. Pete Giorgio, David Jarvis, Brooke Auxier, Hannah Bobich, and Kat Harwood, “[2023 sports fan insights: The beginning of the immersive sports area](#),” Deloitte Insights, June 26, 2023.
15. Katelin Kharrati, “[Global smart stadium market size likely to expand at a compound annual growth rate of 22.5% by 2033](#),” press release, Custom Market Insights, June 28, 2024.
16. Ibid.
17. Ed Dixon, “[Study: Nine in 10 Gen Z sports fans use social media to consumer content as consumption habits shift](#),” SportsPro, June 28, 2023.
18. Giorgio, Jarvis, Auxier, Bobich, and Harwood, “[2023 sports fan insights](#).”

19. Ibid.
20. Gary Drenik, “*Stadium of the future: Emerging game day technologies for engaging fan experience*,” *Forbes*, Aug. 18, 2022.
21. Ohm Youngmisuk, “*Storms, stats, and T-shirt cannons: LA Clippers’ Halo Board goes all out*,” ESPN, Aug. 16, 2024.
22. Deloitte, *2024 Sports Industry Outlook*, March 12, 2024.
23. Drenik, “*Stadium of the future*.”
24. Mercedes-Benz Stadium, “*Delta Fly-Through Lanes*,” accessed Nov. 5, 2024.
25. Deloitte Sports Business Group, *Annual Review of Football Finance 2024*, June 2024.
26. Bryan Toporek, “*The NBA’s new TV deals are poised to send the salary cap skyrocketing*,” *Forbes*, May 30, 2024.
27. Deloitte Sports Business Group, *Annual Review of Football Finance 2024*.
28. Sports Business Institute Barcelona, “*Commercial revenue: Increasing financial power of football clubs and leagues*,” July 11, 2024.
29. Deloitte Sports Business Group, *Annual Review of Football Finance 2024*.
30. Gavin Hamilton, “*Real Madrid announces record turnover as stadium rebuild nears completion*,” SportBusiness, July 18, 2023.
31. Guillermo Rai, “*Real Madrid surpass €1bn in revenue for 2023 to 2024 season*,” *The Athletic*, July 23, 2024.
32. Real Madrid, “*Real Madrid becomes the first football club to exceed 1 billion euros in revenue*,” July 23, 2024.
33. Conor Laird, “*The staggering sum Real Madrid earned from four Karol G concerts*,” Yahoo Sports, July 24, 2024.
34. Ibid.
35. Kansas City Current, “*CPKC Stadium and University of Kansas Health System Training Center*,” accessed Nov. 5, 2024.
36. Kevin Collison, “*Port KC approves massive project next to KC Current Stadium*,” Flatland, April 23, 2024.

37. Ibid.

38. Women’s National Basketball Association– Las Vegas Aces, “*Home sweet home! Aces take up residence in first-of-its-kind Women’s National Basketball Association practice facility and team headquarters,*” press release, April 29, 2023.

39. Morgan Ofori, “*Brighton ready to spark revolution with women’s football stadium,*” *The Guardian*, Oct. 29, 2023.

40. Simi Iluyomade, “*Manchester City Women are building a new £10 million training facility,*” *Versus*, May 15, 2024.

41. Samuel Agini and Alice Hancock, “*Sports hospitality shifts focus to fans as UK demand for ‘experiences’ grows,*” *Financial Times*, Aug. 13, 2021.

42. Georgina Yeomans, “*How Formula One transformed its hospitality product,*” *BlackBook Motorsport*, Jan. 6, 2022.

43. Deloitte, *The Future of Sport 2023*, April 2023.

44. Center for Climate and Energy Solutions, “*Global emissions,*” accessed Nov. 5, 2024.

45. Directorate-General for Climate Action, “*Sport—a key player in climate action?*” European Union, July 26, 2024.

46. Elite sports are defined as the highest level of competition, which may or may not be classified as “professional” sports where participants are paid for their performance.

Acknowledgements

The authors would like to thank **Tim Bridge, Jeff Harris, James Savage, Brooke Auxier,** and **Dhruv Garg** for their contributions to this article.

Cover image by: **Jaime Austin**

Autonomous generative AI agents: Under development

Autonomous gen AI agents—agentic AI—could increase the productivity of knowledge workers and make workflows of all kinds more efficient. But the “autonomous” part may take time for wide adoption.

ARTICLE • 16 MINUTE READ

Autonomous generative AI agents, referred to as “agentic AI,” are software solutions that can complete complex tasks and meet objectives with little or no human supervision. Agentic AI is different from today’s chatbots and co-pilots, which themselves are often called “agents.” Agentic AI has the potential to make knowledge workers more productive and to automate multi-step processes across business functions. Deloitte predicts that in 2025, 25% of companies that use gen AI will launch agentic AI pilots or proofs of concept, growing to 50% in 2027.¹ Some agentic AI applications, in some industries, and for some use cases, could see actual adoption into existing workflows in 2025, especially by the back half of the year.

Their efforts are being aided by startups and established tech companies developing agentic AI, both of which see the technology’s potential to spur revenue growth. Investors have poured over \$2 billion into agentic AI startups in the past two years, focusing their investment on companies that target the enterprise market.² Meanwhile, many tech companies, cloud providers, and others are developing their own agentic AI offerings. They are also making strategic acquisitions, and increasingly licensing agentic AI technology from startups and hiring their employees, instead of buying the companies outright.³

Agentic AI puts the “agency” in agent

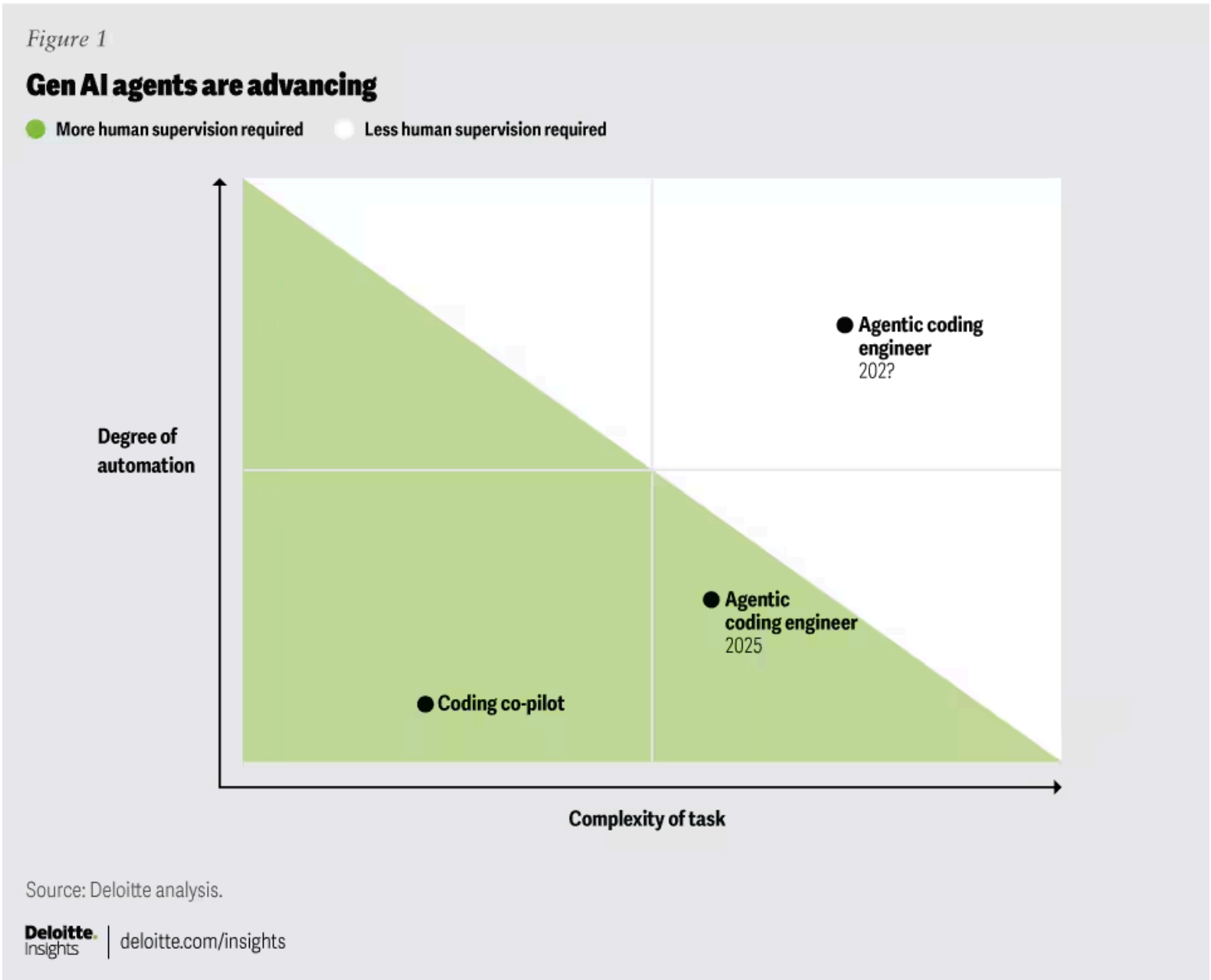
Gen AI chatbots and co-pilots are sophisticated; they can interact intuitively with humans, synthesize complex information, and generate content. But they lack the degree of agency and autonomy that agentic AI promises. While chatbots and agents share the same foundation—large language models (LLM)—additional technologies and techniques enable agents to act independently, break a job down into discrete steps, and complete their work with minimal human supervision or intervention. AI agents don’t just interact. They more effectively reason and act on behalf of the user.

As its name suggests, agentic AI has “agency”: the ability to act, and to choose which actions to take.⁴ Agency implies autonomy, which is the power to act and make decisions *independently*.⁵ When we extend these concepts to agentic AI, we can say it can act on its own to plan, execute, and achieve a goal—it becomes “agentic”.⁶ The goals are set by humans, but the agents determine how to fulfill those goals.

An example can illustrate the difference between agentic AI and co-pilots and chatbots. Co-pilots that assist software developers by testing and suggesting code are one of the most successful gen AI use cases to date.⁷ They can make experienced software engineers more productive and increase the effectiveness of junior coders. They can convert natural language prompts (in multiple languages) into suggestions for code, and test code for consistency. But such co-pilots only respond to prompts from engineers and do not show agency. With agentic AI, the software “engineer” takes this a step further. A human coder can enter ideas for software through a prompt, and the agentic AI “software engineer” converts those ideas into executable code, a process that automates multiple steps in the software development process.

For example, Cognition Software launched “Devin” in March 2024 with the goal of creating an autonomous software engineer capable of reasoning, planning, and completing complex engineering tasks that require thousands of decisions.⁸ Devin was designed to perform programming jobs unassisted, based on natural language prompts from human programmers. These jobs include designing full applications, testing and fixing codebases, and training and tuning LLMs.⁹ Competitors like Codeium, which focuses on enterprise software development, and open-source versions of Devin, hit the market in summer 2024.¹⁰

Agentic AI software engineers share similar capabilities and vulnerabilities.¹¹ One vulnerability is that they currently make too many errors to handle full, or even partial, jobs without human oversight. In a recent benchmarking test, Devin was able to resolve nearly 14% of GitHub issues from real-world code repositories—twice as good as LLM-based chatbots,¹² but not fully autonomous. Big tech companies¹³ and startups are striving to make agentic AI software engineers more autonomous and reliable, so human coders—and their employers—can trust them to handle parts of their workload (figure 1).



Augmenting and amplifying labor productivity

Agentic AI software engineers are just one example of how autonomous generative AI agents could transform how work is done (see [Promising use cases for autonomous gen AI agents](#) below). As agentic AI improves, its impact could be enormous. There are over 100 million knowledge workers in the US, and over 1.25 billion knowledge workers globally.¹⁴ Total factor productivity,¹⁵ a useful proxy for knowledge work, has stagnated in the United States, growing 0.8% from 1987 to 2023 and only 0.5% from 2019 to 2023.¹⁶ In most OECD countries, the story is the same.¹⁷ Attempts to increase the productivity of knowledge work by automating tasks have met with only partial success. Many companies also need more knowledge workers. Shortfalls of customer service representatives, semiconductor engineers, and seemingly everything in between persist. When new workers start, they need to be productive quickly.

Expert systems and robotic process automation (RPA) can falter when processes are ambiguous or require multiple steps. Systems based on traditional machine learning require extensive training, which is tailored for specific purposes. Built on LLMs, agentic AI can be more flexible, and it can address a broader range of use cases than machine learning or deep learning.

Agentic AI can significantly advance the capabilities of LLMs and could vindicate the investments companies are making in gen AI. The public release of gen AI tools has quickly captured the attention of executives. It was easy to imagine how their organizations could use the technology. Quantifiable business value from gen AI has often been elusive, however. Challenges with data foundations, risk and governance policies, and talent gaps make it hard for companies to scale gen AI initiatives.¹⁸ Only 30% of gen AI pilots make it to full production.¹⁹ Lack of trust in gen AI output, and potential “real world” consequences from gen AI mistakes, give executives pause.²⁰

Companies that develop and implement agentic AI need to consider the challenges of gen AI, plus the complexity of building bots that can reason, act, collaborate, and create. Most importantly, gen AI agents of all kinds need to be reliable for enterprises to use them: Getting the job right most of the time isn’t enough. There are some use cases and applications in late 2024 that show encouraging signs of being reliable enough for adoption in early 2025.

The potential payoff is worth the effort, however, and early results seem promising. Companies are learning how to boost LLM performance by combining these models with other AI technologies and training techniques. While autonomous and reliable agents are the goal, incremental increases in accuracy and independence could help companies reach their early productivity and efficiency goals for gen AI overall.²¹ With their range of applications—both horizontal and vertical—and clear business goals, agentic AI looks more like the gen AI solutions executives may have expected in the first place.

Gen AI agents explained: A closer look

Generative AI agents can break down a complex task into a series of steps, execute them, and work through unexpected barriers. They can sense their environment, which depending on the use case can be virtual, physical, or a combination of the two. To complete a task, agentic AI can determine which actions to take, recruit assistance from tools, databases, and other agents, and deliver results based on its goals set by humans.

Agentic AI is an emerging technology—and it continues to evolve—but it has some common characteristics and capabilities:

- **Built on foundation models:** Foundation models like LLMs enable agentic AI to reason, analyze, and adapt to complex and unpredictable workflows. This makes it more flexible than RPA and expert systems. LLMs are improving rapidly, with enhanced reasoning and ability to break tasks into smaller steps among the most recent breakthroughs.²² But foundation models alone can’t interact with their environment, make decisions, or execute tasks.²³ They must be augmented by other technologies and capabilities.

- **Acts autonomously:** While the degree of autonomy varies, agentic AI can be trained to plan and execute complex tasks largely on its own. By introducing reasoning tokens, chain-of-thought models can solve more complex challenges than previous LLMs. Chain-of-thought models are slower to respond but are more deliberative in how they reason through a problem; they can correct their own errors; and they can show the steps they have taken to reach a solution.²⁴
- **Senses the environment:** Agentic AI can perceive the environment, process information, and understand the context of the tasks it is given.²⁵ Advanced agentic AI can process multimodal data, such as videos, images, audio, text, and numbers.
- **Uses tools:** Agentic AI interacts with tools and systems to complete tasks, such as software, enterprise applications, and the internet.
- **Orchestrates:** Agentic AI can direct the participation of other systems and bots to complete a task. With multiagent systems, this means collaborating with other autonomous generative AI agents.
- **Accesses memory:** LLMs are stateless: Each interaction is processed independently, and information is not retained when an interaction is complete. With the addition of retrieval mechanisms and databases, agentic AI can access short-term memory to maintain context while performing a specific task, and long-term memory to learn and improve from experience.²⁶

Startups and big tech are developing multiagent gen AI systems, including tools that can help organizations build their own custom agents.

Some of the latest models employ chain-of-thought functions that, while slower and more deliberative than prior large-scale models, enable higher-order reasoning on complex problems.²⁷ Multimodal data analysis can make agentic AI more flexible by expanding the kinds of data that can be interpreted and produced. Multimodal AI also shows that agentic AI can be even more powerful when combined with other kinds of AI technologies such as computer vision (image recognition), and transcription and translation.²⁸ Like agents themselves, multimodal AI is still developing.

True multiagent systems, in which work is orchestrated among a network of autonomous agents, are being developed now, with some pilots being launched in late 2024.²⁹ Multiagent models often outperform single-model systems by distributing tasks, especially in complex environments.³⁰ Startups and big tech are developing multiagent gen AI systems, including tools that can help organizations build their own custom agents.³¹

Promising use cases for autonomous gen AI agents

Big tech companies and startups are developing early-stage solutions that can partially automate functions like software development, sales, marketing, and regulatory compliance. What follows is a snapshot of today's examples, not an exhaustive list of applications. Some are based on proofs of concept and demos that are promising but are not ready for enterprise deployment. While these examples are cross-industry, industry-specific agentic applications are also emerging.

Customer support: Customer service is an essential—and often stressful—job, with an annual turnover rate of 38%.³² Effective automation of parts of the customer support workflow could reduce stress and tedium for staff, and help companies serve more customers.³³ Agentic AI can handle more complex customer inquiries than today's customer support chatbots, and they can act autonomously to resolve issues. In one example, an audio company is using agentic AI to help customers set up new equipment, a multistep process that usually requires a human agent. If a human agent is required, the agentic AI compiles relevant information and summarizes the issue before transferring the customer.³⁴ The next wave of customer support agents will likely integrate multimodal data such as voice and video in addition to text-based chat.

Cybersecurity: Cybersecurity experts epitomize the shortage of skilled knowledge workers: Globally, there’s a shortfall of four million today.³⁵ Meanwhile, malicious actors are using gen AI to infiltrate cybersecurity systems. Emerging agentic cybersecurity systems can make human experts more efficient by automating aspects of their work. They can autonomously detect attacks and generate reports, improving system security and reducing the workload of human experts by up to 90%.³⁶ Agentic AI can also help software development teams detect vulnerabilities in new code. It can run tests and communicate directly with developers to explain how to fix a problem—something human engineers must do manually today.³⁷

Regulatory compliance: Companies across industries, including financial services and healthcare, are required to conduct periodic regulatory compliance reviews. The increased size and complexity of relevant regulations, and the dearth of compliance professionals, makes compliance a growing challenge. Startups are developing agentic AI that can analyze regulations and corporate documents, and quickly determine whether the company is compliant. The agent can cite specific regulations, and proactively provide analysis and advice to human regulatory professionals.³⁸ Companies that use gen AI today cite regulatory compliance as their top barrier to developing and deploying gen AI, ahead of issues like a lack of AI technical talent, and implementation challenges.³⁹ Regulatory uncertainty plays a role, but so does the reach and complexity of new regulations. By helping companies understand and comply with regulations as they’re enacted, a more agentic AI solution could help accelerate wider gen AI adoption across enterprises.

Agent builders and orchestrators: Agentic AI solutions are emerging to help automate other cross-industry and vertical-specific workflows. Companies may not need to wait for the market, however. They can build their own agents and multiagent systems. With Google’s Vertex, companies can use no-code tools to create agents for specific tasks, such as building marketing collateral based on previous marketing campaigns.⁴⁰ LangChain uses open-source technology to help companies construct multiagentic systems. For example, startup Paradigm has launched a “smart spreadsheet” in which multiple agentic AIs partner to collect data from diverse sources, structure it, and complete tasks.⁴¹

Bottom line

Agentic AI has enormous potential to help increase the productivity of knowledge workers by automating entire workflows and discrete tasks. Its ability to take independent action, as single agents or in concert with other agents, sets it apart from today’s chatbots and co-pilots. Yet, agentic AI is in the early stages of development and adoption. As impressive as early agentic examples may be, these agents can make mistakes and get stuck in loops. In multiagent systems, “hallucinations” can spread from one agent to another; they can persuade other agents to take the wrong steps and give incorrect answers.⁴² Although agentic AI can be mainly autonomous, often having a human review decisions after they’ve been made (also known as “human on the loop” rather than the more restrictive “human in the loop”) can make agentic AI more suitable for deployment today. When gen AI agents get stuck, they can consult human experts who help them resolve the challenge and move forward. In this model, agentic AI is like a junior employee who can learn by experience while performing valuable work.⁴³

Because the vision for agentic AI is compelling and the technology is evolving rapidly, companies should prepare themselves now.

While some companies are investing billions to create consistent and reliable agentic AI, it’s not clear when this will happen, or under what circumstances. Will agentic AI reach widespread adoption in 2025, or within the next five years? Will ubiquity require breakthrough innovation, or tweaking current AI technologies and training methods? If the big companies and startups developing agentic AI are successful, the game will change quickly. Imagine autonomous gen AI agents that can process multimodal data, use tools, orchestrate other agents, remember and learn, and execute tasks consistently and reliably. Imagine further that custom agents can be quickly and easily developed by enterprises in “no-code environments” using just conversational text prompts.

Because the vision for agentic AI is compelling and the technology is evolving rapidly, companies should prepare themselves now. As they prepare, they should consider the following approaches.

Prioritize and redesign workflows for agentic AI: Consider which tasks and workflows are well-suited for agentic AI to execute, based on the technology’s capabilities and where the highest value is for your company. Redesign them to remove unnecessary steps. Ensure that agentic AI solutions have a clear goal, and access to the data, tools, and systems they will require. Although these agents can help other agents navigate their environment, cluttered and sub-optimized processes could deliver disappointing results.

Focus on data governance and cybersecurity: For agentic AI to deliver value, it must have access to valuable and potentially sensitive enterprise data, as well as internal systems and external resources. Companies should put strong data governance and cybersecurity in place before getting started with autonomous generative AI agents. For gen AI early adopters, the top areas where they’re increasing IT investment are data management (75%) and cybersecurity (73%).⁴⁴ Despite these investments, 58% are highly concerned about using sensitive data in models and managing data security. And only 23% say they’re highly prepared for managing gen AI risk and governance. In short, many of today’s gen AI leaders seem unprepared for the advent of agentic AI. If these leaders are not ready, companies that are still on the gen AI sidelines surely have further to go.

Balance risk and reward: When starting with agentic AI, companies should consider the level of autonomy and data access agents are permitted. Low risk use cases with non-critical data and human oversight can help companies build the data management, cybersecurity, and governance for safe agentic AI applications. Once these are in place, companies should consider higher value use cases that use strategic data, access to more tools, and more autonomy.

Maintain healthy skepticism: Agentic AI is evolving and will likely be more capable in the next year, and will be applied to more horizontal and vertical-specific use cases. Expect impressive demos, simulations, and product announcements throughout 2025. But the challenges we’ve noted may take some time to resolve. Until these challenges are addressed, agentic AI performance in controlled settings is unlikely to deliver improved enterprise performance. Evaluate and question carefully.

By	Jeff Loucks United States	Gillian Crossan United States
	Baris Sarer United States	China Widener United States

Endnotes

1. According to Deloitte's *State of Generative AI in the Enterprise* survey, 23% of enterprises that currently use gen AI are exploring "gen AI agents" to a "large" or "very large" extent, with another 42% exploring it "to some extent." Given the high interest in agentic AI and the products and services that are being launched by startups and established tech companies, we expect this interest to turn to action, at least on an experimental scale.
2. CB Insights. Gen AI Investment Database, Aug 21, 2024. This data excludes Open AI. It includes funding to companies that are developing agentic AI with "varying degrees of autonomy."
3. Kate Clark, "[Investors undaunted by spate of AI acqui-hires](#)," *The Information*, Aug. 19, 2024.
4. Cambridge English Dictionary, "[Agency](#)," accessed Aug. 26, 2024.
5. Cambridge English Dictionary, "[Autonomous](#)," accessed Aug. 26, 2024.
6. For humans, agency and autonomy are moral and political concepts. In the context of gen AI agents, we are speaking only of the extent to which software-based technology has scope to design and perform tasks without human direction.
7. Faruk Muratovic, Duncan Stewart, and Prashant Raman, "[Tech companies lead the way on generative AI: Does code deserve the credit?](#)" *Deloitte Insights*, Aug. 2, 2024.
8. Scott Wu, "[Introducing Devin, the first AI software engineer](#)," Cognition Software, March 12, 2024.
9. Rina Diane Caballar, "[AI Coding is going from copilot to autopilot](#)," *IEEE Spectrum*, April 9, 2024.
10. Jenna Barron, "[Codeium's new Cortex assistant utilizes complex reasoning engine for coding help](#)," *SD Times*, Aug. 14, 2024; Aswin Ak, "[OpenDevin: An artificial intelligence platform for the development of powerful AI agents that interact in similar ways to those of a human developer](#)," *Marktechpost*, July 28, 2024.
11. Carl Franzen, "[Codium announces Codiumate, a new AI agent that seeks to be Devin for enterprise software development](#)," *VentureBeat*, April 3, 2024.
12. Cognition Software, "[SWE-bench technical report](#)," March 15, 2024.
13. Big tech companies continue to improve their software co-pilots to make them more like gen AI agents. For example, see: Alex Woodie, "[The semi-autonomous agents of amazon Q](#)," *BigDATAWire*, May 3, 2024.
14. Molly Talbert, "[Overcoming disruption in a distributed world: Insights from the Anatomy of Work Index 2021](#)," Asana, January 14, 2024.

15. Total factor productivity, which measures how efficiently both capital and labour are used, can be a proxy for knowledge worker efficiency. Knowledge work requires access to capital-intensive technology, and effectively designed processes.
16. US Bureau of Labor Statistics, “[Table A. Productivity, output, and inputs in the private nonfarm business and private business sectors for selected periods, 1987-2023](#),” March 3, 2024.
17. Organisation for Economic Co-operation and Development, “[Multifactor productivity](#),” accessed Oct. 30, 2024.
18. Jim Rowan, Beena Ammanath, Costi Perricos, Brenna Sniderman, and David Jarvis, [State of gen AI in the Enterprise](#), Q3 report, Deloitte, August 2024.
19. Ibid.
20. Ibid.
21. Ibid.
22. James O’Donnell, “[Why OpenAI’s new model is such a big deal](#),” *MIT Technology Review*, Sept. 17, 2024.
23. Janakiram MSV, “[AI agents: Key concepts and how they overcome LLM limitations](#),” *The New Stack*, June 11, 2024.
24. “OpenAI, “[Learning to Reason with LLMs](#),” Sept. 12, 2024.
25. Anna Gutowska, “[What are AI Agents?](#)” IBM, July 3, 2024.
26. Janakiram MSV, “[AI agents: Key concepts and how they overcome LLM limitations](#).”
27. Simon Willison, “[Notes on OpenAI’s new o1 chain-of-thought models](#),” Simon Willison’s Blog, Sept. 12, 2024.
28. Hamidou Dia, “[So much more than gen AI: Meet all the other AI making AI agents possible](#),” Google Cloud Blog, Aug. 20, 2024.
29. Vivek Kulkarni, Scott Holcomb, Prakul Sharma, Edward Van Buren and Caroline Ritter, “[How AI agents are reshaping the future of work](#),” Deloitte AI Institute, November 2024.
30. *The Economist*, “[Today’s AI models are impressive. Teams of them will be formidable](#),” May 13, 2024.
31. CB Insights, “[The multi-agent AI outlook: Here’s what you need to know about the next major development in genAI](#),” Aug. 30, 2024.
32. Mike Desmarais, “[The call center burnout problem](#),” SQM Group, Feb. 24, 2023.

33. It’s important to balance the work of human agents. When they get only the most complicated and difficult cases, it can lead to burnout. See Sue Cantrell, et al., “*Strengthening the bonds of human and machine collaboration*,” *Deloitte Insights*, Nov. 22, 2022.

34. Sierra, “*Sonos elevates the listener experience*,” Feb. 13, 2024.

35. Michelle Meineke, “*The cybersecurity industry has an urgent talent shortage. Here’s how to plug the gap*,” World Economic Forum, April 28, 2024.

36. Ken Yeung, “*Dropzone AI gets \$16.85M for autonomous cybersecurity AI agents that reduce manual work by 90 percent*,” *VentureBeat*, April 25, 2024.

37. Simon Thomsen, “*Software development cybersec startup Nullify banks \$1.1 million pre-seed round*,” *Startup Daily*, June 26, 2023.

38. Kyt, Dotson, “*Norm Ai raises \$27M to help businesses handle regulatory compliance with AI agents*,” *SiliconANGLE*, June 26, 2024.

39. Rowan, *State of Generative AI in the Enterprise*, Q3 report.

40. Ron Miller, “*With Vertex AI Agent Builder, Google Cloud aims to simplify agent creation*,” *TechCrunch*, April 9, 2024.

41. Iris Coleman, “*Paradigm utilizes LangChain and LangSmith for advanced AI-driven spreadsheets*,” *Blockchain.News*, Sept. 5, 2024.

42. *The Economist*, “*Today’s AI models are impressive.*”

43. Maria Korolov, “*AI agents will transform business processes — and magnify risks*,” *CIO*, Aug. 21, 2024.

44. Rowan, *State of Generative AI in the Enterprise*, Q3 report.

Acknowledgements

Authors would like thank **Chris Arkenberg**, **Duncan Stewart**, and **Ankit Dhameja**.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock.

Deepfake disruption: A cybersecurity-scale challenge and its far-reaching consequences

As the effort to detect and combat fake content escalates, the costs of maintaining a credible internet may fall on consumers, creators, and advertisers alike

ARTICLE • 9 MINUTE READ

Deepfakes—photos, videos, and audio clips that seem real but are generated by artificial intelligence tools—are making it harder for audiences to trust content that they see online. As AI-generated content grows in volume and sophistication, online images, videos, and audio can be used by bad actors to spread disinformation and perpetrate fraud. Social media networks have been flooded with such content, leading to widespread skepticism and concern.¹

In Deloitte’s 2024 Connected Consumer Study, half of respondents said they’re more skeptical of the accuracy and reliability of online information than they were a year ago. Among respondents familiar with or using generative AI, 68% reported concern that synthetic content could be used to deceive or scam them, and 59% reported they have a hard time telling the difference between media created by humans and generated by AI. Eighty-four percent of respondents familiar with gen AI agreed that content developed with gen AI should always be clearly labeled.²

Labeling is one of the ways through which media outlets and social media platforms can flag synthetic content for users, but as deepfake technologies incorporate more advanced models that can generate synthetic content and manipulate existing media, more complex measures may be needed to detect fakes and help restore trust.

Analysts estimate that the global market for deepfake detection—as implemented by tech, media, and social network giants—will grow by 42% annually from US\$5.5 billion in 2023 to US\$15.7 billion in 2026.³ We predict that the deepfake detection market could follow a similar path as that of cybersecurity. Media companies and tech providers will likely work to stay ahead of increasingly sophisticated fakes by investing in content authentication solutions and consortium efforts. Credible content is expected to come at an increased cost for consumers, advertisers, and even creators, potentially.⁴

These efforts currently fall under two broad categories: detecting fakes and establishing provenance.

Detecting fakes

Tech companies often use methods such as deep learning and computer vision to analyze synthetic media for signs of fraud or manipulation, leveraging machine-learning models to recognize patterns and anomalies in deepfakes.⁵ These tools can also detect inconsistencies in video and audio content, such as subtle lip movements or voice-tone fluctuations that are less than human.⁶

Some gen AI tools include functionality that detects whether a piece of content was made with their help, but these may not detect deepfakes created by other models.⁷ Some fake-detection tools look for signs of manipulation—or “fingerprints”—of gen AI tools.⁸ Others use a “whitelist” and “blacklist” approach (maintaining lists of trusted sources and known fakers), while others look for proof of humanity (as opposed to proof of artifice), like natural blood flow, facial expressions, and vocal inflection.⁹

Current deepfake detector tools claim accuracy rates above 90%.¹⁰ One concern, however, is that bad actors may be using open-source gen AI models to generate media that work around these measures. The ability to automate content generation may overwhelm current detectors, for instance, and the subtle adjustments that gen AI tools make to output based on user prompts can also be used to obscure fake content.¹¹

Social media platforms themselves often use AI tools to help detect problematic content in images or videos, score it on a relative scale, and then forward the most suspicious items to human reviewers to make the final designation. This approach can be time-consuming and expensive, however, and efforts are underway to accelerate the process with the help of machine learning.¹²

If this sounds reminiscent of the cybersecurity landscape, that could be because it is. Just as security-conscious companies have adopted layered approaches to data and network protection, we expect news outlets and social media companies will likely need multiple tools—along with content provenance measures—to help determine the credibility of digital content.

Establishing provenance and trust

The other path that some companies are exploring involves cryptographic metadata (or digital watermarks) added to a media file when it’s created. This data is attached to the media, detailing its provenance and maintaining a record of all modifications.¹³

Social platforms are collaborating with media outlets, device-makers, and tech companies in cross-industry consortia to help perpetuate standards for content authenticity. Various tech and media organizations, including Deloitte, have joined the Coalition for Content Provenance and Authenticity (C2PA), pledging to use the C2PA metadata standard so that AI-generated images can be verified more easily.¹⁴ C2PA technology records every step of the life cycle of an image, from initial creation through the editing process, by creating a detailed log of alterations and modifications.¹⁵ With the C2PA record available for perusal, content outlets and users can check the source of visuals and consider their trustworthiness.

In another effort to differentiate accounts run by humans, some social media platforms are beginning to implement verification options for creators. These may require submitting forms of identification and paying a fee for creators to prove their identity. Platforms may also incentivize verification by requiring it for creators to be included in revenue-share programs.¹⁶

Certifying the authenticity of human-operated accounts can help platforms improve trust and credibility as AI content grows ubiquitous.¹⁷ Platforms may have to evaluate whether passing these certification costs on to creators, advertisers, or users is sustainable.

Waiting for regulation

Although some governments have initiated measures to regulate content authenticity,¹⁸ more comprehensive and globally coordinated legislation may be beneficial. Public awareness campaigns can be equally crucial in informing users about the dangers of deepfakes and helping teach them to verify before accepting media as authentic.

In the United States, legislation requiring digital watermarks on AI-generated content has been introduced and is now under consideration by the Senate Committee on Commerce, Science, and Transportation.¹⁹ The state of California is considering AB-3211, a law requiring device-makers to update their firmware such that it attaches provenance metadata to photos, and ordering online platforms to disclose provenance metadata for online content. If passed, this law will take effect in 2026.²⁰ Other individual states have enacted legislation criminalizing the production and distribution of nonconsensual deepfakes intended to spread misinformation.²¹ The US Federal Trade Commission is formulating new regulations aimed at prohibiting the creation and dissemination of deepfakes that mimic individuals.²²

Revisions to the EU AI Act primarily emphasize transparency, mandating clear labeling of AI-generated content and deepfakes. This strategy helps support the ongoing advancement of AI technologies, while ensuring that users know the nature of the content they encounter. The European Commission has established the AI Office to foster the development and application of AI and to promote the effective labeling of artificially generated or manipulated content.²³

The swift progression of deepfake technology demands regulatory frameworks that are both flexible and adaptive, capable of evolving in tandem with technological advancements.

Bottom line

The authenticity of a photo, video, or audio clip may be established through analysis and through verification of its provenance. It's likely that media companies and social networks will invest in both approaches as gen AI continues to be used to create more synthetic media and bad actors adjust their models and output to evade detection.

Staying ahead of bad actors is important as gen AI grows more powerful and versatile. More sophisticated technologies like blood-volume detection and facial analysis can help distinguish real from manipulated content. As with cybersecurity tools, however, these measures should be as unobtrusive as possible for end users and consumers, ensuring content integrity without compromising user experience. Techniques like digital watermarking can help verify authenticity without affecting quality or requiring real-time computing cycles to analyze.²⁴

One leading practice could be critical for companies that use trained machine learning models (or pay third parties) to help detect fake content. Such organizations should prioritize tools and vendors that leverage diverse, high-quality data sets spanning images, video, and audio. These data sets should incorporate a broad array of demographic groups to help promote fairness and minimize bias in detection accuracy.²⁵

Tech and media companies alike should collaborate with peers²⁶ across industries to help create and support standards for deepfake detection and content authentication. Digital watermarking, for example, can be more effective when device-makers *and* media outlets co-sign content to affirm its creation and publication. This collective effort can lead to more robust and universally accepted practices, enhancing overall security and trust in digital content.

On the enterprise security side, companies across industries should be aware that gen AI can make social engineering attacks more effective and can compromise some authentication measures.²⁷ It may be necessary to implement additional verification layers, especially for video and audio-based processes. End users should be encouraged to cross-check information with reliable sources and utilize multi-factor authentication to help mitigate risks associated with deepfakes. Because of ever-evolving dynamics, user education (along the lines of cybersecurity awareness training) may also be an important measure for companies to consider.

These strategies can not only safeguard against the threats posed by deepfake technology but also help position tech and media companies as leaders in maintaining digital content integrity and trust. At this pivotal time, companies can shape the trusted content space and position themselves as reliable sources in an increasingly uncertain digital landscape.

By	Michael Steinhart United States	Bree Matheson United States
	Ankit Dhameja India	Gillian Crossan United States

Endnotes

1. Margaret Talev and Ryan Heath, “[Exclusive poll: AI is already great at faking video and audio, experts say](#),” Axios, accessed Oct. 28, 2023.
2. Susanne Hupfer, Michael Steinhart, et.al, “2024 Connected Consumer Survey,” Deloitte, December 2024
3. Vivaan Jaikishan, Cameron D'Ambrosi, Jennie Berry, and Stacy Schulman, “[The rising threat of deepfakes: Detection, challenges, and market growth](#),” Liminal, May 7, 2024.
4. Ian Shepherd, “[Human vs. machine: Will AI replace content creators?](#)” Forbes, April 26, 2024.
5. Analytix Labs, “[Detecting deepfakes: Exploring advances in deep learning-based media authentication](#),” Medium, January 4, 2024.
6. For example, see: Intel, “[Trusted media: Real-time FakeCatcher for deepfake detection](#),” accessed Oct. 28, 2024.
7. Cade Metz and Tiffany Hsu, “[OpenAI releases deepfake detector to disinformation researchers](#),” *The New York Times*, May 2024.
8. Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm, “[Beyond deepfake images: Detecting AI-generated videos](#),” Drexel University, accessed Oct. 28, 2024.
9. Alex McFarland, “[5 best deepfake detector tools & techniques](#) (October 2024),” Unite.AI, Oct. 1, 2024.
10. Konstantin Simonchik, “[Deepfake detection: Accuracy of commercial tools](#),” LinkedIn, February 2024
11. Jiansong Zhang, Kejiang Chen, Weixiang Li, Weiming Zhang, and Nenghai Yu, “[Steganography with generated images: Leveraging volatility to enhance security](#),” *IEEE Transactions on Dependable and Secure Computing* 21, no. 4 (2024): pp. 3994–4005; see also: Mike Bechtel and Bill Briggs, “[Defending reality: Truth in an age of synthetic media](#),” *Deloitte Insights*, Dec. 4, 2023; and, Loreben Tuquero, “[AI detection tools for audio deepfakes fall short. How 4 tools fare and what we can do instead](#),” Poynter, March 21, 2024.
12. Barbara Ortutay, “[Content moderation in the AI era: Humans are still needed across industries](#),” Fast Company, April 23, 2024; also see: Meta, “[How review teams work](#),” Jan. 19, 2022.
13. Glenn Chapman, “[Meta wants industry-wide labels for AI-made images](#),” *AFP News*, Feb. 6, 2024; also see: Nick Clegg, “[Labeling AI-generated images on Facebook, Instagram and Threads](#),” Feb. 6, 2024; Sasha Luccioni et al., “[AI watermarking 101: Tools and techniques](#),” Hugging Face, Feb. 26, 2024; and Partnership on AI, “[Building a glossary for synthetic media transparency methods, part 1: Indirect disclosure](#),” Dec. 19, 2023.
14. Ryan Heath, “[Inside the battle to label digital content as AI-generated media spreads](#),” Axios, accessed Oct. 28, 2024.

15. Demian Hess, “[Fighting deepfakes with content credentials and C2PA](#),” CMSWire, March 13, 2024.
16. Andrew Hutchinson, “[X will require ad revenue share participants to confirm their ID](#),” Social Media Today, May 22, 2024.
17. Guy Tytunovich, “[The future of trust and verification for social media platforms](#),” Forbes, May 22, 2024.
18. Amanda Lawson, “[A look at global deepfake regulation approaches](#),” Responsible Artificial Intelligence Institute, April 24, 2023.
19. US Congress, “[S.2765—Advisory for AI-Generated Content Act](#),” Sept. 12, 2023.
20. California Legislative Information, “[Assembly Bill 3211—California Digital Content Provenance Standards](#),” Aug. 24, 2024.
21. Kevin Collier, “[States are rapidly adopting laws regulating political deepfakes](#),” NBC News, Aug. 7, 2024.
22. Federal Trade Commission, “[FTC proposes new protections to combat AI impersonation of individuals](#),” Feb. 15, 2024; also see: Michelle M. Graham, “[Deepfakes: Federal and state regulation aims to curb a growing threat](#),” Thompson Reuters, June 26, 2024.
23. Melissa Heikkilä, “[Five things you need to know about the EU’s new AI Act](#),” MIT Technology Review, Dec. 11, 2023.
24. Deloitte, “[How to safeguard against the menace of deepfake technology](#),” accessed Oct. 28, 2024.
25. AI Index Steering Committee, “[The AI Index 2024 Annual Report](#),” accessed Oct. 28, 2024.
26. AI Election Accord, “[A tech accord to combat deceptive use of AI in 2024 elections](#),” accessed Oct. 28, 2024.
27. Stu Sjouwerman, “[The growing threat of AI in social engineering: How business can mitigate risks](#),” Fast Company, April 8, 2024.

Acknowledgements

The authors would like to thank **Jeff Loucks, Susanne Hupfer, Duncan Stewart, Jeff Stoudt, Jason Williamson, Tim Davis, Gopal Srinivasan, Shreeparna Sarkar, and Andy Bayiates** for their contributions to this article.

Cover image by: **Jaime Austin; Getty Images, Adobe Stock**

Cloud gets lean: 'FinOps' makes every dollar work harder

Enterprise cloud spend is growing, and using FinOps strategies can make each dollar work harder. Companies can save money, boost value, and build cross-functional cohesion.

ARTICLE • 8 MINUTE READ

Global cloud spending will likely top US\$825 billion in 2025, but ask an organization's leadership what they spend, and it might be difficult for them to answer.¹ Lots of companies either don't know, or struggle to explain it.

As organizations increasingly rely on cloud services, the need for an effective strategy to manage cloud investments becomes paramount. Enter "FinOps" (a mash up of finance and DevOps), a set of strategies to help track and optimize cloud spending. Deloitte predicts US\$21 billion may be saved by companies implementing FinOps tools and practices in 2025 alone, and this could grow in subsequent years. Some may even cut cloud costs as much as 40%. Going forward, we expect companies without a FinOps team to act swiftly to implement first steps, and we expect FinOps veterans to develop more sophisticated optimization strategies.

Cloud is complex, and complexity often creates waste

Cloud is now indispensable for enterprises. Spinning up new cloud environments often takes just a few clicks, whereas building private physical infrastructure can include procuring and installing servers and can take weeks or months to complete. It has helped democratize convenience and scalability. Cloud powers innovation at pace without an army of PhDs on staff; it helps enable industries like video on demand, ride sharing, challenger banks, and telehealth to disrupt their respective markets;² and it underpins applications like data analytics, remote working, and of course AI.

Today, the organizations investing in software engineering are increasingly diverse. Commercial off-the-shelf products may no longer meet their needs. Automotive companies, like Daimler, have assembled teams of developers to build software platforms for electric vehicles.³ Even in industries without digital heritage, like wooden pallet distribution, there is also a growing need for custom software and expertise.⁴ All of this increases the cloud bill.

However, cloud is getting complex. Companies tend to juggle their private computing resources with public cloud services—a hybrid cloud infrastructure—that 73% of companies now have. Also, more than half (53%) of companies source cloud from multiple providers to take advantage of promotions, specific capabilities or avoid vendor lock-in.⁵ It's also common for individual departments (think finance, HR, or marketing) to buy cloud software applications without the knowledge of the central engineering team. All of this can create complexity in areas like company data integration, compliance, and security.

In general, companies are not good at sticking to their cloud budget. Half of organizations surveyed overspent last year, and the average overrun was 15%.⁶ One factor is pay-as-you-go billing, which means cost is variable, and can make forecasting a greater challenge. In extreme cases, cloud engineers have inadvertently triggered thousands of dollars of cloud spend overnight.⁷

Cloud is not cheap; it can cost more than the equivalent private infrastructure,⁸ and is fast becoming a company's largest IT line-item bill. Coca-Cola, for example, recently signed a \$1.1 billion deal for its cloud needs.⁹ Crucially, though, 27% of spend is wasted, according to cloud leaders.¹⁰ Cloud customers have started to recognize this, and half now have a dedicated FinOps team, while 20% more should form a team within a year.¹¹

Getting started with FinOps

FinOps is a financial management discipline. It can range from the technical, like rearchitecting cloud workloads and reviewing what can go into less accessible long-term storage, to much less technical, like negotiating discounts and credits. Its long-term impact, however, is cultural change. At its core, it's about cross-organization responsibility and financial accountability, and aligning each cloud dollar spent with the business value it generates.

Starting with FinOps is all about planning: reviewing the current strategy, evaluating any tagging and alerting structures, and then defining key performance indicators.¹² A practical first step is to focus on visibility, which can mean cataloguing current cloud resources and exploring how they may align with organizational needs. For this, cloud providers offer resource monitoring tools as well as specific tools focused on cost, or there are third-party FinOps platforms like that can provide more granular metrics.

However, interpreting dashboard data may require dedicated FinOps specialists and practitioners; who are often in-demand people. Also, companies with multiple cloud providers may need a dashboard for each. A single integrated portal can be challenging as data feeds from each provider will vary. Finally, FinOps tools can bear a considerable cost (around 3% to 5% of the cloud bill at the high end), so a company should understand their cloud economics before deploying one.

Starting out with FinOps: Preliminary measures

Organizations starting their FinOps journey will likely focus more on preliminary measures to help cut waste, ensure that resources are allocated optimally, review their contracts, and take advantage of potential credits and discounts.

Waste and consumption: We expect that FinOps novices will find getting rid of waste a great place to start. FinOps tools and dashboards might help a company pinpoint underutilized or idle resources to be right-sized or shut down, for potentially instant cost-savings. Examples of waste can include oversized virtual machines, redundant storage instances, orphaned resources, or duplicate data. The companies most adept at FinOps may use predictive analytics to forecast usage, and automated governance scripts that can dynamically adjust capacities. Plus, this can often be done by the central cloud engineering team, so the task can be completed quickly.

Structure and tiers: Cloud services are not all created equal. Computing and storage instances can span a range of qualities and price points. Companies could assess the caliber of their provisioning (allocating and managing cloud resources effectively), and whether it adequately fits the need of an application. It may be that some applications perform well on more cost-effective instances. For example, to help mitigate seasonal volatility, an event ticketing website may choose an instance which does not use the full CPU continuously, but occasionally needs to burst to align resourcing with web traffic.¹³

Incentives: Cloud platforms offer discount programs that can lead to substantial savings. Some platforms allow users to commit to a consistent amount of usage in exchange for lower rates. For some companies, directly renegotiating with their cloud service provider can be a fruitful approach. Cloud companies tend to be receptive to this, welcoming the chance to exchange discounted rates for multi-year contract commitments.

Getting serious: Refining with broader approaches

In 2025, experienced FinOps practitioners may continue to refine the above, but are also expected to advance their approaches to cost observability and control.

Accountability: As cloud is vital across all parts of many businesses, each department (and team) should be financially accountable for their spend. Departments should be given oversight and responsibility for costs that can be directly attributed to them, via either a chargeback model (charging a department directly) or “showback” model (showing a department its cost burden).¹⁴ This can require a robust tagging strategy, which assigns resource costs to specific teams or projects, ideally auto-tagging based on predefined rules. Ideally, this could create a culture where teams throughout the organizations feel engaged in cloud cost reduction.

On-prem: FinOps communities like the FinOps Foundation have started to encourage discourse around on-premises infrastructure, which is often opaque to users, as part of the overall equation.¹⁵ Companies should be thinking about cost across their entire IT estate. But this can be complex, as the central cloud team may need to liaise with branch offices and infrastructure sites and may find that they use a variety of hardware and software tools for local needs. Strategies for on-prem cost reduction can include cancelling redundant licenses and extending lifecycle of hardware.

Sustainability: FinOps also intersects with the growing “GreenOps” movement. GreenOps describes a set of cloud management strategies which optimize for sustainability. Granular metrics delivered by FinOps reporting tools can help with measurement of energy consumption, carbon emissions, and other sustainability goals.¹⁶ In the wake of major reporting regulations like the EU’s Corporate Sustainability Reporting Directive,¹⁷ tracking energy and carbon metrics and subsequently improving them can be an incredible by-product of investing in FinOps.

FinOps makes a tangible difference

FinOps practices have been instrumental for many companies working to achieve cloud cost savings:

- **Airbnb:**¹⁸ Travel app Airbnb generated a \$63.5 million saving in cloud costs. Part of its approach was to shift storage to a lower-cost service tier and switch out its homegrown backup system for a cloud provider’s alternative.
- **Sky Group:**¹⁹ Media and entertainment company Sky discovered it had spent a full year’s cloud budget within six months. It deployed a first-party FinOps tool to identify \$1.5 million in savings and implemented visibility dashboards, which enabled \$3.8 million in savings in the subsequent year.
- **The Home Depot:**²⁰ The home improvement retailer built a dedicated cloud cost team in 2022 and identified “tens of millions of dollars” in savings compared with the previous year.
- **Lyft:**²¹ Ride-sharing app Lyft cut cloud costs per ride by 40% in six months, with a spreadsheet-based software tool to track billing data made available to the entire company. This led to a wave of right-sizing programs.
- **WPP:**²² Advertising firm WPP saved \$2 million after just three months of FinOps deployment, which eventually scaled to a 30% annual cost reduction on its yearly cloud spend. It leveraged a range of tools and techniques, including autogenerated sizing recommendations.

A broad range of organizations are now actively investing in FinOps. For example, members of the FinOps Foundation—a non-profit organization promoting best practices in cloud financial management—include Walmart, Mastercard, and American Airlines.²³

Bottom line: Cloud unit economics

The emergence of FinOps reflects a need for better visibility, improved budgeting, and proactive control of cloud expenditures, which has continued to grow as organizations increase their reliance on cloud.

Going forward, global IT spending is set to rise (exceeding \$5.1 trillion in 2025,) ²⁴ in part driven by digital transformation and AI. On top of that, private infrastructure still accounts for around half of all workloads, which, if migrated to public cloud, could significantly inflate the cloud bill. Additionally, high interest rates (at least compared to the last decade), have caused companies to focus on profitability and cost reduction, with a particular appetite to remove cost variability. In other words, the stage is set for FinOps’ growth.

FinOps should be viewed as a long-term practice, integral to operational strategy. It should not be seen as a simple fix. It starts with cost reduction, but it can eventually transform cloud spending from a mere line item into a strategic asset and enabler.

For some of the most advanced companies, the end-goal may be to create a “cloud unit economics” model. This approach quantifies the costs associated with each unit of cloud service used—per application, workload, or gigabyte of data processed—and aligns this with the resulting business metrics, such as revenue, cost per delivery, cost per booking, and cost per ride. This more granular insight can help companies make effective decisions about IT in the context of their whole business, helping to ensure each unit of spend is trackable to the bottom line.

For some companies, costs saved may be reinvested in new growth opportunities, such as scaling through new cloud services, or accelerating a product roadmap.

Cloud will always be complex, and it may never be inexpensive, but companies that can apply FinOps should make cloud more valuable to the bottom line than ever.

By	Ben Stanton United Kingdom	Adam Gogarty United Kingdom
	Paul Lee United Kingdom	Gillian Crossan United States

Endnotes

1. Gartner, “[Gartner forecasts worldwide public cloud end-user spending to surpass \\$675 billion in 2024](#),” press release, May 20, 2024.
2. Yury Izrailevsky, Stevan Vlaovic, and Ruslan Meshenberg, “[Completing the Netflix cloud migration](#),” Netflix, Feb. 12, 2016.
3. Douglas Busvine, “[Daimler to hire 1,000 programmers in Germany](#),” Reuters, April 18, 2021.
4. CHEP, “[CHEP uses ‘track and trace’ technology on its reusable pallets](#),” press release, April 8, 2022.
5. Flexera, “[2024 State of the cloud report](#),” 2024.
6. Ibid.
7. Parshv Jain, “[Avoiding costly cloud mistakes: Lessons learned from a \\$72K bill](#),” Medium, June 12, 2023.
8. Owen Rogers, “[Reports of cloud decline have been greatly exaggerated](#),” Uptime Institute, Jan. 18, 2023.
9. The Coca-Cola Company, “[The Coca-Cola Company and Microsoft announce five-year strategic partnership to accelerate cloud and generative AI initiatives](#),” press release, April 23, 2024.
10. Flexera, “[2024 State of the cloud report](#),” 2024.
11. Ibid.
12. Nikhil Roychowdhury, Nik Jethi, Farhan Akram, and Rishabh Kochhar, “[Optimizing the value of cloud: A guide to getting started](#),” Deloitte, March 30, 2023.
13. Amazon Web Services, “[TicketSwap tames demand ups and downs with AWS](#),” 2021.
14. FinOps Foundation, “[Invoicing & chargeback](#),” accessed Nov. 4, 2024.
15. For example: The Linux Foundation, “[FinOps across public cloud and on-prem](#),” accessed Nov. 4, 2024.
16. Meredith Shubel, “[What is GreenOps? Putting a sustainable focus on FinOps](#),” The New Stack, Sept. 22, 2023.
17. Magda Puzniak-Holford, Adithya Subramoni, and Simon Brennan, “[EU Corporate Sustainability Reporting Directive \(CSRD\) - Strategic and operational implications](#),” Deloitte, Sept. 8, 2023.
18. Belle Lin, “[Airbnb details road map to lower cloud costs](#),” The Wall Street Journal, Nov. 7, 2022.

19. James Ma, “*How Sky saved millions with Google Cloud*,” Google Cloud Blog, July 19 2021.

20. Angus Loten and Isabelle Bousquette, “*Amazon warns of weaker cloud sales as businesses cut spending*,” *The Wall Street Journal*, April 13, 2023.

21. Amazon Web Services, “*Lyft uses AWS Cost Management to cut costs by 40% in 6 months*,” 2020.

22. IBM, “*How the world’s largest ad company optimizes FinOps*,” accessed Nov. 4, 2024.

23. FinOps Foundation, “*FinOps Foundation Members*,” accessed Nov. 4, 2024.

24. Gartner, “*Gartner forecasts worldwide IT spending to grow 8% in 2024*,” press release, Oct. 18, 2023.

Acknowledgements

The authors would like to thank **Nikhil Roy Chowdhury, Mitesh Gursahani, Nik Jethi, Rebecca Wood, Avishek Swain, Sophia Atkinson,** and **Vipul Mehta** for their contributions to this article.

Cover image by: **Jaime Austin**

On-device generative AI could make smartphones more exciting—if they can deliver on the promise

With specialized chips and extensive mobile OS integration, smartphones could become smart—even intelligent. Will users embrace the new approach?

ARTICLE • 10 MINUTE READ

Smartphones have become the most widely used piece of consumer technology in the world. They have absorbed many other devices, and their advanced and miniaturized components have flowed downstream into innumerable consumer and industrial devices. Their at-hand convenience and utility has reshaped behaviors and the competitive landscape. Yet, despite this, recent smartphone innovations seem to have failed to excite the market, appearing more incremental than revolutionary.

Now, dominant mobile ecosystem providers are starting to reorganize their devices around next-generation operating systems and advanced chips that aim to bring generative AI into the center of the smartphone experience.³ More original equipment manufacturers are now shipping gen AI-capable smartphones.⁴ Looking to bottle the lighting of generative AI, providers could make smartphones exciting again, but it may not be without some risk.

Deloitte predicts that in 2025, global smartphone shipments will see a modest lift to around 7%, up from about 5% annual growth in 2024.⁵ Some of this lift could be due to resetting the typical device upgrade cycle as consumers upgrade to the latest models. And some may be from early adopters and developers enthusiastic about next-generation phones shipping with chips designed to support generative AI locally on-device. Deloitte further predicts that the share of shipped gen AI-enabled smartphones could exceed 30% by the end of 2025.⁶

There is excitement about generative AI, but can the technology deliver on its promises, and will users embrace a new way of interacting with the most widely used consumer device?⁷

Can generative AI on smartphones boost the next upgrade cycle?

In the near term, leading smartphone designers may see generative AI integration as a way to stoke demand for their premium models: Sales of smartphones had been down for two years prior to 2024.⁸ In part, this was due to a degree of market saturation: It's estimated that nearly five billion people now own smartphones—more than half of all humans.⁹ Upgrade cycles have been getting longer in recent years: People have been upgrading their phones every two to three years, on average, and more households have reported feeling inflationary pressures that limit their discretionary spending.¹⁰ At the same time, more are opting for higher-end devices, knowing they will be using them for a few years.¹¹ This has likely put more pressure on the need for not just better hardware, but also more compelling value and utility in the smartphone user experience.

In 2025, smartphones are expected to put the utility of generative AI to the test.

The first quarter of 2024 showed stronger growth in smartphone sales, from increased consumer confidence and some apparent early interest in premium generative AI-enabled devices.¹² Deloitte's 2024 Connected Consumer Study found similar evidence: Fewer households now report affordability issues affecting their purchases of connected devices.¹³ This recovery seems evident in Europe as well, which saw continued growth in smartphone sales during the second quarter of 2024.¹⁴ So, in 2025, the upgrade cycle is likely to rebound, more people will likely upgrade their smartphones, and more of those upgrades could be for higher-priced premium devices with onboard generative AI features.

It should be noted that, while generative AI could become a driver for smartphone upgrades, the amount likely varies between markets and age groups. The same Deloitte study shows that 7% of US respondents agree that generative AI features make them likely to upgrade their smartphones sooner than they had planned, but the number jumps to 50% for those between the ages of 24 and 45 years old, who may be more dependent on smartphones and more likely to embrace new tech.¹⁵ In [Deloitte UK's 2024 Digital Consumer Trends report](#), however, only 4% of UK respondents report using generative AI daily, with 23% of respondents saying they don't find it helpful, and 19% saying they're not satisfied with the answers it gives.¹⁶

Will generative AI help create a greater boost to smartphone upgrades? It depends on how much value and utility it delivers. In 2025, smartphones are expected to put the utility of generative AI to the test.

Generative AI in personal computers

The same considerations for user experience, utility, and value, as well as the broader pressures shaping the evolution of hyperscale generative AI, apply to a new generation of PCs shipping with on-device chips dedicated to generative AI.

Results from Deloitte's 2024 Connected Consumer Survey suggest that consumers are interested in buying gen AI-enabled PCs: Thirty-four percent of US respondents planning to upgrade their laptops or PCs agree that generative AI chips are likely to accelerate their purchasing. Deloitte believes that individual consumers will make up about 50% of annual PC sales, so this could be an important factor.¹⁷ For enterprise buyers, there's some uncertainty about which gen AI coprocessor models on PCs make the most sense for businesses, with various PC original equipment manufacturers offering various options at various price points.¹⁸

It's expected that over time, most high-end PCs will have gen AI functionality via special silicon. One estimate is that 80% of all PCs will have these kinds of chips by 2028.¹⁹ Another estimate suggests that nearly 9 million "AI capable" machines were shipped in the second quarter of 2024, although it's unclear how many of these include neural processing units strong enough to run generative AI workloads.²⁰ Indeed, potential customers might wait a year or so for the next generation of machines to deliver greater performance before they upgrade.

Deloitte predicts that roughly 30% of all PCs sold in 2024 will have had some local generative AI processing capabilities,²¹ and we further predict that close to half of all PCs sold in 2025 will have this capacity.

Although the computer market is not as large as the smartphone market—about 261 million units expected to be sold in 2024²² compared to 1.23 billion smartphones²³—the higher average selling price of computers means they often punch above their weight in dollar terms. Computer sales are estimated by Deloitte to be about US\$220 billion in 2024, compared to smartphone sales of about US\$520 billion for the year.²⁴

What is unclear is what effect gen AI-enabled machines could have on the PC sector. We believe that there will be an average selling price increase of PCs caused by these devices, adding a premium of about 15% to each PC.²⁵ However, PC sales are expected to rise only in the single-digit percentage range in 2025.²⁶

For consumers, advancements in components for both smartphones and PCs are likely to shape supply chains and push costs down, enabling such components to move into many more devices. Generative AI capabilities are expected to become more common across connected device categories.

Generative AI could make smartphones intelligent

The term "smart" in smartphones has often meant they're connected and can run apps. Generative AI may offer a way for smartphones to become more personalized and aware of user interactions and intentions, and more intimate through conversational interfaces. Although prior attempts at voice assistants may not have lived up to expectations, some people are already forming relationships with the latest conversational large language models.²⁷ This could become a new interaction paradigm with conversational AI as a way to interface with digital systems, and a new model for trusted intelligent agents that could learn to act on an individual's behalf.

On-device gen AI models could answer questions like, “How early should I leave for my 2 p.m. appointment?” by inferring the user’s intention and understanding the full context of the user’s calendar, their location, and the best route to the destination within the timeframe. They are expected to focus on doing narrow tasks well, leveraging neural processing units that can deliver enough performance—at least 30 tera operations per second, by some estimates²⁸—to support on-device inference. The model could further recognize if a question is beyond the local scope, and then assign the task to larger, cloud-based models better able to answer. This hybrid approach to high-performance mobile computing can allow for more immediate and secure interactions on the device, with direct access to models in the cloud.²⁹

With smaller models running on the device, user interactions and data can be contained and secured locally as needed, and more low-latency operations that might require very fast responses, like real-time translation, could be enabled.³⁰ These features may help secure the trust of users and offer more obvious utility. Providers may also see a new flywheel of data from user interactions that could help inform local and cloud models to deliver better results to users—and greater insights to their business.

A more distant goal is that smartphones—arguably the center of consumer interactions— could become much more personalized and intelligent, tuned to individual behaviors and predictive of our needs. (See our 2025 TMT Prediction on agentic AI.) This kind of “agentic” functionality could push smartphones—and the device ecosystem they often interact with and gradually transform—to evolve from merely “smart” to “intelligent.” (See our [2025 TMT Prediction on agentic AI](#).)

Just as there is strong market pressure to justify the costs of frontier models by establishing their product fit, there is pressure to make them more cost-effective to build and operate.

The coming year will show how quickly users onboard onto the new experience, testing the value—and comprehensibility—of early gen AI features. Providers are expected to roll out new features over the coming months, and are likely assuming that broader adoption will take time.³¹ The year ahead will likely also test the capabilities and limits of small models running on-device rather than going to the cloud. In time, this could change the economics of generative AI. If more generative AI tasks shift from expensive data centers to consumer devices, the capital intensity of the generative AI build-out could be softened

Can the industry spend its way past generative AI hype?

Just as there is strong market pressure to justify the costs of frontier models by establishing their product fit, there is pressure to make them more cost-effective to build and operate.³² Leading model providers have spent billions of dollars to develop current frontier models and are investing billions more to build out the data centers they believe will be necessary to meet demand at scale.³³ By some estimates, US\$600 billion is being spent each year to support generative AI.³⁴ Such capital intensity, however, can only go on for so long before it demands economic value which, in turn, may require better product fit.

Making models smaller, reducing the amount of data they need, and breaking them apart based on the scope of workloads may be ways to potentially reduce their costs, especially for inference tasks that can scale with use. Many tasks for consumers and workers may be exposed to generative AI, and they may be addressable or augmented by cheaper and more energy-efficient small models.

However, it’s unclear how much inference will remain on-device. Current generative AI interactions and expectations have mostly been defined by public cloud-based models. It may take time for users to understand which kinds of tasks and prompts run locally, securely, and for free, and which will traverse networks to models in the cloud. Interacting with a conversational, on-device, and cloud-enabled agent is a new paradigm with unclear implications for adoption and behaviors.

Broad adoption of generative AI still faces challenges

Deloitte's 2024 Connected Consumer Study shows that 38% of US respondents have used generative AI, and 63% of those users say the technologies exceed their expectations.³⁵ The magic may already be there for many who have used generative AI, but providers may need to show broad utility to wider demographics to help justify the cost of a new smartphone for consumers.

Usage of generative AI on smartphones could prove confusing, as users attempt to navigate novel interactions. They may hesitate to cede their own agency to intelligent assistants that seek, for example, to manage their calendars.³⁶ Adoption could reveal trade-offs in battery consumption, costs levied by integrated public models, and unrecognized falsehoods that could undermine high-value use cases. Building trust between users, their personal AI agent, and public models will likely take time; losing that trust could happen very quickly.

Providers likely hope that the next generation of frontier models can unlock greater value, but it remains unclear if frontier models will continue to see such growth in capabilities, or if the curve will flatten or decline. And is there enough data to feed increasingly avaricious training sets?³⁷ Solutions like synthetic data created by models to train themselves may cause the quality of inference to degrade over time.³⁸ Can functionality advance without higher costs in data, training, and inference? Is there a window where functionality could improve while capital and data intensity decrease? Nervous investors could potentially demand greater revenues before the technology is able to deliver them.

Regulators could also impact development of gen AI with a broader approach to safeguarding against emerging ills, like deepfakes, misinformation, and persuasive human-like bots. Conversational bots may be establishing greater rapport and intimacy with users, that are better able to influence their ideas and ideologies.³⁹ Personalized conversational agents could tap into the deeper realms of human interactions, potentially helping more people, but also risking addictions.⁴⁰ Combining on-device generative AI with third-party models could create a larger surface area of security vulnerabilities and exploits.⁴¹ This could further provoke providers to secure their ecosystems, and regulators to install more guardrails.

Bottom line

Despite recurring talk of the “next smartphone”—a consumer device platform with the potential to transform and uplift entire markets—it hasn't come. With billions of users, smartphones still dominate and offer a large test bed for new services and user interactions. In 2025, the number of people interacting with generative AI will likely get a boost through premium smartphones—and through personal computers. They can try it, learn its value, and test for its edges. If it succeeds, smartphones could become more exciting and could help expand their platform to enable entirely new categories of use and opportunity—and drive a new boom in personal devices. But this will likely take time, and the coming year is expected to be an onboarding effort to help introduce users to a new paradigm for personal computing.

In the coming years, the smartphone operating system could capture more interactions, such as the next generation of conversational search that can return more local summaries than remote links, disintermediating service providers and information sources. If users adopt more personalized agentic AI, the nature of digital interactions could change, potentially off-loading more tasks to a user's device rather than demanding direct user interface. In this manner, computing could become more ambient, operating in the background on our behalf, and potentially more spatial—increasingly aware of our surroundings and network interactions.

As providers work to stoke demand, they may find themselves racing against economic pressures to offset the capital intensity and energy costs of training and operating models at scale. The industry could pursue small models, hybrid architectures, and a deeper understanding of which generative AI workloads require which kinds of computational overhead. At a time when climate uncertainty and anxiety is difficult to escape, the gen AI data center buildout is already driving up energy and water usage, as well as the energy costs borne by households and municipalities.⁴² If generative AI surmounts its economic debt, it may yet find itself impaired by energy debt.

As of late 2024, the bet for hyperscalers, smartphone ecosystem owners, and young public models is that the benefit they provide will turn into broad economic value. But how much of that value will they capture? Will generative AI hyperscalers follow the path trod by telecoms and the early internet, spending down their reserves on massive capex just to build the infrastructure that could ultimately power the next generation of innovators?⁴³

Driving the emergence, deployment, and broad adoption of generative AI could constitute one of humanity’s grandest experiments since unleashing the internet to the masses—a moonshot that, destination aside, could deliver a new flood of technologies, behaviors, and business models through its development.

By	Chris Arkenberg United States	Duncan Stewart Canada
	Gillian Crossan United States	Kevin Westcott United States

Endnotes

1. GSM Association, “*Smartphone owners are now the global majority, new GSMA report reveals*,” press release, Oct. 11, 2023.
2. Wolfgang Bock, François Candelon, Steve Chai, Ethan Choi, John Corwin, Sebastian DiGrande, Rishab Gulshan, David Michael, and Antonio Varas, “*The mobile revolution: How mobile technologies drive a trillion-dollar impact*,” Boston Consulting Group, Jan. 15, 2015.
3. IDC Corporate, “*The future of next-gen AI smartphones*,” Feb. 19, 2024.
4. Counterpoint, “*Gen AI-capable smartphone shipments to grow over 4x by 2027*,” April 16, 2024.
5. IDC Corporate, “*Worldwide smartphone market up 7.8% in the first quarter of 2024 as Samsung moves back into the top position, according to IDC tracker*,” press release, April 15, 2024.
6. IDC anticipates a 364% compound annual growth rate in 2024 (from a low base in 2023) for global gen AI smartphone shipments, with 73% growth in 2025. Canalys expects AI-enabled smartphone market share to reach 54% by 2028. Our analysis, for reasons outlined in this paper, is less bullish than the former, and a bit more than the latter. Sources: IDC Corporate, “*The future of next-gen AI smartphones*”; Canalys, “*Now and next for AI-capable smartphones*,” accessed Oct. 30, 2024.
7. Jim Fellingner, “*CTA study: Smartphones most-owned tech, 5G and wireless drive adoption*,” press release, Consumer Technology Association, May 31, 2023.
8. IDC Corporate, “*Worldwide smartphone market up 7.8% in the first quarter of 2024 as Samsung moves back into the top position, according to IDC tracker*.”
9. GSM Association, “*Smartphone owners are now the global majority, new GSMA report reveals*.”
10. Sarah Barry James, “*Consumer checkup: Higher interest rates lead to longer tech replacement cycles*,” S&P Global, March 26, 2024.
11. IDC Corporate, “*Worldwide smartphone market up 7.8% in the first quarter of 2024 as Samsung moves back into the top position, according to IDC tracker*.”
12. Chris Donkin, “*Smartphone sales up again ahead of expected gen AI boost*,” Mobile World Live, July 15, 2024.
13. Susanne Hupfer, Michael Steinhart et al., “2024 Connected Consumer Study,” *Deloitte Insights*, publication forthcoming, 2024.
14. Counterpoint, “*Europe smartphone market recovery continues, shipments up 10% YoY in Q2 2024*,” Aug. 28, 2024.

15. Susanne Hupfer, Michael Steinhart et al., “2024 Connected Consumer Study,” *Deloitte Insights*, publication forthcoming, 2024.
16. Deloitte, “*Generative AI: 7 million workers and counting*,” June 25, 2024.
17. The installed base of PCs is estimated to be about 2 billion, and there are about 1 billion knowledge workers, suggesting that the market is roughly half consumer and half enterprise.
18. Author interviews with enterprise chief information officers in July and August 2024.
19. Canalys, “*AI-capable PCs forecast to make up 40% of global PC shipments in 2025*,” March 18, 2024.
20. Ibid.
21. Deloitte Global analysis of publicly available information for H1 2024, and extrapolation based on usual PC seasonality trends.
22. IDC Corporate, “*PC refresh cycle and tablets in emerging markets expected to spur demand in coming quarters, according to IDC*,” press release, Sept. 23, 2024.
23. IDC Corporate, “*Worldwide smartphone market forecast to grow nearly 6% in 2024, driven by stronger growth for android in China and emerging markets, according to IDC*,” press release, Aug. 27, 2024.
24. Based on quarterly data so far in 2024, Deloitte believes smartphone average selling price is declining and should be roughly US\$425 for the year. PC average selling prices were high during the 2021 chip shortage, but are declining and Deloitte estimates them to be about US\$850 for 2024.
25. Roshan Ashraf Shaikh, “*Analysts expect 15% price hike for AI PCs—60% of PCs will have local AI capabilities by 2027*,” Tom’s Hardware, April 26, 2024.
26. IDC Corporate, “*PC refresh cycle and tablets in emerging markets expected to spur demand in coming quarters, according to IDC*.”
27. Sigal Samuel, “*People are falling in love with—and getting addicted to—AI voices*,” Vox, Aug. 18, 2024.
28. IDC, “*The future of next-gen AI smartphones*.”
29. Baris Sarer, Mark Szarka, Nataliia Bacchus, and Edem Isliamov, “*The world of hybrid AI*,” *The Wall Street Journal* and Deloitte, July 31, 2024.
30. Malik Saadi, “*On-device generative AI unlocks true smartphone and PC value*,” *Forbes*, April 17, 2024.
31. Lisa Eadicicco, “*AI is changing our phones, and it’s just getting started*,” CNET, April 3, 2024.
32. Goldman Sachs, “*Gen AI: Too much spend, too little benefit?*” June 27, 2024.

33. David Cahn, “*AI’s US\$600B question*,” Sequoia, June 20, 2024.

34. Ibid.

35. Susanne Hupfer, Michael Steinhart et al., “2024 Connected Consumer Study,” *Deloitte Insights*, publication forthcoming, 2024.

36. Jon Victor, “*Software firms race to beat OpenAI in AI agents*,” The Information, Sept. 26, 2024.

37. Deepa Seetharaman, “*For data-guzzling AI companies, the internet is too small*,” *The Wall Street Journal*, April 1, 2024.

38. Michael Peel, “*The problem of ‘model collapse’: How a lack of human data limits AI progress*,” *Financial Times*, July 24, 2024.

39. Yuval Noah Harari, “*Yuval Noah Harari argues that AI has hacked the operating system of human civilization*,” *The Economist*, April 28, 2023.

40. CBS News, “*Virtual valentine: People are turning to AI in search of emotional connections*,” Feb. 14, 2024.

41. Matt Burgess, “*Generative AI’s biggest security flaw is not easy to fix*,” *Wired*, Sept. 6, 2023.

42. Camilla Hodgson, “*US tech groups’ water consumption soars in ‘data center alley’*,” *Financial Times*, Aug. 17, 2024.

43. Bryce Elder, “*Gen-AI revisited, by Goldman Sachs*,” *Financial Times*, Sept. 5, 2024.

Acknowledgements

Authors would like to thank **Rohan Gupta** and **Steve Fineberg**.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

Large studios will likely take their time adopting generative AI for content creation. Social media isn't hesitating.

Hollywood (and others) may be cautious about using gen AI for content creation, but they will likely be quicker to adopt it for operations and distribution

ARTICLE • 12 MINUTE READ

Generative AI models for image, audio, and video are advancing, delivering more realistic and creative content that is becoming more controllable over longer sessions. Although studios may have been quick to experiment with gen AI content creation, they will likely be more cautious in moving it into full production. Some reasons for this include the immaturity of the tools and the challenges of content creation with current public models that may expose them to liability and threaten the defensibility of their intellectual property (IP). However, there is growing belief that gen AI applied across their businesses could help studios reduce costs and grow their profitability.

Indeed, leading studios are facing cost pressures, and very few are showing profits.¹ Revenues are high, but operating expenses and the costs of production, marketing, and advertising have typically become higher.² This is often true for many studio streamers that are funding their streaming services without profit while losing revenues from declining cable TV subscriptions and advertising. Inflation, higher interest rates, and the impacts of the COVID-19 pandemic have further inflated costs, and studios now also compete with social media, user-generated content, and video games for consumer attention and revenues.

In 2025, Deloitte predicts that the biggest TV and film studios—especially those in the United States and European Union—will be cautious in adopting generative AI for content creation, with less than 3% of their production budgets going to these tools.³ But we also predict that operational spending will shift about 7% into emerging generative-AI-enabled tools supporting functions like contract and talent management, permitting and planning, marketing and advertising, and localization and dubbing of content that can expand their reach into diverse global markets.

This approach can help studios slow the potential disruptions that gen AI can pose to talent and content, while more quickly adopting gen AI tools that can help reduce costs and accelerate performance across their businesses. However, independent content creators and social media platforms are moving quickly to adopt gen AI into their workflows and content, potentially enabling new forms of media to emerge that could further disadvantage traditional studios competing for scarce attention time.⁴

For Hollywood-level content creation, gen AI tools are still immature

The availability of cheap, off-the-shelf large language models (LLMs) and diffusion models have helped enable studios to experiment with rapid prototyping of scripts, dialog, and story elements, and with early visualization and discovery of character and set design.⁵ Some studios are using generative tools to de-age their celebrities or create digital twins that can be lent to commercials—or to post-mortem productions.⁶ In such cases, studios can help control for potential liabilities by writing protections directly into the contracts with actors. The coming year will likely see more third-party production groups selling services and tools to studios offering such capabilities.

While content creation with generative AI can enable greater creativity in preproduction, it cannot yet deliver Hollywood-level productions.⁷ Although the strongest visual diffusion models are now able to generate photorealistic images, their outputs still seem “uncanny”—too hyper-realistic.⁸ Leading video models can generate short clips, but cannot produce longer, more coherent stories.⁹ Although video-generation models are advancing quickly, it may still take time before they are mature enough to integrate into existing tools and production pipelines.

The year ahead will likely see independent creators leading the way in content creation with generative AI.

However, these limitations may be fine for social media creators, who are often incentivized to create and publish quickly. Fast-paced quick cuts have gained popularity, though this could be changing.¹⁰ Social video lengths are often short, and liabilities are perhaps less concerning. Some early adopters of generative models and tools are regularly publishing their experiments on social media, showcasing the fast-moving advances of video models teased out of third-party solutions.¹¹

The year ahead will likely see independent creators leading the way in content creation with generative AI. This could help studios defer their own risks while they watch to see how the capabilities evolve. But it could also cede more attention time to user-generated content platforms that are becoming highly competitive with traditional media.

Big studios may worry about liability risks of public models

Large studios are also concerned about how gen AI content tools may raise their exposure to IP and liability risks, or make their own generative content indefensible as original works.¹² Some of the most capable publicly available models have been trained on public data, like images and videos from other creatives, making their outputs fundamentally derivative.¹³ If a studio uses outputs from a public model for profit, and that model includes the protected works of other artists in its training set, the studio could be held liable for infringement. With potentially billions of works in a training set, infringement may be nigh impossible to prove—finding a “drop of water in the ocean”—but this uncertainty may be enough to scare off studios whose livelihood is making, securing, and defending their IP. Independent artists and creators are already suing public models for perceived infringements of their works in training sets,¹⁴ as are publishers¹⁵ and music labels.¹⁶

Public models could also make it difficult or impossible for studios to secure their own IP. The US Copyright Act has required sufficient “human authorship” before it will issue a copyright.¹⁷ In recent considerations, the US Copyright Office acknowledges that the degree of human authorship in works that include AI or generative inputs can vary on a case-by-case basis, and such works can receive a copyright provided they meet requirements of “sufficiency.” Which is to say, studios can get copyrights for human works that are supported by generative AI tools, but only to a degree and not for works primarily produced by generative models. This is an ongoing discussion in a maturing space, but the lack of precise definitions introduces further uncertainty and risk.

Hungry for more data to feed their training sets, leading gen AI providers have been courting studios and incentivizing them to license their content archives.¹⁸ However, studios may resist this entirely since their IP is their livelihood, or they may charge onerously high rates to gen AI companies that may already be straining under their own operational costs. Studios could even see an advantage in collectively denying data to training sets in hopes that they might inhibit frontier models—the algorithms being encoded and trained to generate text, audio, image, and video.

Additionally, studios—especially those in the United States—must work with guilds and labor unions that have shown strong resistance to adopting generative AI and have extracted guarantees from studios limiting its use.¹⁹ Similar labor pushback is emerging in the United Kingdom²⁰ and in the European Union where studios will also need to be compliant with regulations like the EU Artificial Intelligence Act governing the safety of models, and the General Data Protection Regulation governing how they collect and store data that may be used for training.²¹

Fully private models are likely too expensive for studios

In Deloitte’s 2024 TMT Predictions, we discussed the rise of private gen AI models to help avoid some of the challenges with public models and to gain more control over outputs.²² Studios could avoid the liabilities and copyright challenges of public models by training their own models on their own IP.

But training generative models has become very expensive—around US\$100 billion to train a leading-edge model, by some estimates²³—and the costs of inference and retraining can grow with usage. Open-source solutions (often more accurately referred to as “open weight”) may defray some costs, but their training sets are opaque, and costs are still high.²⁴ Studios may be challenged to attract expensive technical talent able to build such models—talent that may be more inclined to work with hyperscalers able to pay premium salaries. Additionally, investing in today’s models could require updates within six months due to the rapid pace of model development. To build more effective private models, studios and investors may have to think and act more like tech companies, building and maintaining ecosystem relationships with—and paying rents to—tech providers. For these reasons, studios may be less likely to train their own models without considerable shifts in economics.

However, the year ahead could see a flurry of partnerships between studios and providers that could share the cost burdens more equitably.²⁵ In such partnerships, a third party could provide a pretrained model and interface that can then be further trained and customized with content owned by a studio. The model could then deliver generative content that follows the aesthetic of a studio, for example, or includes their signature characters and set pieces. Additionally, studios might be able to control against potential IP concerns by showing derivations from their own content. Still, providers of such capabilities could be expected to show greater maturity of the tools and outputs.

Using gen AI tools to help optimize the studio business

In the coming year, studios are expected to experiment with generative AI content creation, but they will likely move more quickly to understand how gen AI can better enable and optimize more of their business. Generative AI may be able to help automate and augment contract negotiations, talent and workforce management, finance and accounting, and media operations like localization, marketing and promotions, and storage and distribution.

Studios are likely to absorb some of these capabilities through the software and software-as-a-service solutions they already use. Smaller companies are also emerging to tackle time-consuming—and costly—parts of the preproduction process. Generative AI can accelerate script evaluation, break them down and assign them to production schedules, and even “scout” for potential locations that could support the script.²⁶ Generative AI could also unlock archives of content, for example, by “watching” old films and tagging them for actors, themes, and moods. This could then help enable streamers to dynamically resurface and monetize old content to meet more personalized recommendations or trending moments.²⁷

To help accelerate and amplify content distribution, some studios are now leveraging language and voice models that enhance translation and dubbing so it can reach broader global audiences.²⁸ These tools enable high-fidelity voicings that can be highly expressive and emotional, fine-tuned by users.²⁹ This could be a boon for content creators and distributors that are addressing global markets, both in exporting content to them and importing it from them. Leading platforms for user-generated content creation have also extended these capabilities to their creators.³⁰

Gen AI dubbing and translation could also help enable greater sharing of cultures, potentially generating large hits that might otherwise remain as local phenomena. Analysis of the responses to [Deloitte's 2024 Digital Media Trends](#) survey showed that 66% of Americans surveyed enjoy watching TV shows or movies that help them learn about cultures different from their own.³¹ Generative AI could not only help media companies grow their margins and compete more effectively, but it could also bring audiences closer together.

Smaller companies are also emerging to tackle time-consuming—and costly—parts of the preproduction process.

Bottom line

Like most companies, studios, streamers, and creative talent are both fascinated and concerned about the capabilities of gen AI. For studios exploring generative AI content creation in the year ahead, the dominant driver of adoption will likely be the magic and creativity it seems to enable—the strange fever dream of frontier models remixing human creativity into new forms. They will also be driven by the concern that new forms of media could emerge from outside the Hollywood ecosystem.

More independent content creators are demonstrating what can be done with the latest synthetic media capabilities rapidly entering the market. Hollywood studios once enjoyed controlling the scarcity of content and distribution, but now these are both abundant and democratized.³² In the year ahead, the sense of looming content disruption will likely only grow.

Every month or so there are new developments in frontier models advancing their capabilities further along the path towards human intelligence, creativity, and insight. A year ago, it was thought that, by 2030, a major blockbuster film would likely be generated almost entirely from AI.³³ In 2025, that lofty goal may seem a bit more achievable.

In the meantime, content owners will likely work to shore up the competitive moat around their IP, pursuing more litigation against, and regulation of, public models for perceived copyright violations. Regulators could require leading model providers to prove their training sets do not infringe on pre-existing content rights. Most big studios will likely resist offers from model providers to license their content catalogs into their public training sets, likely preferring to partner with smaller companies able to build more bespoke and protected models around studio IP—if the economics are favorable.

At the macro level, generative AI has required enormous capital intensity that could slow growth if a path to broad economic value isn't revealed within the next year or so.³⁴ (See this year's [TMT Prediction about on-device generative AI](#).) However, if the next generation of frontier models can overcome existing challenges, capabilities could advance quickly. Efforts will likely emerge to reduce the costs to train and run models, and to help reduce the amount of data needed.

Big studios are also large enterprises that will likely adopt more gen AI capabilities that are focused on cutting costs, optimizing their businesses, raising productivity, and expanding and accelerating their reach to customers. In [Deloitte's 2024 State of Generative AI in the Enterprise](#) survey, 42% of executives surveyed report efficiency, productivity, and cost reductions as their single most important benefit achieved from using AI; and 58% reported a range of other benefits such as increased innovation, improved products and services, or enhanced customer relationships.³⁵ There appears to be growing interest in applying these capabilities to modern businesses.

A rising tide lifts all boats, as the saying goes. Gen AI tools seem poised to help more smaller companies and creators to achieve the kinds of productivity and levels of quality once reserved solely for the largest companies. Smaller studios and independent creators could become much more capable, while still being relatively free of the risks and cost overheads born by larger studios. The biggest studios among them may need to lower their costs and accelerate their time to market if they hope to compete—not just with each other, but also with user-generated content platforms, social media, and gaming. Production and distribution may be less scarce, but attention remains a finite resource.

By	Chris Arkenberg United States	Danny Ledger United States
	Ricky Franks United States	Kevin Westcott United States

Endnotes

1. George Szalai, “[Studio profit report: A year of major transition](#),” *The Hollywood Reporter*, April 24, 2024.
2. George Szalai, “[Studio profit report: Disney dives as Sony soars, Paramount rises](#),” *The Hollywood Reporter*, Feb. 24, 2024.
3. This prediction is based on our analysis of earnings reports from leading streaming video providers and other available industry information.
4. Chris Arkenberg, “[Will generative AI challenge authenticity in social media?](#),” *Deloitte Insights*, Dec. 8, 2023.
5. Hannah Murphy, “[Media groups look to AI tools to cut costs and complement storytelling](#),” *Financial Times*, March 26, 2024.
6. David Smith, “[‘We’re going through a big revolution’: How AI is de-ageing stars on screen](#),” *The Guardian*, Feb. 6, 2023.
7. Ibid; Murphy, “[Media groups look to AI tools to cut costs and complement storytelling](#).”
8. Alon Yaar, “[What’s next for AI video generation](#),” *AI Business*, Aug. 6, 2024.
9. Lauren Leffer, “[Everything to know about OpenAI’s new text-to-video generator, Sora](#),” *Scientific American*, March 4, 2024.
10. Taylor Lorenz, “[The ‘Beastification of YouTube’ may be coming to an end](#),” *The Washington Post*, March 30, 2024,
11. Dennis Ortiz and Kenny Gold, “[Gen AI and the creator economy: How creators are looking to leverage AI and what this means for brands](#),” *Deloitte*, accessed Oct. 30, 2024.
12. Paul Sweeting, “[Hollywood’s AI concerns present new and complex challenges for legal eagles to untangle](#),” *Variety*, April 17, 2024.
13. Jennifer Wolfe, “[What would have to happen for gen AI to take over Hollywood? So glad you asked.](#),” *NAB Amplify*, July 5, 2024.
14. Ibid; Sweeting, “[Hollywood’s AI Concerns Present New and Complex Challenges for Legal Eagles to Untangle](#).”
15. Baker & Hostetler, “[Case tracker: Artificial intelligence, copyrights and class actions](#),” accessed Oct. 30, 2024.
16. Natalie Sherman, “[World’s biggest music labels sue over AI copyright](#),” *BBC News*, June 25, 2024.

17. US Copyright Office, “[Copyright and artificial intelligence](#),” March 16, 2023.
18. Lucas Shaw, “[Alphabet, Meta offer millions to partner With Hollywood on AI](#),” Bloomberg, May 23, 2024.
19. Erin Degregorio, “[Hollywood is back to work after strikes, but AI remains in the spotlight](#),” *Fordham Law News*, Jan. 29, 2024.
20. Daniel Thomas and Cristina Criddle, “[UK shelves proposed AI copyright code in blow to creative industries](#),” *Financial Times*, Feb. 4, 2024.
21. Lindsey Wilkinson, “[EU passes AI Act, places first binding rules on generative AI](#),” CIO Dive, March 13, 2024.
22. Chris Arkenberg, Baris Sarer, Gillian Crossan, and Rohan Gupta, “[Taking control: Generative AI trains on private, enterprise data](#),” *Deloitte Insights*, Nov. 29, 2023.
23. Jowi Morales, “[AI models that cost \\$1 billion to train are underway, \\$100 billion models coming — largest current models take 'only' \\$100 million to train: Anthropic CEO](#),” Tom’s Hardware, July 7, 2024.
24. Red Hat, “[What is an open-source LLM?](#),” July 1, 2024.
25. Kyle Wiggers, “[Generative AI startup Runway inks deal with a major Hollywood studio](#),” TechCrunch, Sept. 18, 2024.
26. Lauren Forrestal, “[Filmustage leverages AI to break down film scripts, create shooting schedules and more](#),” TechCrunch, March 20, 2023; Lauren Forrestal, “[Avail rolls out its AI summarization tool to help Hollywood execs keep up with script coverage](#),” TechCrunch, Dec. 7, 2023.
27. Emma Cosgrove, “[Nvidia, Amazon, Microsoft, and Paramount execs discuss the use of AI in Hollywood. Here are 9 startups they're watching.](#),” *Business Insider*, July 24, 2024.
28. *The Economist*, “[The dawn of the omnistar](#),” Nov. 9, 2023.
29. Audrey Shomer, “[The state of generative AI in Hollywood: A special report](#),” *Variety*, June 3, 2024.
30. Andrew Hutchinson, “[YouTube announces expansion of auto-dubbing to more creators and languages](#),” SocialMediaToday, Sept. 19, 2024; Julia Walker, “[How Meta’s AI dubbing breaks down language barriers](#),” PR Week, Sept. 30, 2024.
31. Jana Arbanas, Jeff Loucks, Brooke Auxier, Kevin Westcott, Chris Arkenberg, and Bree Matheson, [2024 Digital Media Trends](#), *Deloitte Insights*, March 20, 2024.
32. Michael D. Smith, “[Lessons from Hollywood’s digital transformation](#),” *Harvard Business Review*, Dec. 16, 2021.

33. Jackie Wiles, “*Beyond ChatGPT: The future of generative AI for enterprises*,” Gartner, Jan. 26, 2023.

34. David Cahn, “*AI’s \$600 billion question*,” Sequoia, June 20, 2024.

35. Deloitte, *The State of Generative AI in the Enterprise*—Moving from potential to performance, June 2024.

Acknowledgements —

Authors would like to thank **Howie Stein** and **Ankit Dhameja**.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

Reevaluating direct-to-consumer: The shift toward video aggregators

Video content creators may need more distributors to reach their total addressable market

ARTICLE • 9 MINUTE READ

Deloitte predicts that streaming on demand “stacking”—the consumer practice of subscribing to multiple, standalone video-on-demand services—will decline in 2025. The average number of subscriptions in each “stack” will likely peak at different levels depending on the market, at about four per consumer in the United States and just over half that in the European market.¹ By implication, the aggregate number of standalone subscriptions per market will likely fall, even if revenues from streaming video on demand (SVOD) may still increase due to price raising, password-sharing crackdowns, and bundling.

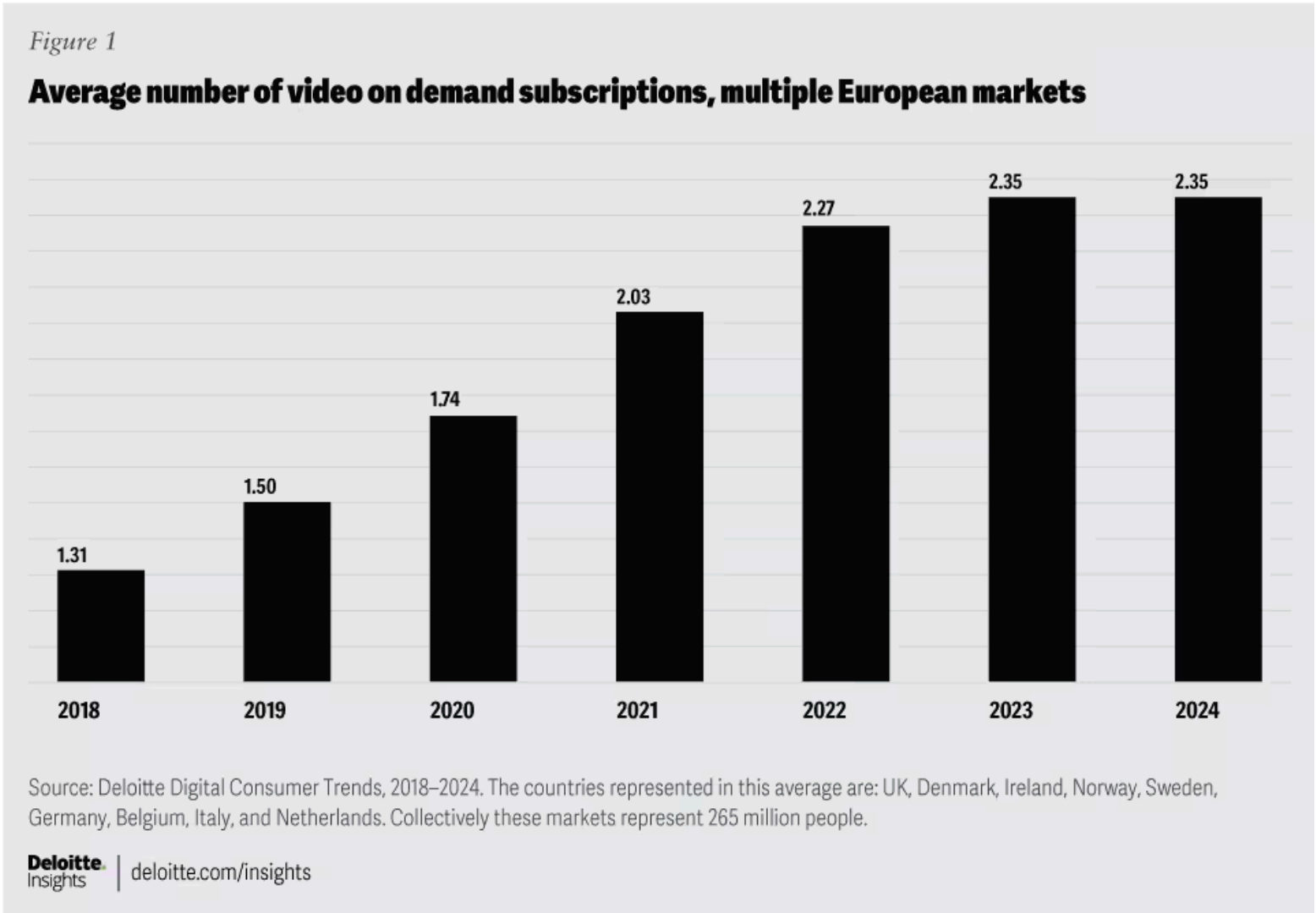
The peak is an expected consequence of a reevaluation of the viability of a video industry marketplace consisting primarily of many dozens of individual direct-to-consumer (DTC) subscription video providers,² with each household purchasing multiple subscriptions in place of a single pay TV subscription. Instead, the likely direction for the industry could be a return to aggregation of content from different providers. This is an approach that pay TV providers have traditionally offered, but which has become considered outmoded.

In the medium term, we expect that the video industry may end up consisting of a handful, most commonly a duo or trio, of standalone SVOD players per national market, along with aggregators. The companies doing the aggregating are expected to be legacy pay TV companies, or telcos, or tech platforms, or the largest SVOD players. In the United Kingdom, as of September 2024, 43% of its SVOD subscribers surveyed purchased at least one of their services via another party (pay TV provider, telco, or tech platform). Among smaller SVOD providers surveyed, close to half of all their subscriptions were via an aggregator; among larger providers, about a quarter of subscriptions were indirect.³

Based on our conversations with those in the industry, Deloitte expects each aggregator will offer a version of the pay TV model including some (and occasionally all) of the following: a single account and bill, standard and optional content channels, 12-month (or longer) contracts, an electronic program guide showing all available content, ad sales and playout, and centralized marketing. The return to aggregation will likely accelerate in 2025 but may not be complete for several years.

The industry is approaching peak SVOD stacking

The 2010s were characterized by two adoption trends: a sustained growth in the number of households with access to SVOD services and a steady increase in the number of services used. Deloitte’s research shows a steady increase in the average number of subscriptions in the European market from 1.3 in 2018 to a plateau of 2.35 in both 2023 and 2024 (figure 1). In the United States, Deloitte’s 2024 Digital Media Trends reported that the average number of SVOD services has been steady at four since 2020.⁴



The rationale for the revival of pay TV elements

Given that some fundamental benefits of standalone SVOD were giving consumers control over content choice and contract length and enabling content providers to bypass distributors, moving back to fewer, larger, pricier, and longer-term bundles may feel like a backward step.

But a return to a version of the aggregation model may offer the equilibrium that balances consumers’ and suppliers’ needs.

For consumers, an idealized version of standalone SVOD—that could include multiple relatively low-cost video services, each readily accessible via an intuitive app and each readily canceled at a few weeks’ notice—may, regrettably, be commercially unviable. The more recent reality has been perennially rising subscription prices and password-sharing crackdowns,⁵ overwhelming content choice,⁶ and user interfaces of varying slickness.⁷ Content providers with decades of profitable experience in creating content to sell to distributors on multiple-year contracts may struggle to pivot effortlessly to running all aspects of an end-to-end DTC business, from setting up credit card payments to managing regulatory compliance required to set up an advertising video on demand tier.⁸

For some viewers who remain subscribers, this may be reminiscent of pay TV.

Pathways to reaggregation

There are multiple pathways that could culminate in the new market structure.

Service bundling: For this type of service, bundled SVOD subscriptions with pay TV, telecom, or financial service contracts could offer discounted rates compared to purchasing each service separately. In exchange, subscribers would commit to a minimum contract duration, typically at least one year. A primary rationale for SVOD companies to become part of a bundle is to help diminish churn, which has remained elevated: at about 40% in the United States and at about 20% in the UK market.⁹ An industry analysis estimates that there were 139.3 million cancellations in 2023 in the United States alone.¹⁰ Among those who canceled, about a quarter of those were “serial churners,” having cancelled a service at least three to four times in the prior two years, up from a mere 3% in 2019.¹¹ A fifth of serial churners had canceled seven or more times within 24 months.¹²

Bundling SVOD into, for example, an 18-month traditional pay TV contract (in exchange for a discount on the total price) defers the possibility of cancellation and diminishes seasonal churn. Over half of US consumers would trade a discount for a year-long subscription,¹³ but as of the start of 2024, only 4% of SVOD contracts in the United States were for a 12-month term.¹⁴ For companies currently running standalone video-on-demand services, another appeal of becoming part of a third party’s bundle could be an opportunity to outsource a range of responsibilities including customer acquisition, billing (and management of bad debt), customer support, and advertising sales.

As of 2024, there was already a fair degree of service bundling, and this is likely to become more extensive through 2025 and beyond. The benefits for pay TV companies to integrate SVOD into their offerings can vary by market, but a common motivation is to lock in popular content packages as part of their offerings to help reduce their own churn. Adding SVOD could also increase gross revenue among companies that are seeing only modest growth for their core services.¹⁵

In the United Kingdom, as of 2019, all of Sky’s pay TV packages include Netflix’s ad-funded tier by default,¹⁶ and a single search bar is available across all subscribed content. In France, pay TV channel Canal+ offers Disney+ and Paramount+ for all subscribers, and there are various packages offering multiple SVOD brands.¹⁷ In Central Europe, it’s been estimated that 25% of all SVOD subscriptions are indirect, sourced via pay TV or telco.¹⁸ In the United States, Xfinity broadband users can add a bundle called Streamsaver, which comprises Apple TV+, Netflix Standard with ads, and Peacock Premium with ads for a 30% saving.¹⁹ Twenty-five percent of online video subscriptions globally are expected to be via telcos by 2028, up from 20% in 2023.²⁰

For telcos, adding popular SVOD services at a discount can also help improve retention, particularly in markets in which there may be little perceived variation in network performance between carriers.

Some banks also bundle SVOD into their subscription services. As of August 2024, Barclays in the United Kingdom offered Bank Account + Blue Rewards customers complimentary Apple TV+.²¹

Some smaller SVOD services are pivoting from standalone DTC to add-on channels distributed by aggregators or exiting some markets altogether.²²

Media aggregation: Services get aggregated, most typically by selling multiple formerly disparate services at a discount relative to individual prices. For example, in the US market, a Disney+ Max and Hulu bundle was made available offering a discount of up to 38%.²³ These bundles may be purely video, or video and other media (music, games, or news). As these become more popular, the individual services may be discontinued or priced to discourage standalone purchases.

Aggregating video services should improve ease of use, for example, by providing a single search bar and a single electronic program guide across all content available to the viewer, based on their subscriptions. By contrast, with standalone services, there can be friction in exiting one service and then opening another, particularly when viewing on an older or budget TV that may have less powerful processors. Deloitte's research has found that almost half of US consumers surveyed would spend more time on streaming services if content was easier to find.²⁴ About three-quarters of surveyed Gen Zers and millennials in the United States say they would like to be able to search across all services they have access to.²⁵ Deloitte UK's research indicates that a major driver of cancellation is not being able to find anything to watch, a paradoxical result given the current all-time abundance of titles.²⁶

Bundling of SVOD services could help mitigate churn. According to industry analysis, subscribers to one bundle (Disney+, Hulu and ESPN+) were 59% less likely to churn than those subscribing to Disney+ alone.²⁷ Deloitte's research suggests that consumer tolerance for further price increases may be reaching its limit: Almost half of US consumers polled in the fourth quarter of 2023 stated that they would cancel an SVOD service if it increased in price by US\$5 per month.²⁸

Permanent churn: Other video-on-demand (VOD) services, including free broadcaster VOD (BVOD) which is popular in Europe, or video-sharing services like YouTube (which is predominantly consumed for free), may, for some households, usurp paid-for SVOD. A major driver of SVOD churn is cost. Over recent years, the cost of subscription has become an increasingly prominent factor with 24% of UK respondents who canceled SVOD services citing "subscription was too expensive" as the reason for cancellation. By 2024, this had risen to 31%.²⁹

There may be an increased offer of FAST (free ad-supported television) services from companies that were formerly subscription-only. For example, Amazon, which started offering SVOD in 2008, launched the FAST channel Freevee in the United States in 2022.³⁰ Crunchyroll, an anime SVOD provider since the mid-2010s,³¹ launched a FAST service in 2023.³² There may also be increased consumption of existing user-uploaded and typically free services such as YouTube. Viewing of YouTube on the television set is expected to increase by 90% between 2024 and 2029, rising from 12 to 22 minutes per day.³³ Some of this growth may come from the displacement of SVOD viewing.

Bottom line: The business of television is evolving ... again

The industry is undergoing a decades-long shift from broadcast to IP-based delivery.

Historically, this was characterized by the genesis and growth of the standalone SVOD market, which enjoyed growth in adoption in the 2010s, a period where the offer was novel, competitors were fewer, and the SVOD subscriptions were modestly priced and often shared.

But the 2020s are proving less benign, churn has been high, and growth has proven more difficult than expected. In a relatively mature market such as the United Kingdom, SVOD's measured share of viewing has grown languidly of late, from 15.8% to 16.4% between 2023 and 2024: Broadcasters' content- and video-sharing sites have markedly higher shares.³⁴ The standalone SVOD market is unlikely to become predominant, and aggregation of a type reminiscent of traditional pay TV is likely to grow over the coming years.³⁵

Standalone SVOD players should consider the role they want to have given this new model. A few players have attained sufficient scale and breadth of capability (from billing to user interface design to compression) such that they may want to remain focused on full-service SVOD (that is with one player being fully integrated, from content creation to ad sales to customer management). The exception may be in markets where third-party distribution makes better commercial sense.

Some SVOD players may want to take the lead on aggregating other players’ content in addition to their own: They can create bundles with their content at the core. They could also market and host other companies’ channels like traditional pay TV. However, many players may determine that their focus may simply be on selling content to the highest bidder—a role that they’ve historically been successful at.³⁶

The nub of this bottom line is that pure direct-to-consumer (DTC) is often challenging to deliver, particularly for companies that may need to add in capabilities and adjust corporate culture: A company that has spent decades thriving at generating content for others to distribute may not immediately be able to thrive at DTC. The difficulty of DTC as a sole business model is a constraint that applies to most industries and is not specific to video. There are very few companies of scale that have managed to pivot entirely to DTC. Distributors are key to most major consumer brands, and this tenet is unlikely to change.

The fact that standalone SVOD may not be the predominant industry model should not be cause for complacency for broadcasters, which, in many markets, retain the majority of viewing hours across all measured viewers. In Europe (based on 42 markets), the average viewing of traditional TV was at 3 hours 16 minutes per day in 2023; but among younger viewers, it was just 1 hour 12 minutes.³⁷ Broadcasters should work together to ensure their offer has a resilient and growing appeal to younger audiences.

By

Paul Lee
United Kingdom

Rupert Darbyshire
United Kingdom

Eliza Pearce
United Kingdom

Kevin Westcott
United States

Endnotes

1. Kevin Westcott, Jana Arbanas, Chris Arkenberg, and Jeff Loucks, “[Streaming video at a crossroads: Redesign yesterday’s models or reinvent for tomorrow?](#)” *Deloitte Insights*, March 20, 2024; Deloitte, “[Generative AI: 7 million workers and counting](#),” June 25, 2024.
2. Ampere Analysis, “[Analytics - SVoD](#),” accessed November 2024.
3. This is based on a nationally representative survey of 4,000 respondents ages 16 to 75, commissioned by Deloitte UK and undertaken in August 2024.
4. Westcott, Arbanas, Arkenberg, and Loucks, “[Streaming video at a crossroads](#).”
5. David Pierce, “[Streaming services keep getting more expensive: all the latest price increases](#),” The Verge, Sept. 23, 2024; Emma Roth, “[Disney’s password-sharing crackdown starts ‘in earnest’ this September](#),” The Verge, Aug. 7, 2024.
6. As of June 2023, Nielsen’s Gracenote reported that the number of individual titles available on streaming services was 2,346,171, up from 1,882,401 in July 2021. These numbers span the range of content available in the United States, the United Kingdom, Canada, Mexico, and Germany; see: Nielsen, “[Data-driven personalization: The future of streaming content discovery](#),” accessed November 2024.
7. Selome Hailu and Jennifer Maas, “[From ‘glitchy’ HBO Max to ‘overwhelming’ Amazon Prime Video, Hollywood insiders spill on their \(least\) favorite streaming interfaces](#),” Variety, April 11, 2023.
8. Andrew Blustein, “[How GDPR, ad fatigue and content costs are complicating the now-global streaming wars](#),” The Drum, Sept. 4, 2019.
9. Westcott, Arbanas, Arkenberg, and Loucks, “[Streaming video at a crossroads](#)”; Deloitte, “[Generative AI](#).”
10. Antenna, “[The rise of the show chaser](#),” accessed November 2024.
11. Ibid.
12. Ibid.
13. Westcott, Arbanas, Arkenberg, and Loucks, “[Streaming video at a crossroads](#).”
14. Antenna, “[Understanding the relationship between annual plans and promotions](#),” accessed November 2024.
15. IDC Research, “[IDC forecasts slower growth for global telecommunications market: Could AI help telcos to maintain healthy margins](#),” May 3, 2024.
16. Sky, “[Official website](#),” accessed November 2024.

17. Canal Plus, “*S'abonner à Disney+ avec les offres CANAL+,*” accessed November 2024; Georg Szalai, “*Paramount+, Canal+ expand partnership in France,*” The Hollywood Reporter, Aug. 21, 2024.
18. Omdia, “*Omdia unveils surging SVOD growth in CEE through strategic pay TV and telco partnerships,*” June 12, 2024.
19. Xfinity, “*Streaming services: Stream one, stream all,*” accessed November 2024.
20. Bango, “*Super bundling: What telco leadership needs to know about securing a wider role in the subscriptions market,*” accessed November 2024.
21. Barclays, “*Current accounts: Barclays Bank Account + Blue Rewards,*” accessed November 2024.
22. Viaplay Group, “*Q2 2024 interim report January-June,*” press release, July 18, 2024.
23. Disney Plus, “*New Disney+, Hulu, Max bundle is now available in ad-supported and ad-free plans,*” July 25, 2024.
24. Westcott, Arbanas, Arkenberg, and Loucks, “*Streaming video at a crossroads.*”
25. Ibid.
26. Deloitte, “*Generative AI.*”
27. Ampere Analysis, “*Subscribe, cancel, repeat: 42% of US consumers are SVoD ‘resubscribers’,*” July 8, 2024.
28. Westcott, Arbanas, Arkenberg, and Loucks, “*Streaming video at a crossroads.*”
29. Deloitte, “*Generative AI.*”
30. Christian de Looper, “*Amazon Freevee: Everything you need to know about the free streaming service,*” Amazon UK, May 10, 2023.
31. Janko Roettgers, “*Chernin, AT&T Set Brand for New Online Video Venture: Ellation (Exclusive),*” Variety, Aug. 3, 2015.
32. Crunchyroll News, “*Crunchyroll launches 24/7 anime channel in the US,*” Oct. 11, 2023.
33. Enders Analysis, <https://www.endersanalysis.com/reports/video-viewing-forecasts-slowdown-change>
34. Share of viewing is calculated based on 12-month rolling averages from October 2022 to September 2023 through August 2023 to July 2024. All data is from Barb’s monthly viewing summary; see: Barb, “*Monthly viewing summary,*” accessed November 2024. Barb’s methodology for capturing viewing patterns is explained in: Barb, “*What is the Barb panel and why is it important?*” accessed November 2024.

35. Richard Waters, *The next phase of the streaming wars*, Financial Times, June 6, 2024 (subscription required)

36. Etan Vlessing, “*Sony CFO: Without a streaming platform, we’re free to sell films and shows ‘to the highest bidder’*,” The Hollywood Reporter, March 6, 2023; Diane Haithman, “*Why Sony’s streaming deals with Netflix and Disney make sense for everyone*,” The Wrap, May 10, 2021.

37. *Audience Trends Television 2024*, European Broadcasting Union Media Intelligence Service, August 2024 (registration required)

Acknowledgements

The authors would like to thank **Stacy Hodgins, Beth Rae Rosenstein, Helen Rees, Ben Stanton,** and **Duncan Stewart** for their contributions to this article.

Cover image by: **Jaime Austin; Getty Images, Adobe Stock**



Wireless telecom consolidation speeds up ... where regulators allow

In many markets, smaller wireless telecoms see slow growth, low profits, and have debt to repay. M&A, specifically combining assets or even entire consumer-facing companies, may help where it gets approved by regulators.

ARTICLE • 7 MINUTE READ

Deloitte predicts that more in-market telecom mergers will get approved in 2025 and beyond, at first led by the European Union.¹ In many regions and countries, wireless telecom markets are fragmented, and some players are subscale. Historically, regulators have focused on maximizing competition by encouraging as many consumer-facing players as possible, which helps keep consumer prices lower. Increasingly, however, those who advise regulators are suggesting that future network growth, features, security, and resilience might be better maintained by permitting consolidation within markets.

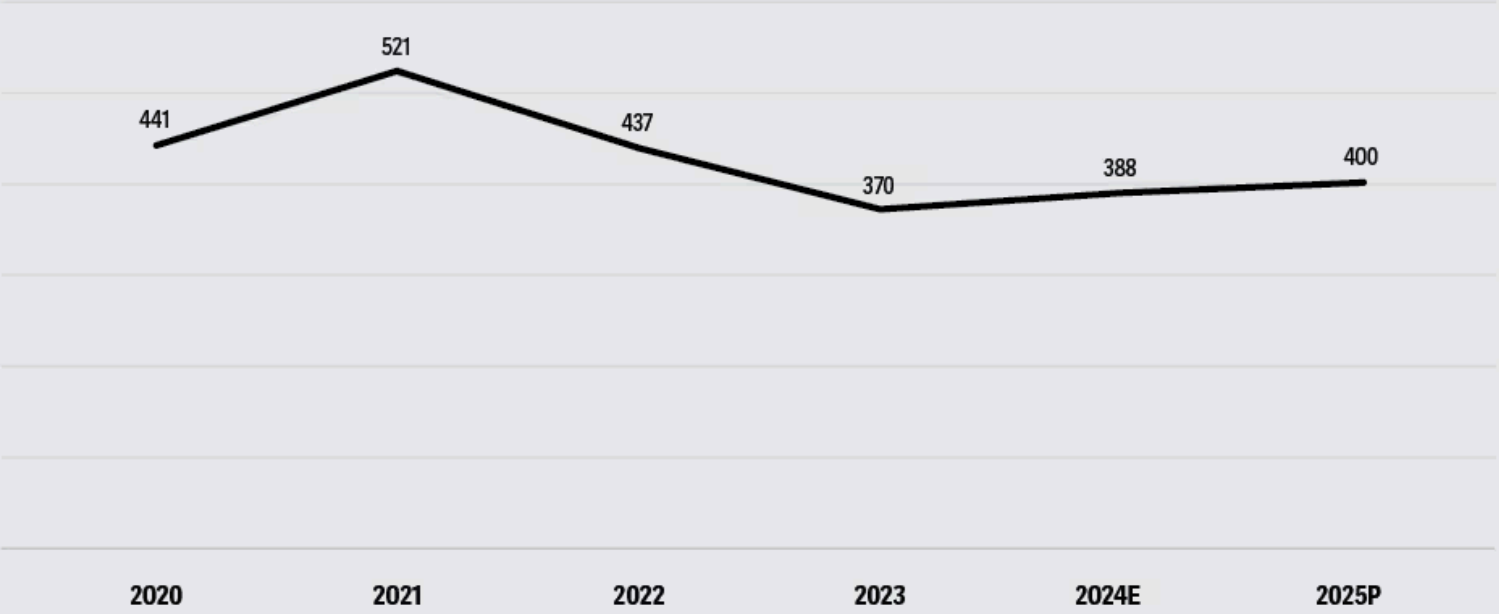
Deloitte predicts that there will be about 400 telecom mergers and acquisition deals in 2025, more or less in line with the deal volume over the last five years (figure 1).² That may not be that interesting—what is interesting is the kind of M&A deals we predict we'll see more of: actual operator consolidation.

There are many kinds of telecom M&A deals, but at a high level, no single kind of M&A dominates (figure 2).³ There are both wireless and wireline deals, and as a percentage, the various deal types are fairly consistent over time, although the number of data center deals has picked up recently, likely driven by activity around AI data centers.⁴

Figure 1

The volume of telco M&A in 2025 will be in line with the previous five years

Average number of deals



Notes: E denotes estimated value for year in progress, and P is the Deloitte Global predicted value.

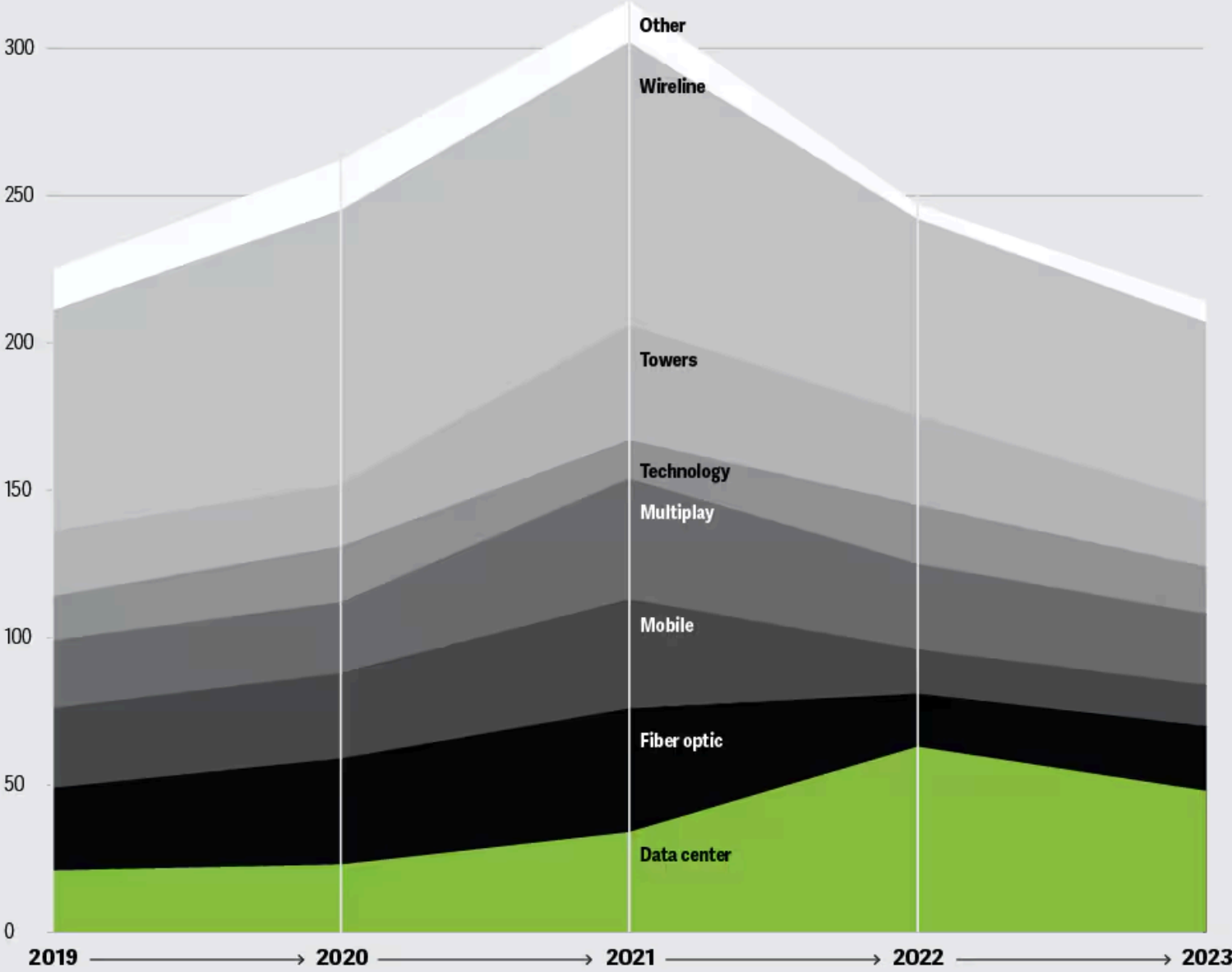
Source: Graphic prepared by Deloitte based on the data from S&P Global Market Intelligence - S&P Capital IQ and CB Insights. The above figure calculates the average number of M&A deals in the telecom industry annually. The data encompasses full-year figures for 2020, 2021, 2022, and 2023, along with Deloitte's extrapolated estimates based on partial data from 2024.

Deloitte Insights | deloitte.com/insights

Figure 2

Diverse M&A activity spans multiple telecom subsectors, with a notable recent growth of data center deals

Communications provider M&A deal volume, 2019 to 2023 (number of deals annually)



Source: Omdia via CSI Magazine, July 23, 2024.

Deloitte Insights | deloitte.com/insights

Some forms of consolidation or carve-outs have been underway for years and are likely nearing the end of their growth phase: As an example, 97% of all cellphone towers in the United States and Mexico in 2023 were run by standalone tower companies rather than the telcos (up from 65% in 2016), while in Europe, tower companies nearly doubled their share in 2023 to 70%, up from 36% in 2016.⁵

Equally, for years, there has been consolidation and M&A activity of wireline networks (copper, fiber optic, and coaxial), and back-office software systems such as billing and operations, field service fleets, and data centers.⁶ Increasingly, there are wireless consolidations of various kinds.

As examples of wireless network consolidation, in Canada, two major wireless providers have shared the radio access network (RAN) since 2009.⁷ In Malaysia, where there used to be three separate and distinct wireless networks, the government decided to have a single national 5G network in 2021, although they have now decided to have a second network.⁸ In Brunei, there are three mobile companies offering services to consumers and enterprises, but all three of them use the same radio network provided by Unified National Networks Sdn Bhd.⁹ In 2024, two Australian operators agreed to share the 4G and 5G RANs.¹⁰

However, in these cases and others, the number of “retailcos” (companies that offer communications services to consumers and enterprises) have stayed more or less constant: For buyers of telecom services, there are still three or more (sometimes many more) companies competing with each other.¹¹

What is relatively new and is the core of our prediction: Governments and regulators globally have been allowing mergers. Since 2020, there have been 13 telecom mergers or joint ventures that have decreased the number of customer-facing players, which have been approved or are in the process of approval by governments and regulators:

- **The Americas.** Six mergers (three in the United States and one each in Canada, Chile, and Colombia)¹²
- **Asia Pacific.** Five mergers (Indonesia, Malaysia, Thailand, Taiwan, and Australia)¹³
- **Europe.** Two mergers (the Netherlands and Spain)¹⁴

Observers are awaiting a final decision from the United Kingdom’s regulator on the proposed merger of Vodafone UK with Three UK.¹⁵ It may be worth noting that proposed mergers in Italy and Denmark were denied in recent years.¹⁶

Further, former Italian PM Enrico Letta submitted a report to the EU in April of 2024, explicitly calling for telecom consolidation.¹⁷ In part, he based his recommendations on an EU white paper that discussed the challenges telecoms have in getting returns on their investments in the highly fragmented European markets.¹⁸ One data point suggested why Europe may lead the way in approving consolidation: The average number of subscribers per mobile operator in Europe is 4.5 million, compared to 95 million in the United States, 300 million in India, and 400 million in China.¹⁹ Even more recently, in September of 2024, former European Central Bank president Mario Draghi (who is also a former Italian prime minister) submitted a 69-page report, in which there was a section supporting market telecom consolidation.²⁰

In most countries, there are usually financially strong wireless telecoms in the No. 1 and No. 2 slots, as measured by revenues and subscribers. The third player is often less financially strong, but the fourth operator (or fifth or sixth, depending on the market) can be even less financially strong. These operators in the lower slots often caution that it may not be possible to continue network investments going forward. In these markets, proponents argue to regulators that three robust competitors can benefit consumers, enterprises, and the overall competitive landscape more than two dominant players and two or more weaker ones.

There appear to be a number of wireless telecom markets in Europe and elsewhere that could see customer-facing consolidation as a result.

One important factor behind the regulatory stance shifting is that connectivity choice has never been higher. As we have written about in the 2024 and 2023 Telecommunications Outlooks, consumers and enterprises have seen a surge in choice over the last few years. These choices include but are not limited to:

- **Fixed wireless access competition for home broadband service.** Deloitte predicts that over 30 million homes will connect via fixed wireless access in 2025, up 20% from the current year.²¹
- **Low Earth orbit satellite competition for home broadband, especially in rural and remote areas.** Over three million homes are currently connected worldwide by this approach, with multiple new networks expected to launch in 2025 and 2026.²² It should be noted that low Earth orbit satellites are less of a factor in countries with high population density and unchallenging geography such as that found in some of Europe. They're more of a factor in less dense markets, or markets with mountains, deserts, or many small islands such as those in Asia Pacific, Africa, and the Americas.²³
- **3G, 4G, and 5G still in use.** In some markets, there are multiple networks in use, which provides more choice and competition for consumers.²⁴
- **Mobile virtual network operators on the verge of change.** Mobile virtual network operators have been around for 25 years but have recently reached an inflection point and started taking more of a share of recent mobile subscriber additions.²⁵ As an example, there are about 14 million US wireless customers for cable mobile virtual network operator offerings as of 2024. It's worth noting that these are succeeding in part due to the viability of Wi-Fi. Both home Wi-Fi and Metro Wi-Fi hotspots allow US cable companies to offload 87% of all wireless data consumption.²⁶

Bottom line

For telecoms, maintaining a robust and competitive wireless network over the next few years is likely to be less expensive than in the past few years. A costly part of the 5G network build, buying new equipment and purchasing spectrum, is now mainly over for many operators in developed world countries. RAN spending, after peaking in 2022, is now dropping at double digits for the foreseeable future.²⁷ Most global telcos who built 5G non-stand-alone networks initially are not spending as much to upgrade to 5G stand-alone networks.²⁸ And there are no signs that 6G is coming before 2030, if then. As a result, the annual capex intensity for the industry (which hit a 10-year peak in 2022 at 17.8%)²⁹ is predicted by Deloitte to decline further to the 15% to 16% range from 2025 to 2029. This is positive for network operators, although may be a challenge for the RAN original equipment manufacturers.

On the other hand, telecoms often struggle to make money from 5G and other new services, except for fixed wireless access. As stated in TMT Predictions 2024, consumers have likely reached an “era of enough,” and most are unwilling to pay more for higher speeds.³⁰ Further, potential monetization sources such as premium services to support consumer virtual reality or augmented reality glasses, private 5G networks for enterprises, self-driving cars, or telesurgery are niche at best. Some telecoms have started getting into telecom-adjacent, value-added services (healthcare, agtech, security, and more), but so far, the impact of these on most bottom lines appears to be fairly small. Some telecoms could consider getting into the gen AI data center business as a way of generating additional revenues and profits, but these are usually the larger players in the market and not an option for the usual third- and fourth-place players who are more likely to merge.³¹ Further, many value-added services often require scale to succeed, and as noted earlier, most European and smaller Asian telecoms lack that scale due to the fragmented market.

At a high, global level, there are often two different regulatory authorities who need to approve wireless mergers. There is an industry regulator (Ofcom in the United Kingdom, the Federal Communications Commission in the United States, and both EU-level and national-level industry regulators in the EU)³² and a competition regulator (the Competition and Markets Authority in the United Kingdom, the Federal Trade Commission in the United States, and the Directorate-General for Competition in the EU, plus national-level competition authorities). There are similar divisions of regulatory authority across much of Asia Pacific.³³

Once again, at a high level, the industry regulators have generally been more open to merging wireless players within a country, while the competition regulators have been the larger challenges. We believe that may be changing in some jurisdictions, especially spurred on by recent papers and letters in Europe encouraging the merger of subscale wireless players.

That said, regulators will likely take their time and investigate closely: Multiple, recent mergers took 24 to 36 months to close.³⁴ However, the percentage of mergers that get approved overall could increase, if our prediction is correct.

Sometimes regulators approve mergers unconditionally, but they also sometimes have conditions, such as divestitures, pricing guarantees, or commitments for future investments or providing 5G coverage, prior to approving a merger.³⁵

By

Duncan Stewart
Canada

Dan Littmann
United States

Jack Fritz
United States

Ariane Bucaille
France

Endnotes

1. Deloitte analysis of recent events in Europe, specifically recent letters and white papers from Draghi and Letta (see endnotes 16 and 19).
2. Deloitte analysis of historical merger and acquisition trends, combined with early publicly available signs of deal activity.
3. *CSI Magazine*, “[Telecom consolidation: Over 500 M&A deals in five years](#),” July 23, 2024.
4. Duncan Stewart, Dan Littmann, Girija Krishnamurthy, and Matti Littunen, “[Telecoms tackle the generative AI data center market](#),” *Deloitte Insights*, Sept. 16, 2024.
5. Stephanie Price, “[For sale: Canadian read-throughs from global telecom asset sales](#),” CIBC Capital Markets, Aug. 14, 2024.
6. ArdorComm News Network, “[Telecom M&A activity witnesses surge: 514 deals from 2019 to 2023](#),” ArdorComm Media Group, July 25, 2024.
7. Sue Marek, “[Marek’s take: 5G network sharing may be the answer](#),” Fierce Network, May 14, 2021.
8. Affandy Johan, “[5G in Malaysia – Single wholesale network driving regional leadership](#),” Ookla, March 17, 2024.
9. Digital Regulation Platform, “[Changing the operating model: the creation of UNN in Brunei Darussalam](#),” Aug. 27, 2020.
10. Australian Competition & Consumer Commission, “[Optus Mobile Pty Ltd and TPG Telecom Limited proposed network and spectrum sharing](#),” Sept. 5, 2024.
11. Megan Emfosi Meena and Jiaying Geng, “[Dynamic competition in telecommunications: A systematic literature review](#),” *Sage Open*, April 26, 2022.
12. Detecon Spotlight, “[Telco mergers and acquisitions](#),” 2022; M&A Community, “[12 significant telecom mergers and acquisitions over the last 20 years](#),” Sept. 17, 2024; Henri Capin-Gally, Sergio Márquez García Moreno, and Bill Parish, “[Energy and telco deals power Mexican M&A](#),” White & Case, March 6, 2024; Geussepe Gonzalez, “[Access Alert: Colombian telecoms industry faces shakeup with potential Millicom-Telefonica merger](#),” Access Partnership, Aug. 8, 2024.
13. Julber Osio, “[Asia-Pacific telcos consolidate to compete with market leaders](#),” S&P Global, May 25, 2023; Tom Leins, “[Deal-making Down Under: Oceania’s wave of telecom M&A](#),” TeleGeography, March 24, 2021.
14. Jan Frederik Slijkerman, “[Telecom Outlook: Will we see more mergers and buyouts in 2022?](#)” ING, Jan. 28, 2022; Jan Frederik Slijkerman and Diederik Stadig, “[Telecom tycoons on the move?](#)” ING, Feb. 1, 2024.

15. Competition and Markets Authority, “[How we are investigating the Vodafone / Three potential merger](#),” Sept. 13, 2024.
16. Slijkerman and Stadig, “[Telecom tycoons on the move?](#)” ING, Feb. 1, 2024.
17. Enrico Letta, “[Much more than a market](#),” Consilium, April 2024.
18. European Commission, “[White paper - How to master Europe’s digital infrastructure needs?](#)” Feb. 21, 2024.
19. Hamish White, “[Europe’s looming mobile crisis](#),” Technative, April 2024.
20. Mario Draghi, “[EU competitiveness: Looking ahead](#),” European Union, Sept. 9, 2024.
21. See section “Fixed wireless access: Contrary to popular opinion, adoption may continue to grow” in chapter [Updates: Past TMT Predictions’ greatest hits and \(near\) misses](#).
22. Ongoing Deloitte analysis of current and proposed low Earth orbit satellite networks.
23. Deloitte author conversations with communications providers in North America, Europe, Africa, and Asia.
24. Deloitte analysis of developing world wireless networks.
25. Piran Partners, “[Seizing the future of telecoms: The continuing ascendancy of MVNOs](#),” April 12, 2024; Puneet Takyar, “[The journey of MVNOs](#),” Comviva, April 1, 2021.
26. Jeff Baumgartner, “[Cable snared nearly half of US mobile line adds in Q3 – analyst](#),” Light Reading, Nov. 16, 2023.
27. Juan Pedro Tomás, “[Global RAN market faces challenging scenario in Q2: Dell’Oro](#),” RCR Wireless News, Aug. 19, 2024.
28. Deanna Darah, “[5G NSA vs. SA: How do the deployment modes differ?](#)” TechTarget, July 25, 2024.
29. Matt Walker, “[Telco capital intensity hits 10 year peak in 2Q22](#),” MTN Consulting, Sept. 6, 2022.
30. Paul Lee, “[No bump to bitrates for digital apps in the near term: Is a period of enough fixed broadband connectivity approaching?](#)” *Deloitte Insights*, Nov. 29, 2023.
31. Stewart, Littmann, Krishnamurthy, and Littunen, “[Telecoms tackle the generative AI data center market](#),” *Deloitte Insights*, Sept. 16, 2024.
32. DataHub, “[Regulatory authority - Institutional structure](#),” accessed October 2024; Policies, “[Telecommunications national regulatory authorities](#),” European Commission, Jan. 11, 2023.

33. Lynn Robertson, “*Interactions between competition authorities and sector regulators – contribution from business at OECD (BIAC)*,” Organisation for Economic Co-operation and Development, Nov. 18, 2022.

34. Research Notes, “*5G rollout slowed while mobile operators await merger approval. Case in point: the slow rollout in UK*,” Strand Consult, Oct. 5, 2024.

35. Deloitte author’s assessment of multiple approved mergers in North America, Europe, and Asia in the period of 2015 to October 2024.

Acknowledgements

The authors would like to thank **Dieter Trimmel, Hugo Santos Pinto, Prashant Raman, Pankaj Bansal, and Akshay Jadhav** for their contributions to this article.

Cover image by: **Jaime Austin; Getty Images, Adobe Stock**

Updates: Past TMT Predictions' greatest hits and (near) misses

TMT Predictions revisits previous predictions on enterprise edge computing, 5G telecom, women's sports, quantum-resistant encryption, and the geopolitics of open-source semiconductors

ARTICLE • 37 MINUTE READ

New for 2025 is our series of shorter articles on topics we've covered in previous editions. Sneak peek: We did well with past predictions, but nobody bats 1.000. Read on to hear what's new with these oldies:

- Generative AI comes to the enterprise edge: 'On-prem AI' is alive and well
- (Re)defining the investment case for women's sports
- Fixed wireless access: Contrary to popular opinion, adoption may continue to grow
- 5G standalone appears to be at a standstill: Will 6G run late?
- Open RAN mobile networks and vendor choice: Single vendor now, multivendor when?
- Despite quantum's slow start, don't be slow to start your defense with post-quantum cryptography
- RISC-V: Closing the geopolitical gen AI loophole

Generative AI comes to the enterprise edge: 'On-prem AI' is alive and well

Owning their own servers gives companies a more private, secure, flexible, and possibly cheaper IT environment for AI

Duncan Stewart, Karthik Ramachandran, Gillian Crossan, and Jeff Loucks

Deloitte predicts that in 2025, although generative AI via the cloud will continue to be a dominant option, about half of the enterprises worldwide will add AI data center infrastructure on-premises (on-prem)—an example of enterprise edge computing. This could be, in part, to help protect their intellectual property and sensitive data and comply with data sovereignty or other regulations, but also to help them save money. This is a continuation of what we observed in late 2024: About 45% of one source of gen AI chips was going to the hyperscalers, while 55% were from a mix of consumer, internet, and enterprise players.¹ According to our 2024 State of Generative AI in the Enterprise Q2 report, 80% of companies with very high AI expertise report spending more on AI in the cloud. However, 61% are investing more in their own hardware.² We predict enterprise on-prem AI servers could approach a US\$100 billion market in 2025.³

The 2021 and 2023 editions of TMT Predictions discussed how enterprises were expected to invest in edge AI solutions to perform and run computational tasks faster.⁴ Latency (the time it takes for an AI to see an input and get a response back) was a key driver of earlier enterprise edge applications.⁵ Latency does not seem to be a significant driver this time around: Most gen AI requests take thousands of milliseconds to process, so turnaround time isn't usually a challenge.⁶

Instead, the demand for private, sovereign, and secure gen AI is now driving enterprise edge to an altogether new wave of growth.

As enterprises scale their investments and efforts in gen AI,⁷ the cloud providers, hyperscalers, telecom companies, and AI and tech companies are building out data centers to help meet the demand surge and are expected to spend over US\$200 billion in chips and data centers in 2025.⁸ But many global enterprises seem to be adopting a hybrid approach: using both third-party cloud solutions and investing in hardware to do some portion of the training and inference on their own premises, which may provide a more secure, controlled, sovereign, and a flexible IT environment.⁹ Processing data locally versus relying on an external cloud infrastructure can enable enterprises to accelerate response times and address privacy and security issues with regard to gen AI implementations—which Deloitte uncovered in the State of Generative AI in the Enterprise Q3 report.¹⁰

A variety of enterprise use cases and opportunities make gen AI at the edge viable and relevant. For example, banks and financial services companies often prefer to keep their volumes of sensitive data on-premises to address data security concerns and have better control of gen AI models.¹¹ Media and entertainment companies are already using natural language AI-infused solutions to spur creativity in the fields of animation and content creation (for example, writing first draft scripts or prose), gaming, and entertainment (for example, using patterns in movies' subtitle content to enhance recommendation engines for end users).¹²

What does the enterprise edge look like for gen AI in 2025? Some portion of enterprise spending for on-prem gen AI will be toward devices such as employee smartphones and PCs that increasingly are expected to have specialized gen AI chips on them.¹³ But there are other options, and they have implications for how much enterprises might spend for on-prem solutions, as well as adaptations they might need to make in their data closets or data centers, as the new versions tend to use more power, are physically bigger, and sometimes require liquid cooling, which is a relatively new solution.

In 2024, many enterprises had a gen AI box (there were many possible options and vendors) that was about the size of a home printer, weighed about 300 pounds, consumed roughly 10 kW, cost just under US\$500,000, and was capable of approximately 30 PetaFLOPS in processing power.¹⁴ In 2025, some companies will likely buy similar boxes, but some will buy larger and more powerful boxes—larger than a (big) refrigerator, weighing over 3,000 pounds, drawing 160 kW and needing liquid cooling, costing more than US\$3 million, and over 1.3 ExaFLOPS.¹⁵

To be clear, enterprises are not building these machines themselves. Both gen AI chipmakers and multiple-server original equipment manufacturers are offering these rack-scale servers and will install them on-prem.

Non-AI on-prem computing can often be cheaper than cloud, and there's no reason to expect gen AI computing to be different over time.

Spending US\$3 million on an exascale gen AI supercomputer may seem like a lot of money in one way, but it may not be. Large enterprises across all industries could have IT budgets in the billions of dollars. Training large language models can cost from US\$1 million to US\$100 million.¹⁶ And cloud gen AI computing power, whether for training or inference, is hardly free.

It's unlikely that an enterprise would solely use an on-prem gen AI solution. It's likely that many will use the cloud at least some of the time. Further, it's likely that cloud could be the bigger portion of total AI computing. As is often common in IT architectures, hybrid is likely the ultimate outcome: On-prem may offer benefits around security, latency, resilience, and privacy, while cloud may offer benefits around choice, flexibility, scalability, and experimentation.

But for those who want to own their own hardware, does it make financial sense to buy an on-prem AI solution from a return-on-investment perspective? It might: Non-AI on-prem computing can often be cheaper than cloud,¹⁷ and there's no reason to expect gen AI computing to be different over time. Even if the "hard" ROI is not there, the advantages around ownership of IP, privacy, security, and resilience may make on-prem worth it. Moreover, new types of training and inference techniques like split learning and split inference can distribute gen AI workloads across edge devices, optimizing computational needs and reducing latency, as well as addressing privacy and security issues.¹⁸

Bottom line

Some companies may believe that although they have no short-term need for on-prem gen AI computing, over the long term, on-prem may be inevitable. In that case, the time and money they spend now in learning how to best use on-prem processing as part of a hybrid cloud and on-prem approach could justify the cost of the hardware.

One question concerns the possibility of a gen AI bubble. With many players investing in hardware at the same time, with a view that underinvesting may be a bigger risk than overinvesting, it's possible that there will be significant overcapacity, at least in the short term. If that happens, what would enterprise companies that have invested in on-prem gen AI servers do? It seems more likely that they would use their own hardware (which is already paid for and being depreciated), rather than spend on cloud gen AI computing.

Most of the prediction is anchored on how, for example, a major company might have its own on-prem gen AI IT infrastructure to run a part of its processes and operations (banks, auto manufacturers, health care companies, public sector agencies, and so on). That's what we (mainly) mean by edge. But there is an additional market, which some consider to also be edge: the telco edge. As we wrote in September, over 15 telecom companies globally have announced that they are building gen AI data centers, some with capacity they'll use for their own needs, but some of which they're aiming to sell to enterprises: gen AI as a service.¹⁹ Therefore, enterprises could (1) buy from cloud providers, (2) have on-prem hardware, or (3) buy from a gen-AI-as-a-service provider.

Finally, there is the question of sustainability and the inherent trade-offs. An enterprise edge gen AI server is the same in many ways as a box in a hyperscaler's data center. But having a thousand servers in a thousand enterprises can spread out the electrical load and could put less stress on the electrical grid than deploying a thousand (or ten thousand) servers in a single building.²⁰ On the other hand, the companies that run hyperscale data centers tend to be more efficient in running them with power usage effectiveness (PUEs) of 1.2 or 1.3 (and they tend to be able to buy low-carbon energy at scale, which may be harder for a smaller enterprise to access²¹). Meanwhile, the average enterprise data centers' PUE ranges from 1.67 to 1.8, suggesting that a thousand different gen AI boxes in a thousand different enterprises could have a larger carbon footprint than the same number in a hyperscaler data center.²²

(Re)defining the investment case for women's sport

Rising women's sports revenue fuels investor interest and valuation records

Jennifer Haskel, Pete Giorgio, and Amy Clarke

The increasing professionalization and commercialization of women's sports around the world is garnering the attention of fans, sponsors, and—importantly—investors. In last year's edition of this report, Deloitte predicted the women's elite sports market would generate over US\$1 billion in revenue in 2024, a rise of more than 300% on the previous valuation of revenue earned in 2021.²³

This growth in revenue has led to record valuations and a rise in the scale of capital flowing into the sector,²⁴ which has helped increase the visibility, regulation standards, and sponsorship innovation in elite²⁵ women's sports.

In 2025, we expect to see an expanding group of investors—including institutional investors, private equity, and high-net-worth individuals—take note.

In Deloitte UK's report, *Future of Sport 2024: Seizing the moment*, 65% of global sports leaders surveyed pointed to women's sports as the largest growth opportunity in the sector. Women's sports are rapidly evolving, garnering more attention, viewership, revenue, and investment than we had initially predicted.²⁶

And while future growth is expected, it is not a given.

Rising valuations influenced by structural setups

North American sports leagues, including the Women's National Basketball Association (WNBA) and National Women's Soccer League (NWSL), are setting a high bar for franchise valuations.

In 2024, the "Caitlin Clark Effect" swept through the WNBA, attracting new fans and sponsors to the league.²⁷ Following growth in viewership, app downloads, and engagement, the WNBA reportedly agreed to a US\$2.2 billion new media rights deal—more than triple the previous agreement.²⁸

As visibility increases and more fans engage with the league, some investors are capitalizing on the growth opportunity. In August 2024, the Dallas Wings lingered at the bottom of the league record but were valued at US\$208 million after two investors bought a 1% stake in the team at US\$2.08 million.²⁹ Prior to this deal, the Las Vegas Aces held the crown for most valuable franchise with a US\$140 million valuation.³⁰ The Aces were bought by Mark Davis in 2021 for US\$2 million, and after investing heavily into the franchise, including a US\$40 million investment into a team-specific practice facility, his team may now be worth more than 70 times what he paid to acquire it.³¹ Part of this supporting investment thesis is a belief in the continued growth of WNBA fandom, with tailwinds looking favorable. The Golden State Valkyries, an expansion team slated to start play in 2025, have already sold a record 17,000 season ticket deposits, proving that demand is expected to continue into the foreseeable future.³²

In September 2024, the WNBA announced an additional expansion franchise in Portland, Oregon, which will start play in the league in 2026.³³ The team will be owned and operated by Raj Sports, and led by Lisa Bhathal Merage and Alex Bhathal, who also own the Portland Thorns of the NWSL.³⁴ The group paid US\$125 million for the franchise, a steep increase from the US\$50 million paid for the Golden State and Toronto franchises entering play in 2026.³⁵

As visibility increases and more fans engage with the league, some investors are capitalizing on the growth opportunity.

In the NWSL, clubs are experiencing the benefits of a new broadcast deal as well. The league agreed to a new four-year media rights deal beginning in 2024 with broadcasters ESPN, CBS, Amazon Prime Sports, and Scripps' ION Network, worth a reported US\$60 million annually (US\$240 million total).³⁶ This deal represents an increase from the previous cycle, a three-year agreement with CBS worth US\$4.5 million total.³⁷

Fueling this growth may be the increased levels of sponsorship and sophisticated investment across the league. In 2024, multiple clubs changed hands, with valuations increasing throughout the year. In June, private equity firm Carlyle, in conjunction with the ownership group of Seattle Sounders FC, completed the purchase of NWSL club Seattle Reign for a reported US\$58 million.³⁸ The previous ownership, OL Groupe, paid approximately US\$3.5 million when it acquired the club in 2019.³⁹ In July, Willow Bay, dean at the University of Southern California, and her husband Bob Iger, CEO of Disney, were announced as the new controlling owners of Angel City FC following an investment deal that valued the club at US\$250 million, the highest valuation ever for a women's professional sports team.⁴⁰ Angel City FC reportedly paid an expansion fee of US\$2 million when it joined the NWSL in 2022.⁴¹

European women's soccer has grown over the past few years. Revenues across some of Europe's top clubs increased 61% in the 2022–23 season compared to the prior season.⁴² The opportunity to invest into standalone women's soccer entities is often rare across Europe, with the current combined structure limiting focused investment into the women's side as prospective investors would need to invest through a men's team.⁴³

However, some investors are challenging the status quo by investing in some of Europe's elite women's teams. In 2024, Michele Kang completed her acquisition of a 52.9% stake in eight-time UEFA Women's Champions League winners, Olympique Lyonnais Féminin.⁴⁴ In this groundbreaking transaction, Kang agreed to a 50-year licensing fee for use of the Olympique Lyonnais intellectual property (IP) and branding.⁴⁵ The acquisition adds to Kang's already growing multiclub ownership model across global soccer, having acquired the NWSL's Washington Spirit in 2022 and one of the few standalone English women's clubs, London City Lionesses, in 2023.⁴⁶

In the English Women's Super League (WSL), Chelsea Women also announced a new strategic growth plan that would see the team repositioned to sit alongside, rather than beneath, the men's side in the club's overall business structure.⁴⁷ This can allow for capital investment directly into the women's side, as opposed to investing through a men's team, and it's possible that other affiliated women's sides could follow suit to allow for direct investment channels.

Bottom line

Women's sports are expected to continue their growth trajectory into 2025. Last year, a handful of bold organizations and investors made moves across the market. More stakeholders may need to follow suit, to help capitalize on the opportunity and push the market past one-off examples of investment. Select women's sport properties have been traded at higher valuation multiples than is typically seen across the broader industry, but as revenue grows, these multiples are likely to decrease in future transactions. In the next year, more rights holders could evaluate their investment structures, as market appetite is expected to grow in line with accelerated revenues and fan engagement.

Fixed wireless access: Contrary to popular opinion, adoption may continue to grow

With US FWA net adds likely being slightly lower than last year, and some markets not expected to take off until 2026 ... there may be pockets of unrealized or potential growth out there, both in the US and globally

Duncan Stewart, Dan Littman, Jack Fritz, and Hugo Santos Pinto

Fixed wireless access—when consumers get their home broadband over a fixed cellular device (mainly 5G) rather than via wires—has been a big 5G growth story over the last few years in the United States, with well over 10 million homes expected to be connected by the end of 2024.⁴⁸ Consumers seem to have been attracted by competitive prices and speeds that are “good enough,”⁴⁹ and carriers find it more affordable to provide broadband in areas where population densities make fiber optic less attractive.⁵⁰

But with fixed wireless access (FWA), net additions (also referred to as net adds, or quarterly new subscribers less those who have canceled) in the United States during the first quarter of 2024 were less than in Q1 2023,⁵¹ and with the Indian FWA market (which is the other global market expected to drive a large number of net adds, with some projections of 100 million subscriptions by 2030) still nascent, many are expecting 2025 to be a year of lower growth for FWA.⁵²

Deloitte's 2022 TMT Predictions forecasted FWA growth at almost 20% annually to reach under 100 million subs globally (mostly 4G) in 2023.⁵³ Actual growth has been roughly in line with that, with more than 150 million total subscriptions worldwide expected by the end of 2024 (with about 30% coming from 5G).⁵⁴

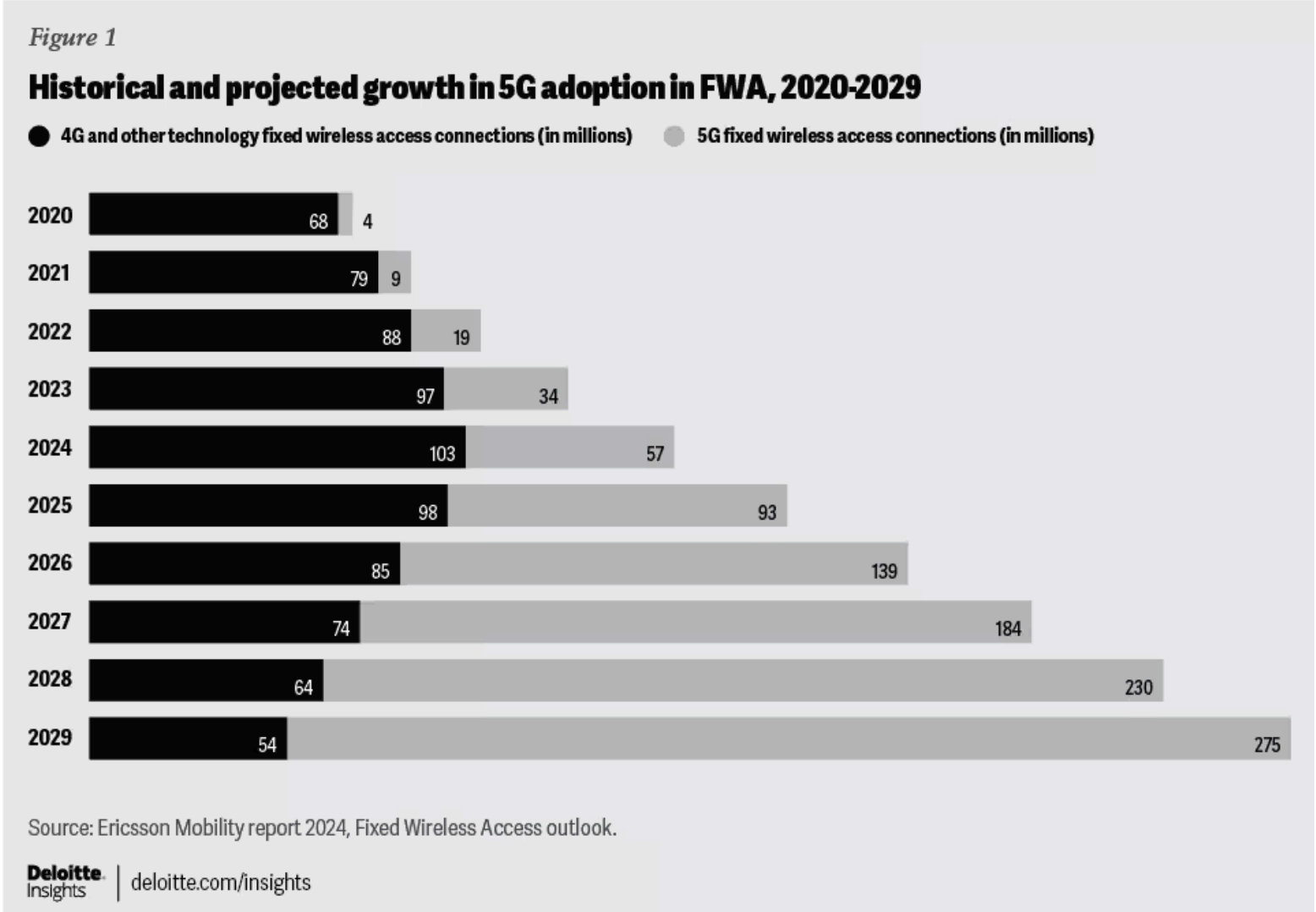
Deloitte predicts that global FWA net adds in 2025 and 2026 will continue to rise by about 20% each year. This expectation is driven by three trends identified:

1. **Some FWA growth is escaping the global headlines.** The United States and India are both large markets. FWA net additions in the United States are nearly one million per quarter.⁵⁵ As FWA numbers in India grow, quarterly additions may be even larger.⁵⁶ But there is FWA growth in other markets, too—for example, quarterly FWA net adds of 100,000 in Italy may make few headlines outside of that market,⁵⁷ but on a per household basis, Italy's FWA growth rate is only marginally lower than that of the United States.⁵⁸ Italy's FWA rate is high relative to most countries in the rest of Europe, Latin America, Southeast Asia, and Africa,⁵⁹ but even so, millions of additional net adds in 2025 and 2026 are expected outside of the United States and Indian markets.
2. **Enterprises are increasingly opting for FWA.** So far in the United States, most of the growth in fixed access wireless has come from consumers. Recently, there's been a shift to more enterprise (mainly small and medium businesses) FWA connections.⁶⁰ Some enterprises are finding FWA increasingly interesting, offering a single point of contact, consolidated billing, and more robust security.⁶¹ Deloitte predicts over a million enterprises will connect to FWA in each of 2025 and 2026 in the United States alone.⁶²
3. **New tech raises the ceiling on the number of customers that can use 5G in each market.** For the last few years, some believed that there was a natural ceiling on US FWA: Given the existing 5G technology and available spectrum (mainly at 2.5 GHz and 3.5 GHz), it might be difficult for there to be more than about 10 million 5G FWA subscriptions in the United States without affecting both the fixed and mobile wireless subscriber experience.⁶³ It was thought that, in the near term, in some areas, operators would be forced to stop offering FWA to additional subscribers.⁶⁴ However, even as that number of subscribers becomes closer, customer downstream speeds are improving.⁶⁵ A number of 5G technology upgrades, mainly around the advanced use of radio technologies, indicate that up to 20 million US homes (19% of about 106 million US broadband homes)⁶⁶ could be connected to FWA using current spectrum, suggesting the current US FWA run rate of about 3–4 million new subscribers per year might be able to continue for at least a few more years.

Bottom line

The continued growth of FWA could be a tailwind for telecom companies seeking to monetize 5G investment. Other than FWA, most 5G services with the potential to grow revenues are small as of 2025.⁶⁷ For context, the average price of FWA in the United States is about US\$50 per month and adding 4 million more FWA subs in 2025 translates to an additional US\$2.4 billion in incremental revenues, plus the benefit or reduction of mobile churn via bundling.⁶⁸

On the other hand, those who offer other kinds of broadband access (coaxial cable, copper DSL, fiber optic) are likely to continue to see competition from FWA, which could result in subscriber losses or difficulty in maintaining or increasing prices. FWA has been a competitor to this access in multiple markets, and those hoping FWA was nearing the end of its growth phase may have to wait a while yet.



The continued growth of FWA could be a tailwind for telecom companies seeking to monetize 5G investment.

5G standalone appears to be at a standstill: Will 6G run late?

Telecoms reassess investments in 5G standalone and delay 6G progress amid ROI concerns

Prashant Raman and Duncan Stewart

The deployment of 5G standalone networks is progressing more slowly than initially expected.⁶⁹ Telecom companies may be hesitant to invest heavily in this next generation of 5G in part due to underwhelming returns on their existing 5G investments. For now, the rollout of 6G seems further away than ever.

As of March 2024, 49 operators (out of 585 that have launched 5G globally) have deployed, launched, or soft-launched 5G standalone (SA) networks, or 8%.⁷⁰ This slow adoption rate suggests that carriers may be reluctant to spend on SA given uncertain returns on capital, at least in the near term. Additionally, the lack of monetization of 5G SA use cases may pose a challenge for adoption. Operators are putting the brakes on 5G SA deployment and may be hesitant to invest heavily in 5G SA infrastructure without clear revenue streams from its applications.⁷¹ As a result, Deloitte predicts that fewer than 20 additional networks are expected to be upgraded to standalone in 2025, keeping 5G SA at around 12% of all 5G deployments.⁷²

5G was launched in 2018, and during the initial launch, most operators opted for non-standalone (NSA) architecture, where the 5G network is built on top of an existing 4G infrastructure, to help expedite the rollout of their services.⁷³ This approach primarily stemmed from the unavailability of fully developed 5G SA equipment and the desire to leverage existing infrastructure.⁷⁴ Consequently, the NSA rollout allowed for a quicker and more cost-effective deployment of 5G services, enabling operators to help provide enhanced network speeds and connectivity, although without the full feature set of 5G SA.⁷⁵ In 2022, Deloitte Global predicted that the number of mobile network operators investing in 5G SA networks—with trials, planned deployments, or actual rollouts—would double from more than 100 operators in 2022 to at least 200 by the end of 2023, but that has not happened.⁷⁶

Delays in 5G SA deployment may impede the development and testing of these technologies, affecting progress toward 6G.

Current 4G and 5G NSA networks can efficiently manage the most heavily used current consumer and enterprise applications, which may be diminishing the urgency for operators to invest heavily in 5G SA.⁷⁷ On the enterprise side, despite the longer-term potential for revolutionizing different industries, the slower-than-expected pace of development and adoption of emerging 5G-based solutions that require SA has led to a more cautious approach to 5G SA deployment by operators.⁷⁸

5G SA infrastructure is essential for testing and validating the technologies that will be foundational for 6G.⁷⁹ It supports key use cases envisioned for 5.5G and 6G that require low latency, high reliability, and network slicing.⁸⁰ Delays in 5G SA deployment may impede the development and testing of these technologies, affecting progress toward 6G.

Telecom companies are struggling to achieve ROI from 5G due to high costs and slow monetization and consumer and enterprise adoption of 5G.⁸¹ Additionally, the anticipated revenue streams from additional 5G services such as 5G network application programming interfaces haven't materialized yet.⁸² With 6G use cases still under development, mobile network operators are likely focusing on maximizing 5G potential and recouping investments, which could push 6G development and deployment further down the timeline. Although various studies and standard setting are expected to occur before then, it has generally been accepted that the first commercial 6G rollouts at scale would occur around 2030.⁸³

Bottom line

As mobile network operators plan for 5G SA, it's important to consider how to continue optimizing the performance of their existing 5G NSA networks. A phased investment approach for 5G SA, targeting regions or sectors with higher demand, can help manage costs and focus resources more effectively. Additionally, exploring new revenue models, such as offering 5G SA as a service to specific industries or bundling it with cloud, AI, and edge computing, could open new opportunities. Collaborating with enterprise clients to help develop industry-specific solutions, particularly in sectors like manufacturing, health care, and logistics, can also help drive the adoption of 5G SA and support further investment in this technology.

Governments and regulators should be aware that the deployment of SA is an ongoing process and is taking longer than expected. This likely has implications for the move to 5.5G and 6G, and they may want to reconsider their approach to planning for next-generation network requirements and spectrum. Telecom equipment manufacturers were, to some extent, counting on deploying SA networks to help provide ongoing revenues in the anticipated spending trough between peak 5G investment in 2021 and the next wave of spending in the late 2020s or early 2030s.⁸⁴ If SA continues to occur slowly, that trough may become deeper or longer than expected, with potential implications for equipment makers' revenues and profitability.

Open RAN mobile networks and vendor choice: Single vendor now, multivendor when?

Open RAN's journey toward a diverse, multivendor ecosystem is marked by slow growth and complex challenges

Prashant Raman and Duncan Stewart

Open Radio Access Network (Open RAN) aims to democratize networks by providing mobile network operators (MNOs) who build RANs with greater choice and more flexibility, hopefully leading to better networks and lower prices. Despite high expectations and endorsements, the transition toward a diverse, multivendor ecosystem is proving slower and more complex than some initially anticipated.⁸⁵ Realizing true multivendor Open RAN may take a while, as Deloitte predicts that there will be no additional multivendor Open RAN networks deployed or even announced in 2025.⁸⁶

The Radio Access Network (RAN) is an important component of cellular networks that manage radio communications between mobile devices and the core network, and historically it was a closed and proprietary solution. The entirety of a network was from one vendor, and no other vendor's gear would work with it.⁸⁷ Open RAN emerged as a transformative concept in the telecommunications industry, aiming to standardize and democratize the design and implementation of network components.⁸⁸ It gained momentum in the late 2010s with the establishment of the O-RAN Alliance in 2018, which was formed to help promote more open and interoperable RAN architectures.⁸⁹

By 2028, single-vendor Open RAN solutions are expected to comprise 15% to 20% of total RAN revenues, while multivendor Open RAN solutions are expected to make up only 5% to 10%.

In 2021, Deloitte predicted that global active public network Open RAN deployments would double from 35 to 70.⁹⁰ The forecast was too optimistic: As of March 2024, the ongoing public network Open RAN deployments and trials stand at 45, with only two networks globally being multivendor Open RAN.⁹¹

A goal of Open RAN was to provide operators with a more open alternative, fostering competition and diversity in the market.⁹² However, the reality has not yet lived up to the aspirations. Many operators continue to purchase radios and baseband products for any given site from the same vendor, and the impact of pure-play Open RAN companies has been minimal so far; the five biggest RAN vendors currently account for around 95% of the market.⁹³

In 2024, Open RAN is projected to represent only 7% to 10% of the total RAN market, or about US\$2.5 billion to US\$3.5 billion, of a projected US\$35 billion.⁹⁴ Moreover, single-vendor solutions are anticipated to continue to be more widely used than Open RAN revenues in the near future. By 2028, single-vendor Open RAN solutions are expected to comprise 15% to 20% of total RAN revenues, while multivendor Open RAN solutions are expected to make up only 5% to 10% of the market, with traditional RAN still making up 80% to 85% of the market.⁹⁵ Although many in the industry still think the long-term potential of multivendor Open RAN remains strong,⁹⁶ achieving this vision could take more time.

Navigating the Open RAN landscape can involve integrating various hardware for high-capacity performance and cost efficiency. Operators historically preferred one-stop shopping from traditional RAN vendors who took care of everything and provided a single source of accountability.⁹⁷ Sourcing from multiple providers could pose challenges, especially for smaller operators.

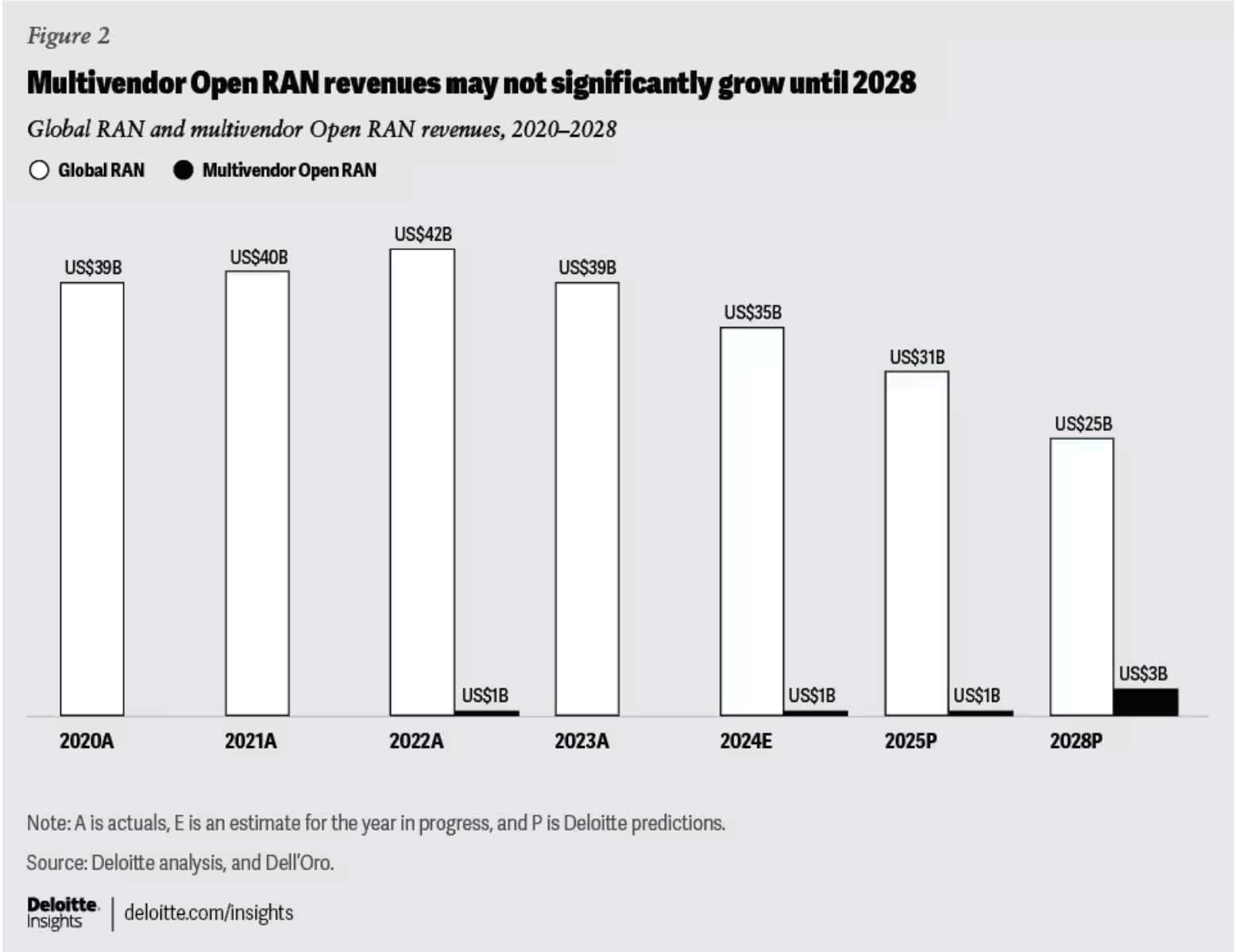
However, the Open RAN market is part of the broader US strategy to help enhance economic security and reduce reliance on foreign supply chains, particularly in critical sectors like telecommunications.⁹⁸ Open RAN technology may be pivotal in this effort, aimed at breaking the current market concentration dominated by a few major players, mostly non-American.⁹⁹ Geopolitical tensions have led to the exclusion of some of these companies from significant markets, including the United States and Europe, thus limiting vendor options and underscoring the need for supply chain diversification.¹⁰⁰ Open RAN helps address this by enhancing competition and innovation within the domestic market. Reshoring through Open RAN could be a strategic opportunity to help encourage domestic production of telecom equipment and maintain its leadership in the global telecommunications infrastructure market.¹⁰¹

Bottom line

Most MNOs already have Open RAN on their planning agenda. It’s been talked about for years now.¹⁰² Some have started the steps below; others are expected to start in the next year or two. MNOs can explore Open RAN in a low-risk and experimental way by gradually integrating its components into less critical areas of their networks. Collaborative testing with other operators and vendors or alliances can help address the technical challenges of deploying multivendor solutions. Working closely with emerging Open RAN vendors as well as existing RAN vendors could also provide an opportunity to help develop solutions that meet specific network needs. Additionally, investing in building internal capabilities and expertise around Open RAN technologies, whether through staff training or hiring specialists, could likely help navigate the complexities of a multivendor environment.

Existing large RAN vendors will likely want to balance being “open.” At one level, Open RAN could be considered a threat to their existing business, but if Open RAN is inevitable, they may want to consider disrupting themselves, before someone else does it. Original equipment manufacturers (OEMs) are already focusing on developing interoperable products that can seamlessly integrate with existing network infrastructures¹⁰³—and they will likely continue to partner with smaller players and startups in the Open RAN space. Staying engaged with industry alliances can help OEMs and emerging Open RAN vendors keep their products aligned with evolving standards and trends.

One of the benefits of multivendor Open RAN was increased vendor diversity, especially vendors outside Europe and Asia.¹⁰⁴ If Open RAN continues to proceed slowly for now, that geographic balance of vendors is unlikely to change materially in the near term, and other tools may be needed to help diversify the existing RAN supply chain.



Despite quantum’s slow start, don’t be slow to start your defense against it

Quantum drug discovery and financial modeling are likely several years away, but the time needed to upgrade cyber defenses for the quantum age likely necessitates prompt action

Colin Soutar, Duncan Stewart, Scott Buchholz, and Gillian Crossan

The 2019 and 2022 TMT Predictions reports discussed quantum computing¹⁰⁵ and mentioned the attendant cryptographic threats in the 2022 edition.¹⁰⁶ Some of those expectations have become actualities. Current quantum computers are not yet reliable enough for real use cases;¹⁰⁷ heightened attention is being placed in the cybersecurity domain;¹⁰⁸ and as predicted by Deloitte in December 2023, various aspects surrounding cyber standards came to fruition this year.¹⁰⁹ For example, the National Institute of Standards and Technology (NIST) recently issued post-quantum cryptography standards.¹¹⁰ (NIST is widely considered the gold standard for cybersecurity and cryptography, particularly in the development and adoption of encryption and digital signature algorithms, protocols, and frameworks that ensure secure data communication and transaction protection.) One cybersecurity risk relates to a specific algorithm—Shor’s algorithm, an algorithm developed in 1994 specifically to harness quantum effects to crack public key encryption schemes rapidly¹¹¹—being implemented on a quantum computer.

Deloitte predicts that the number of companies working on implementing post-quantum cryptography solutions is expected to quadruple in 2025 compared with 2023, and their spending also is expected to have quadrupled. This is a conservative estimate, based on projected costs between 2025 and 2035 to migrate federal systems¹¹² and ongoing efforts in the financial services industry to mitigate this risk.¹¹³

The timelines for both the positive quantum use cases and the defense against quantum cyberthreats should be considered. The question as it relates to timelines often has less to do with how and when a quantum computer will be used to implement positive use cases and for Shor’s algorithm, but more so: How long will it take for organizations to adopt and implement the recently issued NIST standards, to rely on cloud hyperscalers, specialized security components, and on “homegrown” applications that leverage cryptographic libraries for security features such as confidentiality and trust? To help answer this question, some organizations are starting their journey toward quantum cyber readiness by understanding what their exposure could be to this threat, by planning roadmaps for updates that meet their specific mission statements and business operations, and by working through the potential procurement and contractual requirements.

Contributing to the timeline uncertainty further are state or state-sponsored actors (and others) who are purportedly harvesting encrypted data now to decrypt it later when a cryptographically relevant quantum computer exists (also known as “harvest now, decrypt later” attacks).¹¹⁴ It should be interesting to see how organizations deal with this “sleeping threat” and whether they will proactively adopt the NIST standards to help prevent a future threat of data spills, especially in cases where such data is expected to have a long protection life cycle, such as personal information.

Interestingly, some large-scale providers are starting to introduce post-quantum cryptography into their platforms.¹¹⁵ It’s likely that other messaging services that currently offer end-to-end encryption will also roll out post-quantum cryptography in 2025 and beyond. Further, hyperscalers are offering services that can allow customers to benchmark, prototype, or understand the performance impact of quantum-resistant cryptography on cloud services.¹¹⁶

Bottom line

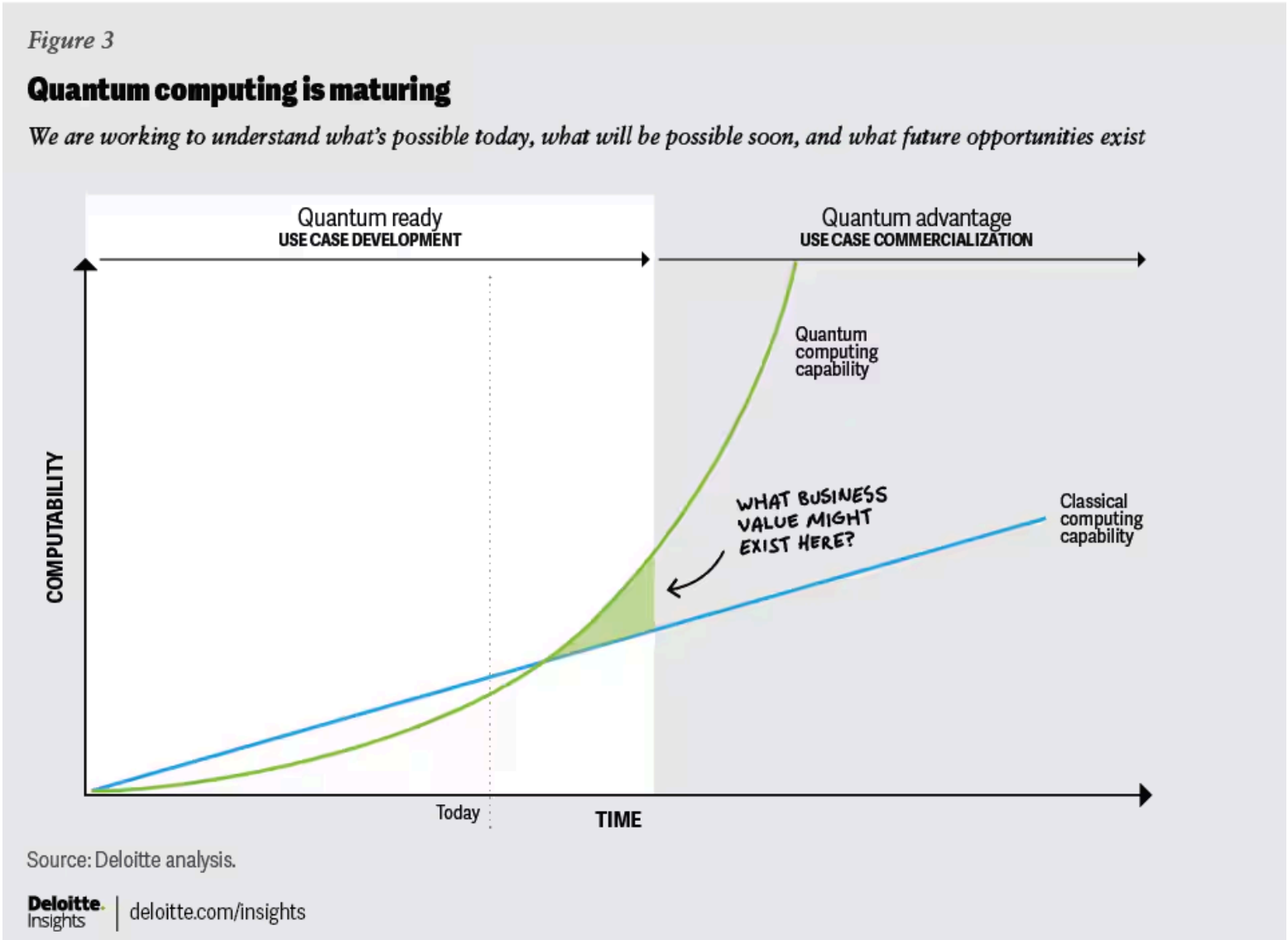
Quantum advantage, in which quantum computers have a large advantage for solving useful problems over classical computers, likely remains a few years off, barring any unexpected breakthroughs (figure 1). As the blue-shaded area in figure 1 gets closer, between use case development and use case commercialization, organizations should consider taking a risk-based approach to help address the threat to cryptography. They should understand how this risk compares with others they’re facing and proactively mitigate the risk accordingly.¹¹⁷

The organization that develops a cryptographically relevant quantum computer will be unlikely to advertise that it has one and can use it. So, there may not be a warning. This should make it more important to act sooner, lest organizations find themselves needing to react to an event at a time when resources may be scarce and systems may need to be shut down.¹¹⁸

Although working on post-quantum cryptography could be the necessary near-term implication of quantum computing in 2025, organizations can still work on their processes in preparation for the day when quantum computers offer quantum advantage and organizations can work on use cases that could bring positive benefit to many industries such as the financial sector and health care.

The organization that develops a cryptographically relevant quantum computer will be unlikely to advertise that it has one and can use it.

Quantum technologies can offer hope for solving long-standing challenges via more effective drug discovery and the predictive modeling of financial or environmental disturbances. The attendant cybersecurity threat shouldn’t overshadow the potential for such positive benefits—especially when cryptographic upgrades that can help protect against the threat can be implemented in a methodical, gradual, even systematic manner.



RISC-V: Closing the geopolitical gen AI loophole

An open-source alternative to proprietary chip design is gaining popularity among CPU designers. Its potential role in gen AI for markets under export restriction is a new twist for a new technology.

Duncan Stewart, Christie Simons, Jeroen Kusters, and Gillian Crossan

The 2022 TMT Predictions report included a chapter on RISC-V, which forecasted that open-source instruction set architecture (ISA), the set of rules and specifications that dictate how a computer’s central processing unit (CPU) operates, would see strong growth. The report predicted that the number of RISC-V chips would double year over year from 2022, and that revenues could “approach US\$1 billion by 2024.”¹¹⁹

The industry organization, RISC-V International, in late 2023 estimated that RISC-V-based system-on-chip (SoC) shipments were 1 billion in 2023 and could be close to 2 billion in 2024.¹²⁰ As predicted, revenues are still nascent, and RISC-V chip revenue is expected to be under US\$1 billion for the calendar year 2024.¹²¹ Given that global chip revenues for the year are predicted to be over US\$611 billion for 2024, RISC-V remains niche for now.¹²² Longer term, some forecasts are optimistic, with one calling for RISC-V to represent 25% of the SoC market by 2030, or more than US\$100 billion.¹²³

While RISC-V has the potential to capture a significant share of the SoC market, there are national security risks associated with this technology. US lawmakers serving on the bipartisan House Select Committee on the Strategic Competition Between the United States and the Chinese Communist Party have urged the Commerce Department to use its export control authorities to regulate or restrict the transfer of RISC-V technology to China. This approach aims to protect and promote collaboration on RISC-V technology between the United States and its allies as well as protect their mutual national security interests.¹²⁴

In order to learn why, we should revisit what ISAs are, where they are found, and how this connects with AI.

ISAs are the heart of CPUs. In many laptops and most data centers, the dominant CPU ISA is the x86 architecture. In most smartphones, the dominant ISAs are CPUs based on the RISC ISA, which is being used in some laptops and data center CPUs, although x86 has larger market share.

Generative AI generally requires special computer processing chips for both training and inferencing of large language models. Most readers may already be aware of the role that graphics processing units (GPUs) can play in performing gen AI computing as their robust computing power handles the intricate calculations required by these models.

Although GPUs do the computational “heavy lifting,” there is also usually a CPU (or two) needed to orchestrate data flows and perform other coordinating tasks, and the “attach rate” (the number of GPUs per CPU) seems to be changing, and CPUs are becoming relatively more important. For example, the first version of gen AI accelerating hardware from Nvidia paired 8 GPUs with 2 CPUs (a 4:1 ratio) with a choice of proprietary x86 ISAs.¹²⁵ And by early 2025, Nvidia is expected to offer a new kind of tray that consists of four next-generation GPUs and *two* CPUs (a 2:1 ratio).¹²⁶ Unlike the first generation, these CPUs were designed by Nvidia itself and are based on the Arm ISA.¹²⁷

The United States has imposed multiple export restrictions on IP, talent, chipmaking equipment, and certain chips. In early 2023, the United States restricted the export of the most-advanced GPUs at the time (typically, manufactured at process nodes of 10 nm and below and capable of certain levels of chip-to-chip data transfer rates),¹²⁸ and later that year, the United States restricted the export of additional advanced CPUs, both x86 ISA and Arm ISA.¹²⁹

China currently cannot import or manufacture GPUs or CPUs that are state-of-the-art advanced process nodes at 7 nm and below and required for gen AI training or inference.¹³⁰ Although RISC-V is a niche ISA today, it’s possible that over time, future generations of gen AI computing infrastructure could use RISC-V CPUs as controllers. China would still need to make state-of-the-art GPUs; restricting RISC-V exports to China could be considered an extension of current US restrictions. And if RISC-V is not restricted by the United States, it could be a loophole or workaround for other restrictions including the ones on CPUs and GPUs at advanced process nodes. Further, China is exploring RISC-V in other areas, for example, developing CPU controllers for gen AI processing—which could be something the United States restricts as well.

Bottom line

Gen AI hardware solutions use GPUs plus CPUs (or GPUs with CPU cores on the same die). Most CPUs are based on x86 ISA, and the rest are Arm ISA, but none seem to be using RISC-V as the main ISA, for now (some use RISC-V cores for other functions). As such, export restrictions in 2025 may be more focused on closing future loopholes.

RISC-V International moved from the United States to Switzerland in 2020,¹³¹ so it is still unclear exactly how US export restrictions might work or be enforced. However, the US government is reported to be reviewing potential risks and assessing what actions could both address potential concerns and not harm US companies that are part of the international groups working on the technology.¹³²

China has already made headway in RISC-V and appears well-positioned in this space.

Finally, while China’s use of RISC-V could be restricted, it is noteworthy that China has long been interested in and a leading proponent of RISC-V, with over half of premium-level RISC-V International members being from China (12 out of 22).¹³³ Further, as of 2022, over half of the 10 billion RISC-V cores made that year were from China.¹³⁴

China has already made headway in RISC-V and appears well-positioned in this space. At the same time, it will be interesting to see how it navigates any potential adverse impact that could stem from possible RISC-V curbs in the future.

By	Duncan Stewart Canada	Karthik Ramachandran India
	Prashant Raman India	Jennifer Haskel United Kingdom

Endnotes

1. Investing.com, “[Earnings call: NVIDIA posts record revenue, bullish on data center growth](#),” Aug. 29, 2024.
2. Nitin Mittal et al., “[Now decides next: Getting real about generative AI](#),” Deloitte’s State of Generative AI in the Enterprise Q2 report, April 2024.
3. Deloitte global analysis based on current market share of gen AI chip sales, combined with our survey data, and a US\$200 billion to US\$300 billion gen AI server market.
4. Jack Fritz et al., “[Battle for the enterprise edge: Providers prepare to pounce on the emerging enterprise edge computing market](#),” Deloitte 2023 Global TMT Predictions, November 30, 2022; Chris Arkenberg et al., “[Gaining an intelligent edge: Edge computing and intelligence could propel tech and telecom growth](#),” Deloitte 2021 Global TMT Predictions, December 7, 2020.
5. Sending to and from the cloud took too much time for real-time visual inspection of soft drink bottles on a high-speed line, for example, requiring AI decision-making on-prem, or at least within a few kilometers: milliseconds matter.
6. Dr. Lance B. Eliot, “[Speeding up the response time of your prompts can be accomplished via these clever prompt engineering techniques](#),” *Forbes*, June 12, 2024.
7. Rowan et al., “[Now decides next: Moving from potential to performance](#).”
8. See our related 2025 Global TMT Predictions chapter on “Gen AI, data centers, and the electricity grid.” Further, see: Datacenter BMO report, Communications Infrastructure, “1Q24 data center leasing: Records are made to be broken,” April 28, 2024; CBRE, [North America data center trends H2 2023](#), March 6, 2024.
9. Fifty-five percent of organizations reported avoiding certain generative AI use cases because of data-related issues. Top data-related concerns include using sensitive data in models and managing data privacy and security. To read further, see: Deloitte’s State of Generative AI in the Enterprise Q3 report, August 2024.
10. Organizations were much more worried about using sensitive data (for example, customer or client data) including IP. Using sensitive data in models (58% had at least a high level of concern), data privacy issues (58%), and data security issues (57%) were the top three issues that execs noted in the survey. To read further, see: Deloitte’s State of Generative AI in the Enterprise Q3 report, August 2024.
11. Andy Lees et al., “[Financial services: Scaling gen AI for maximum impact](#),” Deloitte, accessed October 2024; Bijit Ghosh, “[The rise of small language models—efficient & customizable](#),” *Medium*, November 26, 2023.
12. Jana Arbanas et al., [2024 media and entertainment outlook: Generative AI](#), Deloitte, 2024; Ghosh, “[The rise of small language models—efficient & customizable](#).”

13. See our related 2025 Global TMT Prediction on the topic of “Gen AI in consumer devices.”
14. Deloitte analysis of publicly available specifications for various gen AI small servers, as well as conversations with companies globally.
15. Deloitte analysis of publicly available specifications for proposed rack-scale gen AI servers likely to be sold in the first half of 2025, although some may be available in limited quantities in late 2024.
16. Jennifer L, “*The carbon countdown: AI and its 10 billion rise in power use*,” Carboncredits.com, February 28, 2024.
17. Diana Goovaerts, “*Could AI drive a new cloud repatriation wave?*,” *Fierce Network*, June 6, 2024.
18. Ramesh Raskar, *Split Learning and Inference*, MIT Media Lab’s Split Learning Project, accessed September 24, 2024.
19. Duncan Stewart et al., “*Telecoms tackle the generative AI data center market*,” Deloitte Insights, September 2024.
20. See the related 2025 TMT Prediction chapter on “Gen AI, data centers and the electricity grid.”
21. Jacqueline Davis, “*Large data centers are more efficient, analysis confirms*,” Uptime Institute, February 7, 2024.
22. Duke Robertson, “*What Is a hyperscale data center? Overview and comparisons*,” *Enconnex Blog*, March 27, 2023.
23. Deloitte, “*Breaking the billion-dollar barrier: Women’s elite sports to generate more than \$1 billion in revenue in 2024*,” press release, December 1, 2023.
24. Priya Oberoi, “*Investors have their eyes on women’s sports as profitably soars*,” *Forbes*, June 21, 2024.
25. Elite sports is defined as the highest level of competition, which may or not be classified as “professional” sport where participants are paid for their performance.
26. Deloitte, “*Breaking the billion-dollar barrier: Women’s elite sports to generate more than \$1 billion in revenue in 2024*.”
27. Brandon Drenon, “*The Caitlin Clark Effect has made women’s basketball the hottest ticket around*,” BBC News, April 5, 2024.
28. Doug Feinberg, “*WNBA announces landmark 11-year media rights deal with Disney, Amazon Prime and NBC*,” AP News, July 24, 2024.
29. Alex Schiffer, “*\$2M cash injection sends WNBA’s worst team to league-leading valuation*,” *Front Office Sports*, August 12, 2024.

30. Josh Sims, “[Las Vegas Aces valued at US\\$140m as average WNBA team hits US\\$96m](#),” *SportsPro*, June 18, 2024.
31. Abhimanyu Chaudhary, “[Tom Brady’s WNBA investment sees dizzying 6,900% appreciation as Las Vegas Aces’ valuation touches \\$140,000,000](#),” *Sportskeeda*, June 18, 2024.
32. Dee Lab, “[WNBA expansion team Golden State Valkyries breaks season-ticket record](#),” *Just Women’s Sports*, September 17, 2024.
33. Doug Feinberg, “[WNBA awards Portland an expansion franchise that will begin play in 2026](#),” *AP News*, September 18, 2024.
34. Ibid.
35. Shawn Medow, “[Tanenbaum’s KSV to shell out \\$50m for WNBA expansion franchise in Toronto](#),” *SportBusiness*, May 13, 2024.
36. Cesar Hernandez, “[NWSL announces new 4-year rights deal with ESPN, CBS, Prime and Scripps](#),” *ESPN*, November 9, 2023.
37. Ibid.
38. Meg Linehan, “[Seattle Sounders and Carlyle ownership group completes purchase of Seattle Reign in NWSL](#),” *The New York Times*, June 17, 2024.
39. Ibid.
40. Angel City, “[Willow Bay and Bog Iger to become Angel City’s new controlling owners](#),” press release, July 17, 2024.
41. Kurt Badenhausen, “[NWSL team value 2024: Angel City, KC lead, average up 57% to \\$104M](#),” *Sportico*, September 25, 2024.
42. Timothy Bridge et al., “[Deloitte Football Money League 2024](#),” *Deloitte UK*, January 12, 2023.
43. Samuel Agini, “[English women’s football clubs hope new league will kick start investment](#),” *Financial Times*, January 26, 2024.
44. Charlotte Harpur, “[Lyon women’s team bought by Washington Spirit owner Michele Kang](#),” *The Athletic*, February 9, 2024.
45. Meg Linehan, “[Spirit owner Michele Kang joins OL Groupe to create new global women’s football organization](#),” *The New York Times*, May 16, 2023.
46. ESPN, “[Washington Spirit owner Kang buys London City Lionesses](#),” December 15, 2023.

47. Chelsea Football Club, “[Chelsea Women announces strategic growth plan](#),” May 29, 2024.
48. Deloitte extrapolation based on 7.8 million as of Q1 2024, and 2–3 million additional subscribers for the balance of the year. Masha Abarinova, “[Fixed wireless continues to climb US broadband charts—Parks](#),” *Fierce Network*, June 13, 2024.
49. Paul Lee, Dieter Trimmel, and Eytan Hallside, “[No bump to bitrates for digital apps in the near term: Is a period of enough fixed broadband connectivity approaching?](#),” *Deloitte Insights*, November 29, 2023.
50. Zply Fiber, “[Fiber internet in rural areas: When will it be here?](#),” January 18, 2024.
51. Abarinova, “[Fixed wireless continues to climb US broadband charts—Parks](#).”
52. Gagandeep Kaur, “[Why 5G is failing to gain momentum in India](#),” *Light Reading*, April 25, 2024.
53. Naima Hoque Essing et al., “[Fixed wireless access: Gaining ground on wired broadband](#),” *Deloitte Insights*, December 1, 2021.
54. Ericsson, [Ericsson mobility report](#), June 2024.
55. Nick Ludlum, “[5G home broadband continues to bring real competition to cable](#),” *CTIA Blog*, January 31, 2024.
56. Jolanta Stanke, “[Global broadband subscriber growth in Q4 2023 slowest since 2019](#),” *Point Topic*, April 22, 2024.
57. Andrey Popov, “[5G FWA is a game-changer for broadband services in Italy](#),” *Opensignal*, May 16, 2024.
58. Deloitte calculation based on 2024 population and household estimates.
59. Deloitte analysis of publicly reported information from wireless network operators.
60. Jericho Casper, “[Fixed wireless subscriber growth solid in Q2](#),” *Broadband Breakfast*, August 5, 2024.
61. Bob Wallace, “[Exploring telco enterprise fixed wireless access \(FWA\) services](#),” *Network Computing*, August 1, 2024.
62. Based on 2024 publicly reported business FWA additions from the major wireless providers, combined.
63. Robert Wyrzykowski, “[5G fixed wireless access \(FWA\) success in the US: A roadmap for broadband success elsewhere?](#),” *Opensignal*, June 6, 2024.
64. Ibid.
65. Ibid.

66. Abarinova, “[Fixed wireless continues to climb US broadband charts—Parks.](#)”
67. Ericsson, “[FWA is the largest 5G use case after mobile broadband,](#)” February 2023.
68. Ericsson, [Ericsson mobility report.](#)
69. It was initially anticipated that at least 200 operators would have launched standalone (SA) 5G networks by the end of 2023. However, as of March 2024, only 49 operators have successfully deployed 5G SA networks.
70. GSA, “[GSA Market Snapshot March-2024,](#)” March 2024.
71. James Kirby, “[5G standalone networks: How vendors can accelerate adoption,](#)” Analysys Mason, September 2023.
72. Based on lead author conversations with multiple operators globally between January and October 2024.
73. Deanna Darah, “[5G NSA vs. SA: How do the deployment modes differ?,](#)” TechTarget, July 25, 2024.
74. Ibid; GSMA, [The 5G guide: A reference for operators,](#) April 2019.
75. Darah, “[5G NSA vs. SA: How do the deployment modes differ?,](#)”
76. Naima Hoque Essing et al., “[5G’s promised land finally arrives: 5G standalone networks can transform enterprise connectivity,](#)” Deloitte Insights, November 30, 2022.
77. Darah, “[5G NSA vs. SA: How do the deployment modes differ?,](#)”
78. Ibid.
79. Philippe Poggianti and Pratik Das, “[It’s time for 5G to standalone,](#)” Qualcomm, July 6, 2023; Gavin Horn, “[6G foundry: Make the migration from 5G to 6G a rewarding experience,](#)” Qualcomm, May 14, 2024; Roger Billings, “[What is 5G standalone? 5G SA means network slicing, security, and automation,](#)” Ericsson Enterprise Wireless Blog, June 27, 2023.
80. GSMA, [The state of 5G 2024,](#) February 2024; Sylwia Kechiche, “[Will 5G Advanced deliver on the 5G promise?,](#)” Opensignal, April 4, 2024.
81. Andrew Wooden, “[The telecoms industry’s biggest problem? Failure to monetise 5G,](#)” Telecoms.com, March 14, 2024.
82. Mike Dano, “[A deeper dive into the 5G network API opportunity,](#)” Light Reading, April 1, 2024.
83. Global 6G Conference, “[Three 3GPP chairs clarify 6G standard release timeline at Global 6G Conference,](#)” April 23, 2024.

84. *Communications Today*, “[5G investment cycle tapering off and another one not on the horizon](#),” March 30, 2024; Mary Lennighan, “[5G growth comes with new operator spending requirements—GSMA](#),” *Telecoms.com*, February 29, 2024.
85. Caroline Gabriel, “[OpenRAN: Increased collaboration and a realistic view of timing will be needed to deliver the full benefits](#),” Analysys Mason, April 2023.
86. Based on the authors’ review of public announcements and multiple conversations with network operators globally.
87. Naima Hoque Essing et al., “[The next-generation radio access network: Open and virtualized RANs are the future of mobile networks](#),” *Deloitte Insights*, December 7, 2020.
88. Zineb Gdali, “[The future of mobile networks: Exploring the benefits of Open RAN 5G](#),” Firecell, December 15, 2023.
89. Parallel Wireless, [Everything you need to know about Open RAN](#), 2020.
90. Essing et al., “[The next-generation radio access network: Open and virtualized RANs are the future of mobile networks](#).”
91. TeckNexus, “[Current state of Open RAN—countries & operators deploying & trialing Open RAN](#),” March 10, 2024.
92. O-RAN Alliance, [About](#) page, accessed October 2024.
93. Dan Jones, “[Dell’Oro: Don’t expect 5G-Advanced to fuel a RAN resurgence](#),” *Fierce Network*, July 26, 2024.
94. Ray Le Maistre, “[Single vendor solutions to dominate Open RAN sales—Dell’Oro](#),” *TelecomTV*, February 7, 2024.
95. Ibid.
96. Ibid.
97. Iain Morris, “[Telcos doubt open RAN challengers will have a role](#),” *Light Reading*, March 15, 2024.
98. Dan Oliver, “[Open RAN: Everything you need to know](#),” *5Gradar*, July 23, 2021; Mike Dano, “[How American 5G operators learned to love open RAN](#),” *Light Reading*, February 12, 2024.
99. Wes Davis, “[The US government makes a \\$42 million bet on open cell networks](#),” *The Verge*, February 12, 2024.
100. Ibid.

101. Oliver, “[Open RAN: Everything you need to know.](#)”
102. David Debrecht, “[Building a smarter 5G future through Open RAN development](#),” CableLabs, November 29, 2023.
103. Ibid.
104. Hosuk Lee-Makiyama and Florian Forsthuber, “[Open RAN: The technology, its politics and Europe’s response](#),” European Centre for International Political Economy (ECIPE), October 2020.
105. Scott Buchholz et al., “[Quantum computing in 2022: Newsful, but how useful?](#),” *Deloitte Insights*, December 1, 2021; Duncan Stewart, “[Quantum computers: The next supercomputers, but not the next laptops](#),” TMT Predictions 2019, *Deloitte Insights*, 2018, pp. 96–103.
106. Deborah Golden et al., “[Preparing the trusted internet for the age of quantum computing](#),” *Deloitte Insights*, August 6, 2021.
107. Alex Wilkins, “[Useful quantum computers are edging closer with recent milestones](#),” *New Scientist*, September 30, 2024.
108. Filipe Beato et al., [Transitioning to a quantum-secure economy](#), World Economic Forum, September, 2022.
109. Nancy Liu, “[Deloitte predicts 2024 will be a breakthrough year for post-quantum cryptography](#),” SDxCentral, December 14, 2023.
110. National Institute of Standards and Technology (NIST), “[NIST releases first 3 finalized post-quantum encryption standards](#),” August 13, 2024.
111. P.W. Shor, “[Algorithms for quantum computation: Discrete logarithms and factoring](#),” *Proceedings 35th Annual Symposium on Foundations of Computer Science* (1994): pp. 135–42.
112. Executive Office of the United States President, [Report on post-quantum cryptography](#), July 2024.
113. FS-ISAC PQC Working Group, [Building cryptographic agility in the financial sector](#), October 2024.
114. John Potter, “[Deloitte: Companies face harvest now, decrypt later quantum threat](#),” *IoT World Today*, September 27, 2022.
115. Apple Security Engineering and Architecture (SEAR), “[iMessage with PQ3: The new state of the art in quantum-secure messaging at scale](#),” *Apple Security Research Blog*, February 21, 2024.
116. AWS, “[Post-quantum cryptography](#),” accessed September 24, 2024.
117. Katherine Noyes, “[NIST’s postquantum cryptography standards: ‘This is the start of the race’](#),” *Deloitte’s CIO Journal for The Wall Street Journal*, June 11, 2024.

118. Ibid.
119. Duncan Stewart et al., “*RISC-y business: Could open chip standard RISC-V gain traction against dominant incumbents?*,” *Deloitte Insights*, December 1, 2021.
120. Ishika Setia, “*RISC-V: Projected growth to over 16 billion chips by 2030*,” *TechnoSports*, November 9, 2023.
121. As an example, large player SiFive had only US\$38 million in revenues in 2023, although it forecasts 2024 revenues of over US\$241 million. Mavis Tsai and Judy Lin, “*SiFive projects 6x growth in 2024 revenues, optimistic over RISC-V and AI demand*,” *DigiTimes Asia*, March 18, 2024.
122. WSTS, *WSTS semiconductor market forecast spring 2024*, June 2024.
123. Belle Lin, “*Open-source chip design takes hold in Silicon Valley*,” *The Wall Street Journal*, December 14, 2023.
124. Stephen Nellis and Max A. Cherney, “*RISC-V technology emerges as battleground in US-China tech war*,” Reuters, October 7, 2023; Select Committee on the Chinese Communist Party (CCP), “*Gallagher, Rubio, Select Committee members urge Sec. Raimondo to safeguard American chip design from Chinese Communist Party threat*,” press release, November 2, 2023; Select Committee on the CCP, “*Reset, prevent, build: A strategy to win America’s economic competition with the Chinese Communist Party*,” December 12, 2023.
125. Nvidia, “*Nvidia DGX H100*,” accessed September 24, 2024
126. Nvidia, “*Nvidia GB200 NVL72*,” accessed September 24, 2024.
127. Nvidia, “*Nvidia Grace CPU*,” accessed September 24, 2024.
128. Christie Simons et al., *2024 global semiconductor industry outlook*, Deloitte, 2024.
129. Bureau of Industry and Security, “*Public information on export controls imposed on advanced computing and semiconductor manufacturing items to the People’s Republic of China (PRC) in 2022 and 2023*,” US Department of Commerce, November 6, 2023.
130. Alex He, “*In the global AI chips race, China is playing catch-up*,” Centre for International Governance Innovation, September 18, 2024.
131. Jacob Feldgoise, “*RISC-V: What it is and why it matters*,” Centre for Security and Emerging Technology (CSET), January 22, 2024.
132. Stephen Nellis, “*US is reviewing risks of China’s use of RISC-V chip technology*,” Reuters, April 23, 2024.
133. Sunny Cheung, “*Examining China’s grand strategy for RISC-V*,” *The Jamestown Foundation China Brief* 23, no. 23 (December 15, 2023): pp. 15–21.

134. Ma Si, “*Open-source chip platform to cut reliance on proprietary tech*,” *China Daily*, August 25, 2023.

Acknowledgements

Authors would like to thank **Hugo Pinto, Dan Littmann, Jack Fritz, Paul Lee, Itan Barmes, Casper Stap, Ben Shapiro, Adnan Amjad, Mark Nace, Emily Mossburg, Deborah Golden, Joe Mariani, Adam Routh, Ankit Dhameja, Zoe Burton,** and **Lizzie Tantam.**

Cover image by: **Jaime Austin;** Getty Images, Adobe Stock

Rising trends: The new and next technologies worth putting on your radar

TMT Predictions explores AI's impact on cyber defense, chiplets for higher semiconductor yields, telcos modernizing with cloud-based systems, and silicon photonics for AI data centers

ARTICLE • 20 MINUTE READ

New for 2025 is our series of shorter articles on emerging technologies. These trends tend to be earlier in adoption and smaller in revenues than our traditional Predictions topics, but we're betting they will grow faster than average and make it to the big leagues in the next year or two:

- Generative AI and cyber: Big risks, but big opportunities too
- Silicon building blocks: Chiplets could move Moore's Law forward
- B/OSS: Telcos modernize their business and operational support systems software
- Silicon photonics: Gen AI communicates at lightspeed

Generative AI and cyber: Big risks, but big opportunities too

Recognizing generative AI's potential for enabling both threats and cyber solutions, cybersecurity professionals are exploring ways to harness its power to counter emerging risks and help fortify the technology environment

Arun Perinkolam, Sabthagiri Saravanan Chandramohan, Alison Hu, and Duncan Stewart

AI, including gen AI, is an increasingly important part of growing cyberthreats: Seventy-one percent of US state chief information security officers characterized AI threat levels as “very high” or “somewhat high” in a 2024 survey.¹ AI is creating a mix of challenges including regulatory changes, undermining the effectiveness of current solutions, and AI-armed adversaries—complicated by the pace of enterprise AI adoption.

Deloitte expects that gen AI-based cyberattacks, already occurring more frequently in 2024 (doubling or even tripling), will continue to grow in 2025.² There are multiple ways in which gen AI can be used in a cyberattack, but one example would be in writing malicious phishing emails: These attacks were up more than 856% as of Q1 2024 compared with the same period in 2023.³ Threat actors are already using gen AI tools to write code for malware attacks.⁴

But gen AI tools can also be a force for good, defending against or ameliorating the new generation of AI-backed cyberthreats.

Some inside the cyber industry fear that gen AI, in addition to offering various benefits, can increase cyber risk, as it is a new attack vector that increases the attack surface.⁵ There are many ways in which gen AI can be used in cyberattacks. It can be used to generate sophisticated and high volume, text-based phishing attacks, as well as deepfake images and videos used to impersonate CEOs or other C-level executives: Sixty-one percent of organizations surveyed have experienced a deepfake attack in the last year, with 75% of those being executive impersonations.⁶ In response, several gen AI solution providers are putting up guardrails in an effort to prevent their tools from being used for generating these text and video attacks, as well as embedding digital watermarks so that gen AI images or text can be detected, flagged, or blocked (see the [2025 TMT Prediction about watermarks and AI detection](#)).

Next, many industries, led by the tech industry, have been increasingly using gen AI coding tools, which can help coders to write more code faster.⁷ However, the code the gen AI tool creates could have security issues, which more than half (56%) of developers surveyed in late 2023 said happens sometimes or even frequently.⁸ Further, the same survey showed that some developers often bypassed company policies about using coding tools, were routinely overconfident in the security of the code being generated, and were not always scanning the generated code for security issues.⁹ While there may be security concerns, gen AI coding and security large language models (LLMs) are helping to accelerate the maturity, efficiency, and efficacy of security processes, such as auto generation of monitoring rules in security information and event management (SIEM) technologies, use cases in the identity and access management arena around access workflows, provisioning, and third-party risk management.¹⁰

Currently, there are regulatory and geopolitical developments relevant to the intersection of gen AI and cyber. As an example, Article 15 of the EU AI Act explicitly addresses cybersecurity issues for high-risk AI systems.¹¹ Further, there have been export restrictions on various technologies that have been implemented for AI, especially gen AI, since 2022, such as advanced node semiconductors necessary for training and inference of gen AI models, the equipment used to make those advanced chips, and the design tools for those chips.¹²

Bottom line

As gen AI becomes more integrated with businesses in general, companies providing AI solutions should continue to focus on making secure products for end users. It isn't just the products themselves that need to be secure, but companies should be careful when sharing their own or others' customer data with fourth parties, who are usually the providers of LLM services.

There is increased regulatory complexity due, in part, to the EU's Digital Markets Act,¹³ Digital Services Act,¹⁴ and the new AI Act.¹⁵ Tech companies are not only the lead developers, but also are among the biggest deployers of broad use AI models. Therefore, there is likely a higher level of expectations that tech companies should play a larger and a more meaningful role in working to ensure trust and safety of gen AI being implemented in the products and solutions that they develop and sell to enterprises. The use and abuse of gen AI technology by threat actors, particularly in times of heightened risk, divisive geopolitical matters, elections, wars, etc., is likely to become an increasingly critical defense and strategic consideration in 2025 and beyond (see the [2025 TMT Prediction about trust in AI and the tools that companies are often using](#)).

Silicon building blocks: Chiplets could move Moore’s Law forward

Chiplets promise to deliver more flexible, scalable, and efficient systems for AI and high-performance computing environments, at higher yields

Karthik Ramachandran, Duncan Stewart, Christie Simons, and Dan Hamling

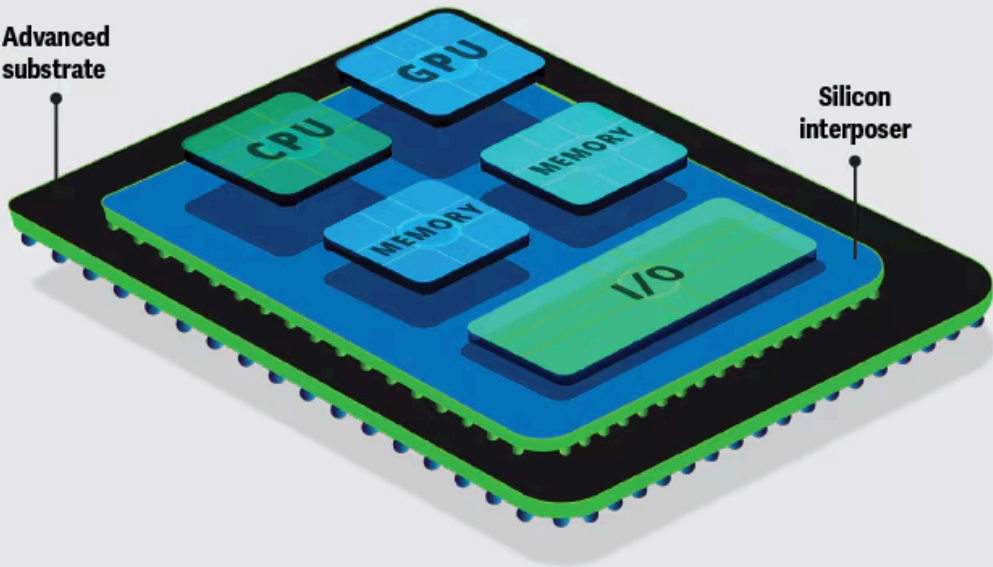
Deloitte predicts that worldwide advanced packaging revenue based on “chiplets,” the building blocks of today’s most advanced systems in a package (SiPs), will more than double from an estimated US \$7 billion in 2021 to reach US \$16 billion in 2025.¹⁶ Compared to more traditional architectures that rely on separate interconnected chips on a printed circuit board (PCB), chiplets offer high-speed data transfers, reduce latency, and optimize PPA (power, performance, and area), and even extend Moore’s Law.¹⁷ Chiplets are often being used and explored in some of the fastest-growing markets such as AI accelerators (especially generative AI), high-performance computing (HPC), and telecommunications applications.

What are chiplets?

How is a chiplet different than chips? Typically, chipmakers use a 300 mm silicon wafer (about 70,000 mm²) and make a number of monolithic dies, which we call chips, once they are packaged. A high-end advanced chip generally measures 20 mm by 20 mm (or 400 mm²), resulting in approximately 175 dies per 300 mm wafer. But a chiplet is not monolithic: It is a heterogeneous architecture, where smaller dies are packaged in such a way that they work together akin to a monolithic die. Moreover, those dies and modules can come from various chip manufacturers.¹⁸

Figure 1

Chiplet is a small, modular integrated circuit that can be combined with other chiplets to form a larger, complete system



Source: Deloitte analysis.

Deloitte Insights | deloitte.com/insights

Why are chiplets important now? Chiplets have been around since the 1980s. But there has been a renewed interest and a large-scale shift toward chiplets in the past four to five years, largely because of the need for improved yield at leading-edge manufacturing nodes.¹⁹ Making advanced chips is becoming more challenging, as the industry is getting closer to the physical limits of Moore’s Law. But chiplets are making it possible to advance the miniaturization of semiconductors, and SiP-based chips are delivering performance that is comparable to the traditional system-on-chips designed using monolithic dies.²⁰

The smaller and more complex chips get (e.g., advanced processing nodes at 5 nm and 3 nm), the more likely they are to have a defect rate across the 300 mm wafer that would affect yield.²¹ An 800 mm² die, which is used to produce the most-advanced AI chips at 3 nm or 5 nm will likely only have 50% to 55% yield when assembled and packaged using a traditional monolithic approach, with a defect density of 0.1 defects per cm².²² For context, normal yields for a mature semi process node (at 90 nm and 130 nm) would be on the order of 90% to 95%.²³ To address this problem, chiplets combine several smaller, higher-yield chips into a larger system that functions as one: Smaller die of 180 mm² size each (with a yield of 95% each) packaged using chiplet-based architectures can create more efficient and powerful AI processors at a lower cost and enhance product/functional flexibility and configurability to meet dynamic market needs.²⁴

As chiplet adoption grows, industry players are finding creative ways to improve design processes, increase connection speeds and bandwidth, and improve energy efficiency. For example, the industry is looking at digital twins to emulate and visualize complex design processes step-by-step, including the ability to move around or swap chiplets to measure and assess performance of a multi-chiplet system.²⁵ Some companies have introduced a range of interconnection techniques to assemble and stack discrete components on a chip, improving efficiency over traditional, large-sized monolithic designs.²⁶ There is also ongoing R&D to explore using glass, which is proving to be a more flexible and scalable organic substrate, as well as superior in thermal conductivity and performance-per-watt as a substrate in chiplets packaging, targeted at HPC and AI environments.²⁷ Even photonics, using light for data transfer, is being explored as an interconnect solution to provide optical input/output (I/O), especially for HPC and AI workloads. This technology could deliver energy-efficient and high-speed data transfer and processing (see the rising trend about [silicon photonics](#), below).²⁸

At the same time, chiplet architectures continue to deal with unique challenges. For example, stacking multiple dies that are connected by thin substrates can create thermal management problems, leading to potential circuit malfunctions and power loss.²⁹ Additionally, as more intellectual property gets integrated into these complex packages, sourcing components from various vendors across different regions could increase the risk of cyberattacks and expose underlying systems to new security threats.³⁰

Bottom line

To help unlock business value from chiplets, participants across the semiconductor value chain should consider working together to close the gaps and address challenges, while exploring avenues for growth.

Equipment makers, foundries, integrated device manufacturers, fabless companies, and outsourced semiconductor assembly and test providers could further bolster their fab-to-packaging partnerships and co-development efforts. Power and logic integrated circuit manufacturers and designers can consider nuances related to thermal and heat management.³¹

Companies should also consider pivoting toward establishing standards for chiplet interconnects and data interoperability—building on and advancing their early efforts such as the Universal Chiplet Interconnect Express standard, the High Bandwidth Memory Protocol, and the Bunch of Wires interconnect technology.³²

Electronic design automation (EDA) companies, chip designers, and security experts can devise ways to develop in-built functionality that could sense potential IP theft and cyber infringement at the chiplet level and work with the rest of the supply chain to help address the broader threat and attack parameters that could affect chiplets. Additionally, designers should work with EDA and other computer-aided design and computer-aided engineering companies to strengthen the design, simulation, and verification and validation tools and capabilities for hybrid and complex heterogeneous systems—including applying AI techniques to chip design.³³

B/OSS: Telcos modernize their business and operational support systems software

Telcos' back-end business and operations software market is growing slowly but modernizing it—by adopting SaaS and microservices architecture, moving to the cloud and more—is a hot spot of growth for software vendors and an opportunity for telcos to do more with 5G, fiber, and AI

Amit Kumar Singh, Duncan Stewart, Hugo Santos Pinto, and Dan Littman

Historically, telcos had two separate but important suites of telecom-specific IT systems. Business support software (BSS) mainly consisted of capturing customer orders, customer relationship management, and billing. Operational support software (OSS) handled service order management, network inventory management, and network operations.³⁴ These were usually two distinct systems, often custom, mainly on-premises, mainly hardware defined, and composed of a series of individual, specialized solutions targeting specific service lines (fixed and mobile) or technological areas such as access, core, and transmission, creating a fragmented and complex infrastructure.³⁵ In 2025 and beyond, telcos are expected to modernize and infuse next-level automation and intelligence into these systems, which could lead to an acceleration of growth. Longer term, there may even be an integration of BSS and OSS into a single platform.

Why is this happening now? The advancement of on-demand access to services and products can require businesses to reimagine their customer experience, redefine offerings, reinvent business models, and reprogram sales channels. Within the BSS domain, specifically within billing, evolving customer expectations and new digital revenue streams may require new capabilities and support for offer-centric and customer-centric billing. This impact may be felt throughout the B/OSS life cycle.

In line with analyst estimates, Deloitte predicts that the global revenues from the combined OSS and BSS market, or B/OSS, will be about US\$70 billion in 2025, up from an estimated \$63 billion in 2023, or about a 5% annual growth rate.³⁶ That said, telcos may want to take advantage of potentially new revenue-generating features delivered via 5G stand-alone, fiber, and more. Plus, the maintenance cost of legacy infrastructures (telcos can spend up to 80% of IT budgets on integrating and customizing legacy B/OSS systems)³⁷ could drive telcos to modernize their B/OSS software at a more rapid pace. The B/OSS software-as-a-service offering is expected to grow at about 18% annually, and moving to the cloud (aka “cloudification”) is expected to grow at 21% annually, which suggests that the various subtypes of BSS/OSS modernization are growing at triple to quadruple the growth rate of the overall BSS/OSS software industry of 5%.³⁸

As modernization accelerates, BSS and OSS can achieve more effective integration through tools such as APIs and microservices, leveraging cloud-based, software-defined solutions that are available in off-the-shelf standard configurations and offer modularity. Furthermore, these integrations may allow telcos opportunities for greater efficiencies (including lower costs), new revenue streams, a more resilient network, greater cyber and operational security, as well as the ability to better leverage gen AI technologies as they merge over the coming years. Service-centricity could be key, as next-gen OSS systems work to orchestrate the provisioning, fulfillment, and assurance on a per-service level, as opposed to a per-tech-domain level. This could help make the processes for OSS systems more service-centric, along with the horizontal consolidation (integrating various technology platforms and supporting software into a unified system, mainly on the cloud) of technology and supporting software.

Many European telcos have already modernized their B/OSS software in the last few years, with some financial gains.³⁹ However, a goal of new deployments will likely be service-centricity. Most of the growth in the next few years is expected to come from the Americas, Middle East and North Africa, and emerging Asia Pacific.⁴⁰ Also, BSS (specifically within customer engagement systems) have been moving to the cloud more recently, while OSS have been slower to shift due, in part, to telcos being cautious about moving important functions to these newer systems; however, that seems to be changing.⁴¹

Bottom line

One question might be “Who provides these new services?” Historically, there were various BSS or OSS solution providers and integrators, or companies built their own solutions in-house. As B/OSS modernizes, established enterprise software vendors and the hyperscalers are attempting to introduce their own offerings.⁴² Their success will likely require modern system integration with a cloud and AI-enabled mindset.

A billing transformation (a subset of B/OSS modernization) can have significant implications as billions of dollars flow through the legacy billing systems. Some telecom executives have had trepidation about putting this revenue at risk with a billing transformation.⁴³ Balancing this financial cornerstone, aligning business objectives, and minimizing business disruption while evolving billing often requires a delicate balancing act. Considerations and options for navigating this gauntlet are further explored in *Navigating the complexities of billing transformation*.⁴⁴

Telcos should spend money to both save money and make money during B/OSS modernization. Cost reduction can be an important part of the business case for modernization, but the opportunity to grow revenues through offering products such as network-as-a-service or converged offerings could help enhance the ability of telcos to monetize fixed wireless access services.

Further, modernizing B/OSS could require telcos to integrate their OSS and BSS, work with industry-standard APIs,⁴⁵ focus on DevOps for cost control, and work with emerging AI and machine learning technologies.

Finally, as telcos look to move from a relatively fragmented B/OSS environment to a more monolithic model, the governance model should also evolve. Where B/OSS was once the exclusive purview of engineering, new stakeholders who should be involved at all points in the modernization process could include the HR, IT, and finance functions.

Silicon photonics: Gen AI communicates at lightspeed

Propelled by the demanding requirements of gen AI, optical devices on silicon are stepping out of research labs and into the limelight of data centers

Duncan Stewart, Karthik Ramachandran, Jeroen Kusters and Christie Simons

Deloitte predicts that sales of silicon photonics chips used as optical transceivers will grow from US\$0.8 billion in 2023 to US\$1.25 billion in 2025, a compounded annual growth rate of 25%.⁴⁶ Although a fraction of estimated global 2026 chip sales of US\$687 billion,⁴⁷ these chips allow generative AI data centers—which need to move around much more significant amounts of data at higher speeds than other data centers—to communicate at lightspeed, using smaller, cheaper components, and less energy, and producing less heat (better thermal management) than the traditional alternatives.⁴⁸

Silicon chips work in the electrical domain, either communicating with other chips using electrical signals over wires or needing to be attached to or combined with external lasers and modulators that use photons traveling over fiber optic cables. Fiber usually has higher bandwidth than copper wires; signals can travel longer distances using less energy. Moreover, fiber optic cables are immune to electromagnetic interference, which can be a problem for copper wires. Fiber optic cables are often more challenging to tap into or intercept than copper wires, making them more secure. However, traditional photonics have limitations, mainly around cost and size, which silicon photonics hope to overcome.⁴⁹

In 2025, photonic devices are expected to increasingly be made:

- Using the same material as many electronic chips—silicon.
- Using a substrate of the same material as many electronic chips—silicon.

- Using the same manufacturing fabrication techniques as many chips.
- Using a mature ecosystem for silicon photonics across design, manufacturing, foundry, testing, packaging, and assembly; and should be compatible with the comparable ecosystem for making silicon chips today.

This can allow chip companies to integrate electronic and photonic components on a single chip. Over time, this could matter in many different use cases. However, in 2025, the main driver of silicon photonics adoption is expected to be in data center applications, specifically for those running gen AI training and inference. Where most data centers have chips, trays, and racks that communicate with each other at speeds of less than 100G (100 gigabits per second), gen AI equipment needs to move more data faster—speeds of 400G or even 800G are required—and photonics is the optimal solution.⁵⁰

Some necessary, detailed data center context may be needed. Within gen AI data centers, there are many server racks. The standard size is 24 inches (600 mm) wide, 42 inches (1066.80 mm) deep, and 73.6 inches (1866.90 mm) tall: This is called a 42U (each U is a “rack unit” that is 1.75 inches [44.45 mm] in height).⁵¹ Many different types of chips and racks need to talk to each other over different distances and speeds, which are determined partly by rack dimensions.

As a result, there are opportunities for silicon photonics in 2025 and beyond. Different technologies have different sweet spots, primarily driven by the distance between components, with silicon photonics having an optimal zone that is not too short, not too long of more than 10 cm and less than 10 meters, where it could have the greatest near-term advantages versus copper or traditional photonics and therefore may have more opportunities for near-term revenues.

Chip to chip, on a tray: One configuration for gen AI rack-scale servers features trays of 2 GPUs and 1 CPU that are either 1U or 2U in height, depending on the cooling technology chosen. In 2025, the tray-level communication between chips (distance is less than 10 cm) is done electrically, but this could evolve and be done optically over time. Given the limited space available (two to four inches in height) and an effort to keep costs low, this may need to be done with integrated silicon photonics rather than discrete photonic devices. However, given the very short distances, electrical signals could be adequate in 2025.

Tray to tray, on a rack: There can be 18 1U trays as described above in a single server rack. This is the densest possible configuration. Each tray needs to talk to all the others by communicating from a distance of no more than a meter or two vertically.⁵² In early 2025, this will be doable optically for about US\$144,000 per rack, according to one estimate.⁵³ By the back half of 2025, or early 2026, silicon photonic devices could start gaining traction in this application.

Rack to rack, but close: For various reasons (power, cooling, cost) there could be many paired server racks with half the density, sitting beside each other and needing to communicate over a meter or two. Communication between two server racks could, at some point, almost entirely be done optically, and this may be the largest near-term opportunity for silicon photonics in 2025.

Rack to rack, but less close: Each server rack (or pair of server racks) needs to talk to all the other rack servers (and memories and processors of various kinds) across a full hyperscale data center: These can involve fiber optic cables that are tens or even hundreds of meters long. Silicon photonics can offer very high bandwidths and long reach, reducing the costs and power consumption due to the high level of photonic device integration.⁵⁴ Although the higher cost is a consideration, silicon photonics are not expected to displace traditional photonics for this application in the near term.

Bottom line

One additional prediction around silicon photonics: M&A. If there is continued growth in gen AI data centers, and especially in the need for high speeds and reduced power consumption, both of which may be likely, then large companies may spend money, billions, on acquiring silicon photonics startups, companies, or divisions of other companies who are leaders in what could increasingly be seen as a critical emerging technology.⁵⁵

Although this article is focused on the importance of gen AI data centers in accelerating the demand for silicon photonics, it’s important to note that the technology is of potential interest in other use cases. Perhaps the most notable near-term opportunity is to make on-chip LIDAR units for advanced driver assistance systems (near term) and autonomous driving features (long term).⁵⁶

By

Duncan Stewart
Canada

Prashant Raman
India

Karthik Ramachandran
India

Endnotes

1. Srini Subramanian and Meredith Ward, “[2024 Deloitte-NASCIO Cybersecurity Study](#),” *Deloitte Insights*, Sept. 30, 2024.
2. The Deloitte authors make this prediction based on what they are seeing in the market and what their clients are telling them.
3. Duncan Riley, “[Generative AI services have driven a huge surge in phishing attacks](#),” *Silicon Angle*, May 22, 2024.
4. Michael Crider, “[Hackers are now using AI-generated code for malware attacks](#),” *PCWorld*, Sept. 25, 2024.
5. Tiernan Ray, “[Generative AI is new attack vector endangering enterprises, says CrowdStrike CTO](#),” *ZDNet*, June 30, 2024.
6. Ian Barker, op. cit.
7. Faruk Muratovic, Duncan Stewart, and Prashant Raman, “[Tech companies lead the way on generative AI: Does code deserve the credit?](#),” *Deloitte Insights*, Aug. 2, 2024.
8. Snyk, [2023 Snyk AI-generated code security report](#), accessed Aug. 11, 2024.
9. Ibid.
10. Mandy Address, “[Generative AI for cybersecurity: Is it right for your organization?](#),” *Fast Company*, June 17, 2024.
11. EU Artificial Intelligence Act, “[Article 15: Accuracy, Robustness and Cybersecurity](#),” accessed Aug. 28, 2024.
12. Christie Simons et al., [2024 global semiconductor outlook](#), Deloitte, Jan. 22, 2024.
13. European Commission, “[The Digital Markets Act: Ensuring fair and open digital markets](#),” accessed Oct. 6, 2024.
14. European Commission, “[The Digital Services Act](#),” accessed Oct. 6, 2024.
15. European Commission, “[AI Act](#),” accessed Oct. 6, 2024.
16. Deloitte analysis based on data and chart presented in “[Semiconductor – A treasure trove for private equity investors](#),” (June 2024, p. 11). We used chiplet packaging baseline market share for 2021 (24% of the US\$30 billion total market) and applied 22% CAGR to arrive at the 2025 predicted value of US\$16 billion.

17. Moore's Law notes that the number of transistors on an integrated circuit would double every two years but with a smaller increase in cost—translating into nearly twice the superior performance over the previous generation (because of doubling of transistors by shrinking the linewidths) at a marginal additional cost. However, shrinking transistor linewidths is reaching its physical limit. To read further, see: Alchip's "[Moving from SoCs to chiplets could help extend Moore's Law](#)," Sept. 26, 2022.
18. Deloitte's analysis of multiple publicly available sources including product information published by chip companies, as well as articles from sources such as *EE Journal*, *Semiconductor Engineering*, *EE Times*.
19. Based on our analysis of chiplet-based new product announcements and launches from major semiconductor companies (including IDMs, fabless, and chip design players). Chiplets allow multiple functionalities such as GPU, CPU, and memory components to be densely packed on a single chip. Moreover, chiplets have helped deal with the complexity involved in integrating the diverse components with varying manufacturing and packaging technologies coming in from IDMs, foundries, and other component manufacturers from various regions worldwide. To read further, see: Dr. Uwe Lambrette et al., "[Semiconductor – A treasure trove for private equity investors](#)," Deloitte, June 2024.
20. TrendForce, "[\[News\] Understanding 3DIC, heterogeneous integration, SiP, and chiplets at once](#)," March 19, 2024; Alchip, "[Moving from SoCs to chiplets could help extend Moore's Law](#)."
21. Max Maxfield, "[Are you ready for the chiplet age?](#)," *EE Journal*, July 27, 2023.
22. Yinxiao Feng and Kaisheng Ma, "[Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration](#)," Institute for Interdisciplinary Information Sciences (Tsinghua University, China), April 9, 2024.
23. "[Test & reliability challenges in advance semiconductor geometries](#)" (presentation at 2013 Semiconductor Wafer Test Conference), June 9, 2013. Data for 130 nm and 90 nm based on the chart on page 22 titled "Dramatic rise in systematic yield issues."
24. Deloitte analysis based on our conversations with subject matter experts in the areas of advanced packaging, as well as data and research from publicly available sources, including Feng et al., "[Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration](#)"; Maxfield, "[Are you ready for the chiplet age?](#)"
25. Ann Mutschler, "[Digital twins gaining traction in complex designs](#)," *Semiconductor Engineering*, June 27, 2024.
26. Eric Beyne, "[Chiplet interconnect technology: Piecing together the next generation of chips](#)," *3D InCites*, July 3, 2024.
27. Bilal Hachemi, "[Glass Core substrates: The new race for advanced packaging giants](#)," Yole Group, June 17, 2024; Anton Shilov, "[Intel's glass substrates advancements could revolutionize multi-chiplet packages](#)," *Tom's Hardware*, Sept. 18, 2023.
28. See section Silicon photonics: Gen AI communicates at light speed.

29. Karen Heyman, “*Thermal challenges multiply in automotive, embedded devices*,” *Semiconductor Engineering*, July 2, 2024.
30. Saman Sadr and Richard Lin, “*Securing the new frontier: Chiplets & hardware security challenges*,” *Universal Chiplet Interconnect Express*, Feb. 7, 2024; Nitin Dahad, “*Chiplets are the latest buzz, but many challenges lie ahead*,” *Embedded*, March 10, 2024.
31. Thermal and heat management are noted as one of the major roadblocks to commercializing 3D ICs. To read further, see: Brian Bailey, “*Why there are still no commercial 3D-ICs*,” *Semiconductor Engineering*, Jan. 29, 2024.
32. As noted in *Deloitte Global’s 2024 semiconductor industry outlook*, not only the traditional OSATs but even major IDMs, foundries, fabless companies, EDA vendors, and startups are making the moves and ramping up solutions based on chiplets architectures to push the bar on advanced packaging technologies. Also, see: Ann Mutschler, “*Chiplet IP standards are just the beginning*,” *Semiconductor Engineering*, March 6, 2024; Majeed Ahmad, “*A sneak peek at chiplet standards*,” *EDN*, Sept. 4, 2023.
33. Ann Mutschler, “*Chip design digs deeper into AI*,” *Semiconductor Engineering*, June 3, 2024.
34. Andrew Wooden, “*The evolution of BSS and OSS in the telecoms sector*,” *Telecoms.com*, Aug. 15, 2023.
35. Ibid.
36. Deloitte analysis, based on Alex Bilyi, “*CSPs’ spending on telecoms-related OSS/BSS software and services will reach USD80 billion by 2028*,” *Analysys Mason*, Nov. 13, 2023.
37. Nia Batten, “*The hidden costs of legacy tech*,” *Data Centre Review*, Sept. 1, 2023.
38. Deloitte analysis, based on Alex Bilyi, “*CSPs’ spending on telecoms-related OSS/BSS software and services will reach USD80 billion by 2028*.”
39. Chris Silberberg and Chris Barnard, “*How telcos are transforming in Europe: Technology, services and customers*,” *IDC*, Sept. 1, 2022.
40. Deloitte analysis, based on Alex Bilyi, “*CSPs’ spending on telecoms-related OSS/BSS software and services will reach USD80 billion by 2028*.”
41. Mark Mortensen, Andy He, and John Abraham, *Market pulse: Digital transformation of BSS/OSS to the cloud & DevOps*, *Analysys Mason*, Jan. 2018.
42. Ryan, “*OSS/BSS in the clouds*,” *Passionate about OSS*, July 20, 2020; Anjali Mishra, “*OSS/BSS market players are building robust systems to support next-gen networks*,” *Global Market Insights (GMI)*, May 6, 2022.

43. Amit Kumar Singh et al., “*Navigating the complexities of billing transformation*,” *Deloitte Insights*, 2024.

44. Ibid.

45. TM Forum, “*Introduction to Open APIs*,” accessed Sept. 24, 2024; GSMA, “*GSMA Open Gateway API descriptions*”

46. Deloitte analysis and interpolation of *Light Trends Newsletter*, “*Sales of silicon photonics chips will reach \$3 billion by 2029*,” *LightCounting*, May 2024.

47. WSTS, “*WSTS Semiconductor Market Forecast Spring 2024*,” press release, June 4, 2024.

48. Adam Carter, “*Silicon photonics key to unlocking AI’s full potential*,” *EE Times*, Aug. 18, 2023.

49. Deloitte analysis based on publicly available third-party sources, including Karen Heyman, “*Transitioning to photonics*,” *Semiconductor Engineering*, April 13, 2023; Maxime Fazilleau, “*What makes optical fibre immune to EMI?*,” *Tiny Green PC*, Jan. 23, 2017.

50. FiberStamp, “*Driving the future of high-speed data transfer: The role of PAM4 and silicon photonics in the age of AI*,” *Medium*, Nov. 8, 2023.

51. Christopher Tozzi, “*A guide to server rack sizes for data centers*,” *Data Center Knowledge*, Jan. 8, 2024.

52. Mary Zhang, “*Data center racks, cabinets, and cages: An in-depth guide*,” *Dgtl Infra*, Sept. 28, 2023; Tozzi, “*A guide to server rack sizes for data centers*.”

53. Dylan Patel and Daniel Nishball, “*Nvidia’s optical boogeyman – NVL72, Infiniband Scale Out, 800G & 1.6T Ramp*,” *SemiAnalysis*, March 25, 2024.

54. M. Duranton, D. Dutoit, and S. Menezo, “*3 - Key requirements for optical interconnects within data centers*,” in *Optical Interconnects for Data Centers*, Tolga Tekin et al. (eds) (Sawston, UK: Woodhead Publishing, 2017), pp. 75–94.

55. Contributions from Deloitte subject matter specialists in July and August 2024.

56. Eric Walz, “*Stellantis invests in lidar startup SteerLight*,” *Automotive Dive*, April 2, 2024.

Acknowledgements


The authors would like to thank **Jack Fritz, John Levis, Stephen Winsor, Sandy Lawrence-Morgan, Essaki Velusami, Kannan Ramakrishnan, Nina Zhang, Gautham Dutt, and Dan Hamling** for their contributions to this article.

Cover image by: **Jaime Austin; Getty Images, Adobe Stock**

1/7/25, 3:09 PM

TMT Predictions 2025 | Deloitte Insights

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.



Let's make this work.

[Change your Analytics and performance cookie settings](#) to access this feature.