

Generative AI

2025

Assessing Opportunities and Disruptions in an Evolving Trillion-Dollar-Plus Market

The adoption of generative artificial intelligence (AI) and large language models (LLM) is already rippling through every segment of the technology sector, as incumbents and new entrants reimagine existing end markets to potentially generate \$1.6 trillion in revenue by 2032.

- **Bigger Clusters for Training, Inferencing:** The demand for larger sized training and inferencing compute clusters is set to continue for improved model reasoning and capabilities. Among the foundational model companies, OpenAI-Microsoft, Google, Meta and Anthropic continue to release new LLMs with more features.
- **Gen AI Agents Set to Explode:** In addition to the established use cases around coding copilots and customer-service chatbots powered by LLMs, the integration of writing assistants, and image and video creation tools using text and voice prompts will power gen AI agent deployments for consumer and enterprise uses.
- **AI Inference Aided by Model Efficiency:** Most LLM companies are focused on driving model efficiency for achieving scale in inferencing. Instead of processing everything on the cloud, on-device AI adoption will be driven by a focus on privacy for features like writing text, translating language, generating images and videos with text-based prompts, and voice assistants.

Featured in This Report: Beginning in Section 3 and throughout this report, Bloomberg Intelligence's interactive market-sizing models are used to forecast growth potential for total generative-AI spending and segments including, hardware, software, training, inference, advertising and services. These calculators are available on the Terminal at [BI INET <GO>](#) and [BI GENAG <GO>](#).

Nov. 20, 2024

Contents

Section 1.	Executive Summary	2
Section 2.	Catalysts to Watch	3
Section 3.	AI Overview	4
Section 4.	Market Disruption	16
Section 5.	Segment Analysis	21
Section 6.	Expanding Uses	39
Section 7.	Capital Spending Outlook	54
Section 8.	Processing, Memory Chip Demand	56
Section 9.	Regulatory Landscape	64
Section 10.	ESG Outlook	68
Section 11.	Performance and Valuation	70
Section 12.	Company Impacts	73
Section 13.	Methodology	77
Section 14.	Glossary of Terms	82
	Bloomberg Intelligence Research Coverage	84
	Copyright and Disclaimer	87
	About Bloomberg Intelligence	88

Lead Analysts

Mandeep Singh	Software, Internet, Hardware Americas	msingh15@bloomberg.net
Anurag Rana	Software, IT Services, Americas	arana4@bloomberg.net

Contributing Analysts

Masahiro Wakasugi	Japan Technology, Global Semis	mwakasugi4@bloomberg.net
Poonam Goyal	E-commerce & Athleisure	pgoyal4@bloomberg.net
Steven Tseng	Technology, Hardware, APAC	htseng18@bloomberg.net
Sunil Rajgopal	Software, Americas	srajgopal4@bloomberg.net
Charles Shum	Semiconductors, Electronic, APAC	cshum2@bloomberg.net
Woo Jin Ho	Hardware, Networking, Americas	who88@bloomberg.net
Kunjan Sobhani	Semiconductors, Americas	ksobhani@bloomberg.net
Tamlin Bason	Tech Policy, Litigation, EMEA	tbason3@bloomberg.net
Breanne Dougherty	Equity Strategy - Thematics	bdougherty25@bloomberg.net

Editorial & Visuals

Rik Stevens, Justin DeVoursney, Philippe Tardieu

More detailed analysis and interactive graphics are available on the Bloomberg Terminal

Section 1. Executive Summary

\$1.6 Trillion

AI-driven sales by 2032
from about \$93 billion in
2023

37%

Projected CAGR
through 2032

\$646 Billion

AI Training Market

Racking Up 14-16% of Technology Spending

Generative artificial intelligence is poised to produce \$1.6 trillion in revenue, or 14-16% of all technology spending, through 2032 in hardware, software and services and more as businesses supercharge their products and more consumers embrace the shift. The interfaces and tools leveraging gen AI are early, but we believe some common themes include generating summaries, personalized recommendations, image and video content using conversational user interfaces, and built-in language translation. Training AI through machine learning and neural network algorithms using massive datasets (the large language model, or LLM) will be a huge market, reaching \$646 billion in sales by 2032 and boosting demand for accelerators on servers and storage units at data centers. Companies will use the public cloud to deploy generative AI, benefitting hyperscalers like Meta, Microsoft, Amazon and Alphabet, with a projected CAGR of 54% to \$231 billion.

Key Research Topics

- **Training Aided by LLM Scaling:** Training will remain the largest segment of the gen AI market, aided by growing GPU cluster sizes for scaling large language models. Hyperscale vendors are likely to see new AI workloads deployed on top of the enterprise data in their cloud infrastructure.
- **Narrowing the Gap Between LLMs:** A convergence in model functionality between OpenAI GPT, Google Gemini, Meta Llama, Anthropic Claude and Mistral has increased the likelihood that increased commoditization in LLMs with the lowest-cost provider will be based on open-source model.
- **Inference on Applications and Devices:** Smartphone makers such as Apple and auto OEMs like Tesla could reap benefits from the demand for inference-based conversational AI products and vision AI offerings tied to generative AI. Quantization is increasingly being used to shrink the size of trained models for inference on edge devices.
- **High-Bandwidth Memory (HBM) Chips:** As AI models become more complex and training more demanding, HBM chips, which generate high operating profit margins, should be more widely adopted. The speed of chip-performance improvement required for AI is faster than the evolution of miniaturization and advanced packaging, aiding the role of chip testers.
- systems that are deemed to pose a "systemic risk." OpenAI's GPT-4 and Google's Gemini are likely to be the first to fall into that category.

Performance and Valuation

AI (BAIAET Index) was among the strongest performing BI themes in the first half of 2024. The index includes 119 companies, and its top-tier early 2024 performance was driven not just Nvidia but also other hyperscalers, as well names such as Super Micro Computer, AMD and Broadcom.

Section 2. Catalysts to Watch

Model Scaling Driving Adoption

Spending on generative AI has quickly become non-discretionary for enterprises, and we expect sharp growth to be driven by steady hardware investments, uptake of chatbots and attached subscriptions for copilot-type offerings. Retrieval-augmented generation (RAG) allows enterprises to leverage proprietary data lakes to enhance accuracy and reduce hallucinations within LLM responses. RAG will likely become a driving factor for enterprise AI adoption, and we may see many hyperscalers intertwine the technology within their own LLM offerings. Already, companies such as Nvidia have seen huge moves in growth forecasts as a result of the push into AI, while others, like Microsoft (Azure consumption and copilots) anticipate robust gains. Over the last 12 months, a majority of the catalysts have been met, including continued architectural advancements in GPUs and improvement in reasoning capabilities of foundational models, which has in turn led to the adoption of copilots for enterprise use cases. On the consumer side, Google Overview was rolled out at scale while Meta has 500 million MAUs for its Meta AI offering.

Critical Milestones:

- **2024:** GPU, accelerator-chip availability improves for training LLMs. **Surpassed.**
- **2024:** New versions of foundational LLMs show greater accuracy. **In progress.**
- **2024:** Strong attach rates for copilots launched by software companies. **Not met.**
- **2024:** Chatbots disrupt customer service and help save operations costs. **In progress.**
- **2024:** RAG makes enterprises comfortable with deploying foundational LLMs. **In progress.**
- **2024:** New content-generating tools, ad-targeting improvement from large internet companies. **Surpassed.**
- **2025:** On-device AI gains steam with new features for mobile, PCs, AR/VR
- **2025:** Multimodal LLMs become a feature across internet apps, enterprise software
- **2025:** EU on course to adopt first comprehensive regulations through the AI Act.
- **2023-27:** TSMC's generative AI segment reaches compound annual growth of 50%
- **2027:** AI networking could expand by 5x, driven by specific accelerator requirements
- **2030:** Software spending on generative AI hits \$309 billion (10% of the total) from \$1 billion in 2022

Section 3. AI Overview

Addressable Markets Appear Ready to Expand

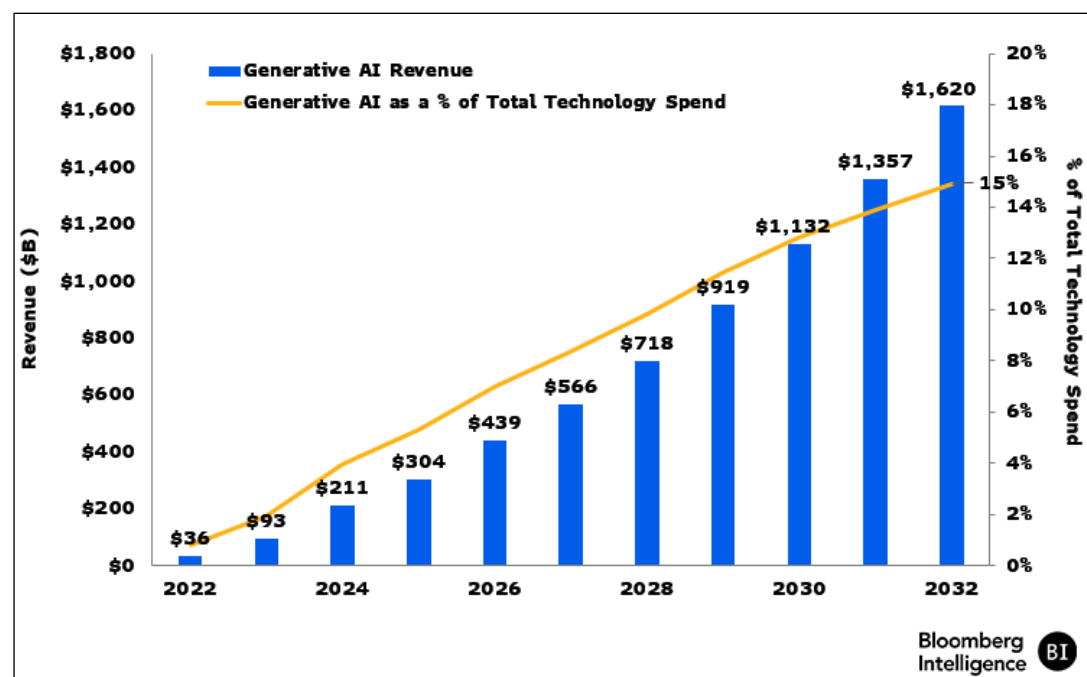
Training foundational LLMs is still the biggest driver of incremental revenue for generative AI, yet traction with GitHub Copilot and growing interest in new apps like Perplexity for consumer search and Sora for prompt-based video generation may continue to expand the addressable market. Generative AI is poised to be a \$1.6 trillion market by 2032 (see Figs. 1 and 2), as it boosts sales for tech's hardware, software, services, ads and gaming segments, growing at a compound annual rate of about 37%, based on BI's interactive market-sizing model. As the revolutionary technology changes how businesses operate and enhance their products and services, generative AI could expand to 14-16% of total information-technology spending in such segments from less than 2% today.

Figure 1: Generative AI Revenue Potential

Generative AI Revenue Projections	2023	2028E	2032E	2023-32E CAGR
Infrastructure (Training)	\$75,519	\$329,393	\$645,962	27%
AI Servers	\$56,642	\$190,373	\$317,419	21%
AI Storage	\$10,893	\$34,267	\$59,055	21%
Training Compute (Cloud Workloads)	\$3,050	\$26,652	\$110,728	49%
Networking	\$3,268	\$19,037	\$36,909	31%
LLM Licensing Revenue	\$1,667	\$59,063	\$121,852	61%
Devices and Applications (Inference)	\$10,451	\$209,159	\$486,651	53%
Inference/Fine-Tuning (Cloud Workloads)	\$2,179	\$49,497	\$121,062	56%
Drug Discovery Software	\$32	\$13,571	\$41,203	121%
Computer Vision AI Products	\$2,813	\$27,147	\$63,517	41%
Conversational AI Products	\$3,750	\$70,583	\$114,330	46%
Chatbots/Copilots	\$1,421	\$41,545	\$134,282	66%
Enterprise Applications	\$257	\$19,085	\$51,315	80%
Cybersecurity Copilots	\$21	\$5,351	\$14,133	107%
Coding and Devops Copilots	\$237	\$13,734	\$37,181	75%
Consumer Applications	\$1,163	\$22,459	\$82,967	61%
Educational Copilots	\$617	\$6,303	\$20,995	48%
E-Commerce Copilots	\$258	\$3,278	\$11,743	53%
Social Media Chatbots	\$52	\$3,722	\$13,048	85%
Customer Service Chatbots	\$237	\$9,156	\$37,181	75%
Educational Content Creation	\$123	\$700	\$2,333	39%
Data Protection	\$134	\$6,115	\$9,923	61%
Generative AI Driven Ad Spending	\$4,638	\$70,179	\$208,987	53%
Search	\$2,472	\$27,959	\$70,174	45%
Videos	\$1,667	\$30,962	\$100,783	58%
Messaging	\$500	\$11,259	\$38,031	62%
IT Services	\$167	\$43,072	\$79,926	99%
Business Services	\$79	\$14,970	\$30,782	94%
Workload Monitoring Software	\$1,214	\$20,046	\$80,797	59%
Entertainment/Media	\$831	\$31,248	\$86,700	68%
Gaming	\$522	\$26,109	\$71,042	73%
Virtual Goods	\$130	\$9,791	\$26,641	81%
Game Design	\$391	\$16,318	\$44,401	69%
Image/Video Generation Tools	\$258	\$4,430	\$13,048	55%
Film Production/Music Generation Tools	\$52	\$709	\$2,610	55%
Total	\$92,901	\$718,067	\$1,619,805	37%

Source: BI's forecasts based on data from IDC, eMarketer, Statista

Figure 2: Generative AI Spending



Source: BI's forecasts based on data from IDC, eMarketer, Statista

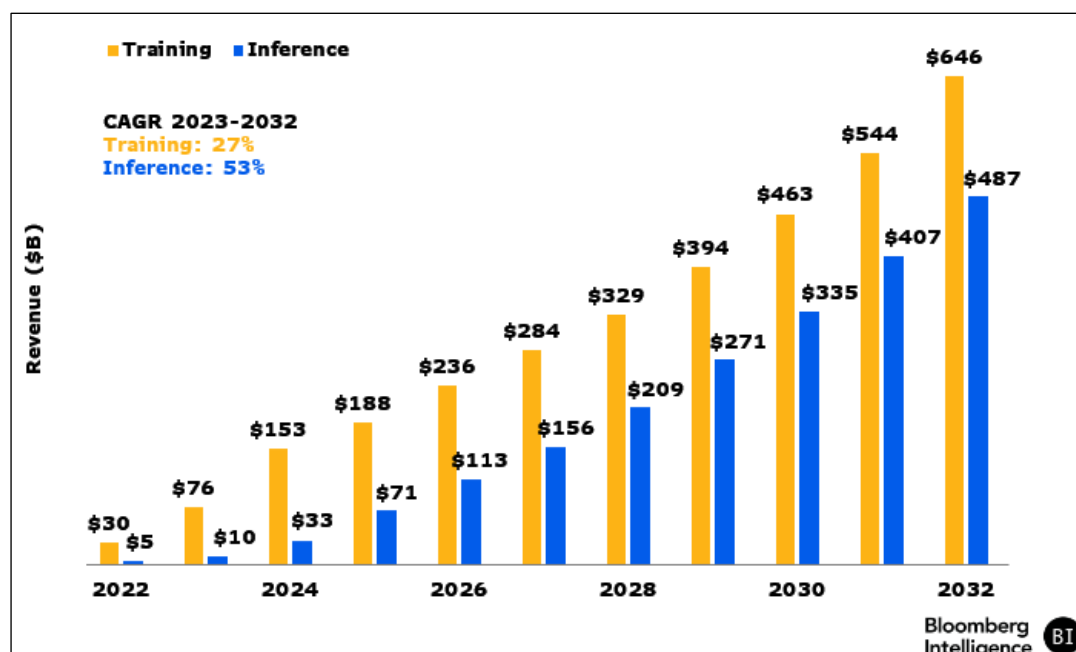
Semiconductors, hardware, cloud software, IT services and advertising companies may be the vanguard of these changes. Yet we also could see new products and services, a displacement of incumbents and the emergence of new categories.

3.1 Training, Then Inference Offer Market Opportunity

Training AI platforms through LLMs based on neural networks with billions of parameters will likely be a bigger part of the market than inference (using previously built models to make predictions or decisions) in the near-term, driving demand for accelerators for servers and storage units at data centers. Training could be the field's largest source of added revenue and nearly a \$650 billion market by 2032, encompassing servers, storage, networking, service offerings on the cloud, as well as LLM Licensing. By 2032, 17% of the training segment revenue will be related to cloud workloads, growing faster than the overall infrastructure market at a 49% CAGR. The high growth expectations have so far been supported by the rapid change in data-center compute and storage to handle AI workloads across both consumer and enterprise apps. Nvidia has been the most significant beneficiary of the trend, while hyperscalers have all boosted capex to ensure GPU availability for both internal consumption and their cloud businesses.

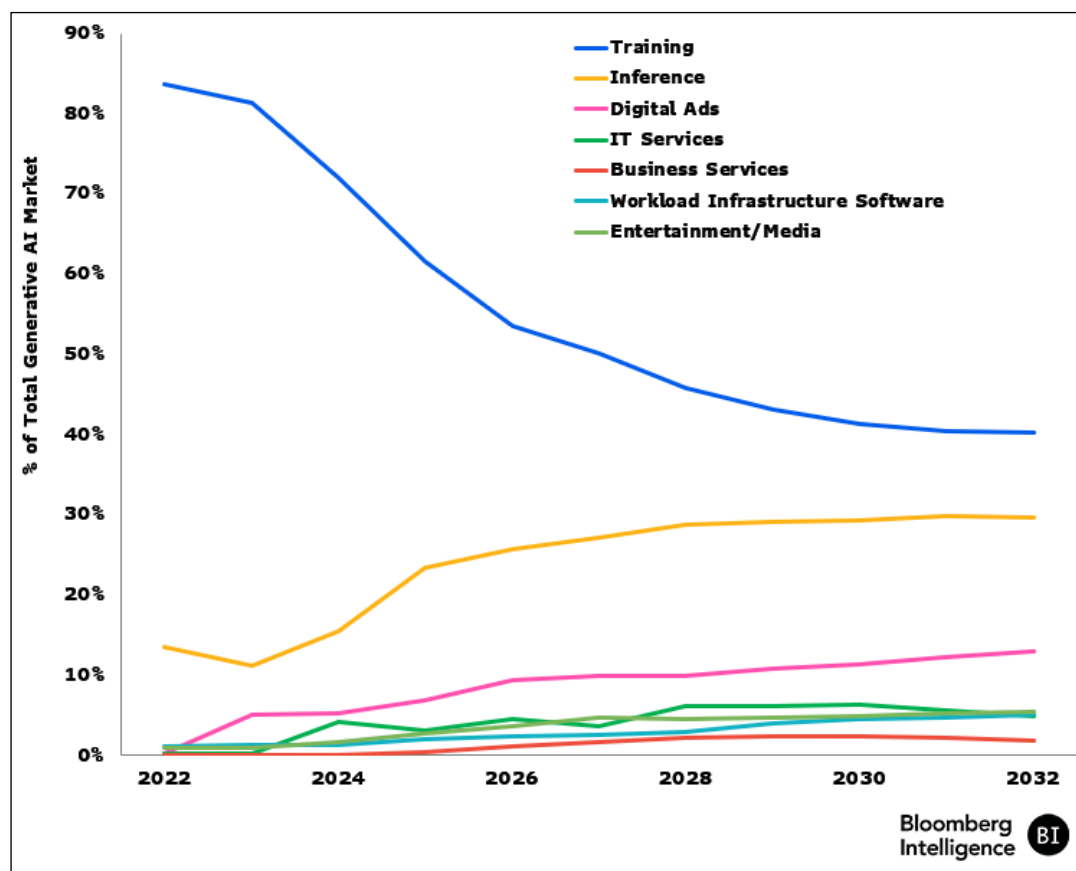
As for inference, computer vision and conversational AI products may emerge as new categories with the availability of LLMs for domain-specific predictions and applications. Such new items can help accelerate growth in the \$1 trillion devices market, which has blossomed with smart speakers and wearables.

Figure 3: Training vs. Inference Forecasts



Source: BI's forecasts based on hardware and software data from IDC

Figure 4: Generative AI Market Share



Source: BI's forecasts based on data from IDC, eMarketer, Statista

Within hardware, infrastructure expenditures (for training) will likely be about triple the size of those for devices (inference) as companies spend on servers and storage to manage the intensive workloads required. Nearly 75% of chief information officers in the US surveyed by Bloomberg Intelligence indicated plans to increase their IT-infrastructure budgets in 2024, with about 38% expecting to boost spending by over 11%. Nvidia and Dell may remain among the top server providers for generative AI workloads, with about 51% of CIOs selecting them. Generative AI infrastructure as a service (IaaS) will be key to training LLMs and could add \$232 billion in sales over a decade. Within generative AI IaaS, the market for compute resources for training will likely be the largest at around \$111 billion, while inferencing workloads run on the cloud could result in a \$121 billion market through 2032. Networking can add about \$37 billion and fine-tuning clouds for specialized use may grow to a \$121 billion opportunity. The market for computer-vision AI products is set to grow to \$64 billion, while sales of conversational AI products could hit \$114 billion. We expect that AI may add \$646 billion to the total AI training market by 2032 from roughly \$75 billion last year.

In software, generative AI products may add about \$309 billion in spending by 2032, growing at a compound annual rate of 71%, with cybersecurity, drug discovery, AI assistants and coding workflow among the top beneficiaries. Many software peers will likely introduce their own AI copilots to enhance the user experience, with specialized assistant software poised to log \$95 billion in sales by the end of the decade. Spending on educational software could be strong in an effort to improve existing learning tools and build new ones. We also expect generative AI to expedite development of gaming and creative software, reducing barriers to entry and creating opportunities for disruption.

On the internet side, generative AI can improve ad targeting and spur the creation of new formats to drive user engagement and increase conversion of ad views to sales. Large companies like Meta and Alphabet rely less on open-internet corpuses than other companies developing foundational LLMs, given their abundance of first-party data to deploy, along with robust capacity to spend that can help train models to improve ad targeting and efficiency. Such enhancements may drive an additional \$210 billion for the digital-ad sector by 2032.

In IT and business services, we estimate that generative AI products and tools can add about \$111 billion in sales as companies look for new products to drive top-line growth and trim unnecessary costs.

3.2 Cloud to Overtake Server Development

Though servers and storage could be the most prominent segments for generative AI services in the near term, many enterprises doubtlessly will leverage public cloud deployment eventually. We believe hyperscalers will develop in-house foundational LLMs, which will work best on their own cloud infrastructures. Meta, Microsoft, Alphabet, Nvidia, Amazon and other such suppliers may be among the main facilitators for training LLMs. These companies have access to the capital needed to set up the training infrastructure while keeping usage high for their servers to sustain healthy profit margins.

In time, generative AI as a service should be a much bigger market than servers and storage, as seen in Fig. 1, logging 49% growth compounded annually through 2032 as gains for stand-alone

servers and storage taper off. The trend favors expansion for hyperscale cloud suppliers over smaller infrastructure-software peers, mirroring the evolution of the software-, platform- and infrastructure-as-a-service portions of the roughly \$663 billion public cloud market.

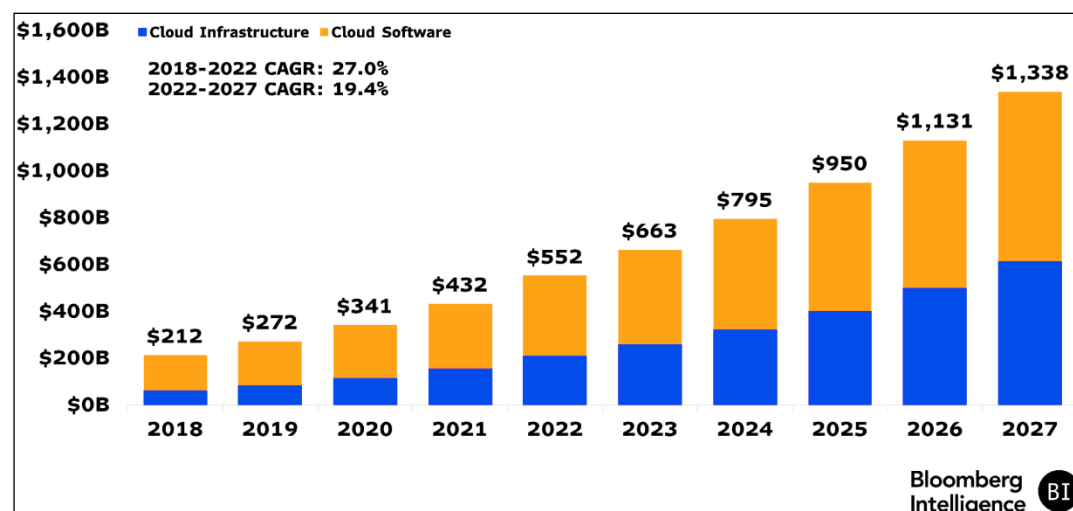
Figure 5: Generative AI Market Overview

Infrastructure (Training)	AI Servers		
	AI Storage		
	Training Compute		
	Networking		
	LLM Licensing Revenue		
Devices and Applications (Inference)	Inference/Fine-Tuning Cloud		
	Drug Discovery Software		
	Computer Vision AI Products		
	Conversational AI Products		
	Chatbots/Copilots	Enterprise Applications	Cybersecurity Copilots
			Coding Devops Copilots
		Consumer Applications	Educational Copilots
			E-Commerce Copilots
			Social Media Chatbots
			Customer Service Chatbots
Educational Content Creation			
Data Protection			
Generative AI Driven Ad Spending	Search		
	Videos		
	Messaging		
IT Services			
Business Services			
Workload Infrastructure Software			
Entertainment/Media	Gaming	Virtual Goods	
		Game Design	
	Image/Video Generation Tools		
	Film Production/Music Generation Tools		

Source: Bloomberg Intelligence

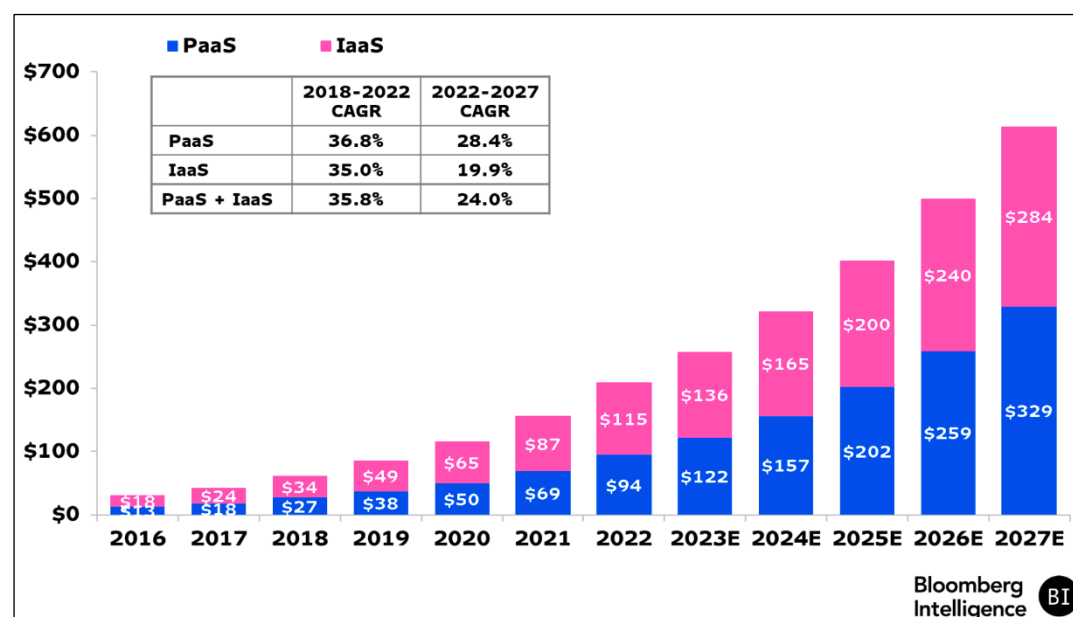
The forecasts in the accompanying graphic are conservative. Though it's highly likely that the enterprise demand shift toward the cloud will gather speed in the coming years, that isn't included in our assumptions.

Figure 6: Total Public Cloud Spending Forecast (\$ Billion)



Source: BI's forecasts based on hardware and software data from IDC

Figure 7: IaaS, PaaS Revenue Forecast (\$ Billion)



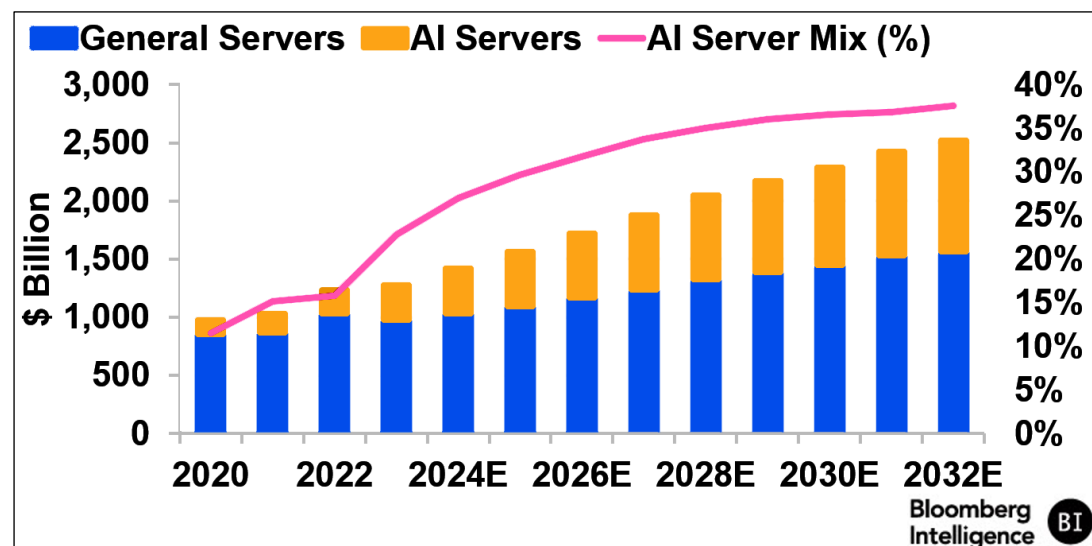
Source: BI's forecasts based on hardware and software data from IDC

3.3 Near-Term Server Demand Should Stay Healthy

The eventual shift to cloud deployment notwithstanding, the explosive demand for generative AI – as evidenced by the insatiable demand for Nvidia GPUs – should fuel significant growth in the infrastructure hardware market, particularly servers (Fig. 8), that provides the necessary computing power. The global market for AI servers is set to roughly double to \$39.4 billion in 2024 from 2022, according to our estimates, notching 41% average annual growth. AI is poised to contribute more than 20% of global server revenue starting this year, from 15% in 2021. Despite

economic headwinds in 2023, spending on AI servers could remain robust, thanks to the ChatGPT-fueled arms race in generative AI.

Figure 8: Worldwide AI Server Market Forecast



Source: BI's forecasts based on hardware and software data from IDC

The lion's share of server demand might go to original design manufacturers (ODM) building customized models for major cloud service providers like Microsoft and Google that are providing significant backing and development for AI applications. Their public cloud infrastructures also offer the necessary scalability for AI development in terms of computing and storage capability. Microsoft is a key investor in OpenAI, the owner of ChatGPT, and Microsoft Azure is the exclusive cloud platform for ChatGPT. Wiyynn, a major ODM server maker based in Taiwan, indicated that AI servers accounted for 20% of its revenue. It expects that revenue contribution to rise further in the coming quarters, given the steady launches and ramp-up in shipments of new AI server projects.

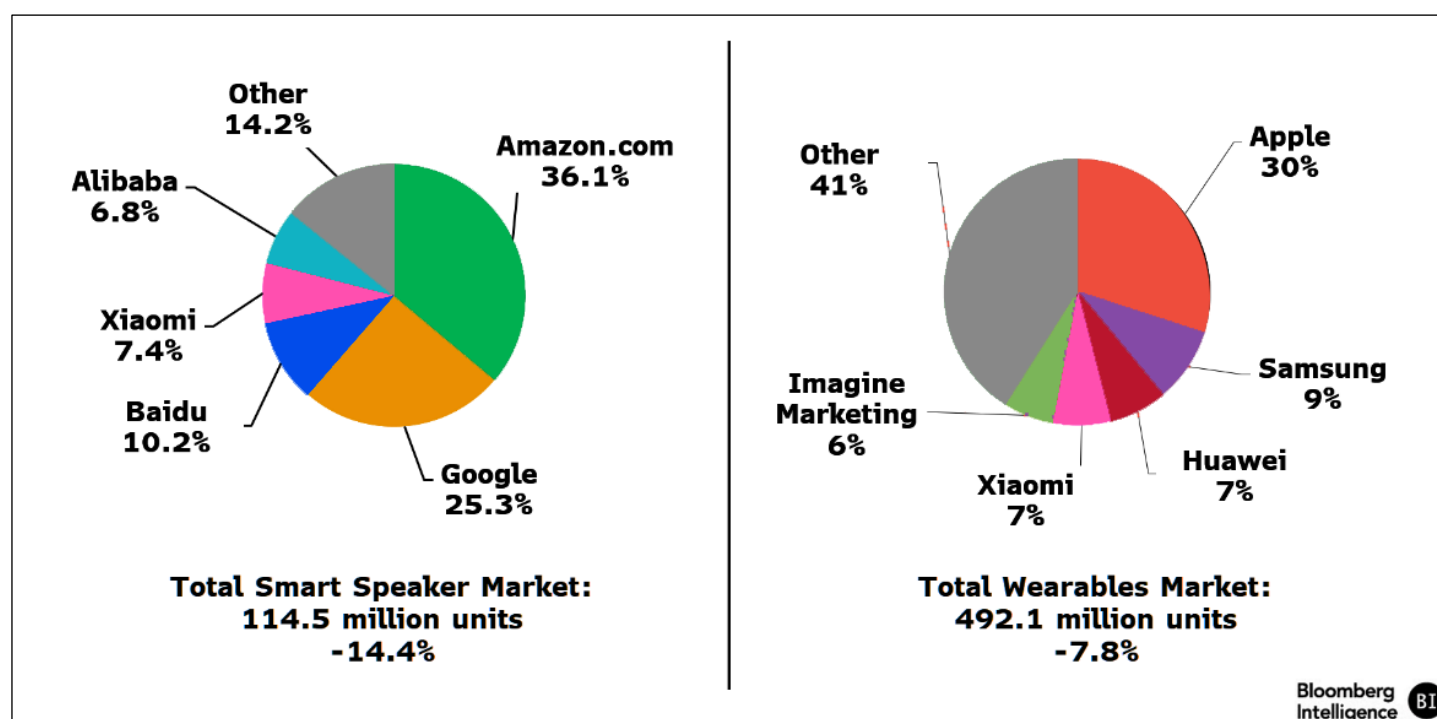
The complex designs of AI servers may help boost profit for related suppliers. While typical servers mainly use Intel and Advanced Micro Devices' x86 central processing units, AI servers use so-called heterogeneous computing architecture, which requires the combination of different processors – such as CPUs, graphic processing units and advanced RISC machine-based (ARM-based) chips – or proprietary application-specific integrated circuits. That mix-and-match approach can optimize system performance and power efficiency but poses a challenge to server designs as each processor has different instruction sets and data transmission cycles. As a result, ODMs with advanced design expertise would have a competitive edge over rivals that don't and might be able to charge more, expanding profitability.

The shift from using general-purpose CPUs to custom accelerators for large dataset workloads is key to why training is poised to be a bigger market (accounting for 35% in 2032) than devices (12%) in generative AI. The use of semiconductor accelerators will likely increase as more companies train their own LLMs, similar to OpenAI's GPT, Meta's Llama and Alphabet's Gemini.

3.4 Faster Hardware Refreshes; Networking a Key

The need for inferencing on so-called edge devices (hardware that controls data flow across the boundary between two networks) could speed refreshes of personal computers and smartphones – which currently aren't well-suited for the heavy processing, memory and storage requirements for AI's LLMs – while spawning new categories beyond wearables and smart speakers. Demand for inference is expected to ramp up as more applications are developed on top of foundational models such as OpenAI's GPT, Google's Gemini and Meta's Llama. We expect more compact versions of these models, such as the Gemini Nano, to be released in the next 1-2 years, allowing edge devices to perform AI workloads natively. These models will likely require less computational power and be less expensive to train.

Figure 9: Smart Speakers, Wearables Market 2022

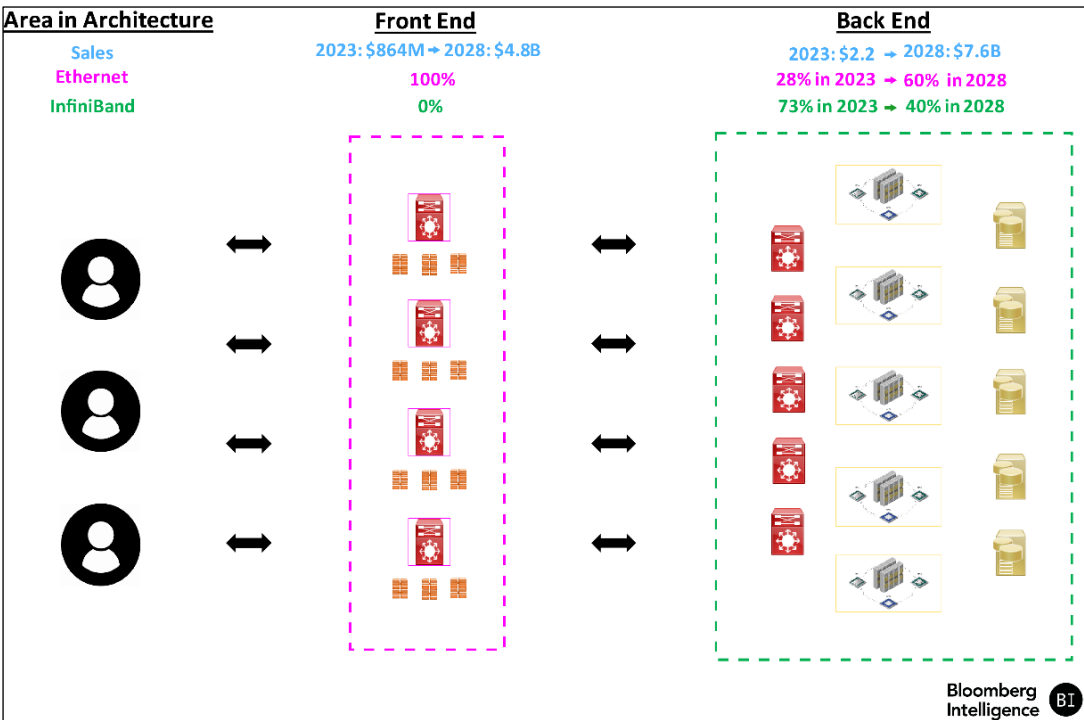


Source: IDC

Networking has emerged as a strategic area of AI infrastructure, along with servers and graphics processing units. It's typically been a bottleneck for hyperscale cloud infrastructure, which companies have aimed to resolve with higher capacity gear. Public cloud generative AI workloads are expected to grow faster than general cloud due to the computing intensity needed to ingest a burgeoning amount of structured and unstructured data for LLMs, supported by clusters of GPU-powered servers that can number in the tens of thousands. This complexity and density of AI architecture requires a separate "back end" AI network that supports high capacities – 800 gigabits or greater – while delivering low-latency, error-free data traffic separate from the general purpose, user-facing "front-end" networks. Given the rapid rise of AI architectures, roughly 18-20% of total Ethernet data center ports are projected to support AI traffic by 2028. AI networking

will likely be a \$19 billion market within the next 5 years, from \$3 billion in 2023, as illustrated in Fig. 14, according to our market forecast.

Figure 10: AI Network Architecture Overview



Source: Bloomberg Intelligence

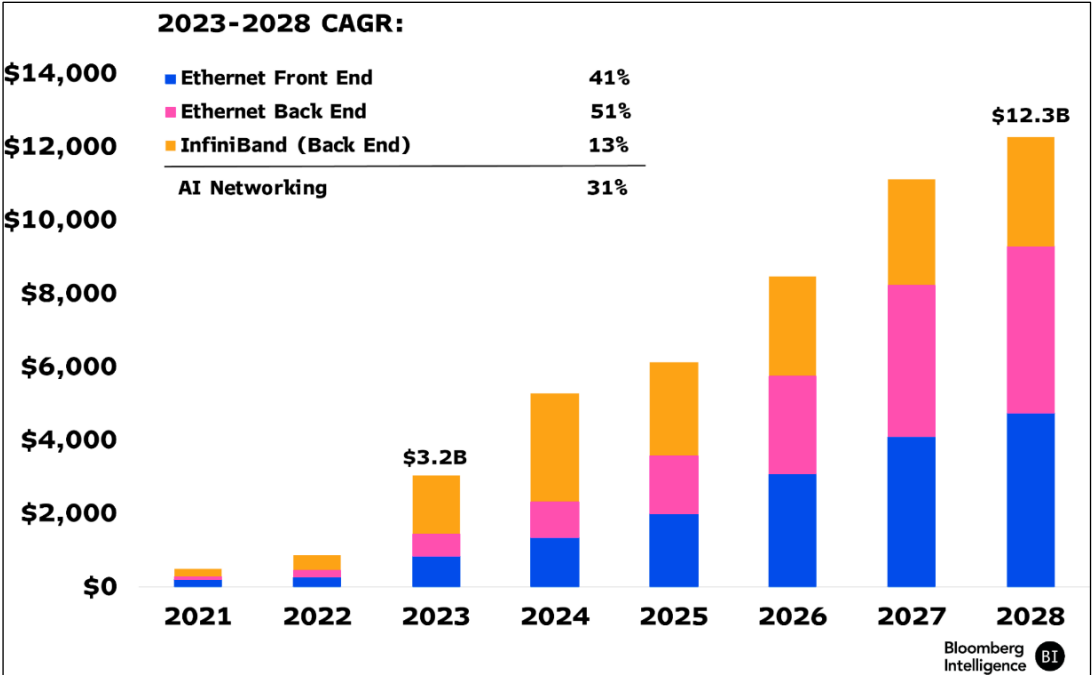
Today, InfiniBand has emerged as the preferred connectivity technology for backend networks, contrasting the ubiquity of the Ethernet protocol powering most of a cloud and corporate data center network. InfiniBand has lineage to high-performance computing and supercomputing environments and the ability for the technology to reliably transport data at high speeds with little data loss, allowing the proprietary technology to account for 73% of AI back-end networking sales in 2023.

Nvidia accounts for nearly all the InfiniBand market through its 2020 acquisition of Mellanox. It leveraged leadership in AI infrastructure – software, GPUs, data processing units (DPUs) and interconnects – to bundle InfiniBand as an integrated product. While Nvidia’s bundled approach sharply contrasts with hyperscale clouds’ disaggregated approach, a tightly integrated, turnkey AI system could fit well with enterprise infrastructures over the long term.

BI

AI networking sales could grow at 31% a year through 2028

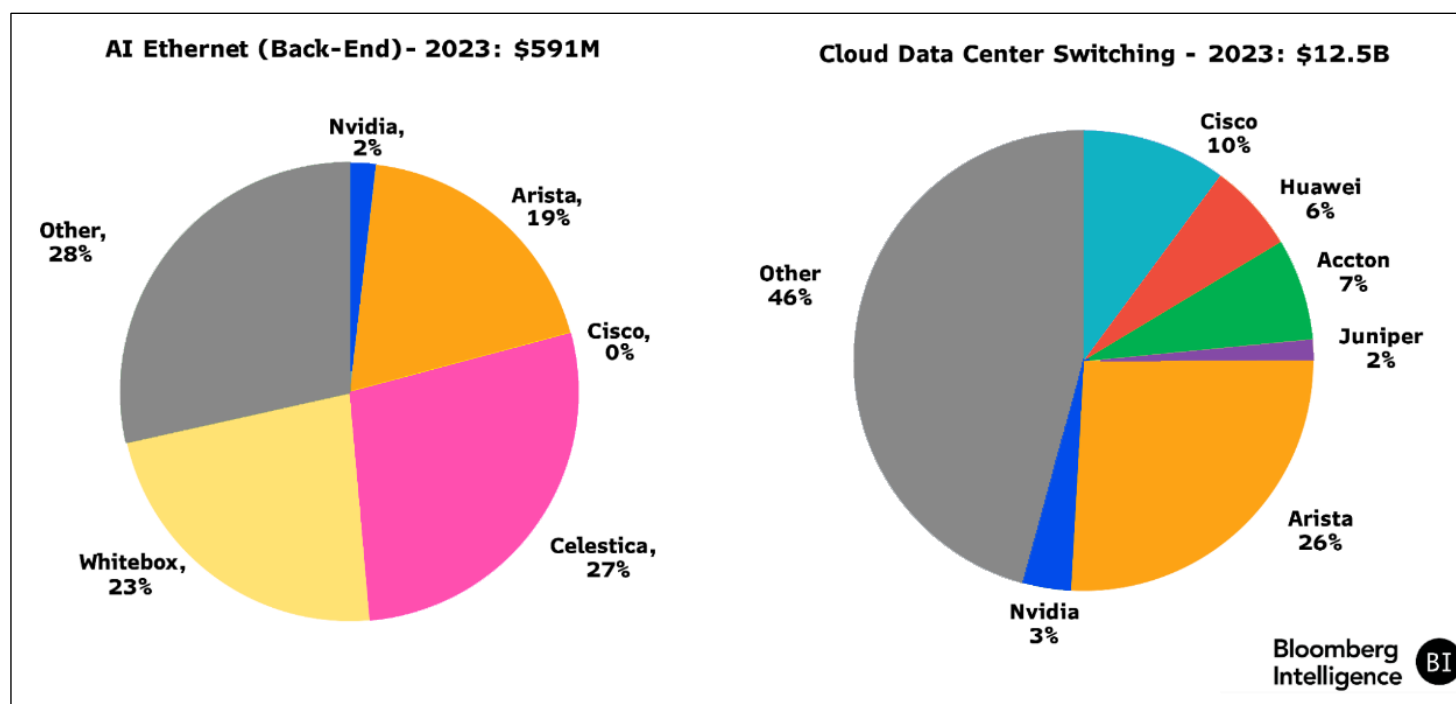
Figure 11: Cloud AI Networking Sales Forecasts



Source: 650 Group

Yet interest in Ethernet technologies is high, especially among cloud customers, and is enhanced by a growing ecosystem of products led by the Ultra Ethernet Consortium. As a result, back-end AI-related Ethernet sales are expected to grow at a 50% compound annual rate for 2023-28 to \$4.6 billion. The latest chip and hardware innovations solve the “bursty” and “lossy” nature of Ethernet data traffic, which could make it more attractive than InfiniBand. Hyperscale cloud interest in adopting Ethernet may be high in part because of familiarity with the technology but also because it allows each cloud to create differentiated AI architectures and helps avoid lock-in with the Nvidia ecosystem. Arista’s strength in high-speed networking gear positions it well to be a leading beneficiary of the shift toward Ethernet by cloud providers. Nvidia meanwhile is well situated to shift customers to Ethernet for AI networks, thanks to the Spectrum switching gear it gained through its Mellanox acquisition.

Figure 12: AI Ethernet, Total Cloud Switching Market Share



Source: 650 Group, Dell 'Oro

3.5 Digital Transformation Initiatives Spill Over

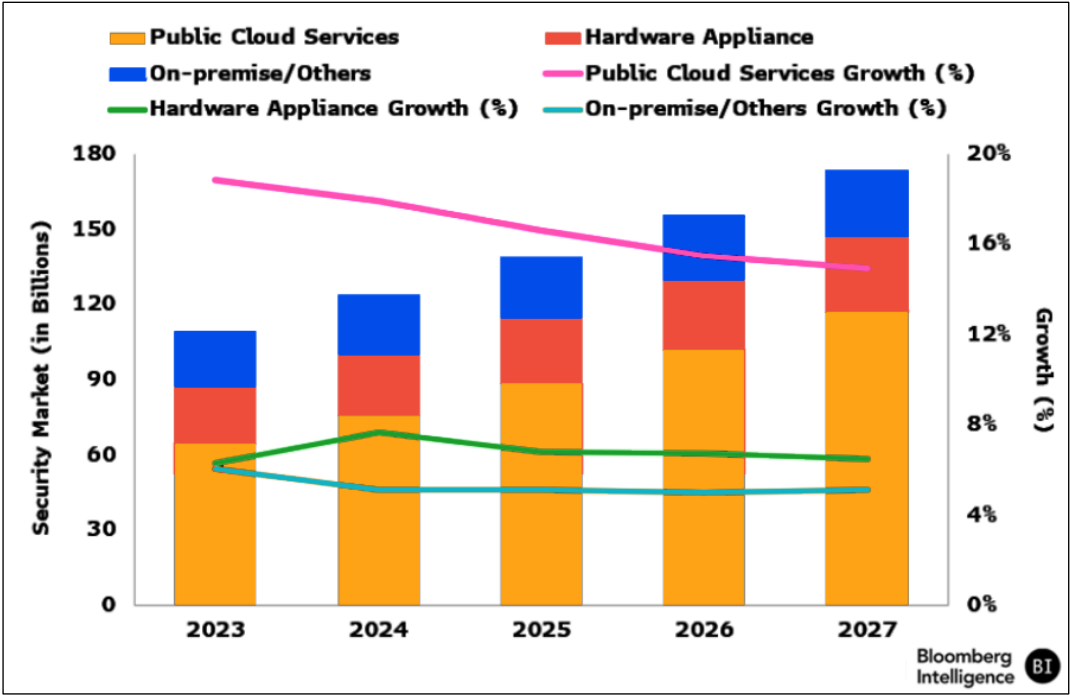
With the rapid development of cloud-based AI technologies like ChatGPT, the significance of edge AI development is growing and represents a significant step in advancing the AI ecosystem. Edge AI is particularly valuable for real-time decision making and cost savings, which are crucial in areas such as health care, manufacturing and transportation and could lead to a larger user base than cloud-based AI. Our scenario analysis finds that the edge AI semiconductor market could be as much as 3.37x the size of the cloud-based AI market by the end of 2032. Adoption of edge AI can drive significant growth in uptake by the consumer (projected to lead other segments with a 10-year CAGR of 39%), industrial and automation sectors over the next 10 years.

Beyond generative AI, advancements in machine learning and other aspects of artificial intelligence appear likely as well. Oracle has been pushing its autonomous databases for the last few years, which could gain from increased budget allocations to AI. We expect increased availability of such features from other software providers, too, where product patching, security updates and tasks usually assigned to a database administrator are automated using machine learning. It could play a much bigger role in cybersecurity in the coming years, as well, especially in event management, in analyzing irregular patterns inside organizations and in the form of copilots to augment security professionals and boost automation.

BI

Public cloud services captures largest share of security market

Figure 13: Market Size by Deployment Type



Source: IDC

Section 4. Market Disruption

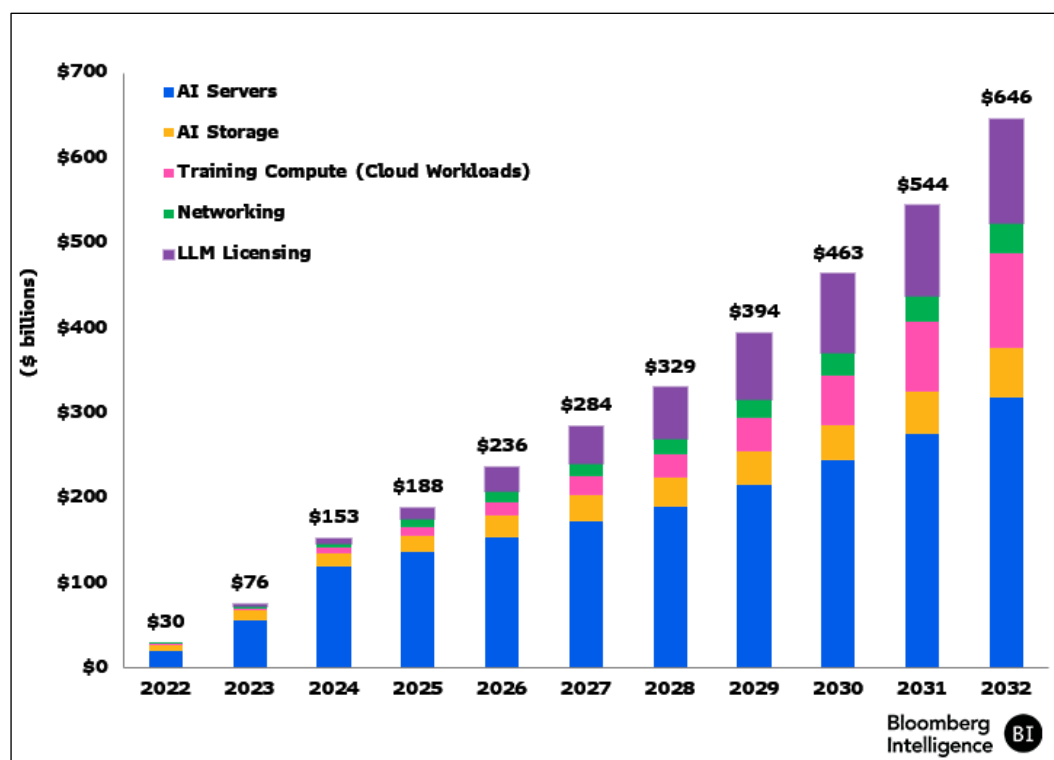
A Shift Coming in Hardware, Ads, Gaming

Generative AI offers opportunities for disruption, especially within hardware, digital advertising and gaming. The computational intensity of training large language models may spark a market-share shift toward advanced RISC machines, potentially making this category the fastest growing within hardware. Alphabet, Meta and other digital-ad giants can improve targeting and brand conversions by implementing machine-learning models based on their vast library of first-party data. Sony, Google, Unity and others in the gaming segment may leverage AI to facilitate development and make the user experience more engaging.

4.1 LLM Training Favors Move to Accelerators

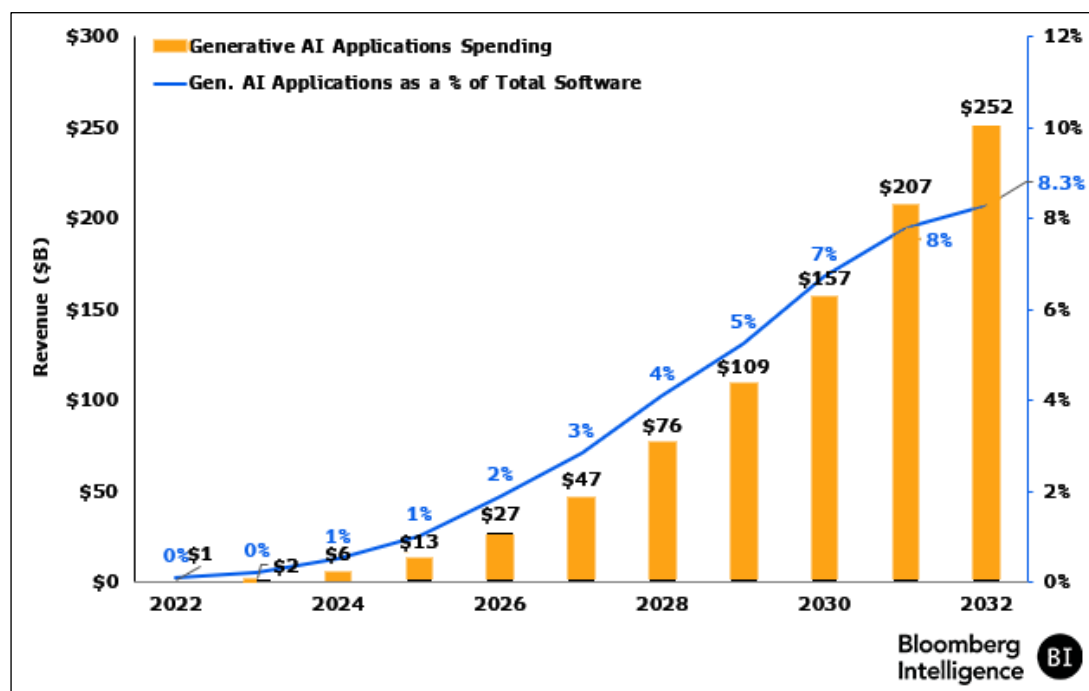
Training LLMs remains a significant driver for the generative-AI market as demand for AI servers and storage units at data centers grows (see Fig. 14). Nvidia may surpass \$100 billion in data-center chip sales in 2024, driven largely by hyperscale customers like Meta, Microsoft and Alphabet. Those companies have said they aim to boost capital spending in the next 1-2 years to meet demand. Companies are leveraging public cloud for deploying LLMs, and we calculate that training compute on cloud could reach \$110 billion in spending by 2032. Alphabet and Meta, along with foundational model companies, may see incremental revenue from licensing their LLMs to smaller enterprises that lack the resources to build their own.

Figure 14: Generative AI Training/Infrastructure Spending Forecast



Source: BI's forecasts based on hardware and software data from IDC

Figure 15: Generative AI Software Spending Forecast

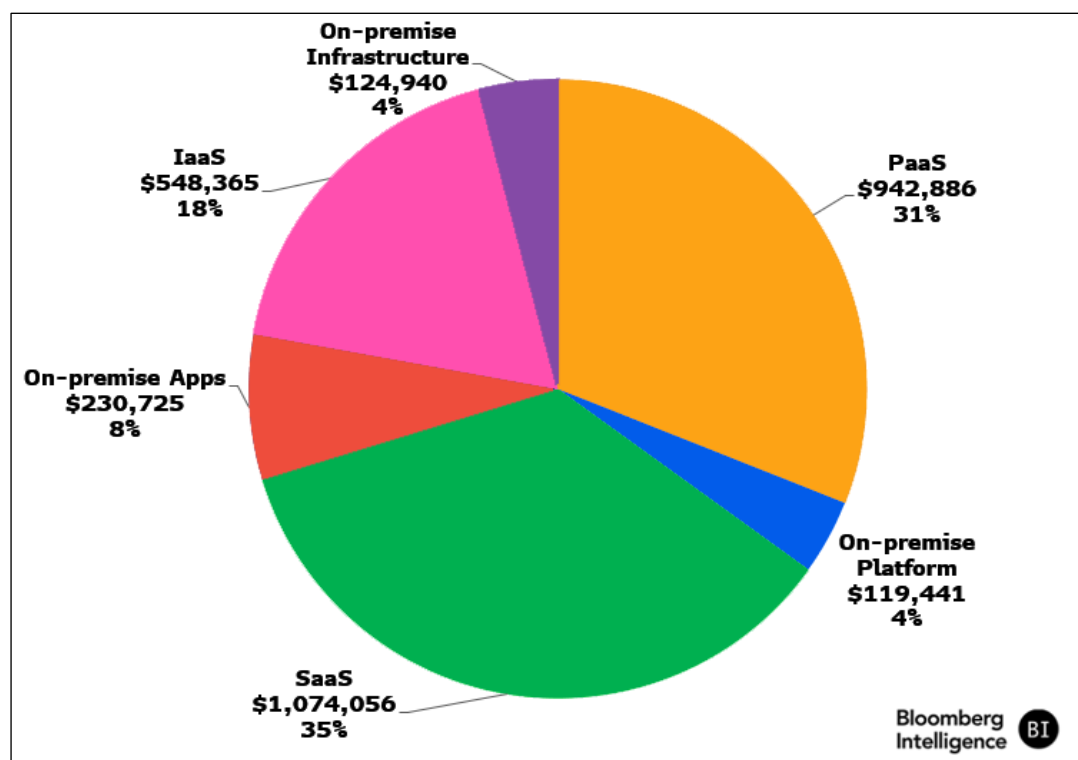


Source: BI's forecasts based on hardware and software data from IDC

BI

Software spending growth at 8.3% by 2032; PaaS, SaaS are market leaders

Figure 16: Software Spending Forecast Breakdown, 2032



Source: BI's forecasts based on hardware and software data from IDC

LLMs have extensive computing and storage needs, central reasons that we expect the first phase of experimentation to be performed with hyperscale cloud providers like Google, Microsoft and Amazon Web Services. Even at maturity, such companies are likely to gain the most market share, given the scale and cost needed to develop an infrastructure in-house.

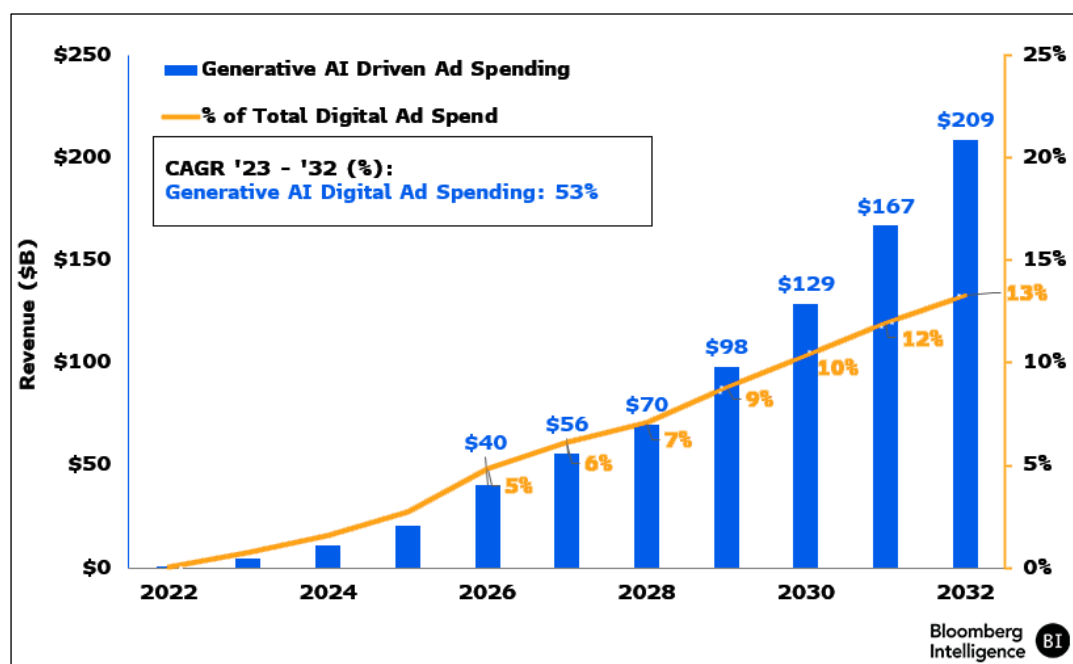
OpenAI's lead in training transformer models and its early partnership with Microsoft has aided ChatGPT's adoption over text-based LLMs from Meta, Amazon, Alphabet and Anthropic. With most hyperscalers investing in developing their own foundational LLMs, we believe OpenAI will need to maintain its lead on the algorithm side while ensuring access to training data from the open-internet corpuses from companies such as Wikipedia, Reddit and Stack Overflow. Alphabet's merger of its DeepMind and Google Brain AI units, can spur faster changes that leverage LLMs to maintain user engagement across revenue generators, like its Search, Chrome and Maps applications. Amid the rapid shift to generative AI, large players like Meta, Adobe, Microsoft, Alphabet and Salesforce are better positioned than smaller rivals for two reasons: they have scads of first-party data in hand and ample ability to deploy capital. Each has a leading market share in its category, offering access to large amounts of information to train AI models, driving more accurate and efficient results.

Social-media platforms like Meta should see a lift as AI-generated content rises rapidly, helping to boost engagement and monetization. LLMs and generative AI can accelerate the shift to digital ads from linear TV. We calculate that more time spent online, coupled with ad targeting and personalization could add over \$209 billion to the market through 2032 (Fig. 17). Conversion rates for ads on these platforms might be aided by the growing capabilities of LLMs, which

should favor companies with strong presences in cloud infrastructure and that have the most first-party data.

Given the high costs of generative AI infrastructure, OpenAI, Anthropic and Google currently have paid subscriptions for their LLM offerings for heavy users. An ad-supported model is unlikely to be profitable for online search and new tools that leverage deep learning and generative AI due to the higher costs of LLM-based queries. A recent Bloomberg Intelligence survey found that only 13% of respondents were willing to pay for a subscription to use a generative AI tool like ChatGPT. Of those, just 1% said they would spend \$20 a month for ChatGPT, with the rest amenable to \$6-\$10. Among all participants, 93% indicated that they wouldn't lay out more than \$10 monthly. The results suggest that lower prices could help boost adoption by 10x for a generative AI subscription. As an illustration: though a freemium version of ChatGPT helped propel it to 100 million monthly active users faster than any consumer app at the time, its conversion to paid subscribers remained in the low single digits.

Figure 17: Generative AI Digital Advertising



Source: BI's forecasts based on digital advertising data from eMarketer

4.2 Sony, Google Parlay New Interfaces for Gaming Design

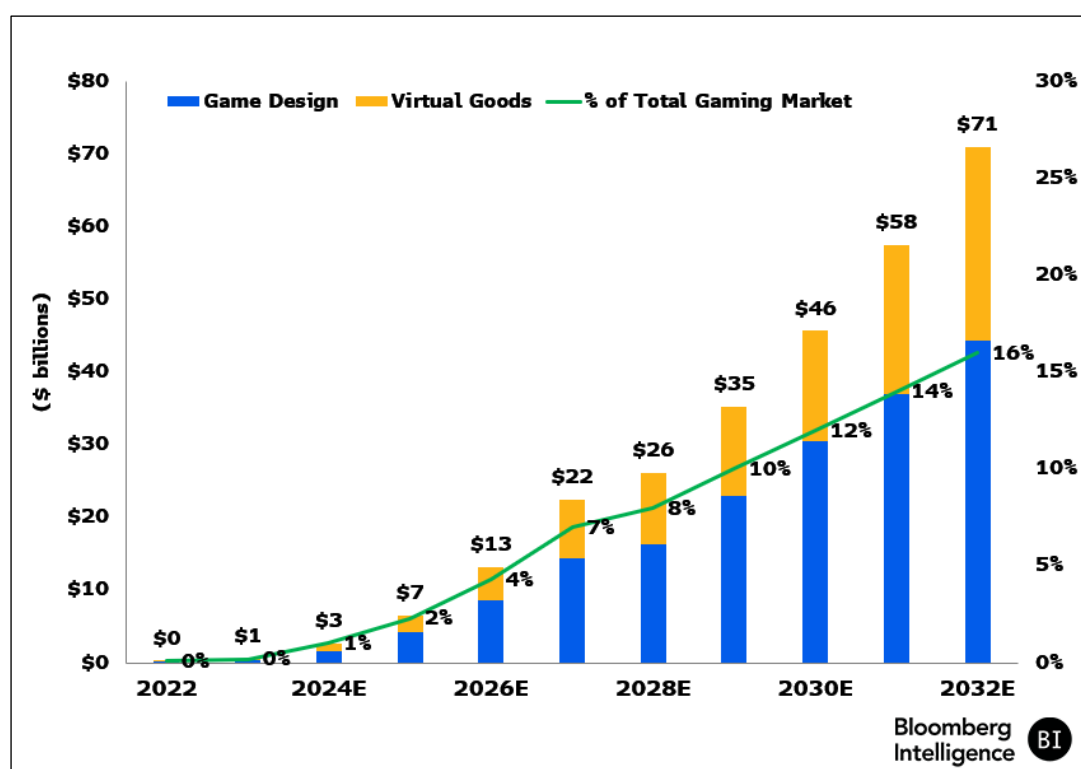
Some startups have already showcased interesting techniques to produce synthetic content – computer-made data that mimics real-world information – based on descriptions and the large amounts of training data available on the open internet. Generative AI can speed that creative process for mobile games, social media and virtual- and augmented-reality applications.

AI tools may rapidly increase gaming data available beyond the high-budget, high-profile major publishers, including those made by users. Developers remain key to gaming and the metaverse

beyond the foundational models provided by tools such as OpenAI's GPT, Google's Gemini, Meta's Llama and Anthropic's Claude. Apple and Google's Android, as well as gaming ecosystems like Sony's PlayStation could offer software development kits to leverage LLMs to ease creation of new content on their platforms. Generative AI might help creative software tools shift to description- and voice-based content from point-and-click user interfaces.

Though Google and Meta have developed LLMs for image generation, they have trailed Stability AI, Midjourney and OpenAI's Dall-E in terms of adoption. OpenAI just released its video-based LLM, Sora, as foundational models become multimodal. Most image-based generative models rely on a diffusion technique and the quality of images rendered depends on training data and the weights assigned to the parameters used. While Adobe has been investing in developing its own generative AI capabilities with the introduction of its Firefly offering, we expect other design- and gaming-software companies to invest in their own generative AI models to leverage proprietary data and distribution.

Figure 18: Generative AI Gaming



Source: BI's forecasts based on hardware and software data from IDC

Section 5. Segment Analysis

Popularity Hinges on Adoption in Existing End Markets

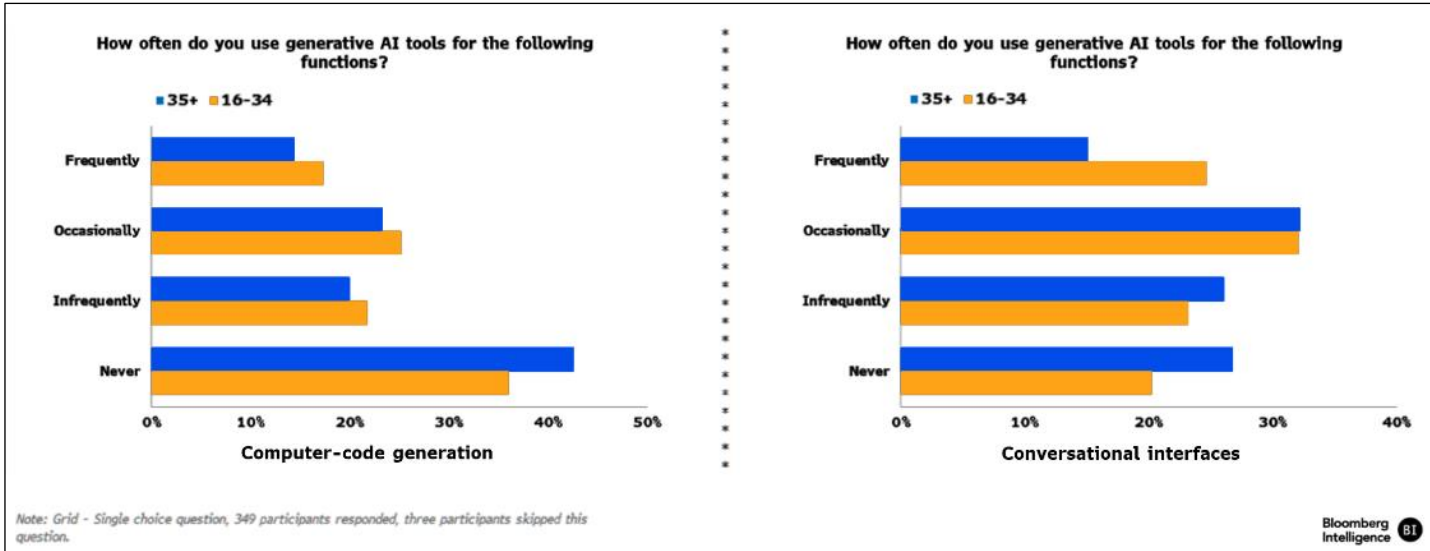
Generative AI is likely to have a far less pronounced effect on the application software industry than infrastructure software in terms of new revenue generated. Within application software, however, we’re already starting to see the rise of AI copilots, with companies like Microsoft, Adobe, Snap and others introducing their own versions of this technology in recent months. Bloomberg Intelligence market opportunity analysis shows that a majority of the \$309 billion of new software sales tied to generative AI is likely to fall in the infrastructure bucket.

5.1 Copilots Lead the Way Into New Endeavors

Education, drug discovery and specialized AI assistants could be the bigger contributors to new revenue streams within application software. Gaming, IT and business services may be smaller categories. The customer-service and business-process outsourcing subsegment of business services might be heavily affected by AI tools and sales could shrink.

Microsoft, Salesforce and Adobe stand out as some of the early adopters of enterprise-AI copilots and agents that focus on sales, customer service and content creation, which Bloomberg Intelligence calculates could grow into a \$83 billion market by 2032. Companies with the largest market shares and capital are likely to garner a bigger share of this pie, widening the gap over smaller vendors.

Figure 19: Coding vs. Conversational Interfaces



Source: Bloomberg Intelligence

The generative-AI copilot and agent market, consisting of AI chatbots that can assist in various ways, from customer service to customer prospecting and content generation, was at roughly \$1.1 billion in 2023, and our analysis of its end markets suggests this could grow 61% a year to \$83 billion by 2032 (see Fig. 20). Customer service will likely be the largest use case, expanding about 75% a year to \$37 billion, given the more straightforward application of copilots into a primarily chatbot setting.

We view "agents," or autonomous copilots that can now take specific actions, as the next evolution of AI assistants, which could accelerate the market further by removing the need for human intervention for certain tasks, thereby increasing return on investment for customers.

Early adopters in enterprise software of these gen-AI chatbots, such as Microsoft and Salesforce, should stay atop peers in the midterm based on their leading market shares, large customer bases and ability to deploy vast resources to consistently develop new capabilities. Their large customer bases not only provide better upselling opportunities, but also create larger datasets for training and fine-tuning the foundational models that power their chatbots.

Adobe is uniquely positioned in creative software, given its vast stock image and video library, which it uses to train its content-generation models. We expect this to be one of the hardest competitive moats to overcome for competitors, as this provides Adobe clients with copyright-safe image and video generation.

As AI-copilot adoption rises, we expect companies to price these either on a subscription or consumption-based model, or leverage these tools indirectly to fortify retention and justify price hikes. Applications designed primarily for assisting a user, such as document summarization or email generation, could be more apt for a subscription plan, such as Microsoft 365 Copilot, whereas more compute-intensive uses may lean on consumption pricing. Examples of the latter may include video generation or customer service inquiries.

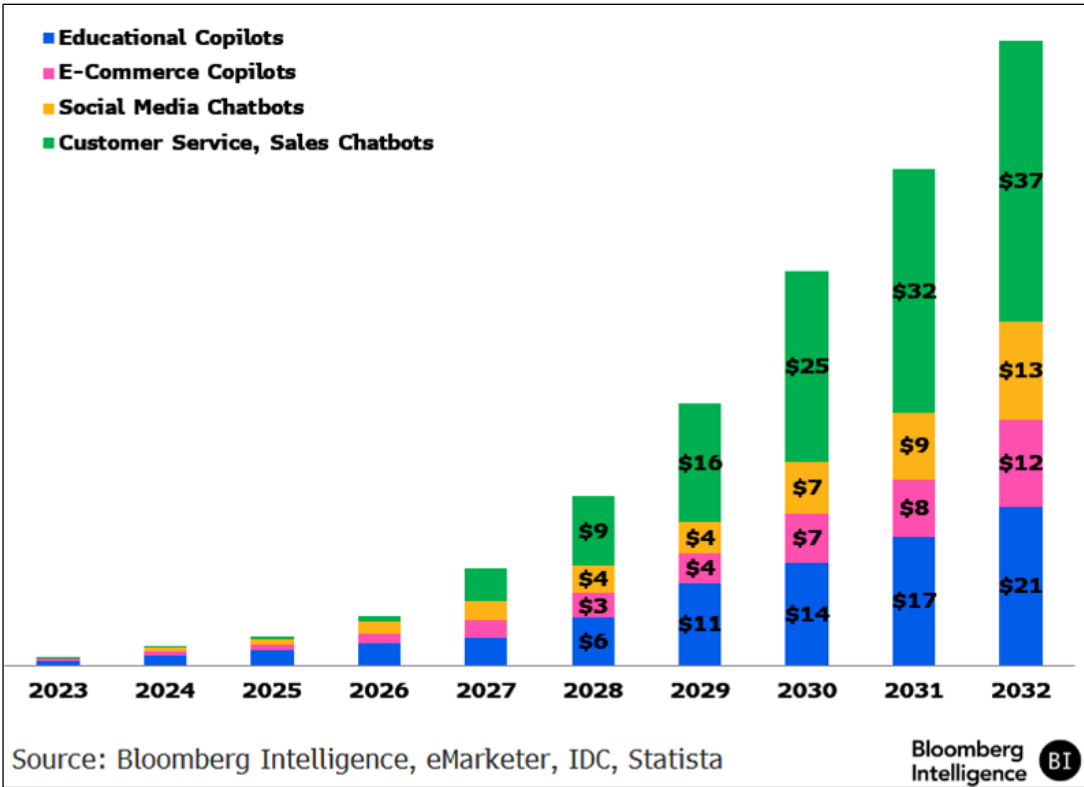
Generative AI has also led to broad-based subscription pricing increases in cases like Adobe who lifted pricing for all Creative Cloud plans by roughly 9-10%. In other cases, such as ServiceNow, pricing for its generative AI Pro-Plus SKU was based on customer savings due to AI, in which ServiceNow only reclaimed 10% of the perceived customer savings, equating to roughly a 60% price uplift over their Pro SKU.

Software providers like Workday have also taken the approach of embedding most gen-AI tools in their products at no extra cost to the customer. We view this as a good way to boost client retention that could lead to future subscription-price increases.

BI

Copilot spending
from \$1.1 billion in
2023 to \$83 billion
in 2032

Figure 20: Generative AI Consumer-Facing Copilot Software Spending Forecast



5.2 Apple Intelligence

Apple's use of its own on-device foundational model that can identify information on the user's screen and take related actions shows the company wants to maintain control of inferencing on its devices. At its WWDC event, Apple showcased App Intents and a semantic index framework for on-device execution, which we believe will be mostly used for tasks like email summaries, image editing and calendar scheduling. Given the widespread use of search, maps and YouTube videos in terms of daily time spent on iOS devices, we believe a large language model integration with Google Gemini would enhance the user experience. Among Android-based smartphone makers, Samsung is utilizing Gemini Nano for on-device AI and Gemini Ultra for larger models on the cloud. The company also showcased its Private Cloud Compute infrastructure for certain types of tasks and an integration with ChatGPT for more complex queries.

Apple showcased AI features such as visual intelligence at its iPhone 16 launch, highlighting AI use cases that are driven by Siri's on-screen awareness, similar to Alphabet's Pixel demo last month. Though Apple is employing small AI models using a semantic index framework vs. Google's Gemini large language model, we believe Apple's iOS 18 software is likely to further narrow the feature gap with its Android rival.

Apple Intelligence will boast a number of new features within iOS 18 as it is released in stages, including text summaries for emails and messages powered by on-device LLMs and an image generator based on prompts, which is already available on Android OS used for the latest Pixel

and Samsung devices. The release of Apple's visual intelligence will help the company narrow the gap around on-device search features powered by gen-AI models, as well as potentially increase Siri usage to better compete with Gemini Live Assistant on Android devices.

5.3 Enterprise CIO Survey

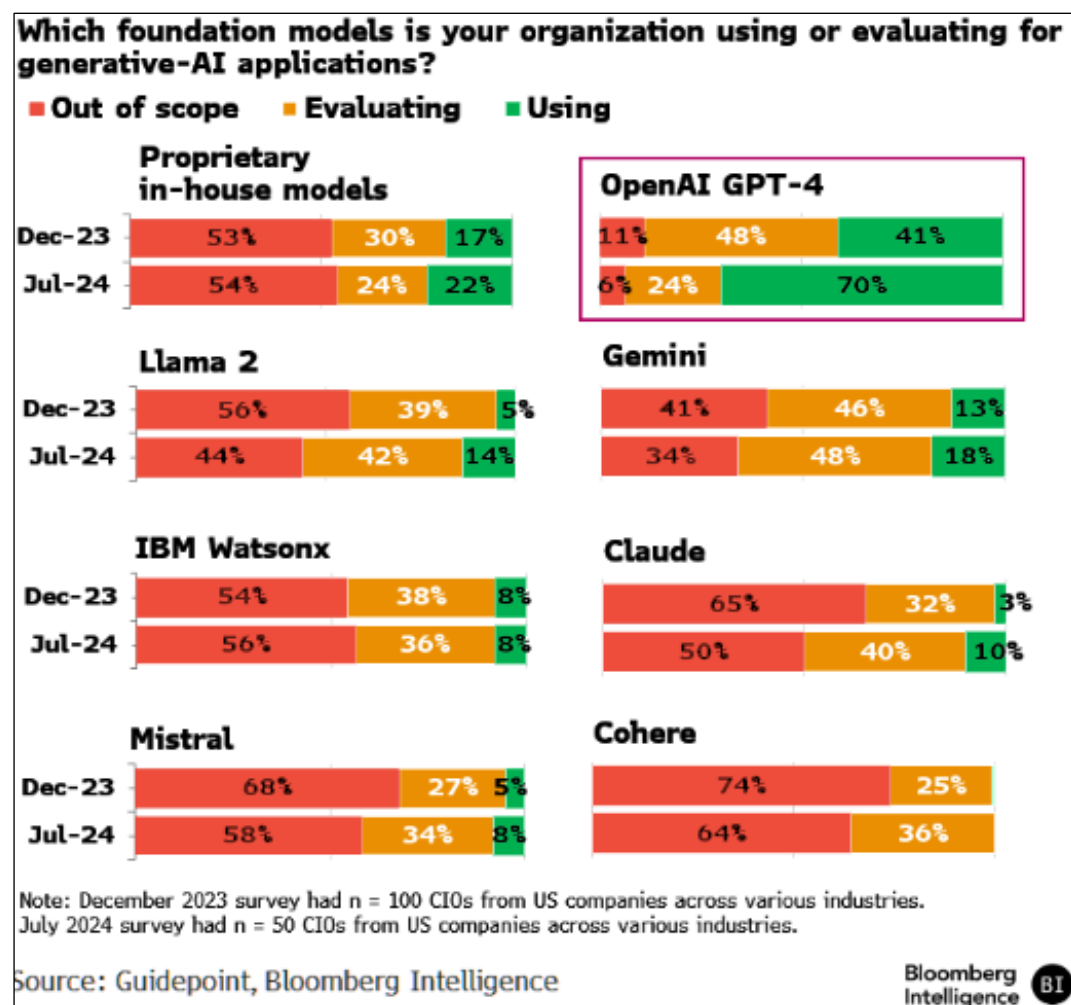
The use of large language models for copilots and chatbots was the most significant driver of enterprise spending on IT infrastructure in our CIO survey. Participants in generative AI signaled a preference for Microsoft Azure cloud and OpenAI's GPT model over other foundational LLMs for inferencing workloads.

Some 66% of respondents said they were working on deploying generative AI copilots, compared with 32% in December's survey. Focus on data validation also increased as companies sought to fine-tune foundational LLMs based on their proprietary data.

The integration of Microsoft's Azure platform with OpenAI models continues to be an advantage over public cloud rivals for hosting inferencing workloads. Others, such as Google Cloud with its Vertex AI, offer application programming interface integration for all LLMs, including Anthropic's Claude, Mistral AI and Meta's Llama. OpenAI GPT isn't offered on Google Cloud or Amazon Web Services.

Microsoft's platform was cited by 28% of respondents for ease of deployment, while 18% named Google and 17% picked Amazon Web Services. Microsoft's lead is probably due to its early tie-up with OpenAI, which can help further bridge the gap with AWS in the cloud infrastructure services market. At the end of 2023 Microsoft had 16% of the market, compared with AWS at 47%. In 2018 Microsoft had 12%, while AWS had 48%. We expect the gap to narrow further by the end of 2025 as Azure's sales growth may outpace AWS' by 1.5x-2.5x.

Figure 21: Foundational Model Preference



Oracle and Salesforce are well-placed to gain cloud sales as enterprises look beyond hyperscalers on AI. Our latest CIO survey indicates that Oracle could become the No. 4 cloud infrastructure provider, after AWS, Microsoft and Google, while Salesforce taps its strong CRM market share.

Oracle Cloud Infrastructure's demand among survey respondents increased about 4 percentage points for AI training and roughly 7 points for inferencing workloads, supporting our thesis that the product eventually could become the fourth-largest cloud provider, behind Amazon Web Services, Microsoft and Google. In July, 11% of responses indicated plans to increase spending on OCI for AI inferencing, compared with 4% in December's survey. For training, the figure rose to 7% from 3%, possibly aided by access to Nvidia GPUs, Oracle's competitive pricing and its ability to increase capital spending significantly.

CIO spending plans for Salesforce regarding AI training, inferencing and copilots logged among the largest increases outside the Big Three hyperscale cloud providers, suggesting that the company's new Data Cloud and adjacent Einstein One products are gaining momentum. Data Cloud enables customers to combine and optimize data for training and inferencing various generative AI models on the Salesforce CRM platform.

The increases in our survey also could indicate customers' preference for having their data aligned between the platform and the AI workload application, likely resulting in a more secure and better end-user experience.

Our survey reveals a sustained rise in AI training workloads, driven by growing availability of data. This trend is likely to boost demand for data-center graphics processing units (GPUs) and custom AI training accelerators over the next 6-12 months. Forty-four percent of surveyed US enterprise CIOs report over half of their data is now ready for AI training, up from 33% six months ago. Additionally, 40% of respondents are currently training foundation models, an increase from 26% six months earlier.

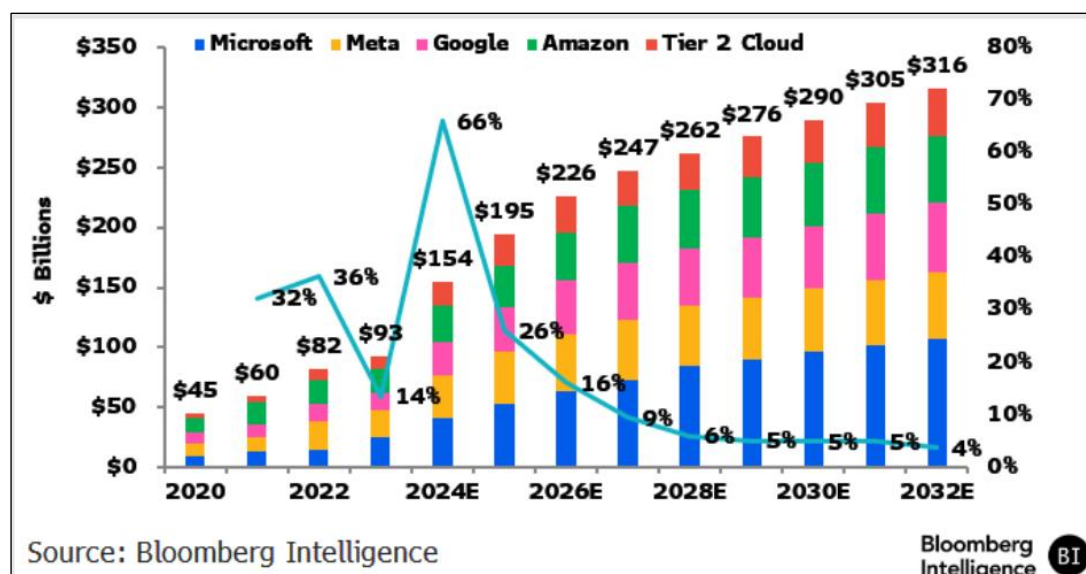
Compared to AI inference tasks, model-training tasks will require high-performance chips deployed in data centers that have stronger computational power, precision and larger memory capacity to handle larger datasets' parallel calculations.

5.4 LLM Training

The continuous scaling of foundational LLMs will be a tailwind for both cloud hyperscalers, which have already seen a multi-billion sales contribution from AI workloads, and companies licensing their LLMs using APIs and broader enterprise use. We expect capex spending on LLM training to remain robust over the medium-term given the high barriers to train a foundational model, which has already driven consolidation among LLM vendors.

Training of LLMs remains the largest portion of the Bloomberg Intelligence Gen AI forecast through 2032. With the five leading foundational model players – OpenAI, Google, Meta, Anthropic and Mistral – focused on scaling the next version of their LLM using the latest GPUs, we expect training costs to expand in proportion with the growth in LLM parameters. While small language model variants such as OpenAI GPT-4o mini and Gemini and Llama lighter version have been released for on-device and edge deployment, the recent release of its chain-of-thought reasoning o1 model from OpenAI suggest the investments in scaling of frontier models is likely to continue in the near to medium-term. Training is expected to reach \$646 billion, while inferencing could be around \$487 billion market with the growth more back-end weighted for the latter, based on Bloomberg Intelligence market sizing model.

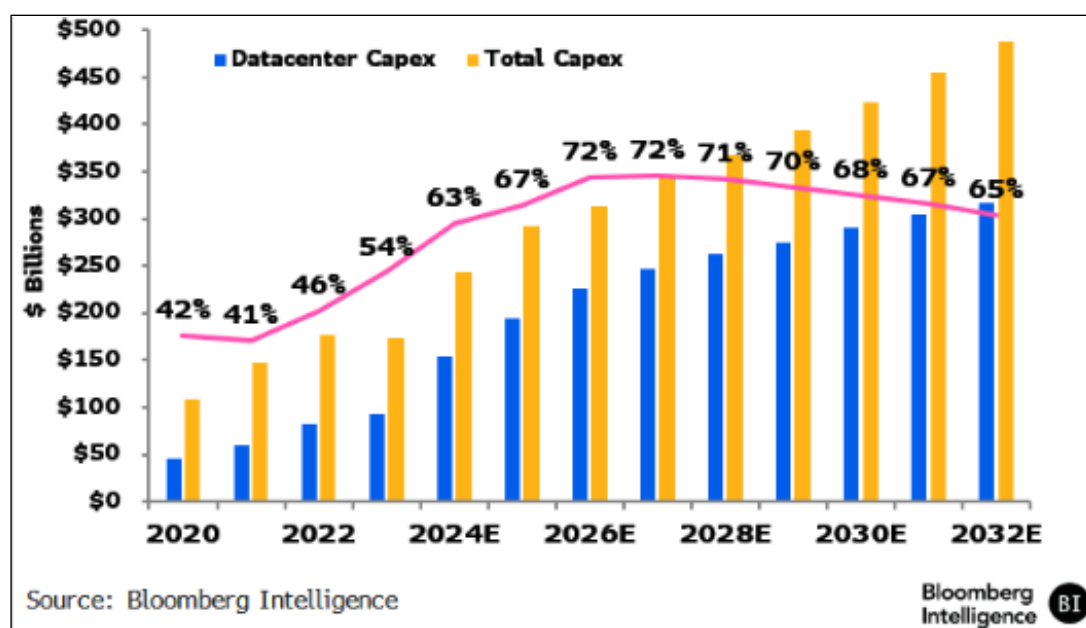
Figure 22: Hyperscaler Capex



AI model-training costs should be similar for the leading LLM players including OpenAI, Anthropic, Meta, Google and Mistral, which all have a multimodal focus. Some of the image and video-specific LLM companies may have lower training costs, though we believe monetizing via APIs will involve leveraging the distribution of hyperscale cloud platforms.

Capex spending on datacenters (see Fig. 23) is likely to stay robust for hyperscale cloud vendors including Microsoft, Amazon, Google and Oracle, which have already seen a multi-billion revenue contribution to their cloud infrastructure business from AI workloads. Microsoft's Azure AI business is already at a run-rate of \$5 billion, about a high single digit contribution to its Azure segment sales of around \$80 billion. Similarly, Amazon and Google have called out multi-billion sales contribution from AI infrastructure. While new capex for datacenter expansion is likely behind the above 50% hyperscale capex growth for 2024, we believe growth may taper gradually than abruptly given the faster architecture releases for GPUs and strong demand for both AI training and inferencing.

Figure 23: Inside Datacenter Spending % of Total Capex



While GPT-4 was trained using a cluster of 25,000-30,000 A100 GPU chips, costing about \$300 million for one training run, the newer models will likely require a bigger-sized cluster proportional to the growth in model parameters. Xai is building a 100,000 H100 cluster, while Nvidia CEO Jensen Huang has painted a vision for connecting one million GPUs in an AI factory. OpenAI expects to spend more than \$200 billion through 2030, based on reporting from The Information, the majority of which will be spent on training its LLM.

SLMs require significantly fewer resources to train and fine-tune than LLMs, which could lessen capacity constraints seen by Microsoft and many hyperscale-cloud providers over the past year. In an example offered by Sebastien Bubeck, former vice president of gen-AI research at Microsoft, Phi-1 needed only 8 GPUs over the course of a week to train vs. an LLM that could require thousands. IBM reiterated this sentiment, stating that SLMs may be roughly 10% of the cost compared with LLMs. SLMs employ synthetically created data in their training sets that facilitate learning without the need for massive amounts of input data.

Lower costs are also likely to trickle down to the end customer. In the case of Khan Academy, Microsoft's latest Phi-3-medium was able to produce similar results to ChatGPT-4 at almost no cost vs. the \$30/month price of the latter.

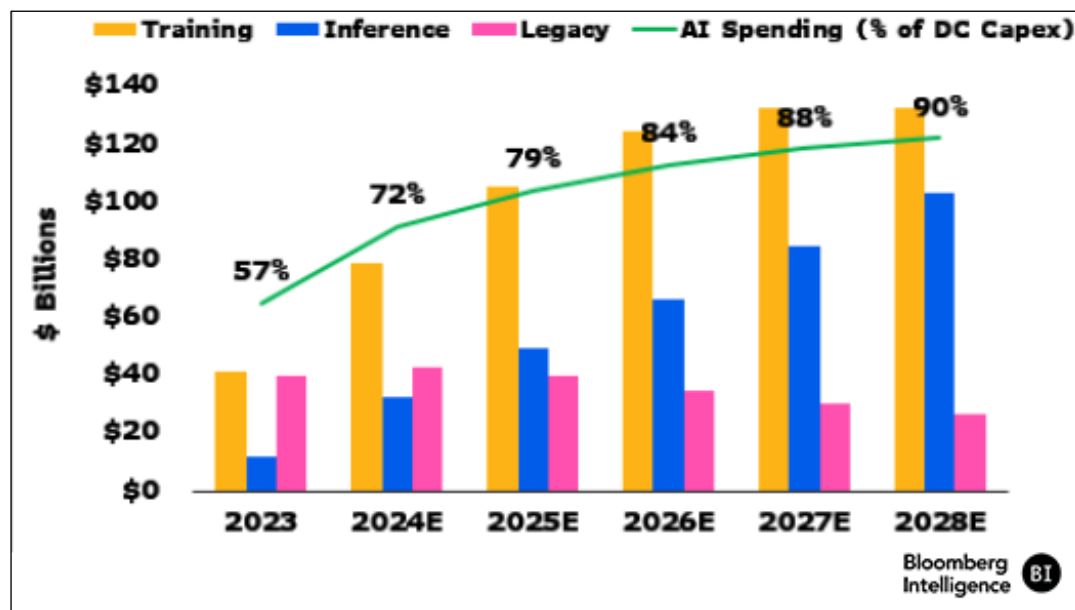
The lightweight nature of SLMs also unlocks a wide range of industry-related use cases that require a foundational model to run on-device or in highly regulated sectors where data can't be sent to cloud servers. Additionally, given the lower requirements for training data, an SLM could be tuned to a far wider array of cases. Among industry verticals, we expect the financial, health-care and public sectors to be the most likely big adopters of SLMs given their unique regulatory constraints and sensitive data. Mobile applications or gaming companies could also benefit from SLMs as they prioritize the lower latency of an on-device foundational model.

Meta is using its growing GPU compute clusters to power the training and deployment of its Llama LLM across its family of apps. Hyperscalers including Microsoft, Amazon, Google and Oracle have seen their cloud segments' growth accelerate, driven by AI-infrastructure workloads.

BI

AI spending can reach 90% of all data-center capex in four years

Figure 24: Training vs Inferencing AI Spending



Source: Bloomberg Intelligence

Recent talent acquisitions by hyperscalers including Google-Character AI, Microsoft-Inflection and Amazon-Adept AI suggest continued consolidation among the foundation model companies. Given the continuous scaling of LLMs and elevated capex requirements for data centers needed to deploy generative AI, we expect very high barriers to entry for companies looking to train their own foundational models. With the supply constraints around the latest GPUs, which have at least 2-3x performance and computation advantage over the prior generation, we believe hyperscalers are unlikely to pull back on data-center spending through 2026, barring a plateauing of LLM parameters or model performance.

Apple maintains the lowest capex intensity among big tech peers, partly due to the fact that it isn't a hyperscale cloud vendor that can leverage its distribution to support multiple LLMs. Capex intensity for Google and Meta could remain around 15-20% for the medium term, as these companies seek to scale the next version of their AI models. Though capex growth is likely to slow gradually over the next five years, a boost in cloud revenue could offset any headwinds to free cash flow. Pure-play gen AI LLM vendors like OpenAI, Anthropic and Mistral are likely to rely on licensing sales while partnering with hyperscale cloud providers to offset pressure from the elevated costs of AI training.

5.5 Inference Efficiency

Inference efficiency will likely become a big focus given the higher compute costs for OpenAI's o1 model, which the company is seeking to mitigate with an o1-mini version. Though the o1 model may offer an improvement in reasoning capabilities, both Meta's Llama and Mistral's latest large

language model have incorporated multimodal reasoning and offer cost advantages for inferencing vs. OpenAI's models. Given the difference in pricing for different versions of the frontier LLMs, we believe companies are likely to mix LLMs from different providers depending on the type of query and to implement checks and guardrails by cross referencing each other LLMs.

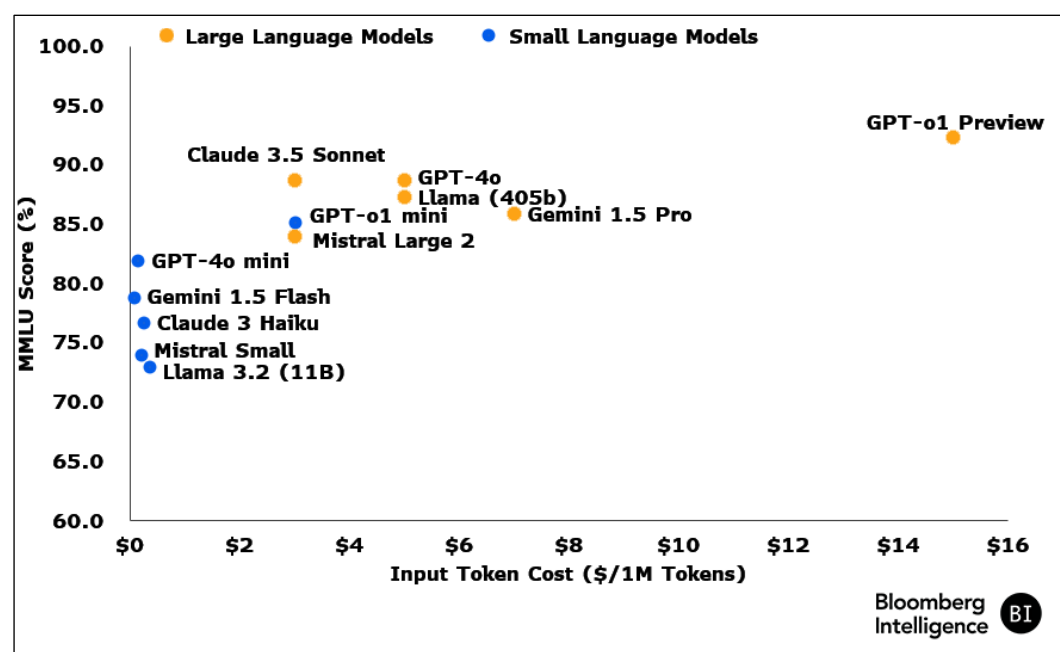
OpenAI's release of its o1 model may further aid adoption of open-source models like Meta Llama and Mistral as enterprises increasingly focus on privacy, cost and fine-tuning LLMs with their own data. Given the likely convergence in model functionality between OpenAI GPT, Google Gemini, Meta Llama, Anthropic Claude and Mistral, we believe the lowest-cost model provider will be based on open source. OpenAI's o1 models performed better on certain benchmarks such as reasoning and math, suggesting different models may be employed depending on the query type.

Meta also disclosed its Llama model family has reached 350 million model weight downloads on Hugging Face, with 20 million in just the past month. Access to Llama models via API at cloud partners grew 10x from January to July.

A pivot to smaller models with fewer parameters that can be used for specific tasks vs LLMs is likely to drive a secular shift in applications to agent functionality. OpenAI released its "chain-of-reasoning" o1 model, along with an o1-mini version, to showcase improvements in reasoning capabilities. Given the difference in pricing across different versions of LLMs, we believe companies are likely to mix LLMs from different providers depending on the query.

A small, low latency AI model (5-10B parameters) will be included in iOS18 as part of the Apple Intelligence framework, that will be able to understand user commands, the current screen, and take actions on apps. It can handle tasks like summarization, as well as powering the "AI agent" features of Siri including user commands that require utilizing multiple apps.

Figure 25: LLM Costs vs Performance



Source: Bloomberg Intelligence

The mid-tier Sonnet model of Anthropic's Claude (AI assistant) showed a notable performance improvement while lowering input and output token costs to below that of GPT-4o, suggesting competition is likely to increase around reducing AI-inferencing costs. This could give an advantage to hyperscalers such as Google, with its Vertex AI -- where it offers both its Gemini model and other foundational models through the Vertex AI platform. Amazon Bedrock also offers Anthropic's Claude models for inferencing workloads on its platform.

As LLMs grow exponentially larger, the number of floating point operations (FLOPs) generally scale with the parameter counts demanding significantly greater computational resources. To offset rising LLM costs, foundational model companies will seek to shrink the size of trained models for lowering inferencing costs for broader use cases across enterprise and consumer applications.

Most foundational model providers including OpenAI GPT, Anthropic Claude, Google Gemini, Meta Llama and Mistral have released smaller versions using quantization and distillation for edge use cases to run LLMs natively on PCs and smartphone devices. We believe shrinking the parameter size of models will become more important amid continuous scaling of datasets and tokens used in the transformer architecture that underpins most foundational LLMs.

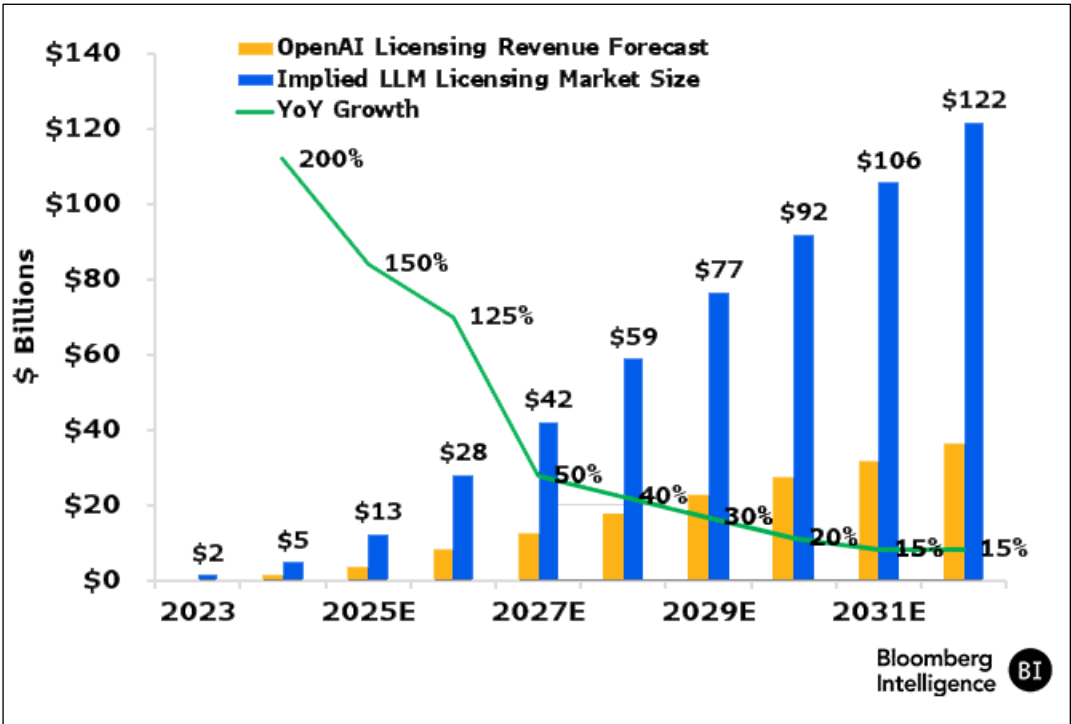
5.6 Open Source / Commoditization of LLMs

With most model providers having access to a similar corpus of web data, it will be hard for them to maintain differentiation based on dataset and compute. Model parameters are likely getting better with every new version, and inference costs could come down over time. OpenAI's o1-mini and o1 models are aimed at addressing different types of use cases involving chain-of-thought reasoning, though higher inference costs can be a deterrent.

Given the likely convergence in model functionality between OpenAI GPT, Google Gemini, Meta Llama, Anthropic Claude and Mistral, we believe the lowest-cost model provider will be based on open source. OpenAI's o1 models performed better on certain benchmarks such as reasoning and math, suggesting different models may be employed depending on the query type. Both Meta Llama and Mistral's latest open-source Pixtral 12B parameter model have closed the gap in multimodal functionality vs. closed-sourced LLMs like OpenAI's.

OpenAI recently raised \$6.6 billion, implying a \$157 billion valuation, putting a focus on the company's monetization as pure-play generative AI player and licensing sales for its foundational LLM. Though training costs are relatively in the same range for most foundational AI model companies, greater monetization of OpenAI's technology and subscription products suggests room for Meta and Google to bolster their licensing API revenue, given their distribution advantage as they narrow the functionality gap vs. OpenAI.

Figure 26: OpenAI Revenue Expectations



Source: Bloomberg Intelligence

OpenAI's API revenue of \$1 billion (30% of total sales) has been driven by its leading large language models and close integration with Microsoft Azure AI. OpenAI generates API revenue from both enterprise and individual application programming interface calls, as well as consumption through Azure AI. This is in contrast to Meta, which has open sourced its Llama foundational model and requires a license fee only for scale deployments. A narrowing of the functionality gap between the latest version of LLMs creates an opportunity for Meta and Google to similarly monetize their models via API and licensing revenue.

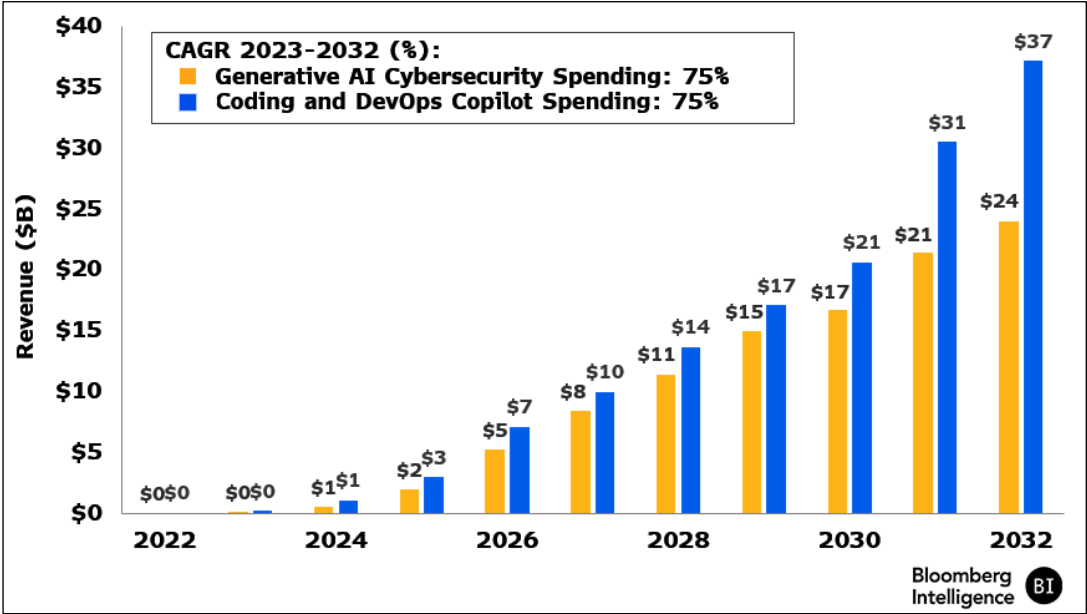
Hyperscaler distribution has so far aided Anthropic's API sales, at about a \$500 million run rate, supported by the company's alliances with Amazon and Google. In contrast, OpenAI's licensing revenue from Azure may be about \$200-\$300 million, a mid-single-digit percentage of

Microsoft's Azure AI workload sales. A lack of multi-cloud options beyond Microsoft Azure could limit OpenAI's indirect API sales, whereas other foundational models might see faster growth. Both Meta and Google have invested aggressively in building their own generative-AI infrastructures for model training and deployment. We expect API usage and other consumption-based contracts will be a big driver of LLM licensing sales in the near-to-medium term.

5.7 Gen-AI Coding Copilot Adoption Likely Above Other Uses

Coding and DevOps copilots may be one of the largest areas of generative-AI customer spending, and we anticipate adoption to be higher in these products than other gen-AI tools. These possess the ability to significantly shorten the time to a return on investment for customers, as they help increase a software engineer's productivity and ease pressures companies face from the global developer shortage. Tools such as Microsoft's GitHub Copilot and Amazon's Q Developer have claimed over 40% greater coding efficiency. We expect that, even if half of that metric were realized by companies, it would be enough to drive outsized adoption.

Figure 27: Generative AI Cybersecurity, DevOps




Source: BI's forecasts based on hardware and software data from IDC

GitHub is the clear leader at well over 100 million developers, and is uniquely positioned as the biggest code repository for developers irrespective of the cloud provider, with annual recurring revenue potentially reaching \$1.5-\$2 billion by the end of 2025. Amazon's Q Developer and Google's Gemini Code Assist are also likely to expand within their developer bases, while growing integrations with development environments like Visual Studio could serve to broaden their compatibility and reach.

Figure 28: Copilot Offerings

	GitHub Copilot Business	Amazon Q Developer
Price	\$19/Month/User	\$19/Month/User
Productivity Boost	55% Faster Coding	Up to 40% Increased Productivity
Features	<ul style="list-style-type: none">• Code Completions• IDE/Mobile Chat• CLI Assistance• Security Vulnerability Filter• Code Referencing• Public Code Filter• IP Indemnity• Enterprise-grade Security	<ul style="list-style-type: none">• Code Completions• Support for IDEs, Command Line• Optimized for AWS Services• AI Powered Code Remediation• Reference Tracker• Bias Avoidance

Source: Bloomberg Intelligence

Bloomberg Intelligence 

Greater adoption of these coding copilots -- which we expect to be more than 75% for GitHub and Q Developer over the long run -- could also accelerate the modernization of on-premise applications to the cloud, thereby also benefiting hyperscale cloud providers. An internal Amazon study noted that Q Developer upgraded over 1,000 Java applications in two days, a process that could take a team of developers roughly a year to complete, according to AWS Vice President Deepak Singh, who was featured on BI's Tech Disruptors podcast.

We anticipate that more than half of all cloud applications were brought over from on-premise without optimization, a process known as "lift and shift," which may also accelerate developer copilot sales as teams look to modernize these apps.

5.8 Data Comes at a Premium

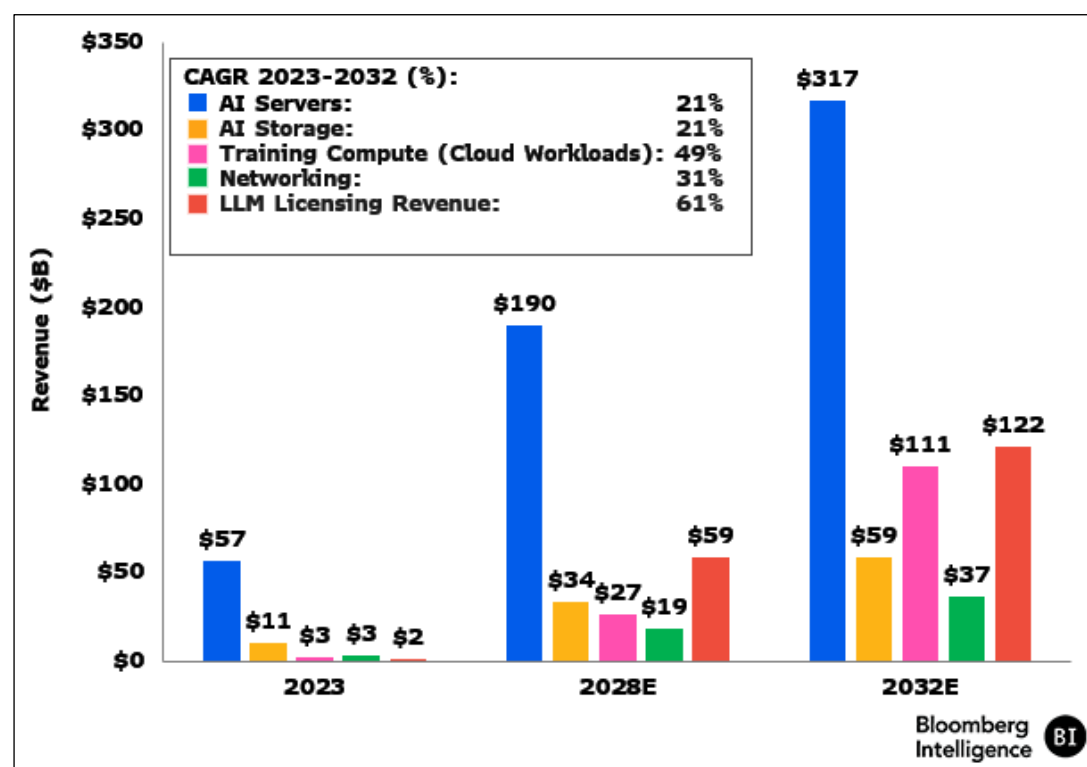
The size and complexity of large language models makes the training process extraordinarily data intensive. Though OpenAI's ChatGPT reached a partnership with Microsoft, it still could be disadvantaged compared with internet enterprises in the volume of training data available.

ChatGPT's initial application was mainly focused on changing the nature of search, which has been dominated by Alphabet's Google. ChatGPT's primary use was to analyze, generate and edit text based on user input. Yet within just a few months, OpenAI realized how powerful generative AI can be, and the platform's scope quickly expanded beyond traditional search. The latest version of ChatGPT can handle data including images, audio and video. Such inputs require vastly more computing resources than text-based LLMs.

The size and complexity of LLMs based on transformers architecture are likely to grow as a result of multimodal input, which can help hyperscale companies including Microsoft-OpenAI, Meta, Google and Amazon to maintain their lead over other foundational LLMs from peers. Given training of LLMs is a recurring process, we expect foundational LLM companies will consolidate to the ones that have a hyperscale infrastructure and large amounts of first-party data to improve the accuracy of the models.

Retrieval-augmented generation (RAG) techniques will likely become key for developing enterprise chatbots. This approach features RAG model that is trained on proprietary enterprise data, adding an element of customization. When a user inputs a query, the RAG model can identify relevant information from the enterprise database and append it to the user query, which then gets passed on to the LLM. The additional context provided by the RAG model helps reduce irrelevant answers and hallucinations from the LLM. Given the recent scrutiny around Alphabet Gemini's image generator, we anticipate hyperscalers will have a heightened focus on reducing historical inaccuracies and questionable responses. As a result, these companies could be first in line to develop a RAG counterpart to their existing models.

Figure 29: Training Forecasts, 2023-32



Source: BI's forecasts based on hardware and software data from IDC

5.9 AI Fuels a Fifth of Global Server Revenue

Robust gains in the number of ChatGPT active users indicate that generative AI could be among the most important catalysts to growth for the server supply chain in coming years, driving over 20% of global server revenue by 2024 from 15% in 2021, by our calculations.

After its November launch, OpenAI's ChatGPT amassed a base of 1 million users in a week and exceeded 100 million in just two months. OpenAI introduced a subscription service at \$20 a month and offered businesses paid access to ChatGPT to expand commercial applications. Companies including Snap, Shopify and Instacart already have integrated ChatGPT into their products.

The server supply chain's original design manufacturers could reap the most demand, since cloud service providers are integrally involved in AI development. AI servers could also drive robust sales for other suppliers with design expertise.

5.10 IT, Business Services To Benefit in Long-Run

Fine-tuning of large language models, application modernization and data-related services are a few ways that IT services companies can leverage increased generative-AI spending. We anticipate consulting-heavy firms like Accenture, IBM and Capgemini to see more of the outlay than offshore providers like Infosys and Wipro.

The services industry will likely play a major part in fine-tuning of large language models (LLMs), as end markets look to maximize the benefits provided by copilots. We see this work as akin to customization of packaged software, which has been a key backbone of the industry for the past 20-30 years. Data-related services is another area that can play a big role, which includes data aggregation, cleanup, creating common data lakes and using that final information to create custom copilots for clients.

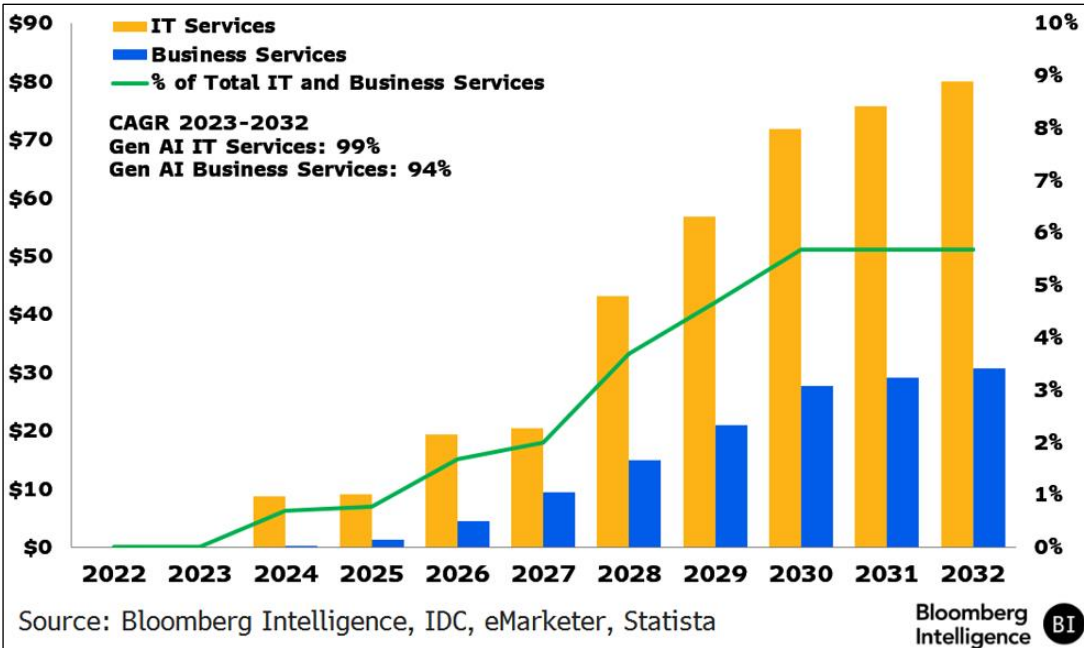
Application modernization could also see increased use of gen-AI technology. The IT services industry will likely lag behind AI infrastructure and software, like the trends observed in past emerging technologies. Our analysis indicates meaningful growth starting in 2026 onward.

The total market for IT and business services is valued at \$1.2 trillion and could reach \$1.9 trillion by 2032, assuming 5% yearly growth. This implies generative AI could eventually represent 5.7% of the market by then, which we think is a conservative estimate.

BI

Gen AI IT services seen growing by 99% compounded annually through 2032

Figure 30: Generative AI for IT & Business Services (\$ Billion)



The shift to cloud and embracing more digital workflows is another area which will likely see increased adoption because of gen AI, especially as majority of the IT spending still resides on-premise, where data is siloed across multiple systems. The recent success of Salesforce's Data Cloud product highlights the need for cleaner data to train LLMs. Cloud is likely to be another essential precursor to AI workloads, which could result in an accelerating shift from on-premise architectures. We see the services industry playing a big role in this effort.

Accenture, IBM and Capgemini are well-positioned to capture more generative-AI projects than the India-based IT services companies due to their larger consulting units. All three have invested in emerging technologies, with Accenture committing \$3 billion to expand its data and AI practice, Capgemini making a similar €2 billion investment and IBM enhancing its software portfolio, which could drive greater work for its consulting unit. These investments position the companies well for when corporate IT spending improves, especially compared with peers with less robust gen-AI offerings.

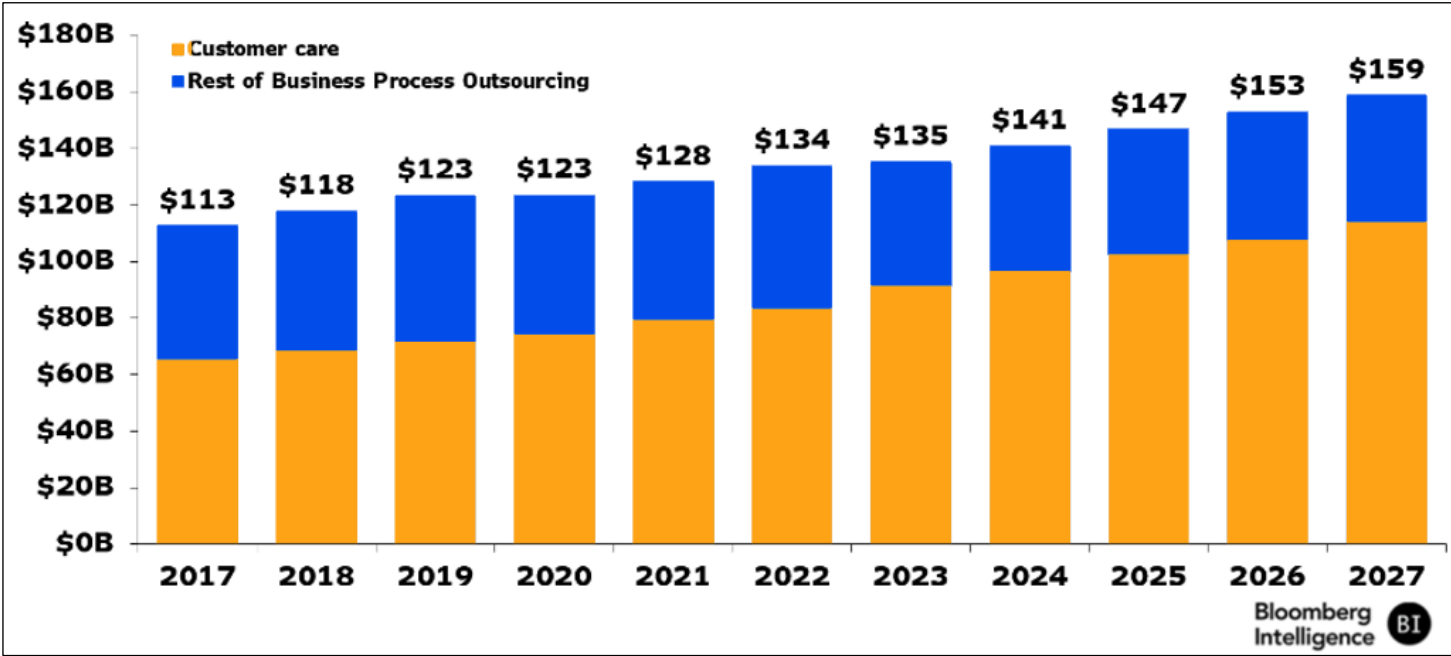
Services spending could increase about \$113 billion over the decade, doubling annually from 2023. Consulting, data-related services, custom application development and creation of new chatbots would drive the additional expenditures. The current total market for IT and business services combined is roughly \$1.2 trillion, which may reach \$2.1 trillion in 10 years, assuming 6% annual growth.

Among IT services peers, Accenture stands out given its propensity to invest ahead of the curve in emerging technologies. In 2023, the company announced a three-year \$3 billion investment in generative AI that would include training 80,000 workers on AI as well as offering greater solutions and models. We calculate generative AI will contribute roughly \$1.5-\$2 billion to Accenture's revenue by 2025, after booking \$450 million alone in 1Q. Outside of Accenture, practices such as those from IBM, Infosys and Tata Consultancy Services will also likely benefit.

Business-process outsourcing services (see Fig. 31) may be more heavily disrupted than IT services, with jobs in customer service and back-office areas displaced by AI assistants. That could lead to near-term pricing pressures, particularly in customer care, which sits at the bottom of the BPO value chain yet is still the largest and fastest growing of its segments. Customer care is forecast to expand 6.5% annually through 2027, compared with 3.5% for all other BPO services, according to IDC.

Generative AI will be more of a revenue tailwind for BPO companies that have minimal exposure to customer-care services, like Genpact and EXL Service. Companies that specialize in greater value-added services might leverage efficiency gains from AI to expand their total addressable markets, particularly in areas like data analytics.

Figure 31: Business Process Outsourcing Forecasts



Source: BI's forecasts based on hardware and software data from IDC

Section 6. Expanding Uses

Enterprise Gen-AI Use Cases Rising, Adding Productivity

Generative AI has the potential to boost productivity across multiple areas, including software development, customer service, operations and security. As language models evolve, costs are likely to drop and enterprise use cases expand. Though much of the quoted AI success is from companies developing these tools, more customer testimonials might drive greater adoption.

Figure 32: Generative AI Use Case Tracker

End Market Impacted	Quote	Company	Area of Company
Software Development	With Q's code transformation capabilities, Amazon has migrated over 30,000 Java JDK applications in a few months, saving the company \$260 million and 4,500 developer years compared to what it would have otherwise cost.	Amazon	Back End Operations
Software Development	In our engineering organization, our developers now save more than 20,000 hours of coding each month through the use of our AI tools. These innovations are helping us drive continued efficiencies across the business and accelerate our product road maps.	Salesforce	Back End Operations
Software Development	We've done an awful lot to digitize many parts of our business and we're now applying Gen AI to it. The places that we're seeing tremendous -- tremendous results on are coding. We need 70% less coders from third parties to code as the AI handles most of the coding, the human only needs to look at the final 30% to validate it. That's a big savings for the company moving forward.	BP	Back End Operations
Security	implementing and managing a scalable platform powered by Gen AI, enabling the agency to act on evolving cyber threats up to 60% quicker than with traditional technologies, including detection, response, and containment models.	Accenture	Back End Operations
Security	Initial modelling shows AI enhancements boost fraud detection rates on average by 20% and as high as 300% in some instances.	Mastercard	Back End Operations
Sales	AI-powered personalization can increase revenue per passenger by 10 to 15%	Microsoft	Front End Operations
Sales	Typically our sales cycle was 90 days and now with these experiences that HubSpot AI has helped us to build that, has gone down to 30 days to 45 days. We had people spending a minute or less on our experiences and our pages, now they are spending 3 minutes to 4 minutes.	HubSpot	Front End Operations
Sales	81% of sales teams are either experimenting with or have fully implemented AI. The results speak for themselves: 83% of sales teams with AI saw revenue growth this year vs. 66% without AI. Sales Teams Using AI 1.3x More Likely to See Revenue Increase	Salesforce	Front End Operations
Operations	Nurses are spending, on average, 20 minutes an hour on admin tasks that we now with generative AI can help reduce to five minutes.	Koninklijke Philips	Back End Operations
Operations	While on the Lexis Nexus, what they're able to do on They built an AI platform, not just for conversational search, but also for summarization and intelligent legal drafting, which enables lawyers to be more efficient and be productive. They report their legal professionals are saving up to 11 hours per week using their AI platforms	LexisNexis	Back End Operations
Operations	AI has helped us in our fight against financial crime by reducing the processing time required to analyze billions of transactions across millions of accounts from several weeks to a few days.	HSBC	Back End Operations
Marketing and Advertising	Ad campaigns using Meta's generative AI ad features resulted in 7.6% higher conversion rate...More than 1 million advertisers and 15 million ads created with our generative AI ad tools, in the last month	Meta	Front End Operations
Marketing and Advertising	We use AI and tech to boost our agility on advertising and promotional expenses. You're familiar with this BetIQ program that we have started to help us allocate with the best ROI all our A&P which gives a good improvement of 10% to 15% to the productivity of our A&P and we're just beginning to roll it out. It should reach around 60% of our A&P at the end of this year.	L'Oreal	Back End Operations
Customer Service	They recently rolled out service now assist, generative AI product for their customer service piece. And they are now solving customer queries 55% faster because they're using now assist to help with things like case summaries, complex case notes.	BT Group	Front End Operations
Customer Service	Phone Pe leveraged Freshworks customer service platform to automate common service inquiries. Resolving 80% of inquiries without human intervention	Phone Pe	Front End Operations
Customer Service	Ada helped BlueJeans serve 72% of its Chat Users with a Chatbot vs. Live Agent. 75% of support contacts engage with chatbot. 83% reduction in live chats from trial end users	BlueJeans	Front End Operations

Source: Bloomberg Intelligence

6.1 Productivity, Costs, Efficiency All Boosted

AI offers significant productivity benefits in software development, and coding is one area where we see generative-AI technology producing benefits for users (Fig. 33 shows some uses). Amazon Web Services reduced Java application time to a few hours from days with its AI-coding tool Amazon Q Developer. Similarly, Palo Alto reported a 30-40% productivity boost for developers using Copilot tools vs. those without. Companies like BP have seen substantial cost savings, needing 70% fewer third-party coders due to AI handling much of the work.

Customer service is another area well-suited for generative AI, with potential for greater adoption across call centers and support teams, as it improves efficiency and shortens response times. Ada helped BlueJeans automate 72% of its chat interactions using chatbots, resulting in an 83% decrease in live chats during a trial period, and BT Group saw a 55% faster resolution of customer queries after implementing ServiceNow's Now Assist, an AI tool that streamlines tasks like case summaries and handling complex notes. Discover also found that AI lets human agents retrieve information 70% faster than traditional search tools, significantly boosting agent productivity and customer satisfaction.

Generative AI is also having a significant effect on operations and back-office functions, driving productivity and reducing time spent on routine tasks. Take the example of Royal Philips, which found that nurses who typically spend 20 minutes each hour on administrative duties, can cut that to just five minutes with AI. Germany's largest health insurer uses AI to process claims by reviewing 800 policy documents, reducing the time required to just three seconds from 23-30 minutes, according to Google.

Similarly, Mastercard reported that after only 28 days of using Microsoft Copilot, it had saved a total of 1,200 hours, or about four hours per employee, in three specific use cases.

AI is making significant strides in improving product capabilities and bolstering security measures across various sectors. HSBC has leveraged AI to drastically cut down the time required to analyze billions of transactions across millions of accounts, reducing the process from several weeks to just a few days, while Mastercard has reported that its early AI implementations have boosted fraud detection by an average of 20%, with some instances showing improvements of up to 300%.

These developments highlight how AI enhances the efficiency and effectiveness of financial services, letting companies proactively combat threats.

Figure 33: Companies Using Generative AI

Examples on how companies are leveraging AI	
Kroger	Deliver the appropriate promotional offers and discounts to customers at the right time
Target	Power product detail pages and provide more friendly and relevant explanations
T-Mobile	Spends around \$2.5 billion on advertising annually, using AI to place ads and optimize media spend
eBay	Enhance sellers' product images to make them more compelling to customers
State Street Corp	Digitize and automate approximately 85% of bank loans settlements
Edison International	Improve inspections, customer experience, and grid planning; Also using AI for research, workflow automations, and code development
Hershey	Optimize the talent profile for certain key jobs
McKesson Corp	Improve patient intake and workflow, system productivity, and several supply chain use cases, including disruption predictions, forecast accuracy algorithms and fraud detection

Bloomberg Intelligence **BI**

Source: Bloomberg Intelligence

6.2 Base Capabilities Cover Broad Range of Uses

Momentum with gen AI agents is set to accelerate amid successful proof of concepts and ideation across different types of consumer and business apps. Real-time language translation has helped boost engagement and satisfaction, and LLM-based recommendations will be connected to enterprise databases to automate certain business processes across websites, mobile apps and any other touch points. Newsweek is already seeing faster content discovery with its LLM-based search, with longer context windows than what was offered by traditional search.

We expect a broader change in user interface for all types of applications and software as free-form search is likely to play a bigger role replacing the traditional menu- and directory-based navigation. In addition, the use of AI for content ranking and recommendation is likely to drive more personalization across most platforms.

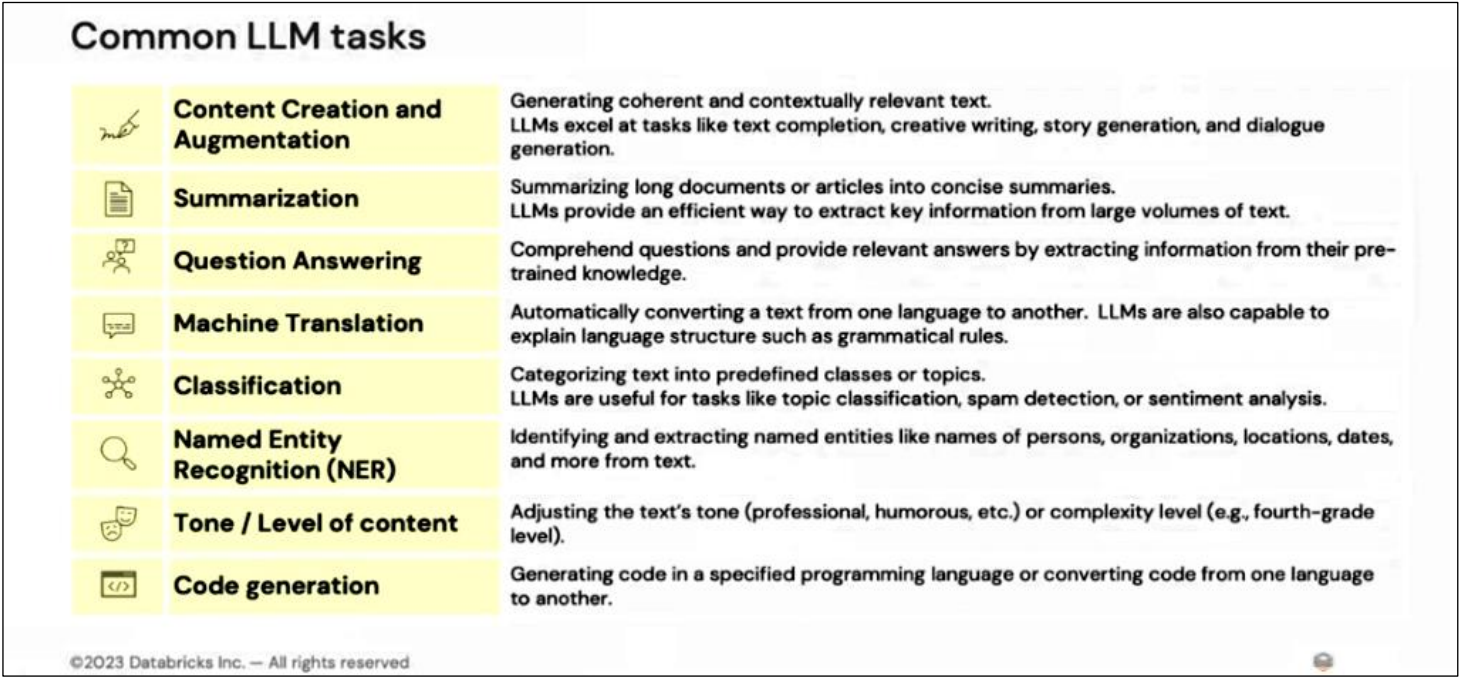
Developer coding copilots, text summarization and customer-service chatbots are among the most prominent use cases for generative AI (Fig. 34), and generating marketing content for websites and presentations with LLM-based prompts promises to be a big productivity driver. The biggest proofpoint is in the ad-targeting improvements that Meta has seen across its advertiser base, driven by its Advantage+ suite that leverages generative AI. Others have done the same, such as Alphabetith its Performance Max offering. Marketing copilots are already deployed at a number of enterprises for drafting documents or generating images based on prompts, which will likely help drive further automation for targeted promotions. Increased model

efficiency and hardware improvements could enable more use cases on the consumer side, with apps likely to deploy user interface changes based on voice and gen-AI assistants to bolster engagement.

Amazon Web Services, Microsoft, Google and Oracle could win increased orders for cloud services, in addition to hybrid cloud providers such as IBM, for which the main source of traction would be in its Red Hat, security services and Watson-related products. Oracle can parlay its leading market share in database-management products, while Cisco, Databricks, Snowflake, VMware and ServiceNow also stand out. The leading cybersecurity players, including CrowdStrike and Microsoft, have launched security copilots.

Similarly, design and gaming companies including Adobe, Unity, Roblox and others are integrating AI into their software to fend off competition from startups that use LLMs.

Figure 34: LLM Uses



Source: Bloomberg Intelligence

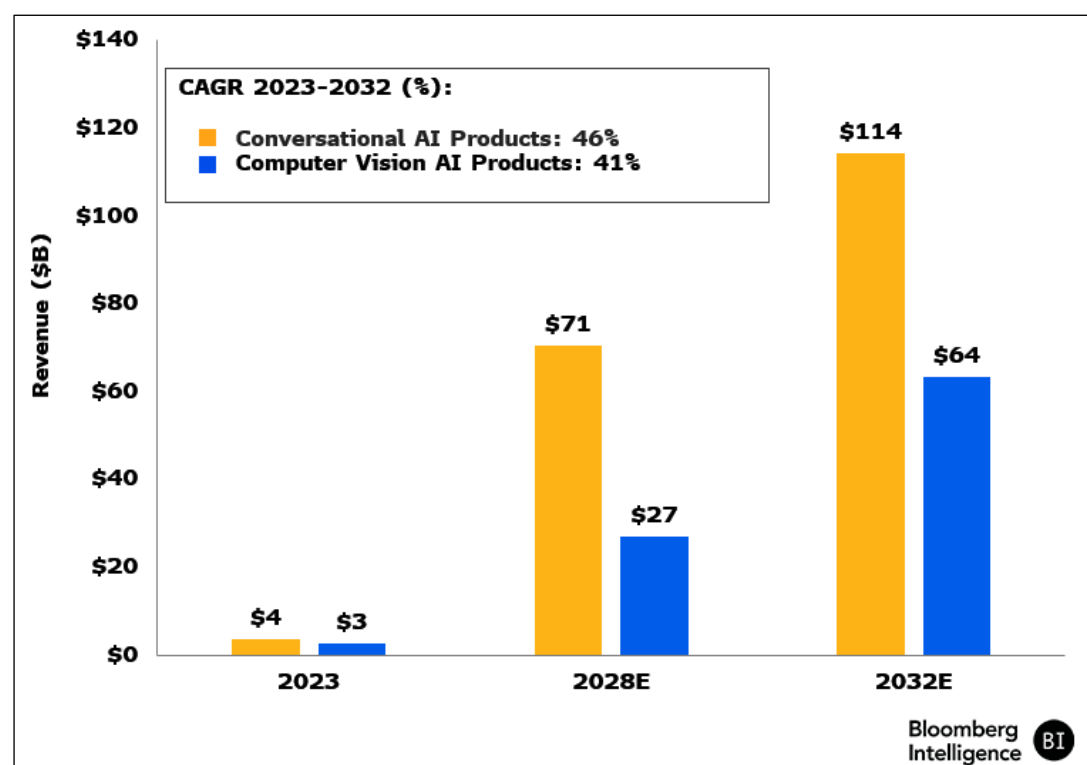
6.3 Apple, Amazon in the Mix as AI-Enabled Tools Expand

Voice assistants based on conversational AI and computer-vision products may emerge as new categories for inference, given the availability of large language models (LLMs) for domain-specific predictions. Apple, Samsung, Amazon and others may look to conversational AI given how well their existing product offerings mesh with the category. The launch of more compact versions of existing LLMs bodes well for consumer devices given the reduced computational intensity required to run AI workloads. Auto manufacturers like Tesla and GM could invest in computer vision research to drive the next generation of AI in vehicles. Advancements in

generative AI and more accurate responses for recently trained LLMs have set the stage for such categories to accelerate the overall \$1 trillion devices market, where smart speakers and wearables already are large categories. Mainstream adoption of generative AI could drive a faster refresh cycle for personal computers and smartphones as new versions of these edge devices may be optimized to run generative AI apps natively, likely leveraging a smaller LLM given the processing, memory, and storage requirements. Adoption in Asia, albeit gradual, is being led by technology giants there. In one instance, Alibaba's open-sourced SeaLLM, despite having fewer parameters than OpenAI's ChatGPT-3.5, can process text in Southeast Asia faster and with more accuracy than the latter, thanks to a custom-built training dataset tailored for the region's diverse language profiles and cultural norms.

Retail is ahead of most sectors in AI investment, representing about 13% of all such spending in 2023, Statista data show, adding capabilities that can aid revenue and margin. We believe that generative-AI spending will increase as retail leans further into driving innovation and speed across search, marketing, product design, supply chains, customer service, seller tools and virtual-shopping assistants. Companies that use AI recorded a 700-bp increase in sales in 2022-23, doubling their profit pace, according to Statista, and its growing use is aimed at boosting conversion -- about 2.5-3% on average, according to Shopify -- and improving productivity and margins. About two-thirds of retailers aim to add Gen-AI features for their customers, according to eMarketer.

Figure 35: Inference Forecasts 2022-32



Source: BI's forecasts based on hardware and software data from IDC

6.4 Getting Personal to Drive Sales

Continued integration of AI and notably generative AI can further strengthen a brand's appeal as it improves marketing messages to better meet and match customer expectations. Tools like augmented reality and virtual reality, tailored recommendations and personalization can all boost conversion, while other companies are using generative AI to quickly formulate marketing campaigns.

Offering AI-driven recommendations and similar-item features can boost conversion as it helps consumers find what they're seeking, and from Gen Z to Baby Boomers (see Fig. 36), consumers are comfortable with buying based on AI-based recommendations, with about two-thirds becoming open to shopping AI-generated suggestions, Statista data show.

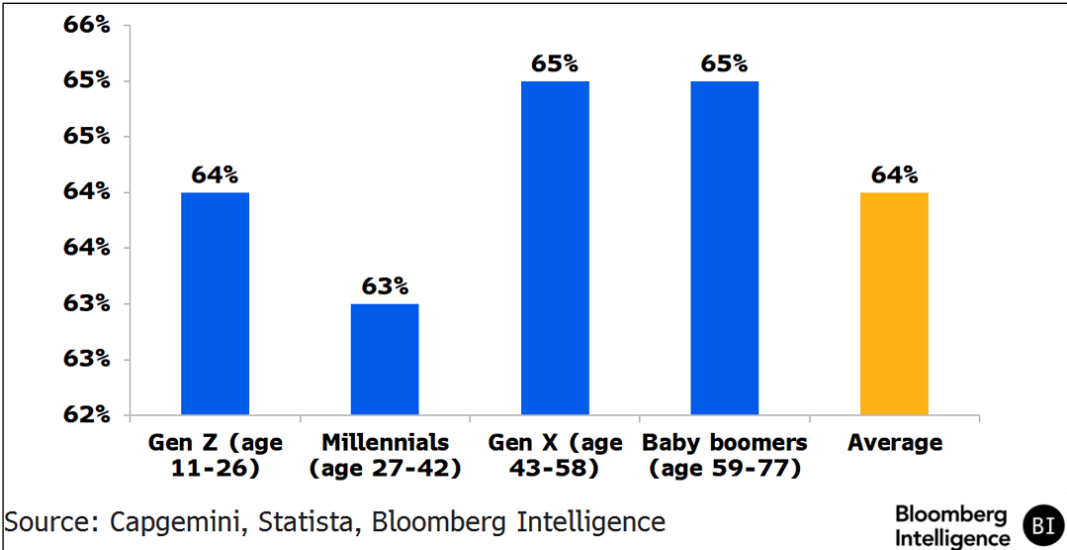
Wayfair's Decorify leverages augmented and virtual reality to recommend items for a specific space and visualize items there, while Warby Parker uses AI to suggest frames based on a shopper's preferences and can show how they'll look on their face. Sephora and Ulta both use color match and virtual try-on technology to help shoppers find the right colors for their skin tones.

Amplifying the allure of product photographs with AI is becoming more common across retail, as it helps boost conversion and save costs. Several online marketplaces are using design features to enhance shopping experiences. Amazon, eBay, Wayfair and Google, among others, have layered in product-imagery tools to save time and money, and boost aesthetics. Revolve has used gen-AI to create its billboard campaign "Best Trip" as well as its capsule collection, which could pare marketing and design costs. Puma is also using AI to create personalized product images, such as customized backgrounds, to improve its marketing campaigns, while Under Armour used AI to create an inspirational "team talk" speech to encourage athletic activity and draw attention to the brand.

BI

Almost two-thirds of consumers say they're willing to use AI suggestions to buy products

Figure 36: % of Consumers Open to Buying AI-Suggested Product



6.5 Conversational AI to Expand Voice-Assistant Category

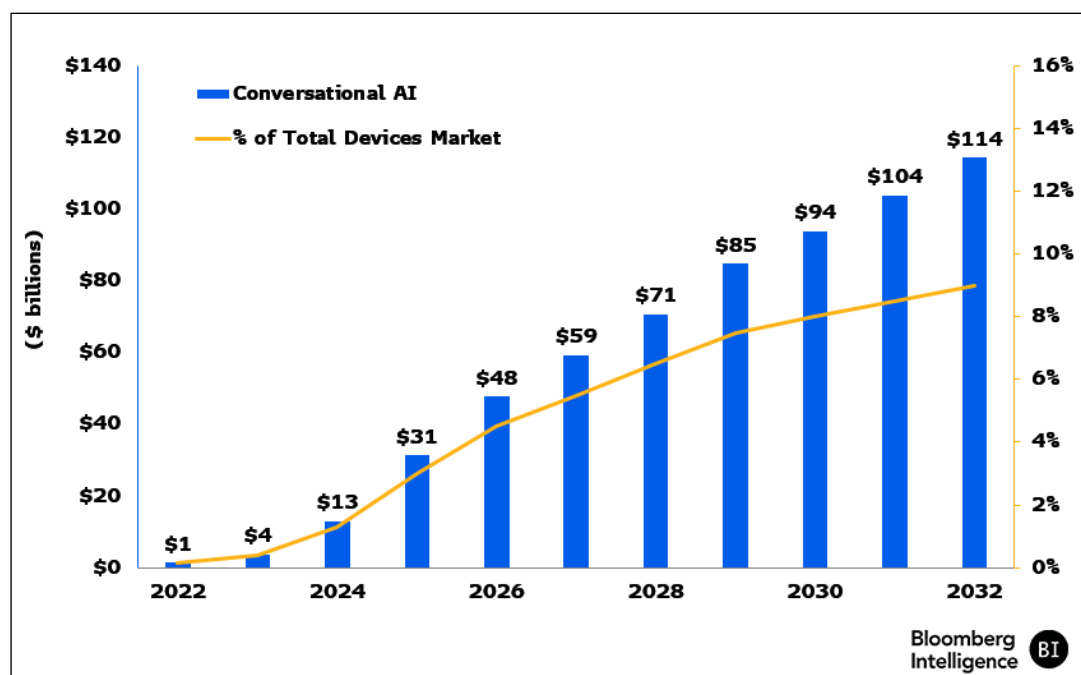
Conversational AI products from hardware makers like Apple and Samsung will likely be tethered to PCs and smartphones, helping to boost upgrades of existing installed bases while driving growth for services. Suppliers like Apple (HomePod), Google (Home) and Amazon (Echo) may enhance their device speakers with assistants, while carmakers including Tesla, BMW, Ford and Volkswagen could incorporate them to boost driver engagement. Conversational AI is much more popular with consumers than generative AI for copilots, according to a recent Bloomberg Intelligence survey, with over 40% of respondents citing frequent use of AI tools for conversational interfaces. We expect these products to grow at roughly a 54% compound annual rate through 2032, in line with the overall generative AI market. Most of the gains will probably come in the latter half of the period as the product category becomes more established.

Conversational generative-AI products could grow at a compound annual rate of 46% to \$110 billion in revenue by 2032. Similar to the Google Assistant on Pixel 8, Apple previewed a revamped Siri powered by Apple Intelligence at its developers conference in June. It also partnered with OpenAI to leverage GPT for more advanced AI capabilities. Apple HomePod, Google Nest and Amazon Alexa could also see similar enhancements to their built-in conversational assistants.

Conversational AI will likely affect many markets. Automakers might implement navigational assistants and companies could use it for HR training and health-care providers may design virtual assistants for patients.

Computer vision also could become a significant application of generative AI tools. Building LLMs will require large amounts of training data and then need generative AI for deployment in automobiles to run inference functions. We expect most incremental revenue from computer vision to come from hardware, with the category expanding to about \$60 billion by 2032, thanks to its application in advanced driver-assistance systems. There may be an even larger impact on related service sales in the medium-to-long term.

Figure 37: Conversational AI



Source: BI's forecasts based on hardware and software data from IDC

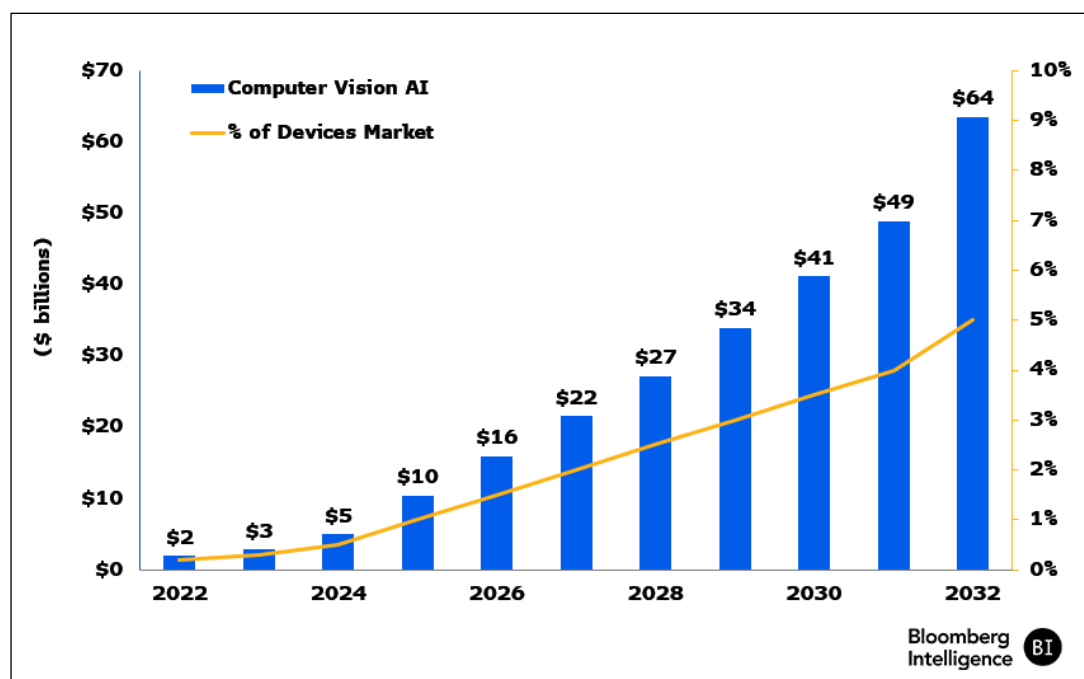
6.6 Better Computer Vision May Boost Autonomous Driving

Computer vision is another major application of generative AI and could reach \$58 billion in revenue by 2032. Autonomous driving may be a catalyst for computer-vision technology, particularly as Tesla aims to launch its robotaxi later this year. We expect multimodal AI's image- and video-processing capabilities to improve, which can be useful for applications like autonomous driving and medical scanning, given the need for accuracy.

Virtual reality (VR) might also benefit from improved computer vision, though VR-headset adoption has struggled amid discomfort with the form and pricing. Augmented reality could see better traction, aided by generative AI.

An AI training infrastructure will be essential to run these heavy workloads, sparking demand for high-capacity servers and storage. Most training-related workloads will be new since enterprises currently use general-purpose CPUs for analytics and transactions.

Figure 38: Computer Vision AI



Source: BI's forecasts based on hardware and software data from IDC

Generative AI can also improve the seller experience on marketplaces, with eBay, Amazon and Depop among those adding features that speed the listing process. AI tools are also making image listings cleaner, which can improve conversion, and AI authentication can also improve a platform's inventory.

Generative AI can help sellers on e-commerce marketplaces, improving the quality of item descriptions and aiding inventory, as seen by eBay's magical listing feature, which uses generative AI to write item descriptions, titles and categories in five markets, and a more than 90% acceptance rate. eBay also has AI-powered apparel authentication through its Certilogo acquisition, which can enhance its inventory assortment, and has integrated generative AI into its parts and accessories app to help shoppers more easily find auto parts.

Etsy's Depop has a generative AI feature that auto-populates listings descriptions and attributes using a simple photo upload.

Amazon's third-party sellers can use AI tools to boost marketplace appeal and listings, bringing efficiencies that may draw even more sellers. The AI assistant Amelia answers sellers' questions, gives access to critical business metrics and cuts time to create and manage listings. Sellers can also use AI to quickly transfer listings to Amazon from their own websites. Enhancing the tool for bulk listings could widen product breadth.

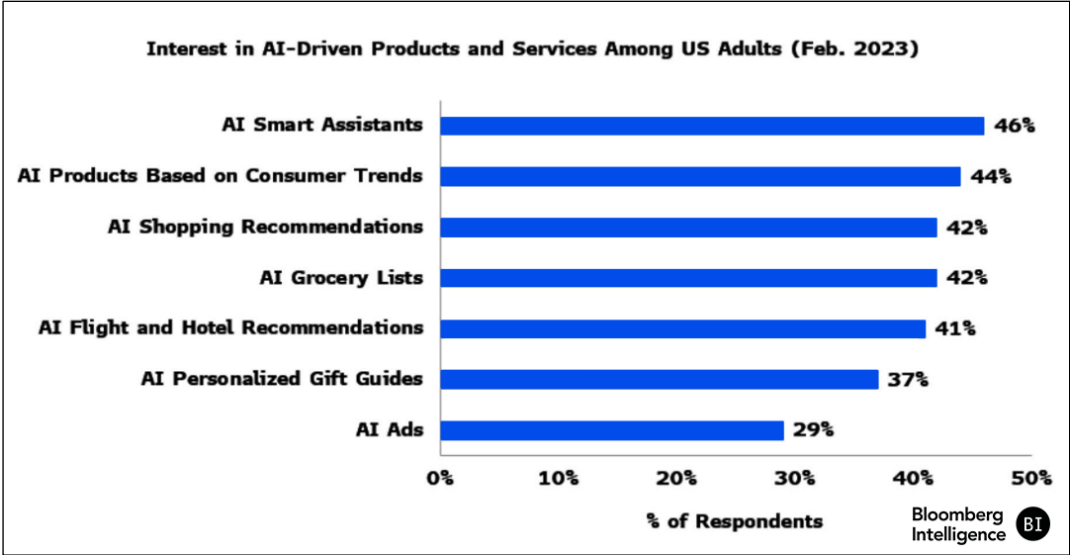
Amazon's AI-generated A+ Content tool can help sellers create more engaging and informative listings, attracting shoppers and saving time and resources previously allocated to photo shoots, drafting and testing. Amazon says it can increase sales by up to 20%. The Video Generator tool can create ads for sellers by using an image to create a custom video.

Generative AI is paving the way for chatbots to become personalized shopping assistants that take consumer requests and display appropriate brands and products. Snap, Meta, Pinterest and

other companies that already have shopping on their platforms are investing in AI chatbots and could implement personalized shopping assistants to boost user adoption of social commerce, fueling monetization opportunities. Enhanced capabilities to handle longer and more flexible search prompts may also benefit retailers. Walmart demonstrated a search-enhancement feature at the Consumer Electronics Show - or CES - in January 2024, which accepts less-specific user input and fetches relevant products. For example, the prompt “help me plan a football watch party” resulted in products like chips, salsa and other game-day essentials being shown. We expect more retailers to add longer search-prompt capabilities to their apps or websites, which may significantly expand the number of search queries on all kinds of digital platforms, especially in e-commerce.

Multimodal search could enhance the user experience beyond text-based functions, which currently dominate the market. We believe the conversational nature of ChatGPT might reduce ad loads in the near term, as summarized responses reduce the need to click on links to find information.

Figure 39: Service Interest



Source: eMarketer

As generative AI and machine-learning algorithms develop and adjust from user input, they cater to the person’s tastes, interests and lifestyle, providing a more customized, unique experience and curating new content for social media and search platforms. That can expand availability and engagement, similar to how TikTok uses AI algorithms to recommend content to users.

LLMs may improve ad targeting for large companies that are rich with first-party data. Meta has already pivoted to AI-based recommendations with its Reels offering, helping to offset some of the headwinds from Apple’s changes to its identifier for advertisers (IDFA) policy. Meta can continue to develop its Llama LLM and enhance the quality of its advertising campaigns.

Generative AI could also accelerate the shift to digital ads from linear TV, especially since offering personalized versions of advertising can increase efficiency and sales conversion. LLMs should also provide added benefit for existing large media companies as more premium content shifts

to streaming from linear TV. Our analysis suggests that the generative AI market may add around \$207 billion through 2032, through time spent on platforms, ad targeting and personalization.

6.7 Search, Recommendations Are Better Aligned

Using AI to improve search can boost gross merchandise value (GMV) and conversion rates as shoppers can more easily find items and use abstract keywords. Sifting through an endless aisle is tiresome, but offering the ability to narrow choices based on preferences and personal characteristics can change shopping behavior and boost sales.

Not only are retailers using large language models to filter across millions of products, but they're also able to request more details before providing search results. Lululemon's enhanced search functionality refines customer queries by prompting users for additional preferences, creating greater accuracy in meeting expectations.

Rent the Runway, Revolve and ThredUp are also using Gen AI to bolster search. Google's enhancements to browse with the technology -- circle to search and multisearch -- lets users remain in an app and search from within by circling a picture or video. In 1Q23, eBay said its search improvement was already boosting conversion and could eventually add about \$1 billion in annualized incremental GMV.

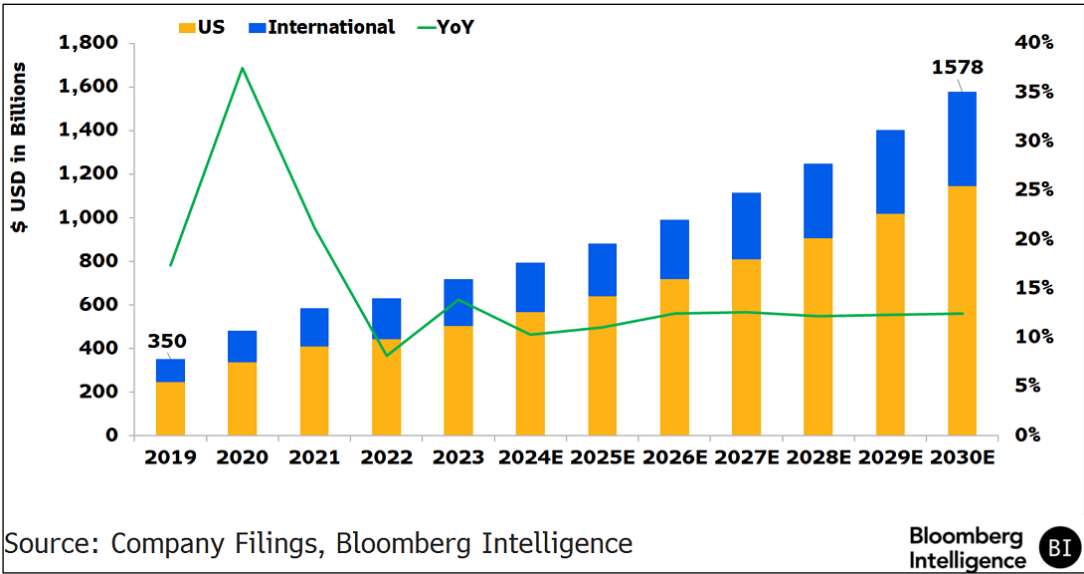
In 2Q, Etsy's gross merchandise sales for demi-fine jewelry grew 9% due to better search results and data-feed curation, and as the company continues to improve search and reduces the dominance of any one seller's item on the first page, it can continue to bolster conversion. Etsy has reduced the percentage of searches with two or more listings that may appear identical by over 70%, while the percentage of searches where a single seller's items dominated the page has been reduced by half.

Etsy's new AI-generated Gift Mode may attract shoppers and aid discovery for specific events and holidays, allowing consumers to enter information about who they're buying for and matching them with gifts on the platform.

Given Amazon.com's large scale (we break out its GMV in Fig. 40), it has a greater number of available data points to enhance search functionality, allowing for more personalized responses and product recommendations, further entrenching shoppers into its ecosystem. Through its large catalog of product attributes and customer-shopping information -- including preferences, search, browsing and purchase history -- the web giant can lean into large learning models to edit product listings to match customer preferences, making them more relevant for shoppers.

For example, if a customer is gluten-free, AI can more prominently display the relevant attributes of a product, which may have otherwise been listed at the bottom of a description.

Figure 40: Amazon US, International GMV



Integrating recommendations and similar-item features can also boost conversion as it helps shoppers more easily find what they're looking for. Amazon's new AI personal shopping assistant, Rufus, can answer shopper's questions, compare products, help consumers find things for a specific occasion and discover information about a specific product without the buyer needing to read all the details. This, coupled with its customer review summaries feature, can help people find items faster.

6.8 Supercharging the Entire Supply Chain

Generative AI offers a big opportunity for retailers to improve supply-chain efficiency and costs, with Amazon.com already quickening inventory processing by as much as 75% and Nordstrom citing a 20% increase in distribution-center productivity. Demand forecasting, warehouse management, transportation and logistics and supplier-relationship management may all benefit, and omnichannel fulfillment is likely a key area for enhancement.

Amazon continues to deploy generative AI across its ecosystem, with fulfillment centers using more robotics systems to improve speed and efficiency. It has partnered with Covariant to build advanced AI models for its warehouse robots and deployed its robotic system, Sequoia, to help identify and store inventory at its fulfillment centers faster. Packages then move through an AI-powered trio robotic arm -- named Robin, Cardinal and Sparrow -- to sort, stack and consolidate items.

Amazon recently launched a Vision Assisted Package Retrieval tool for its delivery vans to find the correct packages faster, shortening routes by 30 minutes.

Nordstrom's chief supply-chain officer said automated fulfillment centers are 5x more efficient than those without the tech upgrades. These enhancements also help reduce out-of-stock items.

The RealReal and eBay are also using AI to streamline the apparel and accessories authentication process, while Warby Parker and Etsy use generative AI to increase the productivity of their software engineers.

As e-commerce sales climb, returns are also rising, which may create bottlenecks in retailers' supply chains, notably when there's no dedicated return center. It often results in a buildup of inventory at warehouses, and as that merchandise waits to be processed, it depletes the value and possibility for a resale at full price. Retailers are leveraging AI to help cut the amount of returned goods, such as H&M analyzing which items get returned the most to help optimize for products that shoppers are more likely to keep, or Amazon's "predict-the-fit" feature to help shoppers select the right size, making them less likely to return it.

Fraudulent returns are a growing problem for US retailers, costing over \$100 billion in 2023, based on the National Retail Federation. About 14.5% of all retail purchases were sent back in 2023.

6.9 Time-to-Market Speed Increase for Product Development

Growing use of generative AI to help with product design can help shorten the window from conception to production, reducing long lead times that often produce inventory imbalances and missteps. Product capsules can then be created much faster. Revolve, Nike and Lululemon, among others are using generative AI to improve and speed up the product pipeline.

In addition to using AI to design marketing ads, notably billboards, Revolve is also using AI to produce new product lines. Last year it worked with winners from the inaugural AI fashion week to create a limited-edition, AI-created capsule collection to be sold on its platform. Though the implementation of generative AI is still in early stages, its growing use can continue to help designers broaden the possibilities more quickly for the future.

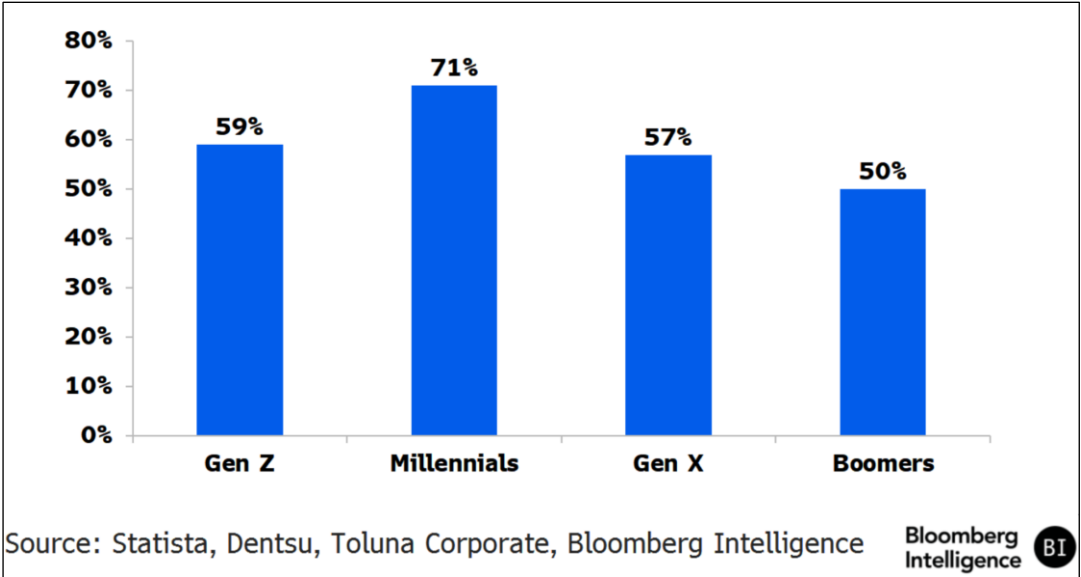
The use of generative AI tools, virtual reality and 3D rendering software has enabled Nike designers to create unique designs that reflect the personality of the athletes, adding newness to the product pipeline. The use of AI isn't yet a means to replace the designer, but rather helps fuel their creativity with results available in minutes and seconds vs. weeks and months. The use of AI models to extrapolate data from texts, images, video and code, to create personalized products that fit the needs and molds of unique feet, can help Nike add more differentiation to its product lineup.

Hyper-personalized shoes may not make their way into mainstream retail just yet, but they can solve athletes' needs and reflect their persona better, helping create a halo effect for a version of the style when sold.

BI

Younger consumers more supportive of brands that use AI-aided designs

Figure 41: Consumer Support for Brands Using AI-Aided Design



6.10 Improving Customer Service and Experience

Retailers, as shown in Fig. 42)are using AI to improve customer-service offerings by leveraging chatbots that can cut expenses and improve the buyer experience. Amazon’s AI shopping assistants can answer personalized questions, while Wayfair uses AI to match shoppers to the right customer service agent.

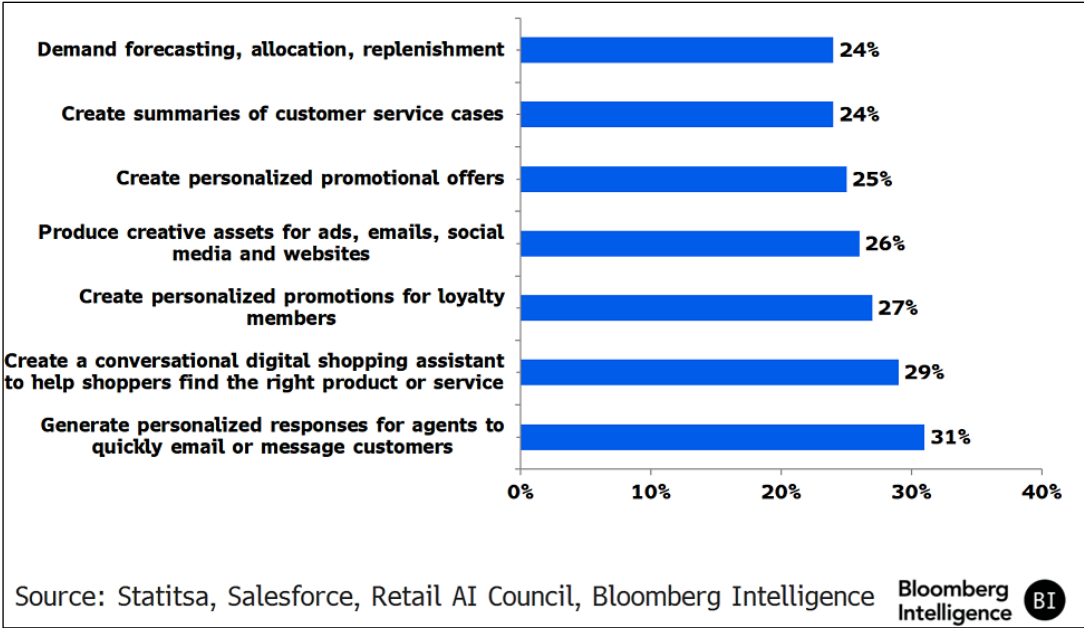
According to IBM, using chatbots can reduce expenses 30% and boost customer satisfaction by 25%. Besides Amazon’s AI shopping assistant, other companies employ chatbots to provide shoppers with immediate help. Etsy is also using the technology to streamline processes, including in trust and safety, for its engineers to drive efficiencies.

Wayfair uses gen-AI to match shoppers to the best customer service agent, predict what kind of service is needed by a customer with a complaint and what kind of discount might persuade a buyer to keep an item. Warby Parker is using gen-AI to quickly transcribe eye-wear prescriptions and help conduct virtual vision tests.

Amazon’s gen-AI-powered shopping assistant, Rufus, can answer customers’ questions to help search for products, product comparisons and reviews to make recommendations to improve conversion. The company is also using AI to summarize customer reviews, which can spur conversion as it saves time for shoppers.

Amazon’s AI-enabled smart-home assistant, Alexa, has been programmed to understand complex natural-language queries and offer personalized recommendations, provide summaries of information and answer open-ended questions like “what to wear to the movies?”

Figure 42: Main Use Cases of Generative AI in Retail Industry



Section 7. Capital Spending Outlook

Appetite for AI to Feed \$2.2 Trillion in Capex

Generative AI workloads are intense, which should spur near-term corporate investment in servers and storage. Growth in global software spending has averaged around 10-12% annually for the past few years and, despite a recent slowdown, prospects for the category are much stronger as companies invest in AI. Software makers, in particular, can burnish their product offerings by adding generative AI. As a result, capital expenditures to deploy these technologies could increase, expanding software spending 13% annually in 2022-32 and reaching \$2.2 trillion by decade's end.

7.1 Hyperscalers Key to Training Projects

In the near term, data centers and cloud operators most likely will tolerate increased costs to ensure high-quality performance for AI workloads since malfunctions and system failures could lead to lawsuits, canceled contracts and financial damages. We believe that eventually, most hyperscalers – like Alphabet, Meta and Amazon – will target capex to develop proprietary, foundational large language models that will work best on their own cloud infrastructures

Hyperscale providers including Meta, Nvidia, Microsoft, Alphabet and Amazon.com will likely be among the main facilitators for training large language models on the public cloud. These companies have the means to invest in infrastructure to handle heavy AI workloads while still maintaining high usage for their servers to sustain healthy profit margins. Many of these hyperscalers already announced plans to ramp up capital spending to avoid underinvesting in the technology. Most of these companies also made multibillion-dollar investments in companies like Anthropic and Runway to leverage existing AI tools to enhance their public cloud offerings.

Our analysis of the top tech companies shows over \$90 billion in incremental capital spending in 2024-25 vs. 2023, dedicated mostly to expanding generative-AI infrastructure. This group, which includes Amazon Web Services, Microsoft, Google, Oracle, Meta and Apple, added an average of \$14 billion annually to capex from 2020-23. The \$90 billion increase over a two-year period illustrates the increased demand and interest by clients, and appears different than other hyped technologies, such as the metaverse.

Our calculations point to an additional \$55 billion in 2024 and \$35 billion the following year, totaling \$200 billion spent in 2025 (Fig. 43) as these enterprises aggressively prioritize data-center expansion and creating copilots.

Microsoft is at the forefront of capital spending among big tech providers, followed closely by Google and AWS, which we believe will mostly go toward expanding their respective data-center footprints and buying more AI-related chips and hardware. We calculate Microsoft's capital outlay could hit \$53 billion in 2024 and \$62 billion in 2025, mostly to meet greater demand for OpenAI workloads, both from consumers and companies. For reference, Microsoft had \$17.6 billion in capital expenditures in 2020. Tjority of incremental spending is likely to be tied to cloud services.

The bulk of this capital spending will likely go toward expanding AI infrastructure, which includes creation or retrofitting of data centers and buying GPUs. A new data center takes roughly 18-30 months to build, according to AWS Global Data Centers Vice President Kevin Miller, who discussed the topic on BI's Tech Disruptors podcast. In addition to building these facilities, the top three cloud providers are also leasing computing capacity from small providers like Oracle. Leases tied to capital expenditures over the past 12 months totaled \$11 billion for Microsoft, roughly 20% of the total outlay of \$56 billion. Our calculations lead to \$13.6 billion of spending tied to leases in 2025, which can be viewed as easier to shut down if demand dries up.

Smaller peers without the ability to construct their own in-house LLMs may consider deploying their workloads on the public cloud using these hyperscalers.

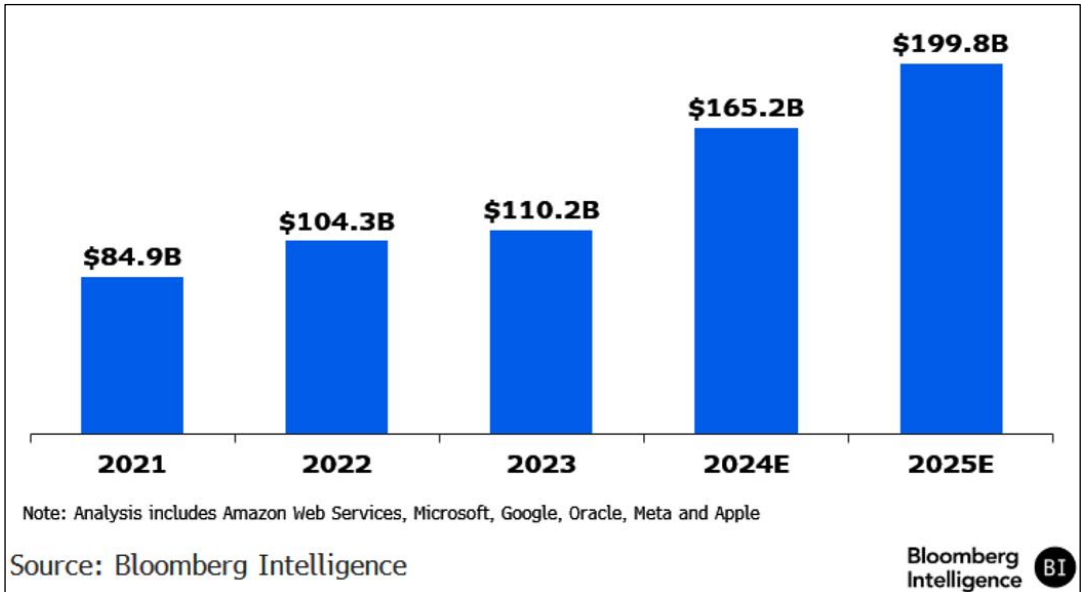
For Microsoft, AWS, Google and Oracle, their respective cloud sales are likely to be the biggest and most direct beneficiaries of increased generative-AI spending, based on our analysis. Some large corporations may pursue buying AI-related servers, GPUs and other hardware for their on-premise IT infrastructure, but we anticipate most clients will embrace a cloud-based model, especially as these big tech companies already have a strong infrastructure in place for inferencing.

Though AI workloads are net new revenue for cloud providers, this revenue stream may cannibalize traditional workloads, given that IT budgets are tight.

BI

Big tech could spend \$200 billion in capex in 2025

Figure 43: Big Tech Capital Spending



Section 8. Processing, Memory Chip Demand

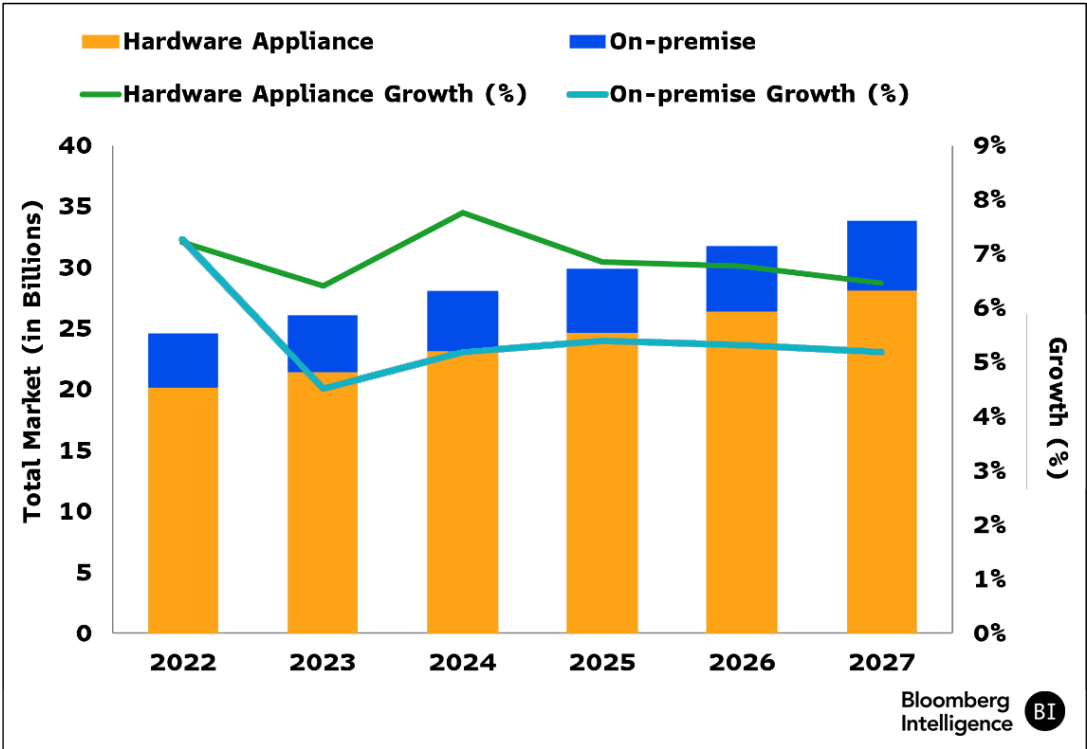
Growth Opportunities Spread Around the Globe

The spread of artificial intelligence could drive demand for graphics processing units (GPUs) and dynamic random-access memory (DRAM), lifting sales at SK Hynix, Samsung Electronics and Micron. We believe memory semiconductors will play a key role in the expansion of the data-center chip market, along with AI accelerators, with each expected to grow more than 15% annually over the next three to five years.

8.1 TSMC Prowess Sets It Apart From Pack of Rivals

Fortinet and Palo Alto Networks may leverage custom semiconductors for generative AI, supporting steady refreshes of their hardware and software firewalls. Fortinet added software-defined wide area network (SD-WAN) functionality to its ASIC chips, helping it to take share from traditional firewall vendors such as Cisco and Check Point. Palo Alto successfully bundled its Prisma, Cortex and virtual firewalls to help enterprise customers secure their on-premise and public-cloud workloads.

Figure 44: Network Security Market



Source: IDC

AMD's energy-efficient AI accelerators position stand to gain amid rising scrutiny over power consumption by data center AI chips.

TSMC is poised for a strong rebound in 2024-25, fueled by robust demand for AI accelerators and significant order gains from leading AI chip designers like Nvidia and AMD. Despite a 9% revenue dip in 2023, we forecast a 22% surge in 2024, buoyed by significant demand for its 3- and 5-nm processing node technologies and 2.5D packaging. AI infrastructure investment and chip demand growth will be a long-term trend, which will help TSMC overcome headwinds spawned by sluggish growth for PCs and smartphones chips.

The burgeoning AI chip market, catalyzed by rapid advancements in generational AI technologies, will notably benefit TSMC, with the server GPU and AI accelerator market expected to quintuple to \$51.6 billion from \$10.5 billion in 2022, according to IDC.

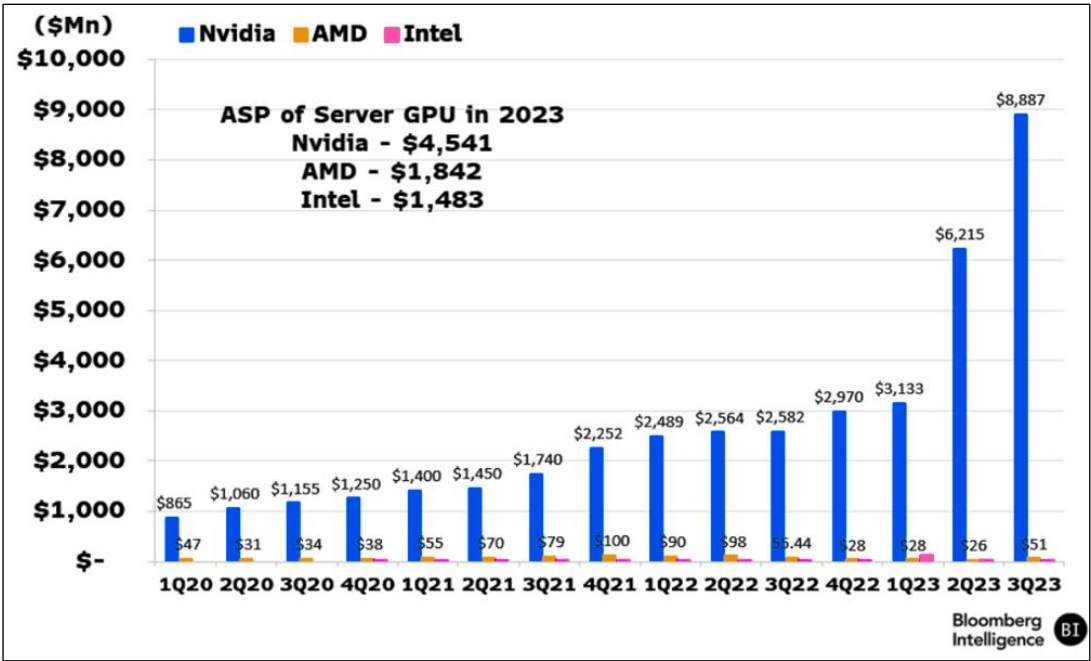
Figure 45: BI Scenario Analysis

	2020	2021	2022	2023	BI Scenario		
					2024E	2025E	2026E
(Revenue By Applications, \$Mn)							
Smartphone	21,917	24,887	29,835	26,146	31,013	35,849	39,434
Sales Mix %	48%	44%	39%	38%	37%	35%	34%
HPC	14,938	21,045	31,295	29,992	39,460	49,416	59,480
Sales Mix %	33%	37%	41%	43%	47%	49%	52%
Others	8,632	10,904	14,856	13,220	13,824	16,101	16,202
TOTAL Sales	\$ 45,487	\$ 56,835	\$ 75,986	\$ 69,358	\$84,298	\$101,367	\$115,117
Growth %		25%	34%	-9%	22%	20%	14%
Sales Consensus Estimation (as of 31/12/23)					84,085	98,783	110,357
					21%	17%	12%
Key Assumptions:							
1. Smartphone related revenue to grow at an 15% compound average rate from 2024 to 2026							
2. High Performance Computing (HPC), which include AI accelerator, server processors, PC will grow at a 26% compound average rate from 2024 to 2026							
3. Total sales contribution from smartphone and HPC chip businesses will be 84% in 2023, 2024 and 87% in 2026							
4. TSMC will still secure over 90% of global AI accelerator production orders							

Source: Bloomberg Intelligence

TSMC's dominance in leading-edge node semiconductor manufacturing processes positions it to maintain its hold on the lion's share of AI chip production orders from key players such as Nvidia and AMD. That advantage is expected to continue, thanks to the company's strong production yield. Also, many AI chip designers prefer TSMC's CoWoS packaging for its superior interconnection density, larger package sizes and cost effectiveness.

Figure 46: AI Server ASPs

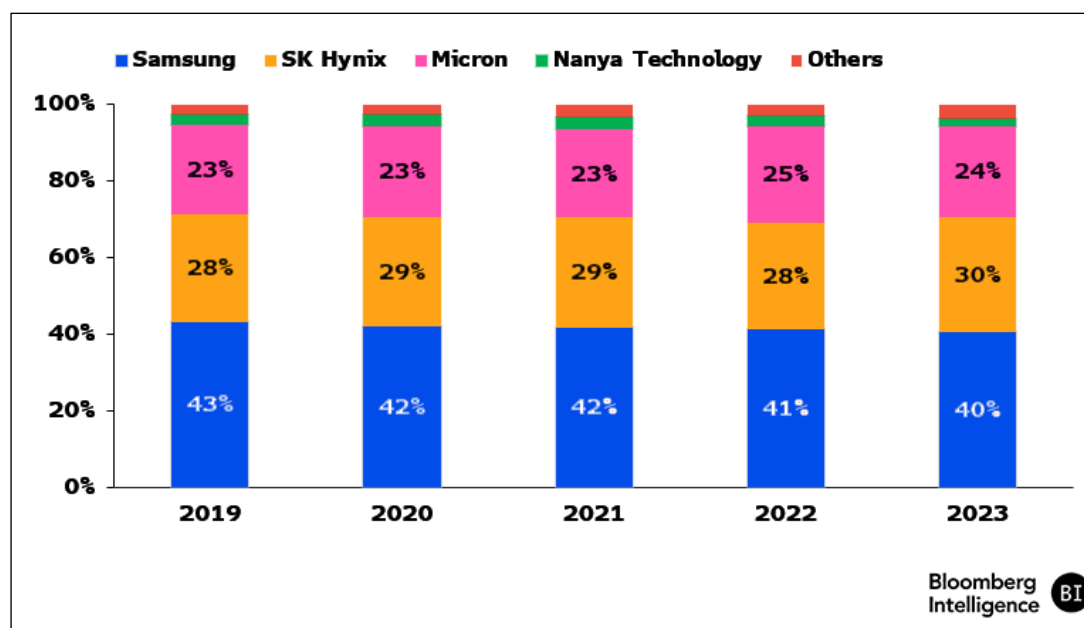


Source: IDC, Bloomberg Intelligence

8.2 Need for Speed Fuels Rapid Performance Gains

High-bandwidth memory (HBM) chips are due for a significant role since rapid performance improvements of GPUs can only be fully realized if memory can supply it with large volumes of data at high speed. As AI models become more complex and training becomes more demanding, HBM chips, which have high selling prices and operating profit margins, are expected to be more widely adopted. Since SK Hynix has built a good relationship with Nvidia on the current generation of GPUs, it can benefit from increasing demand.

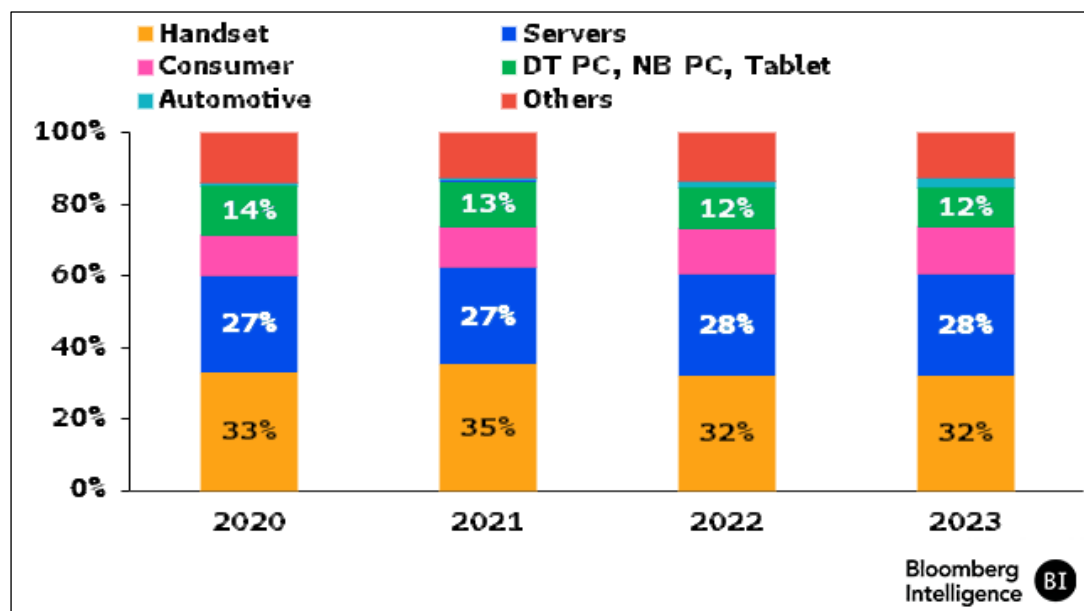
Figure 47: DRAM Bit Demand by Application



Source: Gartner, Bloomberg Intelligence

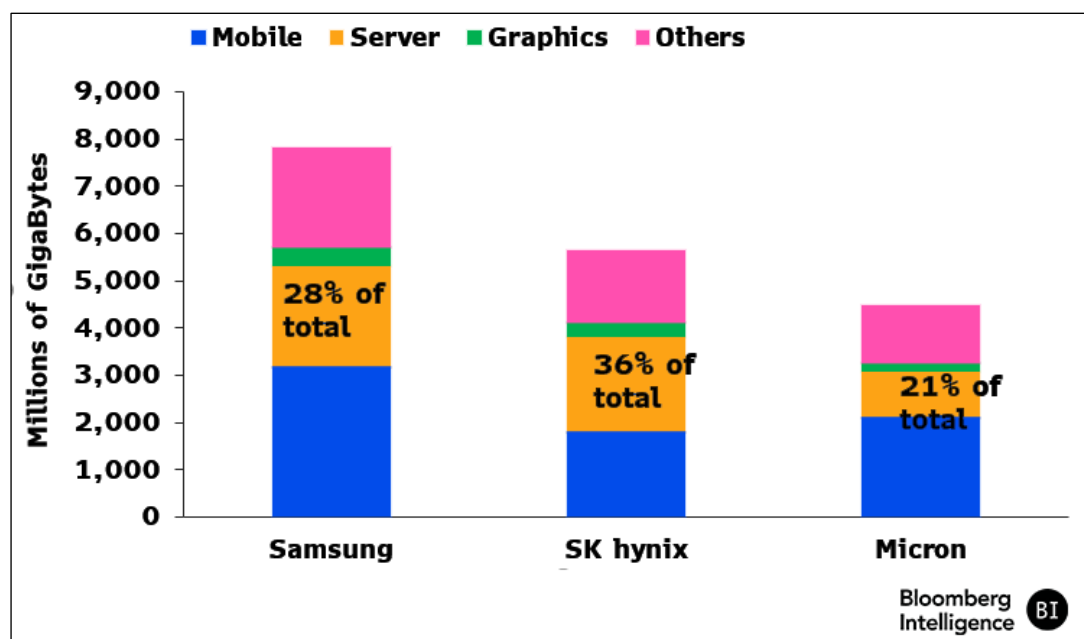
Nvidia's next generation GPUs. Micron could also grow revenue, as it started volume production of HBM chips for Nvidia in early 2024. Results at Samsung, the world's largest DRAM maker, also may rise amid increasing use of GPUs and HBMs.

Figure 48: DRAM Bit Demand by Application



Source: Gartner, Bloomberg Intelligence

Figure 49: Leading DRAM Makers Gigabytes Shipped by Use



Source: IDC, Bloomberg Intelligence

Escalating use of AI for inference applications might heighten the value of DRAM formats like graphic double-data rate (GDDR), which is used in retail PC graphic boards where cost is critical, and low-power-consumption double-data rate (LPDDR), which is mainly deployed in smartphones.

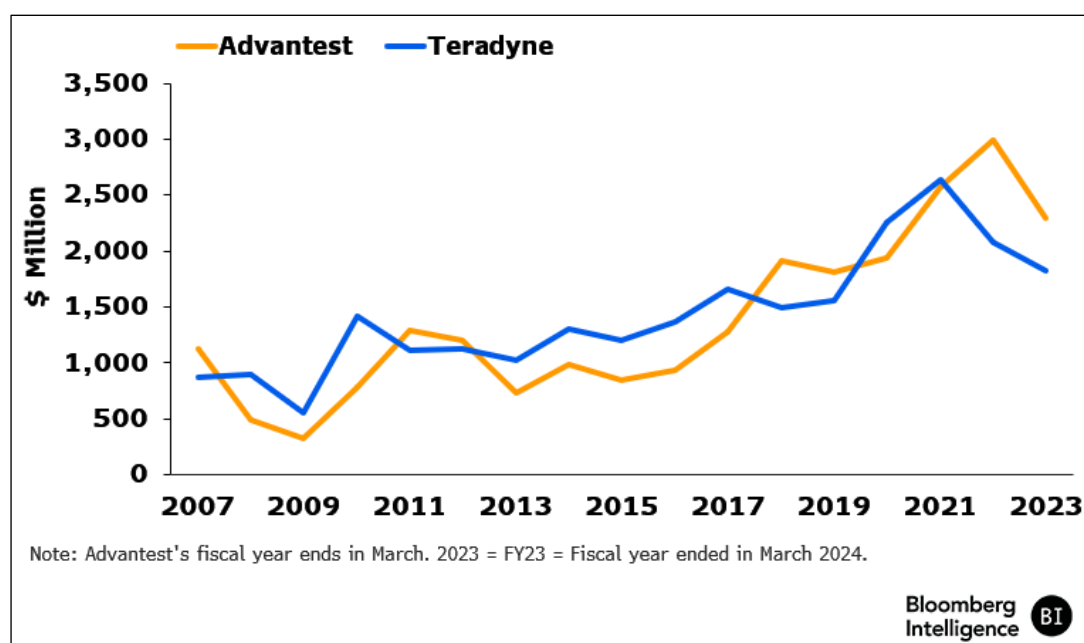
Power consumption and data telecommunications would become excessive if all processing were conducted on servers. That will make it necessary to perform AI tasks on edge devices, which could fuel a surge in DRAM orders for products including PCs, autos, robots, smartphones and security cameras, buoying sales for Samsung, SK Hynix and Micron.

The volume of DRAM used on servers to perform AI's large-scale calculations is smaller than smartphones and PCs. Smartphones and PCs account for about 40% of global DRAM bit demand. It is true that servers account for about 30% of overall DRAM demand and artificial intelligence is just a small portion of that. However, demand for specific tools to make AI processors, HBM chips and AI chip packages could grow robustly and start to contribute to sales and profit growth, as chip customers are expected to rapidly expand production capacity.

The speed of chip performance improvement required for AI is faster than the evolution of miniaturization and advanced packaging, meaning quality isn't certain. As a result, the role of chip testers (see Fig. 50) that can accurately assess performance and quality might rise sharply.

Teradyne has a strong competitive edge in the field. And customers have given high marks to Advantest's T5000 series of memory chip testers and V93000 series of system-on-a-chip testers.

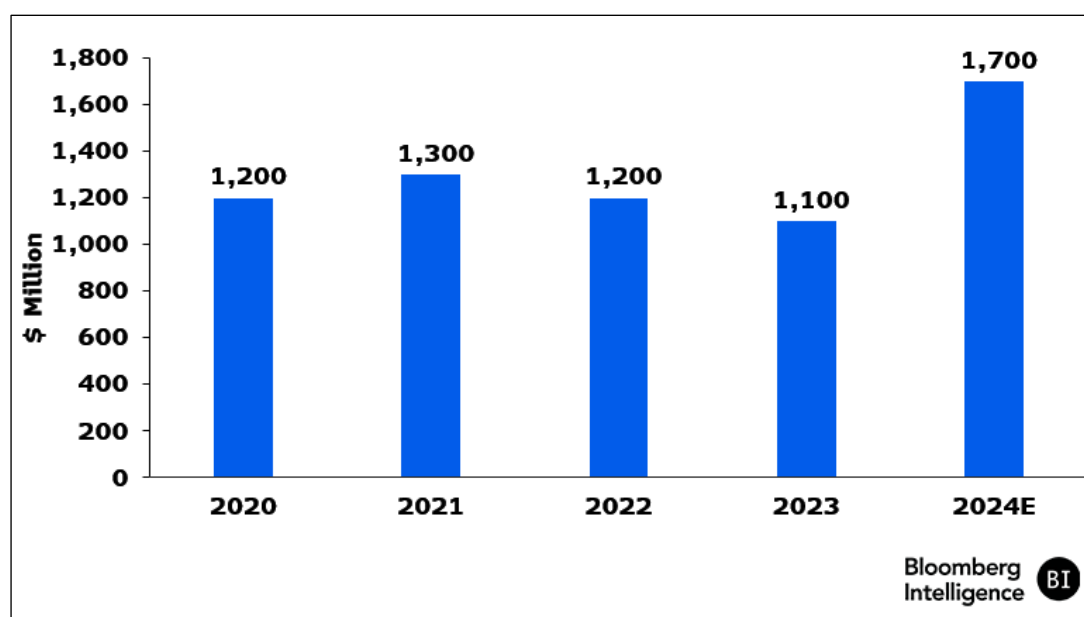
Figure 50: Chip Tester Sales for Advantest, Teradyne



Source: Company data, Bloomberg Intelligence

Quality assurance will be necessary not only for single chips, but also for chip packages, and entire system-level tests (SLTs) will become more important. Three-nanometer technology is taking off and chiplet packages, which place several chips or dies on a substrate vertically, could be adopted in a few years. Beyond identifying defective products, tester makers like Teradyne need to reduce the risk of misclassifying good products as deficient, requiring substantial technological capability. Advantest strengthened its SLT business by acquiring Astronics in 2019 and Essai in 2020.

Figure 51: Memory Tester Market Size



Source: Advantest, Bloomberg Intelligence

8.3 User-Based Software Growth May Shrink

The onset of AI-powered agents and copilots could enable companies to optimize resources, build more applications in-house and that could reshape the seat-based businesses that sell to them. As AI-driven productivity gains broaden across industries and functions, job growth could slow, pressuring seat-based revenue models. We're also doubtful of a revenue uplift from price increases as AI features standardize.

Managed-services providers like Confluent that have built their business around open-source models may need to beef up their product innovation. The ease of code generation through simple natural-language commands -- which has become possible with the use of large foundational models and applications such as OpenAI's ChatGPT -- potentially brings application development capabilities to the masses. It's similar to the content generation that became possible with the confluence of mobile phones, social-media platforms, broadband connectivity and easy content-editing tools. With code-generation workloads potentially easing, developers' time could be used to build more applications in-house, using open-source software and engines rather than through third parties or proprietary products.

Figure 52: Shifting Business Models

		Companies Exposed To The Theme
Seat Based Models	AI could act as double-edged sword. While it does bring prospects for streamlining cost base, the productivity gains that come with AI and the standardization of AI features across technology platforms could potentially narrow product edge (differentiation) and curb the seat and headcount expansion led growth.	Atlassian GitLab Monday Asana Smartsheet
Opensource, AI Disruption	Adoption rates on solutions like Grafana in observability, OpenTofu in infrastructure provisioning, Apache Icerberg, PostgreSQL, Apache Spark, Trino for data workloads, Apache Cassandra, Weaviate and Milvus for Vector search -- need monitoring and could potentially become more pervasive and limit incremental growth fro managed service solutions.	Confluent MongoDB
Point Solutions	The integration of AI capabilities into existing technology solution and as large technology vendors enter into adjacent verticals, they could stand to reap from customer moves to consolidate technology stack and vendors.	C3.ai UiPath
AI Infrastructure	The rising capex towards AI infrastructure could also drive up demand for infrastructure provisioning and solutions that help monitor and optimize those costs.	Datadog Dynatrace Elastic Cisco HashiCorp Flexera IBM (Apptio) ServiceNow

Bloomberg
Intelligence

BI

Source: Bloomberg Intelligence

We see risk to revenue consensus for point-solutions software companies, seat-based revenue models and those that lack scale and face narrowing product edges like Asana, MongoDB, C3.ai, Confluent, Rubrik and UiPath. The changing competitive landscape in the observability sector, with open-source solutions like Grafana gaining scale, could force companies like Datadog, Dynatrace and Elastic to sharpen their execution and demands monitoring.

The criticality of digital assets -- application, compute, storage and data layers -- to modern businesses demands always-on monitoring, to prevent operational disruptions and to remediate at speed. The recent incidents involving CrowdStrike and Snowflake reflect the rising risk of bad actors and potential implication of disturbances to the highly interconnected digital ecosystem.

Rising capital spending toward AI infrastructure could also drive up demand for infrastructure provisioning and products that help monitor and optimize those costs.

Section 9. Regulatory Landscape

EU Farther Ahead Than US; Big Tech Exposed

Trust and content safety for generative AI needs to be strengthened, and large participants including Snap, Meta, TikTok and Alphabet are well positioned to detect and prevent deep fakes created by its misuse. Doing so should improve brand safety for advertisers and increase conversion rates for ad spending on these platforms. Any potential increase in AI-related regulation also could contribute to greater outlays on securing and encrypting data. Europe is farther ahead than the US on regulating AI, and given the tool's rapid development, we believe that creating a dedicated agency is one of the few workable regulatory models.

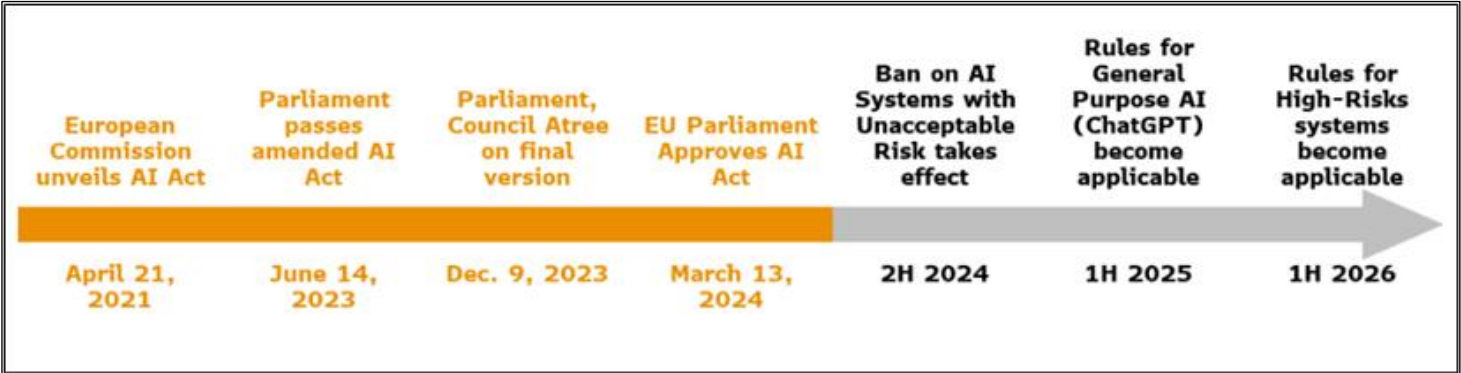
9.1 EU AI Rules Add to Regulatory Thicket

Apple's delayed rollout of Apple Intelligence in Europe isn't due just to the EU's AI rules, but also a thicket of regulations the EU has imposed on tech in recent years that can make clearance for new product launches onerous and, if done incorrectly, costly. The Digital Markets Act is the primary cause of Apple's delay, while the General Data Protection Regulation (GDPR) has halted Meta, X and Google from training their large-language models on EU user data

The AI Act, which will subject general-purpose AI models like ChatGPT to heightened transparency rules from early 2025 (Fig. 53), by itself won't pose a significant obstacle to the rollout of AI in Europe, but there could be near-term pressure on AI investment and innovation until there's more clarity on the interplay between the AI Act and other regulation

The transparency obligations will only apply to general-purpose AI systems that are deemed to pose a "systemic risk." The EU will use training computing power as a basis for bucketing general-purpose AI models into the systemic-risk category, with an initial cutoff of 10 to the 25th power of floating-point operations per second (FLOPs). OpenAI's GPT-4 and Google DeepMind's Gemini are likely to be the first to fall into that category, according to the European Commission. The FLOPs threshold could be adjusted upward or downward over time. Meta has said it won't release its most-powerful AI system in the EU due to regulatory uncertainty. Unless it reverses course, it could avoid systemic-risk designation.

Figure 53: Timeline of EU Act Legislative Process



Source: Bloomberg Intelligence

The EU's AI Act buckets systems into categories based on the type of risks their applications pose. The rules will outlaw all "unacceptable risk" applications by the end of the year. That includes use of AI for things such as behavioral manipulation and biometric identification. For "high-risk" applications, which comprise large platforms' recommendation systems, a multistep approach to get approval will take effect in 2026. The European Parliament sought to include general AI systems as high risk, but, in a boon to the industry, that was rejected. Those systems will instead be subject to transparency requirements from next year. Limited-risk applications (such as chatbots) would simply require disclosure, while minimal-risk applications (like the use of a spam filter) wouldn't have any restrictions.

The AI Act could impose fines of up to 7% of annual turnover, exceeding the 4% maximum under the current General Data Protection Regulation. In its first five years of enforcement, the GDPR resulted in cumulative fines of almost €4 billion, with penalties on Meta making up 64% of the total. Amazon accounted for 19% and Google, 5%. We don't believe the threat of penalties would deter near-term investment in generative AI, given the potential to shape the market.

Figure 54: Rules Slowing Rollout of AI Products

Firm	Action	Regulation
Apple	Will not roll out Apple Intelligence in Europe at same time features are offered in other regions	Digital Markets Act
Meta	Will not train large-language-model using user-generated content from EU users	General Data Protection Regulation
Google	Being probed by data privacy regulators	General Data Protection Regulation
Microsoft	Will not train LinkedIn's large-language-model using user-generated content from EU users	General Data Protection Regulation
X	Will not train large-language-model using user-generated content from EU users	General Data Protection Regulation

Bloomberg Intelligence **BI**

Source: GDPR Enforcement Tracker, Bloomberg Intelligence

Transparency obligations for general-purpose AI models under the EU's AI regulatory package are significantly less onerous than the multistep approval process contemplated for advanced AI systems under an earlier EU proposal. Models that pose "systemic risk" will have additional rules, the most burdensome of which are to assess and mitigate risks. Even that heightened level of scrutiny looks more lax than the pre-release conformity assessment contemplated before.

Obligations to develop and adhere to a copyright policy, and to make available a detailed summary about the content used to train the models will likely fuel more licensing agreements between publishers and developers of large language models.

The AI Act would add to the EU's increasingly complex regulatory framework on tech. In recent years, the bloc has sought to rein in the wanton collection and use of personal data (the GDPR) and to impose obligations on platforms relating to content moderation (the Digital Services Act) and abuse of market power (the Digital Markets Act). The rules can impose substantial financial penalties and operational remedies that can materially alter a business. Enforcement of the GDPR has been fractured, with Ireland taking the mantle for regulating most platforms. The European Commission monitors DSA and DMA compliance. Individual national regulatory bodies likely would lead AI Act enforcement.

BI

A look at the European Union's recent tech regulations

Figure 55: EU's Recent Tech-Focused Regulations

Regulation	Compliance Date	Purpose	Mostly Likely Targets	Max Fine
General Data Protection Regulation (GDPR)	May 25, 2018	Provide strict rules on the collection and use of personal data	Meta, Amazon and Google have been the most targeted by GDPR regulators	4% of annual revenue
Digital Services Act (DSA)	August 25, 2023	Impose new obligations on platforms to enhance moderation of illegal content	EC has identified 17 "Very Large Online Platforms" and two "Very Large Online Search Engines" that were required to comply with the DSA from August 25, 2023	6% of annual revenue
Digital Markets Act (DMA)	May 2, 2023; March 6, 2024	Provide new competition rules on large, "gatekeeper" platforms	Gatekeepers include large tech platforms, which will then have to comply by March 2024	10% of annual revenue
AI Act	2026 to 2027	Provide broad safeguards on the development and deployment of AI systems	Potentially any developer or implementer of an AI system, but large tech firms once again likely to be focus given early investments into generative AI models	7% of annual revenue

Bloomberg Intelligence **BI**

Source: Bloomberg Intelligence

9.2 Aggressive Regulatory Tack Unlikely in US

If the US adopted an aggressive regulatory approach – which we doubt will happen – it could dent growth of AI products from a range of companies: chipmakers like Micron and Nvidia; cloud infrastructure providers like Amazon and Oracle; software- and development-tool makers like Adobe, IBM and Microsoft; and platforms that use AI for data, search and ad capabilities, like Alphabet and Meta.

The first federal bipartisan bill – the No Section 230 Immunity for AI Act, which would do little – shows how much work remains before serious AI limits land in the US. The measure would clarify that a federal liability shield, Section 230 of the Communications Decency Act, wouldn't apply to

AI, but we think courts would only rarely apply the provision to the technology anyway. More notably, the legislation wouldn't create a federal right to sue over AI harms. Nor would it attempt to say what AI is, defining "general artificial intelligence" so broadly that the law could remove Section 230's shield from many existing online platforms. The provision most likely won't become law without being narrowed significantly.

A second bipartisan bill, the National AI Commission Act, strikes us as a sensible first step for Congress, with a solid chance of becoming law. It would create an independent, bipartisan commission with 20 members to study AI's risks and suggest guardrails. The panel would release three reports over two years, shaping regulation. The commission would include industry representatives, which should ease opposition.

A Senate committee hearing in May generated surprising support for what we see as the most logical – though probably the most disruptive – approach: creating a dedicated federal agency. Congress is incapable of keeping up with AI's rapid development and though a new agency would struggle as well, it at least would have a shot at monitoring the technology and fashioning legal limits. Such a body could also be more focused than an entity like the Federal Trade Commission, which has oversight of all industries. A new agency would upset the apple cart, however, so it likely would face vigorous industry opposition, and companies that might support it probably would push to narrow the body's powers.

Congress also might consider a licensing model, in which AI applications with a broad reach or the potential to cause severe harm can't be put into operation until they've obtained permission from a new regulator. The approach would face fierce industry pushback, including claims that the move would stifle AI innovation and drive it abroad.

A less intrusive form of regulation would focus on transparency, requiring that AI products be disclosed and labeled, and might require that regulators or researchers be able to monitor data that's collected or used.

Since Congress is unlikely to reach consensus on a disruptive approach – like creating a new agency or developing a licensing model – we expect AI to be governed by existing laws, even if they're not always a good fit. The FTC could monitor for unfair and deceptive practices, for example, and existing sector-specific laws may limit AI application by industry. AI should thrive in the US under such a light touch. Yet unlike social-media companies, AI users probably won't receive broad protection from lawsuits from Section 230, which raises the specter of legal liability.

Section 10. ESG Outlook

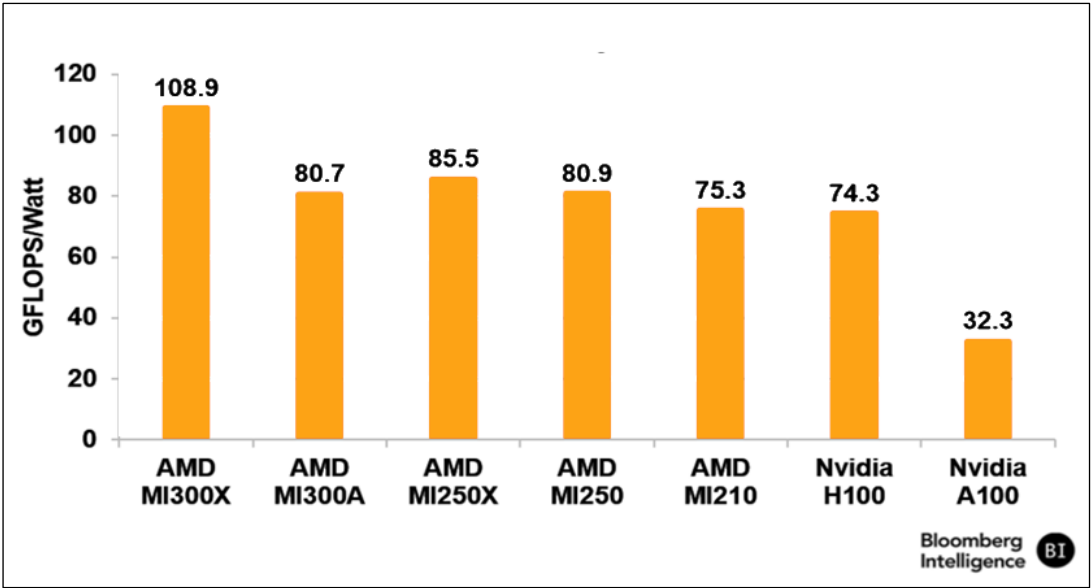
Reducing Power Consumption; Protecting IP and Privacy

Increased use of graphics processing units for AI inference will significantly boost energy consumption in data-center servers, putting a priority on energy savings to maximize operating efficiency while minimizing power and cooling costs. Meanwhile, about 40% of respondents in a BI survey expressed concern over how generative AI's use of information could breach intellectual-property rights, with privacy another concern.

10.1 AMD's Latest Can Challenge Nvidia

That could favor AMD over Nvidia (Fig. 56) as the former's latest MI250X accelerator outperforms Nvidia's H100 by performing a higher peak number of floating-point operations a second per watt. We believe ARMs can continue to gain market share from x86 processors in data centers. Existing internet workloads largely run on x86 architecture, but most generative AI applications will run on chips like GPUs that can conduct massive parallel processing with low power consumption.

Figure 56: Energy Efficiency of Advanced AI Chips



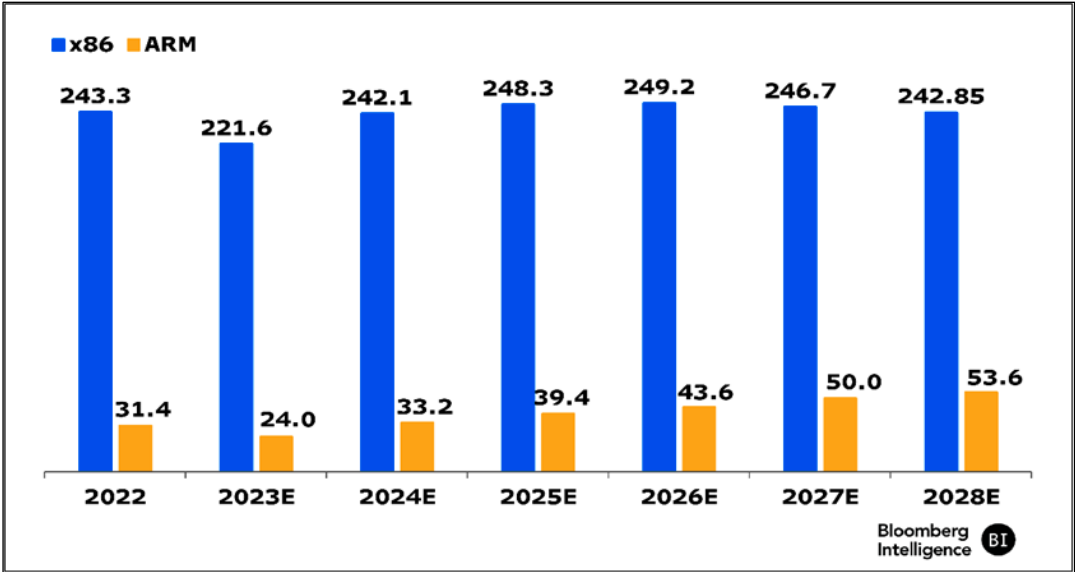
Source: Company filings, Bloomberg Intelligence

Nvidia's advantage with its CUDA interface is likely to remain for ARM-based processors and may help the company gain market share over x86 processors across data centers, networking and edge devices thanks to the ARM design's power efficiency.

BI

x86 shipments seen steady, while ARM climbs through 2028

Figure 57: x86 vs. ARM Shipment Forecast



Source: IDC

A majority of survey respondents said they still would use gen AI tools despite concerns about intellectual property rights, as long as they provided better results than traditional search functions from Google and other websites. IP concerns will likely decrease over time as companies share more information about how their algorithms are trained using proprietary rather than open web data.

Developers of high-level generative AI systems could retreat from the EU, absent further revisions to the bloc's proposed regulations. Mentions of generative AI and ChatGPT by European companies exploded this year, suggesting a strong desire to employ the new technology. Yet the European Parliament on June 14 adopted rules that would subject developers of generative AI models to additional constraints, such as transparency rules related to the data sets used for training.

Since all AI essentially is driven by data – collecting lots and using machine learning to produce outputs based on it – US regulators could address related harms by limiting what can be collected and used. That would be similar to data-privacy regulations that have been proposed for social-media companies. Given the parallels, expect internet platforms to vigorously lobby against such limits.

Section 11. Performance and Valuation

AI Establishes Itself as Dominant Accelerating Tech Theme

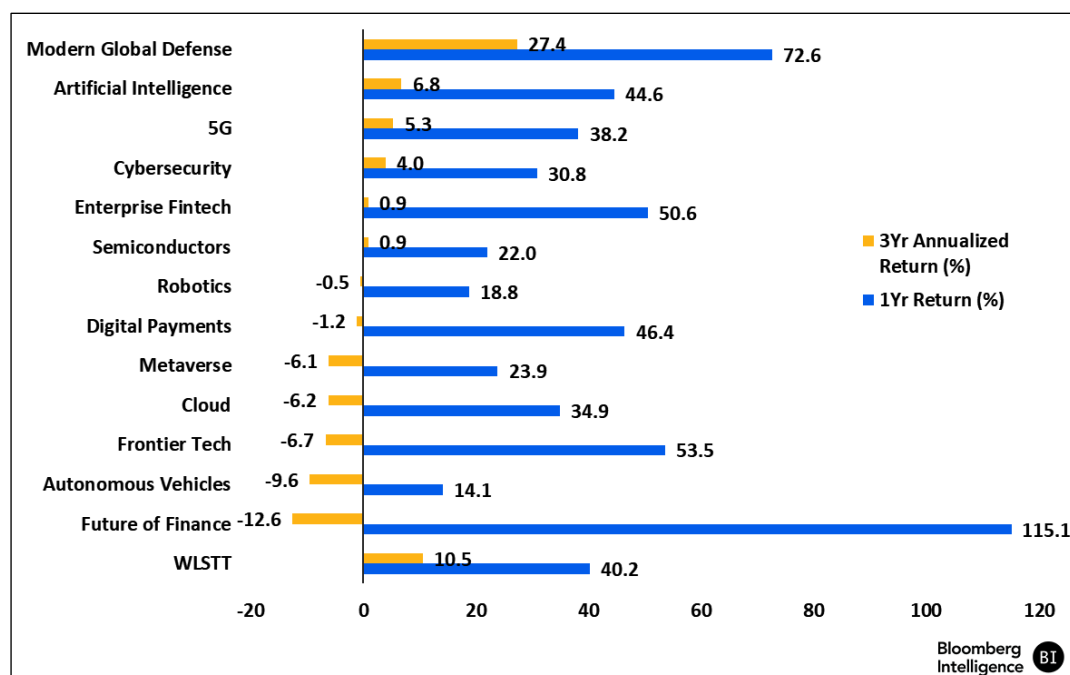
Artificial-intelligence companies appear set to contribute a greater share of S&P 500 total earnings in 2024 than tech, media and telecom did in the late 1990s. Within the Bloomberg AI Aggregate EW Index (BAIAT Index), based on the BI AI theme universe, pick-and-shovel exposures (hyperscalers, semiconductors and hardware) were the earliest tailwind beneficiaries, though as AI is increasingly integrated, we're seeing support broaden to other components as well. With median price-to-sales above its five-year 75th percentile, it's among the few BI themes holding a higher relative valuation heading into 2025. Premium might be warranted as the theme fuels advances in both share prices and valuation multiples across the technology spectrum, from software to hardware, networking, services and more.

11.1 Performance: Maturing to Secular Growth Opportunity

Of BI's 13 accelerating tech themes, AI ranks second to only modern global defense on three-year annualized returns and among the top half of those themes on one-year returns. As proliferation of AI solutions persists, business models across industries stand to see disruption making AI arguably the most disruptive theme in our lineup. One of the most challenging differentiations to make when an investment theme gains traction though is point-in-time enthusiasm versus emerging secular growth opportunity. While stock price performance is one of the most turned to metrics, and AI holds up well broadly on those metrics, for AI, on our radar are fundamental metrics that we see pointing to staying power. In the BI AI theme, hyperscaler and semiconductors exposures have aggregate year-ahead net income growth of 40% and 20%. Notably, BI names that have greater exposure (sum assessment of 2 and 3) show aggregate year-ahead net income growth of 25%, compared with 14% for medium exposure and 11% for low exposure (sum assessment of 4 or 5). For context, the S&P 500 has aggregate year-ahead net income growth of just 11%.

One of AI's most powerful implications is that its proliferation stands to extend its impact to adjacent themes such as 5G, cloud, cybersecurity and semiconductors. AI and AI-adjacent themes, thus, rank at the top of BI's near-term theme radar. AI and its adjacent themes have robust overlap, with 19 names in three or more of those themes (strong AI-adjacent theme breadth stocks). Using AI as the epicenter, most strong theme breadth stocks are seen in cloud and cybersecurity. Broadcom and Cisco appear in four or more themes, making them unique. Broadcom is in AI, 5G, cloud, cyber and semis; Cisco is in AI, 5G, cloud and cyber.

Figure 58: Accelerating Tech Themes Performance



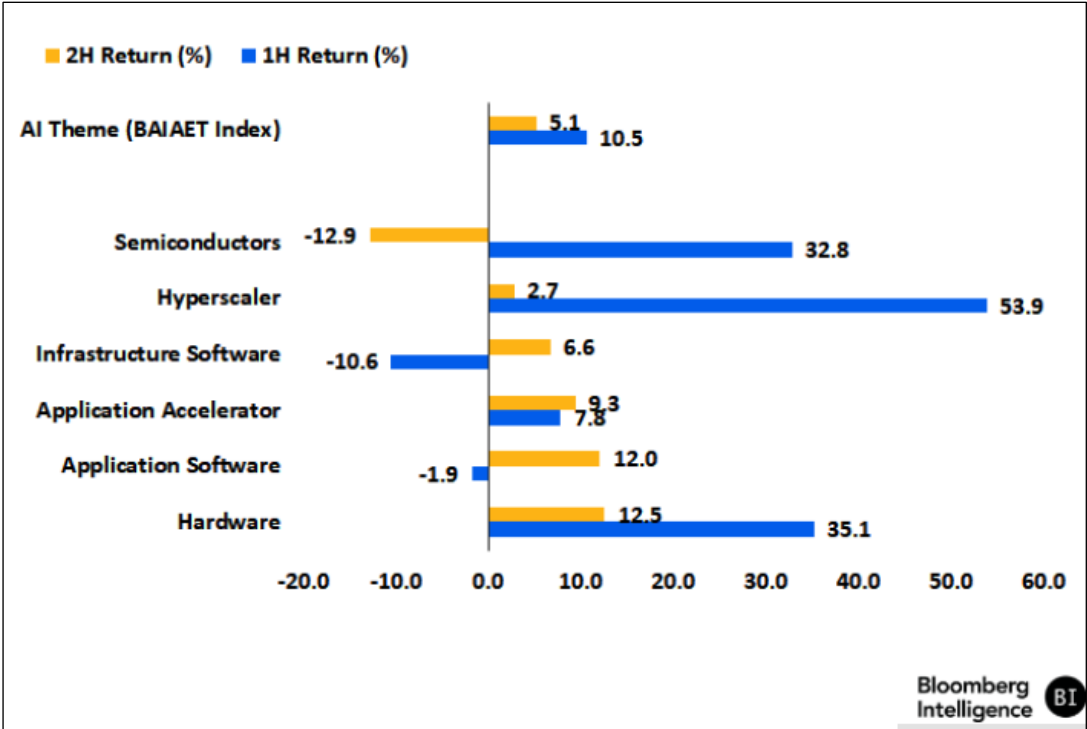
Source: Bloomberg Intelligence

11.2 Valuation: Estimates Climb as Product Horizon Broadens

There's been significant multiple expansion for specific semiconductor companies, with Nvidia again leading the way, with the most visible impact on sales growth expectations from AI.

Momentum in top-line estimate revisions could hinge on the pace of product releases such as Microsoft's release of its GitHub and Office copilots. Alphabet has recently released its Gemini LLM and Duet AI copilot while Meta has open-sourced its Llama foundational models to spur adoption. Database and infrastructure software companies such as Oracle, Snowflake, MongoDB and Databricks have continued to ramp up their offerings with vector search capabilities that could stand to benefit from large amounts of data used for training LLMs, which could help drive positive revisions to consensus. However, the benefits that accrue may not be balanced across the players.

Figure 59: Accelerating Tech Themes Performance



Source: Bloomberg Intelligence

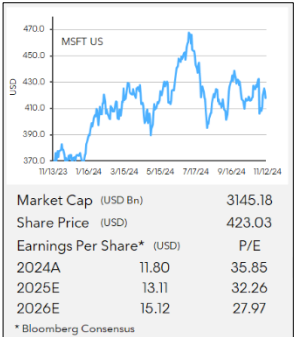
AI is the most held theme in the SPX by number of stocks, and BAIAT Index’s current median forward P/S, which is above its five-year interquartile range, is the fourth strongest in BI’s accelerating tech category that holds 13 themes. There’s high P/S dispersion across AI’s exposure categories, which was most evident from June 2020 to December 2021. The P/S range across exposures is 67% tighter now than at the maximum dispersion. At their December 2020 peak, infrastructure software (22x) and applications (15.6x). Hyperscalers (54%) and hardware (35%) have the strongest 2024 returns, with the former leading in P/S (8.1x) and the latter at the bottom (1.4x); hardware has consistently had the lowest AI P/S.

Section 12. Company Impacts

Seeking a Slice of Generative's AI \$1.6 Trillion Pie

With a projected \$1.6 trillion in spending by 2032, generative AI's effects will ripple through every industry in the technology sector. Here's a look at how some companies are poised to gain over the coming decade.

12.1 Microsoft Among Best Positioned Software Makers



\$6-9 Billion
Gen AI sales impact in 2024

75%
GitHub Copilot adoption

Company Outlook: Microsoft's vast array of software applications makes it a key beneficiary of the growing digital transition as companies upgrade legacy IT systems. A foothold in cloud infrastructure, coupled with its close relationship with OpenAI, puts Microsoft in a strong position to capitalize on rising demand for generative AI.

AI Impact: Microsoft is better positioned than most software companies to capitalize on the increased adoption of generative AI, given a first-mover advantage from its tie-up with OpenAI, and we calculate this could result in \$6-\$9 billion in generative AI revenue in calendar year 2024. It's the first large company to launch AI copilots across its product portfolio, from Office to GitHub. Azure, a cloud-infrastructure product, most likely will be Microsoft's main beneficiary of increased AI demand in the long term. Not only does ChatGPT run on Azure, but Microsoft is also making OpenAI's LLMs available on the platform. Search is another area where we believe Microsoft could gain market share steadily over time.

12.2 Amazon Can Gain From Training, Inference, Creative



100-200 Bps

AWS growth from AI

Company Outlook: Amazon.com's push for speed, convenience and value, coupled with building momentum across retail, cloud and ads, position it well to deliver on its 1Q plan. Cloud-services growth could accelerate in 2024 with margins also expanding. Improving IT budgets and companies' greater willingness to shift infrastructure to the public cloud remain catalysts for AWS in the longer run. Operating margin may continue to expand on cost cuts, efficiencies and rising contribution from cloud and ads. Amazon's push for pharmacy and grocery are large undertakings that we'll monitor closely

AI Impact: AWS should see its fair share of new generative AI workloads from training and inference to creating new applications through its Bedrock offering. Like Microsoft's GitHub, AWS also is offering a generative AI-embedded software development product called CodeWhisperer that significantly shortens the time it takes a developer to code. AWS has the largest market share in cloud infrastructure as a service and more than 100,000 customers using its other AI and machine learning services. Though AWS hasn't combined with OpenAI LLMs yet, it does provide its own foundational models in addition to working closely with other providers, such as Anthropic, Stability AI and AI21.

12.3 Adobe to Parlay Base of 70 Million Creative Professionals



\$0.9-\$1.4 Billion

Gen AI sales gain over 2-3 years

25%-Plus

DF copilot adoption Year 1

100 Bps

Operating margin expansion in 2024

Company Outlook: Adobe's solid portfolio of digital products could lead to organic sales growth of 12-15% in constant currency over the next three years, along with non-GAAP operating margin of about 45%. We see the company as among the key beneficiaries of increased spending on digital transformation, given its focus on data insights, digital commerce, marketing and content creation. Generative AI product Firefly could help drive up average revenue per user, as Adobe differentiates itself with built-in copyright safeguards for immediate commercial use, amid increasing competition from providers like OpenAI.

AI Impact: Adobe's installed base of around 70 million creative professionals, the highest market share in this category, positions it well to reap benefits from generative AI. The recently launched Firefly creative copilot can substantially reduce the time it takes to create images via text and has already helped create more than a billion visuals through generative fill. The troves of data housed in Adobe's creative cloud suite, which includes Photoshop and Illustrator, make it better positioned than rivals to train its LLMs, and rights to the underlying training content in Adobe Stock can also provide creators with legal peace of mind. Gen AI advancements in digital documents as well as its enterprise front-office software can also aid productivity and could lead to pricing improvements.

12.4 Alphabet Leveraging AI Across Product Suite



\$3-\$4 Billion

Boost to Google Cloud

10-15%

Search queries to leverage AI

5-10%

YouTube ad engagement

Company Outlook: Alphabet's improved top-line growth for core search and YouTube segments appear sustainable for the rest of 2024, while demand for generative AI computing bodes well for its Cloud segment profitability. Though Bing-ChatGPT is a risk to the ad business, we believe the company's coming launch of a multimodal large language model and integration of generative AI features into its core search and YouTube products have eased near-term competitive pressure. Network ad sales face pressure amid the removal of cookies, though YouTube ads and subscriptions could see double-digit growth in 2H.

AI Impact: Alphabet is exposed to most segments of the generative AI market, including training, inference and digital ads. Gemini is already being used across the company's offerings, including ad targeting algorithms, Google Cloud Vertex AI, and the Google Pixel 8, the first smartphone with a native generative AI assistant. Cloud AI tools may offer monetization opportunities, while improvements to targeting algorithms can provide a lift to ad pricing. A potential partnership with Apple will further aid Gemini's positioning in the nascent gen AI inferencing market. Though many retailers are rolling out their own enhanced search offerings with longer prompts, the probability of Alphabet losing consumer traffic is relatively low given users' preference for Google's search

12.5 Meta's Llama for Recommendations Draw Enterprises



5-10%

Effect on engagement, impressions growth

\$1-\$2 Billion

LLM licensing sales bump

\$10 Billion

Click-to-message ad run rat

Company Outlook: Meta may continue to see benefits to user engagement from pivoting to AI-based recommendations, driving stronger impressions growth across its family of apps. The company has leveraged its Llama model to improve ad targeting algorithms while also launching new products like AI Studio and Meta Assistant which can aid monetization efforts. Reels, a high-single-digit contributor to sales, and click-to-messaging ads are both above \$10 billion in revenue run rate. User growth will be driven mostly by Instagram and WhatsApp, while Reels could emerge as a contributor to ad pricing this year. Free cash flow may remain squeezed by a slightly higher capex guide and potentially a \$20 billion annual loss from Reality Labs in 2024

AI Impact: Meta's scale in running its own data-center infrastructure, coupled with large amounts of training data from its family of apps, has allowed it to build its own foundational LLM, Llama, to compete with offerings from OpenAI, Alphabet, and others. Llama's open-source nature may be more attractive for enterprises looking to build their own functionality on top of the model. The pace at which new content is created for social media and virtual- and augmented-reality applications for the metaverse could speed up with generative AI. Meta may also implement personalized shopping assistants to boost user adoption of social commerce, increasing monetization opportunities. Our analysis suggests that the generative AI market may add almost \$207 billion in ad spending through 2032 with time spent on platforms, plus ad targeting and personalization.

12.6 OpenAI Supported by Foundational LLM Adoption

\$157 Billion

Valuation

\$3.7 Billion

Annual Sales

Company Outlook: OpenAI recently raised \$6.6 billion for a \$157 billion valuation, putting a focus on licensing sales. OpenAI generates about 70% of its \$3.7 billion in annual sales run rate from ChatGPT subscriptions, which include 11 million subscribers on its ChatGPT Plus plan and another 1 million on higher-priced plans. OpenAI generates API revenue from both enterprise and individual application programming interface calls, as well as consumption through Azure AI.

AI Impact: OpenAI dominant position in generative AI remains aided by its foundational LLM adoption across both enterprise and consumer applications. The company's GPT-4 model came out as the top choice in the Bloomberg Intelligence enterprise CIO survey. OpenAI has also released its o1 model that can reason at the time of inferencing, which other foundational model companies are likely to follow. ChatGPT subscription uptake far exceeds that of rivals Google Gemini and Anthropic Claude aided by higher engagement of the former's chatbot offering.

12.7 Anthropic Leverages Partnerships to Post Gains

\$40 Billion

Valuation

\$500 Million

LLM License Sales

Company Outlook: Anthropic's potential \$40 billion valuation in a new private funding round, based on the Information, suggests the company remains among the leading AI native companies. Hyperscaler distribution has so far aided Anthropic's API sales, at about a \$500 million run rate given the company's alliances with Amazon and Google.

AI Impact: Anthropic's high performance of its Claude LLM across various benchmarks has aided its positioning vs leading rivals including OpenAI, Google and Meta, which all have their own foundational model. Anthropic has partnered with Amazon and Alphabet to deploy its foundational model on public clouds while also sourcing compute for training the next version of its LLM.

Section 13. Methodology

Bloomberg Intelligence's interactive market-sizing model helps provide growth forecasts for generative AI. This model, which will evolve on a regular basis, is still in its early stages, and we have provided Terminal clients an interactive calculator to test scenarios (available at BI INET <GO>).

The methodology is based on a bottom-up approach to forecast revenue for areas within hardware, software, digital ads, gaming, IT and business markets. Our forecasts for new segments are anchored to established end-markets that generative AI is likely to disrupt and create new revenue opportunities. The approximate calculations are driven by our assumptions around how generative AI could disrupt these existing end-markets to varying degrees. Figure 60 illustrates the existing end markets with growth assumptions for 2022-27 and 2027-32. Figure 61 shows BI's base-case penetration rates of new AI segments across these end-markets. Figure 62 depicts BI's base-case generative AI revenue forecasts, driven by AI penetration rates.

Figure 60: Revenue and Growth Forecasts for Existing Technology End-Markets

Worldwide Revenue by Technology Group					
(in billions of \$)	2023	'23-'28 CAGR(%)	2028E	'28-'32 CAGR(%)	2032E
Hardware	\$1,373	6%	\$1,847	10%	\$2,747
Devices	\$938	3%	\$1,086	4%	\$1,270
Infrastructure	\$436	12%	\$761	18%	\$1,476
Software	\$996	13%	\$1,859	13%	\$3,040
Application Development & Deployment	\$237	19%	\$572	17%	\$1,062
PaaS	\$123	28%	\$426	22%	\$943
On-premise	\$114	5%	\$147	-5%	\$119
Applications	\$516	11%	\$886	10%	\$1,305
SaaS	\$297	16%	\$636	14%	\$1,074
On-premise	\$219	3%	\$250	-2%	\$231
System Infrastructure Software	\$243	11%	\$401	14%	\$673
IaaS	\$113	17%	\$248	22%	\$548
On-premise	\$129	3%	\$153	-5%	\$125
IT Services	\$834	5%	\$1,077	5%	\$1,332
Business Services	\$395	5%	\$499	5%	\$616
Digital Ad Spending	\$603	10%	\$993	12%	\$1,571
Search	\$247	10%	\$399	10%	\$585
Display	\$333	11%	\$563	14%	\$951
Classified and Other	\$23	6%	\$30	4%	\$36
Cybersecurity Spending	\$103	13%	\$191	12%	\$301
Gaming Spending	\$261	5%	\$326	8%	\$444
Life Sciences Spending	\$162	11%	\$271	11%	\$412
Education Technology Spending	\$123	14%	\$233	14%	\$389
Total	\$4,850	9%	\$7,298	10%	\$10,852

Source: Bloomberg Intelligence's forecasts based on data from IDC, eMarketer, and Statista

The hardware market is currently valued at \$1.37 trillion, split into devices (\$938 billion) and data-center infrastructure (\$436 billion), based on IDC data. Bloomberg Intelligence expects new segments for generative AI in this category are AI servers, AI storage, Training Compute, and Networking on the training side, and conversational AI and computer vision products on the inference devices side. The assumptions around the shift in spending to AI servers and storage from traditional servers and storage can be changed in the BI interactive calculator, available on the Terminal. Generative AI infrastructure-as-a-service is how we expect the training compute and storage capacity to be consumed on the cloud. Similarly, for inferencing, conversational AI products and computer vision should emerge as new categories in devices that could be used at home and in cars. For the \$996 billion software market, we anticipate new categories to emerge, such as coding copilots, specialized virtual assistants, chatbots and drug discovery software.

Figure 61: Generative AI Penetration Rate (BI Base-Case Assumptions)

AI Penetration (%) Base-Case Assumptions			
	2023	2028E	2032E
Hardware			
Devices (Inference)			
Computer Vision AI Products	0.3%	2.5%	5.0%
Conversational AI Products	0.4%	6.5%	9.0%
Infrastructure (Training)			
AI Server	13.0%	25.0%	21.5%
AI Storage	2.5%	4.5%	4.0%
Training Compute	0.7%	3.5%	7.5%
Networking	0.8%	2.5%	2.5%
Inference/Fine-Tuning Cloud	0.5%	6.5%	8.2%
Software			
Coding, DevOps and Generative AI Workflows Software			
Copilot Spending	0.1%	2.4%	3.5%
Customer Service Chatbots	0.1%	1.6%	3.5%
Specialized Generative AI Assistants Software			
E-Commerce	0.1%	0.4%	0.9%
Social Media	0.0%	0.4%	1.0%
Generative AI Workload Infrastructure Software			
	0.5%	5.0%	12.0%
Drug Discovery Software			
	0.0%	5.0%	10.0%
Cybersecurity Spending			
	0.2%	6.0%	8.0%
Copilot Spending			
	0.0%	2.8%	4.7%
Data Protection			
	0.1%	3.2%	3.3%
Education Spending			
	0.6%	3.0%	6.0%
Copilot Spending			
	0.5%	2.7%	5.4%
Educational Content Creation			
	0.1%	0.3%	0.6%
Image/Video Generation Tools			
	0.1%	0.5%	1.0%
Film Production/Music Generation Tools			
	0.0%	0.1%	0.2%
Gaming Spending			
Virtual Goods			
	0.1%	3.0%	6.0%
Game Design Software			
	0.2%	5.0%	10.0%
IT Services			
	0.0%	4.0%	6.0%
Business Services			
	0.0%	3.0%	5.0%
Digital Ad Spending			
Search			
	1.0%	7.0%	12.0%
Videos			
	0.5%	5.5%	10.6%
Messaging			
	0.2%	2.0%	4.0%

Source: Bloomberg Intelligence

13.1 BI's Market-Sizing Conclusions

The 2022-32 forecast scenario for the generative AI market is based on the size of the end-markets and penetration. The CAGR assumptions for both-end markets and generative AI impact can be modified to come up with your own scenarios. For example, we assume in our base case that the data-center market ("Infrastructure" in Figure 60) is likely to expand at a 12% CAGR from 2022-28 and 18% from 2028-32. The training segment, which is tied to the data-center market, could hit a CAGR of 34% for 2028 and 27% for 2032, in our scenario. The training segment is further comprised of AI Server, AI Storage, Training Compute, Networking, and LLM Licensing. Inferencing includes devices and software applications, such as Chatbots, Copilots, etc.

Figure 62: Generative AI Revenue Base-Case Forecast by Technology Segment

(in millions of \$)	2023	2028E	2032E	Implied 9 yr CAGR (%)
Infrastructure (Training)	\$75,519	\$329,393	\$645,962	27%
AI Servers	\$56,642	\$190,373	\$317,419	21%
AI Storage	\$10,893	\$34,267	\$59,055	21%
Training Compute (Cloud Workloads)	\$3,050	\$26,652	\$110,728	49%
Networking	\$3,268	\$19,037	\$36,909	31%
LLM Licensing Revenue	\$1,667	\$59,063	\$121,852	61%
Devices and Applications (Inference)	\$10,451	\$209,159	\$486,651	53%
Inference/Fine-Tuning (Cloud Workloads)	\$2,179	\$49,497	\$121,062	56%
Drug Discovery Software	\$32	\$13,571	\$41,203	121%
Computer Vision AI Products	\$2,813	\$27,147	\$63,517	41%
Conversational AI Products	\$3,750	\$70,583	\$114,330	46%
Chatbots/Copilots	\$1,421	\$41,545	\$134,282	66%
Enterprise Applications	\$257	\$19,085	\$51,315	80%
Cybersecurity Copilots	\$21	\$5,351	\$14,133	107%
Coding and Devops Copilots	\$237	\$13,734	\$37,181	75%
Consumer Applications	\$1,163	\$22,459	\$82,967	61%
Educational Copilots	\$617	\$6,303	\$20,995	48%
E-Commerce Copilots	\$258	\$3,278	\$11,743	53%
Social Media Chatbots	\$52	\$3,722	\$13,048	85%
Customer Service Chatbots	\$237	\$9,156	\$37,181	75%
Educational Content Creation	\$123	\$700	\$2,333	39%
Data Protection	\$134	\$6,115	\$9,923	61%
Generative AI Driven Ad Spending	\$4,638	\$70,179	\$208,987	53%
Search	\$2,472	\$27,959	\$70,174	45%
Videos	\$1,667	\$30,962	\$100,783	58%
Messaging	\$500	\$11,259	\$38,031	62%
IT Services	\$167	\$43,072	\$79,926	99%
Business Services	\$79	\$14,970	\$30,782	94%
Workload Monitoring Software	\$1,214	\$20,046	\$80,797	59%
Entertainment/Media	\$831	\$31,248	\$86,700	68%
Gaming	\$522	\$26,109	\$71,042	73%
Virtual Goods	\$130	\$9,791	\$26,641	81%
Game Design	\$391	\$16,318	\$44,401	69%
Image/Video Generation Tools	\$258	\$4,430	\$13,048	55%
Film Production/Music Generation Tools	\$52	\$709	\$2,610	55%
Total	\$92,901	\$718,067	\$1,619,805	37%

Source: Bloomberg Intelligence

Section 14. Glossary of Terms

These can help decipher highly technical elements:

Advanced RISC Machines (ARM): A processor architecture based on 32-bit reduced instruction set computer.

AI Assistants: A software agent that can perform tasks for a user based on input such as commands or questions. Think Siri or Cortana.

AI Server: Computers used for AI inferencing and training.

AI Storage: Often a software-as-a-service application that performs analysis in a public cloud.

ChatGPT: A free chatbot that can answer just about any question.

Conversational User Interface: Allows people to interact with software, apps and bots as they would another human being. Amazon's Alexa is an example.

Computer Vision: Enables computers and systems to derive meaningful information from digital information, videos and other visual inputs, then act or make recommendations based on that information.

Corpus: Literally "body," this is the collection of billions of data points used to train a large language model.

CPU: Central Processing Unit. Basically, the semiconductor chip that is the essential logic circuitry in a hardware system.

Edge: Deployment of computing and storage resources closer to where data is produced.

Ethernet: Technology to connect devices in a local area network (LAN) or wide area network (WAN). Slower than InfiniBand.

Generative AI: Uses algorithms, such as ChatGPT, to create new content including audio, code, images, text and videos.

GPU: Graphics Processing Unit. A specialized circuit for image and video display.

Hallucination: A response/output from a large language model that is irrelevant or incorrect

Inference: The process of reasoning and making decisions based on available information or data. It follows training to derive new knowledge or conclusions from existing data.

InfiniBand Network: A high-performance, low-latency way to facilitate high-speed communications. Faster than Ethernet

IaaS: Infrastructure as a Service, a business model that offers computing, storage and networking resources on demand.

Large Language Models: Deep learning algorithms that can recognize, summarize, translate, predict, and generate content using massive datasets.

Machine Learning: The use and development of computer systems that learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

Neural Network: A method that teaches computers to process data the way the brain does. It's a type of machine-learning that uses interconnected nodes in a layered structure.

ODM: Original Design Manufacturer.

PaaS: Platform as a Service. A type of cloud computing service model that offers a flexible, scalable platform to develop, deploy, run and manage apps.

Personalization: Uses AI and machine learning to analyze a user's data, helping to understand their needs and tailor their experience accordingly.

Retrieval-Augmented Generation (RAG): A method by which a company can supply its proprietary data to a model that can append relevant information to a user query prior to feeding it into a LLM

Training: The process of teaching an AI system to perceive, interpret and learn from data. That way, the AI will later be capable of inferencing—making decisions based on information it's provided.

Bloomberg Intelligence Research Coverage

Bloomberg Editorial and Research:

John Micklethwait, Editor-in-Chief; **Reto Gregori**, Deputy Editor-in-Chief

Research Management

David Dwyer, Global Director of Research
Drew Jones, Deputy Global Director of Research
Sam Fazeli, Director of Global Industry Research
Alison Williams, Director of Global Strategy Research

Paul Gulberg, Director of Americas Industry Research
Catherine Lim, Director of APAC Industry Research
Sue Munden, Director of EMEA Industry Research
Frank Jacobs, Global Chief Operating Officer

Content Management

Tim Craighead, Global Chief Content Officer
Karima Fenaoui, Research Content Officer, Communities & EM
John Lee, Research Content Officer, APAC
Renato Prieto, Research Content Officer, FICC
Roger Thomson, Research Content Officer, Americas
Rod Turnbull, Research Content Officer, EMEA

Mariam Traore, Research Digital Content Specialist
Matthew Bloxham, Global Head of Alternative Data & Analytics
Brian Egger, Global Head of Financial Modeling
Donna Weston, Co-head Global Research Editorial
Douglas Zehr, Co-head Global Research Editorial

Equity Strategy

Gina Martin Adams, Director of Equity Strategy Research

Markets

Christopher Cain, Quantitative Analysis, US
Michael Casper, Small Caps and Sectors, US
Marvin Chen, China and North Asia
Nitin Chanduka, Emerging Markets, India
Laurent Douillet, Europe
Anthony Feld, Technical Analysis, Global
Kumar Gautam, Emerging Markets, Global
Gina Martin Adams, Global
Wendy Soong, Americas
Sufianti Sufianti, Emerging Markets, ASEAN
Gillian Wolff, Global

Funds

Eric Balchunas, Exchange Traded Funds, Global
David Cohn, Mutual Funds, Global
Henry Jim, Exchange Traded Funds, Europe
Athanasios Psarofagis, Exchange Traded Funds, Americas
James Seyffart, Exchange Traded Funds, Americas
Rebecca Sin, Exchange Traded Funds, APAC

Thematic

Breanne Dougherty, Global
Andrew Silverman, Tax Policy and Corporate Actions, Global
Shirley Wong, APAC
Lu Yeung, Accounting, Global

FICC Strategy

Noel Hebert, Director of FICC Strategy Research

Markets

Erica Adelberg, MBS, Americas
Negisa Balluku, Litigation-Bankruptcy, Americas
Mahesh Bhimalingam, Credit Strategy, EMEA
Philip Brendel, Distressed Debt, Americas
Rod Chadehumbe, ABS, Americas
Sam Geier, Credit Strategy, Americas
Noel Hebert, Credit Strategy, Americas
Eric Kazatsky, Municipals, Americas
Mike McGlone, Commodity Strategy, Global
Tanvir Sandhu, Derivatives, Global
Damian Sassower, Emerging Markets, Global
Timothy Tan, Credit Strategy, APAC

FX and Rates

Audrey Childe-Freeman, FX, G-10
Stephen Chiu, FX and Rates, APAC
Ira Jersey, Rates, US
Davison Santana, FX and Rates, LatAm
Sergei Voloboev, FX, Emerging Markets, Global
Huw Worthington, Rates, EMEA

ESG Research

Adeline Diab, Director of ESG Strategy Research

Strategy

Manish Bangard, Americas
Adeline Diab, Global
Ortis Fan, APAC
Yasutake Homma, Japan
Rahul Mahtani, Quantitative Analysis, EMEA
Christopher Ratti, Fixed Income, Global

Eric Kane, Director of ESG Company Research

Company and Industry

Shaheen Contractor, Industry, Global
Rob Du Boff, Governance, Global
Gail Glazerman, Integration, Americas
Eric Kane, Global
Grace Osborne, Integration, EMEA
Andrew John Stevenson, Climate, Global
Conrad Tan, Integration, APAC

Market Structure Research

Larry Tabb, Director of Market Structure Research

Jackson Gutenplan, Equities, Americas

Brian Meehan, Fixed Income, Americas

Nicholas Phillips, Equities, EMEA

Larry Tabb, Equities and Fixed Income, Global

Credit Research

Joel Levington, Director of Credit Research

Americas

Himanshu Bakshi, Consumer Finance, Banking, Global
Mike Campellone, Specialty Apparel, Consumer Hardlines, Global
Cecilia Chan, Gaming Lodging & Restaurants, Internet Media, China
Jean-Yves Coupin, Health Care, Corporate Bonds, Americas
Spencer Cutter, Oil & Gas, Global
Stephen Flynn, Entertainment, Cable & Satellite, Americas
Matthew Geudtner, Aerospace, Global, Machinery, Americas
David Havens, Consumer Finance, Investment Mgmt, Americas
Mike Holland, Hospitals, Specialty-Generic Pharma, Americas
Julie Hung, Packaged Food, Beverages, Americas
Arnold Kakuda, Investment Banking, Americas
Joel Levington, Automobiles, Global, Industrials, Americas
Jody Lurie, Gaming Lodging & Restaurants, Americas
Robert Schiffman, Hardware & Storage, Internet Media, Global

Asia

Andrew Chan, Real Estate, Infrastructure, China
Sharon Chen, Telecom Carriers, Infrastructure, India
Pri De Silva, Banking, Aerospace & Defense, APAC
Daniel Fan, Real Estate, China
Rena Kwok, Banking, APAC
Mary Ellen Olson, Metals & Mining, Oil & Gas, APAC

EMEA

Tolu Alamutu, Real Estate, Banking, EMEA
Ruben Benavides, Banking, Europe
Aidan Cheslin, Telecom Carriers, EMEA
Jeroen Julius, Banking, EMEA
Stephane Kovatchev, Industrials, Construction, EMEA, Americas
Paul Vickars, Electric Utilities, EMEA, Oil & Gas, Global

Industry Research

Consumer

Consumer Products & Services

Brian Egger, Gaming & Lodging, Americas
Angela HanLee, Gaming & Hospitality, APAC
Drew Reading, Homebuilders, Americas
Deborah Aitken, Luxury, Personal Care Products, Global
Michael Halen, Restaurants, Americas
Conroy Gaynor, Travel & Leisure, EMEA

Retail & Wholesale

Lindsay Dutch, Consumer Hardlines, Retail REIT, Americas
Tatiana Lisitsina, Consumer Hardlines, Online Apparel, EMEA
Poonam Goyal, E-Commerce, Specialty Apparel Stores, Americas
Catherine Lim, Consumer Goods, E-Commerce, APAC
Charles Allen, Retail Staples & Wholesale, Specialty Apparel, EMEA
Mary Ross Gilbert, Specialty Apparel Stores, Americas
Abigail Gilmartin, Athleisure & Footwear, Americas

Consumer Staples

Jennifer Bartashus, Packaged Food and Retail Staples, Americas
Duncan Fox, Beverages, Packaged Food, EMEA
Kenneth Shea, Beverages, Tobacco & Cannabis, Americas
Diana Gomes, Consumer Health, Household Products, Global
Lisa Lee, Consumer Goods, Health Care, APAC
Ada Li, Consumer Goods, Online Health Care, APAC
Diana Rosero-Pena, Packaged Food, Retail Staples, Americas
Lea El-Hage, Retail Staples, Australia

Energy

Will Hares, Energy Sector Head

Brett Gibbs, Biofuels, EMEA
Talon Custer, Oil & Gas, Americas
Henik Fung, Oil & Gas, Gas Utilities, APAC
Will Hares, Oil & Gas, EMEA
Scott J. Levine, Oil & Gas, Industrials, Americas
Vincent G. Piazza, Oil & Gas, Americas
Salih Yilmaz, Oil & Gas, EMEA
Rob Barnett, Solar Energy Equipment, Americas, EMEA

Utilities

Patricio Alvarez, Electric Utilities, Gas Utilities, EMEA
Nikki Hsu, Electric Utilities, Americas
Kelvin Ng, Utilities, Coal, APAC
Gabriela Privetera, Electric Utilities, Americas

Industrials

Steve Man, Director of Industrial Research

Automotive

Joanna Chen, Automobiles, APAC
Gillian Davis, Automobiles, EMEA
Michael Dean, Automobiles, EMEA
Steve Man, Automobiles, Americas
Tatsuo Yoshida, Automobiles, Japan

Industrial & Industrial Services

George Ferguson, Aerospace & Defense, Global
Will Lee, Aerospace & Defense, Americas
Wayne Sanders, Defense, Americas
Stuart Gordon, Business Services, Europe
Christopher Ciolino, Machinery, Industrials, Americas
Christina Feehery, Industrials, Americas
Takeshi Kitaura, Industrials, Japan
Mustafa Okur, Electrical Equipment, Industrials, Americas
Bhawin Thakker, Industrials, EMEA
Karen Ubelhart, Industrials, Machinery, Americas
Omid Vaziri, Industrials, EMEA
Denise Wong, Infrastructure, APAC

Transportation

Tim Bacchus, Airlines, APAC
Francois Duflot, Airlines, Americas
George Ferguson, Airlines, Global
Conroy Gaynor, Airlines, EMEA
Lee Klaskow, Freight Transportation & Logistics, Global
Kenneth Loh, Marine Shipping, Logistics Services, APAC

Materials

Jason Miner, Agriculture Sector Head
Grant Sporre, Metals and Mining Sector Head

Chemicals

Daniel Cole, Agricultural Chemicals, Americas
Alexis Maxwell, Agricultural Chemicals, Canada
Alvin Tai, Agriculture, Malaysia, EMEA
Jason Miner, Agriculture and Chemicals, Americas
Sean Gilmartin, Specialty Chemicals, Americas
Vivien Zheng, Specialty Chemicals, APAC

Metals & Mining

Richard Bourke, Basic Materials, Americas

Financials

Mohsen Crofts, Metals & Mining, Real Estate, Australia

Financial Services

Ben Elliott, Consumer Finance, Americas
Edmond Christou, Financials, Middle East
Diksha Gera, Global Payments and Fintech, Americas
Salome Skhirtladze, Financials, Middle East

Financial Services (Cont'd)

Sarah Jane Mahmud, Banking, Market Structure, ASEAN and India
Alison Williams, Investment Banking, Global
Sharnie Wong, Investment Banking, Exchanges, APAC
Paul Gulberg, Investment Management, Exchanges, Banks Americas
Ethan Kaye, Investment Management, Americas
Neil Sipes, Investment Management, Investment Banking, Americas
Hideyasu Ban, Financial Services, Japan
Matt Ingram, Financial Services, Australia and Korea

Banking

Francis Chan, Banking & Fintech, China & Hong Kong
Herman Chan, Banking, Americas
Tomasz Noetzel, Banking, EMEA
Philip Richards, Banking, EMEA
Lento Tang, Banking, EMEA
Maryana Vartsaba, Banking, EMEA

Insurance

Steven Lam, Insurance, APAC
Jeffrey Flynn, Life Insurance, Americas
Kevin Ryan, Life Insurance, P&C Insurance, EMEA
Charles Graham, P&C Insurance, Life Insurance, EMEA
Matthew Palazola, P&C Insurance, Americas

Real Estate

Jack Baxter, Real Estate, Australia
Ken Foong, Real Estate, Singapore
Iwona Hovenko, Real Estate, Business Services, EMEA
Kristy Hung, Real Estate, China
Jeffrey Langbaum, Residential REIT, Office REIT, Americas
Sue Munden, Real Estate, EMEA
Patrick Wong, Real Estate, APAC

Health Care

Aude Gerspacher, Director of Health Care Research

Biotech & Pharma

Sam Fazeli, Biotech, Global
Grace Guo, Biotech, Pharma, Americas
Jean Rivera Irizarry, Biotech, Pharma, Americas
Jamie Maarten, Biotech, China
Max Nisen, Biotech, Americas
Michael Shah, Biotech, Specialty-Generic Pharma, EMEA
Leslie Yang, Biotech, China
John Murphy, Large Pharma, Biotech, Americas, EMEA
Aude Gerspacher, Pharma, Biotech, Americas
Justin Kim, Biotech, Pharma, Americas
Ann-Hunter Van Kirk, Specialty-Generic Pharma, Americas
Glen Losev, Hospitals, Managed Care, Americas
Jonathan Palmer, Medical Devices, Supply Chain, Americas
Matt Henriksson, Medical Equipment & Devices, Americas

Michelle Leung, Metals & Mining, China, Japan
Emmanuel Munjeri, Metals & Mining, South Africa
Alon Olsha, Metals & Mining, EMEA
Grant Sporre, Metals & Mining, EMEA

Construction Materials

Sonia Baldeira, Construction, Building Materials, EMEA
Kevin Kouam, Building Materials, Global

Technology

Mandeep Singh, Director of Technology Research

Hardware

Woo Jin Ho, Hardware & Networking, Americas
Ken Hui, Semiconductors, Europe
Charles Shum, Semiconductors, APAC
Jake Silverman, Logic ICs, Americas
Kunjan Sobhani, Logic ICs, Americas
Steven Tseng, EMS/ODM, Consumer Electronics, APAC
Masahiro Wakasugi, Semiconductors, EMS/ODM, Global

Software

Tamlin Bason, IT Services, Americas, EMEA
Robert Lea, Internet Media, Application Software, China
Nathan Naidu, Entertainment Content, Internet Media, Japan
Niraj Patel, Application Software, Americas
Sunil Rajgopal, Application Infrastructure Software, Americas
Anurag Rana, Application Software, IT Services, Americas
Mandeep Singh, Software, Internet, Hardware/Semis, Americas

Communications

Media

Matthew Bloxham, Media, Advertising, Telecom, EMEA
Geetha Ranganathan, Entertainment, Cable, Advertising, Americas
Tom Ward, Media, Advertising, Telecom Carriers, EMEA

Telecommunications

John Butler, Telecom & Towers, Infrastructure Software, Americas
John Davies, Telecom Carriers and Media, EMEA
Erhan Gurses, Telecom Carriers, EMEA
Marvin Lo, Telecom Carriers, APAC
Chris Muckensturm, Telecom Carriers, ASEAN

Litigation and Policy

Nathan Dean, Financials Policy, Americas
Holly Froum, Consumer, Industrials Litigation and Policy, Americas
Josephine Garban, Health Care Patent Litigation, Americas
Jennifer Rie, Antitrust Litigation and Policy, Americas
Matthew Schettenhelm, TMT Litigation and Policy, Americas
Elliott Stein, Financials Litigation, Americas
Justin Teresi, Antitrust Litigation and Policy, Americas
Duane Wright, Health Care Policy, Americas

Copyright and Disclaimer

Copyright

© Bloomberg Finance L.P. 2024. This publication is the copyright of Bloomberg Finance L.P. No portion of this document may be photocopied, reproduced, scanned into an electronic system or transmitted, forwarded or distributed in any way without prior consent of Bloomberg Finance L.P.

Disclaimer

The data included in these materials are for illustrative purposes only. The BLOOMBERG TERMINAL service and Bloomberg data products (the "Services") are owned and distributed by Bloomberg Finance L.P. ("BFLP") except (i) in Argentina, Australia and certain jurisdictions in the Pacific Islands, Bermuda, China, India, Japan, Korea and New Zealand, where Bloomberg L.P. and its subsidiaries ("BLP") distribute these products, and (ii) in Singapore and the jurisdictions serviced by Bloomberg's Singapore office, where a subsidiary of BFLP distributes these products. BLP provides BFLP and its subsidiaries with global marketing and operational support and service. Certain features, functions, products and services are available only to sophisticated investors and only where permitted. BFLP, BLP and their affiliates do not guarantee the accuracy of prices or other information in the Services. Nothing in the Services shall constitute or be construed as an offering of financial instruments by BFLP, BLP or their affiliates, or as investment advice or recommendations by BFLP, BLP or their affiliates of an investment strategy or whether or not to "buy", "sell" or "hold" an investment. Information available via the Services should not be considered as information sufficient upon which to base an investment decision. The following are trademarks and service marks of BFLP, a Delaware limited partnership, or its subsidiaries: BLOOMBERG, BLOOMBERG ANYWHERE, BLOOMBERG MARKETS, BLOOMBERG NEWS, BLOOMBERG PROFESSIONAL, BLOOMBERG TERMINAL and BLOOMBERG.COM. Absence of any trademark or service mark from this list does not waive Bloomberg's intellectual property rights in that name, mark or logo. All rights reserved. © 2024 Bloomberg.

Bloomberg Intelligence is a service provided by Bloomberg Finance L.P. and its affiliates. Bloomberg Intelligence likewise shall not constitute, nor be construed as, investment advice or investment recommendations, or as information sufficient upon which to base an investment decision. The Bloomberg Intelligence function, and the information provided by Bloomberg Intelligence, is impersonal and is not based on the consideration of any customer's individual circumstances. You should determine on your own whether you agree with Bloomberg Intelligence. Bloomberg Intelligence Credit and Company research is offered only in certain jurisdictions. Bloomberg Intelligence should not be construed as tax or accounting advice or as a service designed to facilitate any Bloomberg Intelligence subscriber's compliance with its tax, accounting, or other legal obligations. Employees involved in Bloomberg Intelligence may hold positions in the securities analyzed or discussed on Bloomberg Intelligence.

About Bloomberg Intelligence

Your go-to resource for making better investment decisions, faster.

Bloomberg Intelligence (BI) research delivers an independent perspective providing interactive data and research across industries and global markets, plus insights into company fundamentals. The BI, team of 475 research professionals is here to help clients make more informed decisions in the rapidly moving investment landscape.

BI's coverage spans all major global markets, more than 135 industries and 2,000 companies, while considering multiple strategic, equity and credit perspectives. In addition, BI has dedicated teams focused on analyzing the impact of government policy, litigation and ESG.

BI is also a leading Terminal resource for interactive data. Aggregated, from proprietary Bloomberg sources and 500 independent data contributors, the unique combination of data and research is organized to allow clients to more quickly understand trends impacting the markets and the underlying securities.

Bloomberg Intelligence is available exclusively for Bloomberg Terminal® subscribers, available on the Terminal and the Bloomberg Professional App.

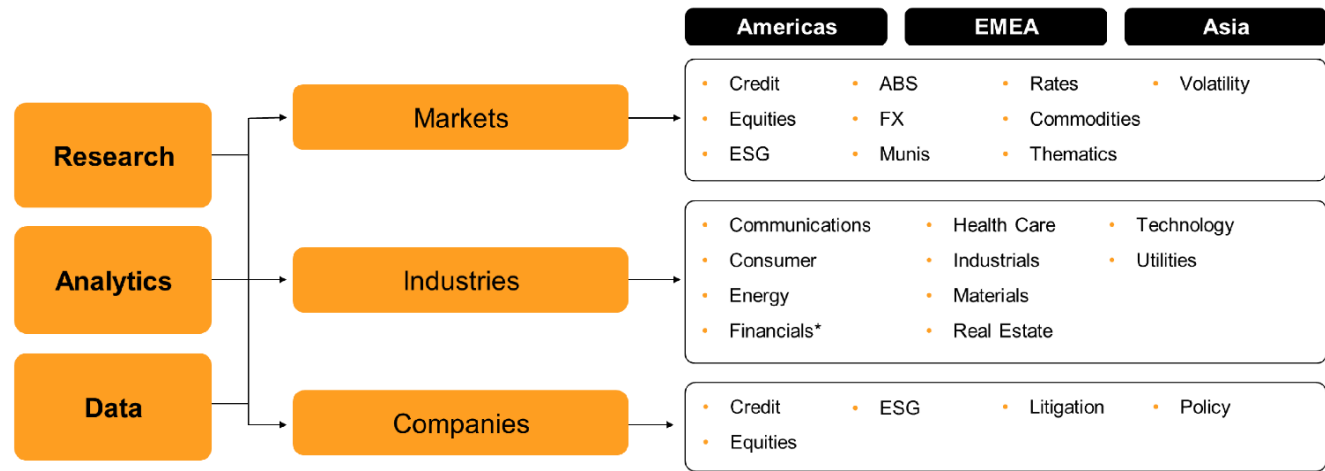
Take the next step.

For additional information, press the <HELP> key twice on the Bloomberg Terminal®.

Beijing +86 10 6649 7500	Hong Kong +852 2977 6000	New York +1 212 318 2000	Singapore +65 6212 1000
Dubai +971 4 3641000	London +44 20 7330 7500	San Francisco +1 415 912 2960	Sydney +61 2 9777 86 00
Frankfurt +49 69 92041210	Mumbai +91 22 6120 3600	Sao Paulo +55 11 2395 9000	Tokyo +81 3 4565 8900

Bloomberg Intelligence

Research, analytics and data tools to help you make informed investment decisions



Bloomberg Intelligence by the Numbers.

500

research professionals

135+

industries

600+

data contributors

2,000+

companies

21

markets covered

