ABIresearch
THE TECH INTELLIGENCE EXPERTS™

GENERATIVE AI IN THE ENTERPRISE SECTOR:

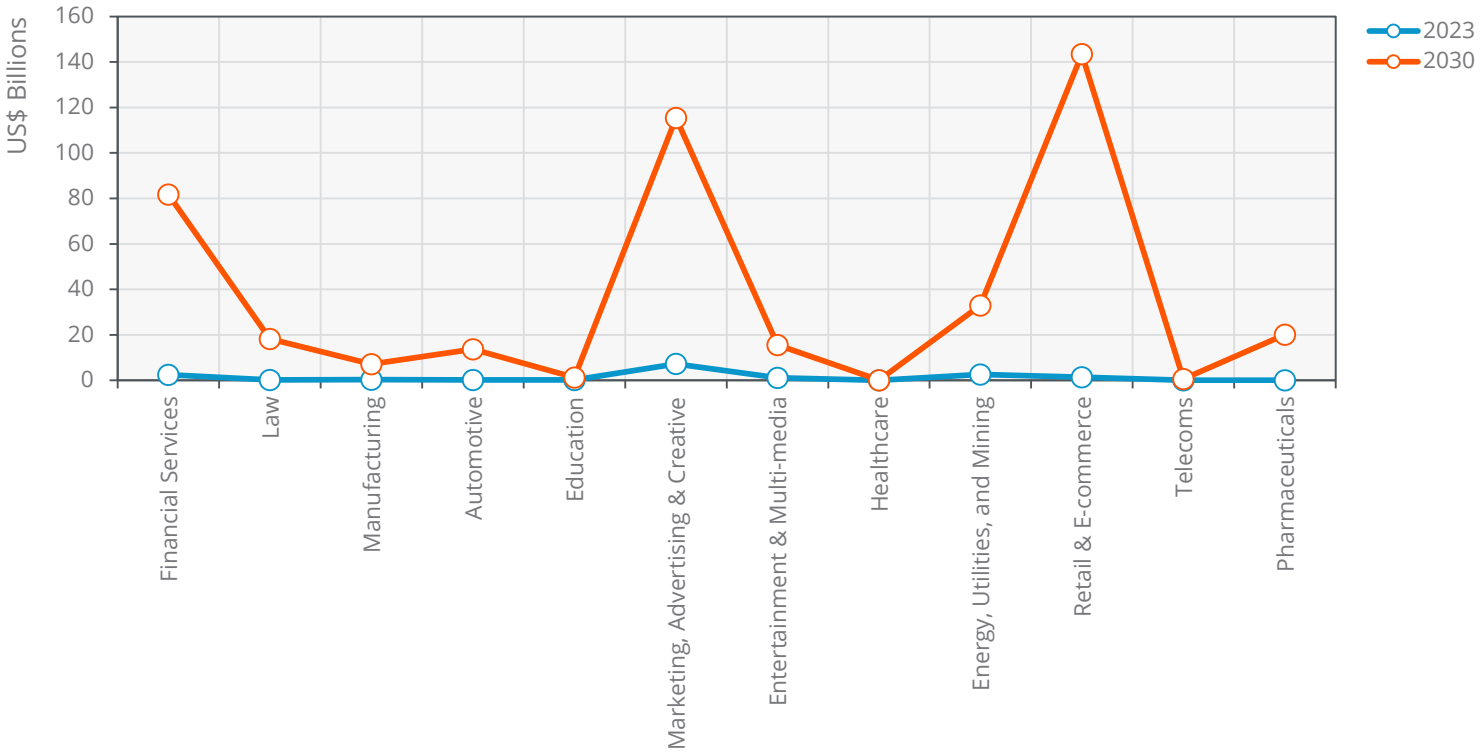
HOW SUPPLIERS CAN MEET THE HYPE

HOW SUPPLIERS CAN CAPITALIZE ON THE INCREASED DEMAND FOR GENERATIVE AI IN THE ENTERPRISE SPACE

Generative Pretrained Transformer (GPT) applications like ChatGPT have ignited massive interest in Generative Artificial Intelligence (Gen AI) since late last year. So far, the opportunities for Gen AI have been constrained to the Business-to-Consumer (B2C) space, with the Business-to-Business (B2B) largely undefined. In the long run, Gen AI has a colossal impact on the B2B space, with ABI Research forecasting it to contribute roughly $450 billion in value across various verticals by 2030, but several enterprise challenges are holding back adoption right now. The value of Gen AI is too great to pass up, from improved employee productivity, and operation efficiency to service augmentation and even widespread automation.

Chart 1: Generative AI Vertical Value Creation World Markets: 2023 to 2030

(Source: ABI Research)



We're still at a very early stage of enterprise generative AI, with the first deployments being reserved for low-hanging use cases where the stakes aren't so high. But further adoption is currently unlikely as risk-reward alarms are ringing across C-suite level executives, such as data security, Intellectual Property (IP) protection, copyright infringement, and the possibility of fragmentation, all underpinned by a significant skills and knowledge gap. Moreover, today's generative AI models are too big and general-purpose, lacking the performance, task efficiency, cost-effectiveness, and security necessary to meet enterprise investment criteria.

But the supply side of the market remains active as they look to build out a successful business-to-business (B2B) proposition. Many juggernaut tech firms, as well as startups, are developing generative AI models, applications, and services. However, as they explore this emerging opportunity, they must continue to grapple with a steep learning curve

as they navigate a new commercial domain. Most significantly, a cost crisis is emerging. Building, training, and running generative AI models have massive overheads. Overcoming this crisis cannot rely on the cash-light B2C market or existing "freemium" revenue models, so they must be proactive and start testing new monetization and product strategies.

WHAT'S HOLDING GENERATIVE AI BACK IN THE ENTERPRISE RIGHT NOW?

Unlike the Business-to-Consumer (B2C) space, enterprise deployment of AI brings significant risks that need to be weighed against potential rewards prior to investment. This risk has led to notable "bans" on third-party, "black box" generative AI services like ChatGPT. While generative AI enterprise use cases continue to emerge, delivering business value cannot rely on large, generalized models as they are slow, insecure, expensive, not adapted for the tasks they service, and subject to dangerous hallucinations. Instead, smaller, contextualized models fine-tuned on specific datasets will offer a much greater Return on Investment (ROI). For example, an investment firm on Wall Street may want to develop an AI-based tool to analyze the stock market and inform users of key trends. Or a utility provider would want a generative AI model that can predict future energy demands.

But a lack of smaller, contextualized models is not the only challenge for enterprise generative AI adoption. Other significant barriers exist, as outlined in the diagram below:

Risk of Vendor Lock-In and Internal Fragmentation:
Closed-source Gen AI models, which are more popular right now, make it difficult to switch between vendors seamlessly. Moreover, when different business units adopt different Gen AI solutions, the likelihood of internal operational challenges emerging increases. These factors dissuade enterprises from investing.

Intellectual Property Protection:
Enterprises need to put safeguards to their intellectual property to avoid any potential leaks through sharing sensitive data over Gen AI platforms.

Skills Gap:
Gen AI s requires extensive engineering expertise in deep ML, while there's a lack of platforms that even support development.

Data Privacy:
Businesses rightfully worry that internal data (e.g., via employees) can be used to train AI models. This could have significant ramifications for intellectual property. E.g., software developers using Gen AI tools to generate new code.

Trustworthiness:
Most Gen AI solutions still don't have the high accuracy, dedicated capacity, and Service Level Agreements (SLAs) that enterprise use cases require—necessitating human oversight still. Moreover, closed-source AI models don't tell developers the "why" behind outputs, but open-source models will add greater transparency. Explainability is also key to building trust. Closed source API-based Gen AI tools are 'black boxes' provide answers but cannot explain why they have given a certain answer. Finally, misinformation must be accounted for, too, as hallucinations continue to raise justified concerns about Gen AI solutions, even in low-risk enterprise use cases.

Operational Change:
Gen AI implementation comes with a host of governance and regulatory frameworks that most enterprises have little knowledge of.

On-Premises Deployment/VPC:
Gen AI solutions will be commonly deployed on-premises and in Virtual Private Cloud (VPN) networks. However, these deployments can be prohibitively expensive and time consuming for enterprises.

Legal and Regulatory Confusion:
Disorientation can quickly settle in when a business is unsure who owns the content generated by Gen AIs and what data can legally be used in AI model training.

Ethical and Social Concerns:
Many ethical and social problems must be confronted before enterprises embark on Gen AI usage. The greatest concerns surrounding generative Gen AI include data privacy, workforce impact, misinformation/ deepfakes, energy footprints, copyright infringement, intellectual property (IP) theft, inadequate accuracy, and questions about the AI regulatory environment.

Lack of Corporate Strategy:
Enterprises do not currently have the governance, strategy, or business outcomes in place yet to effectively deploy Gen AI. Most enterprise deployments are very isolated and could lead to significant siloing if a corporate strategy is not put in place. Until enterprises are well-versed in Gen AI governance and identify the specific business outcomes desired, implementation will be tricky.

Usage Safety:
Due to a lack of adoption of guardrails and content safety features, enterprises cannot centrally control and manage the inputs/outputs of Gen AI.

Generative AI adoption will always come with enterprise risk, but addressing these challenges will help mitigate it and accelerate the B2B market.

WHY SMALLER, "FINE-TUNED" MODELS ARE THE FUTURE OF GENERATIVE AI

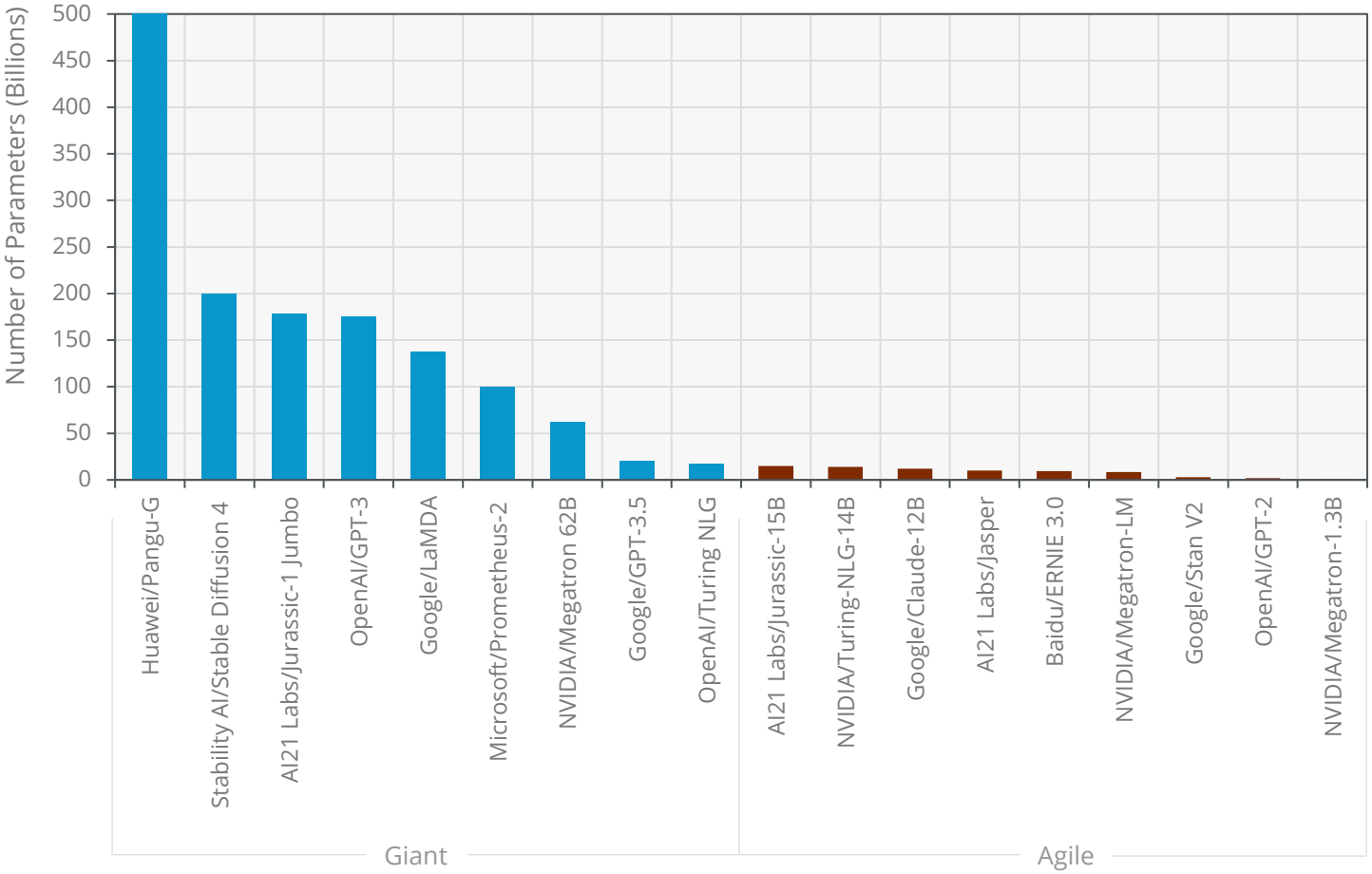
While studies confirm that generative AI models trained on enormous datasets speed up training convergence and improve accuracy, they are not ideal for executing specific business functions. The true inflection point for the adoption of generative AI in the enterprise sector will come when smaller, more finely-tuned generative AI models are built for specific applications or use cases. Smaller models are less resource intensive with lower training and inference costs, provide greater transparency and explainability through retrieval-based inferencing, and offer greater timeliness for enterprise deployment. "Fine-tuning" these "small" generative AI models can enhance performance and trustworthiness. Moving from "giant, generalized" to "smaller, fine-tuned" generative AI models will help alleviate the data privacy, performance, and trustworthiness concerns that enterprises currently have. By deploying "fine-tuned" models, enterprises gain the following advantages:

- **Lower Cost:** Training and inferencing costs are massively lowered, and fine-tuned iterations are much less resource intensive.
- **Explainability and Trustworthiness:** "Fine-tuned" models do not rely on known knowledge, but instead on "retrieval-based" inference models. These can reference sources to back up output. Direct access to original data can limit hallucinations and data approximations.
- **IP Ownership:** Unlike public AI models, fine-tuned models only leverage internal data that the enterprise can be sure they have the rights to.
- **Performance Optimization and Contextualization:** When generative AI solutions are tailored for specific business outcomes, this decreases the number of parameters required and improves performance. As an example, a general ChatGPT can involve up to 170 billion parameters, while a contextualized model may only need about 1 billion parameters or less. These finely-tuned models can also be catered to specific hardware, which decreases cost and improves utilization. Furthermore, because the models are more tailored and use well-curated data, the risk of hallucinations is far lower.
- **Better Economies of Scale:** ML-supported open-source models, which fine-tuned generative AI applications and use cases use, are cheaper than API-based services and better optimized for owned hardware. Scaling an API into GPT-3, for example, will not satisfy an enterprise's economy at scale ambitions. On the other hand, deploying open-source generative AI on owned servers will translate into cost savings, as the enterprise does not need to pay for tokens.

Building generative AI applications based on fine-tuned models has been a significant challenge for enterprises, with internal skillsets and the risk of creating siloed business units being at the forefront of worries. However, advancements in open sourcing generative AI and ML service tools will make fine-tuned models a more realistic opportunity for enterprises.

Chart 2: Number of Parameters (Billions) for Large Language Models

(Source: ABI Research)



The year 2023 kicked off with vendors discussing a trillion+ parameter models, highlighting the industry trend to increase the size—and thus, accuracy—of generative AI. But the problem is that massive generative AI models are expensive, resource-intensive, and time-consuming. The future of generative AI will require a more effective business case, which can only be achieved with "tailored, fine-tuned" models.

OPEN-SOURCE OR CLOSED-SOURCE MODELS? OR BOTH?

Open-source generative AI models are rapidly advancing and are destined to be the future of generative AI; however, closed-source models will still be practical for many enterprises, at least in the short term. Given the significant pros and cons of open- and closed-source models, enterprises and suppliers will refrain from committing solely to one or the other. Ultimately, ABI Research anticipates a meshed, "hybrid" model to induce greater industry value for generative AI. This will enable implementers to leverage federated learning in an economical way, while ensuring data are secure within an enterprise's walled garden.

Table 1: Evaluation of Open- and Closed-Source Models for Enterprises
(Source: ABI Research)

OPEN		CLOSED	
Opportunities	**Challenges**	**Opportunities**	**Challenges**
Enterprises can customize/fine-tune using their own data. External innovation can support improved performance. Low vendor lock-in. More scalable for enterprise use cases. Ecosystem is rapidly expanding. Emerging open-source security frameworks that can be embedded alongside model/ applications.	Requires in-house developmental expertise or third-party support, which can be prohibitive for early-stage or Small and Medium Enterprises (SMEs). Not suitable for mission-critical or sensitive use cases, as security issues are known. Often requires on-premises servers or infrastructure to run models. Heavily reliant on fine-tuning for optimized performance.	Market-leading performance. Security frameworks embedded. Ease of access without any internal skills needed. APIs can be consumed in a flexible model ensuring enterprises of all sizes can access.	Risk of vendor lock-in for users. High API cost and egress fees. Lacks explainability, transparency, and observability. Enterprises can customize/fine-tune using their own data.

HOW CAN SUPPLIERS SEIZE THE GENERATIVE AI MARKET OPPORTUNITIES?

The generative AI supply chain is rapidly evolving. The market sees new foundational models being released, an extensive list of applications/plug-ins being deployed weekly, and new vendors entering this potentially lucrative space through professional services or partnerships. In the following sections, we summarize the activity unfolding in various generative AI market opportunities waiting for each supplier and highlight some "out-of-the-box" monetization strategies.

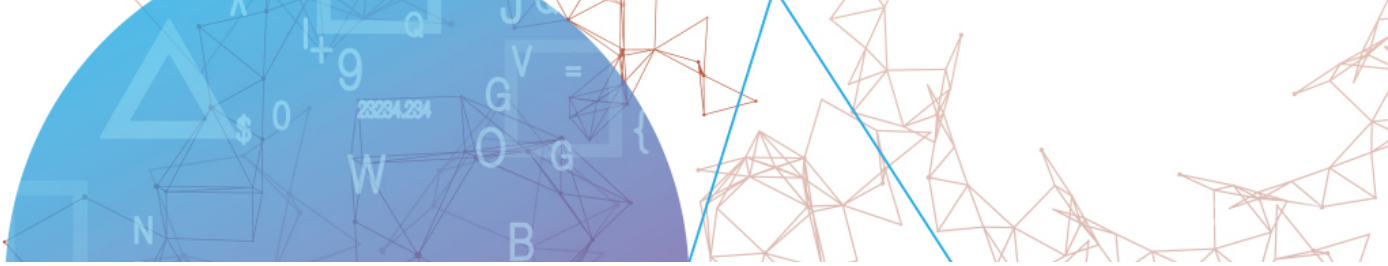


Table 2: Market Opportunities for Gen AI Suppliers

(Source: ABI Research)

Research & Development	Hardware Providers	Foundation Model Providers	Data Services	ML Service Tools	Application Developers	Enterprise Services
Guardrail development can create a competitive advantage in a market concerned about data privacy and environmental impact.	Growing demand for accelerators, developing strong value proposition at the edge, and building full-stack services leveraging hardware innovation.	"Responsible AI," cost offsetting to application developers, and tapping into huge customer bases can establish a market-leading position.	Increased demand for data privacy (data synthesis, curation, and monitoring) and integrating data services into emerging Gen AI platforms.	Embracing and monetizing open-sourced models, profiting from renewed enterprise engagement with AI, and development of no code tools.	Leveraging open-source models to build full-stack applications and seizing the largely untapped space for fine-tuned, contextualized applications.	MSPs can assist smaller enterprises with their Gen AI efforts on day 0, 1, and 2 operations. Build highly secure and transparent implementations.

RESEARCH AND DEVELOPMENT

Hardware vendors, cloud service providers, system integrators, and consultants spend billions on Research and Development (R&D) each year, making it the catalyst for generative AI evolution. Whether it's building/training a new AI model or improving hardware utilization, R&D has an enormous impact across the generative AI realm. Generative AI R&D companies are used to increasing parameters, performance, etc., but ABI Research believes that safety/security will be the main focus going forward, considering tighter regulation and standardization efforts.

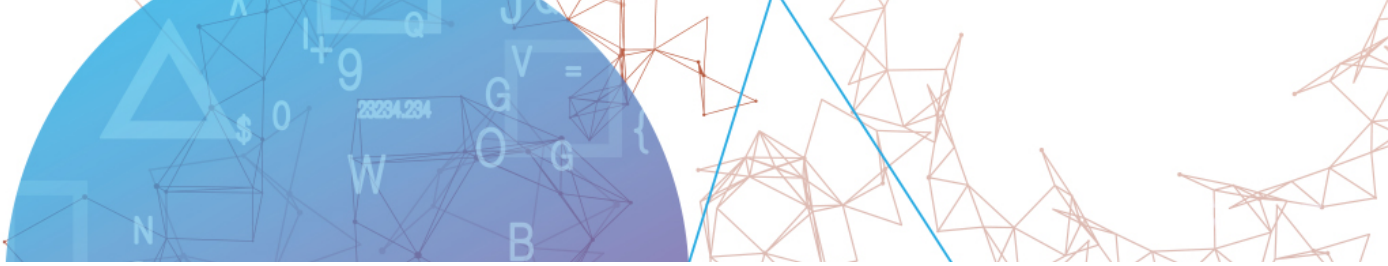
HARDWARE PROVIDERS

The hardware market is dominated by a few vendors with huge budgets, making it the toughest generative AI layer to enter for newcomers. NVIDIA has established itself as the market leader for Graphics Processing Units (GPUs), which are used for AI training and inferencing. NVIDIA has also developed generative AI and ML service tools to augment enterprise adoption of generative AI. Intel, AMD, and Qualcomm are also heavy hitters in the hardware space. Qualcomm has proven its ability to run Stable Diffusion inference (a 1+ billion parameter text-to-image generative AI application from Stability AI) on-device. Meanwhile, Intel and service provider BCG have collaborated to help offer services directly to the enterprise.

FOUNDATION MODEL PROVIDERS

The foundation model market is another tough one to penetrate, requiring high hardware costs, strong AI training expertise, and data access. Building and training a foundation model is a significantly expensive endeavor, so as a result, hyperscalers like Microsoft, Meta, Google, and Baidu have overshadowed other firms. There are also a few startups, such as OpenAI, receiving significant financial backing from Microsoft, Anthropic, Cohere, and AI21. Amazon Bedrock, which will monetize Titan and support enterprise adoption of generative AI-based applications, exemplifies how foundation model providers are now beginning to capitalize on the commercial opportunities of generative AI.

Until now, closed-source models have dominated this space, but that won't always last. The evolution of open-source models, copyright challenges, data constraints, ethical issues, and enormous costs will inevitably cause friction. Foundation model providers must demonstrate their commitment to protecting customer data.



DATA SERVICES

While initial generative AI models have relied on public information (e.g., GPT-3 uses the Internet), data privacy/copyright challenges will make this approach to training more difficult. This will benefit companies offering data services, as enterprises will need synthetic data for training once regulation/cost barriers reduce the quantity of available data. Moreover, enterprises implementing generative AI will find it incredibly valuable to have internal data curated and labeled to fine-tune their models. Showcasing the perceived value in this area, several startups have received investment recently, with Accenture's recent stake in curation and labeling provider Stardog being a good example.

ML SERVICE TOOLS

While ML platforms are more supportive within other AI domains like computer vision, they are increasingly being used in Natural Language Processing (NLP) services. Over time, these ML-based tools and services will be integral to enterprise generative AI deployment because they enable fine-tuning and application development. Some of the tools that ease the development, deployment, operations, and management of generative AI models for enterprises are outlined below:

- **Optimization tools** bolster the overall performance and efficiency of generative AI models/applications by allowing generative AI to function quicker and more accurately.
- **Integration services** ensure that generative AI applications interoperate with existing enterprise processes and workflows.
- **Cloud platforms** act as a central hub where developers and enterprises can deploy, monitor, and manage generative AI deployments across cloud infrastructure.
- **AI security services** are a big deal considering the scrutiny surrounding generative AI legal and ethical concerns. Enterprise adoption will not proliferate until stakeholders are confident that generative AI solutions won't impact data privacy, security, or IP.
- **Low/no-code platforms,** which offer graphical interfaces, drag-and-drop functionality, pre-built application frameworks, generative AI model APIs, etc., enable developers to build generative AI software applications and workflows quickly and easily.

APPLICATION DEVELOPERS

Within the generative AI world, three general generative AI-based application categories have emerged: 1) User Interfaces (UIs) for foundation models (ChatGPT and Bard); 2) application plug-ins based on public generative AI models (WriteMage); and 3) full-stack applications built from fine-tuned generative AI models.

The first two categories don't require extensive AI skills and are well-suited for the B2C domain, but not so much for the B2B market. ABI Research sees the third option, full-stack applications built with specific business use cases in mind, to be the enterprise opportunity. For the next 6 months, application developers will increasingly leverage open-source models to build the finely tuned, full-stack applications that enterprises want. We anticipate a high proportion of these generative AI deployments to be on-premises, particularly when data sensitivity is of concern.

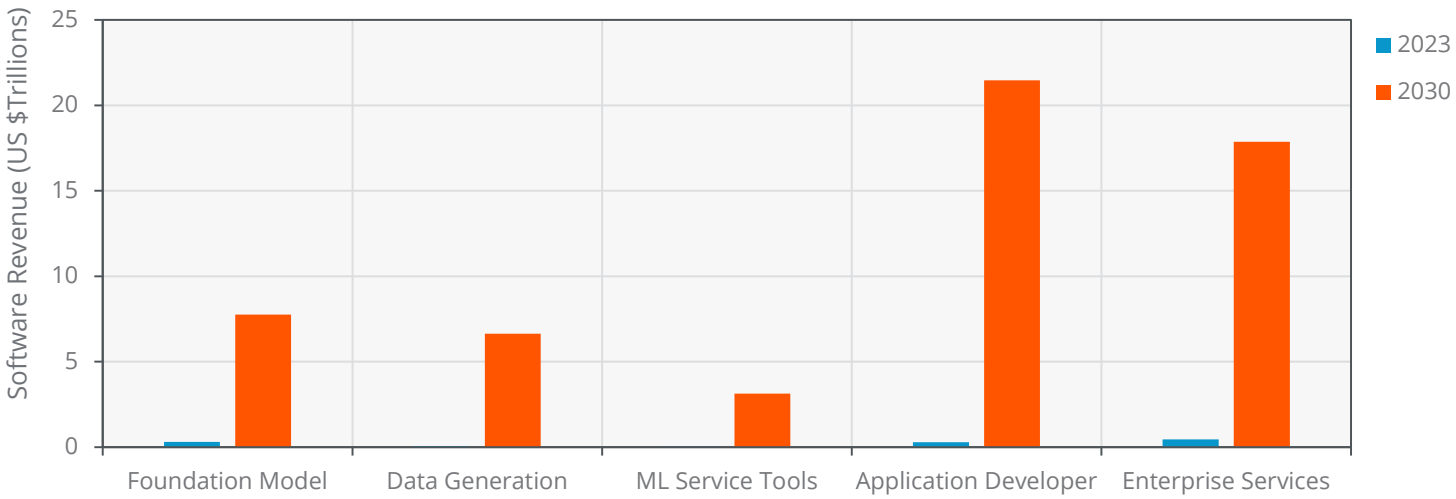
ENTERPRISE SERVICES

Enterprises lack the skills, internal processes, and strategic understanding needed to successfully implement generative AI into business operations and processes. This precipitates a lucrative opportunity for third-party services, such as business consultants, systems integrators, vertical Independent Software Vendors (ISVs), cloud vendors, Managed Service Providers (MSPs), and resellers. This space is mainly fueled by partnerships among business consultants and system integrators, and many examples abound. On the business consultant side, notable collaborations include Bain and OpenAI, BCG and Intel, PWC and Harvey AI, and more. For system integrators, Google Cloud has partnered with Tata Consultancy Service, Wipro, Cognizant, and Capgemini, while Accenture and Scale AI have teamed up.

As depicted in Chart 3 below, the total revenue opportunity for software within the generative AI supply chain will increase from US$1.2 billion in 2023 to nearly US$57 billion by 2030.

Chart 3: Software Revenue Opportunity for the Gen AI Supply Chain World Markets: 2023 vs 2030

(Source: ABI Research)



Fueling revenue creation, going forward, requires suppliers to complement current "freemium," subscription, and consumption-based models with "out-of-the-box" strategies that have proven successful in adjacent markets. Some of these revenue opportunities are as follows:

- **Open-Source Productization:** ISVs, ML tool providers, and enterprise service providers should build products using increasingly competitive open-source models.
- **Advertising:** Successfully used to support monetization of search tools, ISVs, hyperscalers, and ML tool providers could integrate ads into their products.
- **Center of Excellence (CoE) Support:** Startups and ISVs lack the infrastructure and capital to build and scale generative AI applications. Hyperscalers, hardware vendors, integrators, and consultants should look to build accelerators using their tools/infrastructure to support application ecosystem development. This strategy can increase resource and platform usage, while also driving returns through equity.
- **Watermark/Citation Removal:** More applicable in image generation, this strategy can create revenue in both the B2B and B2C markets. It also fits with the wider industry trend toward intellectual property protection.

HOW SHOULD GENERATIVE AI SUPPLIERS SUPPORT THEIR COMMERCIAL PROPOSITION?

The supplier ecosystem still has a way to go before achieving commercial success. The biggest challenges to overcome include safety/security concerns, huge operational costs, and a lack of clearly defined monetization strategies. In this final section, ABI Research provides several ways for suppliers to carve out a leading commercial position in the market:

- **Be a Leader in "Responsible AI":** Establishing a regulatory framework for generative AI will require effort from both enterprises and vendors. A top-down approach must be completed through a bottom-up approach led by vendors. Vendors can lead in "responsible AI" by proactively developing and enforcing safeguards and guardrails to ease enterprise anxiety. This encompasses data used for training purposes, energy usage for generative AI, model accuracy requirements, and the use of watermarks/citing for AI-generated content.
- **Develop Easy-to-Use No-Code AI Platforms:** For enterprise adoption of generative AI to really take off, vendors must offer solutions that enable them to deploy finely-tuned applications (e.g., Harvey AI and Jasper) in a straightforward manner. Current products like GPT are expensive, susceptible to hallucinations, hardware intensive, and lack explainability. To tap into the B2B space, vendors must develop generative AI that provides context and is tailored to specific business outcomes. Meanwhile, open-source models and easy-to-use ML service tools (low/no-code platforms) will enable enterprises to be bolder in their generative AI endeavors, as they will lower barriers to deployment.
- **Harness the Practicality of Commercial Partnerships:** Striking strategic partnerships has been a common trend in the generative AI supply chain. A number of examples abound, such as NVIDIA and Snowflake, Bain and OpenAI, Cohere and LivePerson, and more. These vendors realize that going at generative AI alone will not be commercially advantageous, and it's more efficient to lean on the specialized skills and tools offered by partners.
- **Identify and Deploy a Diverse Range of Monetization Strategies Built around Open-Sourcing:** Monetizing generative AI is a significant challenge for vendors. Venture Capital (VC) funding won't pour in forever, so it's imperative that vendors start creating robust revenue streams for open-source solutions to keep pace with generative AI usage demand. Long-term B2B success requires careful thought into various industry dynamics. From pay-as-you-go to turnkey platforms, vendors can choose from many new revenue models. However, choosing the right one will depend on four main considerations:
  - **Customers:** Customer demand for elasticity strongly influences what revenue models will work for stakeholders.
  - **Core Competencies:** Does the vendor have the necessary skills to deploy enterprise services? And have third parties or ISVs been identified to build an application marketplace?
  - **Partner Ecosystem:** Some revenue models, such as turnkey platforms, revenue share or commission, and transformative consulting, often require partnerships to ensure cross-chain competencies.
  - **Competitor Models:** Creating value for customers is all about differentiating your products from the competition. Therefore, vendors must always be surveying other generative AI solution providers to identify gaps in existing offers.

ABI Research's AI and Machine Learning (ML) market intelligence service assesses the market opportunity created by AI related technology. Click the link below to see our extensive research.

VISIT OUR SITE

If you need help understanding the future of AI and how it intersects with your market, or if you have critical questions that need to be answered, ABI Research can help. Since 1990, ABI Research has partnered with hundreds of leading technology brands, cutting-edge companies, forward-thinking government agencies, and innovative trade groups around the world. Our leading-edge research and worldwide team of analysts enable ABI Research to deliver actionable insights and strategic guidance on the transformative technologies that are reshaping industries, economies, and workforces today.

CONTACT US

ABOUT ABI RESEARCH

ABI Research is a global technology intelligence firm delivering actionable research and strategic guidance to technology leaders, innovators, and decision makers around the world. Our research focuses on the transformative technologies that are dramatically reshaping industries, economies, and workforces today. ABI Research's global team of analysts publish ground-breaking studies often years ahead of other technology advisory firms, empowering our clients to stay ahead of their markets and their competitors.

157 Columbus Avenue
New York, NY 10023 USA
Tel: +1 516-624-2500
www.abiresearch.com