taxnotes® federal

Volume 190, Number 12 ■ March 23, 2026

No Need to Reboot: GenAI Fits the Tax Stack

by Michael Plowgian, Alistair Pepper, Prita Subramanian, and Michael Timmerman

Reprinted from *Tax Notes Federal*, March 23, 2026, p. 1947

No Need to Reboot: GenAI Fits the Tax Stack

by Michael Plowgian, Alistair Pepper, Prita Subramanian, and Michael Timmerman

Michael Plowgian

Alistair Pepper

Prita Subramanian

Michael Timmerman

Michael Plowgian is a principal, Alistair Pepper is a managing director, Prita Subramanian is a principal, and Michael Timmerman is a manager at KPMG LLP.

In this report, the authors explore the taxation of artificial intelligence under existing rules and highlight challenges with two alternative proposals.

The information in this report is not intended to be "written advice concerning one or more Federal tax matters" subject to the requirements of section 10.37(a)(2) of Treasury Department Circular 230. The information contained herein is of a general nature and based on authorities that are subject to change. Applicability of the information to specific situations should be determined through consultation with your tax adviser. This report represents the views of the authors only and does not necessarily represent the views or professional advice of KPMG LLP.

Introduction

Seemingly everywhere you turn, there are grandiose predictions about how generative artificial intelligence (GenAI) will radically change our world:

- Dario Amodei, CEO of Anthropic, has said that a form of artificial intelligence that's "better than almost all humans at almost all tasks" could emerge in the next two or three years. Demis Hassabis, Google DeepMind's CEO, makes the same prediction but over a five-to-10-year time frame.[1] Zoom CEO Eric Yuan, along with other business leaders, has predicted that productivity gains from AI will result in a three- or four-day workweek.[2]
- More pessimistically, Geoffrey Hinton, the computer scientist sometimes called the "godfather of AI," thinks that there is "a 10 percent to 20 percent" chance that AI will

[1] Ryan Browne, "AI That Can Match Humans at Any Task Will Be Here in Five to 10 Years, Google DeepMind CEO Says," CNBC, Mar. 17, 2025.

[2] Hugh Cameron, "CEOs Predict New Type of Working Week," *Newsweek*, Oct. 10, 2025.

> lead to human extinction in the next 30 years.[3]

Maybe it's human nature to try to make things black or white, wholly good or wholly bad — or maybe much ado about nothing. Even leaders in the field have suggested current levels of investment in AI may be indicative of a bubble.[4] OpenAI CEO Sam Altman has asked rhetorically: "Are we in a phase where investors as a whole are overexcited about AI? My opinion is yes."[5] He also compared the current investment in GenAI to the dot-com bubble.[6]

But the starting point for this article is different: We don't know what GenAI will become. Maybe one of the predictions above will be correct. Or maybe GenAI will be a tool that disrupts some occupations and creates others, like many disruptive technologies before.

Whatever GenAI becomes, it is having an impact on the global economy. By July 2025 it was estimated that more than 700 million people, almost 10 percent of the world's adult population, were using ChatGPT each week.[7] In 2024 tech companies including Alphabet, Amazon, Meta, and Microsoft invested $230 billion in data centers, and were forecast to invest a further $320 billion in 2025.[8] NVIDIA, which produces the computer chips that are used in many data centers that support GenAI, has seen its market capitalization grow from less than $500 billion in 2023 to more than $4.5 trillion by the end of 2025.[9]

Despite GenAI's rapid uptake, in many cases the companies involved in the supply chain expect to make much of the profit from GenAI in the future, rather than today. While NVIDIA is profitable, generating over $77 billion in net income on $148 billion in revenue over its first three quarters in fiscal 2026,[10] its price-earnings ratio, which ranged between 30 and 50 in 2025, indicates that it is NVIDIA's future growth that has investors excited.[11]

But GenAI, or AI more broadly, is not just something that affects tech companies. AI is expected to transform a diverse range of industries, from financial services to medical devices.[12] For this reason, businesses in these industries are investing heavily in AI technologies and capabilities.

In light of the significant investments made in GenAI, policymakers and academics are starting to ask whether there is a need to rebalance tax systems in response to the rise of GenAI.[13] In part, this reflects a fear that GenAI will lead to mass unemployment, though we note that fears of mass unemployment have accompanied many other technological changes.[14] Some have suggested that to level the playing field between humans and machines, payroll taxes should be supplemented by a "robot tax," though it is far from clear what exactly a robot tax would entail.[15] The dominance of the AI frontier by U.S. and Chinese businesses also means that some countries are likely to ask whether the way the current international tax system allocates taxing rights between countries is fair, which is the same question that countries asked and struggled to answer in pillar 1 of the OECD/G20 inclusive framework's base erosion and profit-shifting two-pillar initiative.[16]

This article doesn't answer any of those questions. Because we don't know how AI will develop and be used, this article instead looks at where GenAI is today and what the existing tax rules say about how it should be taxed. In fact, this

[3] Dan Milmo, "'Godfather of AI' Shortens Odds of the Technology Wiping Out Humanity Over Next 30 Years," *The Guardian*, Dec. 27, 2024.

[4] *See, e.g.*, Itay Goldstein, "Is There an AI Bubble and What Happens If It Bursts?" *Penn Today*, Dec. 16, 2025 (reporting that many experts "agree that something bubble-like is taking shape").

[5] Alex Heath, "What Even Is the AI Bubble?" *MIT Technology Review*, Dec. 15, 2025.

[6] *Id.*

[7] Aaron Chatterji et al., "How People Use ChatGPT," NBER Working Paper 34255 (Sept. 2025).

[8] Samantha Subin, "Tech Megacaps Plan to Spend More Than $300 Billion in 2025 as AI Race Intensifies," CNBC, Feb. 8, 2025.

[9] Companies Market Cap, "Market Capitalization of NVIDIA (NVDA)" (last accessed Jan. 8, 2026).

[10] NVIDIA release, "NVIDIA Announces Financial Results for Third Quarter Fiscal 2026" (last accessed Jan. 20, 2026).

[11] Wall Street Numbers, "NVIDIA Corporation (NVDA) P/E Ratio" (last accessed Oct. 20, 2025).

[12] McKinsey & Co., "8 McKinsey Insights on How AI Is Reshaping Business" (Aug. 10, 2025).

[13] Cady Stanton, "'Robot Tax' Proposal Sparks Skepticism Over Its Practicality," *Tax Notes Federal*, Oct. 27, 2025, p. 694.

[14] "The Future of Jobs: The Onrushing Wave," *The Economist*, Jan. 18, 2014.

[15] Rossana Merola, "Inclusive Growth in the Era of Automation and AI: How Can Taxation Help?" 5 *Frontiers AI* 867832 (May 31, 2022).

[16] OECD, "G20 Finance Ministers and Central Bank Governors Meeting — Session 5, International Taxation" (July 10, 2021).

article looks only at a subset of transactions involving GenAI. Instead of trying to analyze the full range of transactions in one article, we focus on how our existing tax rules apply to businesses that provide access to GenAI, particularly the hyperscalers, which provide the compute power, storage, and networking necessary to train and deploy GenAI.

The article begins with background on GenAI, explaining the key terms and the supply chain and business models that underpin GenAI. This discussion illustrates that the business models and transactions used to provide access to GenAI are fundamentally the same as the transactions used to provide other cloud services, though we note where potential differences exist. The article then examines how the existing tax rules (both U.S. federal income tax law and guidance from the OECD) characterize (and thus source) transactions involving hyperscalers and GenAI. We then focus on potential tax issues involving the data centers used to provide GenAI and other cloud services, in particular whether data centers create a permanent establishment for multinational enterprises that use them and how data centers should be remunerated under transfer pricing principles. The article closes with a brief overview of some proposals for alternative approaches to taxing GenAI transactions, concluding that radical changes are not necessary.

Background on GenAI

What Is AI?

When considering the taxation of GenAI, it is important to keep in mind that AI is not new. Indeed, AI has been used for several decades.[17] While most people are familiar with GenAI chatbots such as OpenAI's ChatGPT, Alphabet's Gemini, or Anthropic's Claude, AI is a much broader field.[18]

When thinking about AI, it is helpful to be familiar with a few concepts:

- **AI** "refers to the use of technologies to build machines and computers that have the ability to mimic cognitive functions associated with human intelligence, such as being able to see, understand, and respond to spoken or written language, analyze data [and] make recommendations."[19]
- **Machine learning (ML)** "is a subset of artificial intelligence that automatically enables a machine or system to learn and improve from experience. Instead of explicit programming, machine learning uses algorithms to analyze large amounts of data, learn from the insights, and then make informed decisions."[20]
- **Deep learning (DL)** "is a subset of machine learning that uses artificial neural networks to process and analyze information."[21]
- **Large language models (LLMs)** "are a category of deep learning models trained on immense amounts of data, making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks."[22]
- **GenAI** "refers to a specific subset of AI that uses programs to process large data sets, detect patterns, and then create new works of text, imagery, video, and even computer code based on the instructions it's given."[23]

For example, ChatGPT is a GenAI chatbot. It is underpinned by an LLM (GPT-4 or GPT-5). That LLM was trained using techniques referred to as DL, which is a subset of ML, which is itself a subset of AI.[24]

How Are Businesses Using AI?

This article focuses on GenAI and the supply chain and business models that underpin it. However, GenAI is just one example of the ways

[17] Tim Mucci, "The History of AI," IBM (last accessed Dec. 16, 2025).

[18] *See, e.g.*, Michael Wooldridge, "ChatGPT Is Not 'True AI.' A Computer Scientist Explains Why," Big Think, May 17, 2023.

[19] Google Cloud, "Artificial Intelligence (AI) vs. Machine Learning (ML)" (last accessed Oct. 20, 2025).

[20] *Id.*

[21] Google Cloud, "What's the Difference Between Deep Learning, Machine Learning, and Artificial Intelligence?" (last accessed Oct. 20, 2025).

[22] Cole Stryker, "What Are Large Language Models (LLMs)?" IBM (last accessed Oct. 20, 2025).

[23] Oracle, "Understand the Differences Between AI, GenAI, and ML" (Jan. 6, 2024).

[24] Ivan Belcic and Stryker, "What Is ChatGPT?" IBM (last accessed Jan. 5, 2026).

that businesses are using AI today or may deploy it in the future. Other examples include:

- In the pharmaceutical industry, Pfizer used AI to bring its oral treatment for COVID-19 to market more quickly. AI was used to optimize the search for the right molecule to deliver the treatment, to speed up the analysis of clinical trial data, and to optimize the manufacturing process.[25]
- In retail, Walmart has used AI to optimize its supply chain and reduce emissions, enabling it to get its products to its stores more efficiently and at a lower cost. Walmart has made this software available to third parties as software as a service (SaaS).[26]
- In agriculture, John Deere has developed See & Spray technology that identifies weeds and turns farmers' weed sprayers on and off depending on the weeds' presence, saving on spray.[27]

Though it is not the focus of this article, the widespread implementation of a variety of different forms of AI means that businesses beyond the archetypical tech companies will be assessing the income characterization and transfer pricing that this article explores in the context of GenAI, and that any changes to the tax rules to target tech companies are likely to have a much broader effect.

GenAI Supply Chain

Development, Pre-Deployment, and Deployment

The GenAI supply chain can be broken down into three phases:

1. **Development**. The foundation model (defined below) that underpins GenAI is developed. This phase includes gathering and organizing the data and providing compute power to train the model. In the training process, data is fed into the model to adjust its parameters to optimize performance.[28]
2. **Pre-deployment**. The foundation model is incorporated into a platform or applications, both of which are typically hosted by cloud computing providers (hyperscalers), so that the model or applications can be deployed to customers.[29]
3. **Deployment**. The foundation model, or application incorporating a foundation model, is made available to customers.

Key Actors

There are three key actors in the GenAI supply chain:

1. **Model Owners**. Standalone model owners, such as OpenAI, Anthropic, Mistral, and DeepSeek, have developed foundation models that underpin GenAI.[30] Many hyperscalers are also model owners.
2. **Hyperscalers**. Hyperscalers are cloud computing providers that provide the compute power, storage capacity, and scalability necessary to train foundation models and deploy GenAI.[31] Hyperscalers include Alphabet, Amazon, Meta, Microsoft, Tencent, and Baidu.[32] Some hyperscalers have also developed their own foundation models; for example, Alphabet's Gemini and Meta's Llama.
3. **Customers**. GenAI customers include individuals who buy subscriptions, but more importantly for the purposes of this article, business customers that deploy GenAI. Business customers may use off-the-shelf products, customize AI solutions using their own proprietary data, or

[25] Pfizer, "Data and AI Are Helping to Get Medicines to Patients Faster" (last accessed Oct. 20, 2025).

[26] Walmart release, "Walmart Commerce Technologies Launches AI-Powered Logistics Product" (Mar. 14, 2014).

[27] John Deere, "See & Spray™ Technology" (last accessed Oct. 20, 2025).

[28] *See, e.g.*, Amazon, "What Are Foundation Models?" (last accessed Dec. 10, 2025); Stryker and Rina Diane Caballar, "What Are Foundation Models?" IBM (last accessed Dec. 10, 2025).

[29] Amazon, "What Are Foundation Models?" *supra* note 28.

[30] Stryker and Caballar, *supra* note 28.

[31] *See* Karl Montevirgen, "The Rise of Hyperscalers: Reshaping Cloud Computing and Powering AI," Brittanica Money (last accessed Dec. 10, 2025); Melissa Palmer, "Hyperscalers: The Complete Guide to What, Why and How," Solar Winds blog, Jan. 24, 2023.

[32] Synergy Research Group, "Hyperscale Market Tracker" (last accessed Dec. 10, 2025).

develop AI applications that they then sell to their own customers.[33]

The roles performed by model owners, hyperscalers, and customers are similar across different GenAI business models (explored further below). Model owners, which include hyperscalers, develop the foundation models that underpin GenAI. The hyperscalers provide the computational resources necessary to train the foundation model during the development phase and then to deploy the model to customers in the deployment phase. Customers access foundation models in different ways, typically by contracting with either a model owner or hyperscaler.

Increasingly, customers are developing AI apps using the foundation models developed by model owners to provide more domain-specific solutions (that is, focused on specific tasks or industries) to their customers. For example, Harvey is an AI tool that has been developed for the legal industry. This tool started with a foundation model that was then further trained on general legal data, and which can be further trained using its end-customers' (law firms') proprietary data.[34]

What Are Foundation Models?

Foundation models are a type of DL model built on a neural network architecture (such as a transformer model described below) and trained on large data sets, which can perform general tasks and may serve as the "foundation" for building specialized AI apps through further fine-tuning or customization[35]:

- A **neural network** comprises an input layer that receives the input and holds it as raw features; hidden layers that transform input features into new representations by multiplying the features by the weights and adding the bias for each neuron or node; and an output layer that produces a final output or prediction. For example, a neural network designed to detect email spam would take key phrases as inputs, pass this information through various hidden layers, and produce an output — a prediction of whether an email is spam.[36]
- A **transformer model** is a type of neural network that includes a self-attention mechanism, which enables the model to examine an entire sequence of text or pixels simultaneously and thus better contextualize data.

There are also other types of neural networks, such as convolutional neural networks and recurrent neural networks.[37]

The foregoing is a conceptual description of how these models work. Foundation models consist primarily of a computer program (which renders the model architecture, algorithms, and data functionally operational) and the weights and biases file (which represents the accumulated training of the model and provides the values for each node in the neural network).[38] The weights and biases are the result of the training process and determine the accuracy of a foundation model.[39]

There are numerous types of pretrained foundation models, including:

1. **LLMs**. LLMs are designed primarily for natural language processing tasks such as text generation, translation, summarization, and question-answering.[40] Examples include OpenAI's GPT-3[41] and Meta's Llama 2.[42] LLMs are trained on a

[33] *See, e.g.*, Michael Chui et al., "The Economic Potential of Generative AI," McKinsey & Co. (June 2023); Hannah Marlowe et al., "Custom Intelligence: Building AI That Matches Your Business DNA," AWS blog, Oct. 31, 2025.

[34] Louise Donnery, "The Harvey AI Revolution: Changing the Face of Law in the UK," Clio blog, Aug. 25, 2025.

[35] Stryker and Caballar, *supra* note 28.

[36] Fangfang Lee, "What Is a Neural Network?" IBM (last accessed Dec. 1, 2025).

[37] Stryker and Dave Bergmann, "What Is a Transformer Model?" IBM (last accessed Dec. 1, 2025).

[38] OpenAI, "How ChatGPT and Our Foundation Models Are Developed" (last accessed Dec. 10, 2025).

[39] *Id.*

[40] Stryker, *supra* note 22.

[41] *Id.*

[42] Meta, "Llama 2" (last accessed Jan. 9, 2026).

broad spectrum of generalized and unlabeled data.[43]

2. **Multimodal Models (MMMs)**. MMMs are more versatile than LLMs and can handle a broader range of data types, such as text, images, audio, and video, making them suitable for more complex and integrated applications.[44] Examples include models like Alphabet's Gemini,[45] OpenAI's DALL-E 3,[46] and Anthropic's Claude 3.7.[47]
3. **Small Language Models (SLMs)**. SLMs, like LLMs, are designed primarily for natural language processing tasks, but are trained on smaller amounts of data.[48] This means they are more suited for deployment on edge devices or to support domain-specific tasks.[49]

Foundation Model Hardware

High-performance computing hardware resources are necessary to handle the massive amounts of data and complex calculations involved in training and deploying foundation models.[50] Hyperscalers operate the GenAI-capable data centers and strategically build and deploy these data centers, collaborating with chip makers and power companies.[51] Chips that power GenAI are scarce inputs, as evidenced by the significant growth in revenue, orders, and stock prices of chipmakers like NVIDIA.[52]

GenAI-enabled data centers are specifically designed to handle the complex and power-intensive workloads required for GenAI — training of and inference by foundation models.[53] As a result, GenAI-enabled data centers typically use specialized hardware such as graphics processing units (GPUs), tensor processing units (TPUs), and neural processing units (NPUs) that are optimized for parallel processing and handling large-scale computations.[54]

In contrast, data centers that are not optimized for GenAI typically use central processing units (CPUs) and are designed to provide flexible processing power for a variety of different computational tasks.[55] These tasks include web services, database management, and file storage, which do not require the same level of computational intensity as AI workloads.[56] In practice, a single data center often performs both GenAI and non-GenAI tasks and may use different types of processors.

GPUs have significantly more cores than CPUs, allowing GPUs to handle multiple tasks simultaneously through parallel processing.[57] This makes them highly efficient for tasks like rendering graphics, training ML models, and performing complex calculations.[58] TPUs are specifically designed for a range of GenAI workloads. They are adept at training and inference and may be used for code generation, synthetic speech, and media content generation.[59]

Hyperscalers are increasing their investments in their own chips to optimize performance, efficiency, and cost. Instead of relying solely on third-party processors, they are creating custom silicon chips tailored to their specific workloads and infrastructure needs.[60] Custom chips allow

[43] *Id.*

[44] Microsoft, "What Are Multimodal LLMs?" (last accessed Dec. 10, 2025).

[45] Google Cloud, "Multimodal AI" (last accessed Jan. 9, 2026).

[46] *See* Dall-E 3 (last accessed Dec. 10, 2025).

[47] *See* Claude 3.7 (last accessed Dec. 10, 2025).

[48] Caballar, "What Are Small Language Models?" IBM (last accessed Dec. 10, 2025).

[49] *Id.*

[50] Stryker and Caballar, *supra* note 28.

[51] *See, e.g.*, Jesse Noffsinger et al., "The Cost of Compute: A $7 Trillion Race to Scale Data Centers," McKinsey Quarterly (Apr. 28, 2025); Chris Gatch, "Will Utilities Overplay Their Hand in the Age of AI?" Data Center Dynamics, Jan. 29, 2025.

[52] NVIDIA's stock price on Dec. 1, 2022 (when OpenAI released an early demo of its ChatGPT) was $17 (stock split adjusted). As of May 14, 2025, the stock price was $134 — nearly an 8X multiple in less than three years.

[53] Noffsinger et al., *supra* note 51.

[54] Solomon Klappholz, "What Does a Data Center Look Like in the AI Era?" *IT Pro*, Nov. 8, 2024.

[55] *Id. See also* Alexandra Jonker and Alice Gomstyn, "What Is an AI Data Center?" IBM (last accessed Dec. 16, 2025).

[56] Klappholz, *supra* note 54.

[57] Josh Schneider and Smalley, "CPU vs. GPU for Machine Learning," IBM (last accessed Dec. 17, 2025).

[58] *Id.*

[59] Google Cloud, "Accelerate AI Development With Google Cloud TPUs" (last accessed Dec. 17, 2025).

[60] Sally Ward-Foxton, "Why Do Hyperscalers Design Their Own CPUs?" *EE Times*, Apr. 10, 2025.

hyperscalers to fine-tune architectures for their unique workloads, whether it is GenAI, cloud computing, or networking.[61] Many of the hyperscalers' chips focus on reducing power consumption while maximizing compute power, which is crucial for large-scale data centers.[62] Hyperscalers are also designing AI accelerators and networking chips to complement their custom chips, ensuring seamless integration across their infrastructure.[63] This trend is reshaping the semiconductor industry, as traditional chipmakers now face competition from their own customers.[64]

The GenAI capabilities of GPUs and TPUs come at a cost: The cumulative power consumption of the multiple cores of these chips is significantly higher than that of a CPU.[65] CPU-based data centers consume approximately 10 kilowatts per rack, whereas GenAI-enabled data center capabilities may consume five or even 10 times that level.[66]

This increased power demand requires data centers to have access to significant energy resources, with many turning to power from nuclear power stations.[67] Schneider Electric estimates that AI power consumption will grow at a rate of 25 percent to 33 percent annually through 2028, when it could reach as much as 18.7 gigawatts, taking data center power consumption to 93 gigawatts.[68]

In addition, GenAI-enabled data centers need more advanced cooling systems, enhanced networking capabilities, and even reinforced flooring because of the weight of server racks.[69]

How Do Customers Interact With Foundation Models?

A common misconception is that foundation models are trained in real time when users prompt the model to generate an output. Creating a new foundation model, or a new version of an existing model, is a massive undertaking that requires significant computational power and investment. Model owners do not typically publish information on training costs, but training the latest models could cost as much as $1 billion.[70]

When users interact with a foundation model by prompting it, the users do not change the weights and biases or computer code of the foundation model.[71] Instead, the users are simply providing instructions and additional context with the hope that the model provides a more accurate or more appropriate response.[72] That context does not alter the model for use by other users. However, there are ways that customers, typically business customers, can augment or alter foundation models to tailor them to their specific needs, for example, fine-tuning, which we discuss below.

Prompt Engineering and Personas

The simplest way that customers can seek to improve the output of a foundation model is through prompt engineering or creating personas. Prompt engineering refers to the process of refining a prompt to elicit a better response; for example, by providing additional context.[73] Personas are in effect more codified prompt engineering.[74] For example, a document translator persona within a GenAI platform may be specifically designed to handle document translations efficiently, using a dedicated translation tool and providing clear guidance to users on how to proceed with translations. Prompt engineering and the creation of personas do not alter the underlying foundation model, so

[61] *Id.*

[62] *Id.*

[63] *Id.*

[64] Laura Bratton, "Nvidia's Big Tech Customers Might Also Be Its Biggest Competitive Threat," Yahoo Finance, Oct. 20, 2025.

[65] Nick Evanson, "The Rise of Power: Are CPUs and GPUs Becoming Too Energy Hungry?" *TechSpot*, Oct. 24, 2022.

[66] Michaela Galarza, "To Power AI, Data Centers Need More and More Energy," *The Current* (University of California, Santa Barbara), Apr. 15, 2025; Socomec, "Understanding the Power Consumption of Data Centers" (last accessed Dec. 17, 2025); Solartech, "How Much Electricity Does a Data Center Use?" (Oct. 2, 2025).

[67] *See* John Addison, "AI Data Centers Energy Demand Growth," *Energy Central*, Oct. 10, 2024.

[68] *Id.*

[69] "AI-Ready Data Centers: The Infrastructure Behind LLMs, GPUs, and AI Clusters," Data Centers.com (Dec. 9, 2025).

[70] Will Henshall, "The Billion-Dollar Price Tag of Building AI," *Time*, June 3, 2024.

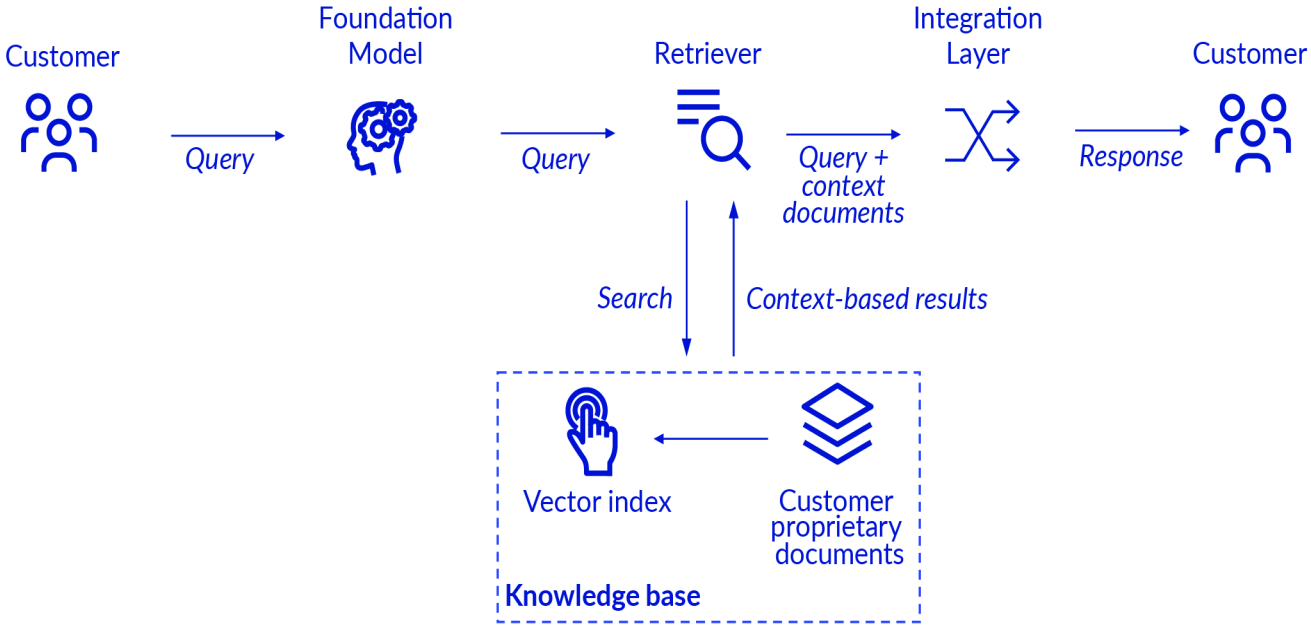
[71] Ali Alemi and Imtiaz Sayed, "Exploring Real-Time Streaming for Generative AI Applications," AWS blog, Mar. 25, 2024.

[72] *Id.*

[73] Vrundu Gadesha, "What Is Prompt Engineering?" IBM (last accessed Dec. 1, 2024).

[74] Nilesh Barla, "What Is Persona-Based Prompting?" Adaline blog, Mar. 14, 2025.

Figure 1. Retrieval-Augmented Generation



they do not make the model more valuable to the model owner, the hyperscaler, or other users.[75]

Retrieval-Augmented Generation

Business customers may combine a foundation model with their proprietary data through a retrieval-augmented generation (RAG) system. A RAG system does not alter a foundation model but provides a database of proprietary data from which a model can pull.[76] A RAG system has four main components (see Figure 1). These are:

1. the **knowledge base**, or customer proprietary documents, which provides the data that underpins the GenAI response;
2. the **retriever**, which is responsible for fetching relevant information from the knowledge base, searching, and ranking proprietary documents based on their relevance to the input query;
3. the **integration layer**, which coordinates the overall functioning of the RAG architecture; and
4. the **foundation model**, which takes the information retrieved by the retriever and generates coherent and contextually appropriate responses.[77]

As with general prompting, RAG does not alter the model or make it more valuable for other users.

Customer Fine-Tuning

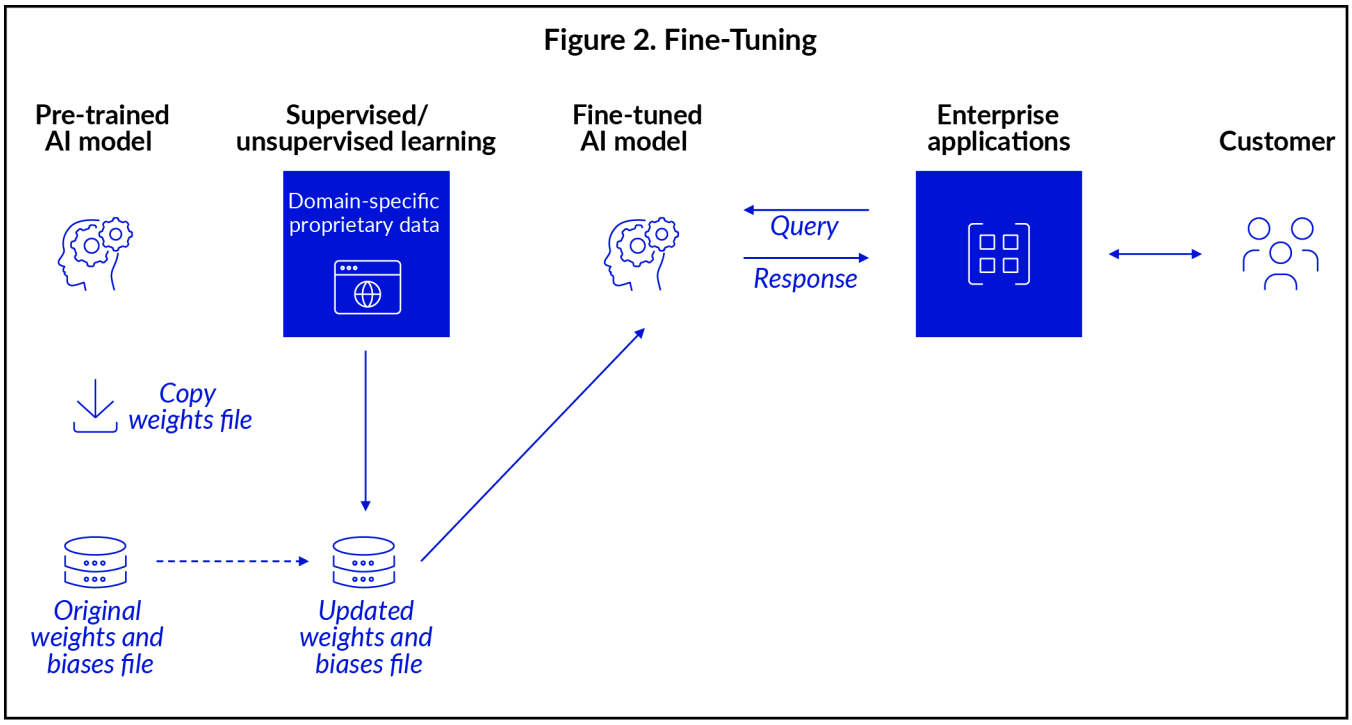
In some circumstances, customers may choose to adjust the weights and biases of a foundation model through "fine-tuning." In this process, a customer-specific copy of the weights and biases file is created, and this file is modified through additional training; for example, based on a customer's own proprietary data (see Figure 2).[78] This process creates a new weights and biases file. The level of fine-tuning can be scaled up or down depending on a customer's requirements. Fine-tuning is, essentially, additional training of the

[75] Thanh Tung Vu, "Understanding Prompt Engineering: From Zero-Shot Prompts to Scalable Systems," Medium, Apr. 21, 2025.

[76] Xin Huang et al., "Question Answering Using Retrieval Augmented Generation With Foundation Models in Amazon SageMaker JumpStart," AWS blog, May 2, 2023.

[77] IBM, "Components of a RAG System" (last accessed Dec. 8, 2025).

[78] Bergmann, "What Is Fine-Tuning?" IBM (last accessed Dec. 8, 2025).



model and is a computationally expensive process.[79]

Business Models for Delivery of GenAI

GenAI can be delivered to customers in a variety of ways. In this article, we consider three types of stylized business models:

1. **First Party Model (1P Model)**, in which the model owner and hyperscaler are the same person (or part of the same group) and sell directly to the end customer.
2. **Second Party Model (2P Model or Reseller Model)**, in which the customer contracts with the hyperscaler, which has acquired the right to sell access to the foundation model from the model owner.
3. **Third Party Model (3P Model or Marketplace Model)**, in which the customer contracts with the model owner, with the hyperscaler acting as a host, and often as a marketplace.

As discussed in more detail later, these business models are essentially identical to SaaS business models, if you substitute "software developer" for "model owner" and "cloud service provider" for "hyperscaler." The similarity between the SaaS model and GenAI business models is a critical point and one we return to throughout this article. These stylized business models reflect several (but by no means all) of the common operating models of which we are aware. Given the speed with which the technology and business practices are evolving, this discussion cannot reflect the full diversity of actual business models and practices used to deploy AI solutions to customers.

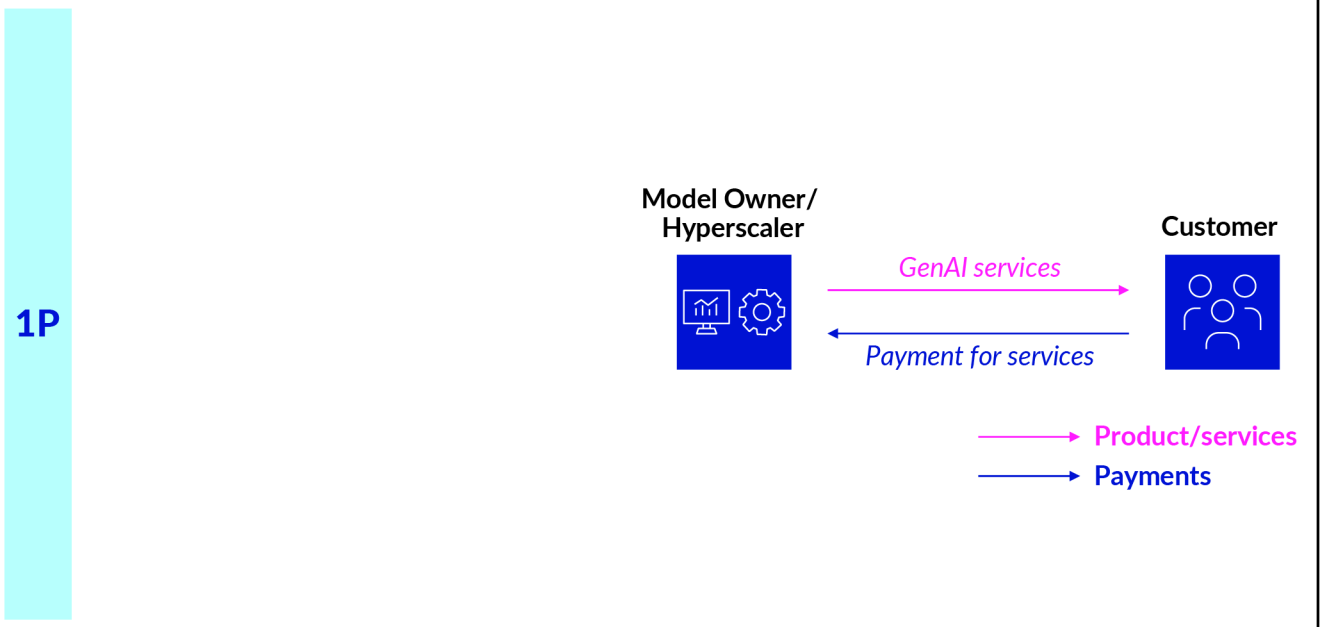
First Party Model

In a first party model, a customer purchases access to GenAI directly from the model owner/hyperscaler, which is the seller of record. In this model, a hyperscaler sells GenAI using its own proprietary foundation model (see Figure 3).

Generally, customer pricing is based on a tokenization metered pricing model — that is, a fixed dollar amount for a fixed number of input and output tokens. Input tokens refer to the amount of text that a model processes as an input, and output tokens refer to the amount of text a model produces as a response. This pricing model is also relevant for second and third party models.

[79] *Id.*

Figure 3. 1P Model



Second Party Model

In the second party model, a customer purchases access to GenAI from the hyperscaler, which is the seller of record. The hyperscaler separately enters into a contract with the model owner, enabling the hyperscaler to provide access to the model to the hyperscaler's customers. Generally, customer pricing is on a tokenization metered pricing basis. Hyperscalers often compensate the model owner on a revenue-sharing basis (see Figure 4).

Third Party Model

In a third party model, a customer purchases access to GenAI directly from the model owner, which is the seller of record. The model owner separately enters into a contract with the hyperscaler to act as a marketplace (facilitating the transactions between model owners and customers); host its model; and provide the cloud computing services necessary to facilitate the delivery of its model to its customer. Both model owners and hyperscalers typically offer their proprietary models on unrelated hyperscalers' marketplaces (see Figure 5).

Typically, the customer pays the hyperscaler, which collects on behalf of the model owner. The hyperscaler then remits this payment to the model owner, less a service fee for facilitating the sale through its platform and a fee for storage and compute consumed by the customer.

Software-as-a-Service Business Models

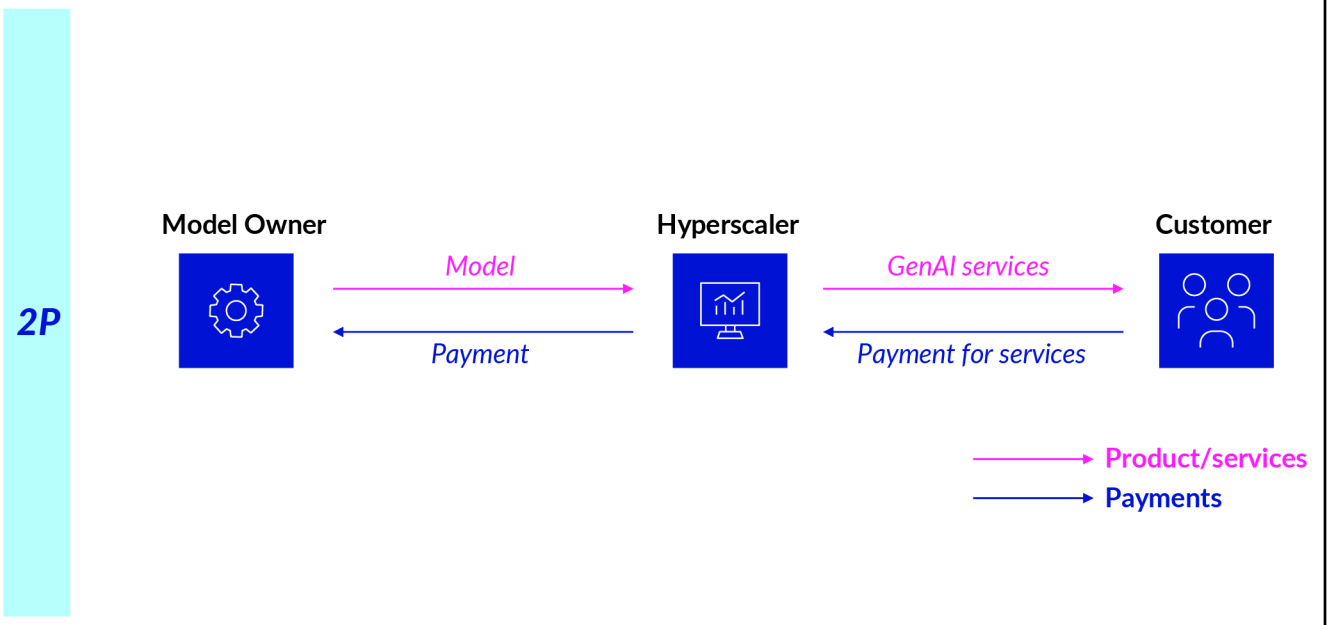
The types of business models used to provide GenAI services are not new; SaaS models are typically delivered in a similar manner. For example, Google sells both AI solutions and software solutions through Google Cloud.[80]

In these models, hyperscalers host software applications on their cloud platforms and make them available to customers online. Customers may purchase access to the hyperscaler's own software (1P model) or software provided by a third party (2P or 3P models). The hyperscaler manages all the hardware and software, including servers, storage, and networking. The customer typically pays a subscription fee to access the software, which can be on a per-user, per-month, or usage basis.

Because the business models used to support the provision of SaaS and GenAI services are very similar, the framework used to characterize the associated transactions for U.S. federal income tax

[80] *See* Google Cloud (last accessed Jan. 8, 2026).

Figure 4. 2P Model



purposes and non-U.S. tax purposes are similar (as we explore further later).

How Are Foundation Models Legally Protected?

One potential difference between GenAI transactions and other types of SaaS models could be the legal protection afforded to GenAI models versus other kinds of software. There is debate over whether the software, weights and biases, and algorithms that make up foundation models are eligible for copyright protection. This debate is distinct from other copyright questions, such as whether the outputs of GenAI are copyrightable.[81]

A comprehensive discussion of whether the different components of foundation models are copyrightable is beyond the scope of this article. The conclusion depends on the specific facts and circumstances of different models and the copyright law under consideration, which varies by country.[82] However, it is possible to make some general observations.

A computer program is eligible for copyright protection as a literary work in the United States.[83] Presumably, then, the computer program that embodies a foundation model is eligible for copyright protection.

For the weights and biases of foundation models to be copyrightable, they would likely need to qualify as original works of authorship. Some commenters argue that weights and biases files are mathematical expressions rather than creative works, making them ineligible for traditional copyright law protection.[84] The U.S. Copyright Office has published a report, focused on the output of GenAI models rather than the models themselves, which affirms that copyright protection in the United States requires human authorship,[85] a view that was supported by a 2023 decision by a district court in Washington, D.C.[86] Arguably, the weights and biases file is an output

[81] S. Alex Yang and Aine Doris, "To Copyright or Not to Copyright: The Big GenAI Question," Think at the London Business School, May 13, 2025.

[82] U.S. Copyright Office, Circular 38A: "International Copyright Relations of the United States" (Apr. 2025).

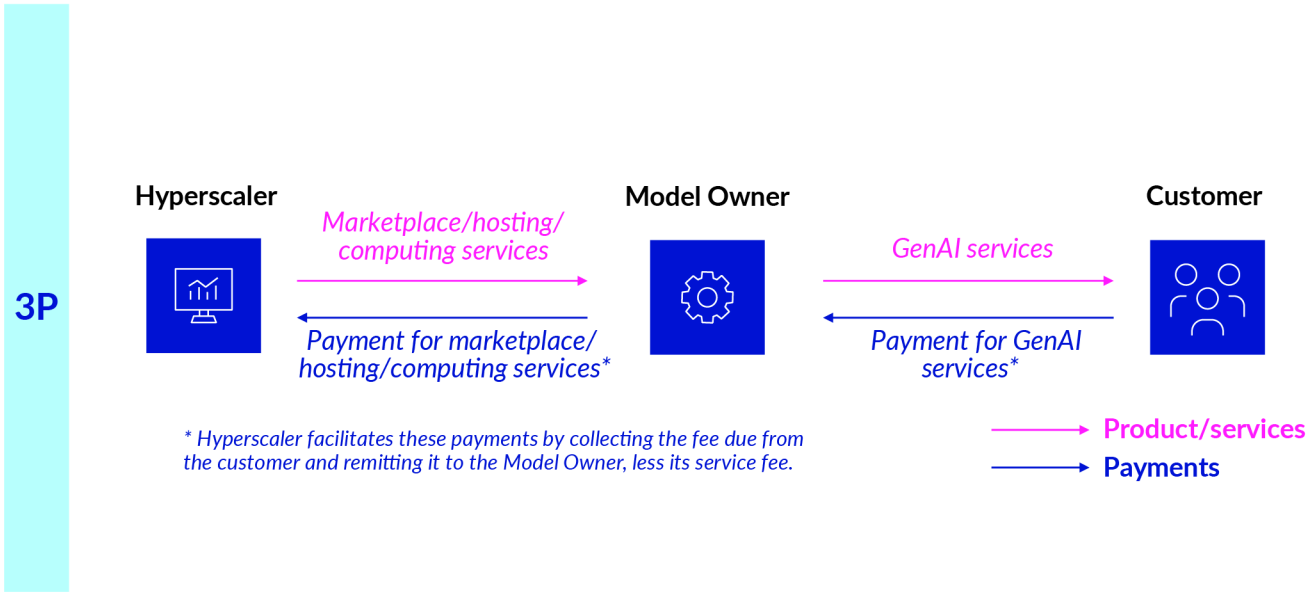
[83] 17 U.S.C. sections 101, 102; U.S. Copyright Office, Circular 61: "Copyright Registration of Computer Programs" (Mar. 2021).

[84] Nuno Sousa e Silva, "Are AI Models' Weights Protected Databases?" Kluwer Copyright Blog, Jan. 18, 2024.

[85] U.S. Copyright Office, "Copyright and Artificial Intelligence Part 2: Copyrightability" (Jan. 2025).

[86] Carl A. Kukkonen III and Emily J. Tait, "Court Finds AI-Generated Work Not Copyrightable for Failure to Meet 'Human Authorship' Requirement — But Questions Remain," Jones Day Insights (Aug. 2023).

Figure 5. 3P Model



of the computer program that embodies the model, because it is the output from the training, so authorities on the treatment of model outputs may be relevant to the weights and biases file.

For purposes of this article, we assume that the weights and biases files of foundation models are not eligible for copyright protection, while the computer program that embodies the model is copyrightable. This issue is discussed further below in the context of characterizing GenAI transactions for U.S. federal income tax purposes.

Note, however, that even if AI components (for example, weights and biases, algorithms) are not copyrightable, some commentators suggest that they may qualify as trade secrets. One law firm highlights that many intrinsic aspects of AI models, such as proprietary algorithms and the architecture of its neural network and weights, are likely ineligible for patent protection but eligible for trade secret protection under trade secrecy law:

> There may be difficulty patenting algorithms alone under [*Alice Corp.*].[87] But the algorithm or [AI model's] neural network design and implementation are eligible for trade secret protection if the statutory requirements are satisfied.[88]

The Copyright Office also acknowledges that computer source code may contain trade secrets.[89]

[87] *Alice Corp. v. CLS Bank International*, 573 U.S. 208 (2014), holding that abstract ideas are not patent-eligible unless there is a significant inventive concept that transforms the abstract idea into a patent-eligible invention.

[88] Quinn Emanuel Urquhart & Sullivan LLP, "The Rising Importance of Trade Secret Protection for AI Related Intellectual Property" (undated).

[89] The U.S. Copyright Office has made special allowances for the requirements to register computer programs containing trade secrets. A valid application for a copyright for a computer program requires submission or "deposit" of a copy of the computer program source code; however, the Copyright Office makes allowances for what must be submitted if the source code contains trade secrets. For example, one option, among several, to protect trade secrets contained in the program's source code is the acceptance of a submission of, "one copy of the first twenty-five pages and last twenty-five pages, blocking out the portions of the code containing trade secret material, provided the blocked-out portions are less than fifty percent of the deposit." Circular 61, *supra* note 83.

To claim trade secret protection for the weights and biases file of a foundation model, certain legal criteria must be met. First, the information must be secret and have economic value — it should derive value from not being generally known or readily ascertainable by others who could benefit from its disclosure or use.[90] In addition, reasonable efforts must be made to maintain secrecy — measures such as using nondisclosure agreements, limiting access, marking documents as confidential, or implementing cybersecurity measures.[91]

The application of trade secret protection to GenAI transactions is discussed in more detail below.

Existing Tax Characterization Rules and GenAI Transactions

We examine how existing tax rules (both U.S. federal income tax law and the 2017 OECD model tax convention on income and on capital (OECD model treaty)) apply to select GenAI transactions to determine their character — a threshold question in determining how transactions are taxed.

In analyzing the tax characterization of transactions required to deliver GenAI, we consider two existing frameworks: the existing Treasury regulations applicable to digital content and cloud transactions regarding U.S. federal income tax characterization; and the OECD model treaty for foreign treaty characterization.

The 2025 Regulations

In January 2025 Treasury and the IRS issued final regulations on the characterization of digital content and cloud transactions (digital content regulations under reg. section 1.861-18 and cloud transaction regulations under reg. section 1.861-19, collectively referred to as the 2025 regulations).[92] We summarize key aspects of them below in the context of GenAI transactions, with a particular focus on transactions involving hyperscalers.

The scope of the digital content regulations is limited to "digital content transactions," defined as any "transaction that constitutes a transfer of digital content, or the provision of services or of know-how with respect to digital content (each a digital content transaction)."[93] The first phrase of what is a digital content transaction — any "transaction that constitutes a transfer of digital content," is generally most relevant to the GenAI transactions discussed in this article. Transactions that are entirely or predominantly (see discussion of the predominant character rule below) digital content transactions are classified as one of the following four possible categories:

1. a transfer of a copyright right in the digital content;
2. a transfer of a copy of the digital content (a copyrighted article);
3. the provision of services for the development or modification of the digital content; or
4. the provision of know-how relating to development of digital content.[94]

Digital Content

The digital content regulations define digital content as:

> (i) . . . a computer program or any other content, such as books, movies, and music, in digital format that is —
>
> (A) Protected by copyright law; or
>
> (B) Not protected by copyright law solely —
>
> (1) Due to the passage of time; or
>
> (2) Because the creator dedicated the content to the public domain.
>
> (ii) Computer program defined. For purposes of this section, a computer program is a set of statements or instructions to be used directly or indirectly in a computer in order to bring

[90] 18 U.S.C. section 1839(3)(B), enacted by the Defend Trade Secrets Act of 2016. *See also* U.S. Patent and Trademark Office, "Intellectual Property Toolkits — Trade Secrets" (undated).

[91] 18 U.S.C. section 1839(3)(A).

[92] T.D. 10022.

[93] Reg. section 1.861-18(a)(1), (b)(1).

[94] Reg. section 1.861-18(b)(1). The provision of information regarding digital content involves the provision of know-how for purposes of this section only if the information is information relating to the development of digital content, furnished under conditions preventing unauthorized disclosure, specifically contracted for between the parties, and considered property subject to trade secret protection (reg. section 1.861-18(e)).

about a certain result and includes any media, user manuals, documentation, data base, or similar item if the media, user manuals, documentation, data base, or other similar item is incidental to the operation of the computer program.[95]

Are Foundation Models Digital Content?

The plain language of the digital content regulations provides that digital content is limited to computer programs and other copyright-eligible digital material; and the preamble to the 2025 regulations makes clear that unless the digital content regulations provide for an exception, regulations should be applied in a manner consistent with U.S. copyright law.[96]

Further support for the view that the digital content regulations are limited to copyrightable digital content is the fact that the regulation defines the term "computer program" using the exact language as the definition under U.S. copyright law.[97] Noncopyrightable elements of the software may be treated as part of the computer program for purposes of the digital content regulations, but only if these nonliteral elements are incidental to the operation of the computer program:

> For purposes of this section, a computer program is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result and includes any media, user manuals, documentation, data base, or similar item *if the media, user manuals, documentation, data base, or other similar item is incidental to the operation of the computer program*.[98] [Emphasis added.]

Therefore, components of a foundation model are treated as digital content for purposes of the digital content regulations only to the extent that those components are either: copyright-eligible (for example, we assume that the computer program that embodies the model is copyright-eligible); or ineligible for copyright, but merely incidental to the operation of the model.

Foundation models are composed of literal elements (such as computer code) and nonliteral elements (such as data, algorithms, and interfaces). As discussed, whether the weights, biases, and algorithms of foundation models are eligible for copyright protection is outside the scope of this article, but for purposes of this discussion, we assume that foundation model weights, biases, and algorithms are not eligible for copyright protection.[99] The weights and biases file could conceivably be treated as a database, but it seems unlikely to be considered incidental to the operation of the foundation model. Rather, the weights and biases file is the brain of the foundation model, critical to its success. Therefore, a foundation model may consist of both digital content (for example, the computer program) and nondigital content (for example, the weights, biases, and algorithms) for purposes of the digital content regulations.

Cloud Transaction Regulations

A cloud transaction is "a transaction through which a person obtains on-demand network access to computer hardware, digital content (as defined in reg. section 1.861-18(a)(2)), or other similar resources."[100] Cloud transactions involving digital content differ from digital content transactions, because in a cloud transaction the customer accesses the digital content remotely over on-demand networks rather than having received a transfer of the digital content. Cloud transactions also include transactions that may not involve digital content at all — that is, that involve only computer hardware and noncopyrightable content. A cloud

[95] Reg. section 1.861-18(a)(2)(i).

[96] In response to a comment requesting expanding of the scope of the digital content regulations beyond copyrightable digital content, Treasury said: "The final regulations do not broaden the definition of digital content beyond content protectable by copyright law. Section 1.861-18, as in effect before this Treasury decision, generally followed copyright law, and the Treasury Department and the IRS are of the view that it is appropriate to continue to apply this longstanding copyright law framework." T.D. 10022, 90 F.R. at 2979.

[97] "A computer program is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result." 17 U.S.C. section 101 (subject matter and scope of copyright); *see also* Circular 61, *supra* note 83.

[98] Reg. section 1.861-18(a)(2)(ii).

[99] Facts of transactions and future legal developments may render this assumption inapplicable to actual transactions.

[100] Reg. section 1.861-19(b).

transaction is classified as the provision of services.[101]

Predominant Character Rule

Under the 2025 regulations, a transaction with multiple elements is characterized based on the predominant character of the transaction. The digital content regulations provide that a transaction that has multiple elements, one or more of which would be a digital content transaction if considered separately, is classified in its entirety as a digital content transaction described in one of the categories of reg. section 1.861-18(b)(1) if the predominant character of the transaction is one of the categories in that paragraph.[102] The cloud transaction regulations provide for a corresponding rule (by cross-reference to reg. section 1.861-18(b)(2)) for transactions that have multiple elements, one or more of which is a cloud transaction.[103]

The digital content regulations and the cloud transaction regulations do not define the terms "transaction" or "element." The preamble to the 2025 regulations says that ordinary tax principles apply to determine the scope of a transaction.[104] Defining a transaction for U.S. federal income tax purposes is well outside the scope of this article, so we note our assumptions as to what constitutes one or multiple transactions. The examples in the 2025 regulations suggest that an element of a transaction may be any property or services that a customer receives as part of the transaction that can be separately identified. Examples in the digital content and cloud transaction regulations indicate that each of the following can be separate elements when transferred together in a single transaction:

- on-demand network access to software;[105]
- downloads of scripting code;[106]
- downloads of software programs;[107]
- website hosting;[108]
- on-demand network access to hardware;[109]
- the right to distribute copies of software components;[110] and
- the right to modify source code.[111]

The predominant character of a transaction is determined under a three-prong test. First, "the predominant character of a transaction is determined by ascertaining the primary benefit or value received by the customer in the transaction."[112] If the primary benefit to the particular customer is not reasonably ascertainable, the taxpayer must look to the primary benefit a typical customer in similar circumstances receives as determined by typical customer use and access data.[113] If this data is not available, the predominant character is determined by "other factors that are indicative of the primary benefit or value received by a typical customer, including: (i) how the transferor or provider markets the transaction; (ii) the relative development costs to the transferor or provider of each element of the transaction; and (iii) the relative price paid in an uncontrolled transaction for one or more elements compared to the total contract price of the transaction in question."[114]

Applying the 2025 Regulations to GenAI Transactions

Below we consider the characterization of different types of GenAI transactions based on the 2025 regulations. We consider two sets of transactions in the 2P model: the hyperscaler-customer transaction and the model owner-hyperscaler transaction.

In the hyperscaler-customer transaction, the hyperscaler, as the seller of record, sells access to the model owner's foundation model for the purpose of providing to the public:

[101] Reg. section 1.861-19(c)(1).
[102] Reg. section 1.861-18(b)(2).
[103] Reg. section 1.861-19(c)(2).
[104] T.D. 10022, 90 *F.R.* at 2978.
[105] *See, e.g.*, reg. section 1.861-19(d)(3), Example 3.
[106] *Id.*
[107] *Id.*, examples 4, 5.
[108] *Id.*, Example 3.
[109] *See, e.g.*, reg. section 1.861-19(d)(3), Example 4.
[110] Reg. section 1.861-18(b)(h), Example 17.
[111] Reg. section 1.861-18(b)(h), Example 18.
[112] Reg. section 1.861-18(b)(3)(i).
[113] Reg. section 1.861-18(b)(3)(ii)(A).
[114] Reg. section 1.861-18(b)(3)(ii)(B).

1. **General Inference Services** (GenAI inference services using the foundation model's original weights and biases);
2. **Customer Fine-Tuning Services** (assisting customers to fine-tune the foundation model's weights and biases through the hyperscaler's AI platform); and
3. **Custom Inference Services** (GenAI inference services using customer fine-tuned weights and biases).

For purposes of this discussion, we assume that customer fine-tuning and custom inference services are not bundled with any other product or service — that they constitute a single transaction and there are no other elements of the transaction that are relevant for characterization.

In the model owner-hyperscaler transaction, the model owner transfers to the hyperscaler the right to use its foundation model in the provision of GenAI services to the public through a 2P model. Below we also consider the arrangement between a model owner and hyperscaler in a 3P model in which the hyperscaler hosts the model owner's foundation model on its platform.

Customer Payments to Hyperscalers

In all three GenAI business models (1P, 2P, and 3P), customers pay hyperscalers or model owners for access to general inference services, customer fine-tuning services, and custom inference services. We examine these transactions below.

General Inference Services

In the provision of general inference services, the customer receives the right to on-demand network access to computer hardware, copyrightable digital content (the computer program that embodies the foundation model), and other similar resources (including the weights and biases file). We understand that it is rare for the foundation model to be transferred to the customer in the provision of general inference services. As such, this discussion assumes that the foundation model is not transferred to the customer; however, it is possible that in the future a transfer of the model to the customer may be more common, if, for example, models become simplified or devices that are powerful enough to run the models become more widely owned.

When the model itself is not transferred, customer payments to hyperscalers should generally be characterized as on-demand network access to computer hardware, digital content, or other similar resources — that is, a cloud transaction — and the transaction should be characterized as a service under reg. section 1.861-19(c)(1).

Customer Fine-Tuning Services and Custom Inference Services

In addition to general inference services, hyperscalers may also provide customer fine-tuning services and custom inference services.

For purposes of this article, we assume that both services together constitute a single transaction, because both elements are required to effectuate the desired outcome — you would not fine-tune a model if you did not intend to use custom inference services enabled by the fine-tuned model, and you cannot use custom inference services if you do not fine-tune the model. Based on that assumption, the question for purposes of characterizing the transaction under the 2025 regulations is what the predominant character of the transaction is. The answer likely depends on whether the customer receives a property right because of the fine-tuning. If the customer fine-tuning does not create a property right for the customer, then it is likely merely on-demand network use of the foundation model by the customer.

In general, customer fine-tuning does not require the customer to have access to the foundation model's source code, and the customer does not alter the computer program, model architecture, or learning mechanisms. Instead, the customer's copy of the model's weights and biases file is modified by further learning on the customer's proprietary domain-specific data.[115] The customized weights and biases file is a newly created digital file that is valuable to the customer and constitutes property created by: the foundation model itself (through its autonomous learning mechanisms); the customer's (and possibly hyperscaler's) efforts in directing the fine-tuning process; and the customer's proprietary training data.

[115] *See* Bergmann, *supra* note 78; Jeffrey Erickson, "Introduction to Fine-Tuning in Machine Learning," Oracle (Feb. 25, 2025).

The relevant legal agreements generally provide that the customer has no copyright rights in the foundation model and has only a nonexclusive right to use the property over an on-demand network. In that case, the only property in which the customer could possibly have a property right is the newly created customized weights and biases file and the outputs from the use of that file.

In many cases when customer fine-tuning services are provided, the customer does not have legal ownership of the modified weights and biases file, the right to possess the file, or the right to transfer the file to another party (including another cloud service provider for the customer's own benefit and use), and the intent of the parties is that fine-tuning does not give the customer ownership of the file. The only indicia of ownership that the customer often obtains is the right to the income the customer might generate by using the file. In light of the many negative factors above, the right to income from use of the file is insufficient to establish that the customer has possession of or other ownership rights in the customized weights and biases file.[116] In other words, the customer typically has only the right to on-demand network access to the modified weights and biases file. The cloud transaction regulations classify any on-demand network access to computer hardware, digital content, or similar resources as a cloud transaction, and therefore a service. Even if a foundation model is not digital content for this purpose, it is clearly a similar resource. A transaction consisting of customer fine-tuning services and custom inference services is therefore likely to be treated as a service transaction in its entirety under the cloud transaction regulations.

Hyperscaler-Customer Transaction Conclusion

We generally expect payments from customers to hyperscalers to be treated as services for U.S. federal income tax purposes. Services income is sourced as either U.S.-source or foreign-source based on the location of the performance of the services.[117] The place where the services are performed does not depend on the residence of the payer, the location the contract was made, or the place or time of payment.[118] In one notable case, the court determined that the location of the capital and the labor used to perform the services generally determines the source of income.[119] Thus, the U.S. rules generally provide that income is sourced, for federal tax purposes, to the jurisdiction in which the productive activities that gave rise to the income took place.[120] Under that approach, in the case of a payment by a customer to a hyperscaler for general inference and customer fine-tuning services, the services income would be sourced based on the location of the hyperscaler's assets and personnel providing them.

Hyperscaler Payments to the Model Owner in a 2P Model

In a 2P model, a model owner typically transfers a copy of the foundation model to the hyperscaler along with the right to use the foundation model in the provision of general inference services, customer fine-tuning services, and custom inference services. We analyze this arrangement below on the assumption that it gives rise to two separate transactions: one for payments for the ability to provide general inference services, and one for payments for the ability to provide customer fine-tuning and custom inference services. However, whether such an arrangement gives rise to one transaction or two depends on the facts.

General Inference Services

For purposes of the predominant character rule of reg. section 1.861-18(b)(2), a transaction granting the hyperscaler the ability to provide general inference services likely consists of four elements:

1. the transfer of a copy of the foundation model;

[116] For a discussion of the factors giving rise to full benefits and burdens of ownership, see, *e.g.*, *Derr v. Commissioner*, 77 T.C. 708, 723 (1981); and Rev. Rul. 72-252, 1972-1 C.B. 193.

[117] Sections 861(a)(3), 862(a)(3).

[118] Reg. section 1.861-4(a)(1).

[119] *Commissioner v. Piedras Negras Broadcasting Co.*, 127 F.2d 260 (5th Cir. 1942).

[120] Note that this is a different question than which jurisdiction may have taxing rights, which we address in more detail in the discussion of the OECD model treaty.

2. the right to make copies of digital content (that is, the foundation model's computer program) to provide general inference services;
3. the right to make copies of noncopyrightable digital material (for example, the foundation model's original weights and biases, which we have assumed are not digital content) to provide general inference services; and
4. the provision of maintenance, support, and integration services by the model owner.

The right to copy the foundation model's computer program, if considered separately, would be a digital content transaction, and assuming the weights and biases file is not eligible for copyright protection, the right to copy the foundation model's weights and biases likely would not be a digital content transaction, if considered separately. The weights and biases file could be treated as digital content, however, if it is merely incidental to the operation of the computer program.[121]

As noted above, the predominant character rule looks first to "the primary benefit or value received by the customer in the transaction."[122] If that is not reasonably ascertainable, the rule then looks to "the primary benefit or value received by a typical customer in a substantially similar transaction . . . as determined by data on how a typical customer uses or accesses the digital content."[123] If that is not possible, the rule then looks to "other factors that are indicative of the primary value or benefit received by a typical customer."[124]

In the case of payments by hyperscalers to model owners, the predominant element may differ depending on the facts. The following examples assume that the transfer of either the computer program or the weights and biases file is the predominant element of the arrangement, but other arrangements are possible. For example, under some circumstances the arrangement could be characterized as the model owner receiving services income and the hyperscaler performing services for the model owner, with the transfer of the model to the hyperscaler treated as merely facilitating the services the hyperscaler provides (similar to the 3P model, discussed below).

Computer Program as the Predominant Element. If the hyperscaler can demonstrate that the predominant element of the transaction is the right to make copies of the model's computer program for use in providing general inference services, the transaction is likely characterized as a transfer of a copy of a copyrighted article (the computer program). This type of transaction is common and is often referred to as a service provider license agreement arrangement.

Under the digital content regulations, the right to make copies of computer programs by itself is not treated as a transfer of copyright rights. It is only when the right to make copies is coupled with the right to distribute copies to the public that there is a transfer of a copyright right for purposes of the digital content regulations.[125]

Assuming there is no transfer of a copyright right under reg. section 1.861-18(c)(2), the transaction would be treated as the transfer of a copyrighted article. If the benefits and burdens of ownership are transferred to the hyperscaler, the transaction would likely be characterized as a sale.[126] If the benefits and burdens of ownership are not transferred, the hyperscaler payments to the model owner would likely be characterized as lease payments.[127]

Weights and Biases as the Predominant Element. If making copies of the weights and biases file for use in providing general inference services is the predominant element of the transaction, it is important to determine the status of the weights and biases file. We assume that the file is not copyrightable, and therefore it would not be a digital content transaction subject to the digital content regulations. However, the file may be eligible for trade secret protection. That question is beyond the scope of this article, and the outcome depends upon specific facts. If the

[121] Reg. section 1.861-18(a)(2)(ii).

[122] Reg. section 1.861-18(b)(3)(i).

[123] Reg. section 1.861-18(b)(3)(ii)(A).

[124] Reg. section 1.861-18(b)(3)(ii)(B).

[125] Reg. section 1.861-18(c)(2)(i).

[126] Reg. section 1.861-18(f)(2).

[127] *Id.*

file is treated as a trade secret or some other type of intangible property, the transfer is likely[128] to be treated as a sale or license, depending on whether the benefits and burdens of ownership have transferred to the hyperscaler.[129]

We note, for completeness, that it seems artificial to treat a foundation model differently from other software for purposes of the digital content regulations. Fundamentally, a foundation model is a computer program that references the data in the weights and biases file, and these two elements are integrally related (both are required for the foundation model to work). While the weights and biases file may not be copyrightable, there are typically contractual arrangements in place between the transferor and transferee that provide the transferor with contractual legal rights that mirror the U.S. copyright protection the weights and biases would have if they were copyrightable. In other words, the weights and biases file may benefit from other IP protection, such as for trade secrets. Under those circumstances, the most appropriate tax authority to govern the transaction may be the digital content regulations. The policy concerns expressed in the preamble to the digital content regulations about expanding their scope beyond copyrightable material seem inapposite in that case. Applying the digital content regulations would treat the weights and biases file in the same way digital content is treated, meaning that the transaction between the model owner and hyperscaler likely would be characterized as a lease of a copyrighted article, rather than as a license, in the same way the digital content regulations characterize time-limited service provider license agreement arrangements for computer software.

Customer Fine-Tuning and Custom Inference Services

The model owner may grant the hyperscaler the right to provide customer fine-tuning and custom inference services to customers. For purposes of the predominant character rule of reg. section 1.861-18(b)(2), the grant of the ability to provide these services likely consist of five elements:

1. the transfer of a copy of the foundation model;
2. the right to make copies of digital content (the foundation model's computer program) to provide custom inference services;
3. the right to make copies of noncopyrightable digital material (for example, the foundation model's original weights and biases, which we have assumed are not digital content) to provide custom inference services;
4. the right to facilitate customer fine-tuning that gives rise to modified weights and biases; and
5. the provision of maintenance, support, and integration services by the model owner.

While the answer may depend on specific facts, we expect that generally the predominant character in this case would be the right to facilitate modifications to the weights and biases file through fine-tuning to provide custom inference services, because we expect that to be the primary benefit received by the hyperscaler.

In that case, the transaction between the hyperscaler and model owner would not be covered by the digital content regulations because the predominant element of the transaction is not a transaction involving digital content, and thus general tax principles would apply to characterize the transaction. As noted above, when the predominant element is the transfer of nondigital content IP like the weights and biases, general tax principles generally characterize the transaction as giving rise to royalties, assuming that the model owner retains significant rights in the model.

Again, it may be possible to argue that the transaction should be governed by the digital content regulations, for the reasons described above. In that case, the transaction would likely be characterized as a lease of a copyrighted article,

[128] Here, again, different facts or context may result in different characterization.

[129] *See* Rev. Rul. 69-482, 1969-2 C.B. 164 (general tax principles apply to transfers of IP outside the scope of section 1235); and *Rollman v. Commissioner*, 244 F.2d 634 (4th Cir. 1957) (the transfer of a manufacturing process gave rise to royalties because the transferor withheld the right to sublicense and retained ongoing control).

unless the right to make modified weights and biases files were treated as the right to make derivative works.

Model Owner-Hyperscaler Transaction Conclusion

Provided the benefits and burdens of ownership of the foundation model are not transferred to the hyperscaler, we generally expect the payments from the hyperscaler to the model owners in the transactions described above to be treated as either rents (if the computer software is the predominant element of the transaction) or royalties (if the weights and biases file is the predominant element of the transaction).

Rents or royalties from property located in the United States or for the use of or for the privilege of using in the United States patents, copyrights, trademarks, and similar IP, are generally treated as income from sources within the United States.[130] A parallel rule applies for property that is located or used, or for use outside of, the United States.[131] Thus, to determine whether rent and royalty income is treated as U.S.- or foreign-source, it is necessary to determine the location where the property is located or used.

There is limited case law and IRS guidance addressing where IP is located or used for this purpose. Case law regarding patents indicates that the place of use is based on the location where the licensee (the hyperscaler in this case) conducts the economic activities giving rise to the royalty payments.[132] The IRS, on the other hand, has taken the position in three revenue rulings that copyright and trademark rights are used in the location in which the copyrighted and trademarked goods are consumed and the copyright and trademark rights are protected by law.[133]

Model Owner Payments to Hyperscaler in a 3P Model

In a 3P model the customer typically pays the hyperscaler, which collects on behalf of the model owner, and then remits the payment to the model owner. The model owner also pays a service fee to the hyperscaler for storage and compute consumed by the customer in operating the model (for inference of tokens) and also often a fee for facilitating the sale through the hyperscaler's platform. Although there is a transfer of the foundation model to the hyperscaler, the hyperscaler is merely acting as host for the foundation model to enable its use. Based on that paradigm, the model owner is paying the hyperscaler to host the foundation model, the hyperscaler is not paying the model owner for the foundation model.

The predominant character of the transaction thus is not a transfer of the foundation model to the hyperscaler; instead, the hyperscaler provides storage and computing services. On that basis, we do not expect the digital content regulations to apply.[134] Assuming the payment to the hyperscaler is treated as one transaction, as relevant here, examples in the cloud transaction regulations indicate that providing storage and computing services is generally considered a cloud transaction,[135] and acting as a marketplace for GenAI services would also be expected to be similar to website hosting such that it is treated as a cloud transaction under the cloud transaction regulations.[136] Thus, in this case, we expect the predominant element of the transaction to be a cloud transaction and therefore a service under reg. section 1.861-19(c)(1).

Hyperscaler-Model Owner Transaction Conclusion

Based on this discussion, we expect model owner payments to hyperscalers in a 3P model to be treated as payments for services for U.S. federal income tax purposes. As discussed, services income is sourced based on the location

[130] Section 861(a)(4).

[131] Section 862(a)(4).

[132] The main case interpreting the place-of-use test is *Sanchez v. Commissioner*, T.C. 1141 (1946), *aff'd*, 162 F.2d 58 (2d Cir. 1947), which addressed the source of royalties paid by a U.S. corporation to a nonresident alien individual for the right to use U.S. and foreign patent rights.

[133] *See* Rev. Rul. 68-443, 1968-1 C.B. 304; Rev. Rul. 72-232, 1972-1 C.B. 276; and Rev. Rul. 84-78, 1984-1 C.B. 173.

[134] *See* reg. section 1.861-18(b)(1).

[135] *See* reg. section 1.861-19(d), Example 7 (stating that the provision of data storage and computing power is a cloud transaction).

[136] *See* reg. section 1.861-19(d), Example 3 (illustrating that hosting of websites is considered a cloud transaction).

of the performance of the services.[137] Thus, in the case of a payment by a model owner to a hyperscaler, the services income may be sourced based on the location of the hyperscaler assets and personnel providing the services.

OECD Model Treaty Characterization

We now turn to the OECD model treaty and how it may characterize the same transactions.

Overview

Analyzing the treatment of GenAI transactions under the laws of all (or even any meaningful subset) of non-U.S. jurisdictions is obviously impractical in an article, so we look to the OECD model treaty for high-level guidance regarding how other jurisdictions may treat the transactions. The OECD model treaty is designed to clarify, standardize, and confirm the fiscal situation of taxpayers involved in international activities, serving as a means to settle common problems arising from international juridical double taxation.[138] The commentaries on the articles of the OECD model treaty serve to illustrate and interpret its provisions. Tax authorities of OECD member countries may follow these commentaries, as modified over time and subject to their observations, when applying and interpreting the provisions of their bilateral tax conventions that are based on the OECD model treaty.[139] The commentaries also set out the position of some non-OECD members, regarding both the OECD model treaty and the commentary.[140]

Given the transactions described above, the articles of the OECD model treaty that are most likely to be relevant to characterizing the income from GenAI transactions are articles 7 (business profits) and 12 (royalties).

Articles 7 and 12

Article 7 establishes the principle that profits of an enterprise are taxable only in the state of residence unless the enterprise carries on business through a PE in the other contracting state.[141] If a PE exists, the other state may tax only the profits attributable to that PE.[142] However, article 7 includes a crucial override — if the profits include items of income dealt with separately in other articles of the convention, then the provisions of those other articles won't be affected by article 7.[143] This means that if an income item clearly fits the definition in a specific article (for example, article 12 for royalties), that specific article's rules apply, not article 7.

Article 12 grants exclusive taxing rights to the state of residence of the beneficial owner for royalties arising in the other contracting state.[144] However, many bilateral treaties provide for positive rates of withholding on royalties. The term "royalties" is defined broadly to mean:

> payments of any kind received as a consideration for the use of, or the right to use, any copyright of literary, artistic or scientific work including cinematograph films, any patent, trade mark, design or model, plan, secret formula or process, or for information concerning industrial, commercial or scientific experience.[145]

Application of Articles 7 and 12 to GenAI Transactions

In this article we consider payments by customers in one jurisdiction (which we refer to as the source state, which should not be confused with the source of the payment for U.S. federal income tax purposes) to the hyperscaler in another jurisdiction, and payments from the hyperscaler to a model owner in another jurisdiction. For purposes of this discussion, we assume that neither the hyperscaler nor the model owner maintains a PE in the source state; both the hyperscaler and model owner derive items of income from the source state in the ordinary course of their business in the resident state; and absent a specific income article applying to the income (for example, royalties under article 12),

[137] Sections 861(a)(3), 862(a)(3).

[138] Paragraph 3 of the introduction to the OECD model treaty.

[139] Paragraphs 29 and 30 of the introduction to the OECD model treaty.

[140] Paragraphs 1 through 5 of Non-OECD Economies' Positions on the OECD Model Tax Convention in the commentary.

[141] Art. 7(1) of the OECD model treaty.

[142] *Id.*

[143] Art. 7(4) of the OECD model treaty.

[144] Art. 12(1) of the OECD model treaty.

[145] Art. 12(2) of the OECD model treaty.

the income derived by the hyperscaler and model owner would be considered business profits under article 7. Therefore, the remaining part of this section analyzes whether the relevant income earned in GenAI transactions would be considered a royalty under article 12.[146]

Customer Payments to Hyperscalers

General Inference Services. In a scenario in which a customer pays a hyperscaler for the receipt of general inference services but does not itself transfer the model to the customer, we expect the payment from the customer to the hyperscaler to be classified as business profits under article 7. This is because the customer is essentially receiving a service for the use of a foundation model to run inference services, without a transfer of the foundation model, and is not acquiring IP rights in the model itself.

As a starting point, under the OECD model treaty, elements within a contract are generally classified separately unless the element is de minimis as compared with the primary element.[147] Note that this approach is fundamentally different from the approach taken in the 2025 U.S. regulations, which characterize a transaction based on its predominant character.

The two most likely characterizations for the payment from the customer to the hyperscaler are as payment for services (which would be business profits under article 7) or as a royalty. Thus, we examine article 12 to determine whether the payment could constitute a royalty. If not, the payment is likely to constitute a payment for services.

As noted above, article 12 defines royalties as "payments of any kind received as a consideration for the use of, or the right to use, any copyright of literary, artistic or scientific work including cinematograph films, any patent, trade mark, design or model, plan, secret formula or process, or for information concerning industrial, commercial or scientific experience."[148] While this definition is potentially very broad, the commentary to article 12 provides additional guidance on the scope of the definition.

In particular, the commentary includes specific guidance on transactions involving computer software, which should be instructive in the case of GenAI, not least because foundation models are embodied in computer software. The guidance on computer software is very similar to the previous version of reg. section 1.861-18, and is focused on the extent to which a transaction involving computer software involves a use of the underlying copyright in the software.[149] Notably, if the payment is "for something other than the use of, or right to use, rights in the copyright" and the use of copyright is required to enable a user to download, store, and operate software on the customer's computer or network, that type of limited use should be disregarded in the determination of whether the payment is a royalty.[150] Moreover, a payment for electronically downloading digital products (for example, software, images, sounds, or text) for the customer's own use or enjoyment is "essentially for the acquisition of data transmitted in the form of a digital signal and therefore does not constitute royalties but falls within Article 7 or Article 13 [capital gains], as the case may be."[151] Even if the act of copying the digital signal involves a use of copyright by the customer, this is considered "merely the means by which the digital signal is captured and stored" and is "not important for classification purposes" because it's not the essential consideration for the payment.[152]

This guidance on digital signals directly applies to GenAI services, in which the customer receives generated text or output (data) for its use, not the underlying model or rights to exploit any copyright in the model. In a typical GenAI inference services transaction, a customer does not download, store, and operate the foundation model. The customer can only use the model on

[146] This discussion is not academic. The Australian Taxation Office has publicly stated that it is exploring the application of withholding tax to transactions involving cloud service providers. See ATO Deputy Commissioner Rebecca Saint, "Key Developments in Tax Administration in Australia," speech to Pacific Rim Tax Conference, highlighting data centers and cloud services as an emerging issue (June 14, 2024).

[147] Paragraph 11.6 of the commentary to art. 12.

[148] Art. 12(2) of the OECD model treaty.

[149] Paragraph 13.1 of the commentary to art. 12.

[150] Paragraphs 14 and 17.2 of the commentary to art. 12.

[151] Paragraph 17.3 of the commentary to art. 12.

[152] *Id.*

the hyperscaler's platform, leaving the customer's interaction limited to operating or using the foundation model's functions, not acquiring the software program or the weights and biases file, or rights to reproduce or distribute the foundation model. Those facts strongly suggest that the customer is paying for services, rather than paying a royalty.

In discussing transponder leasing for satellite capacity, the commentary to article 12 says that payments for merely using a facility's capacity for transmission, without acquiring physical possession or technology transfer, are typically "payments for services, to which Article 7 applies, rather than payments for the use, or right to use, [industrial control system] equipment."[153] This guidance reinforces the idea that consuming a functional output generated by a foundation model, a complex technology, without receiving either a transfer of the model or gaining possession or ownership of the technology itself, constitutes a service.

Some governments are unlikely to be happy with a characterization that treats payments by customers as business profits that are not subject to source-country taxation in the absence of a PE. Because at least some treaties provide for positive rates of withholding on royalties, some governments may try to argue that the payments should be characterized as royalties, and those arguments are likely to focus on the commentary's discussion of know-how. Indeed, we are aware of some governments already making those arguments, but we think they are unfounded.

Paragraph 11 of the commentary to article 12 discusses payments made for information concerning industrial, commercial, or scientific experience. It distinguishes between "know-how" (which generates royalties) and the "provision of services" (which generates business profits under article 7), and clarifies that a contract for the provision of services involves one party undertaking to use its customary skills to execute work for the other party, without transferring special knowledge or experience to that other party to use for its own account.[154] Know-how typically involves the supply of existing, secret, and valuable undivulged information, with little additional effort from the supplier.[155]

In the context of GenAI services, the hyperscaler generally uses its expertise and the foundation model to generate an output for the customer, but the customer does not gain possession of the underlying know-how, secret formula, or any other specialized knowledge of the internal working of the foundation model. This indicates that the customer payments would be considered for a provision of services and business profits under article 7 rather than for know-how and royalties under article 12.

Customer Fine-Tuning. Even if customers are allowed to fine-tune the foundation model, that fact should not lead to a different conclusion. The only difference in this scenario is that with fine-tuning, the output of the foundation model is more tailored to the customer's needs based on the additional data on which the foundation model has been trained. The customer still does not receive a transfer or copy of the fine-tuned foundation model and does not obtain a property or possessory right in the IP of the model.

Hyperscaler-Customer Transaction Conclusion. Overall, we expect the payments from customers to the hyperscaler to be characterized as hyperscaler business profits under article 7. The commentary to article 12 supports the view that for the customer-to-hyperscaler transaction, the customer is paying for a service. This is because the customer does not receive a transfer of the foundation model and does not receive the use of or right to use IP (for example, a copyright in the model or its underlying know-how/secret formula). The hyperscaler is essentially offering a computational service. This characterization of the customer payments aligns with the OECD model treaty's treatment of similar non-GenAI cloud transactions. Even in the case of customer fine-tuning, the customer is paying for the *performance* of this service and the *output* it generates, not for the right to exploit the foundation model's IP.

[153] Paragraph 9.1 of the commentary to art. 12.

[154] Paragraph 11.2 of the commentary to art. 12.

[155] Paragraph 11.1 of the commentary to art. 12.

Hyperscaler Payments to Model Owners

The OECD model treaty approach to characterization of payments made from hyperscalers to model owners to provide access to GenAI in a 2P model may depend on both whether the foundation model is treated as a computer program in its entirety and whether customer fine-tuning services are provided. While we analyze this transaction on the basis of the transfer of the foundation model to the hyperscaler, we note again, for completeness, that under some circumstances the arrangement between the model owner and the hyperscaler may be treated in other ways. For example, the model owner may receive services income and the hyperscaler may perform services for the model owner, the transfer of the model to the hyperscaler treated as merely facilitating the services provided by the hyperscaler.

Foundation Model as a Computer Program. Whether the payments are classified as royalties under article 12 depends on the nature of the rights in the IP acquired by the user. In common software transactions in which the user obtains a program copy, the rights acquired by the transferee or user are limited to those necessary to enable the user to operate the program for its own business purpose.[156] While this may include minimal rights to copy the program, such as copying the program onto the user's computer hard drive or making an archival copy, these types of copying may be regarded as essential steps to enable a user to use the program.[157]

Payments for software are classified as royalties when the consideration is for the granting of rights to use the underlying copyright itself, such as rights relating to the reproduction for public distribution or modification of the program.[158] If the reproduction rights granted do no more than enable the effective use of the program by the user, these rights are limited and may be disregarded for purposes of classifying the payment under article 12.[159] As such, a transfer of a computer program to a user for use by the user in conducting its own business is not a royalty, and is considered business profits under article 7 to the enterprise providing the program copy.

If a jurisdiction viewed a foundation model as a computer program in its entirety, and the hyperscaler were only granted a right to use the program to provide services in its business — without the right to copy and publicly distribute the program — payments for that transaction likely would not be considered a royalty. In other words, if a foundation model is treated in its entirety as a computer program (or software) and is transferred only to be used in the user's business without any further use of underlying copyright rights, that transaction may be analogous to the lease of a copy of a computer program to a user for use in the user's business. Payments made in these types of transactions are viewed as business profits covered by article 7.

Weights and Biases File as Trade Secrets. If, however, the transfer of a foundation model to a hyperscaler is not considered a computer program in its entirety, then it must be determined whether a type of IP right described within article 12 has been provided. Consider a scenario in which the hyperscaler receives a copy of the foundation model and the rights to deploy the model for its own use, providing GenAI services to the public, including general inference services. The weights and biases file conceivably could be considered a scientific work or secret formula or process, or embody industrial, commercial, or scientific experience (often referred to as know-how) covered by article 12(2). Know-how generally refers to "undivulged information of an industrial, commercial or scientific nature arising from previous experience, which has practical application in the operation of an enterprise and from the disclosure of which an economic benefit can be derived."[160]

The weights and biases file within a foundation model could be considered know-how or a secret process or formula of the model under article 12(2), particularly because the actual underlying data of these weights and biases is transferred or licensed for the recipient's own use

[156] *See* paragraph 14 of the commentary to art. 12.

[157] *Id.*

[158] Paragraph 13.1 of the commentary to art. 12.

[159] Paragraph 14 of the commentary to art. 12.

[160] Paragraph 11 of the commentary to art. 12.

and exploitation, and because of the restrictions and limitations on both using and maintaining the confidentiality of the model. The weights and biases file in a foundation model contains the numerical parameters that a model learns during its training process. These parameters are critical to the model's functionality and performance. The specific values of these weights and biases are typically highly confidential and proprietary to the model owner. They are generally not publicly disclosed and may be protected as trade secrets. The development and training of a foundation model, which results in these weights and biases, involves immense computational resources, highly specialized algorithms, and deep scientific and engineering expertise. The final set of weights and biases represents the accumulated experience and logic derived from processing vast amounts of data. These weights and biases are the brain of the foundation model, enabling it to perform its functions (for example, generating text, translating, and answering questions). Their disclosure would allow others to replicate the foundation model's capabilities without the original development cost, thus conferring potentially significant economic benefit to the recipient of the information.

As such, there is risk that a tax administration may assert that the weights and biases file provided by the model owner to the hyperscaler can be interpreted as embodying secret formulas or scientific experience, critical for the model's functionality to generate responses. Payments for the use of such information, which "cannot be separately copyrighted,"[161] may be characterized as royalties.

A potential counterargument is that while the hyperscaler may have possession of and can use the weights and biases file to provide services, it does not receive disclosure or actual knowledge of the numerical parameters in the file or knowledge of the process or techniques used to develop the foundation model. In many cases, these parameters are not viewable, readable, or accessible in a way that conveys understanding — they function as a black box, enabling execution but not comprehension. As such, the hyperscaler is merely using an opaque tool without acquiring any proprietary know-how or secret process or formula. Under this limited use of the weights and biases file, the transfer of a foundation model may be more analogous to the lease of a computer program as opposed to a payment for the use of know-how or secret process or formula under article 12.

Given the general rule under article 12 to bifurcate mixed contracts, if the part of the foundation model that represents computer code is considered separately, a portion of the payment may be considered business profits. The commentary related to computer programs suggests that when the model owner provides the hyperscaler with a copy of the foundation model's computer program and the right to make copies to enable the hyperscaler to use it to provide commercial services, that portion of the transaction should *not* be treated as a royalty. Nonetheless, there is a risk that a tax administration may argue that the transfer of the weights and biases file (which we have assumed is not eligible for copyright protection and is not a computer program) provides rights that go beyond mere operation of a computer program. Under that interpretation, the exception from royalty treatment for operation of a computer program would not apply and the portion of the transaction that represents use of the weights and biases file may constitute a royalty.

The commentary further clarifies that payments for the full ownership of IP are generally business profits (article 7) or capital gains (article 13), not royalties.[162] However, in typical foundation model arrangements, the model owner retains overall ownership, granting only a time-limited right to use and exploit the model's capabilities, not transferring full ownership.

Fine-Tuning Capabilities. If the hyperscaler is allowed to perform or offer fine-tuning for its customers, this may increase the risk of payments being characterized as royalties under article 12(2). This distinction arises from the difference between mere service provision and the transfer of secret processes or formulas or know-how, as

[161] *See* paragraph 14.3 of the commentary to art. 12.

[162] Paragraphs 8.2, 15, and 16 of the commentary to art. 12.

highlighted in the considerations above. Fine-tuning a foundation model involves adapting a pretrained model to a specific task or dataset, which entails modifying its existing weights and biases or adding new, specialized layers that are then trained. This process can be seen as a direct application of industrial, commercial, or scientific experience to refine and customize the core formula of the foundation model for a particular purpose.

Fine-tuning a foundation model enables the hyperscaler to potentially exploit the fine-tuned model by offering its use to the customer. Under this interpretation, it may be argued that the hyperscaler is effectively allowed to create a specialized version of the underlying IP tailored to its customers' needs, which may be analogous to creating a derivative work. Under that interpretation, when the hyperscaler receives the right to copy and deploy the model to provide GenAI services to the public, and when fine-tuning capabilities are included, this may increase the risk of royalty treatment.

On the other hand, fine-tuning the foundation model involves using its existing learning protocols to train the model on additional data. That process could be viewed as simply running the foundation model on new data, with the weights and biases file remaining a black box that neither the customer nor the hyperscaler can see. Under that interpretation, fine-tuning could be viewed as running the foundation model like any other piece of software, in which case the payments would not be a royalty, but would be business profits.

Model Owner-Hyperscaler Transaction Conclusion. The characterization of the hyperscaler payments to the model owner in a 2P model may depend on whether the foundation model is treated as a computer program in its entirety or if instead the weights and biases file is treated as know-how or a secret process or formula. If treated as a computer program in its entirety (as opposed to disaggregating the computer program and the other elements, like the weights and biases file) and the hyperscaler has only a right to use the program to provide services in its business, we expect the payments from the hyperscaler to be treated as the model owner's business profits. If, however, the weights and biases file is considered know-how or a secret process or formula, a portion of the payments made by the hyperscaler may be characterized as royalties. If a hyperscaler is allowed to perform fine-tuning of the model, the risk of royalty treatment may be increased, because tax authorities might argue that the hyperscaler has the right to create a derivative work. On the other hand, taxpayers might argue that fine-tuning the foundation model is simply running it on new data, and the hyperscaler does not have a right to know-how or a secret process or formula, making the payments business profits.

Taxation of Data Centers

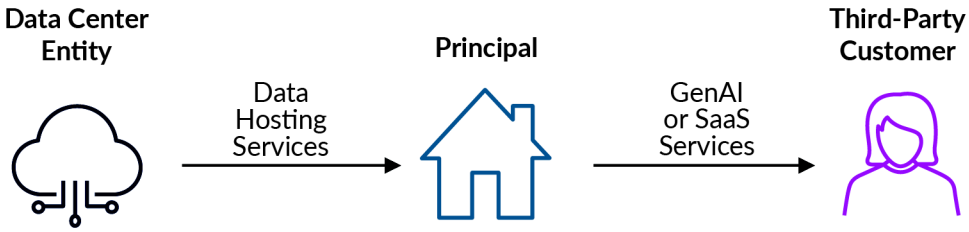
Given the importance of data centers to the delivery of GenAI, we now turn to two key issues involving the taxation of data centers — PE and transfer pricing. We selected these issues because we are aware that tax authorities have already identified them as areas of interest, with some tax authorities advancing novel approaches.[163]

For purposes of this article, we assume the following simplified set of facts for a hyperscaler, which apply both for providing access to GenAI and access to other cloud services (that is, SaaS):

1. The **principal** develops a GenAI or SaaS offering, either by developing its own model or software or contracting with a third party that has already developed a model or software. The principal (the seller of record) contracts with third-party customers to provide GenAI services or SaaS. The principal also contracts with a network of related data center entities that provide data hosting services to support the deployment of its GenAI or SaaS offering. In addition, the principal

[163] *See, e.g.*, Saint, *supra* note 146.

Figure 6. Data Center Entity Facts



performs other activities, such as setting the business strategy.[164]

2. The **data center entity** provides compute power and storage capacity (which we refer to collectively as data hosting services) that are necessary to support the deployment of the GenAI or SaaS offering to the third-party customer (see Figure 6).

Potential PE Issues for Data Centers

When the principal provides services to customers, it may use the compute power and data storage of a data center owned by an affiliated data center entity located in a foreign jurisdiction. The question thus arises whether the principal has a PE in the foreign jurisdiction because of its use of the data center entity's data center.

Fixed Place of Business PE

To answer this question, we turn to article 5 of the OECD model treaty. Article 5(1) defines a PE as a "fixed place of business through which the business of an enterprise is wholly or partly carried on." A place of business can include a facility such as premises, or in certain instances, machinery or equipment.[165] For a place to be fixed, it must be established at a distinct location with a certain degree of permanence.[166] It is "immaterial whether the premises, facilities or installations are owned or rented by or are otherwise at the disposal of the enterprise."[167] While the place of business of an associated enterprise can constitute a fixed place of business of another associated enterprise, to do so the second enterprise's presence at the fixed place of business must not be "so intermittent or incidental that the location cannot be considered a place of business."[168]

The commentary to article 5 does not explicitly address the treatment of data centers that are owned by a data center entity that provides data hosting services to a principal. However, the commentary to article 5 provides detailed guidance on how PE rules apply to e-commerce activities.[169] The commentary focusing on e-commerce distinguishes between a website (software and electronic data), which is not tangible property and therefore not a place of business, and the physical server on which the website is stored and accessible. Paragraph 124 of the commentary on article 5 provides the key guidance on that topic:

> The distinction between a web site and the server on which the web site is stored and used is important since the enterprise that operates the server may be different from the enterprise that carries on business through the web site. For example, it is common for the web site through which an enterprise carries on its business to be hosted on the server of an Internet Service

[164] In common operating models, hyperscalers also have affiliate entities known as distributors in some jurisdictions. Distributors resell GenAI services and SaaS to third-party customers and book the third-party revenue. Distributors typically perform the functions and bear the risks associated with a low-risk distributor and hence earn a return on costs or revenue. The activities and remuneration of distributors are not discussed further in this article, but are mentioned here for completeness.

[165] *See* paragraph 6 of the commentary to art. 5.

[166] *See* paragraph 28 of the commentary to art. 5.

[167] *See* paragraph 10 of the commentary to art. 5.

[168] *See* paragraph 12 of the commentary to art. 5.

[169] *See* paragraphs 122 to 131 of the commentary to art. 5.

> Provider (ISP). Although the fees paid to the ISP under such arrangements may be based on the amount of disk space used to store the software and data required by the web site, these contracts typically do not result in the server and its location being at the disposal of the enterprise, . . . even if the enterprise has been able to determine that its web site should be hosted on a particular server at a particular location. In such a case, the enterprise does not even have a physical presence at that location since the web site is not tangible. In these cases, the enterprise cannot be considered to have acquired a place of business by virtue of that hosting arrangement. However, if the enterprise carrying on business through a web site has the server at its own disposal, for example it owns (or leases) and operates the server on which the web site is stored and used, the place where that server is located could constitute a permanent establishment of the enterprise if the other requirements of the Article are met.[170]

The commentary thus makes clear that data center hosting arrangements do not cause data centers to create a fixed place of business for the recipient of the data hosting services, unless the enterprise owns or leases the server or data center or otherwise has the data center at its disposal.[171] We understand that such arrangements are uncommon; thus, in most cases a data center would not constitute a PE of the recipient of the data hosting services.

In general, a place is at the disposal of an enterprise if the enterprise's employees or subcontractors have "the effective power to use that site" (for example, the enterprise has legal ownership of the site or a right to enter the site along with control of access to and use of the site).[172] For servers or other automatic equipment to be at the disposal of an enterprise, the enterprise generally must own or lease the equipment.[173]

In other words, in-country data centers can only be considered at the disposal of a foreign enterprise if employees of the enterprise have regular and ongoing physical access to, and presence at, the data center, or the foreign enterprise owns or leases the servers. The mere fact that the enterprise uses the data hosting services provided by the data center entity is not sufficient to give rise to a fixed place of business PE.[174]

Dependent Agent PE

Article 5(5) of the OECD model treaty also provides for a dependent agent PE if:

- a person acts in a contracting state on behalf of an enterprise;
- that person habitually concludes contracts or plays the principal role leading to the conclusion of contracts that are routinely concluded without material modification by the enterprise; and
- the contracts are in the name of the enterprise or for the performance of services by the enterprise.[175]

Therefore, a person regularly acting in the jurisdiction on behalf of a foreign enterprise may create a dependent agent PE.[176]

The term "person" includes an individual, a company, or any other body of persons.[177] The

[170] Paragraph 124 of the commentary to art. 5.

[171] *Id.*

[172] *See* paragraph 40 of the commentary to art. 5.

[173] *See* paragraphs 41 and 124 of the commentary to art. 5.

[174] Even if the data center constitutes a fixed place of business PE of the principal, the incremental profits attributable to the data center entity jurisdiction often will be minimal or zero if the PE does not perform functions, use assets, or assume risks beyond those performed, used, and assumed by the data center entity. This issue is discussed in the context of article 5(5) of the model treaty in paragraphs 41 and 42 of OECD, "Additional Guidance on the Attribution of Profits to Permanent Establishments, BEPS Action 7" (Mar. 2018).

[175] *See* art. 5(5) of the OECD model treaty.

[176] Paragraphs 5 and 6 of article 5 of the OECD model treaty were updated in 2015 as part of the OECD/G20 BEPS project; see "Preventing the Artificial Avoidance of Permanent Establishment Status, Action 7 — 2015 Final Report." These changes were not a BEPS minimum standard, so their adoption is at the discretion of jurisdictions as they negotiate or revise their bilateral tax treaties. Notably, the U.S. model income tax convention does not incorporate these changes.

[177] Art. 3(1)(a) of the OECD model treaty.

commentary states that a website is not a person and therefore cannot generally be considered an agent of the foreign enterprise. It notes that "since the web site through which an enterprise carries on its business is not itself a 'person' as defined in Article 3, paragraph 5 cannot apply to deem a permanent establishment to exist by virtue of the web site being an agent of the enterprise for purposes of that paragraph."[178]

This paragraph should also apply to GenAI services or SaaS, which similarly are not a person for the purposes of article 3 of the OECD model tax convention. Hence, the fact that the GenAI services or SaaS provided by the principal to its customers may be purchased through a server in a foreign country, or a contract might be concluded by an AI agent (as discussed further below), operating on a server does not give rise to a dependent agent PE under article 5(5) of the OECD model treaty.

Transfer Pricing for Data Center Services

In the GenAI supply chains outlined above, there are numerous related party transactions that must be priced.[179] In most instances, these transactions are comparable to, and so will likely be priced on a similar basis to, SaaS. We are aware that tax administrations are already questioning the returns allocated to data centers that facilitate the delivery of SaaS, and we expect these challenges (and others based on novel theories) to accelerate. Hence, we have focused on this issue.[180]

Data Center Entity: Functions, Assets, and Risks for SaaS

The centralized operating model described above in a GenAI context is also commonly used to support the provision of SaaS. In these models, the functions performed by the data center entity would typically be limited. McKinsey estimates that a large data center would create more than 50 ongoing jobs, covering roles such as facility management and maintenance, engineers, and technicians.[181] These jobs would typically be viewed as relatively low value or routine roles, as compared with the engineers, architects, and software developers who are responsible for designing the hardware and software used to provide SaaS. The latter workers are typically employed by the principal, not the data center entity.

In terms of assets, the data center entity owns the servers and owns or leases the physical space. It provides data hosting services to the principal. Through remote network access to the servers, the principal runs its software on the data center entity's servers to provide SaaS to the principal's customers. Generally, the data center entity does not own the IP that underpins the hardware or software that it uses, which is owned by another entity in the group; for example, the principal.

The risk analysis is more complex, because there is significant risk in investing in, and running, data centers. When building a data center, an MNE group takes on significant investment risk (the risk that it is unable to earn a return on its investment) and capacity risk (the risk that the capacity is underused). However, it is unlikely that the data center entity would be responsible for making decisions about data center investment, and hence it is unlikely that the investment risk or capacity risk would be appropriately assumed by such entity.

MNE groups also incur significant operational risk from data center downtime, which can occur because of factors like power failures, cooling breakdown, or cybersecurity incidents. Data centers would, to some extent, be responsible for managing operational risk, but the significant risks would typically be controlled and managed through protocols and policies set by other entities within a group, such as the principal. In addition, an MNE group would seek to minimize operational risk associated with a single data center going down for an extended period by operating a network of data centers that can pick up the slack. Again, this means that it is unlikely that, for the purposes of a transfer pricing

[178] *See* paragraph 131 of the commentary to art. 5.

[179] For an analysis of how GenAI will affect value chains and transfer pricing, see Nick Stavrakis et al., "How Will Generative AI Affect Value Chains and Transfer Pricing?" *Tax Mgmt. Int'l J.* (Nov. 17, 2025).

[180] *See, e.g.*, Saint, *supra* note 146.

[181] Adam Barth et al., "The Data Center Balance: How U.S. States Can Navigate the Opportunities and Challenges," McKinsey & Co. (Aug. 8, 2025).

analysis, the data center entity would be found to assume significant operational risk.

Data Center Entity: Remuneration for SaaS Data Hosting

The comparable profits method (CPM)/transactional net margin method (TNMM) is typically used to benchmark the returns earned by entities that perform routine data hosting services. The CPM/TNMM is typically used because of its practical strengths and the limited functional and risk profile of the data center entity, which make it more appropriate than other methods.

To apply the CPM/TNMM, a taxpayer must select the most appropriate profit level indicator, considering a variety of factors. The indicator is a financial ratio, with a measure of profit as the numerator over a denominator that is closely correlated with the profit-generating activity of the business. A taxpayer must be able to reliably determine the denominator for both the tested party and the external comparables, and the denominator must be indicative of the value of the functions performed, assets used, and risks assumed by the tested party.

In our experience, cost-based profit level indicators are commonly selected as most appropriate.

How Do Data Centers Supporting GenAI Services and SaaS differ?

There are many similarities between data centers that support the provision of SaaS and GenAI services. In fact, the same data center frequently supports both types of services.[182]

Data centers perform the same functions when supporting the provision of SaaS or GenAI services. Local employees are responsible for maintaining and monitoring hardware and local networks, managing the facilities, and ensuring they remain secure. While it is possible that the intensity of these functions increases, for example, because data centers supporting GenAI services may require additional maintenance, it is unlikely that there is any meaningful change to the nature of the activities performed.

The tangible and intangible assets used in the provision of GenAI services may vary from those used to provide SaaS. As discussed above, SaaS may be provided using CPUs, while GenAI-enabled data centers rely more on GPUs and TPUs. In simple terms, the computer chips employed in GenAI-enabled data centers are more technologically sophisticated and more expensive.[183] GPUs and TPUs need more power than CPUs, get hotter, and need more advanced cooling systems.[184] GenAI-enabled data centers may use more sophisticated networking technology and software, some of which may be proprietary to the group that owns the data center, but which can also be acquired from third parties. GenAI-enabled data centers may also hold longer-term energy contracts, giving them access to energy at a fixed price, which may be above or below the market rate at a given time.[185]

The risks assumed by groups as they invest in GenAI-enabled data centers are significant, and given higher costs and demand uncertainty, are greater than those associated with investment in data centers in the past. However, the assumption of risks is likely to follow that outlined above for SaaS arrangements, with the principal assuming key risks, such as investment and capacity, and the data center entity assuming less significant risks, such as limited operational risk.

Should SaaS and GenAI Support Services From Data Centers be Remunerated Differently?

The question then becomes whether the transfer pricing methods used to set or test the remuneration of the data center entity should vary if it is to support GenAI services. In short, the answer is likely no. The functional and risk profile of GenAI-enabled data centers is generally the same as other data centers, and hence it is likely to remain appropriate to remunerate these data centers using the same transfer pricing methods and policies that MNE groups have previously used.

[182] Phil Powell and Smalley, "What Is a Hyperscale Data Center?" IBM (last accessed Feb. 5, 2026).

[183] Schneider and Smalley, "What's the Difference Between AI Accelerators and GPUs?" IBM (last accessed Dec. 22, 2025).

[184] Vertiv Group Corp., "The Cost Impact of AI Data Center Design, Build and Operations" (last accessed Dec. 22, 2025).

[185] Alicia M. McKnight and Robert A. James, "Power Purchase and Interconnection Agreements for Data Centers," Pillsbury Law (July 21, 2025).

This does not, however, mean that GenAI-enabled data centers would be allocated the same amount of profit as data centers that primarily support SaaS. Because the size of GenAI-enabled data centers is likely to be larger than data centers that support primarily SaaS, a data center entity that supports GenAI would expect to earn higher operating profit given its larger cost base. In other words, applying an equivalent markup on higher depreciation expenses would give the entity a higher overall return. Thus, while the relative return of GenAI-enabled and non-GenAI-enabled data centers might be the same, the absolute return allocated to GenAI-enabled data centers would typically be higher, reflecting the greater size of these data centers.

Other Transfer Pricing Issues

There are two other concepts included in the OECD guidelines that arise in the remuneration of data center entities.

The first is location savings — cost savings that are attributable to operating in a certain market. Because energy and cooling are an important cost for data centers, centers may be built in locations with cheaper energy or that are naturally cold.[186] In other instances, it is important that data centers are built close to customers to reduce latency. The question arises whether the benefits of a favorable location should accrue to the principal or data center entity.

Before discussing the allocation of any benefits, it is important to quantify what those benefits are. If a site is deemed to be advantageous, the existing landowner is likely to demand a higher price to sell or lease the land, raising the cost of establishing or operating the data center. Hence, it is feasible that the benefits of an advantageous location are captured, primarily or in totality, by the existing landowner. To the extent there is a remaining benefit, that benefit should generally accrue to the principal. With the growth in AI, the competition for data center investment between jurisdictions has intensified, providing the principal multiple choices of location and significant bargaining power. The principal is likely to be responsible for reviewing potential locations for data centers and selecting the location that is most advantageous. The principal may also negotiate long-term energy contracts to guarantee supply at a fixed price. For these reasons, it is difficult to see why the benefits of potential location savings should accrue to the data center entity.

The other concept to consider is the options realistically available to a principal and data center entity when a data center is built.[187] At this point, the principal holds all the cards. It has the know-how required to build the data center, a supply of the necessary computer chips, and responsibility for determining the location of the data center and negotiating any accompanying leases or energy contracts. The principal has a wide range of commercial options and no obligation to contract with the entity that becomes the data center owning entity. This means that the data center entity is a price taker — it would accept a return that remunerates it for its costs and the limited risks that it assumes.

Robot Taxation and Agentic AI

We turn now to a discussion of two proposals to change our taxation system in fundamental ways to address perceived issues with GenAI. We start with a discussion of a proposal on taxation of robots, because of the fear that robots (and now, GenAI) will replace human workers. We then discuss a proposal to treat agentic AI as a legal person. Of course, there are many possible ways to change the tax system, and we do not consider all of them. We have chosen these two proposals as representative of a broader range of proposals. This discussion is merely an introduction to what is likely to be a lengthy debate about potential reforms to our international tax system in response to GenAI.

Both proposals would create significant challenges. Moreover, both are intended to address problems that have not arisen and are not certain to arise, making it difficult to justify the challenges that the proposals would entail.

[186] Eric Olinger, "5 Considerations for Choosing Data Center Locations," Equinix (last accessed Feb. 5, 2026).

[187] Paragraphs 1.38 to 1.40 of the OECD guidelines recognize "options realistically available" to be a core part of accurately delineating a transaction.

Robot Taxation

Discussions of robot taxation have emerged because of concerns that the taxation of labor income, through payroll and income taxes, may lead humans to be priced out of the labor market by robots.

The concept gained some national attention when it was included in presidential hopeful Bill de Blasio's proposed response to automation. In an op-ed published in *Wired*, de Blasio said he would "institute a 'robot tax' on large companies that eliminate jobs through increased automation and fail to provide adequate replacement jobs."[188]

A robot tax could be designed a variety of ways. De Blasio proposed that companies that eliminate jobs through automation be required to pay up to five years of payroll taxes for each job they eliminate. Alternative proposals have suggested that the income attributable to a robot should be taxed as if the robot were a human being.

While a robot tax seems simple in concept, there are significant design questions that would make implementing one challenging:

1. **What is a robot?** The world is filled with machines that perform tasks previously performed by humans. If a robot tax were applied to a coffee shop that automates the job of a barista, the question also arises whether a robot tax should apply to a coffee machine that automates the process of weighing, grinding, and pressing the coffee beans, but still requires a barista to take the customers' orders, hit run, and hand the customers their coffee. And what about a machine that just weighs and grinds the beans? In other words, any machine could be viewed as replacing human labor.
2. **How would the income of a robot be determined?** If a robot were taxed based on the income of the human that it replaced, it would be necessary to determine how this income is determined. However, robots are unlikely to replace individual workers, but rather assume responsibility for specific tasks, with humans continuing to perform other tasks. This would make it difficult to value the contribution of a specific robot and determine the income on which it should be taxed.
3. **Who would be subject to the robot tax?** Because a robot does not earn income, it cannot be subject to robot tax; instead, the tax would necessarily apply to the corporation or individual that owns the robot.
4. **What portion of the income of a robot should its owner be taxed on?** The income derived from a robot by its owner is likely to change over time and is unlikely to reflect the full value of the workers it replaces. For example, a coffee company that automates the job of a barista may initially capture the income that it would have paid to its baristas. However, the company may find that its customers were willing to pay a premium for a barista or that price competition increases as competitors introduce similar robot baristas. In this way, the coffee company's customers would capture some of the economic benefit arising from the introduction of robot baristas, and taxing the company on the entire income of the displaced baristas would be excessive.[189]

The idea of a robot tax has also been criticized as an obstacle to economic growth.[190] Though protecting workers from displacement could be cited as an advantage, in a global economy this could doom a country to underperformance and its workers to lower standards of living.[191]

Taxation of Agentic AI as a Person

Agentic AI "is an artificial intelligence system that can accomplish a specific goal with limited supervision. It consists of AI agents — ML models that mimic human decision-making to solve

[188] Bill De Blasio, "Why American Workers Need to Be Protected From Automation," *Wired*, Sept. 5, 2019.

[189] For further discussion of the challenges of a robot tax, see Yanis Varoufakis, "A Tax on Robots?" Project Syndicate, Feb. 27, 2017.

[190] *See* Robert D. Atkinson, "The Case Against Taxing Robots," Information Technology & Innovation Foundation (Apr. 8, 2019); Robert Seamans, "Tax Not the Robots," The Brookings Institution (Aug. 25, 2021).

[191] Tatiana Falcão, "Should My Dishwasher Pay a Robot Tax?" *Tax Notes Int'l*, June 11, 2018, p. 1273.

problems in real time."[192] A simple example of an AI agent is an AI app that can design and book a holiday based on a traveler's instructions, rather than just make suggestions.

AI agents do not give rise to a PE under the OECD model treaty. GenAI models, which are intangible, do not give rise to a fixed place of business on their own or cause the server on which the model runs to be at the disposal of the owner of the model (assuming the model owner does not own or lease the server).[193] In addition, an AI agent is not a person and therefore cannot give rise to a dependent agent PE.[194]

However, some commentators have argued that AI agents should essentially be treated as persons that create PEs in the jurisdictions in which they operate.[195] This proposal is similar to a robot tax, in that its objective is to extend a jurisdiction's taxing rights over AI agents. On the other hand, the proposal would not impose a new tax on these agents but would seek to assert a taxing right over income that is already taxed in another jurisdiction. The effect of this proposal, assuming that the resulting double taxation is eliminated, would be to transfer taxing rights from one jurisdiction to another. The simplest argument for the approach is that an AI agent performs economic activities that are so connected to a jurisdiction that the jurisdiction in question should have taxing rights.[196]

As foreshadowed throughout this article, we do not think that argument should prevail. The AI agent (and the increased productivity and other benefits it creates) would not exist without the productive activities that developed the AI agent and the enterprises that use the AI agents in their businesses. Those productive activities are enabled by the investments that jurisdictions make in infrastructure, their workforces, legal systems, etc. Divorcing the taxing rights over the outputs of those productive activities reduces the incentive and ability of jurisdictions to make those investments. The existing transfer pricing framework already compensates the jurisdiction in which the AI agent is created, the jurisdiction in which the business customer uses the AI agent, and the jurisdiction in which the AI agent runs on a server based on their relative contributions — that is, it already reaches the correct answer.

Moreover, the same design challenges associated with a robot tax also arise if a jurisdiction were to assert that the use of an AI agent should create a PE:

1. **What is an AI agent?** To introduce an AI agent PE, it would be necessary to define an AI agent and determine when the activities it performs are sufficient that the jurisdictions to which those activities are connected should have a taxing right over the agent PE. For example, it may be necessary to distinguish between an AI-powered travel service that designed but did not book travel and one that also completed the booking process. The distinction would be both difficult to draw but also difficult to rationalize: Why should agents that recommend holidays be taxed differently from agents that book them?
2. **Which contributor to the AI agent would have a PE?** A variety of parties contribute to an AI agent. For example, the provision of an AI travel agent would likely involve a model owner that has developed a foundation model; an AI app developer that incorporates the foundation model with other data and features to deliver the AI travel agent to its customers; a hyperscaler that provides the compute power necessary to deploy the agent in different jurisdictions; and the customers that grant the AI travel agent the authority to conclude contracts on their behalf. In designing an AI agent PE, it would be necessary to determine which party should have a PE. In some instances these parties may belong to the same MNE group, but in others they may not.

[192] IBM, "What Is Agentic AI?" (last accessed Oct. 20, 2025).

[193] *See* paragraphs 123 and 124 of the commentary on art. 5 of the OECD model treaty.

[194] Article 5(5) of the OECD model treaty establishes the concept of a dependent agent PE, when "a person is acting in a Contracting State on behalf of an enterprise and, in doing so, habitually concludes contracts." The OECD model treaty defines "person" in article 3(1)(a), as "an individual, a company and any other body of persons." Because AI does not have legal personhood, it would not be considered a person for the purposes of the OECD model treaty and hence could not create a PE under article 5(5).

[195] *See, e.g.*, Lucas de Lima Carvalho, "AI Agents: Tax Policy Issues," *Tax Notes Int'l*, July 8, 2024, p. 215.

[196] *Id.*

3. **How would the income attributable to the AI agent PE be determined?** This question is tied to the question of which party would be deemed to have a PE, because the amount of income each would earn from a given transaction (for example, the use of an AI travel agent to book a holiday) would be different. In some cases, determining the income a party generates from activities connected to a specific jurisdiction would be impossible. For example, if the app developer pays a flat fee to the model owner to incorporate the foundation model into its app, the model owner would not earn any income that could be directly connected to an individual holiday booking. In these cases, there is also no obvious connection between the activities of the model owner and the jurisdiction of the holidaymaker or destination, and so no obvious policy reason why that jurisdiction should have taxing rights over the model owner.
4. **Where would the AI agent PE be located?** Because AI agents are not tangible, it would be challenging to design rules to determine where an AI agent PE is located. Further, because IP is highly mobile, any jurisdiction with AI agent PE rules may receive less investment, because taxpayers could move the IP to a jurisdiction without these rules. Similarly, such an approach would tax the AI agent's activity in a particular jurisdiction in which the AI agent operates instead of allowing the jurisdiction in which the productive activities and investment that created the AI agent (where the income was generated) to levy a tax. Jurisdictions may have less incentive to invest in the development of further AI technology if they do not have the benefit of increased revenue from its activities.

Conclusion

We do not believe that either a robot tax or AI agent PE are appropriate approaches to tax GenAI business models. First, these proposals suffer from several technical challenges, as described above. They are also designed to address problems that have not arisen and may never arise. That is, like other technologies, GenAI might negatively affect some occupations while generating entirely new ones, on balance being a neutral or positive effect on employment. This would render making a robot tax or AI agent PE unnecessary. Further, as we outline above, the existing tax rules can be applied relatively straightforwardly to GenAI transactions, and the results make sense — they generally impose tax based on where the investment and productive activities that gave rise to the income took place.

While we have focused on two proposals, there are potentially many more proposals to modify taxation to provide jurisdictions — particularly market jurisdictions — with new taxing rights. However, any such proposals that would de-link taxing rights from where the investment and productive activities that created the income took place would need to overcome that detriment with substantial benefits. In addition, any such proposal should consider whether the outcomes it would produce make sense (for example, that the jurisdictions that receive taxing rights have a connection to the earned income). For the reasons we've articulated, the two proposals we examined, regarding robot taxation and AI agent PE, don't meet that test.

More generally, our tax system should be designed to collect the appropriate amount of revenue while doing the least economic harm possible, including by not discouraging innovation. A robot tax is explicitly designed to *reduce* innovation that increases productivity, and an AI agent PE would *de facto* do the same by reducing the incentive and ability of jurisdictions to invest in the conditions that enable innovation. Changing the tax system to reduce economic growth in the absence of evidence of problems that need to be solved is at best premature and more likely foolhardy.[197] ■

[197]The foregoing information is not intended to be "written advice concerning one or more Federal tax matters" subject to the requirements of section 10.37(a)(2) of Treasury Department Circular 230. The information contained herein is of a general nature and based on authorities that are subject to change. Applicability of the information to specific situations should be determined through consultation with your tax adviser. This report represents the views of the authors only, and does not necessarily represent the views or professional advice of KPMG LLP.