OECD publishing

AI OPENNESS: A PRIMER FOR POLICYMAKERS

OECD ARTIFICIAL INTELLIGENCE PAPERS

August 2025 **No. 44**

GPAI

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This report was approved and declassified by written procedure by the Global Partnership on Artificial Intelligence (GPAI) on 25 June 2025 and prepared for publication by the OECD Secretariat. Earlier versions were discussed by the OECD Working Party on AI Governance (AIGO) in outline form in November 2023, and in draft form in June 2024. It was also discussed by the GPAI in December 2024.

Note to Delegations:

This document is also available on O.N.E Members & Partners under the reference code:

DSTI/DPC/GPAI(2024)3/FINAL

Revised version, August 2025

Corrigendum

Page 29:

The bibliographic entry for André, C. et al. (2025) has been amended to reflect the correct author.

Acknowledgements

This report was prepared under the aegis of the OECD Working Party on AI Governance (AIGO) and the Global Partnership on AI (GPAI). Karine Perset, Luis Aranda, and Guillermo Hernández (OECD Artificial Intelligence and Emerging Digital Technologies Division) led the report development and drafting, under the supervision of Audrey Plonk, Deputy Director of the OECD Science, Technology and Innovation Directorate. An earlier version of this report was drafted by Elizabeth Seger (Oxford University).

The paper benefitted significantly from the oral and written contributions of AIGO and GPAI delegates as well as experts from the OECD.AI network of experts. The authors would like to extend their sincere gratitude to the Delegations of Australia, Brazil, Canada, Chile, Croatia, the European Commission, France, Germany, India, Israel, Mexico, New Zealand, Poland, Slovenia, Sweden, Thailand, Türkiye, the United Kingdom, and the United States for their invaluable insights. In particular, they gratefully acknowledge contributions from Simon van Hoeve (Australia); Franklin Rodrigues Hoyer (Brazil); Ricardo Baeza-Yates (Chile); Juraj Bilic (Croatia); Jonas Roule (France); Abhishek Singh (India); Ziv Katzir (Israel); Juraj Corba (Slovakia); Polonca Blaznik and Martin Marzidovsek (Slovenia); Jesse Dunietz and David Turnbull (US); Carlos Muñoz Ferrandis (BigScience); Loise Mwarangu (AI Centre of Excellence, Kenya); Marko Grobelnik (Jozef Stefan Institute); and Stuart Russell (UC Berkeley).

The Secretariat would also like to thank stakeholder groups at the OECD for their input, including Pam Dixon (Civil Society Information Society Advisory – CSISAC); Nicole Primmer and Maylis Berviller (Business at OECD – BIAC); Sarah Jameson and Aida Ponce (Trade Union Advisory Committee – TUAC); and Sebastian Hallensleben and Jibu Elias (Internet Technical Advisory Committee – ITAC).

Finally, the authors thank all those who have contributed to the report throughout its development. This includes Jeff Mollins, Lucia Russo, Kasumi Sugimoto, Sarah Bérubé, and Nikolas Schmidt (OECD/STI); Manuel Betin and Peter Gal (OECD/ECO); and Richard May (OECD/DAF). The authors also thank Anaisa Goncalves, Shellie Laffont and Andreia Furtado for editorial support; the overall quality of this report benefitted significantly from their engagement.

Table of contents

Tables

Figures

Boxes

Abstract

This report explores the concept of openness in AI, including relevant terminology and how different degrees of openness can exist. It explains why the term "open source," a term rooted in software, does not fully capture the complexities specific to AI. This report analyses current trends in open-weight foundation models using experimental data, illustrating both their potential benefits and associated risks. It incorporates the concept of marginality to further inform this discussion. By presenting information clearly and concisely, the report seeks to support policy discussions on how to balance the openness of generative AI foundation models with responsible governance.

Résumé

Ce rapport explore le concept d'ouverture dans l'IA, y compris la terminologie pertinente et comment différents degrés d'ouverture peuvent exister. Il explique pourquoi le terme "open source", un terme ancré dans le logiciel, ne capture pas pleinement les complexités spécifiques à l'IA. Ce rapport analyse les tendances actuelles des modèles de fondation à poids ouverts en utilisant des données expérimentales, illustrant à la fois leurs avantages potentiels et les risques associés. Il intègre le concept de marginalité pour enrichir cette discussion. En présentant l'information de manière claire et concise, le rapport vise à soutenir les discussions politiques sur la manière d'équilibrer l'ouverture des modèles de fondation générative de l'IA avec une gouvernance responsable.

Executive summary

A clearer understanding of AI openness terminology is essential. This report examines key concepts and definitions to foster a common understanding of AI openness, acknowledging that precise terminology is vital for researchers, developers, and policymakers as AI technologies continue to develop.

The term "open source" in AI has limitations: The term "open source" originates from software development and does not accurately describe AI systems. Unlike traditional software, AI "source code" may refer to inference code, training code, or both – and each can be made publicly available independently. Furthermore, AI models can have other critical components such as model weights and training data, which can also be shared or kept private.

AI openness exists on a spectrum: It is not binary but ranges from fully closed systems with restricted access to fully open models that permit unrestricted access, modification, and use. This spectrum encompasses various system components, including data, code, and documentation. Recognising this range is essential for understanding the policy implications of different levels of openness across these components.

Definition of *open-weight* AI models in this report: This report uses the term *open-weight models* to refer to foundation models with publicly available trained weights. These models can generate content and perform a variety of tasks across different applications. While licensing is an important aspect of the discussion surrounding the availability of AI models, this report focuses on open weights due to their growing relevance in policy discussions about the benefits and risks associated with these models.

Market trends and global distribution: Since early 2023, the number of foundation models has surged, with open-weight models now making up over half of the market. The United States leads in the development of open-weight foundation models, followed by the People's Republic of China (hereafter "China") and France. The Netherlands and Singapore serve as key provider hubs due to their advanced cloud capabilities, highlighting the global nature of AI deployment.

Rapid improvements in quality: Open-weight foundation models have significantly improved in performance since early 2024, achieving higher scores on common benchmarks. While these advances offer substantial benefits, they also increase potential risks, underscoring the importance of monitoring openness in AI development.

Benefits of open-weight models: Releasing model weights can provide significant advantages, including enabling external evaluation and accountability, accelerating research and innovation, fostering competition, facilitating access to AI technologies, and supporting sensitive data management. However, realising these benefits often depends on access to sufficient computing resources, data, and skilled talent.

Risks of open-weight foundation models: Open-weight models also pose significant risks, including the potential for malicious activities such as deepfakes, advanced cyberattacks, large-scale generation of child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII). There is also the speculative risk of misuse in areas like biology or chemistry. The availability of model weights can empower malicious actors to fine-tune these models for unintended uses or harmful purposes. Furthermore, modifying model weights can enable malicious actors to bypass some of the safeguards put in place by the original developers.

Deciding on openness: Decisions to release model weights should carefully consider potential benefits and risks. Falling compute costs and more accessible fine-tuning methods lower the barriers to both use and misuse, enhancing the potential advantages of open-weight models while also increasing the risk of harmful applications. It is essential to evaluate the *marginal* risks and benefits of releasing model weights

in relation to closed models and existing technologies. However, this should be done as part of a broader, holistic risk assessment framework that can adapt to evolving capabilities and usage patterns. Developers and other relevant stakeholders should consider whether the benefits outweigh the risks in a given context and consider the potential opportunity costs of not releasing open-weight models.

Introduction

Since the 1990s, open-source software (OSS) has proliferated alongside, and often within, commercial software, encouraging co-operation, promoting software adoption including in developing countries by lowering costs, balancing the market power of major software companies, fostering innovation, enabling upskilling, and improving software quality through community feedback (Langenkamp et al., 2022[1]; Engler, 2021[2]; Engler, 2021[3]). Given these and other benefits, the OSS tradition has been inherited by certain groups in the AI community, where an AI model (or some elements of it) are released publicly for anyone to download, modify and distribute, often under the terms of a licence.

There is ongoing debate about the risks, benefits, and trade-offs of making AI models or their components publicly available, particularly regarding increasingly advanced AI foundation models that exhibit general-purpose capabilities. The debate has gained momentum following recent launches of open-weight models, including Deepseek R1, OpenAI's GPT-OSS, and Alibaba's Qwen.

Numerous beneficial uses of foundation models are developing in healthcare (Fries et al., 2022[4]), customer support (OpenAI, 2023[5]), immersive gaming (Marr, 2023[6]), or personalised education (Marr, 2023[7]). Access restrictions to AI models could stifle innovation, limit external evaluation, hinder the widespread distribution of AI benefits, and concentrate control over future AI technology in the hands of a small number of actors (Goldman, 2023[8]; Creative Commons et al., 2023[9]). However, foundation models can also be misused and deployed by malicious actors, for example, to generate child sexual abuse material, infringe intellectual property and privacy rights, or conduct convincing scams in which victims believe they are interacting with trusted friends and family.

Once a model is open, many safeguards to prevent misuse can be circumvented, and actors with sufficient expertise and computing resources could "fine-tune" it to enhance its propensity for misuse. Furthermore, fully removing open models after their release or adding guardrails retroactively to prevent newly identified risks may prove challenging. Restrictions on releasing AI models and their different components could raise questions about intellectual property rights and may enhance security, incentivising innovation and limiting the proliferation of risks.

This report examines the potential risks and benefits of open-weight models – understood as foundation models for which the weights are publicly available – keeping in mind that as the development and adoption of foundation models continue to advance amidst other technological and societal changes, the balance between the risks and benefits of releasing their weights may change.

Although this report focuses on open-weight foundation models, it is important to note that AI models that are not general-purpose, or "advanced" in the sense described above, can also pose risks. For example, Urbina et al. (2022[10]) showed that standard, narrow AI tools used within the pharmaceutical industry could be repurposed to assist with the design of biochemical weapons.

The report is organised as follows. Section 1 defines key terms and scope, delving into the different levels of openness in AI and explaining why the term "open source", originally used for software, may not fully apply in the AI context. Section 2 analyses current trends in open-weight models using experimental data from the OECD. Section 3 illustrates the potential benefits and risks of open-weight models and presents the concept of marginality. Section 4 concludes.

1. Delving into AI openness

This section delves into the concepts and definitions that are essential for establishing a shared understanding of AI openness. As the field of AI continues to evolve, clear terminology becomes increasingly important for researchers, developers, and policymakers alike.

1.1. The term open-source AI is a misleading legacy

The term “open-source” originates from “open-source software” (OSS). “Open-source” was defined in 1998 as a “social contract” (and later a certification) describing software designed to be publicly available – and released under a license that sets the conditions for using, modifying, and distributing the source code. According to the Open Source Initiative (OSI), an open-source software license must meet ten key criteria, which include free source code access, permission for derived works, and no limits on who may use the software or for what purpose (Perens, 1999[11]; Choose a License, 2023[12]). In principle, “open source” does not necessarily preclude a related commercial activity.

Multiple definitions of open-source AI exist today. OSI released a draft open-source AI definition leveraging the OECD definition of an AI system:

- “An open-source AI is an AI system made available under terms and in a way that grant the freedoms to: use the system for any purpose and without having to ask for permission; study how the system works and inspect its components; modify the system for any purpose, including to change its output; and share the system for others to use with or without modification, for any purposes. These freedoms apply both to a fully functional system and to discrete elements of a system. A precondition to exercising these freedoms is to have access to the preferred form for make modification to the system” (OSI, 2025[13]).

The Linux Foundation has also proposed an open-source AI definition that, unlike the OSI draft, entails sharing information about the underlying components (LinuxFoundation, 2024[14]):

- “Open-source artificial intelligence (AI) models enable anyone to reuse and improve an AI model. Open-source AI models include the model architecture (in source code format), model weights and parameters, and information about the data used to train the model that are collectively published under licenses, allowing any recipient, without restriction, to use, study, distribute, sell, copy, create derivative works of, and make modifications to, the licensed artifacts or modified versions thereof”.

The open-source software concept of “free and publicly downloadable source code” does not translate directly to AI due to differences in how AI models are built compared to traditional software (Finley, 2011[15]; Sijbrandij, 2023[16]). For AI models, “source code” could refer to either the inference code, the training code, or both, and the two can be shared independently. AI models have additional components beyond source code, including model weights and training data, all of which can either be shared or kept private and independent from the source code and from each other.

In particular, source code and model weights represent two distinct concepts. Referring to weights as "open source" is misleading, as they do not constitute source code. Source code comprises instructions

for executing specific tasks, whereas weights are the results of training and fine-tuning processes applied to data. Licenses intended for source code do not directly apply to AI model weights (Sijbrandij, 2023[16]).

For these and other considerations, the term "open-source AI" is currently debated. Some interpret it according to the OSI definition of open source, while others see it as encompassing a range of access options, from "non-gated downloadable" models to fully open models. Fully open models, like GPT-J, make all training and inference code, weights, and documentation publicly available. They can be freely used, modified, and distributed, including for commercial purposes. Non-gated downloadable models provide some components, such as training code and model weights, while withholding others. This is in contrast with gated downloadable models that restrict access to certain users.

This debate is illustrated by the framing of AI models like LLaMA, LLaMA2, Dolly, or StableLM, which use the term “open source” in a way that is inconsistent with OSI’s definition of open-source software (Finley, 2011[15]; Maffulli, 2023[17]). Some developers claim their models are open source simply because their weights are available for download, even though their licenses may restrict certain use cases and distribution.

The Linux Foundation has proposed a Model Openness Framework (MOF) to help evaluate and classify the openness of AI models, by assessing which components of the model are publicly released and under what licenses (Table 1.1). It defines three progressively broader classes of model openness:

- Class III – Open Model: The minimum bar for entry, Class III requires the public release of the core model (architecture, parameters, basic documentation) under open licenses. This allows model consumers to use, analyse, and build on the model, but limits insight into the development process.
- Class II – Open Tooling: Building on Class III, this tier includes the full suite of code used to train, evaluate, and run the model, plus key datasets. Releasing these components enables the community to better validate the model and investigate issues. It is a significant step towards reproducibility.
- Class I – Open Science: The apex, Class I entails releasing all artifacts following open science principles. In addition to the Class II components, it includes the raw training datasets, a thorough research paper detailing the entire model development process, intermediate checkpoints, log files, and more. This provides unparalleled transparency into the end-to-end development pipeline, empowering collaboration, auditing, and cumulative progress (LinuxFoundation, 2024[18]).

Table 1.1. Components of the Linux Foundation’s Model Openness Framework

The framework identifies 17 critical components for a complete model release.

Code	Data	Documentation
Evaluation code	Datasets	Data card
Pre-processing code	Evaluation data	Research paper
Model architecture	Sample model outputs	Evaluation results
Libraries and tools	Model weights and parameters	Model card
Training code	Model metadata	Technical report
Inference code	Configuration file	

Note: For each component, the MOF stipulates the use of standard open licenses based on the artifact type: open-source licenses for code (e.g., Apache 2.0, MIT), open-data licenses for datasets and model parameters (e.g., CDLA-Permissive, CC-BY), and open-content licenses for documentation and content/unstructured data (e.g., CC-BY). Sample model outputs can be code or data.
Source: Adapted from Linux Foundation (2024[18]).

1.2. Degrees of AI openness: The more model components are publicly released, the easier it is for other actors to reproduce, modify, and use the model

Currently, there is a spectrum of model release options available which encompass its various components, ranging from fully closed systems, where access to the model and its underlying data is highly restricted, to fully open models, which allow unrestricted access and modification by users (Box 2.1). Understanding this spectrum is crucial for evaluating the implications of different access levels on innovation, collaboration, and security and other considerations.

Box 1.1. Further research is needed to refine and assess open access gradients of AI systems

The landscape of AI model release options is diverse, encompassing a spectrum from fully closed models to various degrees of openness, including gradual or staged releases, hosted access, cloud-based API access, downloadable models, and fully open models. By systematically examining the different release strategies, researchers and policymakers can identify the associated benefits and drawbacks of each approach, including issues related to transparency and security.

Further research is essential to advance the responsible deployment of AI technologies. This research should aim to establish good practices for model sharing that balance openness and security, protect intellectual property and sensitive data, and promote more equitable access and responsible use of AI systems. Such efforts will contribute to a more informed and strategic approach to AI model dissemination.

Source: Solaiman (2023[19]); Sastry (2023[20]); Liang et al. (2022[21]).

Access to model architecture and trained weights, when combined with inference code, is enough for using a pre-trained model to perform specific tasks. Downstream developers can write their own inference code or even generate it using tools such as ChatGPT, and it does not need to match the original code used by the model developer. Having access to model weights also allows downstream developers to fine-tune and optimise the model for specific tasks and applications or modify its behaviour as needed. Additionally, releasing parts of the training code, such as the code for cleaning and loading training data, can help developers reproduce the training weights and utilise the model more easily.

Sometimes, an AI developer publicly releases the training and inference code for a model, but not the trained model weights (Meta, 2023[22]; Vincent, 2023[23]). In such cases, actors with sufficient computing resources and data access could train an equivalent model and use inference code to run it. However, few actors currently have the computing resources needed to train advanced foundation models.

Recently, several developers have restricted access to some of their models due to concerns about competition and potential misuse. These developers may choose to keep their models completely private, like DeepMind's Chinchilla (Hoffmann et al., 2022[24]) or share them in a controlled manner, such as OpenAI's GPT-4 (OpenAI, 2023[25]) and Anthropic's Claude 2 (Anthropic, 2023[26]) through application programming interfaces (APIs) (Brockman et al., 2023[27]). This approach allows them to enforce user restrictions and maintain better control over features. In contrast, some developers have advocated for more open models. Meta, for example, announced the "open-source" release of I-JEPA (Assran et al., 2023[28]), Llama 2 (Inskeep and Hampton, 2023[29]; Milmo, 2023[30]), and Llama 3, but faced criticism including from OSI for the access restrictions placed on its models (Anandira, 2024[31]). These examples illustrate the difference between the traditional definition of open-source software and the varying degrees of openness in AI models.

1.3. Licensing choices influence access levels, innovation speed, and the potential for beneficial and harmful uses

Licensing in “open-source” AI governs many terms under which AI models, including their weights and associated code, can be used, modified, and distributed. Unlike proprietary models that restrict access to discrete users or that customise licensing access conditions or provide exclusivity, open-source AI offers various licensing schemes that grant different nonexclusive standardised conditions to all users, generally for free. Permissive licenses, such as Apache 2.0 or MIT, typically allow for broad usage, modification, and commercialization with minimal restrictions beyond attribution and disclaimers (OSI, 2025[32]). Conversely, more restrictive or "copyleft" licenses, like certain versions of Creative Commons, may require that any derivative works also be shared under the same or similar open-source terms, aiming to ensure the continued openness of the AI ecosystem (Commons, 2025[33]). The choice of license can significantly impact collaboration, innovation, adoption, and the potential for beneficial and harmful uses of AI models.

While a detailed discussion of specific AI model licensing is outside the scope of this analysis, it's crucial to acknowledge their pivotal role in shaping the current and future landscape. Licensing decisions directly influence who can access and build upon AI models, impacting the pace of development and the potential for both beneficial and harmful applications. For instance, more permissive nonexclusive licenses can accelerate innovation by allowing widespread experimentation and integration into commercial products, but they might also offer fewer safeguards against misuse by malicious actors. Conversely, more restrictive licenses may often be necessary to incentivise model development and investment for specific market needs, but they could also inadvertently limit collaboration. Understanding the nuances of AI licensing is essential for formulating effective policies relating to the release and use of these technologies.

1.4. Clarifying key AI terms: generative AI and foundation models

AI models are actionable representations of all or part of the external context or environment of an AI system (encompassing, for example, processes, objects, ideas, people and/or interactions taking place in context). AI models use data and/or expert knowledge provided by humans and/or automated tools to represent, describe and interact with real or virtual environments (OECD, 2022[34]).

Generative AI (genAI) models create new outputs (e.g., text, code, audio, images, video), often in response to prompts, based on their training data (OECD, 2023[35]).

Foundation models, sometimes referred to as or categorised under “general-purpose” AI models, are machine learning models that can be adapted to perform a wide range of downstream tasks, such as tasks involving text synthesis, behaviour prediction, image analysis and content generation (Bommasani et al., 2022[36]; Jones, 2023[37]). Foundation models can be standalone or integrated into a variety of downstream AI systems and models, either directly or after additional training referred to as “fine-tuning.” There are two primary types of foundation models: generative models, which learn the patterns and distribution of input data to create new, plausible outputs (such as OpenAI’s GPTs), and discriminative models, which predict data labels by distinguishing between different classes in a dataset (like Google’s BERT) (OECD, 2022[34]). It is important to note that not all generative or discriminative models qualify as foundation models.

1.5. This report explores the trends, benefits and risks of open-weight foundation models

For the purposes of this report, “open-weight models” refer to foundation models for which at least the trained model weights are publicly available for download for local deployments. These models are

characterised by their ability to generate content, perform various tasks, and adapt to different applications based on the data they have been trained on.

As noted above, there are discriminative foundation models, and generative AI models that are not foundation models. Additionally, there are benefits and risks associated with openly sharing the weights of machine learning models that are neither generative nor foundation models. The decision to focus on generative AI foundation models responds to feedback from national delegations on earlier versions of this report, the substantial efforts required to collect relevant data (see Section 2), and the relevance of this technology in current policy discussions.

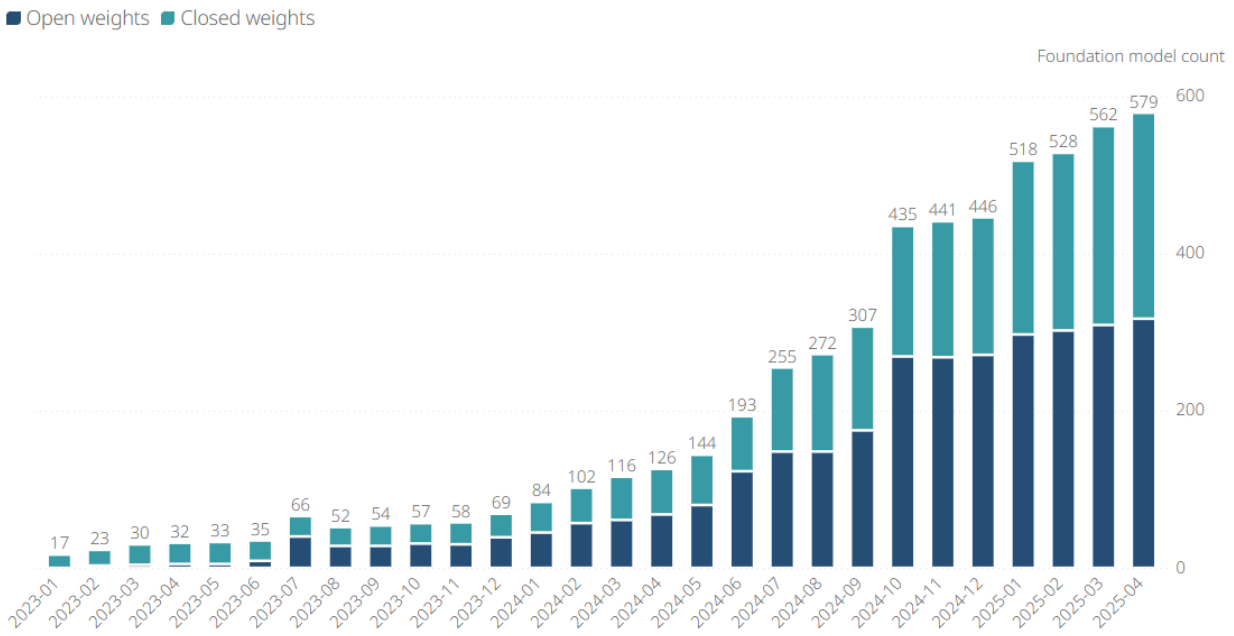
Licensing considerations are beyond the scope of this report. While licensing undeniably shapes how models are used and shared, its legal and jurisdictional complexity makes it impractical to address comprehensively in this report. Nonetheless, licensing remains a critical factor in the deployment and deployment of open-weight models and may warrant dedicated analysis in future work. By narrowing our scope, we aim to provide a clearer, more actionable discussion on open-weight foundation models.

2. Evolution of open-weight models

Since OpenAI launched GPT-3.5 (ChatGPT) in November 2022, the market for generative AI models has experienced significant growth, fuelled by new developers entering the field and existing developers introducing new models. Figure 2.1 reveals a marked acceleration in the global supply of generative AI foundation models, particularly from mid-2024 onward. Notably, open-weight models have not only kept pace with this growth but now account for approximately 55% of all available models as of April 2025. This suggests a trend toward a greater development and provision of open-weight models. This data focusses on foundation models made available commercially by one or more providers through an API endpoint, and is the product of OECD research on developments in AI markets (Box 2.1).

Figure 2.1. The supply of foundation models has increased consistently, with open-weight models representing over half of commercially available models

Number of unique closed and open-weight models made commercially available from providers worldwide each month through an API endpoint.



Note: Open weights indicate that the trainable weights – parameters optimised during training and fine-tuning – are made available for download for local deployments. See André et al. (2025[38]) for more details on the methodology.
Source: OECD.AI (2025), data from the AIKoD experimental database (internal), last updated 2025-04-30, accessed on 2025-05-14, https://oecd.ai/.

Box 2.1. The AIKoD database on active generative AI foundation models

The AI Knowledge on Demand (AIKoD) database is an internal, experimental OECD resource designed to explore developments in the generative AI market. It collects data and information from publicly available websites of cloud providers worldwide to track "active" generative AI models available as AI-as-a-Service. As of May 2025, the dataset includes models from AI developers in 14 countries, offered by 51 cloud providers across 11 countries, along with pricing and quality benchmarks. The number of providers, particularly in Asia and Europe, has been rapidly growing since mid-2024, highlighting the dynamic nature of the supply of AI models.

The database aims to maximise representativity across countries and regions. However, access restrictions in some countries (e.g., China) and the rapidly evolving AI cloud market may result in underrepresentation of new and fast-growing providers. Additionally, the database only includes publicly available models and does not account for strictly on-premise or undisclosed options. In order to collect historical information, the Internet Wayback Machine was consulted to gather data on prices and available models from the past.

The "provider" of the model is the cloud service firm that offers support and hosting services for the model. The "developer" refers to the firm that pre-trains and fine-tunes an AI model, which is then made available commercially by one or more providers through an API endpoint for users. A model's country assignment is determined by the location of the headquarters of the provider or the developer.

The database collects all available model endpoints listed on a provider's pricing webpage and classifies them by their foundation models. Foundation models are identified by three characteristics: the *model family* (e.g., GPT, LLaMA, Claude); the *model variant* (a specific adaptation or configuration of the base model, e.g., GPT-*4*, LLaMA-*3*, Claude-*sonnet*); and the *model version* (a specific release iteration, e.g., GPT-4*o,* Claude-sonnet-*3.7*).

The dataset also indicates whether the model weights – parameters optimised during training and fine-tuning – are available for download for local deployments. A model is classified as open weights in the database if it is specifically labelled as such by the developer or available on platforms like Hugging Face or Ollama. Additionally, models available for commercial use from a cloud provider other than the developer and lacking partnership information (e.g., open-weight models with a permissive license such as Mistral-7B or Deepseek-R1 which are available on Microsoft Azure) are also classified as open weights. A final manual validation based on expert judgment is conducted for models that do not fit these categories, including closed models made accessible from providers other than the model's developer under specific distribution agreements (e.g., Anthropic models available on GCP and AWS).

It is important to note that the term "open-weight model" in the AIKoD database refers only to the foundation model from the original developer and does not include fine-tuned variations released by other developers. This explains why the number of models available in Hugging Face is much larger than those reported in the AIKoD database.

Models are further categorised based on the modality of user interaction, that is, the type of input prompt and generated output. Available modalities include text-to-text, text-to-image, text-to-audio or audio-to-text. Models that support two or more modalities are classified as multimodal.

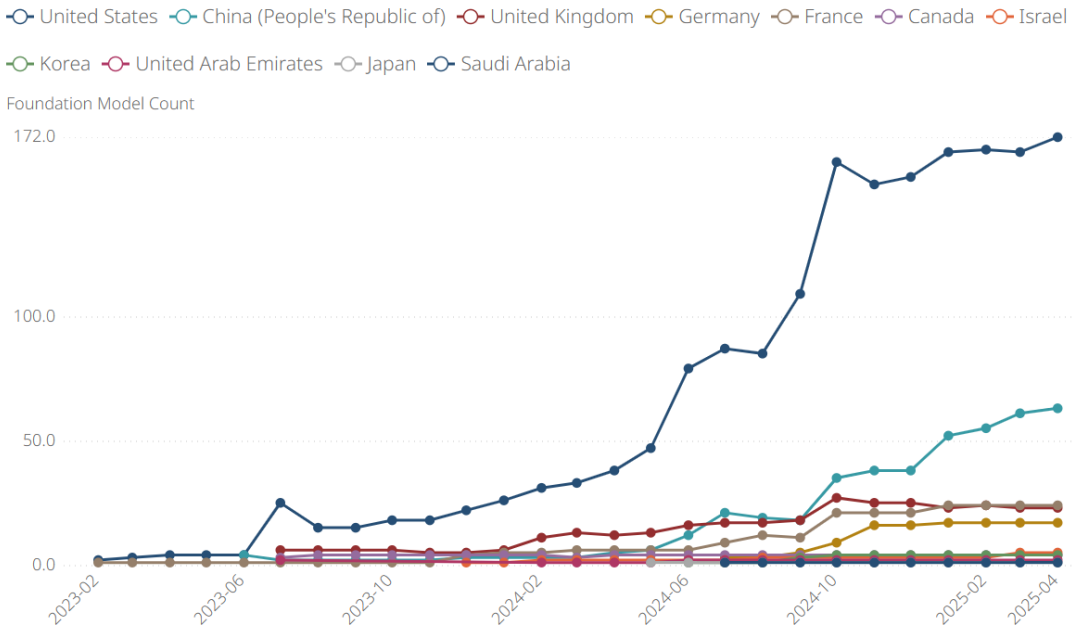
Source: André et al. (2025[38]).

Figure 2.2 illustrates the evolving landscape of open-weight model development and provision across countries. The United States leads in both dimensions, reflecting its robust AI ecosystem and cloud infrastructure. China and France also emerge as key developers, while the Netherlands and Singapore

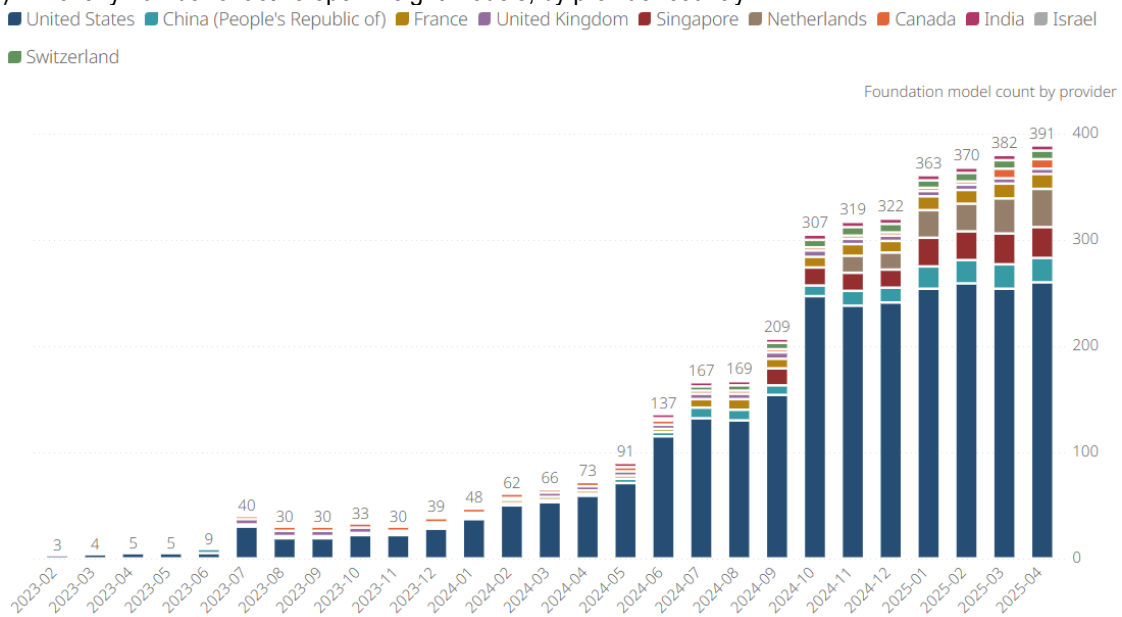
stand out as major provider hubs, despite having fewer domestic developers. This divergence underscores the global nature of AI deployment, where models are often hosted in countries with advanced cloud capabilities, regardless of their origin. The data also reveals a growing international dispersion of model provision.

Figure 2.2. The United States, China and France are at the forefront of open-weight model development, with the largest offerings coming from providers in the US, the Netherlands and Singapore

a) Monthly number of active open-weight models, by developer country.



b) Monthly number of active open-weight models, by provider country.

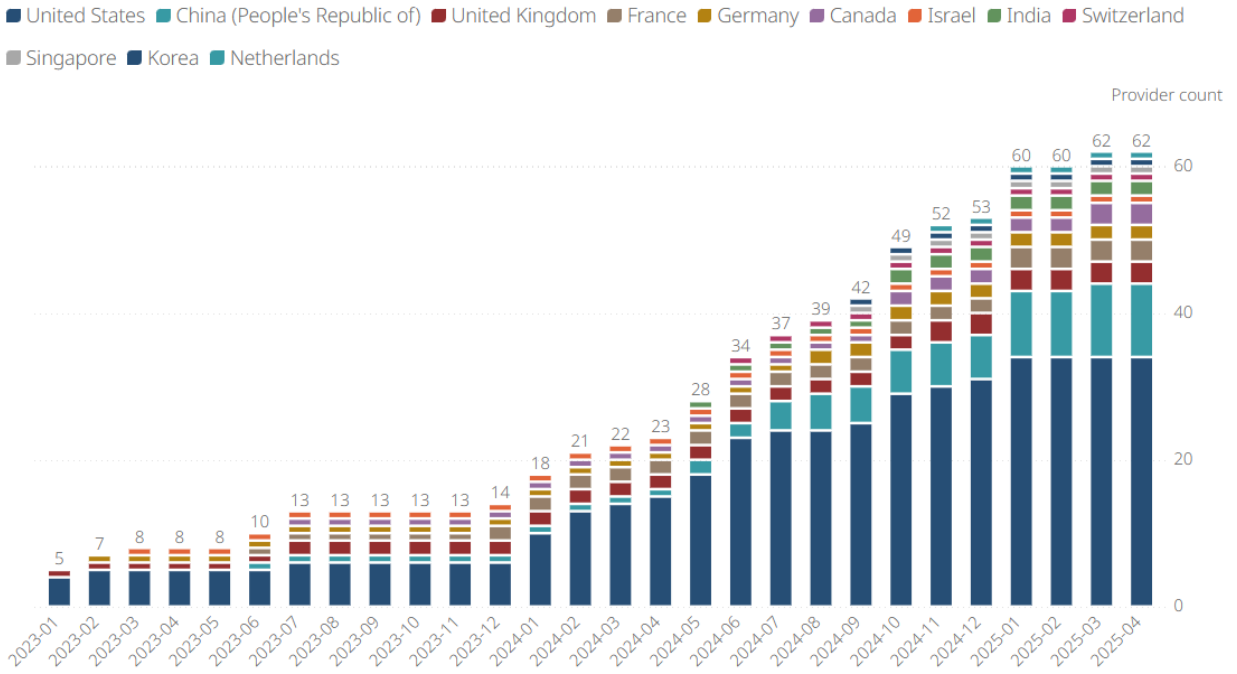


Note: "Developers" are firms that pre-train and fine-tune an AI model, which is then made available commercially by one or more providers through an API endpoint for users, such as OpenAI, Anthropic, and Google. “Provider” refers to the cloud service company that supports and hosts the model. "Country" refers to where either the developers (panel a) or the providers (panel b) are based. Some companies, like OpenAI, Google, Microsoft, and Amazon, both develop and provide models, while others, like Meta, only develop models. Additionally, some companies, such as ReplicateAI, PerplexityAI, and Deep Infra, only provide models created by other developers. A single developer can offer multiple models in different formats and can distribute them via different cloud providers. See André et al. (2025[38]) for more details on the methodology. Source: OECD.AI (2025), data from the AIKoD experimental database (internal), last updated 2025-04-30, accessed on 2025-05-14, https://oecd.ai/.

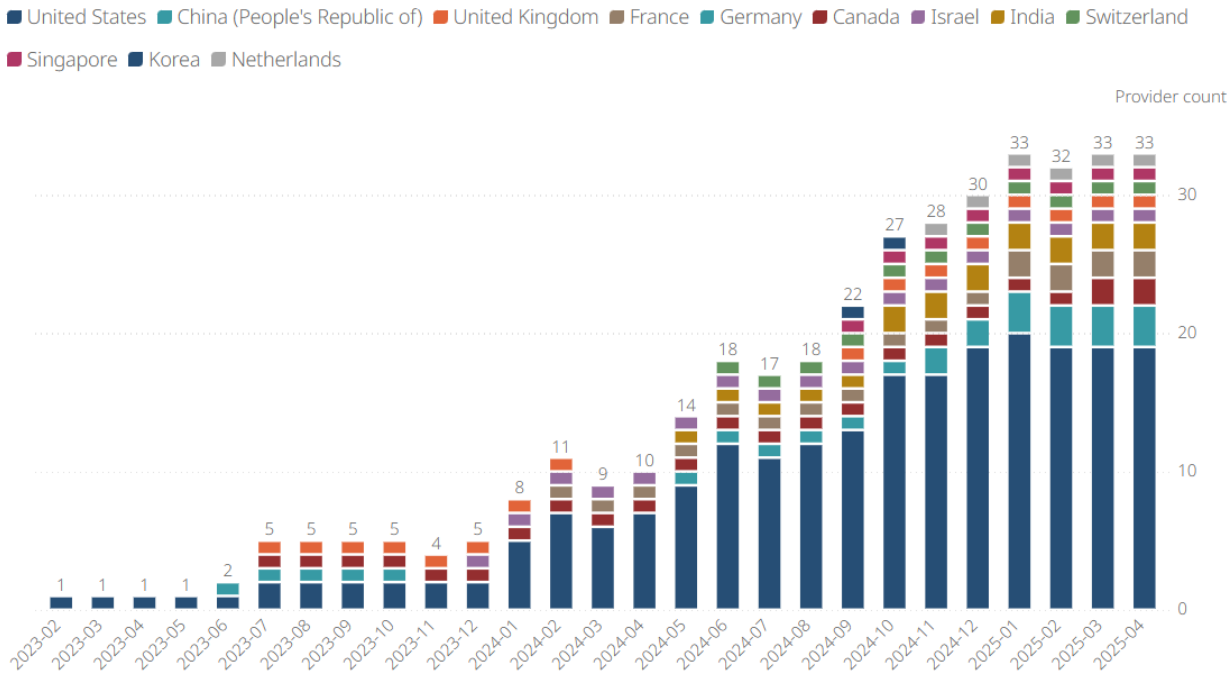
Figure 2.3 highlights the geographic concentration of foundation model providers, revealing United States dominance, which accounts for over half of all providers globally. This leadership is also present in the development of open-weight models. While several other countries such as China, the United Kingdom, France, and Germany also contribute to the ecosystem, their presence is markedly smaller.

Figure 2.3. Over half of foundation model providers are in the United States

a) Provider count per country, all foundation models.



b) Provider count per country, open-weight models.

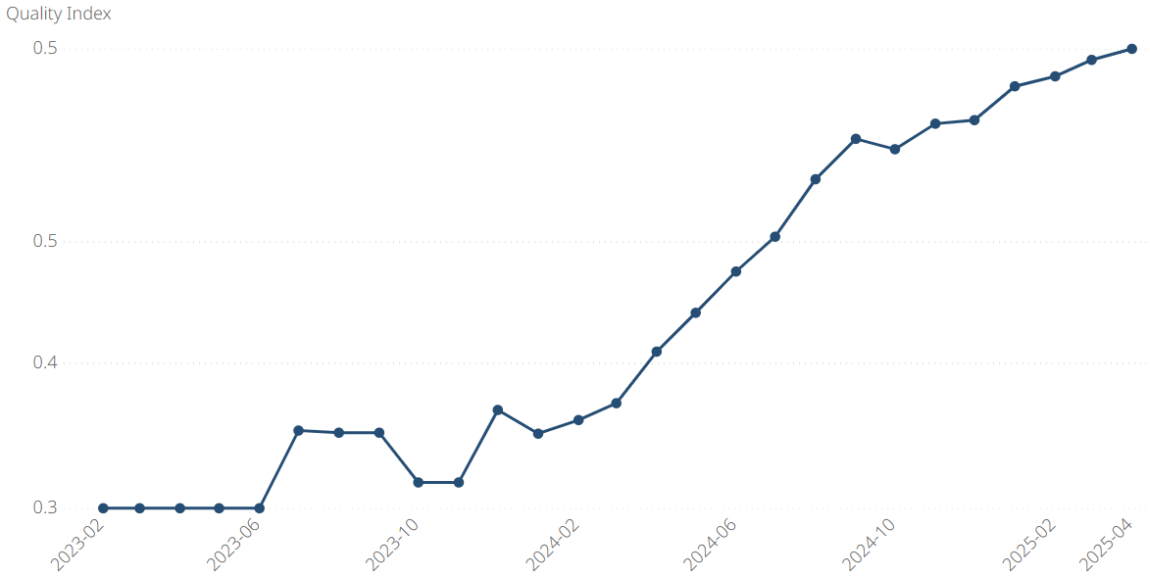


Note: "Provider" refers to the cloud service company that supports and hosts the model. See André et al. (2025[38]) for more details on the methodology.
Source: OECD.AI (2025), data from the AIKoD experimental database (internal), last updated 2025-04-30, accessed on 2025-05-14, https://oecd.ai/.

The supply of generative AI foundation models is predominantly driven by text-to-text models, including coding assistants and models with multi-modal features. These models account for 78% of all offerings. Text-to-image models, which focus on image generation, represent 18% of the total supply, while audio-to-text models make up a smaller share at 2.5%. Notably, the average quality of open-weight text-to-text models – including foundation models and all their variants and updates – has seen rapid improvement since early 2024, as illustrated in Figure 2.4. This trend highlights the significant advancements in the performance of LLMs over a short period of time.

Figure 2.4. Significant gains in the quality of open-weight models

Average quality (performance) of text-to-text open-weight models over time.



Note: The quality index is a normalised weighted measure that reflects how well a model performs a task, based on standard benchmarks of the industry acquired from Artificial Analysis and Hugging Face. Text-to-text models take the average across Hugging Face's MMLU scores, Arena ELO scores, and the GPQA value. Quality is averaged across all model offerings, even when the same model is offered by multiple providers. Please note that this is an average, so a drop in quality may happen if the proportion of lower-quality models increases. See André et al. (2025[38]) for more details on the methodology.
Source: OECD.AI (2025), data from the AIKoD experimental database (internal), last updated 2025-04-30, accessed on 2025-05-14, https://oecd.ai/.

The data in this section indicates that open-weight models are increasing in both quantity and quality. Recent launches of open-weight models, such as OpenAI's GPT-OSS, underscore this point. This highlights the need to monitor openness in AI development and deployment as a key dimension of technological access and policy approaches.

3. Benefits and risks of openly releasing the weights of foundation models

This section highlights some of the benefits and risks associated with open-weight models. Rather than providing a comprehensive overview, it aims to illustrate key advantages, such as enhanced innovation and collaboration, alongside potential challenges, including privacy concerns and the risk of malicious use. By presenting this information clearly and concisely, the section seeks to inform policy discussions on balancing the openness of generative AI foundation models with responsible AI policy approaches.

3.1. Illustrative benefits

The benefits of open-weight models encompass a range of illustrative examples, presented here in no particular order:

- **Facilitate beneficial innovation:** Open-weight models can accelerate AI research and development, driving innovation and the integration of new downstream applications (Creative Commons et al., 2023[9]; Engler, 2021[2]; Engler, 2021[3]). These models allow developers to build on existing technologies, promoting collaboration and experimentation that can lead to significant advancements across industries. This collaborative environment accelerates innovation and encourages the exploration of new use cases, ultimately broadening AI's impact on everyday life (Gans, 2025[39]).
- **External evaluation**: Open-weight models facilitate independent evaluations of projects by wider communities of developers and contributions from individuals, thus enabling more robust evaluations of performance and risk. Involving the wider AI community also facilitates audit and analysis of open-weight models and their components (*e.g.*, training data, weights, documentation) which can help detect bugs, biases, and other issues (Creative Commons et al., 2023[9]; Bommasani et al., 2023[40]). Also, assessing whether the model performs as well as its creators claim is a key form of evaluation.
- **Enable efficiencies in AI development:** Open-weight models enable large-scale collaborative efforts and allow downstream developers to optimise existing models instead of having to start from scratch for each new application – which may help reduce resource consumption and costs associated to AI development (HuggingFace, 2023[41]).
- **Facilitate talent development:** Opening the weights of foundation models could enhance talent development. More people being able to interact with pre-trained cutting edge-models may, over time, lead to a larger AI talent pool and, in the longer run, help address the digital divide by enabling initiatives from across regions.
- **Expand use, adoption, and market choice:** Open-weight models expand the developer community and foster competition by lowering barriers to innovation and entry, enabling collaboration and providing opportunities for skill development. This encourages participation from individuals in new or less common regions and backgrounds, leading to the creation of applications

that address the specific needs of various user groups, such as generative AI tools for various languages and cultural contexts. This could in turn enable a wider range of people to use and benefit from AI applications.

- **Support sensitive data management:** Open-weight models facilitate the adoption of AI by businesses and governments that may lack the resources to develop proprietary AI solutions independently and possess sensitive data that cannot be shared with vendors of closed-weight models (The White House, 2025[42]).
- **Enable on-device solutions:** Direct access to model weights facilitates on-device deployment, crucial for offline functionality or in environments with stringent privacy requirements. This eliminates reliance on internet connectivity and third-party API access and could help to address concerns around data transmission and control (Zheng, Chen and Bin, 2024[43]).
- **Improve digital security and safeguards:** Releasing foundation model weights can strengthen cybersecurity by enabling red teams to legally test and simulate potential attack scenarios using the same tools adversaries might exploit. This enhances adversary emulation, allowing for more realistic and effective defence strategies. Moreover, open-weight models can be more easily customised to implement safeguards tailored to specific operational contexts.
- **Prevent unintended and harmful behaviours:** Open-weight foundation models can help prevent unintended and harmful behaviours, such as the generation of child sexual abuse material (CSAM) and privacy violations. By providing greater access to model weights, architectures and training data and processes, researchers can better identify the causes of such behaviours, align AI outputs with user values, and improve the detection of AI-generated content. This transparency allows for more effective evaluation and fine-tuning of models, ultimately leading to trustworthy AI systems that are less likely to produce harmful results (Al-Kharusi et al., 2024[44]; Thiel, 2023[45]; Hendrycks et al., 2022[46]).
- **Enhance alignment and explainability research:** Alignment research seeks to ensure that AI systems reflect user or developer preferences and values, often requiring model fine-tuning through methods like reinforcement learning. While this fine-tuning can be done via APIs (OpenAI, 2023[47]), these interfaces may not always provide sufficient information about the underlying models for meaningful analysis. Additionally, some aspects of explainability research necessitate direct modifications to model parameters and activation patterns, which require full or nearly full access to the models.
- **Distribute influence:** Open-weight model development allows a wide community to influence AI's evolution. This distribution has economic, social, and political implications, which may enable broader sharing of AI's potential benefits (Bommasani et al., 2023[40]; Howard, 2023[48]; LAION.ai, 2023[49]).

While the benefits of open-weight models are significant, their realisation may be limited by access to computing power, data resources, and available talent.

3.2. Illustrative risks

The risks of open-weight models encompass a range of illustrative examples, presented here in no particular order:

- **Downstream impacts and proliferation of risks**: Both open- and closed-weight foundation models can be used or combined with other tools, models or services (e.g., such as the Internet or third-party APIs) in unintended or harmful ways. However, open-weight models increase the accessibility for a wider range of users – including malicious actors – to fine-tune, modify and deploy models

without the original developer's control. This greater accessibility can accelerate the spread and amplification of existing risks, unresolved issues and vulnerabilities (Bran et al., 2023[50]; Boiko et al., 2023[51]). Additionally, safeguards implemented by foundation model developers – including guardrails to alleviate inaccurate outcomes – may be weakened or removed when models are altered downstream, potentially leading to new or unforeseen issues. While closed models also carry risks, their restricted access can limit the scale and speed at which these risks propagate.

- **Challenges in monitoring and fixing:** While structured API access allows for monitoring and potential harm detection, open-weight models rely on downstream developers to implement fixes. This can hinder effective integration of updates, as developers may avoid updating due to lack of skills or resources or simply to retain certain model functionalities, complicating the resolution of risks and vulnerabilities downstream. Once model weights are publicly available, complete recalling of the model and removal of all copies is unfeasible (Bengio et al., 2025[52]).
- **Model vulnerability exposure:** Releasing the weights of one model can reveal vulnerabilities in other models and enable more sophisticated *jailbreaking*. For example, researchers have developed techniques that leverage the weights of a model to create "adversarial suffixes", which are sequences that, when appended to a query, compel the model to produce harmful content. This method, developed using open-weight models, is transferable and can also be applied to closed-weights models. Thus, releasing one model's weights could expose vulnerabilities in others (Zou et al., 2023[53]).
- **Intellectual property violations:** Releasing the weights of foundation models may present risks to intellectual property rights, primarily concerning the unauthorised reproduction, adaptation, and commercial exploitation of copyrighted material used in their training data. Research has demonstrated that LLMs can indeed memorise and extract copyrighted content to varying extents, with this information being stored within the model parameters (Cooper et al., 2025[54]). Once weights are public, the original developers lose control over how the models are used or altered, making it challenging to track and prevent misuse that could lead to copyright violations.
- **Malicious use:** Releasing the weights of foundation models can increase their vulnerability to malicious use. With access to a model's weights and architecture, individuals with the necessary knowledge, skills, and compute resources can write or modify inference code to run the model without the protocols typically implemented by closed model providers. This allows them to fine-tune the model, potentially allowing or enhancing harmful outputs. Fine-tuning can also be performed on closed models through an API. However, such fine-tuning can be monitored, e.g., the API owner can inspect the contents of the fine-tuning data set, which may allow them to prevent or at least detect malicious activity. Fine-tuning open-weight models generally requires a higher level of technical expertise compared to proprietary, "ready-to-use" AI services. Examples of malicious use include:
    - **Digital security risks**: Releasing model weights could provide both malicious actors and cybersecurity red teams with increased novel means to conduct, analyse, and emulate offensive cyber activities (Mirsky et al., 2021[55]; Buchanan, 2020[56]). While open weights can benefit cybersecurity defenders, they often require complex infrastructure that is not affordable for many cyber actors. However, small open-weight models can help malicious actors automate phishing campaigns, conduct opensource intelligence (OSINT) research, and perform routine programming tasks. Because offensive cyber tasks often look similar to software engineering in defensive cybersecurity, developers of both open-weight and proprietary models find it challenging to effectively prevent misuse. Some researchers argue that because foundation models are more complex than regular software, releasing model weights may favour attackers, as quick access may allow them to exploit vulnerabilities more rapidly, while developing

solutions takes time and resources. Even when solutions are created, they may not fully resolve the issues (Shevlane and Dafoe, 2020[57]).

- **Generation of child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII):** Research indicates that releasing the weights of image-generation models like Stable Diffusion has led to a significant rise in the creation of NCII and CSAM, highlighting the challenges of monitoring the use of open-weight models (Thiel, Stroebel and Portnoff, 2023[58]). While techniques such as prompt filtering and output filtering can reduce the likelihood of harm from generated CSAM or NCII, it is more difficult to enforce the use of such techniques with open-weight models than with models provided as a service. The widespread distribution of CSAM and NCII creates significant harm to women, teenagers, and other vulnerable groups.
- **Privacy risks:** The accessibility of foundation model weights trained on sensitive or personally identifiable information increases privacy risks by enabling more effective membership inference (i.e., attackers exploit open model weights to determine if a specific data point was part of the training set); attribute inference (i.e., attackers leverage open model weights to deduce sensitive characteristics about the aggregate training data); and exploitation of memorisation vulnerabilities (i.e., attackers leverage open weights to facilitate the identification and extraction of specific private data points inadvertently memorised by the model). These methods could result in the extraction of sensitive training data from the model (Kandpal et al., 2024[59]; Nasr et al., 2023[60]; Carlini et al., 2022[61]).

- **Unpredictable agentic deployments**: Some current and emerging applications of foundation models allow these models to access external tools to interact with their environment (Yong, Shi and Zhang, 2025[62]). This approach leverages the broad knowledge and generative capabilities of a foundation model as an autonomous agent. Releasing the weights of a highly capable foundation model could facilitate a substantial expansion in the range of tools and environments it can access, leading to more complex attribution challenges and unpredictable interactions in both physical and virtual environments.

Some of the risks associated with releasing foundation model weights are still speculative. For example, some experts suggest that open-weight models could be misused in biological or chemical contexts, although the evidence remains inconclusive (Mouton, Lucas and Guest, 2024[63]; Peppin et al., 2025[64]). Additionally, the risks posed by open-weight models share similarities with those posed by other technologies, such as Internet search engines or closed-weight models, and these risks could be heightened or reduced by releasing model weights.

3.3. Marginal benefits and risks as part of holistic risk assessments

Emphasising the importance of assessing risks and benefits "on the margin" – the additional risks and benefits associated with releasing foundation model weights, compared to risks posed by closed models or existing technology – is crucial for understanding the true impact of open-weight models (Bengio et al., 2025[52]). This approach allows stakeholders to evaluate how these models compare to existing tools and practices – both AI and non-AI – as well as to consider the potential outcomes in their absence (Kapoor et al., 2024[65]).

For example, if an open-weight model enhances productivity in content creation compared to traditional methods, the marginal benefit presents an argument in favour of open release. Conversely, if the risks associated with malicious use are significantly greater than those posed by existing technologies, it becomes essential to evaluate whether the new models offer improved risk mitigation strategies or greater benefits that justify their release.

It is also helpful for decision-making based on marginal impact to consider alternative methods other than releasing model weights for achieving the same benefits, and alternatives other than withholding model weights for mitigating identified risks (Whittlestone and Ovadya, 2020[66]).

By focusing on marginal assessments, decision-makers can better gauge whether the advantages of open-weight models outweigh the potential downsides. Addressing marginal risk is crucial to ensure that interventions are appropriate and proportional to the level of risk involved (Kapoor et al., 2024[65]). However, it's important to recognise that this is just one approach to risk assessment. Depending on the context, other baselines may be more appropriate. A key concern with relying solely on marginal comparisons is the potential for a “boiling frog” effect, where overall risk tolerance increases as each successive model is compared to an increasingly permissive baseline, especially as model capabilities evolve or usage patterns shift. A more holistic and adaptive risk framework is needed to ensure that AI development and deployment remains trustworthy.

4. Conclusions

Releasing foundation model weights has many benefits, such as allowing external evaluation, speeding up innovation, and spreading control over a potentially transformative technology. Open-source practices in the software industry have shown substantial benefits, distributing influence over the direction of technological innovation and facilitating the development of products that can reach new audiences. However, releasing foundation model weights also presents potential for malicious use and unintended consequences, such as cyberattacks, sexual abuse, and violation of intellectual property and privacy rights. Because of the significant potential risks and benefits, foundation models need careful consideration when they are shared and used.

Opening the weights of foundation models does not inherently result in malicious use, as these models can have both positive and negative impacts depending on the context and application. To better understand the trade-offs of open-weight models, it is important to evaluate each one in the context of their specific applications and compare these benefits and risks to those of existing tools. This approach helps identify any additional – or marginal – benefits and risks that may arise.

While initial assessments suggest that certain risks associated with open-weight models, such as the generation of malicious content, share similarities with existing digital tools, rapid advancements in AI – including significantly decreasing compute costs (Hobbhahn and Besiroglu, 2022[67]) and increasingly accessible fine-tuning techniques (Yang et al., 2024[68]) – could dramatically lower the technical and financial barriers for a broader range of actors, including those seeking to develop harmful applications or bypass security mechanisms.

Therefore, developers should carefully weigh the decision to release model weights, considering the full range of marginal benefits and risks and the full range of options for achieving the benefits or mitigating the risks.

While this report examines the availability of open-weight foundation models, future research could investigate the actual usage of these models and the underlying reasons for their application, particularly focusing on the economic implications and spillover effects associated with their development and deployment. Analysing who produces these models – ranging from large tech companies to academic institutions – and who uses them, including governments, startups and individual consumers, is crucial for understanding the dynamics of the evolving AI ecosystem.

References

Al-Kharusi, Y. et al. (2024), "Open-Source Artificial Intelligence Privacy and Security: A Review", *Computers*, Vol. 13/12, p. 311, https://doi.org/10.3390/computers13120311. [44]

Anandira, H. (2024), "Meta Platforms under fire over open-source AI branding", *Mobile World Live*, https://www.mobileworldlive.com/ai-cloud/meta-platforms-under-fire-over-open-source-ai-branding/. [31]

André, C. et al. (2025), "Developments in Artificial Intelligence markets: New indicators based on model characteristics, prices and providers: New indicators based on model characteristics, prices and providers", *OECD Artificial Intelligence Papers*, No. 37, OECD Publishing, Paris, https://doi.org/10.1787/9302bf46-en. [38]

Anthropic (2023), *Claude 2*, Anthropc, https://www.anthropic.com/index/claude-2. [26]

Assran, M. et al. (2023), *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*, https://doi.org/10.48550/arXiv.2301.08243v3. [28]

Bengio, Y. et al. (2025), "International AI Safety Report (DSIT 2025/001, 2025)", https://www.gov.uk/government/publications/international-ai-safety-report-2025. [52]

Boiko et al. (2023), *Emergent autonomous scientific research capabilities of large language models*, http://arxiv.org/abs/2304.05332. [51]

Bommasani et al. (2022), *On the Opportunities and Risks of Foundation Models*, http://arxiv.org/abs/2108.07258. [36]

Bommasani, R. et al. (2023), *Considerations for Governing Open Foundation Models*, https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf. [40]

Bran et al. (2023), *ChemCrow: Augmenting large-language models with chemistry tools*, http://arxiv.org/abs/2304.05376. [50]

Brockman, G. et al. (2023), *Introducing ChatGPT and Whisper APIs*, https://openai.com/blog/introducing-chatgpt-and-whisper-apis. [27]

Buchanan, B. (2020), *A National Security Research Agenda for Cybersecurity and Artificial Intelligence*, https://cset.georgetown.edu/publication/a-national-security-research-agenda-for-cybersecurity-and-artificial-intelligence/. [56]

Carlini, N. et al. (2022), "Membership Inference Attacks From First Principles", *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897-1914, https://doi.org/10.1109/sp46214.2022.9833649. [61]

Choose a License (2023), *Licenses*, Choose a License, https://choosealicense.com/licenses/. [12]

Commons, C. (2025), *About CC Licenses: The CC License options*, https://creativecommons.org/share-your-work/cclicenses/ (accessed on 10 May 2025). [33]

Cooper, A. et al. (2025), *Extracting memorized pieces of (copyrighted) books from open-weight language models*, arXiv, https://arxiv.org/abs/2505.12546. [54]

Creative Commons et al. (2023), *Supporting Open Source and Open Science in the EU AI Act*, https://creativecommons.org/2023/07/26/supporting-open-source-and-open-science-in-the-eu-ai-act/. [9]

DiBona, C., S. Ockman and M. Stone (eds.) (1999), *The Open Source Definition*, O'Reilly. [11]

Engler, A. (2021), *How open-source software shapes AI policy*, Brookings, https://www.brookings.edu/articles/how-open-source-software-shapes-ai-policy/. [2]

Engler, A. (2021), *The EU's attempt to regulate open-source AI is counterproductive*, Brookings, https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive. [3]

Finley, K. (2011), *How to Spot Openwashing*, readwrite, https://readwrite.com/how_to_spot_openwashing/. [15]

Fries et al. (2022), *How Foundation Models Can Advance AI in Healthcare*, Stanford HAI, https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare. [4]

Gans, J. (2025), *Growth in AI Knowledge*, National Bureau of Economic Research (NBER), https://www.nber.org/system/files/working_papers/w33907/w33907.pdf. [39]

Goldman, S. (2023), *Hugging Face, GitHub and more unite to defend open source in EU AI legislation*, VentureBeat, https://venturebeat.com/ai/hugging-face-github-and-more-unite-to-defend-open-source-in-eu-ai-legislation/. [8]

Hendrycks, D. et al. (2022), *Unsolved Problems in ML Safety*, https://arxiv.org/abs/2109.13916 (accessed on 24 July 2023). [46]

Hobbhahn, M. and T. Besiroglu (2022), *Trends in GPU Price-Performance*, https://epochai.org/blog/trends-in-gpu-price-performance (accessed on 23 July 2023). [67]

Hoffmann, J. et al. (2022), "Training Compute-Optimal Large Language Models", *arXiv*, https://arxiv.org/abs/2203.15556. [24]

Howard, J. (2023), *AI Safety and the Age of Dislightenment*, https://www.fast.ai/posts/2023-11-07-dislightenment.html (accessed on 31 July 2023). [48]

HuggingFace, C. (2023), *Supporting Open Source and Open Science in the EU AI Act*, https://huggingface.co/blog/assets/eu_ai_act_oss/supporting_OS_in_the_AIAct.pdf. [41]

Inskeep, S. and O. Hampton (2023), *Meta leans on 'wisdom of crowds' in AI model release*, NPR, https://www.npr.org/2023/07/19/1188543421/metas-nick-clegg-on-the-companys-decision-to-offer-ai-tech-as-open-source-softwa. [29]

Jones, E. (2023), *What is a foundation model?*, Ada Lovelace Institute, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/. [37]

Kandpal, N. et al. (2024), "User Inference Attacks on Large Language Models", https://arxiv.org/pdf/2310.09266. [59]

Kapoor, S. et al. (2024), *On the Societal Impact of Open Foundation Models*, https://arxiv.org/pdf/2403.07918v1. [65]

LAION.ai (2023), *A Call to Protect Open-Source AI in Europe*, https://laion.ai/notes/letter-to-the-eu-parliament (accessed on 21 September 2023). [49]

Langenkamp et al. (2022), "How Open Source Machine Learning Software Shapes AI", *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, https://doi.org/10.1145/3514094.3534167. [1]

Liang, P. et al. (2022), *The Time Is Now to Develop Community Norms for the Release of Foundation Models*, Stanford CERM, https://crfm.stanford.edu/2022/05/17/community-norms.html. [21]

LinuxFoundation (2024), *Embracing the Future of AI with Open Source and Open Science Models*, https://lfaidata.foundation/blog/2024/10/25/embracing-the-future-of-ai-with-open-source-and-open-science-models/ (accessed on 21 January 2025). [14]

LinuxFoundation (2024), *Introducing the Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI*, https://lfaidata.foundation/blog/2024/04/17/introducing-the-model-openness-framework-promoting-completeness-and-openness-for-reproducibility-transparency-and-usability-in-ai/ (accessed on 21 January 2025). [18]

Maffulli, S. (2023), *Meta's LLaMa 2 license is not Open Source*, open source initiative, https://opensource.org/blog/metas-llama-2-license-is-not-open-source. [17]

Marr, B. (2023), *Digital Twins, Generative AI, And The Metaverse*, Forbes, https://www.forbes.com/sites/bernardmarr/2023/05/23/digital-twins-generative-ai-and-the-metaverse. [6]

Marr, B. (2023), *The Amazing Ways Duolingo Is Using AI And GPT-4*, Forbes, https://www.forbes.com/sites/bernardmarr/2023/04/28/the-amazing-ways-duolingo-is-using-ai-and-gpt-4/. [7]

Meta (2023), *Introducing LLaMA: A foundational, 65-billion-parameter large language model*, https://ai.facebook.com/blog/large-language-model-llama-meta-ai/. [22]

Milmo, D. (2023), *Nick Clegg defends release of open-source AI model by Meta*, The Guardian, https://www.theguardian.com/technology/2023/jul/19/nick-clegg-defends-release-open-source-ai-model-meta-facebook. [30]

Mirsky, Y. et al. (2021), *The Threat of Offensive AI to Organizations*, http://arxiv.org/abs/2106.15764 (accessed on 19 September 2023). [55]

Mouton, C., C. Lucas and E. Guest (2024), "The Operational Risks of AI in Large-Scale Biological Attacks: Results from a Red-Team Study", *RAND Research Report*, https://www.rand.org/pubs/research_reports/RRA2977-2.html. [63]

Nasr, M. et al. (2023), "Scalable Extraction of Training Data from (Production) Language Models", https://arxiv.org/abs/2311.17035. [60]

OECD (2023), "AI language models: Technological, socio-economic and policy considerations", *OECD Digital Economy Papers*, No. 352, OECD Publishing, Paris, https://doi.org/10.1787/13d38f92-en. [35]

OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [34]

OpenAI (2023), *Fine-tuning: learn how to customize a model for your application*, https://platform.openai.com. [47]

OpenAI (2023), *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses'*, OpenAI, https://openai.com/product/gpt-4. [25]

OpenAI (2023), *Inworld AI*, https://openai.com/customer-stories/inworld-ai. [5]

OSI (2025), *The Open Source AI Definition – draft v. 0.0.9*, https://opensource.org/ai/drafts/open-source-ai-definition-draft-v-0-0-9 (accessed on 21 January 2025). [13]

OSI (2025), *The Open Source Definition*, https://opensource.org/osd/ (accessed on 10 May 2025). [32]

Peppin, A. et al. (2025), *The Reality of AI and Biorisk*, https://arxiv.org/pdf/2412.01946v3. [64]

Sastry, G. (2023), *Beyond "Release" vs. "Not Release"*, Stanford CERM, https://crfm.stanford.edu/commentary/2021/10/18/sastry.html. [20]

Shevlane, T. and A. Dafoe (2020), *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*, http://arxiv.org/abs/2001.00463 (accessed on 7 December 2022). [57]

Sijbrandij, S. (2023), *AI weights are not open "source"*, https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/. [16]

Solaiman, I. (2023), *The Gradient of Generative AI Release: Methods and Considerations*, http://arxiv.org/abs/2302.04844. [19]

The White House (2025), *America's AI Action Plan*, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf. [42]

Thiel, D. (2023), "Identifying and Eliminating CSAM in Generative ML Training Data and Models", *Stanford Digital Repository*, https://doi.org/10.25740/kh752sm9123. [45]

Thiel, D., M. Stroebel and R. Portnoff (2023), "Generative ML and CSAM: Implications and Mitigations", *Stanford Digital Repository*, https://purl.stanford.edu/jv206yg3793. [58]

Urbina, F. et al. (2022), "Dual use of artificial-intelligence-powered drug discovery", *Nature Machine Intelligence*, Vol. 4/3, pp. 189-191, https://doi.org/10.1038/s42256-022-00465-9. [10]

Vincent, J. (2023), *Meta's powerful AI language model has leaked online — what happens now?*, The Verge, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse. [23]

Whittlestone, J. and A. Ovadya (2020), *The tension between openness and prudence in AI research*, http://arxiv.org/abs/1910.01170 (accessed on 24 July 2023). [66]

Yang, M. et al. (2024), *Low-Rank Adaptation for Foundation Models: A Comprehensive Review*, arXiv, https://arxiv.org/abs/2501.00365. [68]

Yong, X., G. Shi and P. Zhang (2025), "Towards Agentic AI Networking in 6G: A Generative Foundation Model-as-Agent Approach", *IEEE Communications Magazine*, https://arxiv.org/abs/2503.15764. [62]

Zheng, Y., Y. Chen and Q. Bin (2024), *A Review on Edge Large Language Models: : Design, Execution, and Applications.*, https://arxiv.org/pdf/2410.11845v1. [43]

Zou, A. et al. (2023), *Universal and Transferable Adversarial Attacks on Aligned Language Models*, http://arxiv.org/abs/2307.15043 (accessed on 2 August 2023). [53]