Context: Training Data and AI Model Adaptation

4

KEY MESSAGES

- *Data power artificial intelligence (AI)*. The quantity, quality, and diversity of data are essential to making AI systems effective and adaptable, enabling them to evolve and perform reliably in diverse real-world scenarios.
- *The data divide grows as the AI training data industry booms globally, yet private sector investment in low-income countries and middle-income countries remains negligible.* In 2023, cumulative venture capital investments in the AI training data industry reached US$32 billion, predominantly concentrated in advanced economies—56 percent in the United States, 15 percent in the European Union, and 2 percent in the United Kingdom. China and India contributed 17 percent and 3 percent, respectively, with the rest of the world making up the remaining 6 percent.
- *The dominance of English in text data affects the usefulness of generative AI models in non-English-speaking countries, while opportunities in nontext data are expanding.*
  - English is spoken by 19 percent of the global population, but it makes up 45 percent of global URLs, 56 percent of open-source data sets on the leading AI development platform Hugging Face, and nearly 98 percent of scientific papers.
  - For nontext data, the language distribution tends to be more diverse, offering opportunities for non-English-speaking countries. Only 21 percent of YouTube videos are estimated to be in English, with Hindi and Spanish accounting for 7.6 percent and 6.7 percent, respectively.
- *Market failures highlight the need for government intervention to bridge the data divide.*
  - Although the private sector is adopting innovative approaches such as synthetic data to fill data gaps, market failures such as reusability, vague property rights, and challenges in valuing data often lead to underinvestment and undersharing.
  - Governments play a critical role in addressing these gaps by curating and publishing government data, facilitating secure and ethical data exchange, and developing and promoting interoperable data formats and taxonomies. Civil societies can contribute to enhancing AI data diversity through open-source data sets and community involvement.

- *Open-source AI is proliferating and allows developing countries to adapt global solutions to their local needs.*
    - Developing countries could face challenges in adopting AI because models are misaligned with local needs. Open-source AI models can lower the barrier to entry for developing countries, enabling them to adapt and customize AI solutions to local needs without incurring high licensing costs.
    - Open-source models are growing rapidly and narrowing the performance gaps with proprietary models: Among the 308 notable AI models published from 2022 to June 2025, around half have open weights, and a quarter have both open weights and open-source training code. Open-source models are also increasingly challenging proprietary models in terms of performance.

Introduction

Data are the cornerstone of artificial intelligence (AI) development, profoundly influencing the capabilities and performance of AI models. Three key aspects—quantity, quality, and diversity—are critical in determining how effective and adaptable AI systems are across different applications. Addressing market failures and externalities in training data production is crucial for fostering AI development and digital inclusion. Market failures stem from the positive spillovers and reusability of data, vague property rights, challenges in assessing the value of data, and privacy and security concerns. Although private firms play a key role in training data generation, they often underinvest as market failures diminish their potential returns, particularly for small and medium enterprises (SMEs) in developing economies. Public intervention is essential to address these failures.

For many developing countries, the challenge is not just obtaining large, high-quality data but also adapting AI models to their specific economic, cultural, and institutional contexts. Some countries are exploring local AI models that are better suited to their unique needs. However, the process is often costly and sometimes not feasible because of less-advanced countries' limited resources and expertise. Open-source AI models are more accessible and cost-effective, but they may require more technical resources that countries may lack. The advantages and challenges of developing proprietary systems or adapting open-source models vary depending on each application's requirements, the technological capabilities, and the specific needs, underscoring the importance of a tailored approach to AI development and deployment.

This chapter analyzes the availability and quality of AI training data, particularly text data in developing countries, and explores how various stakeholders can address training data gaps. It also examines the adaptability of AI models in diverse country contexts and governments' considerations between open-source and proprietary AI models.

Effective AI models hinge on training data

The capability of AI is fundamentally shaped by the data used to train the model, much as how the books a person reads deeply influence their thinking. In AI, the term *training* refers to the process by which a model learns from data, adjusting the parameters and structures of algorithms to improve the performance of specific tasks such as classification or prediction. A training data set is comparable to a textbook, providing the foundation for AI models to learn and shaping their ability to recognize patterns, generate content, or make decisions.

AI models heavily rely on the quantity, quality, and diversity of the training data, which we summarize as the three pillars of effective training data.

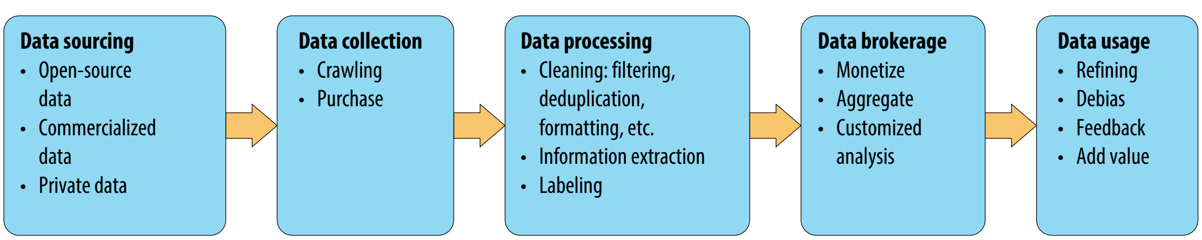
- *Quality* encompasses factors such as accuracy, completeness, granularity, timeliness, intensity of information, being error-free, and ensuring fairness while minimizing bias in the data.
- *Quantity* refers to the scale or size of the data set used for training, which directly affects the model's ability to learn and generalize from data.
- *Diversity,* or *representativeness,* refers to the data's ability to reflect the social, economic, institutional, cultural, and linguistic diversity of populations, aligning with the specific contexts in which AI is applied.

The AI training data industry is growing rapidly. *Fortune Business Insights* (2025) estimated that the market value was approximately US$2.26 billion in 2023, with projections of US$17.04 billion by 2032, representing a compound annual growth rate of 22 percent. In addition, venture capital (VC) investments in data have surged in recent years. The cumulative VC investments from 2012 to 2023 reached US$32 billion, with a significant acceleration after 2020, according to OECD.ai (https://oecd.ai/en/). This influx of capital demonstrates the evolution in the AI training data industry, where vendors strategically respond to the escalating demand for AI capabilities.

The process of obtaining high-quality AI training data involves a comprehensive data production chain. Data, as a vital asset in the AI industry, flow through several key stages: sourcing, collection, processing, brokerage, and usage (refer to figure 4.1). The process begins with sourcing data from various origins, such as open-source, commercialized, or private data, followed by collection through crawling or purchasing. The data then undergo processing, where they are cleaned, filtered, deduplicated, formatted, and sometimes labeled to enhance quality and meet the requirements of AI models. Developers have the option to manage these stages in house or acquire data from brokers who have already processed them. During the usage phase, data are refined by AI developers to fit specific AI models, and this process often loops back to dynamically adjust data production for continuous improvement.

At each step, opportunities and challenges emerge as the industry evolves dynamically. Data sourcing and collection, the pivotal initial steps in AI model development, often involve navigating complex legal and ethical landscapes, especially when using web-crawled content from media sources or

FIGURE 4.1 Production chain for training data



Source: Original figure for this publication.

even private data. The vast amount of text, images, audio, and videos available on social media and traditional media present substantial opportunities for AI training. However, web crawling these sources raises concerns over data ownership, copyrights, use rights, and privacy, underscoring the importance of careful legal and ethical consideration in the sourcing process to maintain the integrity and trustworthiness of AI technologies.

In the era of generative AI (GenAI), data communities have become a key source for AI training data. Unlike private entities, these communities offer valuable data sets as public goods at no cost, with platforms such as Hugging Face hosting data sets comparable in size to those used to train models such as GPT-4 and Llama 3.1. These contributions are often overlooked in traditional market estimates.

Data communities vary widely, from global networks such as OpenStreetMap and Wikipedia, which engage hundreds of thousands of contributors, to localized initiatives such as CocôZap in Brazil and Dakshina in India, advocating for local issues. These communities not only supplement the data sets typically provided by the private sector but also offer diverse, real-world data that enhance AI models' robustness and applicability (Massey, Narayan, and Simperl 2024). Although open-source data are accessible, they frequently lack the economic incentives necessary for comprehensive development and may not meet specific requirements, leading to a demand for customized data solutions.

Data labeling, part of the broader business process outsourcing industry, thrives in low-income countries (LICs) and middle-income countries (MICs) because of low labor costs. *Data labeling* refers to marking or classifying data, such as identifying objects in images or transcribing speech, to enhance machine learning model performance. Data labeling typically requires an intermediate level of education and provides employment opportunities for women, youth, and disabled workers. The earnings are often comparable to, or even more than, those in traditional industries, and these jobs offer flexible work arrangements, driving rapid growth of data labelers in countries such as India, Kenya, and the Philippines.

Companies integrated into the global data production chain in LICs and MICs often enhance their employees' digital skills through targeted training programs. These programs equip employees with the necessary digital skills, covering foundational digital literacy, data processing, and even more advanced technical skills, tailored to meet the demands of outsourcing partners. Integrating local businesses and the workforce into the global AI industry enhances developing countries exposure to advanced technologies and information and communication technology sectors, potentially supporting national industry upgrades and fostering innovation.

However, the rise of GenAI poses a challenge to relatively simple tasks in the data production chain. Although demand is increasing for training data production because of the rise of GenAI, AI systems are also increasingly used to automate repetitive tasks, which complicates the outlook in this industry. These AI applications cut down on labor-intensive work. As a result, more specialized skills are needed for managing complex labeling tools and overseeing quality control. This shift underscores the need for developing countries to invest in higher-level digital skills and adapt to the evolving demands of the AI-driven economy.

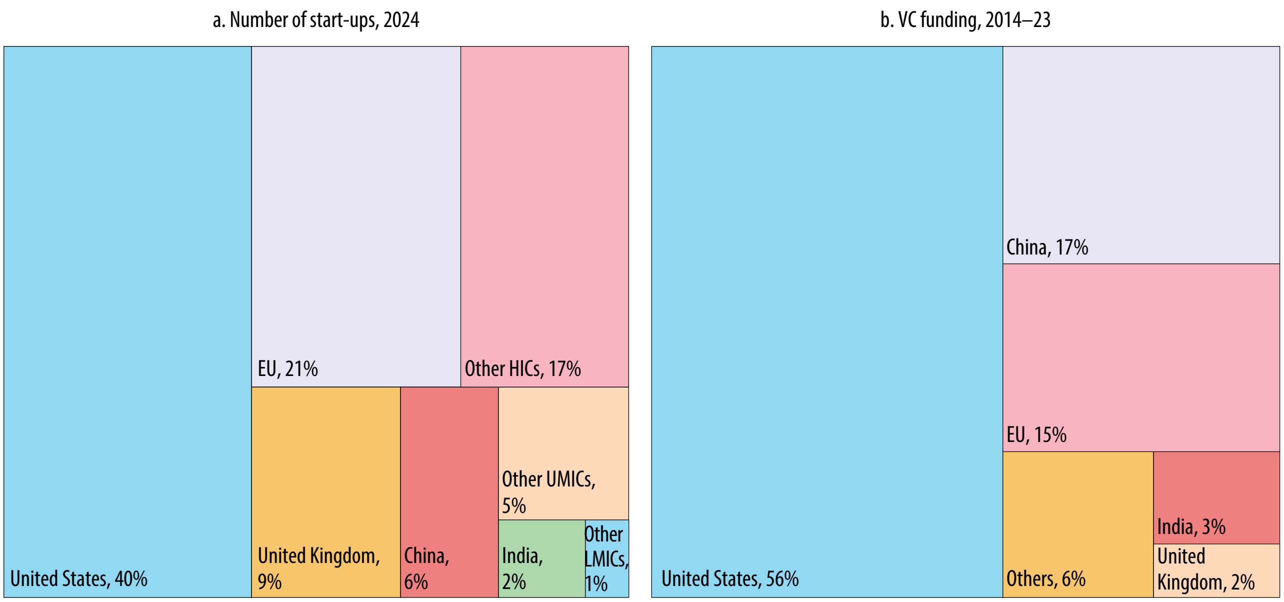
Despite the challenges in data transactions, the growing need for efficient data exchange has catalyzed the development of a distinct market. Issues such as defining property rights, monitoring usage, and verifying data quality distinguish data from traditional commodities and complicate standardization (World Bank 2021). Consequently, a standardized, publicly accessible platform for data trading is lacking, with transactions predominantly occurring bilaterally between data users and providers for specific and customized needs.

Within this evolving framework, specialized firms such as Appen and Scale AI play crucial roles by offering customized data solutions that encompass data collection, enhancement, and labeling. In addition, large technology companies such as Google enhance the efficiency and scope of AI applications by integrating outsourced services into their broader data management and AI development strategies, further illustrating the intricate dynamics of data transactions in the AI industry.

Data divide across countries and languages

Developing countries face a shortage of local data providers, reflected in their marginal share of VC investments and start-up activities in the training data industry. The bulk of VC investments in the training data industry flow to advanced economies, except for China and India (refer to figure 4.2). In 2023, the United States alone reached US$18 billion in VC investments in training data (56 percent of the global total), and the European Union and United Kingdom together attracted US$5.5 billion (17 percent). Outside of this group, China and India emerged as significant players, with investments of US$5.5 billion (17 percent) and US$1.0 billion (3 percent), respectively, with the rest of the world making up the remaining 6 percent.

FIGURE 4.2 Training AI in a data desert: VC investments in training data, by country and region



Sources: Original figures for this publication using calculations from OECD.AI (https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=vc-investments-in-data-start-ups-by-country) and CB Insights (https://www.cbinsights.com/); data extracted in September 2024.
Note: Start-ups in the training data sector were selected because the keywords *data management, data collection,* or *data tracking* were present in their company descriptions. EU = European Union; HICs = high-income countries; LMICs = lower-middle-income countries; UMICs = upper-middle-income countries; VC = venture capital.

Notably, these figures are highly disproportionate relative to their share of global gross domestic product (GDP)—especially for the United States, which accounts for roughly one-quarter of global GDP yet commands more than half of these investments. Similarly, measured by the distribution of start-ups in the training data sector, the United States leads with 40 percent, and China and India trail behind with 6 percent and 2 percent, respectively. This lack of investment not only constrains innovation but also restricts these countries' participation in the burgeoning data economy.

Global data holders, particularly social media platforms, face challenges in regions with limited internet penetration, where low digital engagement restricts the availability of valuable text and nontext data. Social media platforms such as Facebook, TikTok, and YouTube offer abundant resources—text, images, audio, videos, and location data—but poor internet coverage, slow speeds, high costs, and limited digital skills in less-developed countries result in lower computer and smartphone ownership and fewer online activities (World Bank 2024).

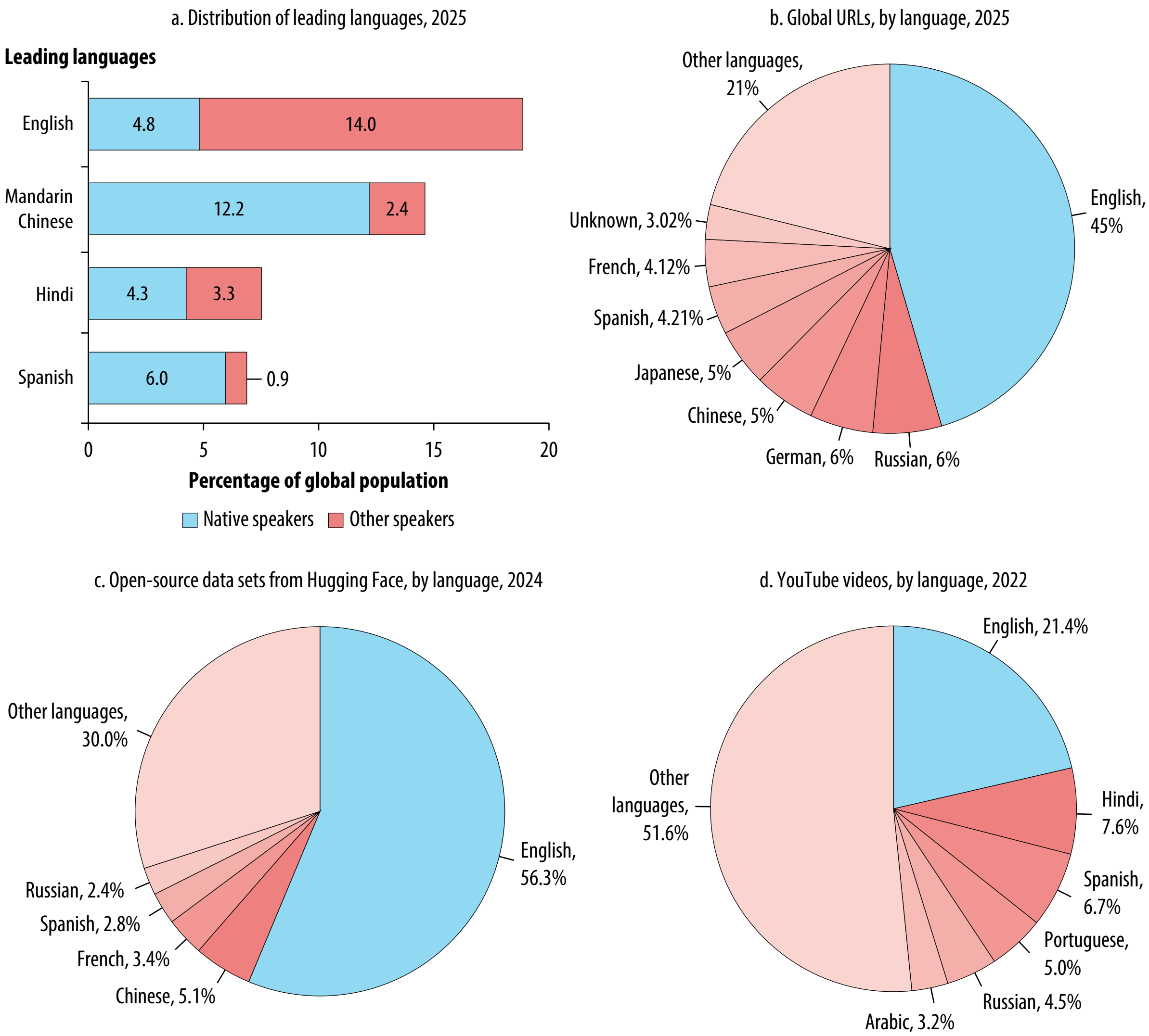
This scarcity of digital footprints hinders data collection in less-advanced economies. The challenge is even more acute in countries with small populations, where the lack of country-specific data compounds the problem. Moreover, although social media provide significant digital footprints, the data are controlled by platforms headquartered largely in the United States, complicating access for developing countries.

The global distribution of spoken languages among populations is uneven, and this natural disparity affects the development and adoption of AI models. More than 7,000 languages are spoken worldwide, with the 20 largest being the native tongues of more than 3.6 billion people, 44 percent of the global population. English, the most-spoken language, is spoken by 1.5 billion people, or 19 percent of the global population (refer to figure 4.3, panel a), and 390 million of these people consider English their first language. The second largest language is Mandarin Chinese, spoken by 15 percent of the global population, followed by Hindi (7.6 percent) and Spanish (6.9 percent).[1]

This uneven distribution is more prominent on the internet and in AI training. A handful of languages dominate most internet content, and English makes up the largest share, 45 percent of global URLs (refer to figure 4.3, panel b). The imbalance in text training data is pronounced. As of April 2024, 56 percent of open-source data sets on Hugging Face, a leading AI model repository, were in English (refer to figure 4.3, panel c). This trend is consistent even among larger data sets; of the 50 exceeding 1 trillion tokens, 28 are English-based. For high-quality text sources, English's dominant position is more pronounced because of historical, economic, and technological factors.

The United Kingdom and the United States have historically led in technological advancements. Their technological prowess, economic heft, and cultural influence have established English as the primary language of science, culture, business, and technology. English has functioned as a lingua franca, or common bridge language, that facilitates communication across diverse linguistic groups. This has resulted in a disproportionate amount of high-quality text data being generated and available in English. Nearly 98 percent of scientific papers are estimated to be written in English, even by researchers from countries in which English is not the native language (Gordin 2015).

FIGURE 4.3 Predominance of English in online content



Sources: Original figures for this publication based on calculations using Ethonologue 2025 data (https://www.ethnologue.com/); CommonCrawl data accessed in July 2025 (https://commoncrawl.org/); OECD.AI data updated on April 22, 2024 (https://oecd.ai/en/); and Figure 18 in McGrady et al. 2023.
Note: AI = artificial intelligence.

For nontext data, the language distribution tends to be more diverse, offering opportunities for non-English-speaking countries. The deepened penetration of mobile internet, increased internet speeds, and widespread use of mobile devices have fueled the fast growth of video-based social media in this decade. This trend has made it easier to produce and share video content on platforms such as TikTok and YouTube, which serve as rich sources of training data for video GenAI. Unlike text-based data and websites, the language distribution on video platforms is more diverse. For instance, only 21 percent of YouTube videos are estimated to be in English, with Hindi and Spanish accounting for 7.6 percent and 6.7 percent, respectively (refer to figure 4.3, panel d). Such diversity presents the potential to mitigate language and cultural biases in training data for nontext GenAI models.

Technology innovation, such as synthetic data, offers promising solutions to bridging data gaps in GenAI models. *Synthetic data* refers to artificially generated data that mimic the characteristics and patterns of real-world data but are created through algorithms, generative models, or simulations (Liu et al. 2024). These data can be used to supplement real-world data or to develop new training data sets altogether. Synthetic text from one large language model (LLM) can train another, potentially mitigating data depletion issues.

For instance, Amazon's multilingual language models have leveraged synthetic data to better understand and generate text across different languages, enhancing model performance where real data are scarce. This approach allows AI systems to function more equitably across various linguistic contexts, reducing the bias inherent in traditional data sets. Another example application of synthetic data is using LLM to write a math proof, with one AI generating multiple initial steps and a helper AI selecting the best one (Luo et al. 2023), where the feedback loop acts as synthetic data and refines the model.

However, this approach also has challenges. Models trained primarily on synthetic data risk "model collapse," in which performance degrades with repeated training on generated outputs (Shumailov et al. 2023). In addition, applying synthetic data in fields such as health care and education is also challenging because of the need for domain-specific knowledge. Nonetheless, ongoing research addresses these challenges to make synthetic data more applicable to AI training. For example, accumulating synthetic data alongside real data during training, rather than replacing them, has the potential to solve model collapse (Gerstgrasser et al. 2024), and physics models have been used to generate synthetic coronavirus disease 2019 lung images to incorporate domain knowledge into AI training (Zhao, Fong, and Bell 2024).

Addressing the data divide across countries and languages requires concerted efforts from the private sector, public sector, and civil societies. The private sector plays a pivotal role in bridging training data gaps by research and investment in collecting and curating data with innovative approaches. The private sector, especially leading tech companies, uses its scale, financial resources, and technological expertise to drive the development of inclusive AI data systems. These efforts, in turn, enable companies to create more inclusive AI models, expanding their reach and effectiveness in global markets.

For instance, Amazon's MASSIVE data set includes more than 1 million utterances in 51 languages, enhancing AI models' multilingual capabilities by incorporating diverse linguistic inputs. Similarly, platforms such as Facebook and Google invest heavily in multilingual data sets and partner with local content creators to enrich their language resources. In addition, firms specializing in providing high-quality training data services, such as Appen and Lionbridge, are emerging as new and promising forces in the AI training data landscape by leveraging global crowdsourcing to gather diverse data sets across various languages and contexts. The sustainability of these efforts relies on collaboration with other stakeholders, including governments and civil societies, to ensure data privacy and ethical standards.

However, market failures and externalities in the production and sharing of high-quality, localized training data arise from several intrinsic characteristics that discourage private investment. First, one special attribute of data is the ability to be reused by multiple parties without diminishing its value, thereby generating positive externalities that the investing firm cannot fully capture. Second, ambiguous property rights and difficulties in valuing data also result in underinvestment and sharing. Third, producing high-quality data involves substantial fixed costs for collection and curation, which can be prohibitively expensive—especially for SMEs in developing economies.

Fourth, although localized data offer significant societal benefits and positive spillover effects for targeted communities and beyond, their direct profitability is often limited, further disincentivizing private investment. Finally, a shortage of necessary skills and digital infrastructure—essential public goods for effective data production—creates additional barriers in many regions. Collectively, these factors highlight the urgent need for public intervention to address these market failures and stimulate investment in high-quality, diverse, localized training data.

Governments can adopt strategic measures to enhance data accessibility and governance. First, by enhancing data governance to ensure privacy and security, governments can create a trustworthy environment that supports the growth of the data sector. For example, initiatives such as the UK Data Service and Health Data Research UK provide secure access to health and social science data to foster innovation (Health Data Research UK n.d.). Second, governments can prioritize the digitization of nondigital records to unlock valuable resources. The Digital India program is another notable example that has significantly increased the accessibility of public records. Third, introducing financial support in the form of subsidies and vouchers can help SMEs access high-quality data for AI training and tools. Programs such as the European Union's Horizon 2020 illustrate how such incentives support a competitive environment for AI developers.

Fourth, fostering international collaboration between countries, and between countries and major tech firms, is crucial. Examples include the Global Partnership on AI, which involves multiple countries working together to guide the responsible adoption of AI, and partnerships between governments and tech giants such as Google and IBM to leverage advanced AI technologies for public sector innovation. In addition, measures to enhance digital infrastructure and skills, as discussed in chapters 2 and 6, are needed to complement efforts to address market failures in data production.

Civil society can contribute to enhance AI data diversity through open-source data sets and community involvement. Engaging diverse communities in AI development ensures language data with fewer biases and blind spots that can arise from homogeneous data sets. This empowers marginalized groups and incorporates their linguistic and cultural nuances into AI systems. Open-source communities such as Hugging Face play a crucial role in diversifying data sharing, a practice that should be encouraged to further enhance data accessibility. Initiatives such as Masakhane (https://www.masakhane.io/), which gathers linguistic data from African speakers, and Universal Dependencies (https://universaldependencies.org/), which offers data sets for nonstandard languages, showcase how community-driven efforts can improve language diversity in training data.

However, investing in AI training data sets for some low-resource languages may not always be financially wise or sustainable. Several factors influence the future of a language, including the size and growth of the population and economy where the language is spoken and its use in academia, business, government, and media. If a language is not widely used in these domains, it may face challenges in attracting and retaining speakers, particularly among younger generations.

In addition, the cost of developing, maintaining, and updating low-resource language data sets can be substantial. As more people learn and use English or other widely spoken languages for online communication, the demand for digital resources in low-resource languages may decline, making it difficult to justify large-scale investments in training data. If the potential benefits of low-resource language data sets do not outweigh the costs, it might be more practical to focus on languages with larger speaker populations and more established digital infrastructure.

Regions with less-available linguistic data have been relying on translation for AI applications. This option will likely be the most feasible, although it comes with caveats. For example, in India, some AI models use a two-step translation process in which user input is first translated into English and then back into the local language. Efforts such as Project Bhashini and AI4Bharat are working to enhance AI technology for Indian languages, aiming to reduce reliance on English as an intermediary and improve the quality of AI translations across diverse linguistic contexts (AI4Bharat 2025). This method leverages the robustness of English-based models to improve accuracy and consistency. However, these approaches have some drawbacks, such as the loss of cultural nuances, increased risk of errors, and additional computation costs.

Adaptation of AI models to the national context

AI models developed in a few leading economies may not address the unique contexts and needs of developing countries, despite their global reach. This mismatch stems from several factors in addition to data limitations, including economic and societal priorities, resource constraints, and divergent AI use cases. Sector-specific challenges further complicate AI adoption: Precision farming models may require significant adjustments for smallholder farmers and different crops; public service AI solutions often do not account for infrastructure gaps and informal economies; and applications in health care, education, and legal services need substantial customization to local protocols and cultural norms.

As Alami et al. (2020) noted, most AI-based health care applications are developed in high-income countries, resulting in a lack of robust adaptations for low-resource settings and making it challenging to tailor decision-making to local needs. These factors collectively highlight the need for context-specific AI development or significant adaptation of existing models to ensure their relevance and effectiveness in developing economies, where the application of AI technologies may be limited without proper customization to accommodate local conditions and constraints.

The development of AI models is characterized by two distinct approaches: open source and closed source or proprietary. For businesses and governments seeking to integrate AI into their operations, open-source and proprietary AI models have pros and cons (refer to the following sections for more information).

Open-source AI models make their underlying code or model parameters and weights accessible to the public, allowing for broad collaboration and innovation. Notable examples include DeepSeek, Meta's Llama 3.1, and Stability AI's Stable Diffusion. These open-source models empower developers to adapt and enhance AI applications for diverse needs, offering transparency, flexibility, and cost-effectiveness. In contrast, closed-source models, such as Anthropic's Claude, Google's Gemini, and OpenAI's GPT-4, keep their code and weights proprietary, offering optimized performance and integrated security for specific business applications.

Although the term *open source* implies full accessibility, in practice, the level of access can differ significantly. Open-source AI not just involves the sharing of code but can also include model architecture, parameters (or weights), training process, training data, application programming interface, and even hardware. Some projects offer complete access, including training data and model weights, and others may share the code but restrict access to essential components, such as training data or model weights, requiring users to exert significant effort to effectively use the AI (Gent 2024; Shrestha, von Krogh, and Feuerriegel 2023). In addition, open-source licenses can vary, imposing restrictions ranging from very permissive to limiting commercial use, potentially involving licensing fees (Liesenfeld and Dingemanse 2024).

Open-source models are growing rapidly and narrowing the performance gaps with proprietary models. Among the 308 notable AI models published from 2022 to June 2025, around half have open weights, and a quarter have both open weights and open-source training code. Open-source models are also increasingly challenging proprietary models in performance. This progress is driven by community-led development, readily available tools such as Hugging Face, and the increasing accessibility of powerful open-source models such as DeepSeek and Meta's Llama series. Although proprietary models still hold an edge in overall capabilities and scale, open-source models are gaining ground and offering compelling alternatives, particularly in areas such as flexibility, transparency, and cost-effectiveness.

For AI solution providers, the choice between open-source and proprietary AI models depends on various strategic, business, and technical considerations and whether they prioritize rapid adoption and community-driven innovation or control, monetization, and competitive differentiation. Open-source models offer broader adoption and a robust ecosystem but require creative monetization strategies, whereas proprietary models allow tighter control over revenue and user experience at the cost of openness.

Open-source models

For open-source models, the pros are as follows:

- *Customization.* Open-source models offer significant opportunities for adaptation; the level of customization goes beyond what is typically achievable with commercial off-the-shelf products, allowing for more targeted applications.
- *Portability.* Often vendor-agnostic, open-source models reduce the risk of vendor lock-in and remain flexible to the changing needs of businesses and governments.
- *Model transparency and risk mitigation.* Their open-source nature allows for a more comprehensive study of their limitations, such as bias, relevancy, and potential degradation. Understanding these issues is crucial for implementing AI responsibly and mitigating risks.

For open-source models, the cons are as follows:

- *High adoption barriers and development costs.* Fine-tuning open-source models for specific contexts requires substantial expertise and access to quality data. This issue can pose a challenge for organizations that may not have the necessary technical capacity or resources in house. Model adaptation also comes with high costs in development and support.
- *Security vulnerabilities.* Open-source code may be more susceptible to security exploits if it is not properly reviewed and maintained.
- *Hidden costs.* Despite being labeled *open source,* many models are not entirely free, with complex licensing requirements that can limit usage and impose costs (Stephens et al. 2024).

- *Potentially lower performance.* Open-source models generally have fewer parameters compared with limited-access and no-access models, which often correlates with a simpler architecture. Although this simplicity can make open-source models more accessible and easier to train and deploy, it also suggests that these models may offer lower performance potential compared with their parameter-richer counterparts.

Proprietary models

For proprietary models, the pros are as follows:

- *High performance.* Proprietary models have superior optimization for complex tasks. For instance, proprietary models such as Google's Gemini and OpenAI's GPT-4 have been widely adopted in sectors requiring robust performance and seamless integration with existing systems.
- *Enhanced security.* Businesses involved in finance, health care, and enterprise-level operations typically opt for proprietary models because of their superior optimization, enhanced security measures, and the ability to handle complex sensitive data securely (Ahmad 2023).
- *Comprehensive support.* Proprietary models provide access to dedicated technical support and updates, ensuring reliable performance and integration.
- *Tailored collaborations.* For governments, collaborating with big technology firms enables low-resource nations to rapidly adopt advanced AI technologies and tailor them to their national needs.

For proprietary models, the cons are as follows:

- *Reliance on external providers.* Such reliance may pose risks to national sovereignty and data security and potentially limit long-term innovation capacity and the ability to independently evolve the technology.
- *High costs.* Such collaborations require significant investment and resources to establish and maintain, which can be a considerable burden for low-resource nations.
- *Limited customization.* Customization is often restricted, and these models may not be as flexible as open-source alternatives.

Conclusion

As AI regulators, governments should promote a balanced approach and avoid unnecessary restrictions on open-source AI. The contours of open and closed systems are fluid, and society benefits from the broad array of constantly evolving systems along that spectrum. The development of more open systems—even when not completely open—is crucial to keep proprietary providers on their toes by infusing more competition and choice into digital systems.

AI regulations should focus on responsible use, promoting innovation while mitigating risks, regardless of whether AI models are open source or proprietary. Policies can encourage open-source AI development to foster wider scrutiny and collaboration. A mix of multistakeholder processes, iterative standards, and existing regulatory remedies could be used to address problems arising from open-source AI. Excess restrictions on open-source AI may slow innovation and curtail options for users.

Open-source AI models can lower the barrier to entry for developing countries, enabling them to adapt and customize AI solutions to local needs without incurring high licensing costs. To capitalize on this, governments should invest in digital infrastructure, promote AI education and skills development, and foster collaborative ecosystems that encourage local innovation and adaptation of open-source AI.

Note

1. Language statistics are based on the 27th edition of *Ethnologue*. For details, refer to https://www.ethnologue.com/.

References

Ahmad, F. 2023. "The Choice for Businesses between Open-Source and Proprietary Models to Deploy Generative AI." *Towards AI*. Last updated November 8, 2023. https://towardsai.net/p/artificial-intelligence/the-choice-for-businesses-between-open-source-and-proprietary-models-to-deploy-generative-ai.

AI4Bharat. 2025. "The IndicLLMSuite—Blueprint for Creating Pre-Training and Fine-Tuning Datasets for Indic Languages." https://github.com/AI4Bharat.

Alami, H., L. Rivard, P. Lehoux, et al. 2020. "Artificial Intelligence in Health Care: Laying the Foundation for Responsible, Sustainable, and Inclusive Innovation in Low- and Middle-Income Countries. *Global Health 16* (52). https://doi.org/10.1186/s12992-020-00584-1.

Fortune Business Insights. 2025. "AI Training Dataset Market Size, Share and Industry Analysis, by Type (Text, Audio, Image, Video, and Others), by Deployment Mode (On-Premises and Cloud), by End-Users (IT and Telecommunications, Retail and Consumer Goods, Healthcare, Automotive, BFSI, and Others), and Regional Forecast, 2025–2032." Last updated August 25, 2025. https://www.fortunebusinessinsights.com/ai-training-dataset-market-109241.

Gent, E. 2024. "The Tech Industry Can't Agree on What Open-Source AI Means. That's a Problem." *MIT Technology Review.*

Gerstgrasser, M., R. Schaeffer, A. Dey, R. Rafailov, H. Sleight, J. Hughes, T. Korbak, et al. 2024. "Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data." Preprint, last revised April 29, 2024. https://arxiv.org/abs/2404.01413.

Gordin, M. D. 2015. *Scientific Babel: How Science Was Done Before and After Global English*. Chicago: University of Chicago Press.

Health Data Research UK. n.d. "Our Strategy." https://www.hdruk.ac.uk/about-us/our-strategy/.

Liesenfeld, A., and M. Dingemanse. 2024. "Rethinking Open Source Generative AI: Open-washing and the EU AI Act." In *Proceedings of the Seventh Annual ACM Conference on Fairness, Accountability, and Transparency,* 1–14. Association for Computing Machinery, New York.

Liu, R., J. Wei, F. Liu, C. Si, Y. Zhang, J. Rao, S. Zheng, et al. 2024. "Best Practices and Lessons Learned on Synthetic Data for Language Models." Preprint, last revised August 10, 2024. https://arxiv.org/abs/2404.07503.

Luo, H., Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, et al. 2023. "WizardMath: Empowering Mathematical Reasoning for Large Language Models Via Reinforced Evol-Instruct." Preprint, last revised June 4, 2025. https://arxiv.org/abs/2308.09583.

Massey, J., V. Narayan, and E. Simperl. 2024. "How Can We Empower Data Communities in the Era of Generative AI?" Open Data Institute (blog). https://theodi.org/news-and-events/blog/how-can-we-empower-data-communities-in-the-era-of-generative-ai/.

McGrady, R., K. Zheng, R. Curran, J. Baumgartner, and E. Zuckerman. 2023. "Dialing for Videos: A Random Sample of YouTube." *Journal of Quantitative Description: Digital Media* 3.

Shrestha, Y. R., G. von Krogh, and S. Feuerriegel. 2023. "Building Open-source AI." *Nature Computational Science* 3 (11): 908–11. https://doi.org/10.1038/s43588-023-00540-0.

Shumailov, I., Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. 2023. "The Curse of Recursion: Training on Generated Data Makes Models Forget." Preprint, last revised April 14, 2024. https://arxiv.org/abs/2305.17493.

Stephens, M., M. Esposito, R. Awamleh, T. Tse, and D. Goh. 2024. "Unraveling Open Source AI." *California Review of Management* 66 (4). https://cmr.berkeley.edu/2024/06/unraveling-open-source-ai/.

World Bank. 2021. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1600-0.

World Bank. 2024. *Digital Progress and Trends Report 2023*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/40970.

Zhao, L., T. C. Fong, and M. A. L. Bell. 2024. "Detection of COVID-19 Features in Lung Ultrasound Images Using Deep Neural Networks." *Communications Medicine* 4 (1): 41.