Mapping the Role of Data Work in AI Supply Chains

ODI Research
ADVANCING TRUST IN DATA

April 2025

Contents

About

This report was researched and produced by the Open Data Institute (ODI), and published in April 2025. It was created by Joe Massey, Sophia Worth and Elena Simperl. If you want to share feedback by email or want to get in touch, contact the research team at research@theodi.org.

We would like to thank the participants in this research project for dedicating their time and expertise.

Executive summary

While AI may often seem to exist abstractly in cyberspace, it is also driving major supply chains of both materials and labour. In this report, we outline the landscape of work behind AI data, which is often invisible to decision-makers and the public. The invisibility of AI data work is a major challenge, since it limits the emergence of debate and setting of best practices that support model developers, researchers, innovators, policymakers and others to play their role in responsible and effective AI data supply chains and AI systems. We aim to outline data's role in AI supply chains, and how its use is evolving within the changing AI landscape.

AI data work enhances and contextualises the datasets used in machine learning. For example, it can be used to moderate whether data scraped from the internet contains harmful images or language, or to prepare datasets of images that accurately represent the presence or absence of a certain disease. In this report, we introduce the importance of data work within the AI lifecycle, including its impacts downstream (Section 1). We then share an accessible review of the landscape of AI data work, and all the components it entails in a 'taxonomy' (Section 2). Finally, we build from interviews with five organisations that demonstrate the cutting edge of different aspects of AI data work, to develop three key reflections on how the landscape of AI data work is evolving, and the likely implications of these changes (Section 3). In this final section, we discuss how:

1. The nature of AI data work is highly dependent on the application domain of the technology. For example, to ensure responsible and safe AI systems, higher-stakes AI application domains will often require bespoke datasets adapted to the application, and built from highly specific domain knowledge among data workers and those deploying the data work. High costs for data work generate different trade-offs between efficiency and quality, depending on the application.

2. A rapidly changing landscape of data work is evolving in parallel to AI innovation, and is changing demands on the data that drives it. This includes the integration of AI tools into the work of data workers, as well as the changing type of human knowledge in demand for integration into AI datasets. There are increasingly challenging questions about what constitutes high-quality data work in an ever-evolving environment.

3. While many organisations are opaque about their AI data practices, some openly share demographic data about data workers to help others to better interpret the model and its outputs. This transparency is helping to set best practices in their respective fields of AI development in terms of developing high-quality, contextually relevant datasets and more interpretable model outputs.

Data work has been the subject of much concern due to significant issues around labour rights. In this report, we take a broader view of data work. We hope this helps readers to understand the importance and evolution of data work in the rapidly developing landscape of AI. Through a better understanding of AI data work, we hope to provide the building blocks to develop best practices in this space, leading to standards including new benchmarks, evaluations and toolkits.

Section 1: Introduction

Chinese AI company Deepseek's launch of the language model R1 in January 2025 sent shockwaves around the world. The stock market even slumped for several days.[1] Deepseek made impressive claims about R1's capabilities and the techniques used to train the model.[2] AI labs have continued to attempt to out-do each other, with incredible speed and expenditure, launching new models of increased capability on a regular basis. While much of the focus and hype around AI has been based on these models, there has been a movement towards focusing on the data that enables AI innovation and impacts the outcomes of AI systems. At the ODI, we have focused a programme of work on data-centric AI. There is now a lot of evidence that without data, there is no AI.[3] Data is crucial to every part of the AI lifecycle,[4] including developing, deploying and monitoring AI systems.

But this does not mean that AI models can be trained with any data. For data to be AI-ready, it should meet basic levels of quality and cleanliness, be saved in optimal file formats for compression and usage, have mature and detailed metadata, and be served with the best infrastructure possible for discoverability and usability.[5] Aspects of data, such as representativeness, validity and reliability, are key to ensure effective and responsible model applications.

While some machine learning techniques involve algorithms identifying patterns across a large unstructured dataset, others require further information or context to be added to a dataset. For example, supervised learning involves an algorithm learning to map input data to correct output labels by training on a labelled dataset, gradually improving its ability to make accurate predictions by minimising the difference between its predictions and the known, pre-labelled correct answers.[6] Fine-tuning in machine learning is the process of adapting a pre-trained model for specific tasks or use cases, using specifically curated and labelled data.[7] Additionally, reinforcement learning with human feedback involves humans providing feedback to the outputs of a model, enabling it to make decisions

[1] BBC (2025). US tech stocks steady after DeepSeek AI app shock.
[2] Azhar, A. (2025). DeepSeek: everything you need to know right now. doi: 10.1145/3025453.302583
[3] Hardinges, J. & Simperl, E. (2025). Open Data Institute. How the data for AI ecosystem is evolving.
[4] Ibid.
[5] Koesten, L. et al. (2017). The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. doi: 10.1145/3025453.302583
[6] Google Cloud (n.d.). What is Supervised Learning?
[7] IBM (n.d.). What is fine-tuning?

that maximise rewards (or benefits).[8] The context provided to these datasets involves the addition of human knowledge, perception and expertise to enhance a datapoint.

There is significant human involvement behind the data used to train the full variety of AI models. This workforce is often dispersed around the world, working on a variety of tasks to enhance AI models,[9] including labelling pedestrians in videos for automated driving algorithms, labelling photos of celebrities, and editing chunks of text to 'ensure language models like ChatGPT don't churn out gibberish'. The need for labelled data is expanding rapidly: the global market for this type of data work was valued at $2bn in 2022 and is forecast to grow to $17bn by 2030.[10] Countries like China have recognised the importance of data work, and expect the industry to grow by more than 20% a year.[11] Leading data annotation platform Scale.AI raised $1bn against a $13.8bn valuation in a recent funding round, demonstrating the immense value of labelled data to the future of AI.[12]

1.1 Importance and challenges of AI data work

Data work and data labelling happens in many different ways. Sometimes the developers and their teammates label data for an AI model, while in other cases, people are hired specifically to label data. Inspired by a definition by James Muldoon et al. (2024),[13] we consider AI data work to be the active human labour that enhances machine learning datasets, with diverse tasks related to preparing and evaluating the data and model outputs. This can include labelling images for a computer vision model, content moderation for social media algorithms, or tagging pages on Wikipedia to improve the site's bots.

Many similar concepts cover different types of work relating to data and AI. The concept of AI data work we use here goes beyond the definition of data work set out by Miceli and Posada,[14] by focusing specifically on data work that contributes

[8] Hugging Face (2022). Illustrating Reinforcement Learning from Human Feedback (RLHF).
[9] Tan, R. & Cabato, R. (2023). Washington Post. Behind the AI boom, an army of overseas workers in 'digital sweatshops'.
[10] Rowe, N. (2023). Wired. Millions of Workers Are Training AI Models for Pennies.
[11] Werner, J. (2025). Babl. China Unveils Comprehensive Plan to Boost Data Labeling Industry Growth.
[12] Sawers, P. (2024). TechCrunch. Data-labeling startup Scale AI raises $1B as valuation doubles to $13.8B.
[13] Muldoon, J. et al. (2024). A typology of artificial intelligence data work. doi: 10.1177/20539517241232632
[14] Miceli, M. & Posada, J. (2022). The Data-Production *Dispositif*. doi: 10.1145/3555561

to AI systems. Microwork[15], a concept coined by Jeff Bezos when setting up Amazon Mechanical Turk refers to a series of small tasks completed independently by different workers. These tasks can pertain to AI models, but are not explicitly focused on AI data tasks. Ghost work[16] – 'the human labour powering many mobile phone apps, websites and artificial intelligence systems [that] can be hard to see' – is broader than AI data work, and includes areas such as content moderation. Crowdsourcing and cloudwork are even broader still, and usually refer to any work that is mediated via a digital labour platform that can be performed remotely by workers in a different location from the task provider.[17]

For Li et al. (2023),[18] the process of generating data is seen as labour, and defined as activities that produce digital records useful for capital generation. Given the role that data scraped from the internet[19] has played in developing this new wave of generative AI models, scholars argue that the effort put into commenting on Reddit, uploading YouTube videos, or posting pictures to Facebook, constitutes labour, as these assets are used as training data for the development of AI models. Other passively-produced data, such as location data, traffic patterns, and private preference information, actively plays a role in the improvement of advertising models, navigation algorithms, and commercial recommender systems. However, they are not commonly seen as digital labour.[20] For the purpose of this paper, we will not cover this definition of data as work.

As Muldoon et al. (2023) point out, such work is often 'outsourced to low-paid and marginalised workers' – the source of human rights issues discussed further in this section and in wider literature. The urgent need for human labour in the AI supply chain has created a booming market, with data work providers competing to reduce costs of labelling datasets for AI labs, academics, and others. However, the demand for cheaper, on-demand labour in the gig economy, with work typically outsourced to Global South workers,[21] has led to significant exploitation.[22] Workers have few rights and operate in harmful,

---

[15] Muldoon, J. & Apostolidis, P. (2023). 'Neither work nor leisure': Motivations of microworkers in the United Kingdom on three digital platforms. doi: 10.1177/14614448231183942
[16] Gray, M. L. & Suri, S. (2019). Ghost Work. ISBN: 9781328566249.
[17] Howcroft, D., & Bergvall-Kåreborn, B. (2018). A Typology of Crowdwork Platforms. doi: 10.1177/0950017018760136
[18] Li, H. et al. (2023). The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. Available at: https://doi.org/10.1145/3593013.3594070
[19] Tiedrich, L. (2024). OECD. The AI data scraping challenge: How can we proceed responsibly?
[20] Massey, J. et al. (2024). Open Data Institute & Aapti Institute. From co-generated data to generative AI: New rights and governance models in digital ecosystems.
[21] Williams, G. (2025). Data Enrichment Work and AI Labor in Latin America and the Caribbean.
[22] Rowe, N. (2023). Wired. Millions of Workers Are Training AI Models for Pennies.

intimidating working conditions for low pay.[23] It is important to understand the role of data work, to address both the rights of the people carrying it out, and its impact on AI systems and their downstream impacts. This report does not focus specifically on these rights issues emerging from AI data work, but looks at the wider evolving landscape of data work, to emphasise its importance in the changing impact of AI.

Beyond serious concerns about the impact of incentives to create high volumes of AI data at low cost, further research demonstrates the importance of data work for the outputs of AI systems, in terms of both their safety and their effectiveness for the task at hand.[24] Much research has discussed the role of data work in introducing subjective elements such as political perspectives, biases and inconsistencies, since 'the tasks that data workers perform are fundamentally about making sense of data'.[25] Still, the responsibilities for this lie with those making decisions about data work,[26] not with the data workers themselves – and there is a need to generate incentives and recognition for appropriate data work, rather than purely a focus on models.

Further, we often know very little about the data used to train specific AI models. This exacerbates issues related to human rights and wellbeing, making workers in the data supply chain 'invisible'.[27] Despite a push from the developer community to progress the conversation on transparency[28] through Dataset Cards, Dataset Nutrition Labels, CrowdWorkSheets and open metadata standards like Croissant, there is still much work to do. In building our AI Data Transparency Index,[29] only five of the 22 models surveyed had high levels of data transparency. One metric of the Index focuses on the human and organisational supply chain of the model, showing that we know very little about the different people involved in building AI models. However, a significant number of people involved in the AI supply chain carry out tasks that computer models find hard to replicate,[30] such as filtering, moderating and labelling data.[31]

---

[23] Privacy International. (2024). Humans in the AI loop: the data labelers behind some of the most powerful LLMs' training datasets.
[24] Sheng, V. S. et al. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. doi: 10.1145/1401890.1401965
[25] Miceli, M. et al. (2022). Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? doi: 10.1145/3492853
[26] Ibid.
[27] Ekbia, H., & Nardi, B. (2014). Heteromation and its (dis)contents: The invisible division of labor between humans and machines. doi: 10.5210/fm.v19i6.5331
[28] Diaz, M. et al. (2022). CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. doi: 10.48550/arXiv.2206.08931
[29] Snaith, B. et al. (2024). Open Data Institute. The AI Data Transparency Index. doi: 10.61557/LPXQ8240
[30] Aapti Institute (2020). AI Data Labelling: Roundtable Readback.
[31] Estampa (2023). Cartography of generative AI.

1.2 What this research achieves

This research explores the work that goes on behind the scenes involving humans to make AI models work. This project builds on existing scholarship to describe a taxonomy of AI data work across different types of data, stakeholders, work and more, to demonstrate the breadth of data work around the world. We then explore upcoming trends in AI data work, complemented by five case studies that demonstrate the practicalities of data work in an ever-moving AI landscape.

This research is focused on creating a better understanding of the realities of AI data work around the world. We hope to generate an understanding of the relationship between concerns about data work and the models themselves, as well as demonstrating examples of best practice. We aim to motivate making the supply chains of models used on a daily basis transparent and clear. This research is for anyone utilising AI models in their day-to-day lives, for policymakers making decisions about AI, and for organisations thinking about implementing AI into their workflows.

We reviewed academic and grey literature, such as podcasts and interviews with AI developers working in this space. For each case study, we conducted a series of semi-structured expert interviews.

Section 2: Taxonomy of AI data work

There has been an expanding field of research into the role that people play throughout the AI supply chain, yet AI developers often obscure this role from view. This section sets out an overview of the different aspects of AI data work, from the tasks that AI data workers do, to the way they are employed. We have built on existing work taxonomising data work in this space,[32] while taking a broader view to give policymakers, researchers and technologists a wider understanding of the diversity of AI data work, across tasks, geographies and impact on AI models. This broad understanding across the supply chain is important for developing a clearer image of best practices.

As discussed in more depth in Section 3, the landscape of AI data work is changing rapidly as AI evolves, so we will seek to continue to iterate on this taxonomy in line with changes. As such, we welcome feedback and comments on this taxonomy, including anything we have missed and how such a taxonomy could be more useful.

2.1 What tasks do AI data workers do?

The concept of AI data work includes many different activities across the lifecycle of AI models. Data workers are engaged to add human context to a data point, and the action of adding this context can happen in a variety of ways. Here we describe the different tasks that AI data workers do, across different types of data.

Most data worker tasks involve tagging, labelling or annotating pieces of data. This activity, also called classifying or attributing, differs depending on the requirements of those developing the AI models, and generates labels to help machine learning algorithms understand and classify the information they process. This activity takes different forms depending on the data type.

- For images, this could mean object detection, where a box is drawn around a certain object in an image and annotated with its name, or image segmentation, where each pixel in an image is annotated to show whether it is part of a certain object.

[32] Muldoon, J. et al. (2024). A typology of artificial intelligence data work. doi: 10.1177/20539517241232632

- For video, just as with images, this can involve object detection (in 2D or 3D), where a labeller draws a box around the edges of an object and annotates it – for example, outlining a car for an autonomous vehicle model. It could involve mapping out the edge of an object in a video using a polygon technique. A similar technique involves annotating key points on an object – for example, to understand how a body moves during sport. Finally, object tracking builds upon object detection by connecting detections across multiple frames of a video into a single track.

- For text, this can involve tagging specific parts of a sentence – for example, selecting the verbs and nouns – or named entity recognition, which classifies text into predefined categories and connects them, as well as classifying whether the sentiment of a sentence is positive or negative.[33] For generative AI, annotating dialogue is crucial,[34] including labelling utterances and intent, as well as classifying errors in generated text or providing preferences between text objects.

- For audio, this can involve labelling particular genres of music, annotating particular musical notes, or detecting specific speakers or sounds within an audio file – for example, a type of bird call. It could also involve adding written translations to spoken word in different languages.

- For multimodal labelling, this can involve linking objects between mediums – for example, attributing a voice to a particular speaker in a text transcript, writing a text caption for an image or video, or linking information about a car from a document to the same car in a video.

AI data workers can be engaged throughout the AI lifecycle, and across many different ML techniques and methods. The input of AI data workers does not necessarily happen linearly through a supply chain, but is iterative, with validation and feedback loops between stages. The stages of the lifecycle where AI data work happens includes:

- **Data preparation and model training**. Many of the tasks described above happen before training the model. AI data workers add context to datasets through annotation and labelling. This can happen as part of supervised learning techniques, where models directly learn the context added through human labelling, or fine-tuning, where smaller, more specific datasets are used to adapt pre-trained models to particular domains or use cases. A big part of this activity is content moderation, where labelling harmful or hateful data points enables AI models to identify and remove future harmful images.

[33] IBM (n.d.). What is sentiment analysis?
[34] Labelforce (2024). Annotating Dialogue Data for NLP – Enabling Conversational AI Advancements.

- **Evaluation and validation** aim to ensure that a machine learning model can make accurate predictions on new data by checking the validity or accuracy of an output. This process often uses benchmarks, which usually include labelled data. However, it can also involve human evaluation of model outputs, identification of edge cases, and model limitations. In pre-training, data workers verify the accuracy of labels or annotations that other labellers have created. Evaluation is also an important part of the ML pipeline. Data workers can support reinforcement learning with human feedback (RLHF),[35] by providing feedback to outputs of models. For example, data workers rank a series of outputs to the same prompt based on certain features, to teach the model what a good output looks like.

- **AI safety** focuses on preventing accidents, misuse, or other harmful consequences arising from AI. Humans play a key role in this process, through red-teaming and adversarial techniques. While data workers themselves might not conduct red-teaming, the process raises areas in which data workers are required to contribute.[36] For example, if certain types of prompts are flagged as producing problematic outputs, RLHF techniques can involve humans in labelling problematic outputs for removal. Adversarial techniques involve attempting to force an AI model into a negative output, and curating datasets that can support training the model to produce better outcomes.[37]

2.2 How are AI data workers engaged or employed?

Crowdsourced or gig-work models use a group of workers to complete tasks. Examples of these platforms include Amazon Mechanical Turk and Sama. These workers are usually paid by the task or per hour, and complete tasks for many different organisations. This method enables model developers to quickly access large amounts of labellers, making it an efficient way to label datasets. Still, in some cases, incentives to drive down data costs can make this working format a source of worker exploitation. Further research suggests the quality of labels can be limited, and crowd work is therefore more suited to less complex tasks,[38] such as labelling basic objects in images.

[35] Hugging Face (2022). Illustrating Reinforcement Learning from Human Feedback (RLHF).
[36] IBM (n.d.) What is red teaming for generative AI?
[37] Parrish, A. et al. (2023). Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models. doi: 10.48550/arXiv.2305.14384
[38] Muldoon, J. et al. (2024). A typology of artificial intelligence data work. doi: 10.1177/20539517241232632

In-house contracts involve companies hiring data workers to annotate and label data in-house, to enable control of the full pipeline of data labelling and annotation. This can be beneficial where the data being labelled is sensitive.[39] It is worth noting that some data science roles will label or annotate data on projects themselves, rather than using external workers. Additionally, AI developers may ask colleagues to test AI systems informally.

Third party organisations, such as Scale.ai or Prolific, hire data workers on a contractual or project basis. This approach enables these organisations to have specific teams focused on specific types of tasks, or with particular technical or domain expertise. These approaches seek to provide a higher standard of accuracy on the labels.

Voluntary data work refers to data work carried out on a voluntary basis. Examples include people labelling images of flora and fauna on iNaturalist, or people donating their voices to the Common Voice project, which seeks to curate datasets of underresourced languages. On Wikipedia, unpaid editors support the quality of AI tools like the Objective Revision Evaluation Service (ORES), which evaluates article quality and detects vandalism. Human contributors support these systems by tagging revisions, reporting vandalism, and providing feedback on flagged content. This work is often based around particular groups, who altruistically dedicate their time to tagging data towards a cause they believe in.

Un-recognised labour refers to data work conducted without the workers' awareness. For example, reCaptcha involves people selecting particular objects in tiles to prove that they are human. This process of classification, while not clearly signposted, is used as part of a data annotation pipeline for Google.[40]

2.3 Who are data workers and where does data work happen?

Data work happens in many ways. In some cases, it may be the team developing the AI models who label a dataset for a specific purpose; in others, it is professional data labellers, or highly skilled individuals with domain knowledge. Since the act of labelling data is not always clearly described by AI developers, we do not necessarily know where this work is happening, nor who exactly is doing it.

---

[39] Tonkin (2020). Keynote: Privacy and Security in Real-World Data Annotation. doi: 10.1109/PerComWorkshops48775.2020.9156138

[40] O'Malley, J. (2018). Captcha if you can: how you've been training AI for years without realising it.

As discussed above, for some AI data workers, this can be a full-time job. Others start out in task-based work as a part-time job to get some extra cash, which can frequently morph into full-time jobs or a primary source of income, despite volatile working conditions. In some cases, particularly in places with high inflation, access to foreign currency is a strong benefit of working part-time.[41]

The demographic data on AI data work is patchy due to a lack of openly available information on both data work platforms and AI model development processes. Existing data focuses predominantly on crowd work research. Much of the existing research[42] focuses on Amazon Mechanical Turk, where around 80% of workers are based in the US and India[43]. More recent research[44], involving 11,000 Appen workers, found that the workers in most countries were male (>60%), between 18 and 34 years of age, and generally well-educated. Still, much AI data work also happens outside of these large organisations, with smaller contractors and often with more specialist labellers.

Much research suggests that a lot of data work happens in the Global South,[45] due to lower costs and labour rights, with the requesters coming from companies based in the Global North. There appear to be particular data work hotspots across the Global South,[46] including Kenya, Uganda, Venezuela, Colombia, India, the Philippines, and even refugee camps in Lebanon.[47]

However, we know that data work is far from limited to the Global South. Platforms like Appen have more than a million data workers, based in more than 170 countries, suggesting a global market of AI data work. Prolific states that its participant pool live in OECD countries (excluding Turkey, Lithuania, Colombia and Costa Rica), Croatia and South Africa. For many of the large-scale voluntary efforts to label data, like iNaturalist, contributors are predominantly based in the US and Europe.[48]

Following China's recent AI boom, there are comprehensive plans to support this growth by boosting the country's data-labelling industry.[49] This plan has led

[41] Tubaro et al. (2025) The digital labour of artificial intelligence in Latin America: a comparison of Argentina, Brazil, and Venezuela. doi: 10.1080/14747731.2025.2465171
[42] Difallah et al. (2018). Demographics and Dynamics of Mechanical Turk Workers.
[43] Posch et al. (2021). Characterizing the Global Crowd Workforce:A Cross-Country Comparison of CrowdworkerDemographics. doi: 10.1145/3159652.3159661
[44] Posch et al. (2021). Characterizing the Global Crowd Workforce:A Cross-Country Comparison of CrowdworkerDemographics. doi: 10.1145/3159652.3159661
[45] Ibid.
[46] Estampa (2023). Cartography of generative AI.
[47] Rowe, N. (2023). Wired. Millions of Workers Are Training AI Models for Pennies.
[48] Zhang et al. (2023). Geographic proximity and homophily effects drive social interactions within VGI communities: an example of iNaturalist. doi: 10.1080/17538947.2023.2297948
[49] Werner (2025). China Unveils Comprehensive Plan to Boost Data Labeling Industry Growth.

to a scenario in which 90% of the people working in AI in China are employed as data annotators.[50]

2.4 What sort of skills do AI data workers need?

AI data work encompasses a dynamic and multifaceted range of tasks that require diverse skills and levels of expertise, the evolving nature of which is discussed further in Section 3. As data work requirements depend on the nature of the model and application, so too do skills requirements.

Low-complexity tasks, such as labelling images of common objects or content moderation, can be performed by workers with minimal specialised training, who are easily replaceable. In contrast, high-complexity tasks, like annotating medical imaging scans or evaluating intricate scientific data, require deep domain-specific expertise. These specialised roles demand professionals with advanced knowledge in medicine, scientific research, or other technical domains, making the workers significantly harder to substitute. Significantly more people qualify for the low complexity tasks (and require less investment in training), compared to the high complexity ones, who are more expensive to hire.

The skills required for effective AI data work extend beyond traditional technical competencies, and encompass linguistic, cultural, and domain-specific capabilities. For some models, language proficiency is crucial, particularly in enabling multilingual large language models (LLMs) through accurate translations and nuanced understanding. This is particularly important for cases of low-resource languages,[51] for which there is currently limited data. Cultural knowledge plays a significant role in tasks like content moderation, where understanding contextual and national variations in what constitutes harmful content is essential. Domain-specific expertise, such as that possessed by medical professionals, legal experts or scientific researchers, allows for precise and sophisticated data annotation that can dramatically improve machine learning model accuracy and reliability. To solicit appropriate data work, AI developers need a level of understanding of the domain and the skills that workers need to complete the task, as well as expertise in defining the tasks they require for labelling.

[50] Weilan, Z. & Mengyang, M. (2024). Data annotation a promising profession, with a talent gap of nearly 30 million: industry insider.
[51] Cohere for AI (2024). The AI Language Gap.

2.5 What infrastructure supports AI data work?

AI data workers operate amongst a suite of tooling and infrastructure to enable them to complete data work tasks as effectively as possible. This includes organisational infrastructure like platforms for labelling tasks, as well as specific tools supporting particular tasks such as quality assurance, moderation, data curation and exploration, and labelling.

Data workers can perform specific tasks like labelling on specialised software. Upon login, labellers are presented with instructions, the data they need to label, and various tasks. These platforms, such as those created by LabelBox, Encord and Argilla, are developed in collaboration with potential data labellers, to ensure they best meet their needs. They are also adaptable for different purposes – different tools are needed to label images and text. For example, when labelling medical images with Encord, users can flip between multiple scans for a patient, adjust the windowing settings to alter the brightness, and also work in dark mode, a preferred setting for radiologists.

These platforms offer AI tools that support data labellers to do specific tasks. In current AI labelling workflows, most will include some form of automation, to maximise the time spent by humans on labelling tasks. AI tools can help with:[52]

- segmenting the dataset, to select data of a certain type for labelling, or to prioritise certain data points for labelling, to maximise the value of human effort. For example, 3D Slicer in the healthcare space.
- suggesting annotations or labels that human reviewers can verify and correct. For example, MONAI Label.
- building on rough guidance from human labellers to complete labels. For example, finding the exact edge of an object in an image, such as those included in CVAT.

[52] Larosa, C. (2022). A Primer on Data Labeling Approaches To Building Real-World Machine Learning Applications.

Section 3: Discussion – The evolving role of AI data work

Data is the cornerstone of AI systems.[53] As new and innovative AI models continue to emerge, the demands on data evolve rapidly alongside the rest of the AI landscape. Where we previously required human labour for simple classification tasks, now automated tasks such as auto-labelling have become increasingly common aspects of data work. This means data workers are more often being asked to validate AI-generated suggestions, use AI themselves to increase productivity in labelling roles, or test AI products[54] as a means to identify harmful prompts and outcomes. Further, as more organisations adopt AI in their products and services, there is more requirement for general purpose AI models to be fine-tuned with internal or context-specific data, which requires more specialised labelling teams.

Having outlined the landscape, we now further discuss three key reflections on the present and evolving state of AI data work. These have emerged from both our desk research and interviews with companies at different stages of the data supply chain and who are part of emerging changes to data work practices. We share five case studies drawn from our interviews with five organisations or project groups – Project MONAI, the LOLA team at Paderborn University, Encord, Prolific and Cohere – that have helped us to establish these reflections. We are conscious that this does not represent the full diversity of AI data work, but together, they offer an interesting set of perspectives to understand recent changes to the landscape.

3.1 Dependency of data work on the application domain

Much of the public understanding of AI data work fits the image of gig work, with workers around the world completing generic tasks via third-party platforms. However, both the nature of the AI application and the technology itself will significantly vary the nature of data work required. The demand for model accuracy, as well as elements such as privacy constraints, sensitivity or

[53] Hardinges, J. & Simperl, E. (2025). Open Data Institute. How the data for AI ecosystem is evolving.
[54] Parrish, A. et al. (2023). Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models. doi: 10.48550/arXiv.2305.14384

context specificity of the data used, dictate which types of AI data work are procured.

There is a significant difference in the data required to develop LLMs, which learn from huge, varied datasets, than that used in smaller, niche models with specific purposes. 'High stakes' AI model applications, such as medical diagnosis, monitoring for environmental disaster, and suicide prevention, will often require more bespoke datasets.[55] Data work will require more particular knowledge of the domain or application setting – and often high levels of accuracy, due to acute negative consequences if model outcomes fail.

This expertise can be particularly costly – for example, in health applications where trained medical experts are required to label medical images. The costs associated with highly-skilled and bespoke data labelling also create demand for more efficient approaches to data work, which maximise the skills and time of high-cost labellers, and increase the accuracy of labels.

Case study 1: MONAI Label – generating labelled medical datasets using a deep learning framework

Project MONAI is an 'open-source PyTorch-based framework for building, training and optimising AI workflows in healthcare'. It was launched in 2019, through a partnership of academic, clinical and private organisations. Project MONAI includes the initiative 'MONAI Label', as well as 'MONAI core' and 'MONAI deploy'.

MONAI Label offers a pipeline for integrating deep learning models into the development of labels for medical images of various types like MRI and CT scans. It provides automated and interactive functions that reduce the manual effort required for labelling. MONAI Label's automated labelling model continually learns from human labels as they are added. This reduces the amount of time required for AI data workers. Where an initial dataset needs to be meticulously labelled by hand by a group of experts, workers can then correct larger sets labels, rather than adding new ones from scratch. The tool integrates with other medical labelling tools, such as 3D Slicer, providing medical AI data labellers with the tools to provide accurate and efficient labels.

Due to high costs, many model developers will adopt existing datasets and will not generate their own (labelled) data, or will do so only for key elements such as fine-tuning or validating models. However, where others have generated the datasets,

[55] Sambasivan et al. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. doi: 10.1145/3411764.3445518

like in the case of the multilingual model LOLA (Case Study 2), data work forms a key part of considerations about data quality.[56]

Issues associated with data labelling can cause major issues, such as high project costs – for example, where eventual failures in model performance can make it expensive to re-label datasets.[57] Organisations have to estimate and balance the quality of the data work when selecting data and considering the initial cost of training. Synthetically-labelled datasets will typically require more intensive training activities due to their inefficiency during training[58] – and organisations therefore often prefer to use human-labelled data. Still, there are clear trade-offs. As we saw in the case of Project MONAI (Case study 1), the use of hybrid data labelling creates the capacity to generate bespoke datasets at more reasonable cost, and to avoid issues associated with data that is not contextually relevant.

Case study 2: LOLA – creating a multilingual model from pre-existing datasets

LOLA is an open-source, multilingual LLM developed by the DICE Research Group at Paderborn University and published earlier this year.[59] Designed to process more than 160 languages, LOLA employs a sparse Mixture-of-Experts (MoE) Transformer architecture, similar to that used by the now renowned DeepSeek models,[60] although multilingual.

The model was trained on the CulturaX dataset,[61] which is produced from raw text documents in 167 languages and totaling more than six trillion tokens. The model performs natural language generation and can understand tasks across multiple languages, and can be fine-tuned to provide functions such as translation. While the CulturaX dataset that LOLA relies upon is not labelled, as the model uses an unsupervised learning technique, the developers needed to use labelled datasets for a range of other functions. For example, given that LOLA is a pre-trained model that functions within a relatively niche field, there is a need for bespoke, labelled datasets to further support fine-tuning the model for specific tasks where it currently underperforms. They further involve language native speakers and language experts within their own organisation for light-touch human evaluation.

---

[56] Government Data Quality Hub (2021). Hidden costs of poor data quality.
[57] Sambasivan et al. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. doi: 10.1145/3411764.3445518
[58] Matta, S et al., (2024). Investigating Cost-Efficiency of LLM-Generated Training Data for Conversational Semantic Frame Analysis. doi: 10.48550/arXiv.2410.06550
[59] Srirastava, N. et al. (2025). LOLA – An Open-Source Massively Multilingual Large Language Model.
[60] Ng, K. et al. (2025). DeepSeek: The Chinese AI app that has the world talking.
[61] Nguyen, T. (2023). CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. doi: 10.48550/arXiv.2309.09400

The challenge of employing appropriate data work will often require significant expertise by those designing and deploying AI models, including significant domain expertise. This should include clarity of the impact of the data labelling approach, including the employment conditions of those hired to complete the work.[62] We amplify others' recommendations for computer science curricula to support this by demonstrating the practical aspects of using real world data for real world applications and 'designing data collection, training raters or creating labelling task designs'[63] – as they often do not. There are major responsibilities in ensuring the role of high-quality and contextually-appropriate datasets in achieving positive outcomes when models are deployed in practice, which are often ignored due to incentives to focus on model-centric metrics of accuracy.

3.2 AI innovation influences the tools supporting AI data work

Much discourse has focused on access to large quantities of data for training AI models, and the demand for new data continues to grow. However, as demonstrated by the recent interest in small language models, the demand for smaller-scale, high-quality, domain-specific datasets[64] is also growing. These datasets enable AI models to learn particular things, such as how to respond to prompts, or how to identify a particular type of vertebrae for diagnosis of medical issues in the spine.[65]

The requirement for high-quality data adapted to diverse specifications, and efficiency in generating this data, as discussed in Section 3.1, has led to an emergence of tools to support data labelling. As we saw in Case study 1, this includes offering labelling 'pipelines', like MONAI Label, that can support the generation of large, bespoke datasets, even for high-stakes environments. Further, whole platforms, like Encord (see Case study 3), provide bespoke tooling for a huge variety of different labelling tasks. Meanwhile, Cohere (Case study 5) developed a bespoke data annotation platform for its Aya model, as a deliberate step in collecting the right type of data and labels for the platform to train a multi-lingual LLM.

---

[62] Jindal, S. (2024). Protecting AI's Essential Workers: A Pathway to Responsible Data Enrichment Practices.
[63] Sambasivan et al. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. doi: 10.1145/3411764.3445518
[64] Pike (2024). LLMs Aren't Just "Trained On the Internet" Anymore
[65] Zhang et al. (2023). Artificial Intelligence in Scoliosis: Current Applications and Future Directions. doi: 10.3390/jcm12237382

Case study 3: Encord – providing the technical tools needed for tailored multimodal data work

Encord is a platform aiming to 'solve data alignment challenges by building a complete data-focused AI infrastructure that's focused on data management and curation, annotation and workforce management tooling, and model evaluation and observability'. Encord has collated and created a range of AI tools to improve the capabilities and efficiencies of teams developing AI models. These tools function across a wide range of data types, including images, videos, documents, and audio files. The platform is focused on the user, both in helping model developers to select the most effective approach to AI data work, and in providing the tools to support annotation tasks.

For those who need data labelled or quality assured, Encord enables the design of labelling workflows and ontologies based on what is of interest in each data file, as well as AI automation tools to segment and prioritise data within a larger dataset. These tools aim to give AI developers the flexibility to manage their labelling projects as effectively and efficiently as possible. AI developers can then collaborate with data workers to annotate the data, depending on their requirements. For the AI data workers, the Encord platform provides tools used by the data workers to make the labelling process easier. This includes tools for automatic object segmentation (to simplify the process of drawing the outline of an object or tracking it through multiple frames), viewing multiple files for concurrent annotation, and to use models to support the annotation process directly from the labelling platform.

Beyond the demand for access to new data, there is an emerging desire to realise the value of existing data through workers adding additional context. For example, beyond classifying a given image as a 'pedestrian', there are many other things about the image that humans understand.[66] The human eye can interpret millions of different data points pera second, innately understanding relationships between objects, meanings of facial expressions, and much more. Therefore, further insight can be attributed to data beyond a high-level classification, or drawing the boundaries of a box around an object. Another route to increase the supply of data, which NVIDIA is exploring,[67] is synthetically-generated data, with labels generated as part of the generation process, to support the training of AI models.

[66] Bouchard, L. F. (2024). The Role of Data in Advancing AI: Insights from Expert Jerome Pasquero [Video].
[67] Guo, P. (2024). Addressing Medical Imaging Limitations with Synthetic Data Generation.

Case study 4: Prolific – working with people, not data, to produce high-quality datasets

Prolific is a company that offers services for developing or acquiring 'human-derived data'. It was founded for academic researchers to build a participant pool for survey research, and now includes more than 200,000 participants. It offers specific services for development of datasets for AI applications across diverse sectors well beyond academia.

In recent years, Prolific has increasingly addressed a need to access expertly labelled data at scale. Beyond delivering generic data tasks, it has expanded the expert pool to include workers of different types, including those qualified with PhDs, mathematicians, lawyers and more. These participants can complete more diverse and niche tasks, which require compensation in line with the hourly rate within their professions. Prolific focuses specifically on the human element of making the most of the data behind AI systems. It aims to develop a platform that is easy for those seeking data labels and the workers to interact with, and a process for ensuring the quality of the data work remains high, as well as assessing whether participant experience is sufficient to complete a task.

Still, with more complex, subjective and high stakes data tasks, there is arguably even more weight added to the question of what constitutes 'high-quality data'. AI data work is often extremely contextual and subjective – for example, what might count as a funny sentence to one data labeller may sound sarcastic to another. While bodies of work explore approaches for averaging the labels of different annotators on the same data point,[68] these remain limited in their ability to address such challenges and there are real concerns about the consequences of skewed or limited datasets. Given the international spread of AI data workers, this can lead to questions of bias in the outputs of AI models as a result of the varied culture of different labellers.[69]

[68] Cook, O. (2025). Efficient Annotator Reliability Assessment and Sample Weighting for Knowledge-Based Misinformation Detection on Social Media. doi: 10.48550/arXiv.2410.14515
[69] Mazeika, M. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. doi: 10.48550/arXiv.2502.08640

3.3 Transparency about who is doing data work is emerging

Given the difficulty of knowing which datasets have been used to train AI models,[70] it is unsurprising that it is difficult to understand the full supply chains and role of workers in AI models. Even when organisations are transparent about the data work behind their models, they will often offer limited information to help others to trace and understand the nature of the data workers behind this work. As such, a distinction must be made between transparency about the process of labelling data and the people doing that work.

For example, in the GPT-4 technical paper from OpenAI,[71] the authors describe how AI data workers ranked outputs of the LLM from best to worst as part of a reinforcement learning with human feedback process. Though they mention that they follow industry best practice when hiring AI data workers, they often give no further information about the people completing the tasks. As such, the role of AI data workers is discussed specifically through the lens of the ML methodology, but with no information about the AI data workers themselves. A May 2024 edition of the Foundation Model Transparency Index[72] based at Stanford highlighted that of the 14 major LLMs developers surveyed, only one (Stability AI) shared information about the wages of data labourers. The Stanford authors highlight that this creates low accountability for the exploitation of these workers.

When organisations are required to make assumptions about the nature of data work, this can lead to serious issues down the line, as discussed in 3.1. Transparency is needed to ensure data quality and data excellence[73] at all stages of the pipeline – for example, ensuring that workers hold sufficient domain expertise to produce sufficiently high-quality data, or that there has been appropriate representation of different native language speakers. The most recent draft of the EU's AI Act Code of Practice offers a range of transparency requirements for documenting AI data practices to downstream providers, but does not specify the need to identify the humans involved. This is a missed opportunity to ensure that AI decision-makers are accountable to all those involved in developing the technologies. Data supply chains should be considered across the frameworks, toolkits, regulatory guidance, and other

[70] Snaith et al. (2024). The AI Data Transparency Index. doi: 10.61557/LPXQ8240
[71] OpenAI (2023). GPT-4 Technical Report. doi: 10.48550/arXiv.2303.08774
[72] Bommasani (2024). Foundation Model Transparency Index. doi: 10.48550/arXiv.2407.12929
[73] Aroyo, L. (2020). Data Excellence: Better Data for Better AI. doi: 10.48550/arXiv.2111.10391

forms of standard practice we set for responsible AI – to ensure that decision-makers consider the human and labour rights associated, as well as the downstream impacts of data work.

While descriptions of the process of AI data work is an important step in AI data transparency, particularly in understanding how a model functions, other organisations are deliberately providing information about who is labelling data. For example, the Aya model, developed by Cohere (Case study 5), includes significant data about the demographics[74] of contributors to the model. This transparency supports a broader understanding of some of the limitations of this approach, where some languages had significantly smaller numbers of contributors than others. There are other ways to increase transparency in this space. Deepmind collaborated with the Partnership on AI to publish a case study[75] about its application of PAI's Responsible Data Enrichment Practices, exploring the steps it took to employ AI data workers responsibly.

Case study 5: Cohere – tackling the need for human-annotated multilingual language datasets

The Aya model,[76] developed by Cohere, is an instruction fine-tuned open-access multilingual language model that covers 101 languages (of which 50% are considered low resource). It was developed with the support of more than 3,000 independent researchers across 119 countries, contributing more than 204,000 original human annotations. The model is open access, meaning all of the data and code, as well as data about the annotators involved in the project, is available for anyone to use online.[77]

To curate such a multilingual dataset, the Cohere team required a bespoke data annotation tool to collect the data they needed. The Aya annotation platform was designed as simply and as low tech as possible, to make it as accessible as possible to all potential contributors and to reduce the opportunity for bias. As a first step, the basic platform enabled people to enter prompts and responses in their native languages. They could then make edits to prompts and responses put forward by other people. Finally, it asked people to rank which of two similar responses they preferred. Each task provided a signal about how good the data is and how much of the data required editing, while the pairwise comparison gave an indication of the reliability of edits from different editors.

[74] Singh, S. (2024). Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning.
[75] Jindal, S. (2022). Protecting AI's Essential Workers: A Pathway to Responsible Data Enrichment Practices.
[76] Üstün, A. (2024). Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. doi: 10.48550/arXiv.2402.07827
[77] Singh, A. (2024). Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning.

A significant part of working on a multilingual LLM is attracting the right people to support data labelling and annotation. For Cohere, it was critical to attract fluent speakers in several languages, to ensure that the data and annotations were organic and representative of the speakers' culture. However, this presents a logistical challenge in attracting participants to take part, offering incentives, and maintaining quality control of data across different speakers. To support these challenges, the team developed the UI described above, as well as curating a gamification element to pit different languages against each other in a race to contribute as much as possible. Ultimately, the main incentive for those taking part was knowing that their contribution would directly support a tool that they could ultimately use in their day-to-day lives.

Section 4: Concluding remarks

Due to the opacity surrounding AI data work within closed platforms, there is limited research into its impact on both AI systems and workers. Our mapping of the different types of AI data work in this report builds on existing research in this space, and represents a first step toward explaining this complex ecosystem that spans multiple geographies and stakeholders with competing interests.

However, this research can only go so far. We cannot paint a complete picture of the AI supply chain while different stakeholders continue to obscure their data work practices. Confidentiality is to be expected in certain parts of research and development, but the widespread lack of transparency about AI data work prevents a comprehensive understanding of AI supply chains, effectively excluding considerations about AI data work from conversations about responsible AI, despite these playing a key part in the development of AI systems.

This gap manifests practically in the absence of practical tools, benchmarks and guidance for labelling data responsibly. For established model developers looking for labelling services, or new entrants into the market, there is a lack of support to implement responsible data labelling. Critically, this also blurs accountability lines, meaning that procurers of data labelling may not have all of the relevant information about working conditions, fair compensation, and wellbeing of individuals throughout the data supply chain. Without greater transparency, we cannot properly integrate AI data work into frameworks and toolkits for responsible AI development, leaving a critical blind spot in our understanding of AI innovation and its broader impacts.

We hope to be part of further work that builds clearer best practices for data work, outlining the roles in achieving this across the AI supply chain for developers, researchers, civil society, AI data suppliers, innovators providing data work services, multilaterals, the media and beyond.

Section 5: Endnotes

Methodology

This project ran from January to March 2025, and builds on our work from the data-centric AI and participatory data programmes. This work was inspired by our Humans in the Loop panel session at the Data Empowerment Fund event in November 2024, which explored the lived experiences of panelist Kauna Malgwi, who chairs the African Content Moderators Union, the work of Soujanya Sridharan with the Aapti Institute and Karya.in, and that of Laurenz Sachenbacher with the data workers inquiry. The work builds on their contributions, and those of others in this space.

We undertook a mixed methodology approach for this research work, combining desk research, in the form of a literature review, with expert interviews. The desk research covered outputs from across academia and civil society, as well as technology and data companies. We additionally explored other mediums, such as podcasts and interviews, to curate a broad overview of the sector. This approach gave us a holistic view of AI data work, taking in perspectives from the AI data workers themselves, the organisations procuring this work, and those providing the tools to make it happen.

The 45 minute semi-structured interviews focused on AI data work practices, decision-making process around AI data work, impact and quality assessment of AI data work, and perspectives on the future of AI data work. We would like to thank the experts who took the time to speak to us as part of this research:

- Jennifer Ding, Encord
- Andres Diaz-Pinto, NVIDIA (MONAI)
- Madeline Smith, Cohere
- Marzieh Fadaee, Cohere
- Nikit Srivastava, University of Paderborn (LOLA)
- Adam Winning, Prolific
- Jenny Huzell, Prolific

Limitations

At the ODI, we want to create a world where data works for everyone. Diversity, equity and inclusion is an important part of how we do research, and we strive to ensure that all of our research adheres to these principles. In practice, this means ensuring that we seek to engage with a diversity of perspectives, and take an inclusive approach to research. We may not always achieve these aims, but we strive to improve on this work in every project.

Since this research project was conducted over a short time frame, we were limited in the methods and data sources we could use. As such we were only able to consume a portion of the literature in this space, and spoke to only five organisations in the AI data supply chain, and these were based in the Global North. This in part reflects the limited timeframes for this project, as well as how data work is changing, and some specific cases, but it will not reflect the wider landscape that is better represented in the taxonomy section (from our literature review). Further research could differentiate different types of data work based on the types of companies, objectives, geographies and more.

We are acutely aware of the challenges facing some AI data workers, including poor working conditions, low pay, and long-term mental health issues. Given the short time frame for this work, we opted to focus on AI supply chains as a whole, attempting to understand the full picture of AI data work, as we did not feel able to create the right environment to work with data workers themselves within the time frame. We hope this research provides a springboard for additional research in this space, focusing on the AI data workers themselves, as well as better understanding what meaningful transparency looks like for AI data work. We would love to see, and be involved in, further future research addressing these issues.