Democratising access to data: Bridging the data divide with generative AI models

ODI Research
ADVANCING TRUST IN DATA

June 2024

Contents

About

This report was researched and produced by the Open Data Institute and Microsoft, and published in June 2024. It was written by James Maddison, Joe Massey, Neil Majithia, Elena Simperl and Sonia Cooper (Microsoft). If you want to share feedback by email or want to get in touch, contact the ODI Research team at research@theodi.org.

Introduction

As technology continues to advance, disparities in access to data and information persist, creating a digital divide[1] that hinders social progress and economic development. This paper focuses on that divide, which concerns access to data, rather than access to technology.[2] The Ada Lovelace Institute defines the data divide as the 'gap between those who have access to – and feel they have agency and control over – data-driven technologies, and those who do not'.[3]

Data is a critical part of society, supporting us to tackle some of the biggest challenges we face, from the climate crisis to the Covid-19 pandemic. Yet because data is not always universally available or accessible, the value of the outcomes we can derive from it are not fully realised. Concerns are emerging about a 'data winter',[4] in which there is less accessible data due to a reduced interest and effort to open it up. This affects the type of products and services created, it impacts decision-making, and ultimately has a negative impact on society. It is particularly limiting to developments involving Artificial Intelligence (AI), which require large amounts of data to both develop and use AI systems. Ensuring that everyone can access and use data effectively is critical for empowering people to make better decisions and create equitable outcomes for society.[5]

Data exists on a spectrum from closed, to shared, to open.[6] Open data is data that anyone can access, use or share.[7] By making a dataset as open as possible – while ensuring people are protected from any potentially harmful impacts – more value can be gained, as this creates more opportunities for data to be used in innovative ways. For example, open data can improve how we access healthcare services, discover cures for diseases more efficiently, understand our governments better, or travel to places more easily.[8] Despite huge increases in the amount of data being collected, data is not always openly accessible to everyone, which can limit its benefits. As open data does not exist on its own, data holders and publishers need to share data openly.[9] There is also publicly available data which includes content published online for anyone to view and analyse, including blogs, images and publications.

---

[1] Van Dijk, Jan A.G.M. (2019). The Digital Divide.
[2] Sudan, R., Hammer, C. & Eferin, Y. (2023). Toward Bridging the Data Divide.
[3] Ada Lovelace Institute. (2021). The data divide.
[4] Verhulst, S. (2024). Are we entering a "Data Winter"?
[5] Walker, J. et al. (2024). The promise and challenge of data discovery with LLMs.
[6] Open Data Institute (2020). The Data Spectrum.
[7] Open Data Institute (n.d.). Open data.
[8] Ibid
[9] Carrara, W. et al., European Data Portal (2018). Open Data Goldbook for Data Managers and Data Holders: Practical guidebook for organisations wanting to publish Open Data.

Microsoft has been working to close the data divide since 2020,[10] focusing on increasing the publication and use of data. Over this time, the ODI has partnered with Microsoft to support its efforts to help close the data divide by supporting the opening up and sharing of more data, to unlock value, share expertise and make data more useful for all. We've collaborated on a variety of projects connected to bridging the data divide, including convening organisations that steward open data to help others address a range of social, environmental and economic challenges,[11] and supporting effective open data publishing for social good through the creation of a toolkit.[12] In the fourth year of our partnership with Microsoft, we are investigating how new conversational generative AI tools can help close the data divide.

Closing the data divide is of particular importance to the world of work. In the 21st century, every organisation is a data organisation and should be thinking about how it uses data, as well as its role in wider data ecosystems.[13] Research from 2022 found that 70% of employees are expected to work heavily with data by 2025, an increase from 40% in 2018[14]. Working with data is becoming a crucial part of the majority of jobs of today and in the future.

For those required to work with data, the development of generative AI models presents an opportunity to enable better use of data. As well as potentially facilitating easier access to data, these models can help users with data comprehension, analysis, visualisation and decision-making, which makes them particularly useful for individuals who may lack technical expertise. Since people interact with many generative AI models via prompting, a written command by the user, interactions are more conversational and less technical than other tools. However, the feasibility and limitations of applying these tools to the workplace are not yet widely understood. There are well-documented drawbacks of these models due to their early stage of development, from hallucinations[15] to a lack of accuracy.[16]

[10] Microsoft (2020). Closing the data divide: the need for open data.
[11] Maddison, J., ODI (2022). Reflections from our second peer-learning network with Microsoft.
[12] Maddison, J. & Parke, S., ODI (2023). User-centric data publishing.
[13] Freeguard, G. et al., ODI (2022). Data literacy and the UK government.
[14] Marr, B. (2022). The Importance Of Data Literacy And Data Storytelling.
[15] IBM (n.d.). What are AI hallucinations?
[16] Miller, K., Stanford (2023). Generative Search Engines: Beware the Facade of Trustworthiness.

This report seeks to explore the potential of generative AI in bridging the data divide, particularly in regards to the publishing and use of data, including open data. While this research identified much excitement around the potential of generative AI tools to close the data divide, there are drawbacks to many of the tools explored here, especially for users with limited data skills.

Generating code and analysing data are two of the areas currently prompting the most excitement and potential, likely due to the utility of the tools for developers. This research found that working with generative AI alters the relationship of the person with the data task, as generative AI has different qualities than traditional tools. We found that these tools work best with a human in the loop. Human expertise is still required to maximise the benefits of these tools, while minimising their drawbacks. Going forward, people will need to adapt and learn to work more effectively with generative AI tools to realise the full potential of the tools, for example, by learning prompting strategies.

In the first section, we set the scene for this research, defining the link between generative AI and data. In the second, we look at four key data publishing and use activities, and explore how generative AI tools can support those working on these activities, the most feasible uses of generative AI for these activities, and the gaps in the current offerings. The conclusion draws together insights from across the data publishing and use activities, and sets out the promise and drawbacks of using generative AI tools.

Generative AI

AI and data are intrinsically linked – without data there is no AI, and access to large amounts of data is crucial for the development of AI. Data guides every stage of an AI system, from conception to operation.[17] Data provides the information that a machine learning (ML) model is trained on and learns from. It is used to test and benchmark a model's success, and data is inputted for utilisation once a model is operational. With generative AI, synthetic data can also be an output.

There are many different definitions of AI. In this research, we use the OECD definition, which defines an AI system as a 'machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.[18]

Generative AI has grown in popularity in recent years, and refers to a category of AI-based algorithms that generate new outputs based on insights and patterns identified during training.[19] Generative AI systems can create new data in the form of images, text, audio and more. These systems are developed on top of large scale AI models like Open AI's GPT4, which power the generative AI applications. Some examples of generative AI applications include ChatGPT, which produces text responses to human prompts; DALL-E, which produces images; and Voicebox, which produces audio responses. In this research, we have focused on generative AI tools, meaning applications built on generative AI models. These tools may come in different forms: some have bespoke user interfaces that enable people to interact with them directly; some are embedded into existing tools; and others come in the form of application programming interfaces (APIs – sets of code-based instructions that enable different software components to communicate and transfer data) or software packages that need to be downloaded via a device's operating system.

Large Language Models (LLMs) are a subset of AI models that can power AI applications that reply to a user's text-based prompt with a text-based output. Under the hood, LLMs are sequential processes with billions of parameters trained on vast amounts of data to understand user-provided prompts and infer the best outputs in response.

[17] Snaith, B., ODI (2023). What do we mean by "without data, there is no AI"?.
[18] OECD (n.d.). OECD AI Principles overview.
[19] Routley, N., World Economic Forum (2023). What is generative AI? An AI explains.

The process involves breaking down a prompt to the words within them, recognising the words' meanings and positional context, processing them with a trained neural network to find appropriate responses, and then translating these inferences to natural language. It's important to recognise that when responding to a prompt, LLMs are not searching their training data for an answer; rather, they are running the prompt through a process that predicts the most likely output in response to the input based on the patterns and concepts it has learnt from the training data. LLMs therefore do not function as a database, they do not store the data on which they have been trained.

Characteristics of generative AI

The generative AI tools explored in this report are all underpinned by generative AI models like OpenAI's GPT-4 model. Standard models are trained on enormous datasets of text collected from the entire internet and covering huge amounts of information, which greatly expands the size of the model's knowledge base, but also means they are subject to the internet's multitude of propensities and biases.[20] The knowledge of standard models will also be constrained by the particular point in time at which the model was trained. Once they have been trained on an enormous amount of data, they no longer use this data and differ from search engines in how they access up-to-date data, or private data that a user might want to access. Therefore, these standard models are desirable for reasoning tasks rather than search. If a user or organisation wants a generative AI tool to help them with tasks across their field of work, a general model may not be optimised to do those tasks. However, a standard model can be enhanced with the use of fine-tuning via Retrieval Augmented Generation (RAG, also referred to as 'grounding'), or search functionality.

- A standard, pre-trained model is **fine-tuned** by retraining it on a custom-made knowledge base.[21] This updates the parameters of the model to fit the new training data, changing the way the model responds to prompts. For example, one could take a standard LLM like ChatGPT 3.5 and fine-tune it with a curated collection of legal documents so that its responses to prompts contain more legal knowledge and language.

[20] IBM (n.d.). Shedding light on AI bias with real world examples.
[21] Ferrer, J. (2024). An Introductory Guide to Fine-Tuning LLMs.

- **Retrieval Augmented Generation (RAG)** is a method by which an LLM's responses are enriched with sources of information that are provided to it as context to the user's prompt, rather than just its pretrained weights that have been learnt from the initial training dataset.[22] Unlike standard LLMs, one that has RAG capability on a specific dataset can answer questions about that dataset with more accuracy and clear citation of its source of information. In practice, this works by adding an extra step to the prompt-response architecture: an 'embedding agent' receives the user prompt, understands the meaning of the words within it, then queries the back-end dataset with those words to find relevant chunks of data that the LLM can utilise to respond to the user with accuracy from the dataset. The result is a valuable tool for people who need LLM functionality to work on private data or statistical/numerical datasets.

- **Search functionality** works on a similar architecture and enables LLMs to base their responses to prompts on internet searches. Here, a 'search agent' takes a user's prompt and performs an internet search using the content within it, providing information it finds from websites to the LLM as context for the user's question. The model then responds to this with an answer based on the search results.[23] With a 'search agent', LLM responses are as up-to-date as possible and, similar to RAG, can provide citations for provenance.

[22] AWS (n.d.) What Is RAG?.
[23] Ribas, J. (2023). Building the New Bing.

Exploring the potential of generative AI tools to support data publishers and users

This section will explore the following activities for data publishers and users, and will detail the activity (and sub activities), current methodologies and tools for achieving those activities. We will explore the availability of generative AI tools, and their added value, as well as offering insights and analysis of the trends of AI tools for each activity. The activities covered in this report are:

- **Collecting and accessing data,** including searching for datasets, evaluating datasets and understanding datasets.
- **Preparing data,** including cleaning data, merging datasets and labelling datasets.
- **Publishing data and improving its accessibility,** including publishing as a file on a website, portal or API, generating accessible language, and identifying accessibility issues.
- **Analysing data and generating visualisations,** including summary statistics, creating insights and generating graphs.

Collecting and accessing data

Collecting and accessing data involves searching for data and understanding the types of data included in the dataset, as well as evaluating, mapping and prioritising datasets. Traditionally, users would search for data using a search engine, via a data portal, or by asking someone. To make sense of the datasets, users are often able to view a preview of the data on a website. Alternatively, they may download the dataset to conduct some initial analysis and querying. It is a very hands-on, manual process.

Using generative AI tools to search for, collect and understand data is a new area of interest. While conversational generative AI chat agents were not specifically designed for search, there is excitement around the opportunity for them to support users in finding and understanding datasets when they are connected to the internet and integrated into search. Earlier versions of conversational generative AI tools were not connected to the internet, but it has become more common for them to be integrated into search; for example, Copilot. There are also tools that support API users to find new datasets, such as HTTPie. These are usually free to use, although some such as ChatGPT-4 charge a monthly fee.

Exploratory research conducted by King's College London and the ODI[24] found that conversational generative AI chat agents (such as Bard and ChatGPT), which were not connected to the internet at the time of study, were able to suggest relevant datasets and support some of the sense-making activities required to understand a given dataset. However, given these tools are not connected to the internet and rely upon the knowledge and patterns learned from datasets they have been trained on, they were not very accurate, often suggesting made-up datasets, and were unable to cite sources. As conversational chat agents' responses to prompts are only inferences, their answers are probabilistic predictions. These inferences rely on many factors, such as the variety of the datasets, the quantity of the dataset, information contained in the dataset, and the design of the architecture under the hood, as they are unable to access additional information.

For data discovery and explainability, connecting a generative AI model to a dataset or the internet has much better results. By connecting them to the internet, tools such as ChatGPT 4.0, Google Gemini and Microsoft Copilot are able to perform the more traditional role of a search engine, making them more successful at finding datasets and providing links to the sources of information they share. However some studies have found that some of the common drawbacks of these chat agents, such as 'hallucinations' and 'inaccuracies' remain; they are still liable to suggest fictional datasets, perform inaccurate analysis and can name wrong sources (if they name any at all).[25]

Training generative AI tools on datasets held by each organisation, via retrieval augmented generation, is likely the way forward in this space. These tools enable users within an organisation to access and explore their organisation's data through a conversational chat interface, facilitating easy data discovery and enabling the democratisation of data throughout the organisation (regardless of data skills).[26] Research backs this up, with one study showing that LLMs can be quite effective in retrieving information from a specific dataset or database.[27] Similarly, they may be able to straddle the line between predictive models and information retrieval systems with their ability to answer questions for proprietary databases.[28] Another recent trend is that of 'copilots' built into tools – for example Excel Copilot enables users to query data in Excel using natural language. There are also bespoke tools, like Wobby which enables users to query public and private datasets, as well as analyse them, via a conversational interface.

[24] Walker, J. et al. (2023). Prompting Datasets: Data Discovery with Conversational Agents.
[25] Miller, K., Stanford (2023). Generative Search Engines: Beware the Facade of Trustworthiness.
[26] Barendse, C. (2023). How Generative AI Will Revolutionize Data Catalogs.
[27] Zhou, X. et al. (2023). LLM As DBA.
[28] Castro Fernandez, R. et al. (2023). How Large Language Models Will Disrupt Data Management.

The feature of multimodality – that these tools can output images, code, tables and text summaries – holds promise for improving general data usability and explainability. Users can query the contents of datasets, and view summaries as graphs and tables. The conversational generative AI tools also provide supplemental information about datasets, such as reminding users to check the terms of the licences associated with the use of the data. But users need to approach conversational dataset search less in the anticipation that they will receive a definitive answer, and more in the expectation of receiving support to find and use an appropriate dataset, which may happen using a different online tool. Similarly, given the potential of conversational generative AI to produce inaccurate information, they can be useful as a starting point for an expert in the field, as long as all results are double-checked through traditional means to ensure accuracy. There is, however, real potential for these tools when trained on specific datasets, which would realise some of the benefits explored above.

Preparing data

Data preparation is the process of gathering, combining, structuring and organising raw data into a format that is suitable for analysis or other uses. Studies indicate that commercial organisations spend as much as 80% of their time preparing datasets for analysis.[29] Within an organisation, the job of preparing data often falls to someone who is also responsible for analysing data, such as a data analyst or data scientist, and can often be a painstaking manual process.

There are already a range of traditional AI tools and platforms that are designed to support data practitioners with the data preparation steps outlined above. Companies like DataRobot and Tamr offer paid services to support various data preparation activities, while others like Trifacta offer a freemium model. OpenRefine is an open source alternative that allows users to clean and transform data for free. Most of these tools are designed to be integrated into existing organisational data systems.

Generative AI has been theorised to help bridge the skills gap required for preparing data in a number of ways, including data cleansing, data transformation and enhancement, and data validation.[30] The general shift towards the adoption of generative AI for supporting data preparation activities seems to be around the integration of generative AI capabilities into existing data management and analytics platforms, such as Copilot for Excel.

[29] Warudkar, H. (2019). AI For Data Cleaning: How AI can Clean Your Data and Save Your Man Hours and Money.
[30] Banerjee, D. (2023). Unleashing Generative AI for Advanced Data Quality Checks: Use Cases and Impact.

Data transformation and enhancement seems to be one of the more promising areas; several commercial generative AI-enabled tools have been developed to support the activity, such as Ydata. These tools offer users an alternative to manual data mapping and transformation, allowing users to more quickly define the relationships between data elements from different sources. Generative AI-enabled tools can also help users to automate data transformation tasks, such as establishing consistent data formats, structures and conversions, without much need for human intervention.

Generative AI tools can support data cleaning activities, by providing conversational interfaces for existing AI-enabled data management and analytics platforms. Generative AI tools can identify patterns and deviations in large datasets and pinpoint errors, in a more time efficient manner than could be achieved through a manual approach. In theory, these tools can also help to limit human error in the data cleansing process. Some researchers have also used Generative Adversarial Networks (GANs), which are more commonly used to develop synthetic datasets, to detect unusual patterns or outliers within a dataset,[31] though the application of this approach is fairly novel and has been observed in a real world context in this research.

Generative AI models can generate synthetic data based on patterns and relationships learned from actual data. Synthetic data has numerous applications. These include creating new data for machine-learning models and creating datasets that preserve the privacy of the data subjects when that data is tested and analysed.[32] Tools like Synthpop, and Faker.js allow users to create synthetic datasets with characteristics similar to real-world data. As these tools come in the form of programmes, they are likely inaccessible to non-technical users, though generalist generative AI chat agents are also able to generate data. However, there are concerns around the ability of synthetic data to protect the privacy of those in the original dataset.[33] In addition, using synthetic data to train models may lead to model collapse, meaning 'models begin to lose information about the less common but still important aspects of the data, producing less diverse outputs'.[34]

[31] Anello, E. (2022). GANomaly Paper Review: Semi-Supervised Anomaly Detection via Adversarial Training.
[32] Stafford, G.A. (2023). Unlocking the Potential of Generative AI for Synthetic Data Generation.
[33] Mustać, T. (2023). Synthetic Data: The Good, the Bad and the Unsorted.
[34] Shadbolt, N. (2023). Why we need to fear the risk of AI model collapse.

Publishing and improving the accessibility of data

Data can be published online either as a file on a website, on a portal, or via an API. Traditionally, users will need to manually upload a file to a website or portal, as well as manually coding an API (this could be done via a third party, such as a website developer, for those without technical skills).

We found there to be limited scope for supporting non-technical data publishers to publish data using AI tools. When publishing data on a website or uploading it to a portal, generative AI chat agents may be able to support new data publishers in walking them through steps to publish data online, and answer questions along the way (as opposed to traditional methods of following a guide). Chat agents may also be able to support exploration and identification of websites where users can publish data. However, our research did not identify any bespoke tools to support publishing data in these ways.

One of the main ways which people make data available is via APIs. These exist as code, written in the programming language of preference of the creator, and usually require some technical coding expertise to create. A specific type of generative AI tool can generate the code needed for an API known as code generators. One of the most popular is GitHub's Copilot, which has been adopted by more than 1.3 million developers and 50,000 organisations. Due to the need for some coding skills, alongside an understanding of data infrastructure, setting up an API is somewhat limited to those with higher-level technical skills.

Tools like RapidAPI say that they 'help users to design and build APIs' via an intuitive user interface that simplifies development workflows. They also enable users to share their APIs in a repository for others to access, expanding the number of potential data users. However, they do not appear to support the development of APIs for those with limited technical skill sets. Other tools, such as APIMatic, support publishers to make their APIs as accessible as possible to other users by enabling integration across different platforms, regardless of coding language used to design the API. On balance, it seems that the development of APIs will remain a task for skilled developers, due to the need to combine the right components for an integrated experience.[35]

[35] Asthana, A. (2023). Generative AI and the impact on APIs and software development.

It is important to ensure that the data being published is both findable and accessible to users with a variety of different needs. The internet is still not accessible to everyone: a recent study found that only 3.7% of webpages[36] were compliant with the Web Content Accessibility Guidelines (WCAG). Enabling these users to access the internet and datasets published online is an important part of bridging the data divide.

There are currently a number of efforts to empower data publishers to publish data in a more accessible fashion.[37] This research found a limited number of bespoke AI tools available to data publishers focused on improving accessibility of datasets, beyond the good practice that we see as part of preparing data for publishing (explored above). Tools like Equally AI and AccessiBle test websites for accessibility purposes and support data publishers to ensure that data listed on their website is findable and accessible to all users. Generative AI can also automatically scan websites and identify accessibility issues, such as poor colour contrast, missing alt text or missing ARIA labels.[38]

A large number of tools that primarily serve other purposes can also help to make data more accessible. For example, tools such as ElevenLabs Generative Voice AI can translate text to speech (and vice versa) providing a service for people who are deaf or hard-of-hearing or who have speech disabilities. Others, such as AltText.AI and ClickUp, can tag images with information about what they show, which can support the development of alt-text. The pricing structures for these tools vary: often they are free to use on a single use basis, but with pricing structures for business-level usage. Others, like OpenAI's Whisper AI, are completely open source, meaning that anyone with coding experience can use the model.

For those with coding experience, Userway has developed a variety of tools to support developers to improve the accessibility of their code, website and data. The generative AI tool FixMyCode, for example, is focused specifically on improving the accessibility of code to be more inclusive, usable, and conformant with WCAG 2.1 AA criteria.

Finally, there are tools that help data publishers in a more general way. For example, Microsoft has a suite of accessible tools, including SwiftKey, which learns about the user's writing style, meaning it make it a supportive tool to help people type; tools for users with neurodiversity; and the Google Live Transcribe app, which captures speech and directly translates it into text.

[36] WebAim (2024). The WebAIM Million.
[37] Bureau of Internet Accessibility (2023). Will Generative AI Improve Digital Accessibility?
[38] ibid

But how useful are these tools for improving accessibility? In an ethnographic study of generative AI for researchers with disabilities, researchers found a mixed picture.[39] While generative AI chat agents were able to provide on-demand support for their accessibility needs, in low-stakes, easily verifiable contexts, they often also 'parroted back' information without completing the task itself.[40] This study shows that the AI tools used in the research (ChatGPT, Dall-e, etc) could provide advice on good accessibility practices, but when asked to review specific content for accessibility issues, could not identify specific issues mentioned in the advice. As with the development of APIs explored above, these generalist tools currently provide advice and information, but are not yet at a stage where they can directly implement actions consistently. While there is promise, there is still a long way to go in this space to provide direct support to those without the relevant skill sets.

Analysing data and generating visualisations

Data analysis is the process of working with data to create useful information to inform decision-making. It involves various techniques and methods to extract insights from raw data, uncover patterns, trends and relationships, and ultimately enable users to make informed decisions. Traditionally, data analysis is conducted by a skilled data professional, such as a data analyst or data scientist, using statistical methods and tools like Excel, SQL, and dedicated statistical software such as R or Python with libraries like pandas and NumPy. This process can be fairly time-consuming and requires organisations to employ skilled data professionals to deliver it. At the same time, many people conduct more basic data analysis as part of their day-to-day work.

For many of the tasks covered in this research, the ability to code is a significant benefit to achieving the activities, and it is particularly important for data analysis. Coding is not yet a widespread skill, with only around 3% of jobs in the UK requiring coding skills in 2016.[41] However this is changing, and around half of young people in the UK consider coding to be as useful as learning a foreign language.[42] Generating code is one of the most promising features of generative AI tools, and can support a variety of technical skill sets.

[39] Glazko, K. et al. (2023). An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility.
[40] ibid
[41] ONS (2017). Are we training enough people to become programmers?
[42] KX (2022). Nearly half of UK students see coding skills as vital as foreign language skills for future career prospects.

For those with limited coding experience, generating code is an important support mechanism to augment their ability and to support learning. For more skilled programmers, a code generator can save them time and mental capacity, enabling them to focus on higher-value work. It can help all users to improve the quality of their code, thanks to more effective testing and debugging. As generative AI tools can generate code in many different languages, including Python, JavaScript and more, they can support all skill levels of developers working with different languages.

There is significant optimism about the potential of the numerous generative AI tools available in supporting developers.[43] Firstly, multi-purpose chat agents like ChatGPT, Bard and Microsoft Copilot can produce code when prompted. In an analysis of the efficacy of using ChatGPT for generating code, researchers found it to be a useful tool for developers, but it struggled with requests to create the complex logic required for many more complex programmes.[44] It was able to support developers with single requests that can be built up prompt by prompt, but required a human touch to steer the development of the code.[45] ChatGPT now has a 'code interpreter' plug-in, which can write and run code more effectively.[46]

One of the most popular code-generating tools is GitHub's Copilot. Powered by generative AI models developed by GitHub, OpenAI and Microsoft, it is trained on all natural languages that appear in public repositories. GitHub Copilot has three subscription plans, one each for individuals, businesses and enterprises, and is freely accessible for teachers, students and maintainers of popular open source projects. AlphaCode, a model developed by DeepMind, is another similar tool that is able to outperform human coders in certain situations.[47] There are also open-source versions of code generators like CodeT5 and Polycoder.

While there is a lot of promise, these tools are not yet perfect. Studies show that they are most effective in helping users with low-complexity tasks such as documentation, generation and refactoring, and their utility reduces as data analysis tasks become complex.[48] This is one of the reasons why people see AI coding tools as a way for developers to become more productive and efficient, rather than to replace them entirely.

[43] Sarkar, A., et al. (2022). What is it like to program with artificial intelligence?
[44] Sadik, A.R., et al. (2023). Analysis of ChatGPT on Source Code.
[45] ibid
[46] Ipsen, A. (2023). How to use ChatGPT's new "Code Interpreter" feature.
[47] Deepmind (2023). Competitive programming with AlphaCode.
[48] McKinsey (2023). Unleashing developer productivity with generative AI.

Beyond bespoke code generators, many of the generative AI tools that exist to support data analytics are designed to embed into particular data analytics platforms[49] and are provided by vendors that specialise in embedding generative AI capabilities into their, or other, platforms. These tools are designed to support skilled data practitioners to make data analytics for organisations more cost-effective and time-efficient. Many businesses are already procuring these services, for example, in the retail sector, to address challenges from enhancing customer engagement and experience to improving employee productivity and managing workflows.[50] Tools like Excel Copilot can help a user with data analysis, such as generating formula suggestions, showing insights in charts and PivotTables, and highlighting interesting portions of data. Many copilot tools are still at an early stage, so there may be limitations in the accuracy of outcomes, but they represent a step towards supporting less technical users to engage more effectively with data.

In addition to analysing data through text-based outcomes, generative AI can present the findings in a more easily accessible way through data visualisation. This approach includes creating charts, graphs and other visual representations to communicate findings and insights effectively. Generative AI algorithms can analyse the underlying patterns and characteristics of the data and automatically generate visually appealing and informative charts and graphs,not only saving time but also ensuring consistency and accuracy in the representation of data.[51] Many traditional data visualisation tools, such as Tableau, are now integrating generative AI capabilities to allow individuals with limited visualisation experience to more easily create their own data visualisations. Some generative AI tools support data visualisation for skilled data practitioners. Sweetviz is one of many Python libraries that generates high-density visualisations for both numerical and categorical features in a dataset.

Data analysis and visualisation are two of the more well provisioned areas for generative AI tooling, but there are still limitations. Generative AI models can have a limited capacity to interpret contextual information and it can be difficult to determine how such models have reached a particular conclusion, which makes it challenging to backtrack and correct any errors in prompting. However, there are tools like Excel Copilot, which explains the relevant formula which it has selected and shows its working.

[49] Pixelplex (2024). Top 12 AI Tools for Data Analytics to Look out for in 2024.
[50] Zaczkiewicz, A. (2023). The Future of Retail: How Generative AI Is Redefining Business Strategy and Brand Experience.
[51] Gill, N. (2023). Generative AI for Data Visualization.

Similarly, code generators lay out the full detail of the code they generate, as well as leaving notes in the code explaining what each chunk of code is doing. Some generative AI models may also struggle to seamlessly integrate and visualise diverse types of data, limiting their effectiveness in handling complex datasets and affecting their outputs.

For non-technical users looking to conduct data analyses or create data visualisations, the available tools may be a good starting point to help them better understand what the data describes about a particular scenario and why something happened. However, with more complex analytics, it is likely that more technical users will rely upon code generators or the RAG data analysis platforms, connected to company data.

What does the future hold for generative AI tools?

A key theme emerging from this report is the *potential* of generative AI tools to bridge the data divide – there is plenty of excitement about what the tools could do, but also some significant hurdles that need to be overcome. For example, knowledge workers using ChatGPT are 37% faster,[52] but when the tool makes mistakes, the people using them are more likely to also make mistakes (around 19%).[53] More can be done to enable human oversight and control to enable these tools to be used effectively when using these tools to bridge the data divide.

Generating code and analysing data are two of the areas with most excitement and potential, likely due to the utility of the tools for developers today. For example, there are currently more than 1.3 million developers already using GitHub's Copilot. A study by GitHub suggests that using Copilot can help existing developers improve productivity, which could contribute an estimated $1.5 trillion to the global economy.[54] Research with developers using Copilot showed that it helped them stay in the flow (73%), preserve mental effort during repetitive tasks (87%),[55] and to complete tasks 56% faster than those without Copilot.[56]

For many of the tools covered above, having a level of technical skill is certainly a benefit to the utility of the tool. For example, being able to identify an incorrect source, some bad code, or incorrect analysis enables the user to work in tandem with generative AI tools, rather than placing a strong reliance upon them. However, this field is moving quickly, and tools to support the identification of errors are being developed to support this issue.[57] There is also potential for generative AI tools to improve the skill sets of low-skilled developers. In one recent study, low-skilled workers improved by 43% by using ChatGPT, while those with more skills improved by about 17%.[58]

---

[52] Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.
[53] Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.
[54] Dohmke, T. (2023). The economic impact of the AI-powered developer lifecycle and lessons from GitHub Copilot
[55] Kalliamvakou, E. (2022). Research: quantifying GitHub Copilot's impact on developer productivity and happiness.
[56] Peng, S., et al. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.
[57] Mündler, N., et al. (2023). Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation.
[58] Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.

Many of the tools covered in this research are free to use at a basic level, but require paid subscriptions to be used with greater functionality. While this cost is likely negligible to a large-scale business, those with smaller budgets, like charities and small businesses, may find the cost of using full versions of these tools to be prohibitive.

There has been much research hypothesising about the impact of generative AI on the future of work, and in particular the future of data work. Working with generative AI can alter the relationship of the person with the task, as generative AI is a new tool; therefore people will need to adapt and learn to work more effectively with it.[59] One skill in particular is prompting, which is different to many of the tasks workers perform today. Prompting refers to the interaction between a user and an AI tool that enables the model to generate the desired result, and usually takes the form of a request written in natural language. Conversational chat agents have the benefit of being able to comprehend a more discursive narrative with the user making them more accessible to a broader range of users. While prompt engineering is at an early stage, people continue to learn and new materials and supporting resources are regularly being developed. Similarly, users will need to utilise AI tools more as a companion rather than relying on the outputs of generative AI with total trust. Because they can incorporate feedback from the user to shape their output, users can refine the output they require, as they would with a companion. One area that will better enable the use of generative AI is explainability of actions. To make decisions on the basis of generative AI tools, it is helpful if the tool can provide additional explanation of the output, to enable a human to be in the loop to verify information and use it to make an informed decision.

Generative AI tools are still at an early stage of development. ChatGPT was only publicly launched in November 2022 and there has been exponential development and improvement since.

Throughout this research, we found that these tools work best with a human in the loop, and human expertise is still required to get the best out of these tools. This combination presents a change to how we currently work, and those working with these tools will have to adapt to the benefits of generative AI tools. According to the Microsoft Research team, we are at a crucial moment in AI's development where 'instead of asking "how will AI affect work?", the question should be "how do we want AI to affect work?'.[60]

[59] Sarkar, A., et al. (2022). What is it like to program with artificial intelligence?
[60] Butler, J., et al. (2023). Microsoft New Future of Work Report 2023.

Overall, these tools have the potential to support more equitable access to and use of data, but in their current form, they are not yet able to support every user in a meaningful way, particularly those with lower technical skill sets. While some of the use cases explored in this report may not be the imagined use of these AI models at the point of development, the potential opportunity to use these tools to close the data divide is huge. To ensure that they can best serve people and society and let them access and use data, additional development and innovation is required for these models. To achieve those developments, increased access to data is crucial, not only to increase innovation, but also to enable society to realise the full value of data.

Methodology

This research focuses on understanding the ways in which generative AI tools can enable people to use data in a more effective way. To do this, we focused on the specific activities of data publishing and use. To curate the list of activities across these functions, we reviewed literature on these activities. Our primary sources were the Open Data Goldbook for Data Managers and Data Holders[61] and The Trials and Tribulations of Working with Structured Data.[62] A set of 48 activities were ranked on a 1-5 scale based on an estimation of the potential of generative AI tools to support these activities. The tools that scored above three were then linked together around four main themes, which form the framework for this report. The process of curating these main activities was done in a collaborative workshop setting amongst the authors and contributors to this report. These four activities formed the basis of the full literature review.

This research was primarily conducted via desk research, primarily covering the topics of the data divide, generative AI and the data publishing and use lifecycle. The literature review, which included academic sources and grey literature, took place between October 2023 and February 2024. We included a broad set of sources to ensure we covered as many perspectives and sources as possible, due to the nascent nature of the field. The sources were found primarily through Google search and Google Scholar, using search terms based on the activities and sub-activities from the data publishing and use lifecycle as well as 'generative AI tools'. We then used a snowball method to explore further sources referenced in the articles we discovered. Overall, we covered more than 100 articles, blogs posts and reports as part of this research.

We complemented the literature review with some exploration of AI tool platforms to look for specific tools, such as Futurepedia. We found it challenging to implement a systematic approach to surveying the tools on Futurepedia due to the ineffectiveness of the search function on the website. However, we were able to find some examples of tools, which feature in this report.

[61] Carrara, W. et al., European Data Portal (2018). Open Data Goldbook for Data Managers and Data Holders: Practical guidebook for organisations wanting to publish Open Data.
[62] Koesten, L. et al. (2017). The Trials and Tribulations of Working with Structured Data: a Study on Information Seeking Behaviour.

Limitations

The desk research methodology used in this research comes with some limitations. Firstly, the field of generative AI is particularly fast moving and constantly evolving. To keep up with developments, this research considered both academic and non-academic literature. Secondly, this team only considered research written in English, narrowing the diversity of information we could cover.