



Published: February 2025

A taxonomy of the data involved in developing, using and monitoring AI systems (‘A data for AI taxonomy’)		
Category	Type of data	Description
Developing AI systems	Data	<ul style="list-style-type: none"> • Most data exists outside of the context of AI development. • Although many new foundation models are able to work with data of various modalities (text, images, audio, video, etc), data is often messy and not necessarily ‘AI-ready’. • Data can be closed (e.g. most enterprise data), shared (e.g. through contract between two organisations) or open (e.g. Wikipedia). • Data that is public rather than openly licensed (e.g. most websites) is not necessarily appropriate and/or legal to use for model training. • Some of this data might only exist in aggregate as the result of creating a training set. It may have been intentionally published and structured with a specific set of use cases in mind that did not originally include AI.
	Training data	<ul style="list-style-type: none"> • Training data is data that has been processed specifically to train an AI model. • Training data is fed to the model to enable the model to recognise patterns within it, refine itself and produce accurate responses to queries. • Training data can be closed, shared or open data. Examples of openly licensed training datasets include CommonCrawl and ImageNet. Examples of shared training data are OpenAI’s Data Partnerships. • There are some new examples of purpose-built, open training datasets, like Common Corpus and Public Domain 12M. BLOOM, BigCode and Argilla are new tools for communities to use to build and iterate on datasets for AI. Platforms like Hugging Face, Dataverse and OpenML have become de-facto homes for training data. • New marketplaces are also emerging to unite supply and demand for training data, such as Human Native AI, Dataset Providers Alliance and Valyu.
	Reference data	<ul style="list-style-type: none"> • Reference data is used in the process of creating a training dataset out of existing data.

		<ul style="list-style-type: none"> • It augments or enriches a training dataset with additional context, including: predefined thesauri, such as Wordnet, which contain linguistic information including words, their synonyms and antonyms; mathematical proofs such as Lean; and knowledge graphs, which are large collections of interconnected facts about topics of interest. Examples of knowledge graphs are Wikidata, DBPedia and Google Data Commons. • Reference data can sometimes support moderating the contents of training data (e.g. Shutterstock's List of Dirty, Naughty, Obscene, and Otherwise Bad Words) or be used to mitigate a model from producing harmful outputs (e.g. SafetyPrompts). • Some reference data can be more specific than others, such as data used in the process of undertaking personalised alignment or examples like community-based collection of linguistic resources to address bias in language technologies.
	<p>Fine-tuning data</p>	<ul style="list-style-type: none"> • Fine-tuning data is used to adapt a pre-trained model to suit more specialised use cases, while maintaining its original capabilities. This data helps produce fine-tuned or custom models. • Fine-tuning data is generally substantially smaller than the training dataset used, and importantly, provides specific knowledge, tasks or use cases to a model. • The data used to undertake fine-tuning depends on the technique used. It can involve: taking portions of a training dataset and reserving them for fine-tuning; labelling a portion of training data with more context; domain adaptation, where the source of data and target tasks are the same as those used in training but the distributions within the data are different; or augmentation, where new data is created by applying transformations to the original training data (like rotations, scaling, and cropping for image data). • There is a significant amount of human input and labour involved in fine-tuning datasets, such as in the process of labelling datasets with contextual information.
	<p>Testing and validation data</p>	<ul style="list-style-type: none"> • Testing and validation data is data used to test a model while it's in development. • Portions of a training dataset can be reserved for testing and validation, but to be useful they need to offer some guarantees of their accuracy, completeness and/or representativeness. The concept of ground truth data has emerged to describe data that has some degree of manual,

		human validation.
	Benchmarks	<ul style="list-style-type: none"> ● Benchmarks are datasets that can be used to evaluate the performance of models. ● Performance in model development seeks to measure the accuracy of the predictions or classifications that a model can make by feeding it with previously unseen data. Here, accuracy is understood as a general term to indicate performance - in practice, depending on the type of model and the task it is applied to, engineers will use a range of metrics. ● Benchmarks can be open, closed or shared. There is momentum behind the creation of more open benchmarks - such as MLCommons Benchmark Datasets, WikiBench and SQuAD2.0 - to enable model comparison across the AI community. ● Benchmarks can also involve evaluations of a model's safety, such as the MLCommons AI Safety Test Specification Schema.
	Synthetic data	<ul style="list-style-type: none"> ● Synthetic data refers to data that is created algorithmically, often using models or simulations, rather than originating from real-world sources. ● Training data, fine-tuning data and benchmarks can be generated synthetically, and synthetic data can be made closed, shared or open (e.g. Awesome Synthetic (text) Datasets). ● There is concern in some parts of the AI community about model collapse, where a model gradually degrades through ingesting too much synthetic data, including data the model may have itself generated.
	Data about the data used to develop models	<ul style="list-style-type: none"> ● Many leading AI firms do not disclose data about the data they've used to develop AI models. ● However, there are various approaches and standards now emerging to disclose data about the data used to develop AI models - including Data Cards and Data Nutrition Labels - as well as an increasing number of mandates from governments for firms to publish or share this data. ● Data about data used to develop AI models involves disclosing aspects such as: the size of the dataset; the source of the data; who created the data; how the data was created; how the dataset has been augmented; whether the dataset includes copyrighted data; what licence the data can be used under; and if personal information is included in the data. ● Data about data is often described as metadata. There are

		<p>many existing metadata standards, such as DCAT and DUO. Croissant is a new, machine-readable metadata format for machine learning-ready datasets.</p>
	Databases of content not available for model training	<ul style="list-style-type: none"> • These are databases of content, or data. that creators have said they do not want to be used for model training. • Spawning's Data Diligence API surfaces all machine-readable rights reservation methods to AI developers, and checks for works in their training dataset against the Do Not Train Registry, which signals the data-use preferences for over 1.5bn individual works. • There can be a legal basis for doing this. For example, in the EU, creators can opt out of the non-research text and data mining exclusion under the EU Copyright Directive.
	Model weights	<ul style="list-style-type: none"> • Model weights are numerical values that a model learns and adjusts to during its development. Each weight represents a specific relationship or understanding the model has captured from the data it was trained on. • Weights encode various contextual relationships, such as linguistic patterns, word meanings and grammar rules. • There is momentum behind making weights openly available. Models with open weights include Mistral 7B and LLaMA.
Deploying AI systems	User data	<ul style="list-style-type: none"> • User data is the data an AI model will ingest and process when it's deployed into a particular context. • The local data involved depends on the purpose of the model, its underlying architecture and the way users interact with it. This data needs to be processed before it can be used by an AI model, just like existing or reference data used in the model's development. • For example, a model that uses retrieval-augmented generation (RAG) accesses a new pool of data just before it responds to a user's query, in order to enhance the accuracy and reliability of its response. A law firm, for example, could use RAG to grant a model access to its internal case databases and email system to generate accurate and relevant outputs, such as draft contract terms or emails, for its staff. • Techniques like RAG make it difficult to draw a tight distinction between development and deployment phases, given they enable models to 'learn on the job'.
	Reference data	<ul style="list-style-type: none"> • As well as in model development, as described above, reference data can be used when deploying models.

		<ul style="list-style-type: none"> For example, by using retrieval-augmented generation (RAG) to retrieve and use up-to-date facts from Google Data Commons to respond to queries, rather than generating the response from the trained model alone.
	Prompts	<ul style="list-style-type: none"> A prompt is an instruction or query, typically written in natural language, inputted to an AI system to generate a response. Prompts are used in many generative, consumer-facing AI tools, such as ChatGPT or Midjourney There are different prompt techniques that can be used, such as chain of thought or few shot learning. A number of databases of and platforms for prompts have emerged, including PromptBase, Hugging Face and AI Prompt Database. Platforms such as SafetyPrompts maintain catalogues of prompt databases that can be used by developers to mitigate the use of prompts that elicit sensitive or unsafe responses.
	Outputs from models	<ul style="list-style-type: none"> The outputs from AI systems can be considered data. This includes statistics or analytics from predictive models, text, audio and video outputs (e.g the CIFAKE dataset, which includes 60,000 images generated by the Stable Diffusion model), and most recently, structured data outputs from generative models (e.g. OpenAI's GPT models can now produce outputs that adhere to JSON Schemas).
Monitoring AI systems	Data about models	<ul style="list-style-type: none"> Many leading AI firms have not disclosed data about the models they've developed. However, there are various approaches and standards now emerging to disclose data about models, such as Model Cards, as well as an increasing number of mandates from governments for firms to publish or share this data. Data about models involves disclosing aspects such as: the model version; the person or organisation developing the model; the licence it is made available under; its intended users and uses; technical attributes; model performance metrics, the data it has been trained on, and ethical considerations for its use.
	Data about model usage and performance in context	<ul style="list-style-type: none"> Many models collect a vast amount of data as they are used, such as the logs of queries made by users, as well as data about its speed and performance. This data is used by the company supplying the model to make improvements. There are arguments and proposals for researchers to have greater access to data about model usage and performance in context (e.g. from the Centre for Democracy and Technology, the Centre for the Governance of AI and The

		<p>Collective Intelligence Project). In the absence of data about model usage and performance in context, members of the AI community sometimes undertake their own evaluation campaigns and publish the results openly (e.g. DecodingTrust). More data about model usage and performance in context will be generated as third-party testing, red-teaming and auditing of AI systems scales.</p> <ul style="list-style-type: none"> • Some are concerned that a new data asymmetry is emerging between the firms supplying models and the consumers that use them. Microsoft, for example, has added restrictions to its consumer AI product's terms and conditions that ban users from scraping or otherwise extracting data from them. ChatGPT has begun to introduce user controls such as the ability to turn off 'conversation' history and export data out of the system.
	<p>Registers of model deployments</p>	<ul style="list-style-type: none"> • Registers of model deployments are authoritative lists of AI models that have been used in a particular context. • This includes lists of models deployed by governments (such as Scotland's mandatory national register of public sector AI and the register of artificial intelligence systems used by the City of Helsinki), lists of models maintained by advocacy organisations (such as the Public Law Project's Tracking Automated Government database register and AI Localism's repository of AI models deployed at local levels) and lists of models approved for use in particular ecosystems or markets (such as the FDA's list of AI-enabled medical devices marketed in the United States).
	<p>Data about the AI ecosystem</p>	<ul style="list-style-type: none"> • There is a growing volume of data being collected and maintained about the AI ecosystem. • Data about the AI ecosystem includes: <ul style="list-style-type: none"> ○ data about AI models, companies and developers (e.g. EpochAI's Notable AI Models and Large-Scale AI Models datasets, the Ethical AI Database, the AI Index and data about the use of platforms like Hugging Face) ○ registers of model incidents and risks (e.g. the AI Incident Database, the OECD AI Incidents Monitor, the AI Intersections Database and the AI Risk Database) ○ databases of AI policies, strategies and regulation (e.g. the OECD's Observatory of National AI Policies & Strategies, the Global AI Law and Policy Tracker and the US State AI Governance Legislation Tracker)

		<ul style="list-style-type: none">○ databases of legal cases (e.g. the Generative AI Intellectual property cases and policy tracker)○ catalogues of AI standards and tools (e.g. the AI Standards Hub and the OECD catalogue of tools and metrics for trustworthy AI)○ data about the workforce involved in the AI supply chain.
--	--	--