NO PREP

How AI Learns

Training Data Explained

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

How AI Learns: Training Data Explained

Learning Without a Teacher (Sort Of)

When people talk about AI 'learning,' they are using a word that sounds familiar but describes something very different from how humans learn. A human child learns to recognize a dog by seeing a few real dogs, maybe getting corrected once or twice, and building a mental category through lived experience. An AI learns to recognize dogs by processing millions of labeled photographs until the patterns in the pixel data become statistically associated with the label 'dog.' Same outcome, completely different process. Training data is the raw material that AI systems learn from. It is the collection of examples, text, images, audio, or other inputs that a model processes during the training phase. The quality, quantity, and diversity of training data have a direct impact on how well the AI performs. A model trained on poor data will make poor predictions. A model trained on biased data will produce biased outputs. Understanding training data is one of the most important concepts in AI literacy.

Name: **Period:** **Date:**

How AI Learns: Training Data Explained

Learning Without a Teacher (Sort Of)

When people talk about AI 'learning,' they are using a word that sounds familiar but describes something very different from how humans learn. A human child learns to recognize a dog by seeing a few real dogs, maybe getting corrected once or twice, and building a mental category through lived experience. An AI learns to recognize dogs by processing millions of labeled photographs until the patterns in the pixel data become statistically associated with the label 'dog.' Same outcome, completely different process. Training data is the raw material that AI systems learn from. It is the collection of examples, text, images, audio, or other inputs that a model processes during the training phase. The quality, quantity, and diversity of training data have a direct impact on how well the AI performs. A model trained on poor data will make poor predictions. A model trained on biased data will produce biased outputs. Understanding training data is one of the most important concepts in AI literacy.

What Training Data Looks Like

For a text-based AI like ChatGPT, training data consists of enormous collections of written text. This includes books, articles, websites, social media posts, academic papers, code, legal documents, and virtually any other written material that could be gathered at scale. OpenAI has confirmed that GPT models were trained on datasets called Common Crawl (a snapshot of billions of web pages), along with books, Wikipedia, and other curated sources. For image-recognition AI, training data consists of millions of labeled images. Researchers at universities and tech companies spend years building datasets like ImageNet, which contains over 14 million images labeled with categories like 'goldfish,' 'rocking chair,' or 'volcano.' The model processes each image and adjusts its internal settings to better match

the label. After millions of iterations, it can reliably identify objects it has never seen before, as long as those objects resemble what it saw in training. For voice AI, training data is recordings of real people speaking. For self-driving cars, it is video footage of real roads. For medical AI, it is patient records and scan data. Every type of AI needs training data that matches the task it is being built for.

The Training Process: Loss and Optimization

The actual process of training an AI model involves a cycle of prediction, measurement, and adjustment. The model makes a prediction on a piece of training data. The prediction is compared to the correct answer using a mathematical formula called a loss function, which measures how wrong the prediction was. Then, using a process called backpropagation, the system calculates how each of the model's internal settings contributed to the error and adjusts them slightly to reduce the error next time. This cycle runs billions of times. Each iteration is called a training step. Over billions of steps, the model's parameters, which are its internal numerical settings, gradually move toward values that produce better and better predictions. The goal is to minimize the loss function, which means getting the predictions as close to the correct answers as possible. This process is called optimization, and the specific algorithm most AI systems use is called gradient descent. For large models, this training process can take weeks or months running on thousands of specialized chips simultaneously. The final result is a trained model with billions of parameters set to values that encode the patterns the model learned from the training data.

Labeled vs. Unlabeled Data

There are two main types of training. Supervised learning uses labeled data, meaning each example in the training set comes with the correct answer attached. An image of a cat is labeled 'cat.' A loan application that ended in default is labeled 'default.' The model learns by comparing its predictions to these labels. Supervised learning is used for tasks like spam detection, medical diagnosis, and image classification. Unsupervised learning uses unlabeled data. The model is not told what the correct answer is. Instead, it discovers patterns on its own. LLMs largely use unsupervised learning: they are given raw text with no labels and learn to predict the next word in a sequence. The 'correct answer' is simply whatever word actually came next in the original text. This is why LLMs can train on essentially all written human text

without requiring humans to manually label each example. A third approach called reinforcement learning from human feedback (RLHF) is used to fine-tune LLMs after initial training. Human raters compare different model outputs and rank which ones are better, more helpful, or safer. The model learns from these rankings to produce outputs that humans prefer. This is a major part of why modern chatbots like ChatGPT feel more helpful and polite than earlier AI systems.

The Garbage In, Garbage Out Problem

There is an old computing principle that applies powerfully to AI: garbage in, garbage out. If the training data contains errors, biases, or harmful content, the model will learn those patterns and reproduce them. Early versions of image recognition AI failed to accurately identify people with darker skin tones because training datasets were not diverse enough. Early language models reproduced harmful stereotypes because the internet text they were trained on contained those stereotypes. This is not a bug that can be patched with a simple fix. It is a consequence of training on data that reflects the imperfections of the real world. AI companies now invest heavily in data curation, which means carefully selecting, cleaning, and filtering training data to reduce harmful content. They also use RLHF to train models to avoid producing harmful outputs. But these efforts are ongoing and imperfect, which is why AI bias remains a major concern in the field. Fine-tuning is a crucial step that happens after initial pre-training. While pre-training teaches a model general language patterns from massive datasets, fine-tuning trains the model on a smaller, curated dataset specific to the desired application. A general LLM fine-tuned on medical literature becomes much better at answering medical questions. One fine-tuned on legal documents performs better at legal tasks. Fine-tuning is far less computationally expensive than pre-training because the model already has extensive general knowledge; fine-tuning merely specializes it. This makes fine-tuning accessible to organizations that cannot afford to train foundation models from scratch. Synthetic data is an emerging approach to addressing training data limitations. Rather than relying entirely on real-world data that may be scarce, biased, or difficult to obtain, AI companies increasingly generate synthetic training data: data created by AI systems themselves or by carefully controlled procedures. For example, a medical AI company might use simulations to generate thousands of synthetic patient records covering rare conditions, conditions that would be nearly

impossible to find enough real examples of in available medical databases. Synthetic data is not without risk, since models trained heavily on synthetic data can develop subtle errors if the synthetic data does not accurately reflect real-world distribution, but it represents an important frontier in addressing data scarcity and diversity problems.

Key Information: AI Training Data

Detail	Information
Definition of training data	The collection of examples an AI processes to learn patterns
Text AI training sources	Web pages, books, Wikipedia, code, academic papers
Image AI training example	ImageNet: 14+ million labeled images
Key training metric	Loss function: measures how wrong a prediction was
Adjustment process	Backpropagation + gradient descent
Supervised learning	Uses labeled data with correct answers attached
Unsupervised learning	Model discovers patterns without labeled answers
RLHF	Reinforcement Learning from Human Feedback — fine-tunes models using human ratings
Garbage in, garbage out	Biased or flawed training data produces biased or flawed AI
Training duration	Weeks to months for large models on thousands of chips

Case Study: Priya's Face Wasn't Recognized

Priya is a 16-year-old who signed up for a new AI-powered app that her school piloted for attendance tracking. The app used facial recognition to automatically mark students present when they walked into the room. For most of her classmates, it worked instantly. For Priya and two other students, who all had darker skin tones, the system consistently failed, requiring them to manually check in at the front office every day. When Priya asked her teacher why the system didn't work for her, the teacher didn't have a good answer. The app's company said the system had been 'tested extensively.' But the testing had used a dataset that was predominantly made up of lighter-skinned faces. The AI had learned patterns from that data and performed well on faces similar to what it had seen during training. It had not seen enough examples of Priya's face type to recognize her reliably. Priya felt embarrassed and singled out every single day. The problem wasn't that the AI was programmed to discriminate. The problem was that the training data didn't represent her.

1. Using what you learned about training data, explain why the facial recognition system worked for some students but not for Priya.

2. What changes to the training process could have prevented this problem from occurring?

Multiple Choice

1. What is a loss function used for during AI training?

    a) Measuring how wrong the model's prediction was compared to the correct answer

    b) Measuring how much electricity the entire training process uses for each epoch

    c) Counting how many times the model has been updated

    d) Calculating the total size of the training dataset

2. What is the main difference between supervised and unsupervised learning?

    a) Supervised learning is run by computers; unsupervised learning depends entirely on human reviewers

    b) Supervised learning is faster; unsupervised learning is slower

    c) Supervised learning uses labeled examples; unsupervised learning discovers patterns without labels

    d) Supervised learning works only for images; unsupervised works only for text

3. What does RLHF stand for, and what is it used for?

    a) Reinforcement Learning from Human Feedback; used to fine-tune models using human ratings

    b) Rapid Learning from Hierarchical Features; used to speed up AI model training

    c) Random Learning from Historical Files; used to build the original training set

    d) Recursive Logic with Human Filters; used to remove harmful content from multiple databases

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Training data	A) A training method that uses labeled examples with correct answers attached
2. ___ Loss function	B) A fine-tuning technique where human raters rank AI outputs to improve model behavior
3. ___ Backpropagation	C) A formula that measures how wrong a model's prediction was
4. ___ Supervised learning	D) The process of adjusting model parameters to reduce prediction error
5. ___ RLHF	E) The collection of examples an AI processes to learn patterns

Fill-in-the-Blank

1. The principle that poor quality training data produces poor AI output is called _____ in, garbage out.

2. LLMs use _____ learning, discovering patterns from raw text without labeled examples.

3. The internal numerical settings of an AI model that are adjusted during training are called _____.

Short Answer

Answer each question in 2-4 complete sentences.

1. Describe the cycle of prediction, measurement, and adjustment that happens during AI training.

2. Why did early AI image recognition systems perform worse on darker-skinned faces, and what does this reveal about the importance of training data diversity?

3. Why does training a large AI model take weeks or months instead of just a few hours?