AI inference vs. training: What is AI inference?

AI inference is when an AI model produces predictions or conclusions. AI training is the process that enables AI models to make accurate inferences.

Learning Objectives

After reading this article you will be able to:

- Define and explain AI inference
- Distinguish between AI training and AI inference
- Compare the amount of compute power used by AI inference vs. training

RELATED CONTENT

What is artificial intelligence (AI)?
What is machine learning?
What is a large language model (LLM)?
Predictive AI
What is deep learning?

Copy article link

What is AI inference?

In the field of artificial intelligence (AI), inference is the process that a trained machine learning model* uses to draw conclusions from brand-new data. An AI model capable of making inferences can do so without examples of the desired result. In other words, inference is an AI model in action.

An example of AI inference would be a self-driving car that is capable of recognizing a stop sign, even on a road it has never driven on before. The process of identifying this stop sign in a new context is inference.

Another example: A machine learning model trained on the past performance of professional sports players may be able to make predictions about the future performance of a given sports player before they are signed to a contract. Such a prediction is an inference.

**Machine learning is a type of AI.*

AI inference vs. training

> **Training** is the first phase for an AI model. Training may involve a process of trial and error, or a process of showing the model examples of the desired inputs and outputs, or both.
>
> **Inference** is the process that follows AI training. The better trained a model is, and the more fine-tuned it is, the better its inferences will be — although they are never guaranteed to be perfect.

To get to the point of being able to identify stop signs in new locations (or predict a professional athlete's performance), machine learning models go through a process of training. For the autonomous vehicle, its developers showed the model thousands or millions of images of stop signs. A vehicle running the model may have even been driven on roads (with a human driver as backup), enabling it to learn from trial and error. Eventually, after enough training, the model was able to identify stop signs on its own.

What are some use cases for AI inference?

Almost any real-world application of AI relies on AI inference. Some of the most commonly used examples include:

Large language models (LLMs): A model trained on sample text can parse and interpret texts it has never seen before

Predictive analytics: Once a model has been trained on past data and reaches the inference stage, it can make predictions based on incoming data

Email security: A machine learning model can be trained to recognize spam emails or business email compromise attacks, then make inferences about incoming email messages, allowing email security filters to block malicious ones

Driverless cars: As described in the above example, inference is hugely important for autonomous vehicles

Research: Scientific and medical research depends on interpreting data, and AI inference can be used to draw conclusions from that data

Finance: A model trained on past market performance can make (non-guaranteed) inferences about future market performance

How does AI training work?

At its essence, AI training involves feeding AI models large data sets. Those data sets can be structured or unstructured, labeled or unlabeled. Some types of models may need specific examples of inputs and their desired outputs. Other models — such as deep learning models — may only need raw data. Eventually the models learn to recognize patterns or correlations, and they can then make inferences based on new inputs.

As training progresses, developers may need to fine-tune the models. They have it provide some inferences right after the initial training process, then correct the outputs. Imagine an AI model has been tasked to identify the photos of dogs from a data set of pet photographs. If the model instead identifies photos of cats, it needs some tuning.

How does AI compute power usage compare for inference vs. training?

AI programs extend the capabilities of computers to far beyond what they were able to do previously. But this comes at the cost of using much more processing power than traditional computer programs — just as, for a person, solving a complex mathematical equation requires more focus and concentration than solving "2 + 2."

Training an AI model can be very expensive in terms of compute power. But it is more or less a one-time expense. Once a model is properly trained, it ideally does not need to be trained further. If the model does need to be adapted to a new use case, developers can use less-intensive techniques like low-rank adaption (LoRA) instead of retraining the model from scratch.

Inference, however, is ongoing. If a model is actively in use, it is constantly applying its training to new data and making additional inferences. This takes quite a bit of compute power and can be very expensive.

How does Cloudflare allow developers to run AI inference?

Cloudflare Workers AI offers developers access to GPUs all over the globe for running AI tasks. This pairs with Vectorize, a service for generating and storing embeddings for machine learning models. Cloudflare also offers cost-effective object storage for maintaining collections of training data — R2, a zero-egress-fee storage platform.

Learn more about how Cloudflare enables developers to run AI inference at the edge.