

Open4Tech Summer School 2025

How to Train Your Personal AI Assistant



Razvan Tudosie Software Developer, Syncro Soft



Alexandru Smarandache Software Developer, Syncro Soft



Agenda

- What is AI?
- Al Core Components
- The AI Model Training Process
- Fine-Tuning
- Demo
- Conclusions



What is Al

- Artificial Intelligence (AI) is a broad collection of techniques and algorithms developed to solve specific problems.
- The history of AI is long and tumultuous, marked by periods of rapid advancement as well as times when progress stalled almost completely.



Most important "AI" related events (credits digitalwellbeing.org):

1950

TURING TEST

Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955

A.I. BORN UNIMATE Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961 1964

ELIZA

First industrial robot. Pioneering chatbot Unimate, goes to work developed by Joseph Weizenbaum at MIT at GM replacing humans on the holds conversations assembly line with humans

1966

SHAKEY

The 'first electronic person' from Stanford, Shakey is a generalpurpose mobile robot in the cold that reasons about its own actions

A.I.

WINTER

Many false starts and dead-ends leave A.I. out

1998

DEEP BLUE

Deep Blue, a chessplaying computer from

KISMET

Cynthia Breazeal at MIT introduces KISmet, an IBM defeats world chess emotionally intelligent. robot insofar as it detects and responds to people's feelings

champion Garry Kasparov

1997





2017

ALPHAGO

Google's A.I. AlphaGo beats world champion Ke Jie in the complex

1999

Sony launches first consumer robot pet dog autonomous robotic AiBO (Al robot) with

2002 ROOMBA

First mass produced Apple integrates Siri, an intelligent virtual vacuum cleaner from assistant with a voice

2011 2011

WATSON

IBM's question answering computer Watson wins first place

2014

Turing Test with a third

Eugene Goostman, a chatbot passes the

ALEXA

Amazon launches Alexa, Microsoft's chatbot Tay an intelligent virtual assistant with a voice

2014

2016

goes roque on social

media making



What is Al(now)

 Now, AI is capable of performing a wide range of tasks—from simple ones like predicting weather patterns, to more complex activities such as image recognition and powering personal assistants. The field of AI is evolving rapidly, growing at a pace no one could have anticipated.



AI Core Components

Artificial Intelligence Natural Language Problem Solving Processing (NLP) & Search Strategies Knowledge Intelligent Robotics Representation **Machine Learning** Automated Planning and Principal Component K-Nearest Scheduling Programming Support Vector Analysis (PCA) Neighbours (KNN) Machines (SVM) Speech Visual Reinforcement Naive Bayes Recognition Perception Classification Learning Neural Networks Linear/Logistic Automatic Expert Recurrent Neural Adaptive Resonance Regression Reasoning Systems Networks (RNN) Theory (ART) K-Means Multilaver **Radial Basis** Autoencoders Clustering Ensemble Perceptrons (MLP) **Function Networks** Methods Decision Self Modular Trees Organising Anomaly Neural Deep Learning Maps Detection Networks Random Boltzmann Generative Long Short-Term Hopfield Forest Adversarial Machines Memory Networks Networks Networks (GAN) (LSTM) Transformer Models Convolutional Neural (BERT, GPT etc) Networks (CNN) Recurrent Neural Deep Reinforcement Networks (RNN) Learning Deep Belief Deep Networks Autoencoders @robmay70 (DBN)

 "Al" represented as an onion of different algorithms and techniques ->



Goals

- The goal of this course is to familiarize participants with the AI field and guide them through the process of training a personal AI model to solve a specific task.
- We will also investigate some practical examples.



• The process of "training" an AI model is usually done in 3 steps.

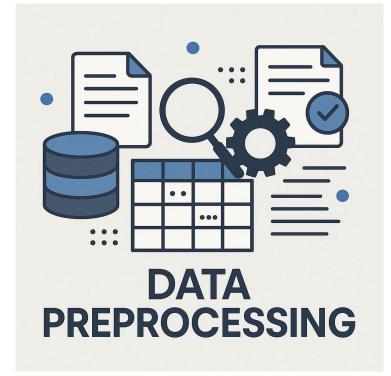




Step I: Data preprocessing



 The "Data preprocessing" step is the first stage in the machine learning workflow. It involves creating or collecting relevant data that the AI model will train on.





- During this phase, specific data processing tasks are essential, such as cleaning the data, handling missing values, normalizing or standardizing features, and sometimes transforming data into formats suitable for the model.
- Proper data preprocessing is critical, as the quality and structure of the data directly impact the performance and reliability of the AI model.
 Data Preprocessing
- This is the step where most mistakes are made!





- When searching for data to train your AI model, two main concerns often arise: privacy and cost.
- Many of the best sources of data are hidden behind paywalls, or are simply proprietary assets of other companies.

But... what about open source datasets?





- There are a number of open-source datasets hosted on sites like HuggingFace or Kaggle.
- These datasets range from useless, to mildly interesting (such as a "Zara sales analysis" or "English Premier League Player Stats 24/25" :)), to even game-changing, depending on what we want our AI to accomplish.







https://huggingface.co/d

How to train Your Personal AI Assistant

Main Tasks Libraries Languages Licenses Other	Datasets 433,000 Filter by name	Full-text search 11 Sort: Trending	
Iodalities	EssentialAI/essential-web-v1.0	institutional/institutional-books-1.0	
3D Audio Document Geospatial	● Preview - Updated 3 days ago - ± 75.5k + ♡ 148	■ Viewer + Updated 9 days ago + □ 983k + ± 38.2k + ♡ 184	
🖪 Image 🔲 Tabular 😝 Text 📮 Time-series			
	fka/awesome-chatgpt-prompts	nvidia/AceReason-1.1-SFT	
Video	\boxplus Viewer \cdot Updated Jan 6 \cdot \boxplus 203 \cdot \pm 22.9k \cdot \heartsuit 8.03k	■ Viewer + Updated 7 days ago + □ 3.96M + ± 2.95K + ♡ 42	
ize (rows)	nvidia/OpenScience	BAAI/CCI4.0-M2-CoT-V1	
< 1K > 1T	E Viewer + Updated 7 days ago + □ 4.48M + ± 435 + ♡ 23	Wiewer + Updated 16 days ago + □ 157M + ± 5.45k + ♡ 39	
•	nvidia/Nemotron-Personas	Lixsp11/Sekai-Project	
ormat	■ Viewer - Updated 16 days ago - □ 100k - ± 18.5k - ♡ 146	Elispii/Sekai-Fioject Every Sekai-Fioject Every Sekai-Fioject	
(→ json 💷 csv 🗇 parquet 🕞 imagefolder			
🕼 soundfolder 🐐 webdataset 😑 text 🚿 arrow	lingshu-medical-mllm/ReasonMed	Team-ACE/ToolACE	
	■ Viewer + Updated 1 day ago + □ 1.11M + ± 1.04k + ♡ 44		
	Josephgflowers/Finance-Instruct-500k	<pre>open-r1/Mixture-of-Thoughts</pre>	
	\boxplus Viewer \ast Updated Mar 1 \ast \boxplus 518k \ast \pm 1.72k \ast \heartsuit 105	\boxplus Viewer + Updated 30 days ago + \boxplus 699k + \pm 41.4k + \heartsuit 240	
	openbmb/Ultra-FineWeb	LLM360/guru-RL-92k	
	■ Viewer + Updated 9 days ago + □ 1.29B + ±47k + ♡ 197	Preview + Updated 5 days ago + ± 428 + ♡ 14	
	cognitivecomputations/dolphin-distill	APRIL-AIGC/UltraVideo	
	■ Viewer - Updated 10 days ago - □ 11.6M - ± 180 - ♡ 14	■ Viewer - Updated 5 days ago - □ 58.8k - ± 795 - ♡ 13	
	🔲 cais/mmlu	HuggingFaceFW/fineweb	
	Wiewer + Updated Mar 8, 2024 + □ 231k + ≥ 164k + ♡ 495	■ Viewer • Updated Jan 31 • \Box 258 • \pm 210k • \heartsuit 2.21k	

 However, you might also need to look for specific scientific papers. In many cases, the data published in peer-reviewed journals is made openly available after the article is published.

𝖓 Competition: Datasets Discussions Your Work ASL Fingerspelling Im. Google - American Sig American Sign Langua Sign Language Hand View Active Events

+ Create

(Home

& Model

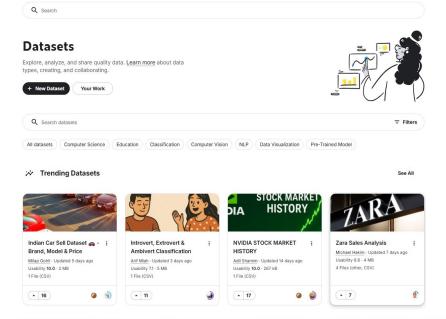
<> Code

A Learn

V More

VIEWED

seals. YOLOV • A typical search for a specific dataset involves looking for relevant keywords and/or intended uses.





- A common way to improve an AI model is to create a training dataset by merging two or more open-source datasets.
- By combining different sources, you can expose your model to a wider variety of data, increasing its robustness and ability to perform well even in unexpected edge cases *.

*There might be some drawbacks if the merging isn't done carefully, which we will discuss in the "Testing and Validation" phase.

For example, improper merging can lead to issues such as overfitting, where the model learns patterns that do not generalize well to new, unseen data.



← premier league

<> Notebooks 1,477	← Comments 839	Datasets 743 X Topics 215 🛞 Hugging Face Models 10 & People 5	Models 4 Competitions 3
Filter by	743	Results	Relevance -
DATE			
O Last 90 days	49	Premier League	× 313
O This week	5	Content The data was acquired from the Premier League website and is representative of seasons 2006/	17,139 downloads
O Today	3	Indian Premier League (Cricket)	* 894
VIEWED BY YOU		Dataset · 8y ago · by <u>Manas</u> Source: Content All Indian Premier League Cricket matches between 2008 and 2016.	84,320 downloads
O Viewed		English Dromier League	- 172
O Not Viewed	742	English Premier League Dataset - 3y ago - by SAIF UDDIN Barclays premier league	25,648 downloads
CREATOR			
O You	o f	Premier League Dataset - 24d ago - by IvanRamosDataTech	48
O Others	743	The English Premier League is the most popular and highest level football league in the world.	3,605 downloads
		Indian Premier League 2008-2019	· 417
	C	Dataset · 6y ago · by <u>Navaneesh Kumar</u> Context Indian <u>Premier League</u> (IPL) is a Twenty20 cricket format <u>league</u> in India.	34,938 downloads
	4	English Premier League (EPL) Results	* 152
	4	Dataset - 3y ago - by <u>Alvin</u> Premier League Results from 1993-94 to 2021-22	8,935 downloads



How to train Your Personal Al Assistant(I) But what if there are no open source datasets? Then what?





How to train Your Personal Al Assistant(I) If you can't find any open-source datasets, there are generally two alternatives:

- 1. You can *purchase data* from a data broker, which can be expensive but provides instant access;
- 2. You can *collect and curate* your own dataset, which is often time-consuming but gives you full control over data quality and relevance.



Step II: Training



• The "Training" step is often the most straightforward one. At this stage, we use the dataset we have collected to feed a training algorithm, which adjusts the model's parameters in order to learn useful patterns from the data.

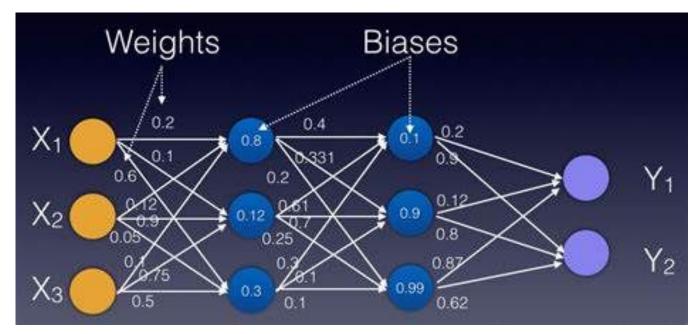




- This step involves writing some code (usually in Python) and running it either on our local machine or in the cloud.
- Without going into much detail about how AI works, this is the stage where we train the "intelligent" part of the artificial "brain" (similar to training the "neurons" in a human brain).



- The "training" step focuses on adjusting the weights and biases between the nodes in the network.
- An AI model (specifically, a machine learning or neural network model) consists of interconnected nodes, each with associated values that are updated during training.



Introduction to Deep Learning



- Usually, the dataset is split into three parts: "training", "validation", and "testing".
- This allows us to properly evaluate our model by measuring how well it performs on data it hasn't seen during training.



- For a lone developer, a common way to train a model is by using Google Colab.
- You need to create a <u>.ipynb</u> file (which is a combination of Python code and documentation, and can be executed step by step).
- The code will then run in the cloud.





• This is an example of a running training process. Depending on the size of the dataset, training can take anywhere from a few minutes to several hours.

O gesture_recognizer.ipynb Save in GitHub to keep changes File Edit View Insert Runtime Tools Help	🔅 🖙 Share 🔶 Gemini 🌘	
Q. Commands + Code + Text > Run all Copy to Drive from the dataset. The resulting dataset will contain the extracted hand landmark positions from each image, rather than images themselves. Image: The HandbataPreprocessingParams class contains two configurable options for the data loading process: Image: Shuffle: A boolean controlling whether to shuffle the dataset. Defaults to true. • min_detection_confidence: A float between 0 and 1 controlling the confidence threshold for hand detection. Split the dataset: 80% for training, 10% for validation, and 10% for testing. • Image: data = gesture_recognizer.Dataset.from_folder(dirname-dataset_path, hparams=gesture_recognizer.HandDataPreprocessingParams()) • Image: data , rest_data = data.split(0.8) validation_data, test_data = rest_data.split(0.5) •	task	esture recognition guide example n <u>ai.google.dev</u>
Train the model Train the custom gesture recognizer by using the create method and passing in the training data, validation data, model options, and hyperparameters. For more information on model options and hyperparameters, see the Hyperparameters section below.	↑ ↓ ♦ ☞ ✿ ▮ 🔟 :	
<pre>Evaluate the model performance After training the model, evaluate it on a test dataset and print the loss and accuracy metrics. [] loss, acc = model.evaluate(test_data, batch_size=1) print(f"Test loss:{loss}, Test accuracy:{acc}")</pre>		
() Variables 🗈 Terminal) Executing (13s) 🛛 📙 Python 3	



 For large datasets—such as those used for enterprise or production-ready models-cloud-based training can be quite expensive, and training locally may be a more cost-effective solution.



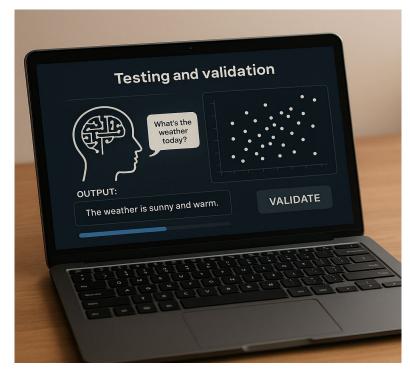
NVIDIA Volta, a special architecture for training deep learning models.



Step III: Testing and validation



- The "Testing and Validation" step is where we analyze how well our trained model performs.
- There are several ways to measure its performance, but the most important metrics are accuracy, precision, recall, and F1-score.





- Accuracy: The percentage of all predictions that the model got correct.
- **Precision**: Of all the items the model identified as positive, how many actually are positive. (It measures how many selected items are relevant.)
- **Recall**: Of all the actual positive items, how many did the model identify as positive. (It measures how many relevant items are selected.)
- **F1-score**: The harmonic mean of precision and recall. It provides a balance between precision and recall, especially useful when you need to balance both.



- After analyzing these metrics, the next step is the manual testing phase.
- At this stage, we might be pleased with the results and consider the training process a success. 🙂
- But more often than not, we find that the model is... lacking. The most common problems are:
 - 1) Underfitting
 - 2) Overfitting



1) *Underfitting*: The model is too simple and fails to capture the underlying patterns in the data. Possible solutions:

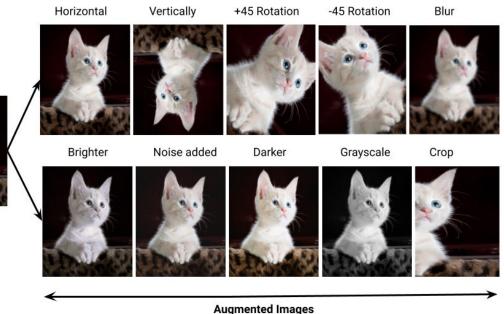
- Use a **more complex model** (e.g., add more layers or nodes to a neural network).
- Increase the duration or number of training epochs (how many times the model "trained" on the data).
- Provide more or better quality data. <-usually the culprit :(
- Reduce regularization or <u>avoid</u> techniques that artificially reduce the amount of training data.



What is Data Augmentation? A technique used to artificially expand your dataset by transforming existing examples — improving generalization and reducing overfitting.

Original Image

- 📝 For Text:
 - Synonym replacement
 - Back-translation
 - Prompt rewording
 - Sentence shuffling
- 🔄 For Images:
 - Rotation & flipping
 - Cropping & zoom
 - Brightness/contrast changes
 - Noise or blur injection





2) Overfitting: The model performs well on training data but poorly on new, unseen data because it has "memorized" the training examples, including noise.

Possible solutions:

- Use a **simpler model** with fewer **parameters**.
- Use regularization techniques (like dropout, L1/L2 regularization).
- Gather more data to help the model generalize better.
- Perform early stopping during training (stop training when validation performance stops improving).



- Once you have identified the problem, apply the appropriate solution and repeat steps I and II.
- After a few iterations, you will have an efficient AI model tailored to your task.







How to train Your Personal Al Assistant But...what about the costs?

- To fully train a *good* working model from scratch it could cost "tens of thousands to hundreds of millions of dollars" (based on estimates made by <u>epoch.ai</u>)
- Think about the data licensing costs, computing costs and other such expenses...



Solution: Fine-tuning



What is fine-tuning?

Fine-tuning is the process of adapting a pre-trained AI model to a specific task.

Instead of training a model from scratch, we use an existing one as a foundation.

This drastically reduces:

- . 🕒 Time
- E Compute power
- . 💸 Cost





Why Fine-Tuning Instead of Full Training?

- Pre-trained models already understand language patterns and general knowledge.
- You only need a small amount of task-specific data to guide them.
- Ideal when:
 - You have limited data or resources.
 - You want to focus on a niche domain (e.g. legal, medical, banking).
 - You need fast iteration and deployment.

"Pretraining costs millions. Fine-tuning costs hundreds."



OpenAl Fine-Tuning

\$

\$

0

Create a fine-tuned model

Method

Specify the method to be used for fine-tuning

Supervised

Base Model

gpt-4.1-2025-04-14

Suffix

Add a custom suffix that will be appended to the output model name.

my-experiment

Seed

The seed controls the reproducibility of the job. Passing in the same seed and job parameters should produce the same results, but may differ in rare cases. If a seed is not specified, one will be generated for you.

Random

Training data

Add a jsonl file to use for training. By providing the file, you confirm that you have the rights to use the data.



Example for Fine-Tuning using OpenAI API



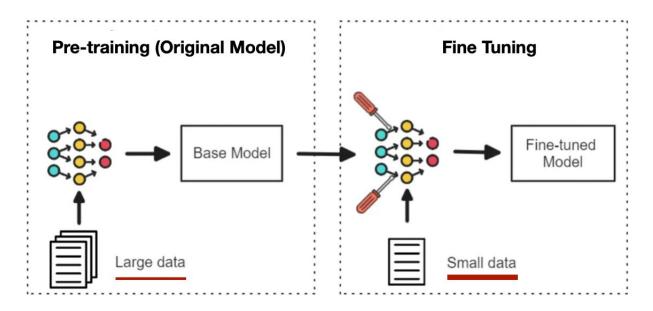
Fine-Tuning on LLMs (Large Language Models)

Fine-tuning an LLM means adapting a general-purpose language model (e.g., GPT, LLaMA, Falcon) to a specific domain or task using a smaller, specialized dataset.

Why do it?

- Makes the model more accurate and relevant to your use case (e.g., banking, legal, medical).
- Reduces the need for prompt engineering.
- Achieves better performance with fewer inputs.

Large Language Model





Fine-Tuning on LLMs (Large Language Models)

Resources:

https://cookbook.openai.com/examples/how_to_fin etune_chat_models

https://www.datacamp.com/tutorial/fine-tuning-lar ge-language-models



Recap...

- The three main steps in training an AI model are: collecting and preprocessing data, selecting and configuring the model architecture, and iteratively training and evaluating the model for optimization.
- Common pitfalls include using insufficient or low-quality data, overfitting the model, and neglecting proper validation and testing processes.
- Fine-tuning involves adapting a pre-trained model to a specific task with less data and resources, while training from scratch requires building and training a model entirely with new data.
- Useful resources for training AI models include open-source datasets, tutorials, online courses, and extensive documentation found on platforms like TensorFlow, PyTorch, Hugging Face, and academic repositories.

THANK YOU!

Any questions?

Razvan Tudosie razvan_tudosie@sync.ro

Alexandru Smarandache alex_smarandache@sync.ro

© 2025 Syncro Soft SRL. All rights reserved.