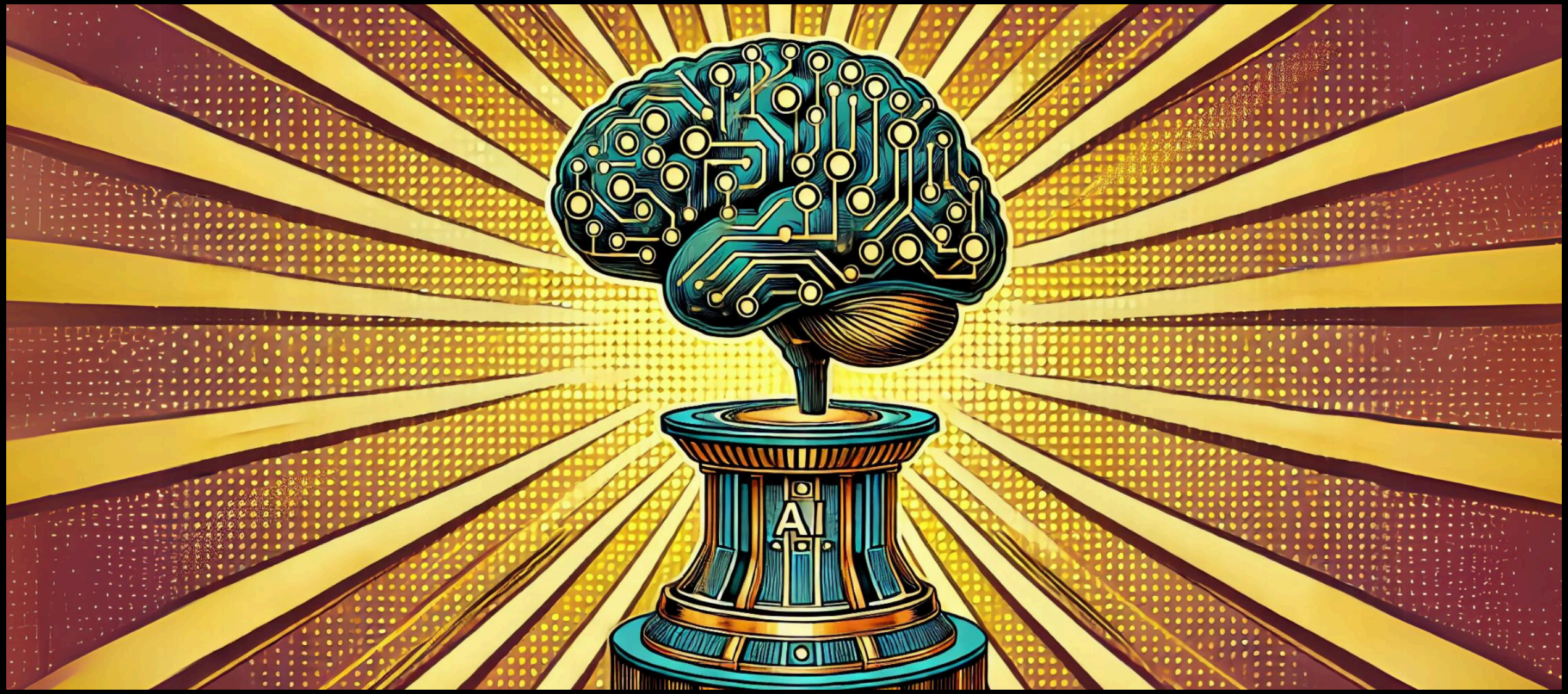
LLM Engineering

MASTER AI & LARGE LANGUAGE MODELS



WEEK 6

Now it gets real: from inference to TRAINING

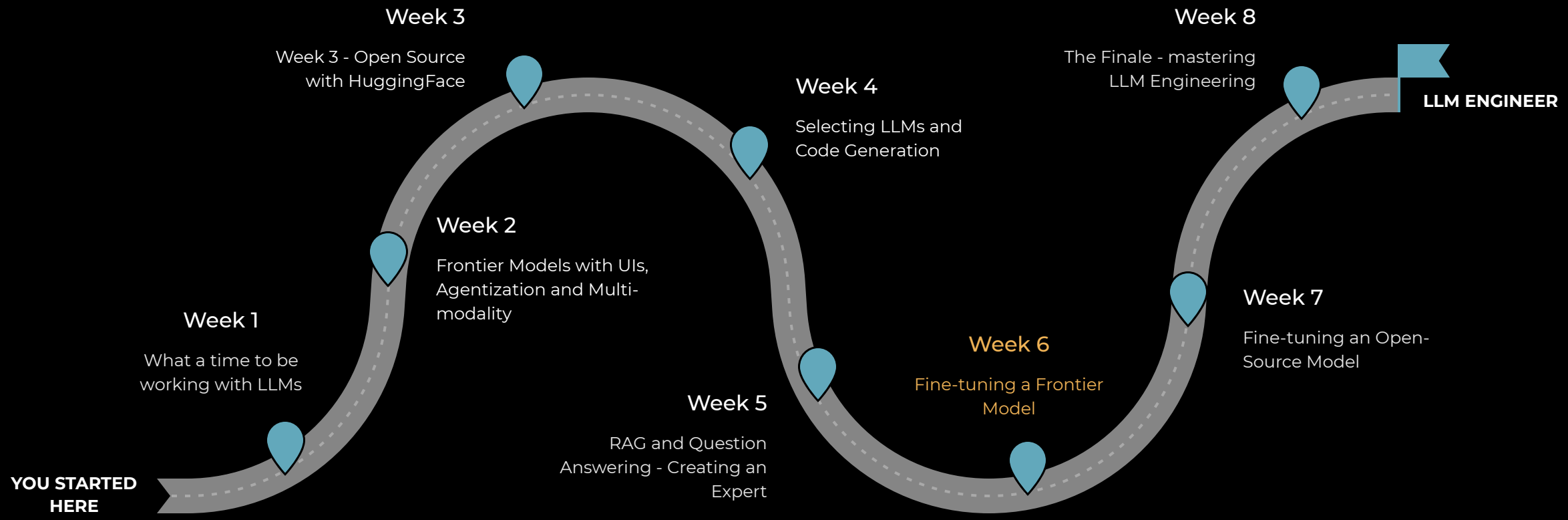
What you can now do

- Generate text and code with Frontier Models including AI Assistants with Tools, and with open-source models with HuggingFace transformers
- Confidently choose the right LLM for your project, backed by metrics
- Create advanced RAG solutions with LangChain

Today we start on the major project; very shortly you'll be able to

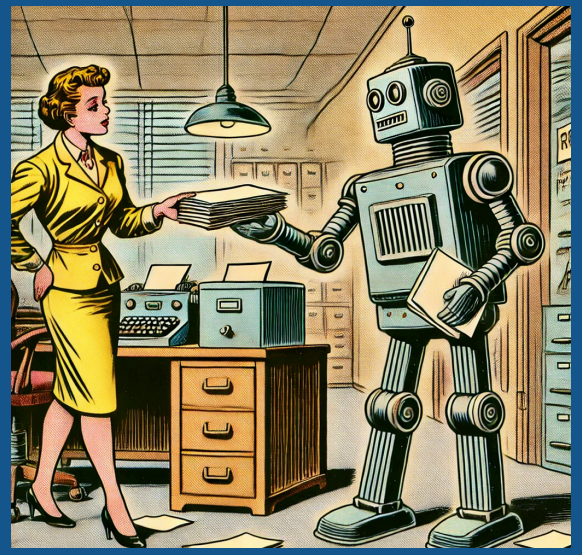
- Download a Dataset from the HuggingFace hub
- Examine a dataset
- Identify evaluation criteria for judging success

Reminder of the 8 weeks to mastery

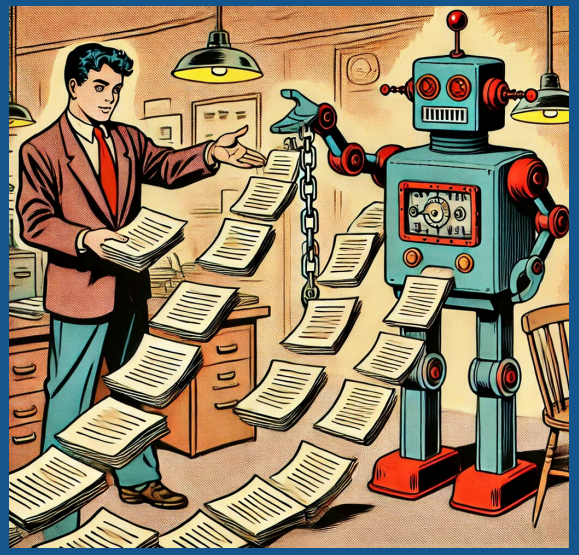


So far we have focused exclusively on INFERENCE

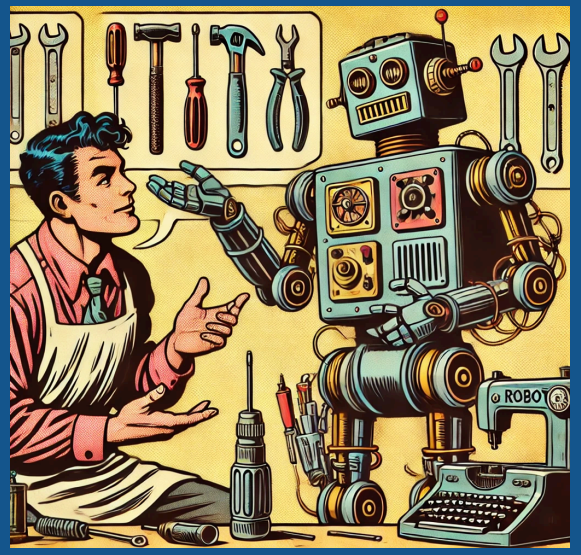
Techniques to improve results at run-time with Closed and Open-Source models



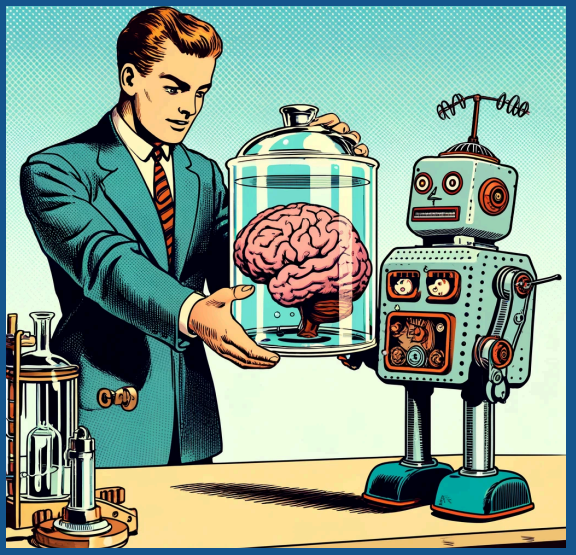
Multi-shot prompting



Prompt Chaining

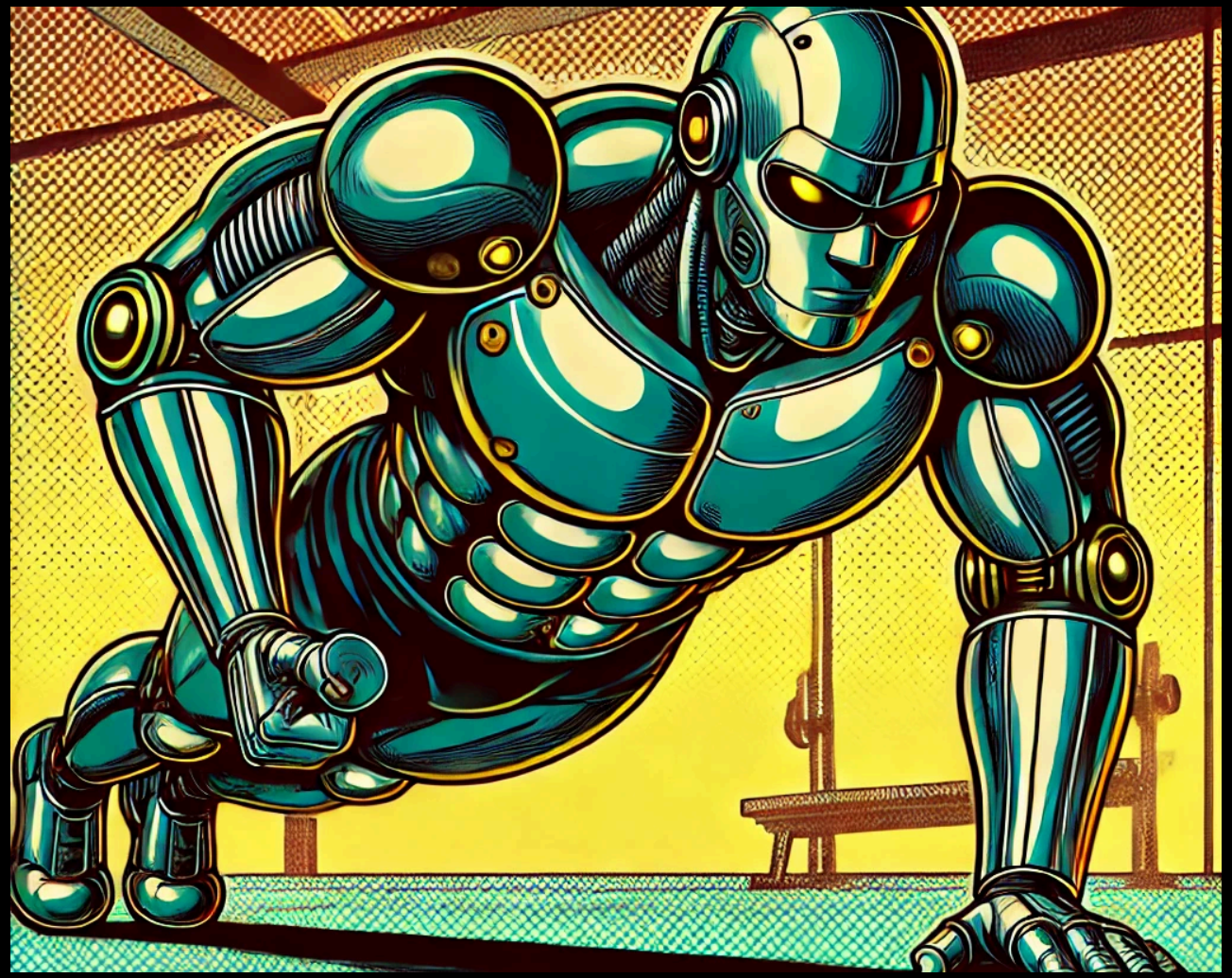


Tools / Function calling



RAG / Knowledge Base

This week we turn to TRAINING



- Training a multi-billion parameter model from scratch would cost tens to hundreds of million $
- Instead, we take advantage of **Transfer Learning**
- We take a pretrained model as base, and use additional training data to fine-tune it for our task

BUT FIRST, TO INTRODUCE

A juicy commercial problem

Given a description of a product, predict its price

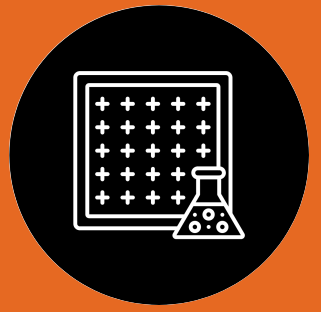
- For a marketplace to estimate prices of goods
- Future versions should be able to write and improve descriptions too
- We'd typically use a Regression model to predict prices, but there are good reasons to try Gen AI
  - We can train an LLM and evaluate it very clearly
  - It means we can battle with GPT-4o
  - Spoiler alert: the frontier models are already great at this!

Finding datasets

Your own proprietary data

Kaggle

HuggingFace datasets



Synthetic data

Specialist companies like Scale.com

HuggingFace is a Treasure Trove of Data

Datasets: McAuley-Lab / **Amazon-Reviews-2023** like 58

Languages: English Size: 10B<n<100B Tags: recommendation reviews

Dataset card Files Community 7

Dataset Viewer View in Dataset Viewer

The viewer is disabled because this dataset repo requires arbitrary Python code execution. Please consider removing the loading script and relying on automated data support (you can use convert to parquet from the datasets library). If this is not possible, please open a discussion for direct help.

Amazon Reviews 2023

Please also visit amazon-reviews-2023.github.io/ for more details, loading scripts, and preprocessed benchmark files.

Downloads last month 24,354

Edit dataset card

Models trained or fine-tuned on McAuley…

hyp1231/blair-roberta-base

Feature Extraction • Upd… • 3.57k • 1

Digging into the data

We'll do some work today, and some refinement tomorrow

Investigate

Parse

Visualize

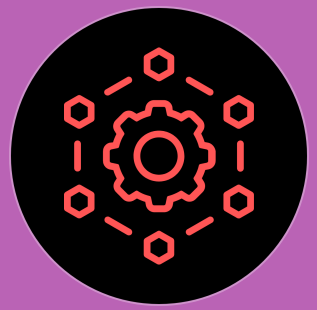
Assess Data Quality

Curate

Save

How will we evaluate performance?

From our Predicted Prices versus Actual Prices



Model-centric or Technical Metrics

Training loss

Validation loss

Root Mean Squared Log Error (RMSLE)

Business-centric or Outcome Metrics

Average price difference

% price difference

% estimates that are "good"

PROGRESS REPORT

Just before we get coding

What you can now do

- Generate text and code with Frontier Models including AI Assistants with Tools, and with open-source models with HuggingFace transformers
- Confidently choose the right LLM for your project, backed by metrics
- Create advanced RAG solutions with LangChain
- Select, investigate and curate a Dataset

After next time you'll be equipped with new important skills

- Lay out a 5 step strategy for selecting, training and applying an LLM
- Contrast the 3 techniques for improving performance and give use cases
- Curate and upload a dataset that's ready for training