ARTIFICIAL INTELLIGENCE

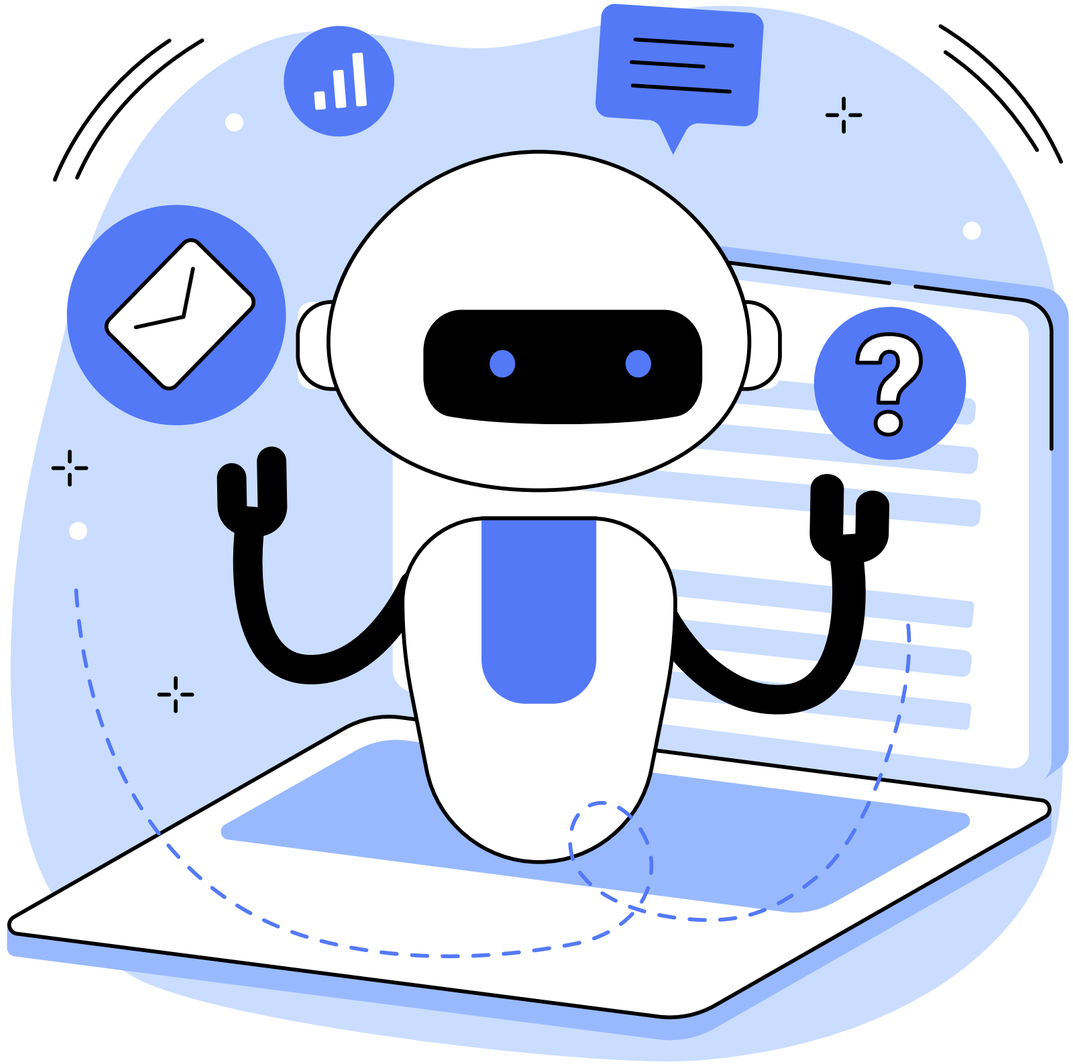
Master Class

CONTENTS

INFERENCE

Inference is when an AI model makes predictions using new data like a chatbot answering questions or software flagging fraud.

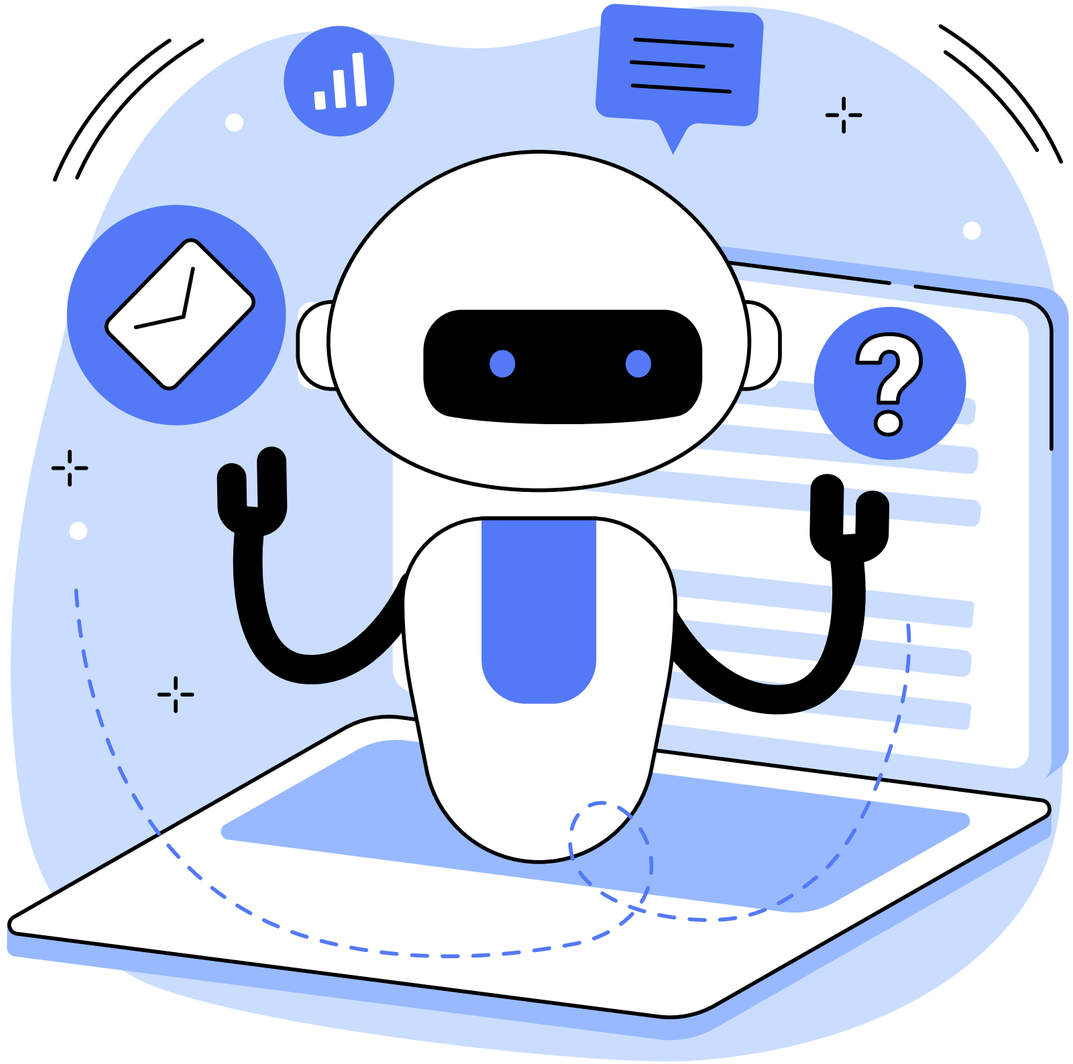
It is basically a model that is runnable, accepts input and produces predictions.



OPEN WEIGHTS

Open weights are model files freely shared so anyone can use or improve the AI.

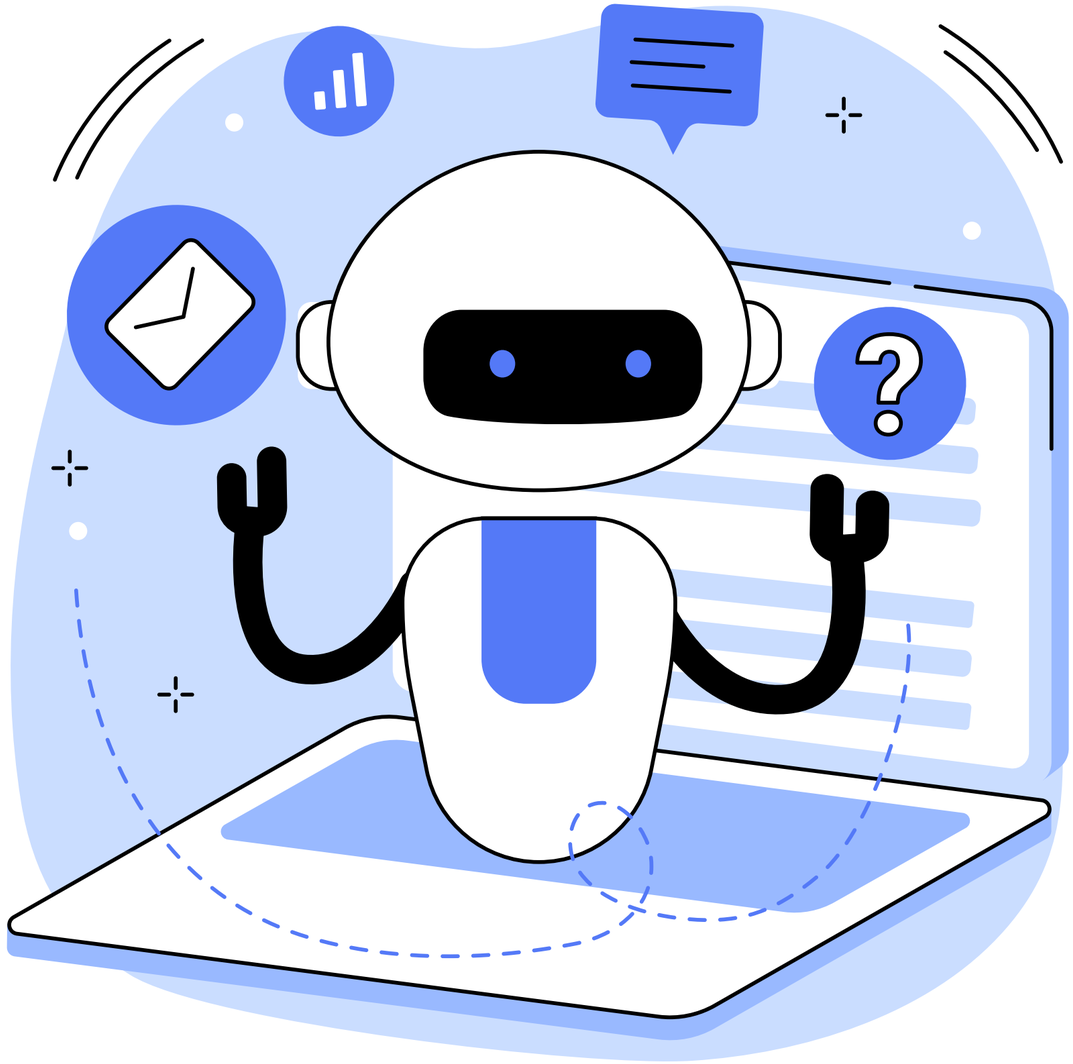
Meta's Llama and Mistral's models let researchers and startups build new apps with full access.



WEIGHTS

Numerical values assigned to connections between neurons that determine how strongly an input feature influences the output prediction.

Good weights lead to accurate predictions and better generalization on unseen data.



EXAMPLE WEIGHTS

Predicts house prices based on features like size, number of rooms and age.

The network assigns weights to each feature to indicate its influence on price.

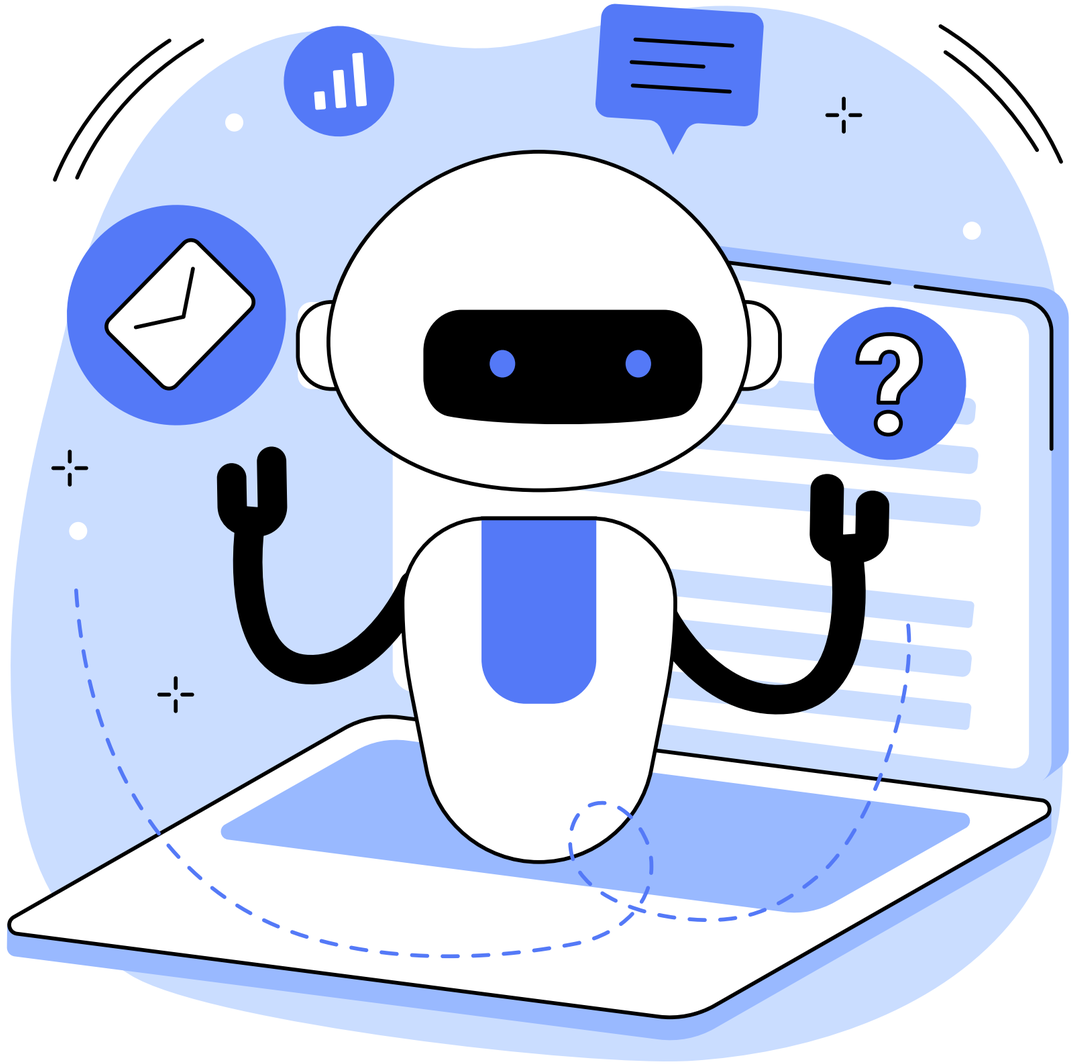
- Size weight: 3.5 (means size strongly increases price)
- Bedrooms weight: 1.2 (moderate influence)
- Age weight: -0.9 (older houses reduce price)

The model multiplies each input by its weight, sums them, and adds a bias (say, 50k) to adjust the baseline price independent of inputs.

BIAS

Bias in AI means systematic, unfair errors often baked in from data, design, or human choices, that skew an AI system’s results.

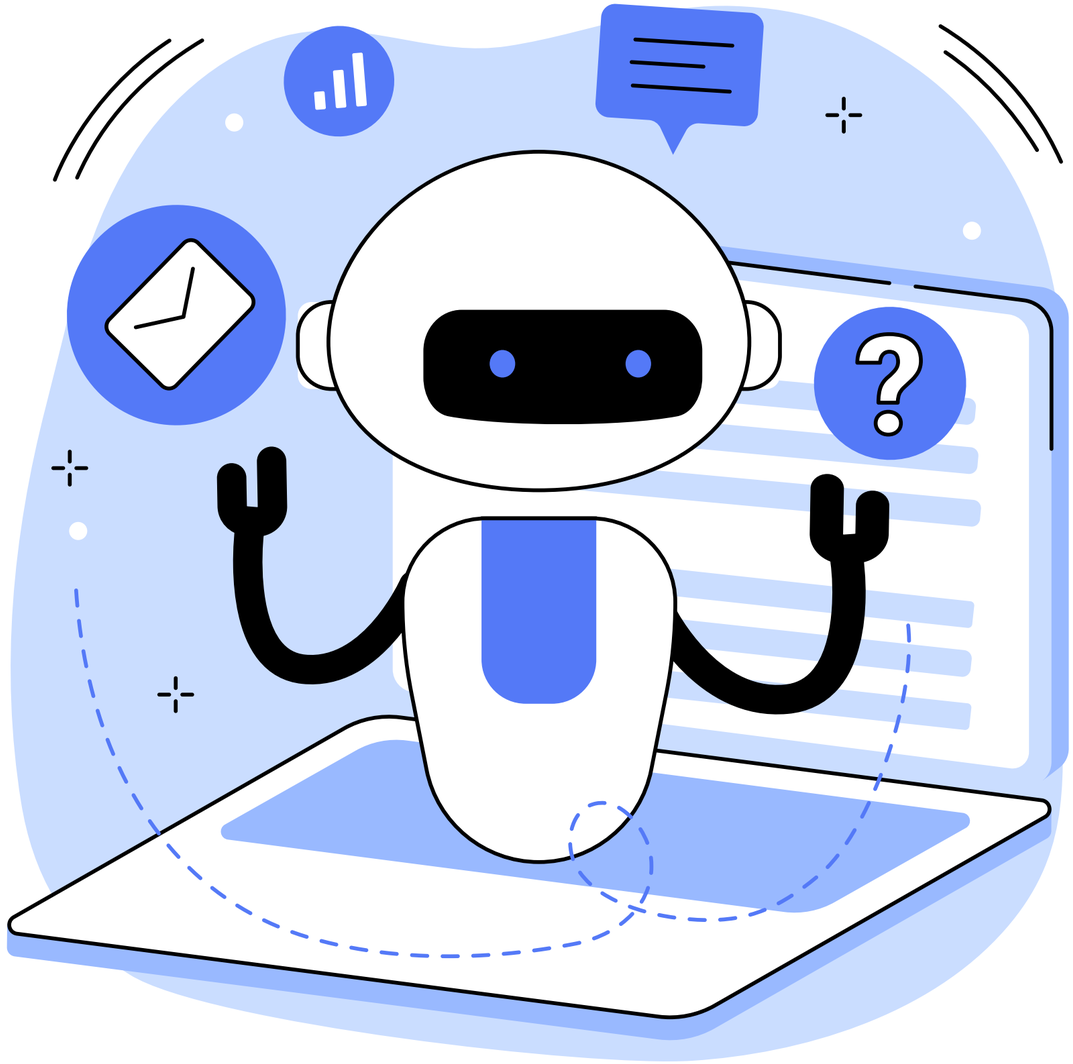
A hiring AI trained mostly on resumes from men may unfairly favour male applicants, perpetuating gender inequality.



FEATURES

Features are the input variables or attributes given to an AI model, describing the characteristics of the data.

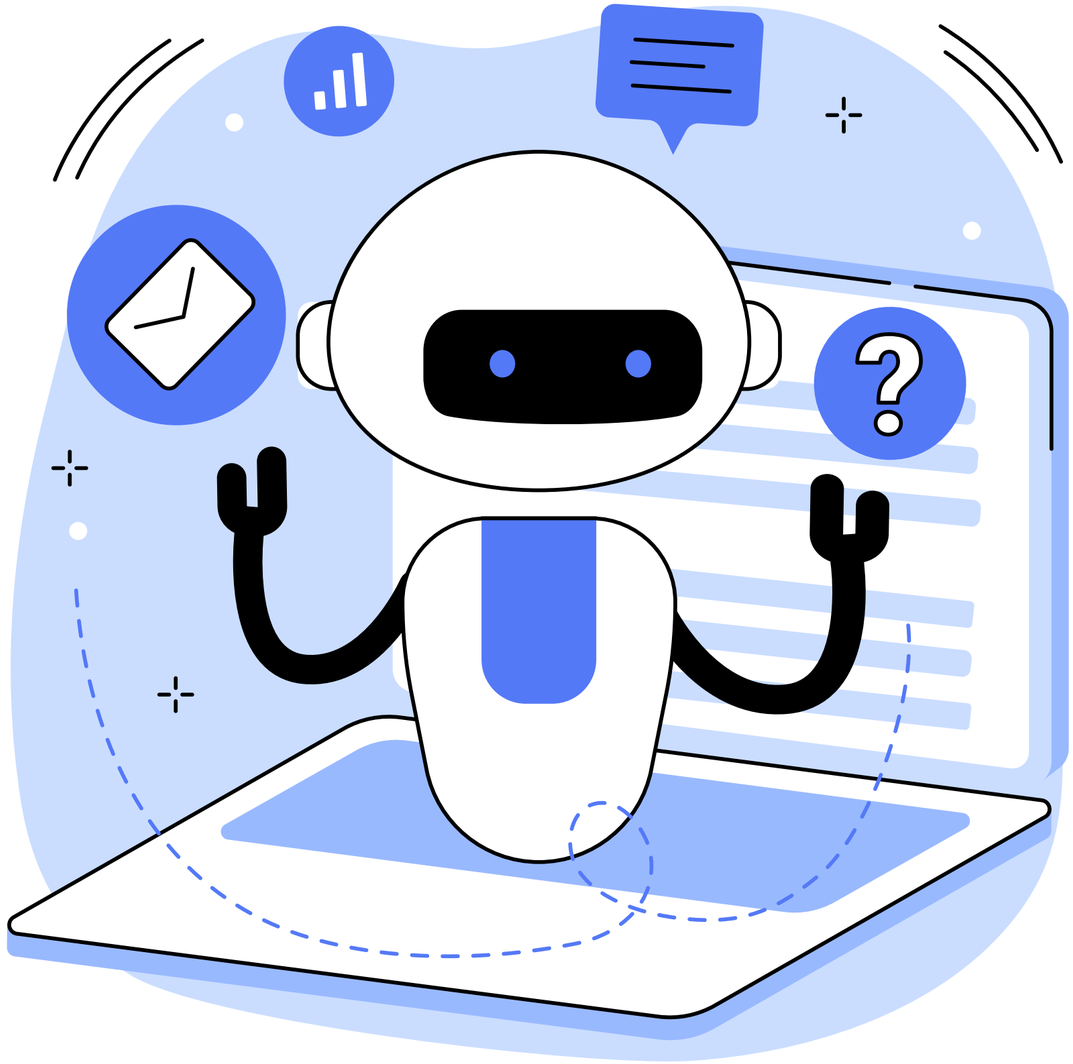
In house price prediction, common features include size, location, number of bedrooms, age, and nearby amenities.



PARAMETERS

Parameters are the learnable settings, mostly weights and biases, that determine an AI model’s behavior.

GPT-3 has 175 billion parameters, each helping the system compose meaningful sentences.

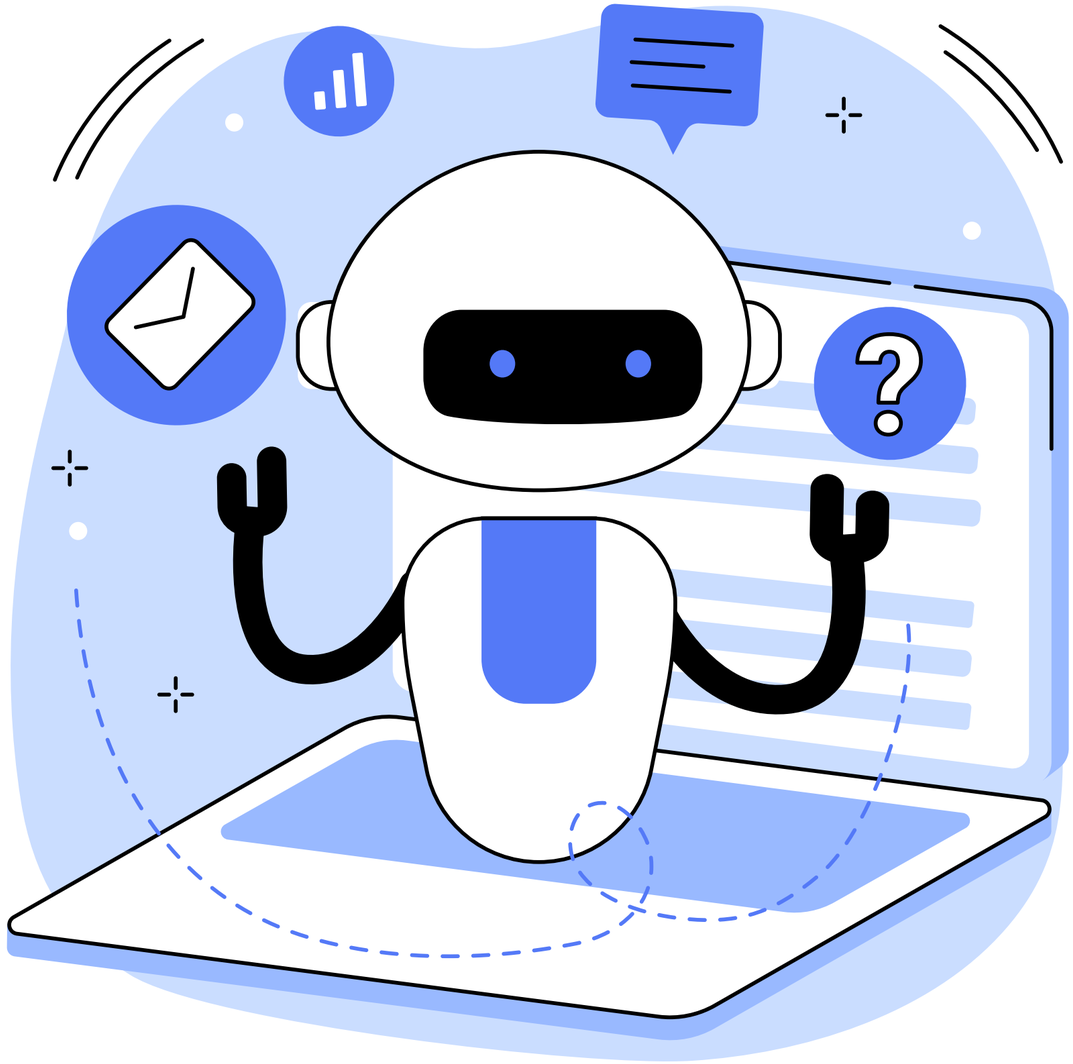


INFERENCE PROVIDERS

Companies who host and serve AI models for fast predictions.

APIs, fast setup, scaling, so you get AI answers without running your own servers.

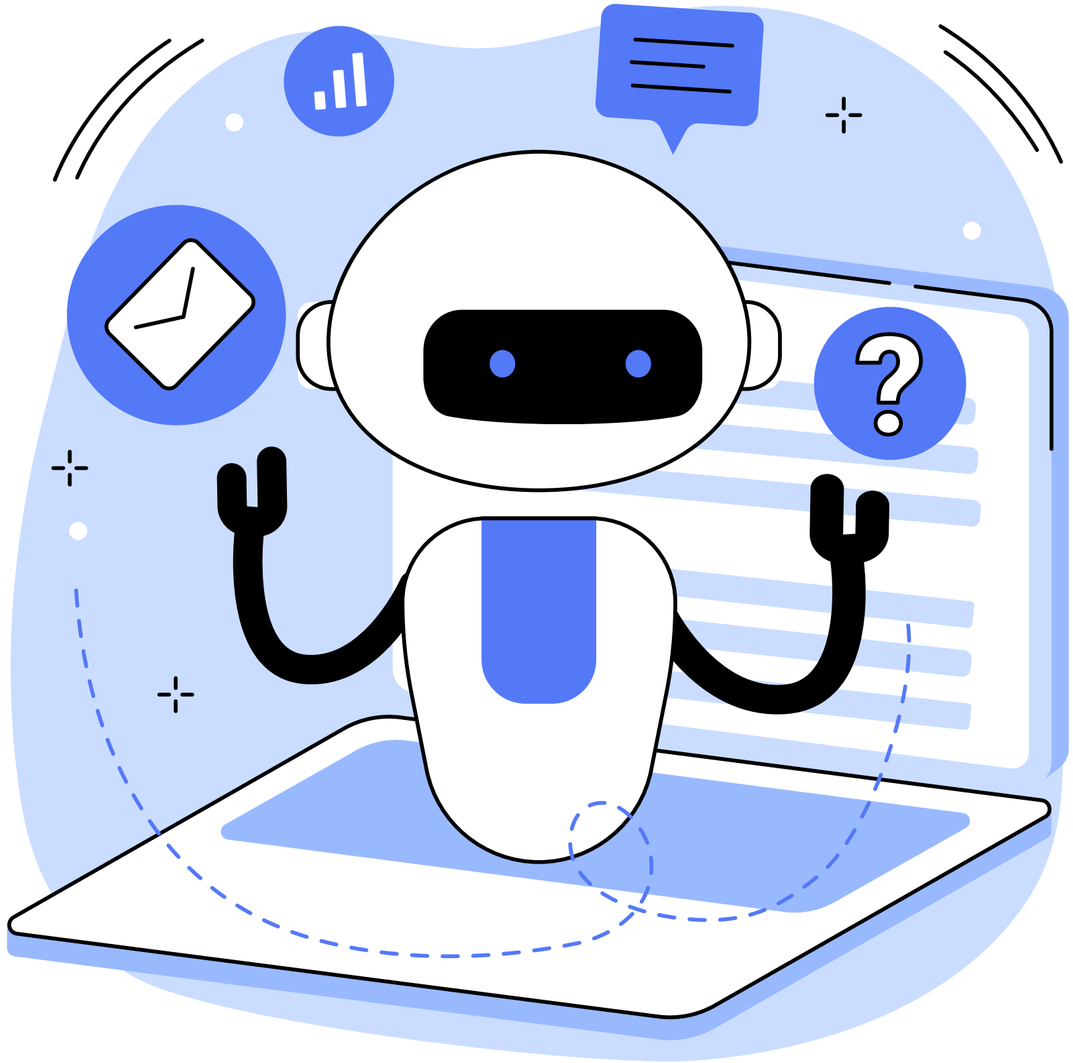
Hugging Face, Together AI, NVIDIA NIMs etc



TRANSFORMER

A transformer is an AI model using self-attention to “read” context, this is the backbone of advanced search, translation, and chatbots.

ChatGPT and Google’s BERT use transformers to understand and answer complex questions.



THANK YOU FOR LISTENING!