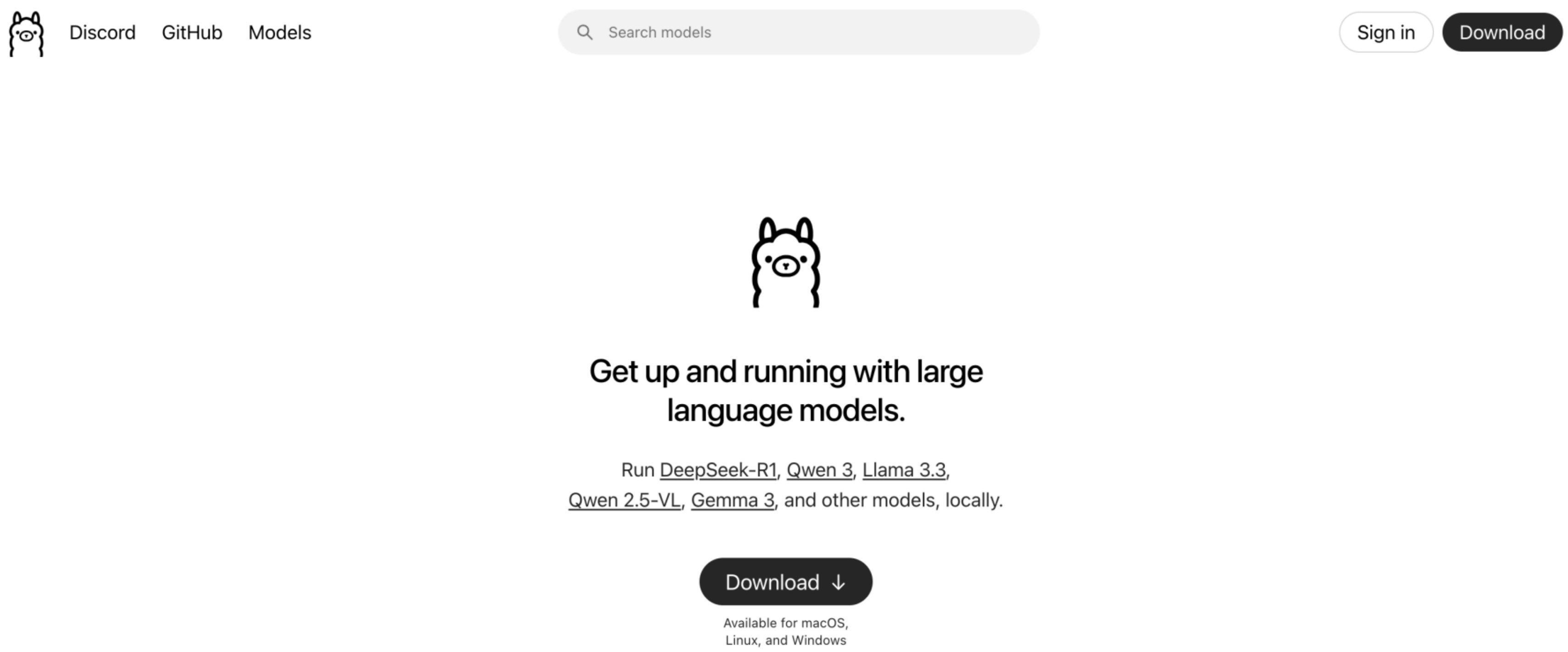
AI Labs

Index

- Download and Run a Model on Local
- What is Model, FM, LLM ?
- What is Foundation Model ?
- How to create it ?
- How it works ?
- What are weights in Model ?
- What are technologies ?
- How/Where to deploy a model ?
- A Cause

Demo

- https://ollama.com/

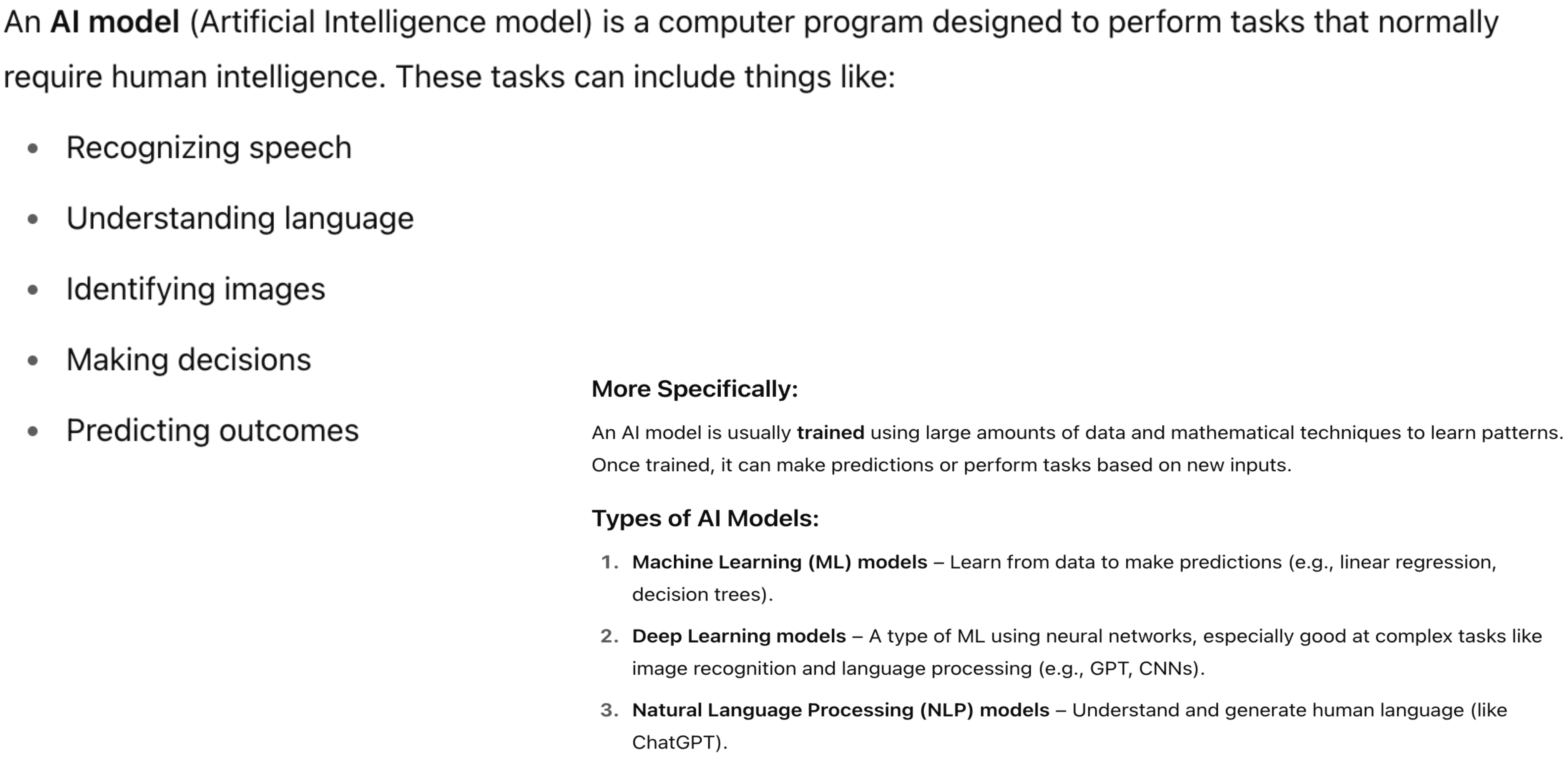


Get up and running with large language models.

Run DeepSeek-R1, Qwen 3, Llama 3.3, Qwen 2.5-VL, Gemma 3, and other models, locally.

Available for macOS, Linux, and Windows

What is AI Model



An **AI model** (Artificial Intelligence model) is a computer program designed to perform tasks that normally require human intelligence. These tasks can include things like:

- Recognizing speech
- Understanding language
- Identifying images
- Making decisions
- Predicting outcomes

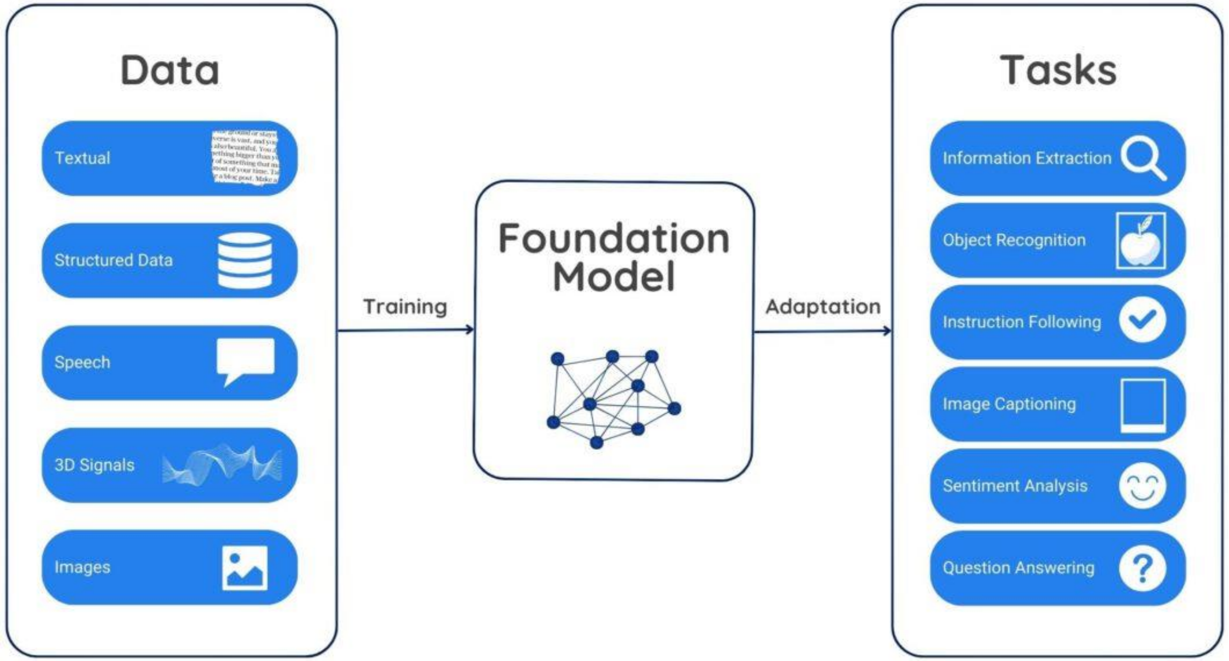
More Specifically:

An AI model is usually **trained** using large amounts of data and mathematical techniques to learn patterns. Once trained, it can make predictions or perform tasks based on new inputs.

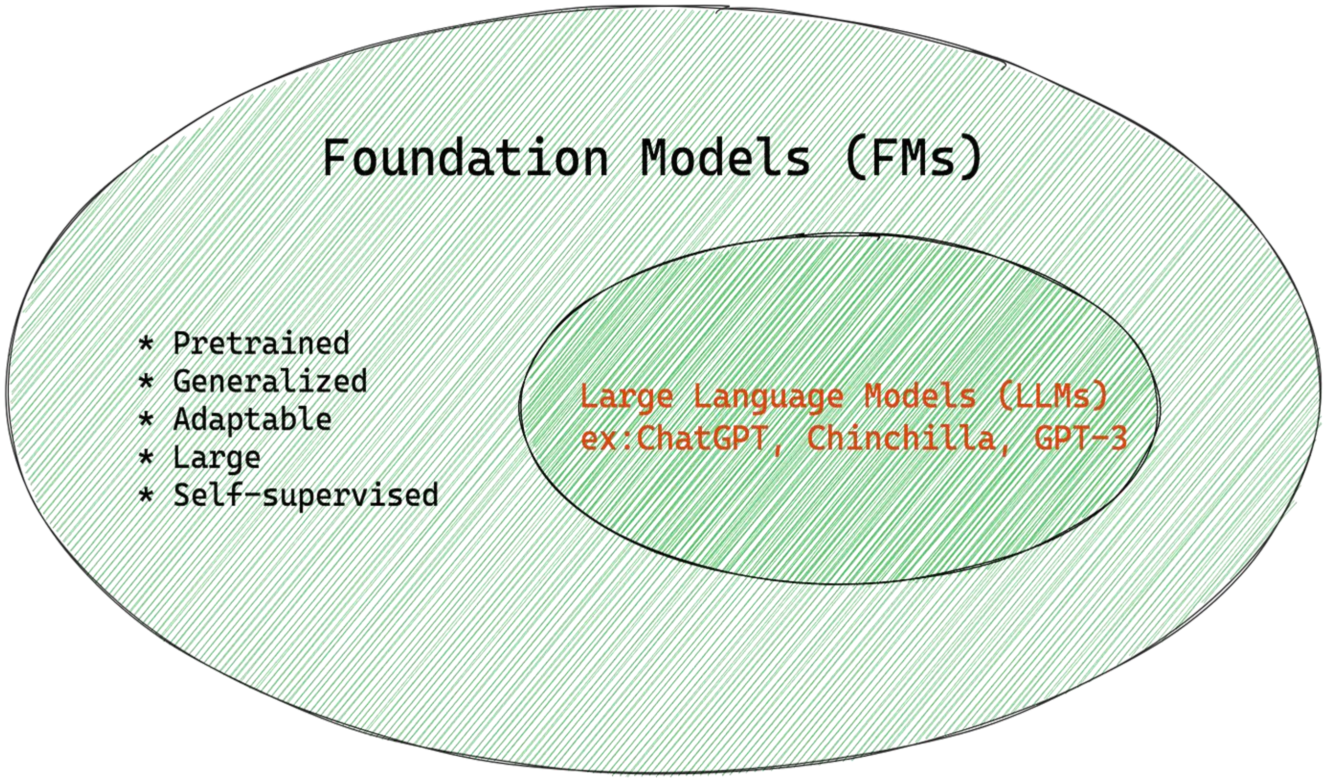
Types of AI Models:

1. **Machine Learning (ML) models** – Learn from data to make predictions (e.g., linear regression, decision trees).
2. **Deep Learning models** – A type of ML using neural networks, especially good at complex tasks like image recognition and language processing (e.g., GPT, CNNs).
3. **Natural Language Processing (NLP) models** – Understand and generate human language (like ChatGPT).

What is a FM


Data
Textual
Structured Data
Speech
3D Signals
Images
Training
Foundation Model
Adaptation
Tasks
Information Extraction
Object Recognition
Instruction Following
Image Captioning
Sentiment Analysis
Question Answering

What is LLM



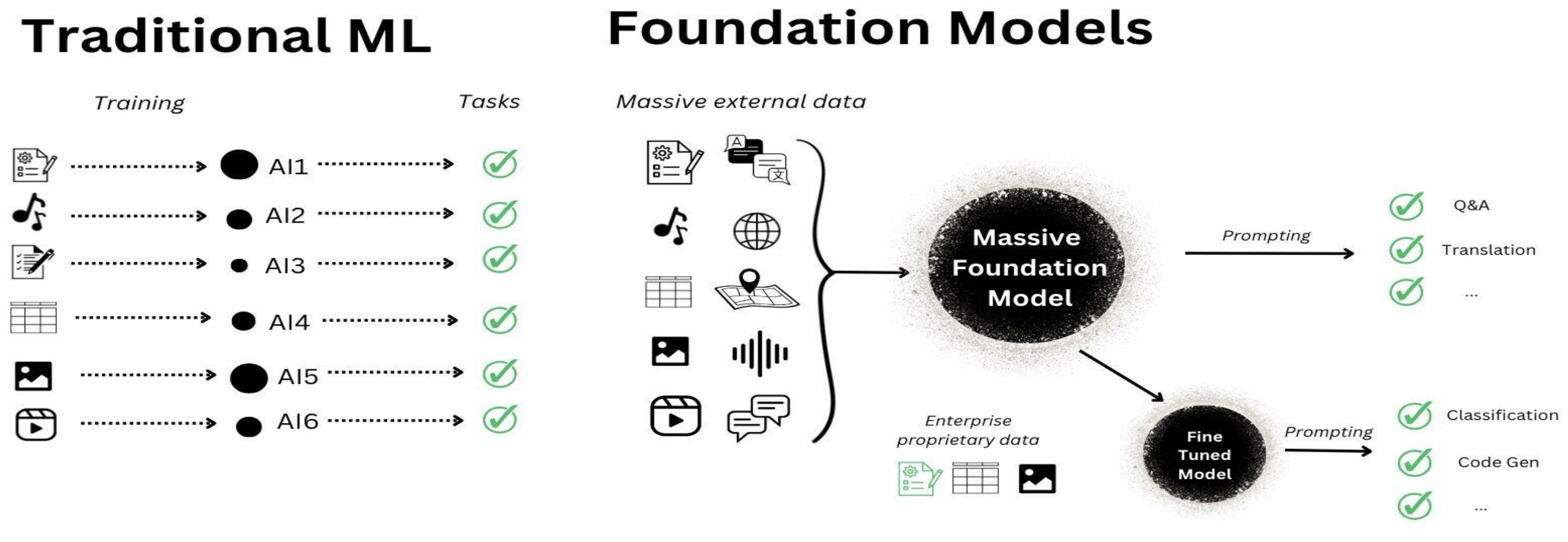
Foundation Models (FMs)

* Pretrained
* Generalized
* Adaptable
* Large
* Self-supervised

Large Language Models (LLMs)
ex:ChatGPT, Chinchilla, GPT-3

FMs are models trained on broad data (using self-supervision at scale)
that can be adapted to a wide range of downstream tasks.
https://hai.stanford.edu/news/reflections-foundation-models

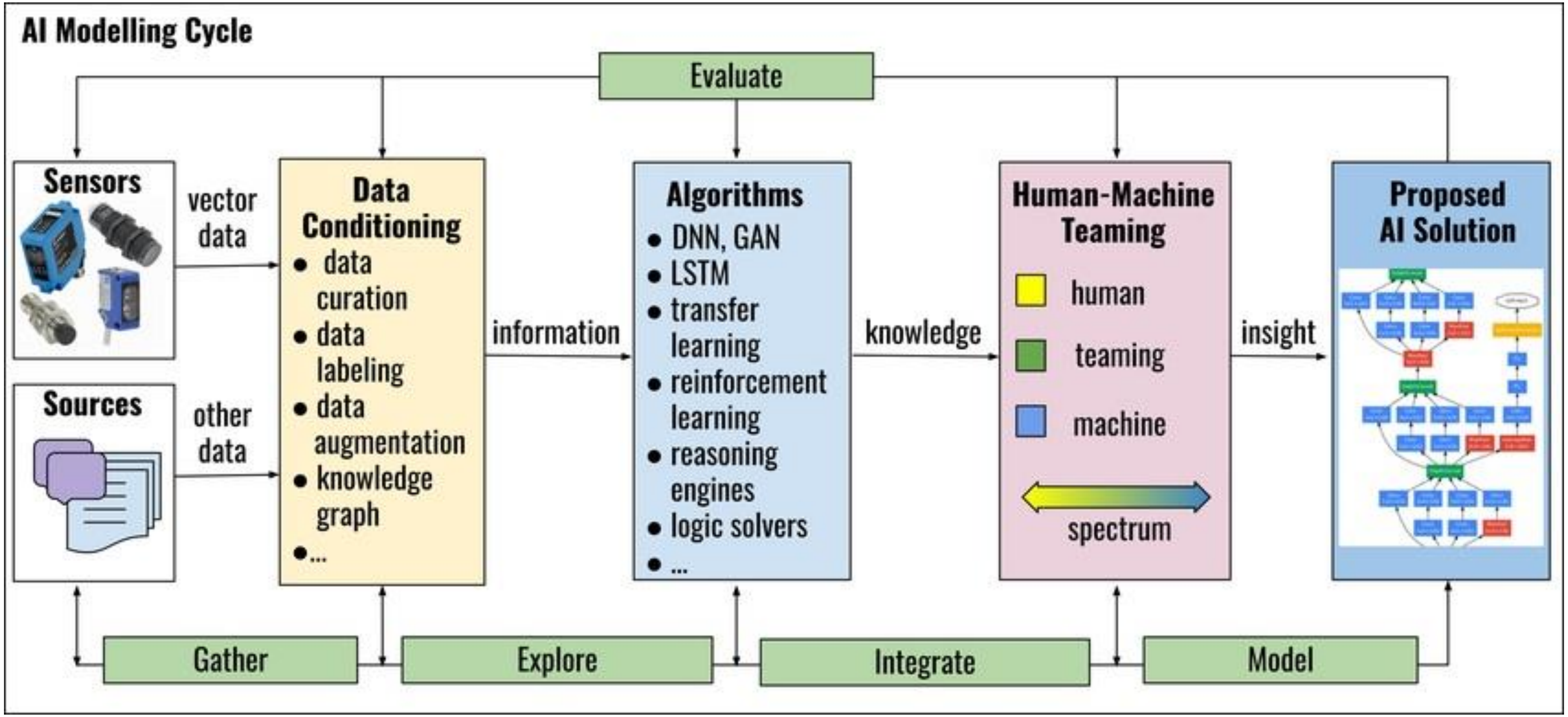
Tell me more



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

Step to make a model



AI Modelling Cycle

Evaluate

Sensors

vector data

Sources

other data

Data Conditioning

- data curation
- data labeling
- data augmentation
- knowledge graph
- ...

information

Algorithms

- DNN, GAN
- LSTM
- transfer learning
- reinforcement learning
- reasoning engines
- logic solvers
- ...

knowledge

Human-Machine Teaming

🟨 human

🟩 teaming

🟦 machine

spectrum

insight

Proposed AI Solution

Gather

Explore

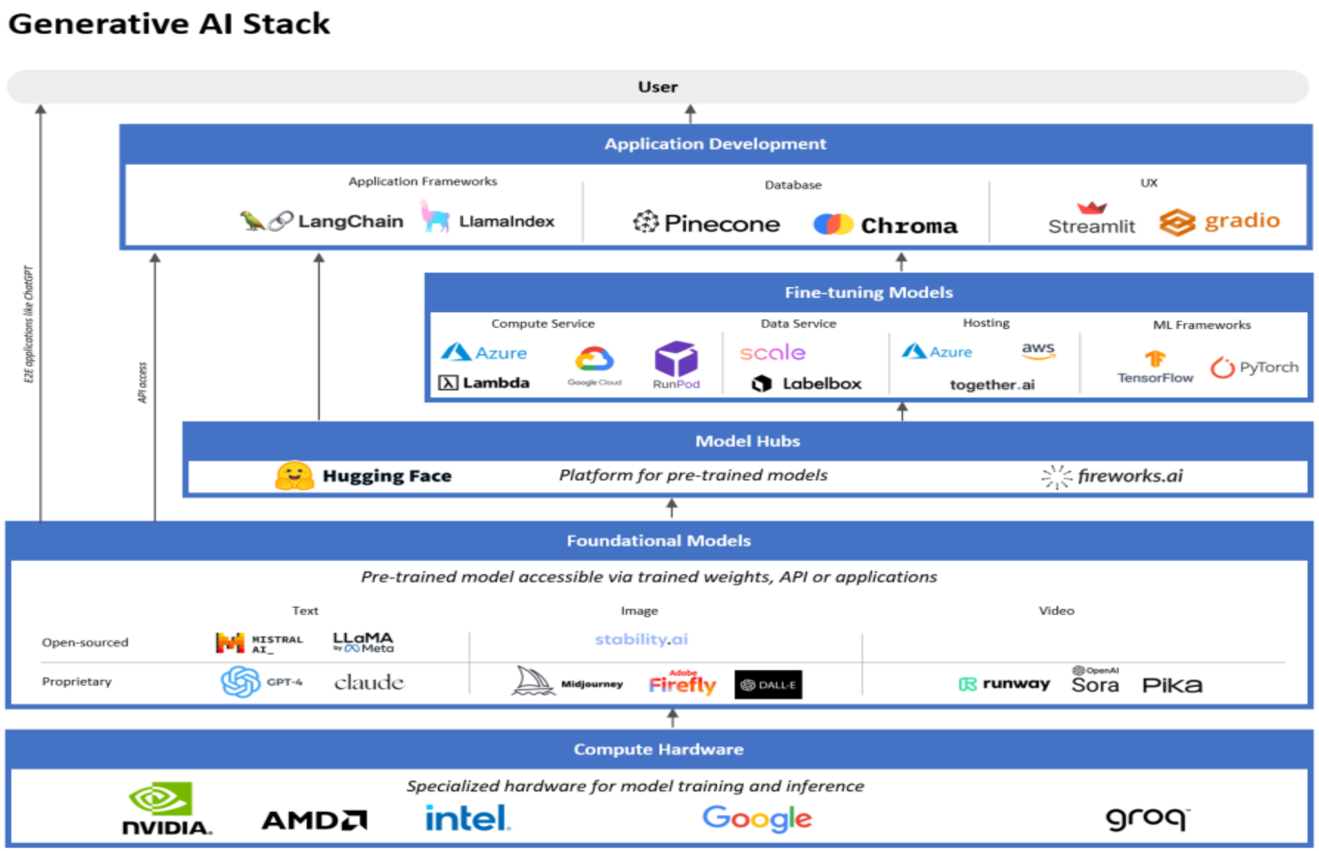
Integrate

Model

What is Weight in LLM

- https://www.youtube.com/watch?v=LPZh9BOjkQs

Technologies



UI/UX

- Stremlit, Chainlit
- Angular, React etc

Programming Languages

- Python – Popular.
- R, Java, C++ etc

Libraries & Frameworks

- TensorFlow
- PyTorch
- Scikit-learn
- Keras etc

Data Handling Tools

- Pandas
- NumPy
- SQL / NoSQL/Vector databases

Compute Platforms

- GPUs .
- TPUs.

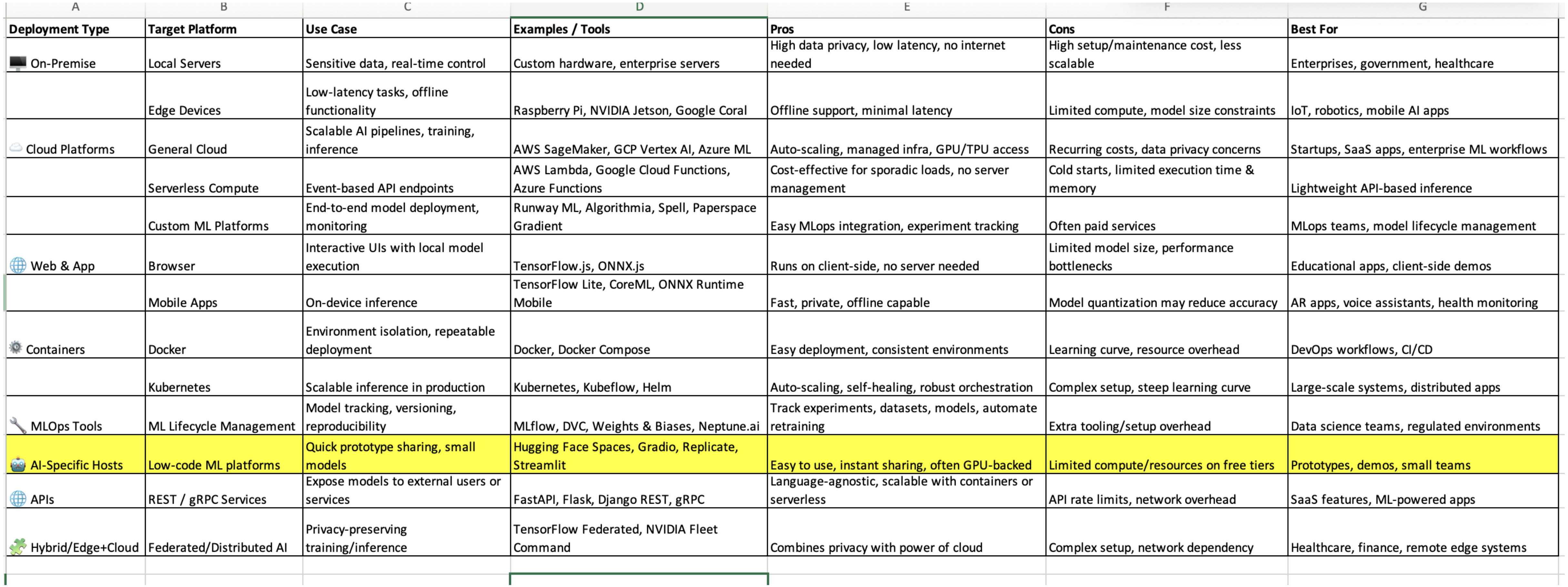
Cloud Platforms

- AWS, GCP, Azure etc.

Model Training & Experiment Tracking

- Jupyter Notebooks
- MLflow / Weights & Biases
- Docker / Kubernetes

Where to deploy and get more details



A	B	C	D	E	F	G
Deployment Type	**Target Platform**	**Use Case**	**Examples / Tools**	**Pros**	**Cons**	**Best For**
🖥️ On-Premise	Local Servers	Sensitive data, real-time control	Custom hardware, enterprise servers	High data privacy, low latency, no internet needed	High setup/maintenance cost, less scalable	Enterprises, government, healthcare
	Edge Devices	Low-latency tasks, offline functionality	Raspberry Pi, NVIDIA Jetson, Google Coral	Offline support, minimal latency	Limited compute, model size constraints	IoT, robotics, mobile AI apps
☁️ Cloud Platforms	General Cloud	Scalable AI pipelines, training, inference	AWS SageMaker, GCP Vertex AI, Azure ML	Auto-scaling, managed infra, GPU/TPU access	Recurring costs, data privacy concerns	Startups, SaaS apps, enterprise ML workflows
	Serverless Compute	Event-based API endpoints	AWS Lambda, Google Cloud Functions, Azure Functions	Cost-effective for sporadic loads, no server management	Cold starts, limited execution time & memory	Lightweight API-based inference
	Custom ML Platforms	End-to-end model deployment, monitoring	Runway ML, Algorithmia, Spell, Paperspace Gradient	Easy MLops integration, experiment tracking	Often paid services	MLops teams, model lifecycle management
🌐 Web & App	Browser	Interactive UIs with local model execution	TensorFlow.js, ONNX.js	Runs on client-side, no server needed	Limited model size, performance bottlenecks	Educational apps, client-side demos
	Mobile Apps	On-device inference	TensorFlow Lite, CoreML, ONNX Runtime Mobile	Fast, private, offline capable	Model quantization may reduce accuracy	AR apps, voice assistants, health monitoring
⚙️ Containers	Docker	Environment isolation, repeatable deployment	Docker, Docker Compose	Easy deployment, consistent environments	Learning curve, resource overhead	DevOps workflows, CI/CD
	Kubernetes	Scalable inference in production	Kubernetes, Kubeflow, Helm	Auto-scaling, self-healing, robust orchestration	Complex setup, steep learning curve	Large-scale systems, distributed apps
🔧 MLOps Tools	ML Lifecycle Management	Model tracking, versioning, reproducibility	MLflow, DVC, Weights & Biases, Neptune.ai	Track experiments, datasets, models, automate retraining	Extra tooling/setup overhead	Data science teams, regulated environments
🤖 AI-Specific Hosts	Low-code ML platforms	Quick prototype sharing, small models	Hugging Face Spaces, Gradio, Replicate, Streamlit	Easy to use, instant sharing, often GPU-backed	Limited compute/resources on free tiers	Prototypes, demos, small teams
🌐 APIs	REST / gRPC Services	Expose models to external users or services	FastAPI, Flask, Django REST, gRPC	Language-agnostic, scalable with containers or serverless	API rate limits, network overhead	SaaS features, ML-powered apps
🧩 Hybrid/Edge+Cloud	Federated/Distributed AI	Privacy-preserving training/inference	TensorFlow Federated, NVIDIA Fleet Command	Combines privacy with power of cloud	Complex setup, network dependency	Healthcare, finance, remote edge systems

Donate for India

- https://indianarmy.nic.in/about/dg-1b-ii--departments-dgafms-directorates-and-branches/army-welfare-funds-for-donation---contributions-cw-directorate-directorates-and-branches
- https://www.pmindia.gov.in/en/national-defence-fund/
- https://ndf.gov.in/en/online-donation

End