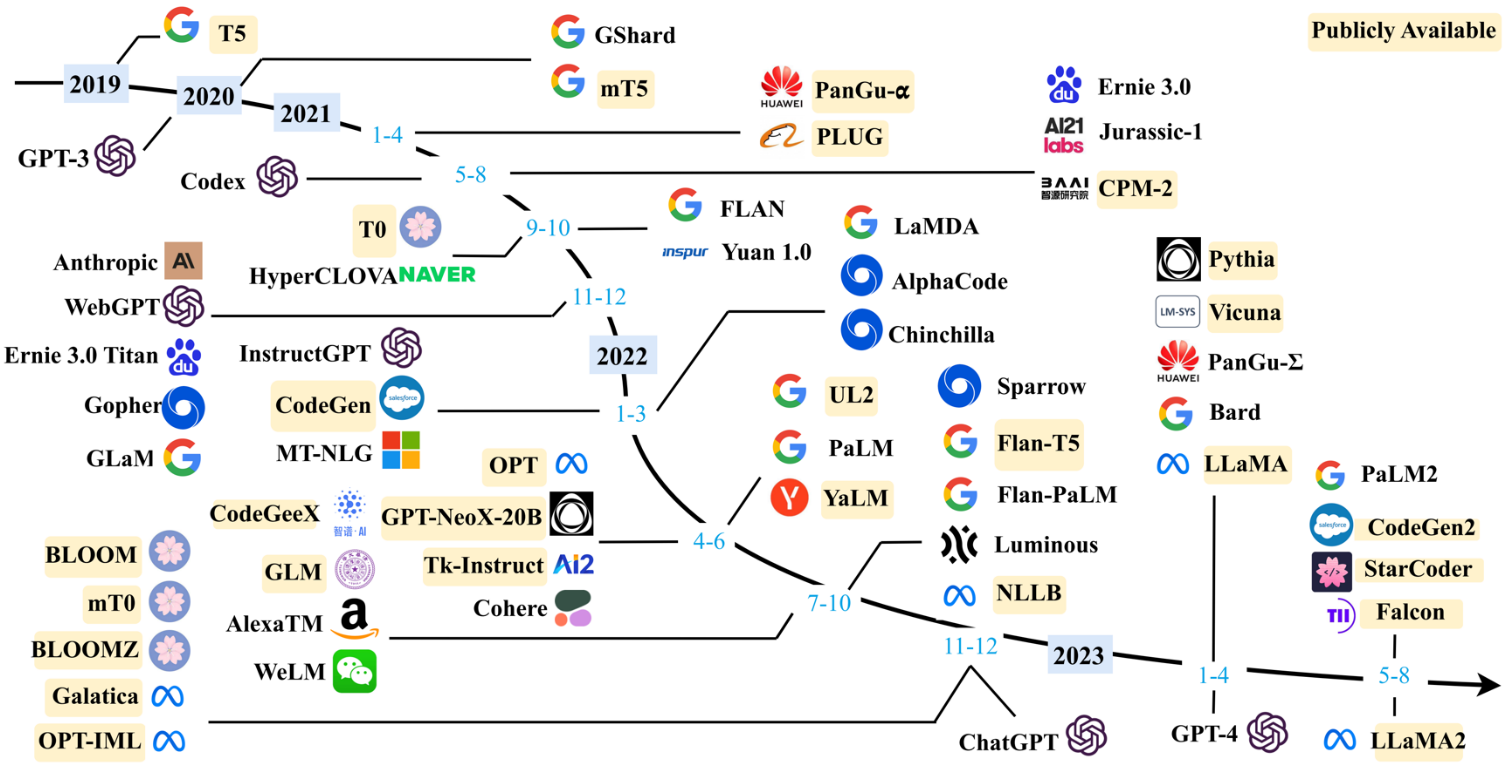
What’s out there?

ACCELERATE PROGRAMME
FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Timeline of >10B parameter models



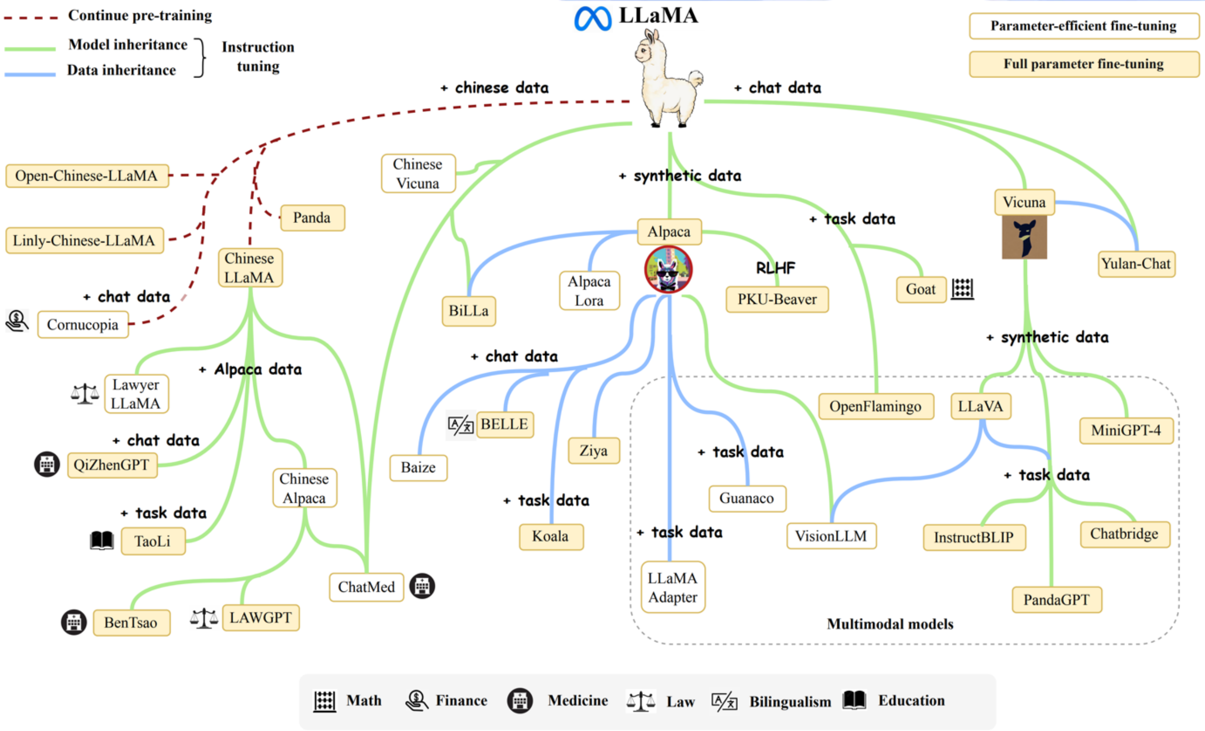
ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

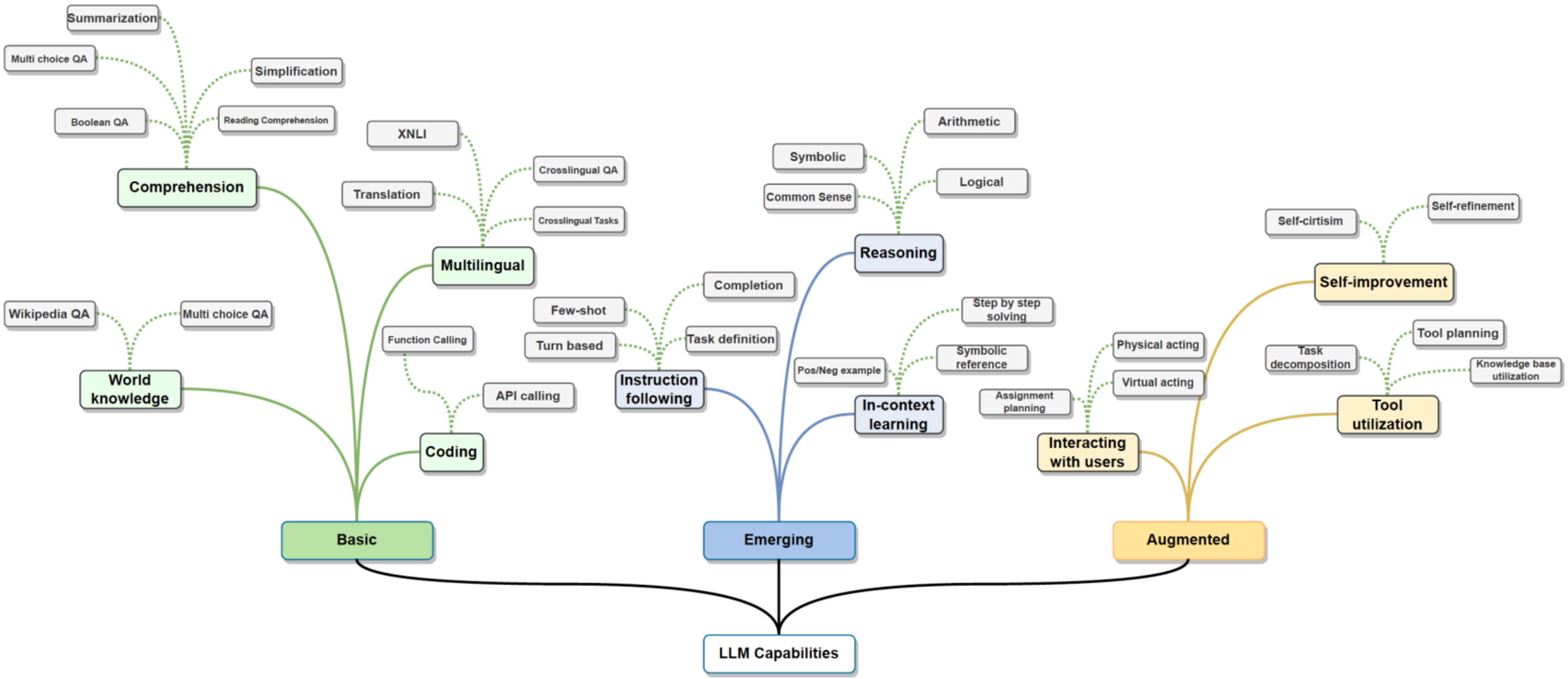
How can we categorize LLMs?

- Architecture
- Objective or use-case
- Scale
- Modalities
- Language
- Availability

Sometimes, one base model can evolve into all of the above!

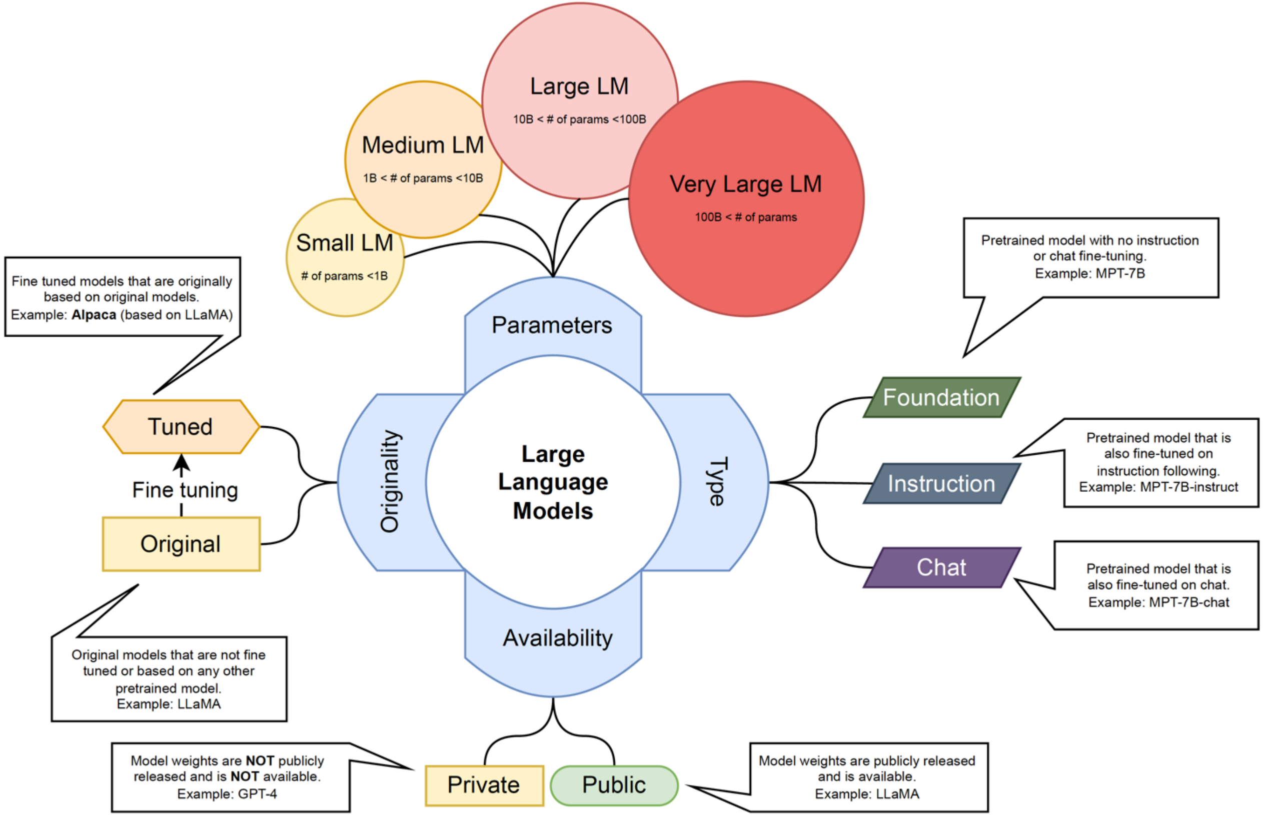


How can we categorize LLMs?



- LLM Capabilities
  - Basic
    - Comprehension
      - Summarization
      - Multi choice QA
      - Simplification
      - Boolean QA
      - Reading Comprehension
    - Multilingual
      - XNLI
      - Crosslingual QA
      - Translation
      - Crosslingual Tasks
    - World knowledge
      - Wikipedia QA
      - Multi choice QA
    - Coding
      - Function Calling
      - API calling
  - Emerging
    - Instruction following
      - Completion
      - Few-shot
      - Task definition
      - Turn based
    - Reasoning
      - Arithmetic
      - Symbolic
      - Logical
      - Common Sense
    - In-context learning
      - Step by step solving
      - Symbolic reference
      - Pos/Neg example
  - Augmented
    - Self-improvement
      - Self-cirtisim
      - Self-refinement
    - Interacting with users
      - Physical acting
      - Virtual acting
      - Assignment planning
    - Tool utilization
      - Tool planning
      - Task decomposition
      - Knowledge base utilization

How can we categorize LLMs?



Large Language Models

Parameters

Small LM
# of params <1B

Medium LM
1B < # of params <10B

Large LM
10B < # of params <100B

Very Large LM
100B < # of params

Type

Foundation

Pretrained model with no instruction or chat fine-tuning.
Example: MPT-7B

Instruction

Pretrained model that is also fine-tuned on instruction following.
Example: MPT-7B-instruct

Chat

Pretrained model that is also fine-tuned on chat.
Example: MPT-7B-chat

Originality

Tuned

Fine tuning

Original

Fine tuned models that are originally based on original models.
Example: **Alpaca** (based on LLaMA)

Original models that are not fine tuned or based on any other pretrained model.
Example: LLaMA

Availability

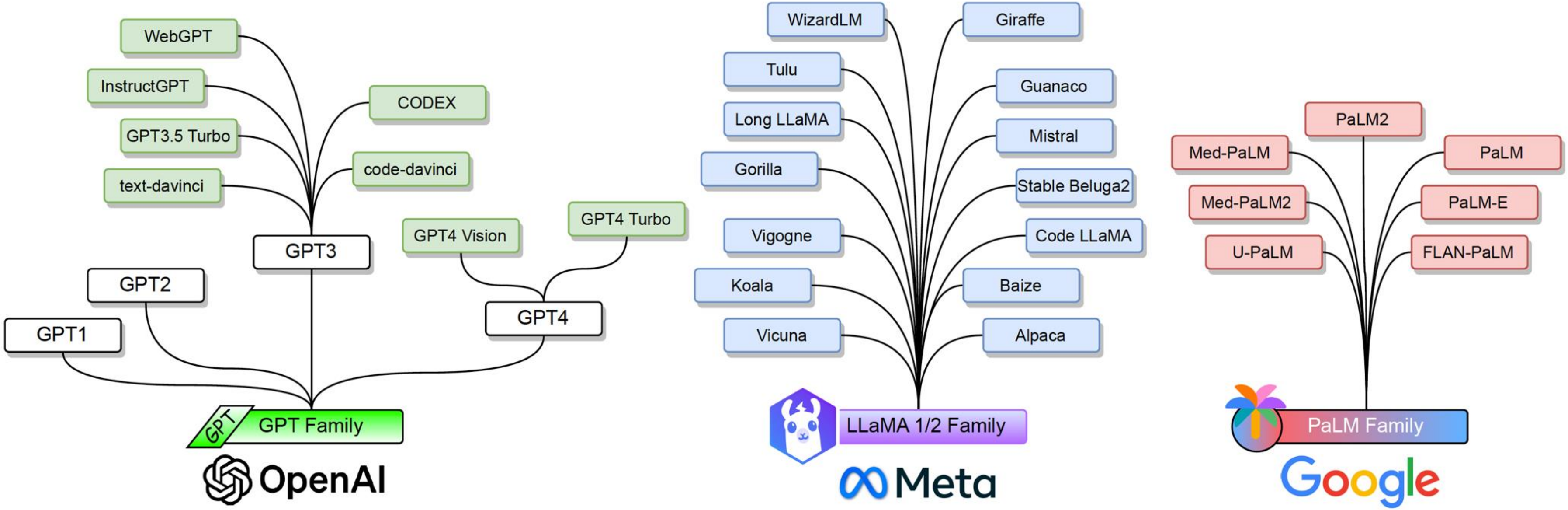
Private

Public

Model weights are **NOT** publicly released and is **NOT** available.
Example: GPT-4

Model weights are publicly released and is available.
Example: LLaMA

Popular LLM Families

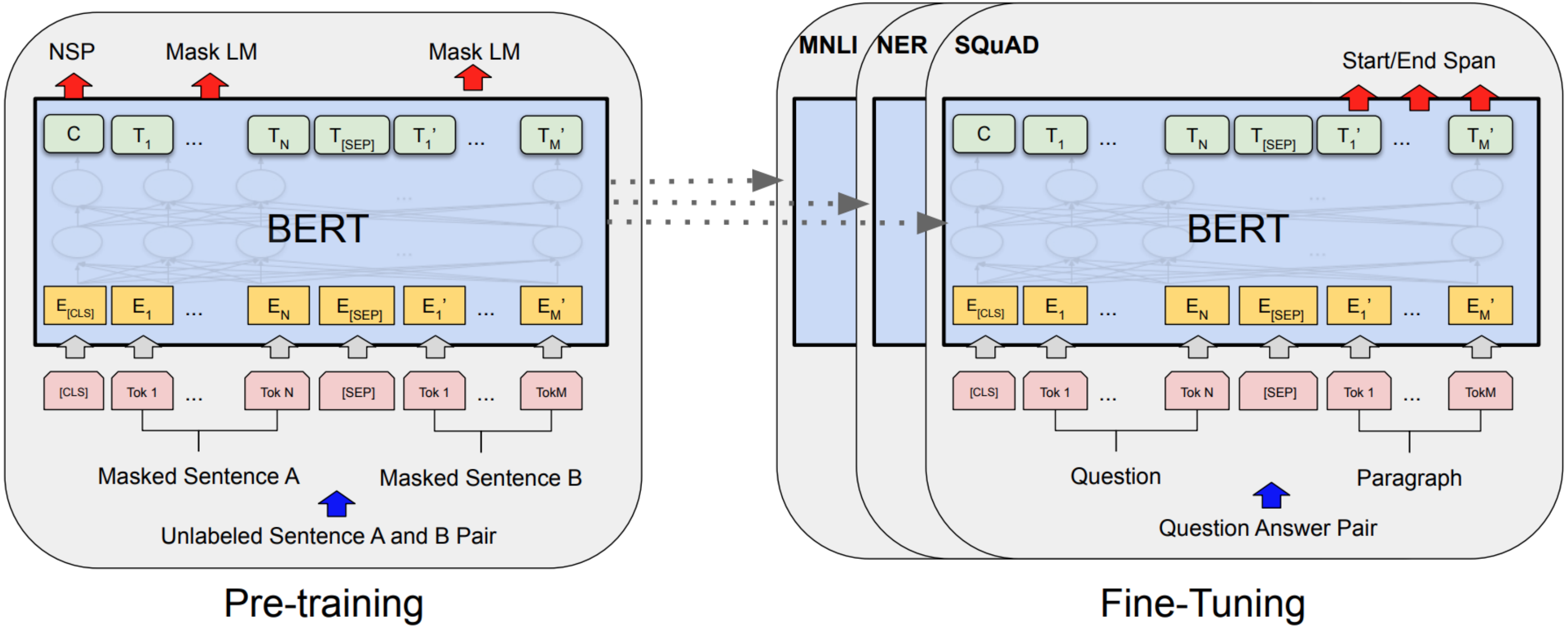


APSCI ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

BERT

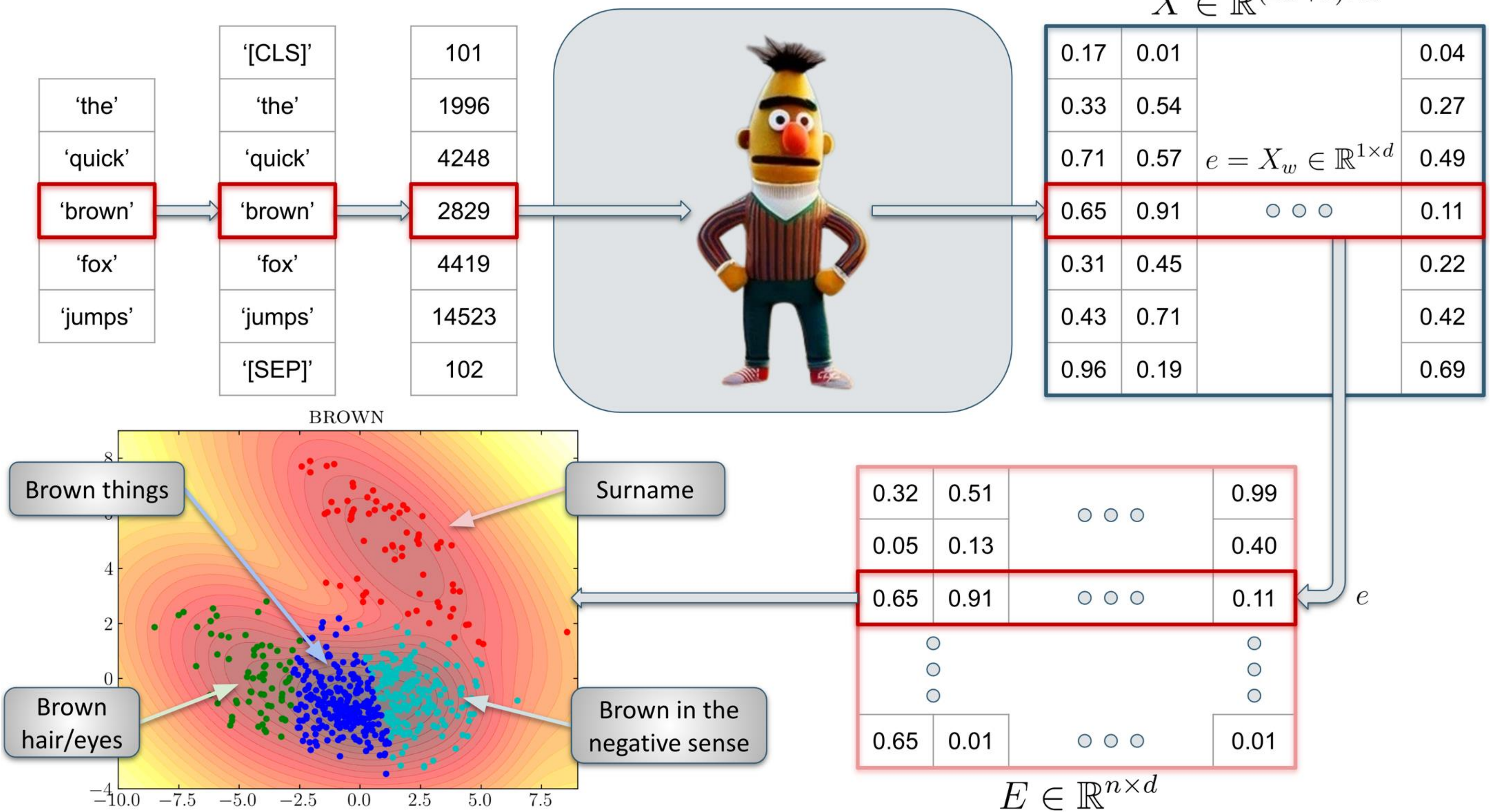
Bidirectional Encoder Representations from Transformers



Pre-training

Fine-Tuning

Haket and Daniels, *“BERT’s Conceptual Cartography: Mapping the Landscapes of Meaning.” EMNLP2024, under review*



	‘[CLS]’	101
‘the’	‘the’	1996
‘quick’	‘quick’	4248
‘brown’	‘brown’	2829
‘fox’	‘fox’	4419
‘jumps’	‘jumps’	14523
	‘[SEP]’	102

$X \in \mathbb{R}^{(2C+1)\times d}$

0.17	0.01		0.04
0.33	0.54		0.27
0.71	0.57	$e = X_w \in \mathbb{R}^{1\times d}$	0.49
0.65	0.91	○ ○ ○	0.11
0.31	0.45		0.22
0.43	0.71		0.42
0.96	0.19		0.69

e

0.32	0.51	○ ○ ○	0.99
0.05	0.13		0.40
0.65	0.91	○ ○ ○	0.11
○ ○ ○			○ ○ ○
0.65	0.01	○ ○ ○	0.01

$E \in \mathbb{R}^{n\times d}$

Brown things

Surname

Brown hair/eyes

Brown in the negative sense

BERT

BERT is a very versatile family

DistilBERT

CamemBERT

ERNIE

XML-RoBERTa

RoBERTa

ALBERT

BART

Q-BERT

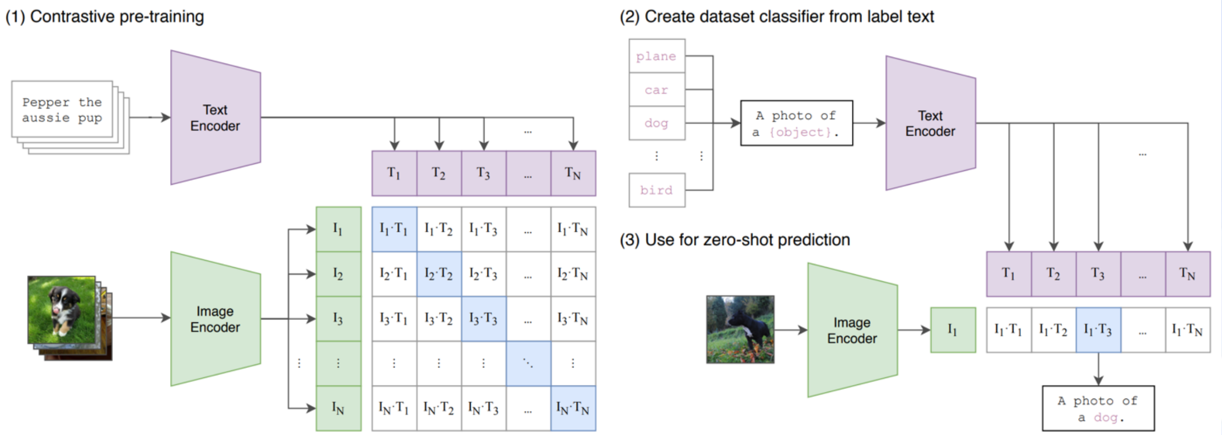
APSCI ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

CLIP

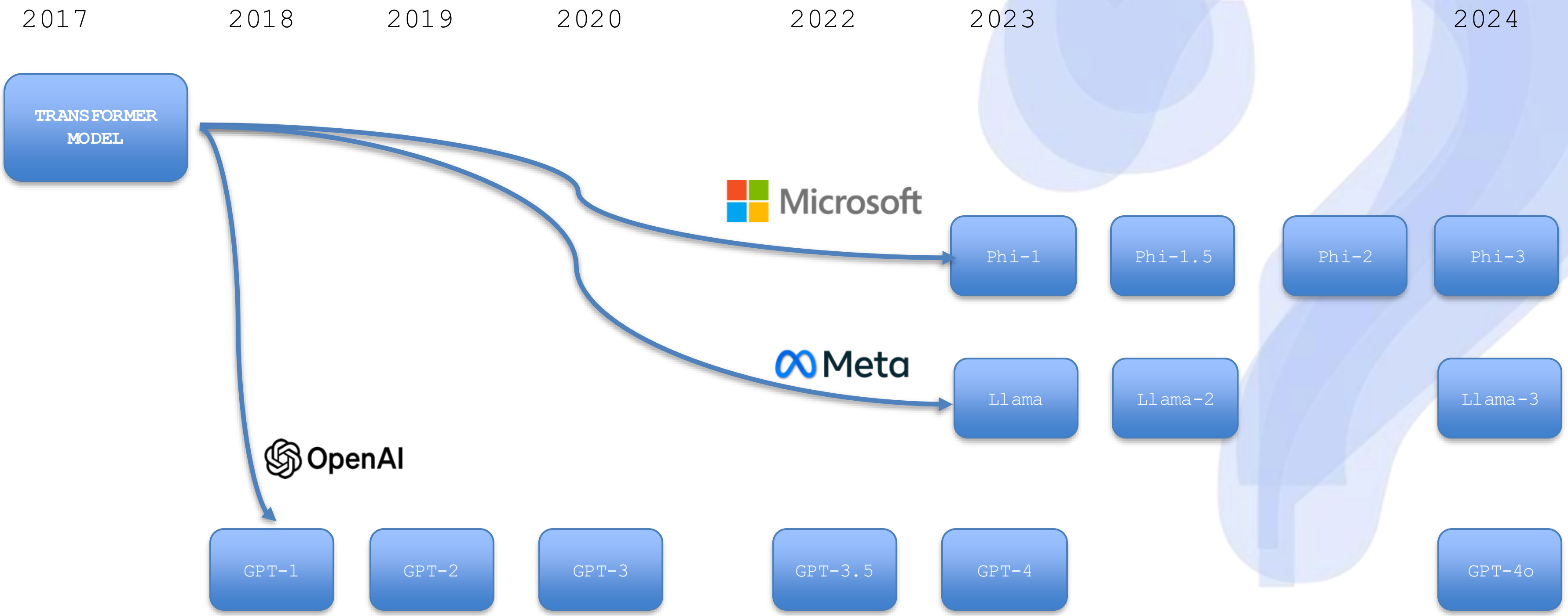
Contrastive Language-Image Pre-training

- *Learning Transferable Visual Models for Natural Language Supervision*, Radford et al, 2021, ~17k citations
- Takes in images and texts and connects them together


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
⋮ ⋮ ⋮ ⋱ ⋮
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
T1 T2 T3 ... TN
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

How are companies updating their models?



2017 2018 2019 2020 2022 2023 2024

TRANSFORMER MODEL

Microsoft

Phi-1 Phi-1.5 Phi-2 Phi-3

Meta

Llama Llama-2 Llama-3

OpenAI

GPT-1 GPT-2 GPT-3 GPT-3.5 GPT-4 GPT-4o

APSCI ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

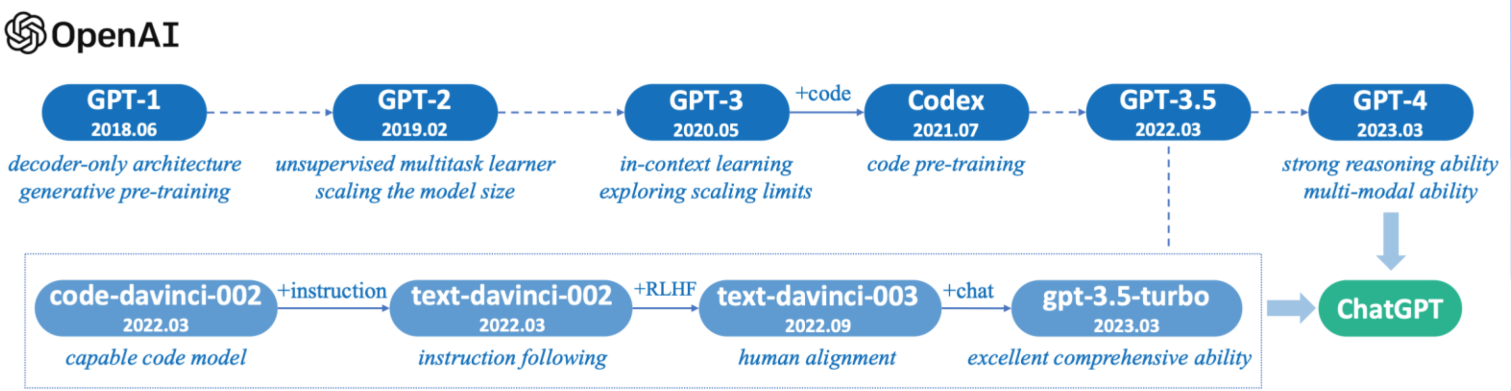
What does OpenAI offer?

Browser-based service:

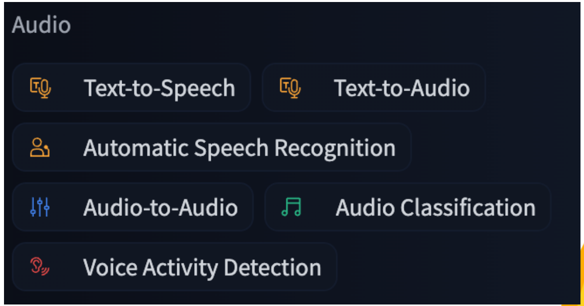
- ChatGPT and ChatGPT Plus
- DALL-E
- GPTs

API-based service:

- Embeddings
- Fine-tuning
- Image generation
- Vision
- Text-to-speech (and speech-to-text)

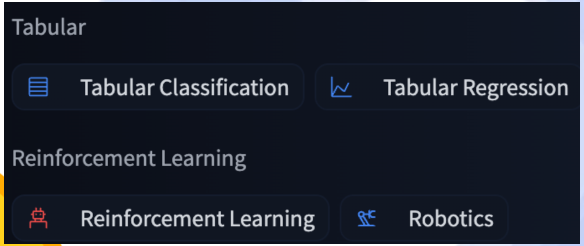


Open source LLMs



Audio

- Text-to-Speech
- Text-to-Audio
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

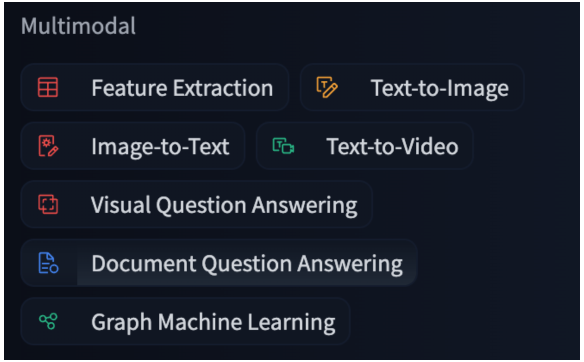


Tabular

- Tabular Classification
- Tabular Regression

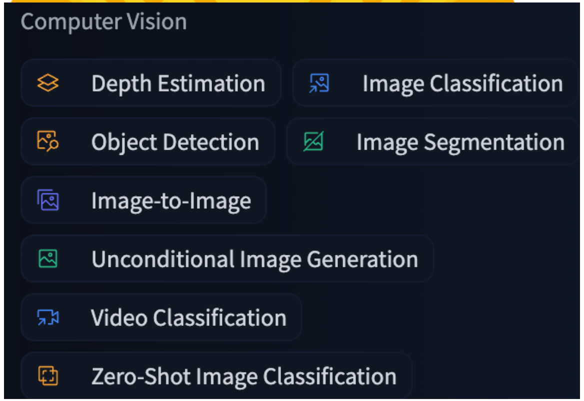
Reinforcement Learning

- Reinforcement Learning
- Robotics



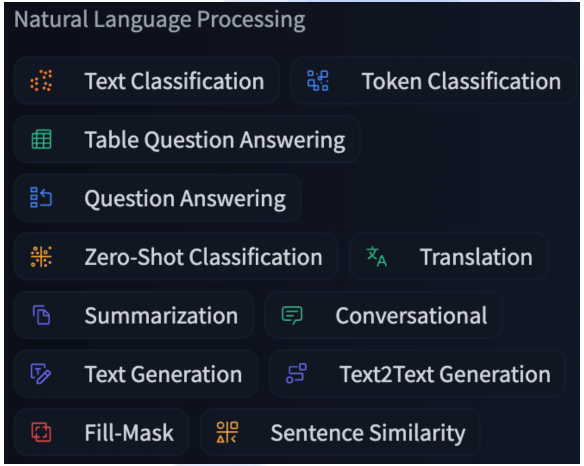
Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning



Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification



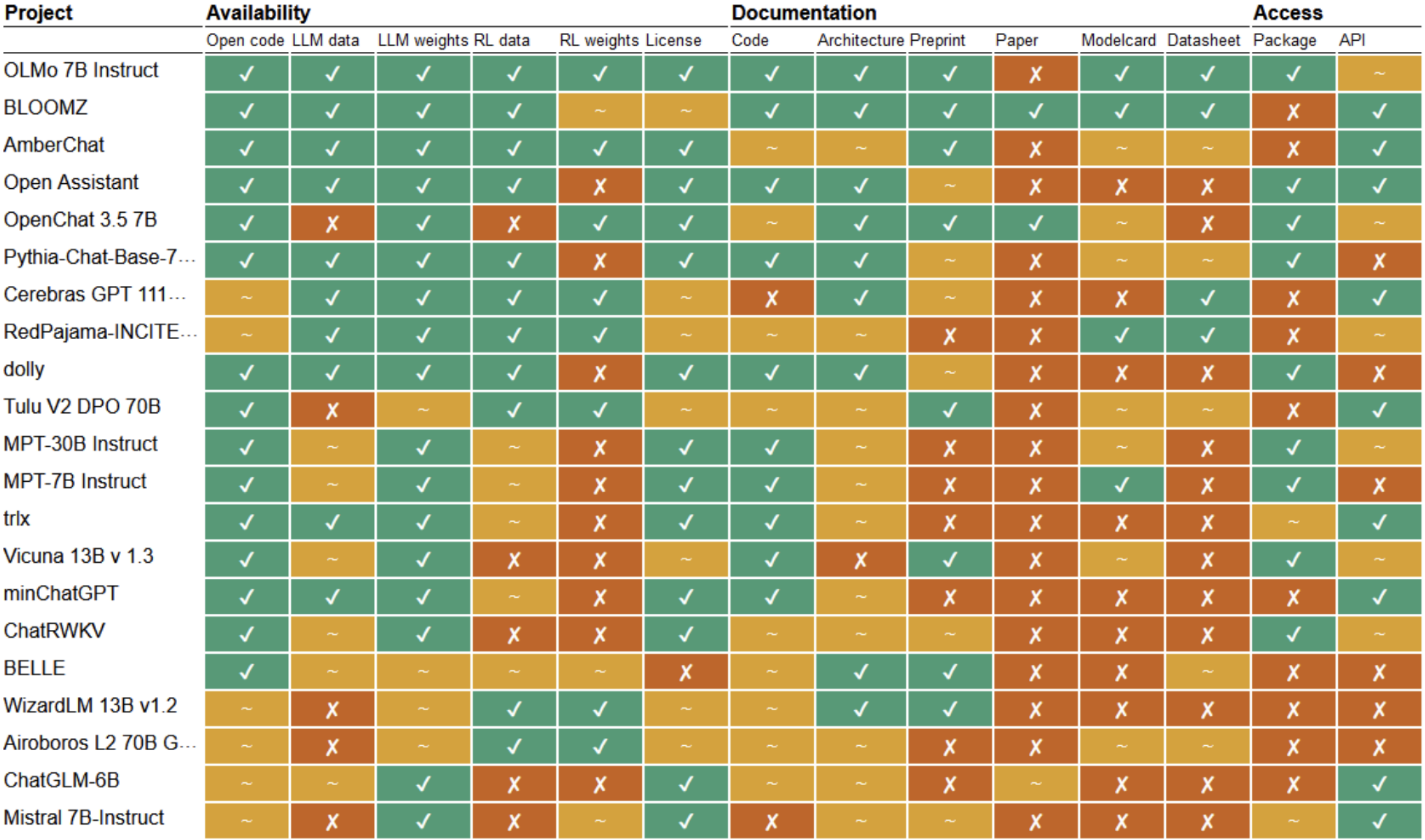
Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

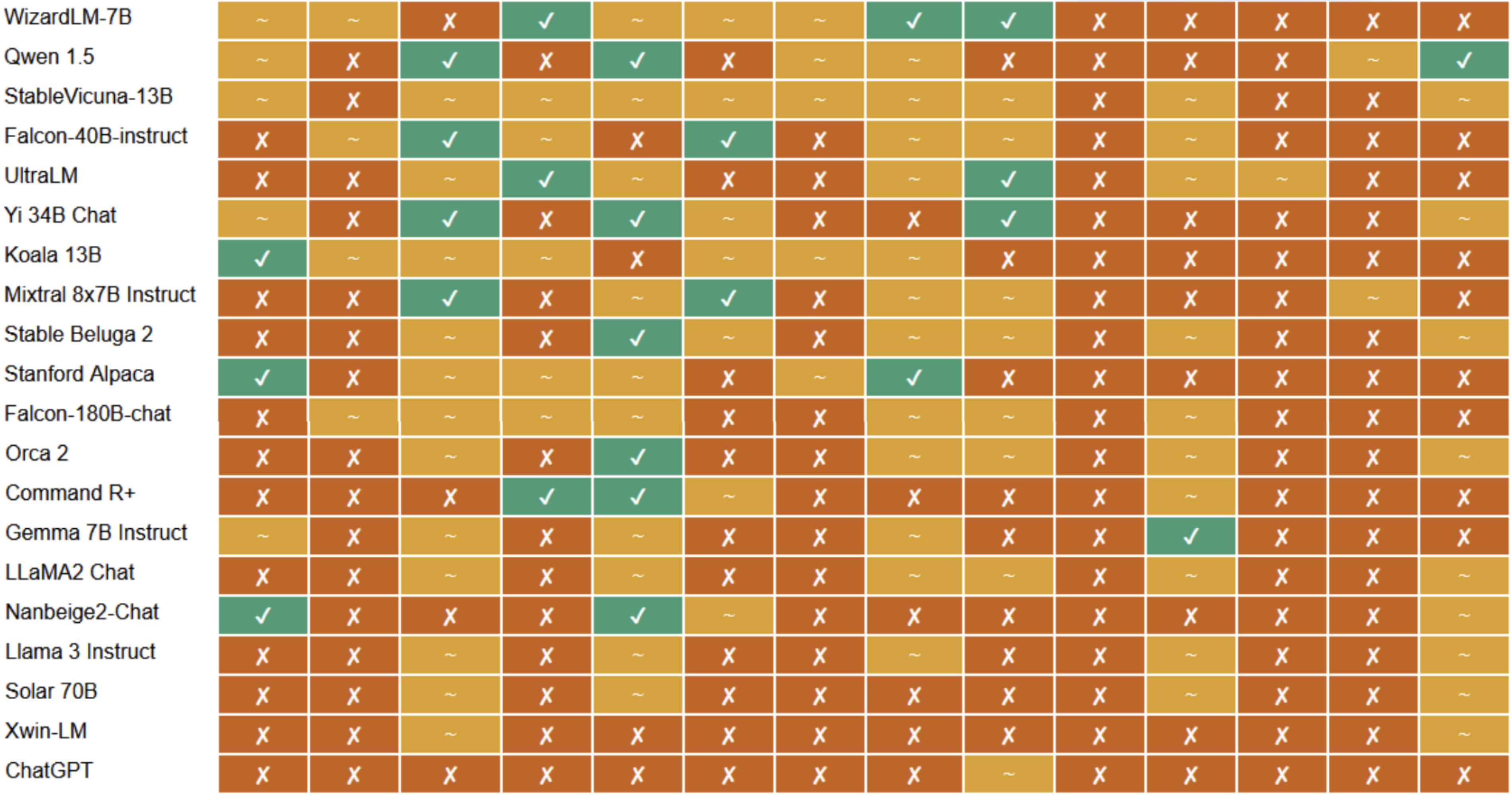
But is it really open source…?

Here are some things to look out for:

- Weights
- Training data (and RL data)
- Training code
- License
- Architecture
- Preprint/paper
- Modelcard
- Package
- API



Project	Availability						Documentation						Access	
	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
OLMo 7B Instruct	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	~
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
AmberChat	✓	✓	✓	✓	✓	✓	~	~	✓	✗	~	~	✗	✓
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✓
OpenChat 3.5 7B	✓	✗	✓	✗	✓	✓	~	✓	✓	✓	~	✗	✓	~
Pythia-Chat-Base-7…	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	~	~	✓	✗
Cerebras GPT 111…	~	✓	✓	✓	✓	~	✗	✓	~	✗	✗	✓	✗	✓
RedPajama-INCITE…	~	✓	✓	✓	✓	~	~	~	✗	✗	✓	✓	✗	~
dolly	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✗
Tulu V2 DPO 70B	✓	✗	~	✓	✓	~	~	~	✓	✗	~	~	✗	✓
MPT-30B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	~	✗	✓	~
MPT-7B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	✓	✗	✓	✗
trlx	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	~	✓
Vicuna 13B v 1.3	✓	~	✓	✗	✗	~	✓	✗	✓	✗	~	✗	✓	~
minChatGPT	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	✗	✓
ChatRWKV	✓	~	✓	✗	✗	✓	~	~	~	✗	✗	✗	✓	~
BELLE	✓	~	~	~	~	✗	~	✓	✓	✗	✗	~	✗	✗
WizardLM 13B v1.2	~	✗	~	✓	✓	~	~	✓	✓	✗	✗	✗	✗	✗
Airoboros L2 70B G…	~	✗	~	✓	✓	~	~	~	✗	✗	~	~	✗	✗
ChatGLM-6B	~	~	✓	✗	✗	✓	~	~	✗	~	✗	✗	✗	✓
Mistral 7B-Instruct	~	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✓



WizardLM-7B	~	~	✗	✓	~	~	~	✓	✓	✗	✗	✗	✗	✗
Qwen 1.5	~	✗	✓	✗	✓	✗	~	~	✗	✗	✗	✗	~	✓
StableVicuna-13B	~	✗	~	~	~	~	~	~	~	✗	~	✗	✗	~
Falcon-40B-instruct	✗	~	✓	~	✗	✓	✗	~	~	✗	~	✗	✗	✗
UltraLM	✗	✗	~	✓	~	✗	✗	~	✓	✗	~	~	✗	✗
Yi 34B Chat	~	✗	✓	✗	✓	~	✗	✗	✓	✗	✗	✗	✗	~
Koala 13B	✓	~	~	~	✗	~	~	~	✗	✗	✗	✗	✗	✗
Mixtral 8x7B Instruct	✗	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✗
Stable Beluga 2	✗	✗	~	✗	✓	~	✗	~	~	✗	~	✗	✗	~
Stanford Alpaca	✓	✗	~	~	~	✗	~	✓	✗	✗	✗	✗	✗	✗
Falcon-180B-chat	✗	~	~	~	~	✗	✗	~	~	✗	~	✗	✗	✗
Orca 2	✗	✗	~	✗	✓	✗	✗	~	~	✗	~	✗	✗	~
Command R+	✗	✗	✗	✓	✓	~	✗	✗	✗	✗	~	✗	✗	✗
Gemma 7B Instruct	~	✗	~	✗	~	✗	✗	~	✗	✗	✓	✗	✗	✗
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~
Nanbeige2-Chat	✓	✗	✗	✗	✓	~	✗	✗	✗	✗	✗	✗	✗	~
Llama 3 Instruct	✗	✗	~	✗	~	✗	✗	~	✗	✗	~	✗	✗	~
Solar 70B	✗	✗	~	✗	~	✗	✗	✗	✗	✗	~	✗	✗	~
Xwin-LM	✗	✗	~	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	~
ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	✗	✗	✗	✗

Liesenfeld, Andreas, and Mark Dingemanse. "Rethinking open source generative AI: open washing and the EU AI Act." The 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024.

How can you run open source models?

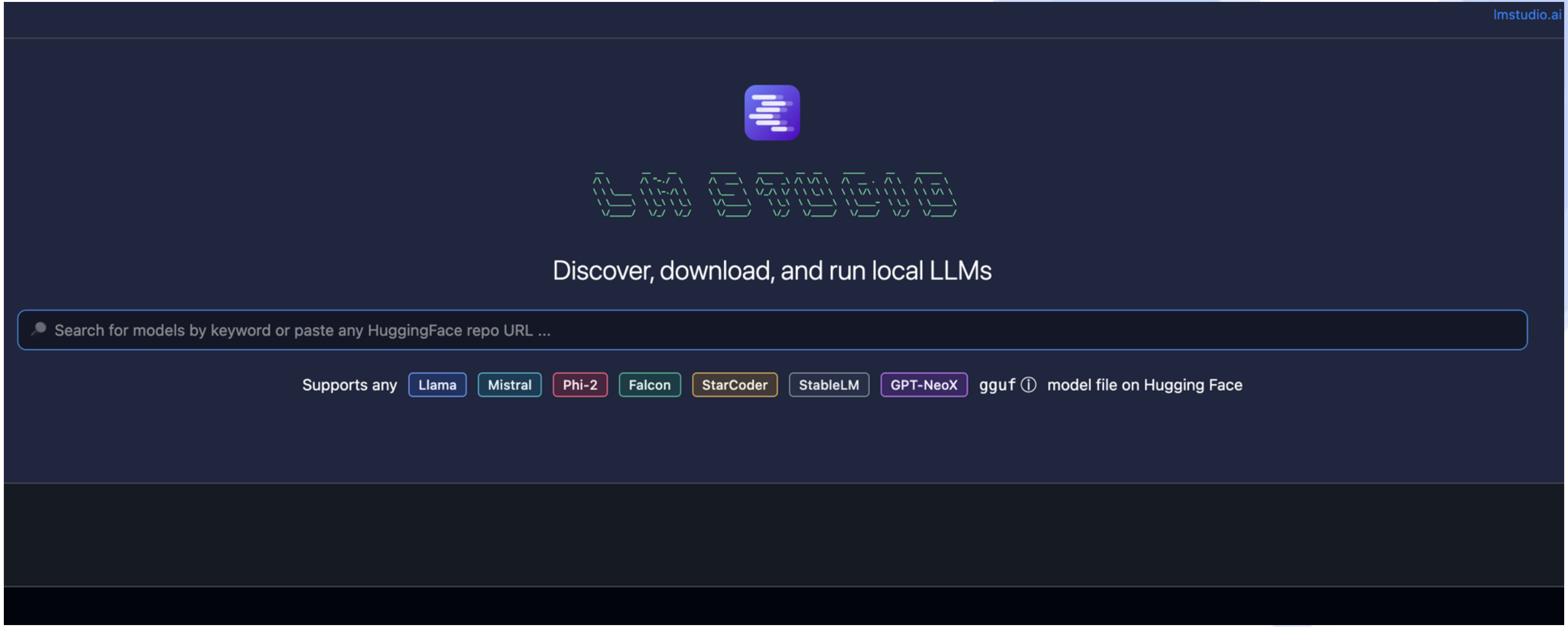
There are two main ways to access open source models:

- API
  - Huggingface Hub API
  - Individual company APIs
  - No compute required but restrictions on requests

- Cloud
  - Google Colab
  - Commercial cloud providers are expensive
  - Steep learning curve

- Locally
  - Code
  - No code
  - Requires compute…

No code



lmstudio.ai

Discover, download, and run local LLMs

Search for models by keyword or paste any HuggingFace repo URL ...

Supports any Llama Mistral Phi-2 Falcon StarCoder StableLM GPT-NeoX gguf ⓘ model file on Hugging Face

No Code

Some no code options for running LLMs locally

- LMStudio
  - Not open source
  - Desktop client
- GPT4All
  - Open source desktop client
  - Upload documents for question answering
  - Bindings for Python and NodeJS, and LangChain itegration
- Textgen-webui
  - Open source
  - Runs in the browser which means you can spin it up on a remote GPU
  - Requires some effort to install and run on Linux machines.
  - Ability to fine-tune models

Resources

LMStudio

GPT4All

Textgen-webui

OpenAI API Usage, Pricing, and Policy

HuggingFace Tutorials

How to keep up?

- LinkedIn
  - OpenAI
  - Meta and AI at Meta
  - Hugging Face
  - Microsoft and Microsoft Research
  - Apple
  - PyTorch
- YouTube
  - Fireship
  - Two Minute Papers
  - Joma Tech
- Reddit
  - r/LocalLLaMA
  - r/MachineLearning and r/learnmachinelearning
  - r/OpenAI
  - r/homelabs
  - r/LLMDevs
  - r/StableDiffusion