HANDS-ON MACHINE LEARNING (HOME)

Lecture/Lab – 2

An Introduction to Machine Learning Using LLMs as an Example

Goal: Run an LLM (Mistral 7B or Llama 2) on a local machine and chat with it

Ghulam Rasool, PhD

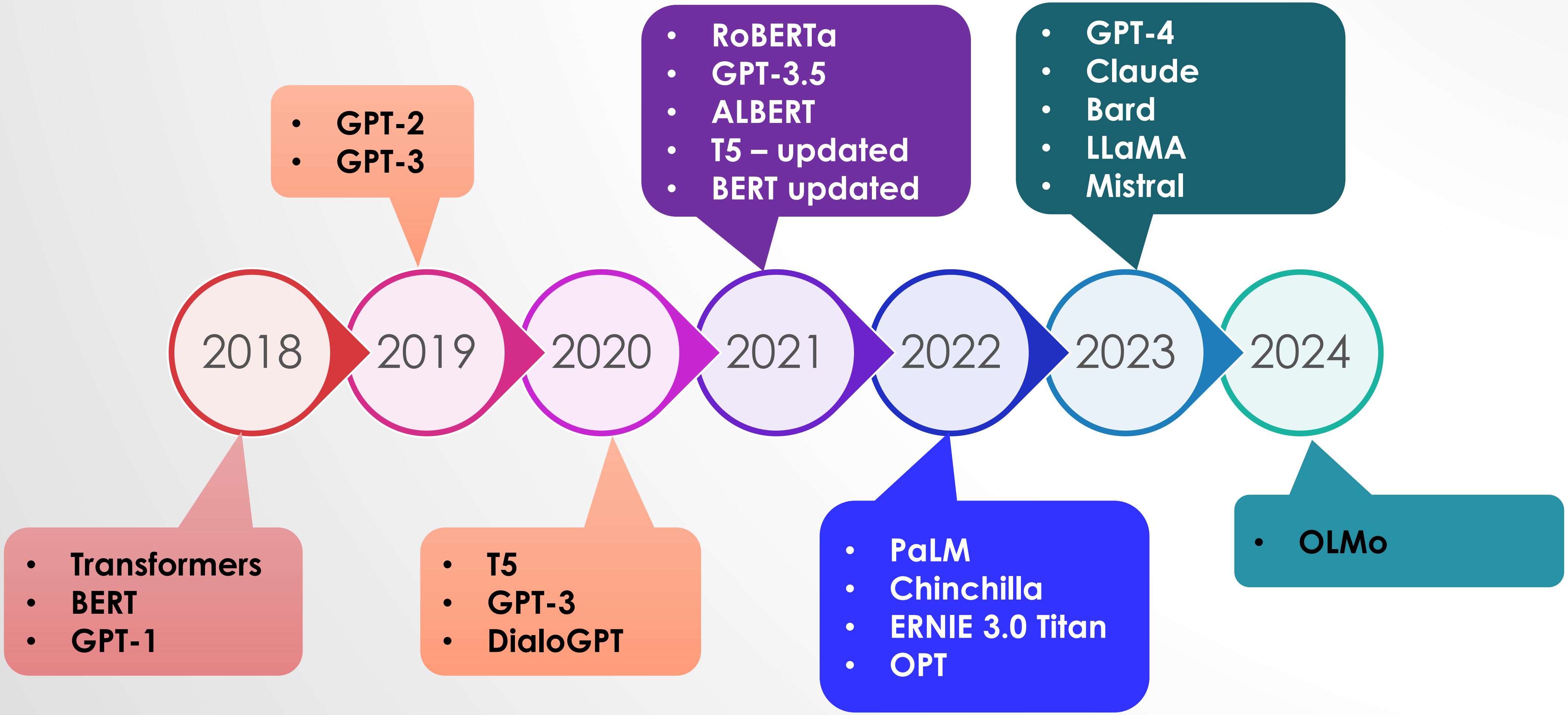
Department of Machine Learning, Moffitt Cancer Center

OVERVIEW

- **Review of LLMs**
- **Running LLMs on Local Machines**
  - LM Studio
  - Ollama
- **Graphical User Interface**
  - Open WebUI
  - Docker
- **Llama.cpp and gguf**
- **Prompting and Retrieval Augmented Generation (RAG)**
  - LangChain
  - LlamaIndex

LARGE LANGUAGE MODELS (LLMs)

Brief History



2018
- Transformers
- BERT
- GPT-1

2019
- GPT-2
- GPT-3

2020
- T5
- GPT-3
- DialoGPT

2021
- RoBERTa
- GPT-3.5
- ALBERT
- T5 – updated
- BERT updated

2022
- PaLM
- Chinchilla
- ERNIE 3.0 Titan
- OPT

2023
- GPT-4
- Claude
- Bard
- LLaMA
- Mistral

2024
- OLMo

LARGE LANGUAGE MODELs (LLMs)

Large

- **Higher capacity to learn**

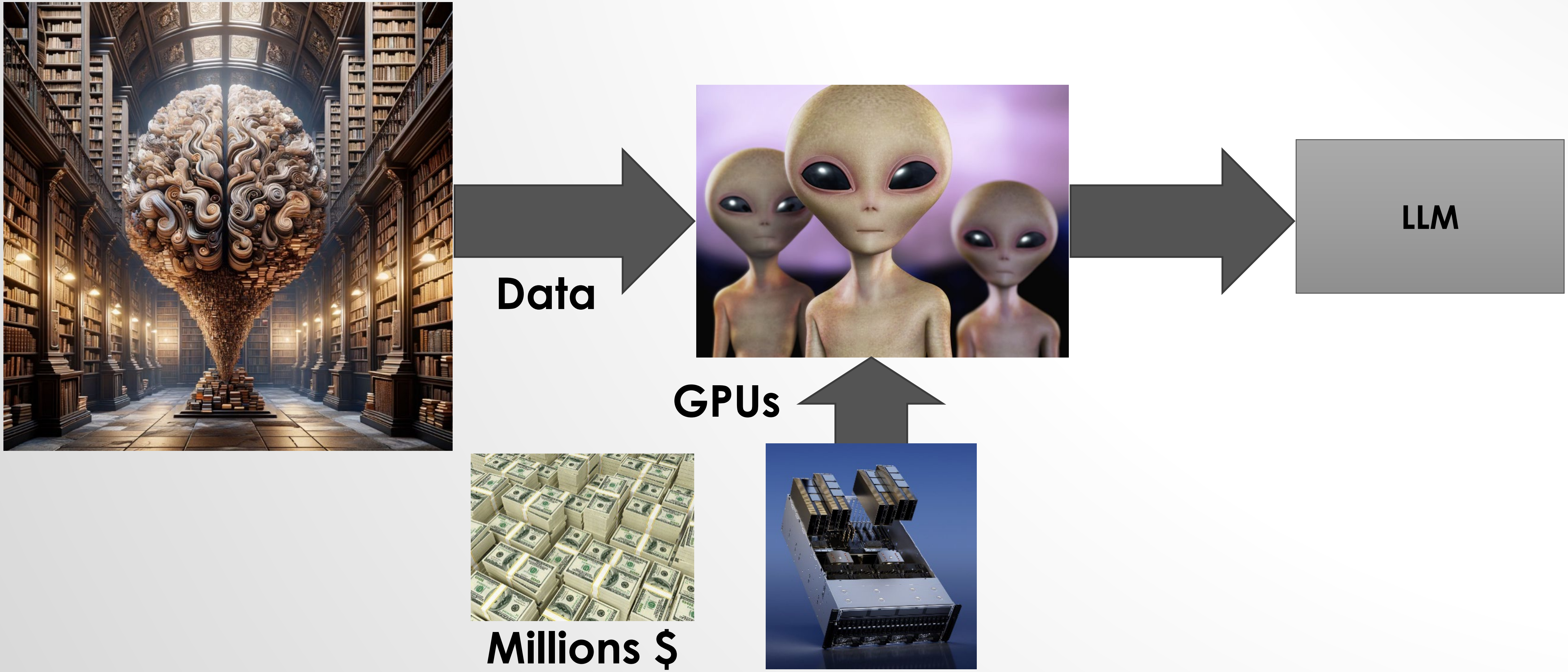
Billion of variables to store the learned knowledge

- **Larger datasets**

Trillions of bytes of data to learn from

LARGE LANGUAGE MODELs (LLMs)

How to train your LLM



Data

GPUs

Millions $

LLM

LARGE LANGUAGE MODELs (LLMs)

What does training mean?

- Predict the next word

She is moving to Tampa for her job.

Gradient Descent Algorithm

She is moving to Tampa for her job.

She is moving to Tampa for her family.

- Correct – Pat
- Incorrect – Punish

She is moving to Tampa for her car.

She is moving to Tampa for her alligator.

LARGE LANGUAGE MODELs (LLMs)

As a result of the training, we have

She is moving to Tampa for her _____.

What is the most probable next word?

Job 49%

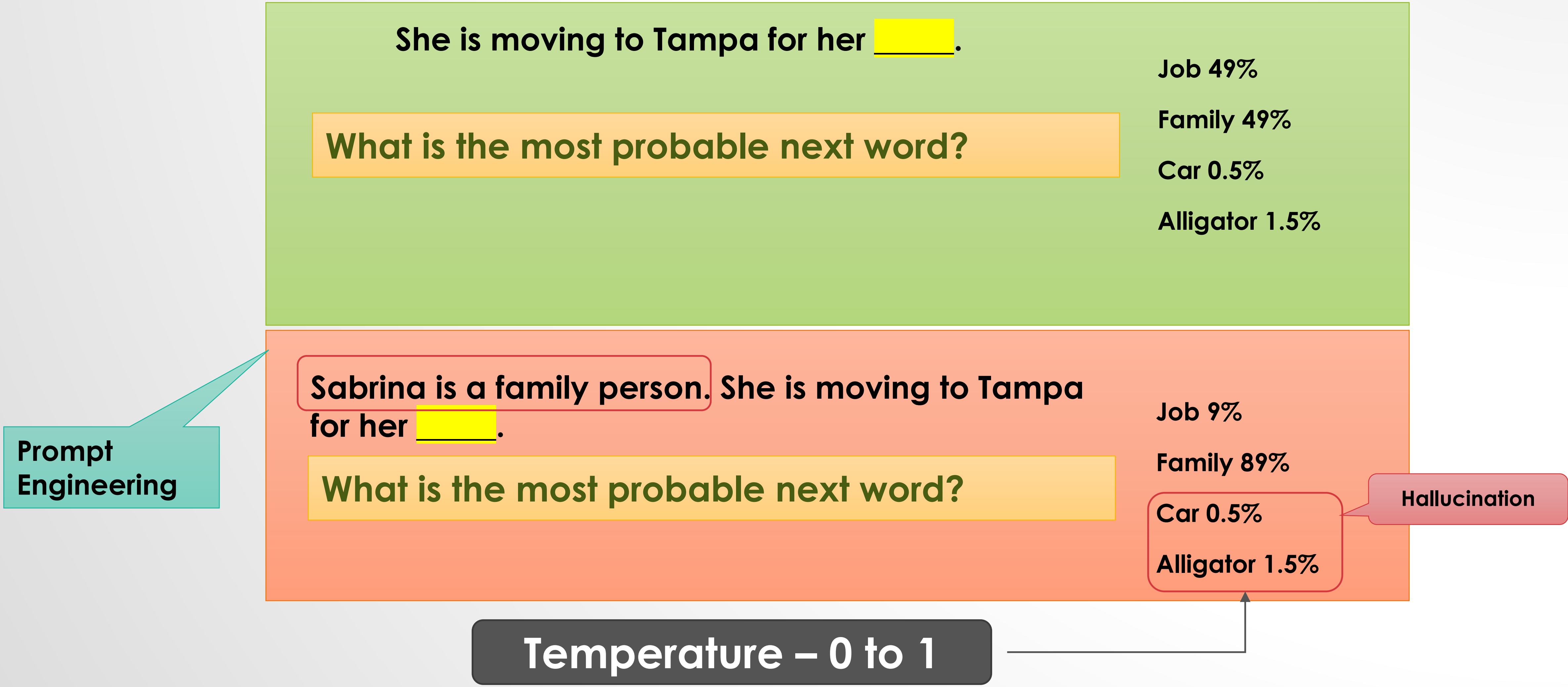
Family 49%

Car 0.5%

Alligator 1.5%

LARGE LANGUAGE MODELs (LLMs)

As a result of the training, we have



She is moving to Tampa for her ____.

What is the most probable next word?

Job 49%

Family 49%

Car 0.5%

Alligator 1.5%

Sabrina is a family person. She is moving to Tampa for her ____.

What is the most probable next word?

Job 9%

Family 89%

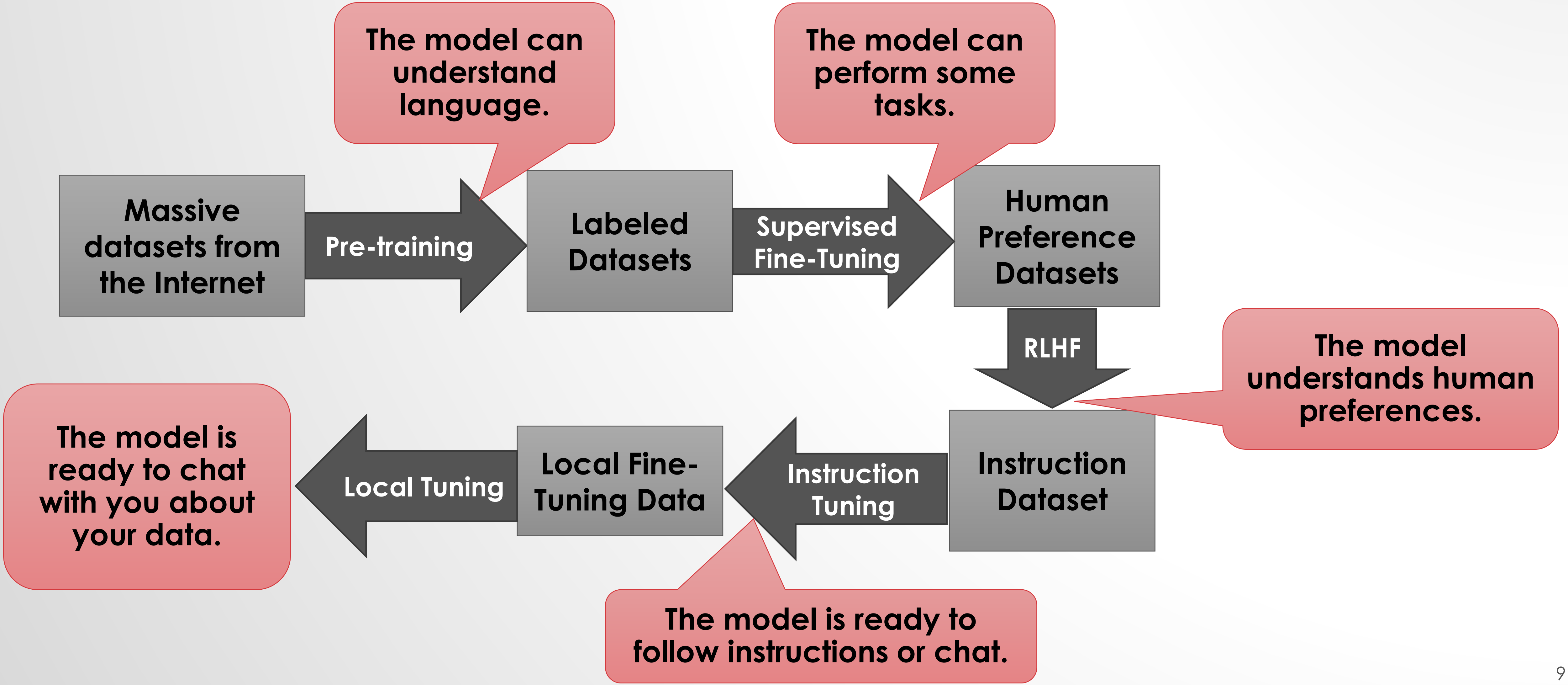
Car 0.5%

Alligator 1.5%

Temperature – 0 to 1

LARGE LANGUAGE MODELs (LLMs)

How to train your LLM



Massive datasets from the Internet

Pre-training

The model can understand language.

Labeled Datasets

Supervised Fine-Tuning

The model can perform some tasks.

Human Preference Datasets

RLHF

The model understands human preferences.

Instruction Dataset

Instruction Tuning

The model is ready to follow instructions or chat.

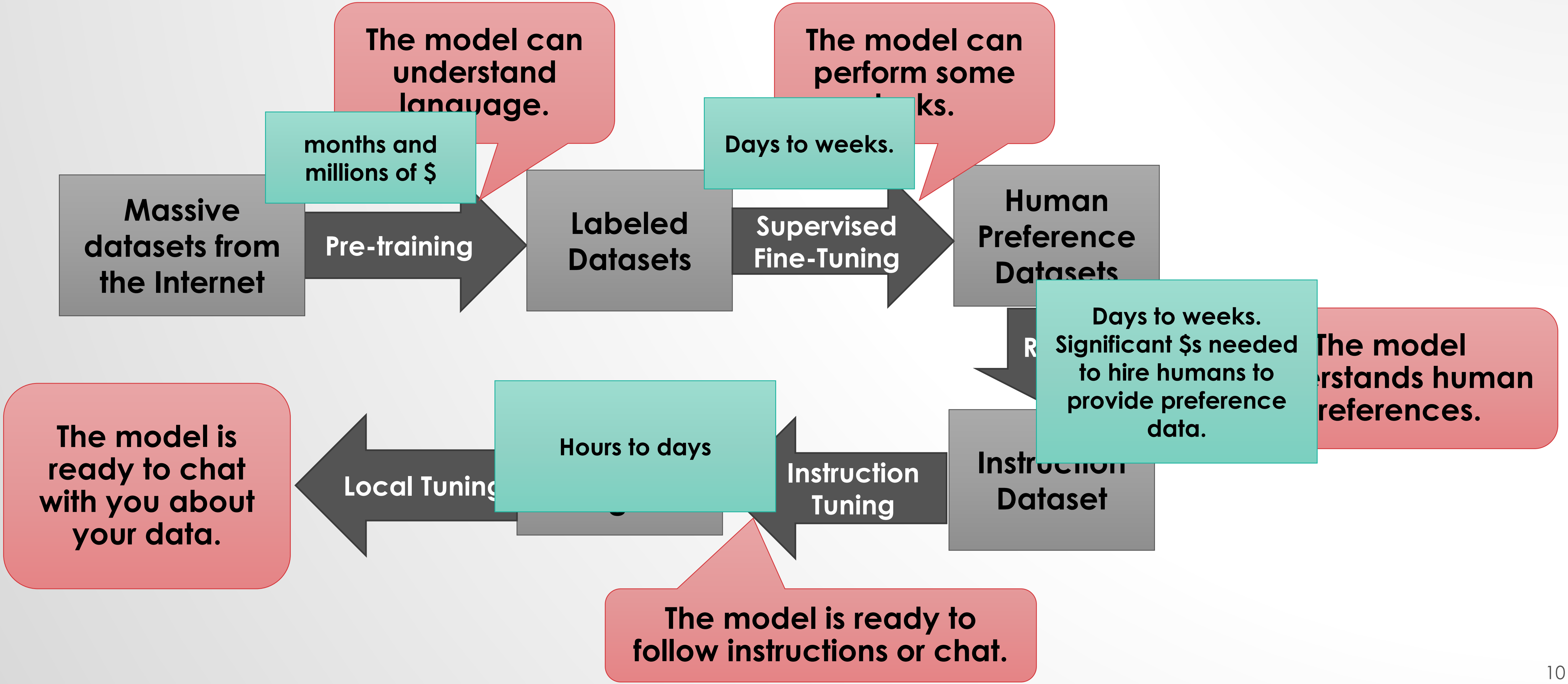
Local Fine-Tuning Data

Local Tuning

The model is ready to chat with you about your data.

LARGE LANGUAGE MODELs (LLMs)

How to train your LLM



Massive datasets from the Internet

months and millions of $

Pre-training

The model can understand language.

Labeled Datasets

Days to weeks.

Supervised Fine-Tuning

The model can perform some ks.

Human Preference Datasets

R

Days to weeks. Significant $s needed to hire humans to provide preference data.

The model rstands human references.

Instruction Dataset

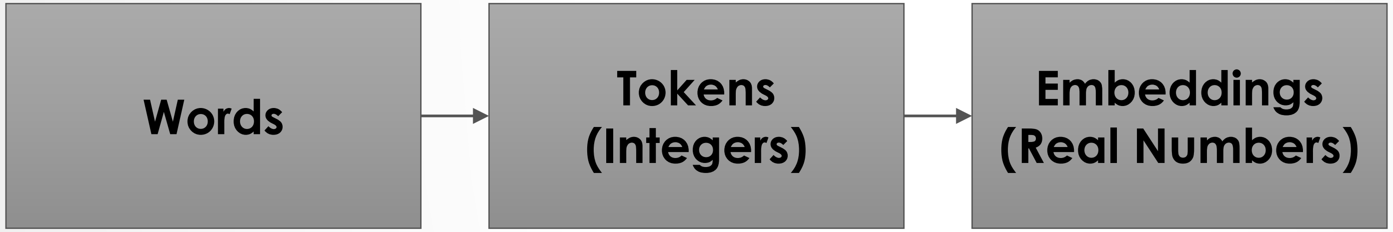
Instruction Tuning

The model is ready to follow instructions or chat.

Hours to days

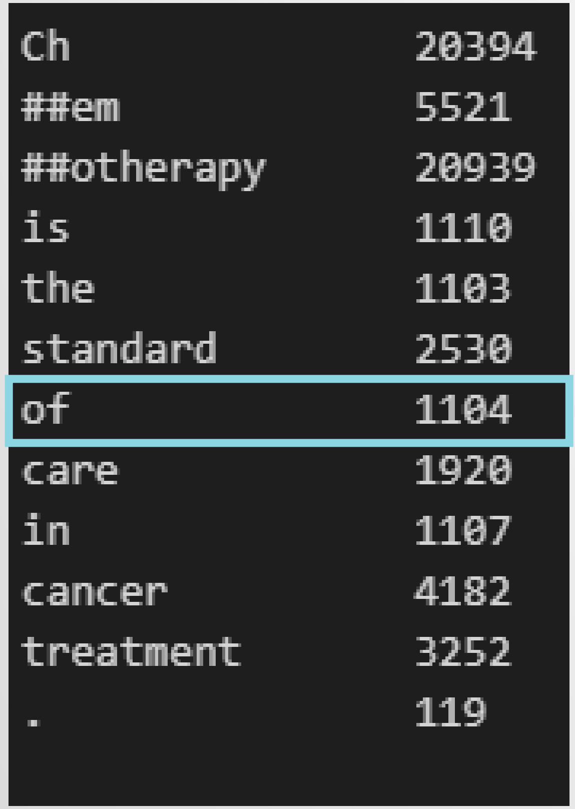
Local Tuning

The model is ready to chat with you about your data.

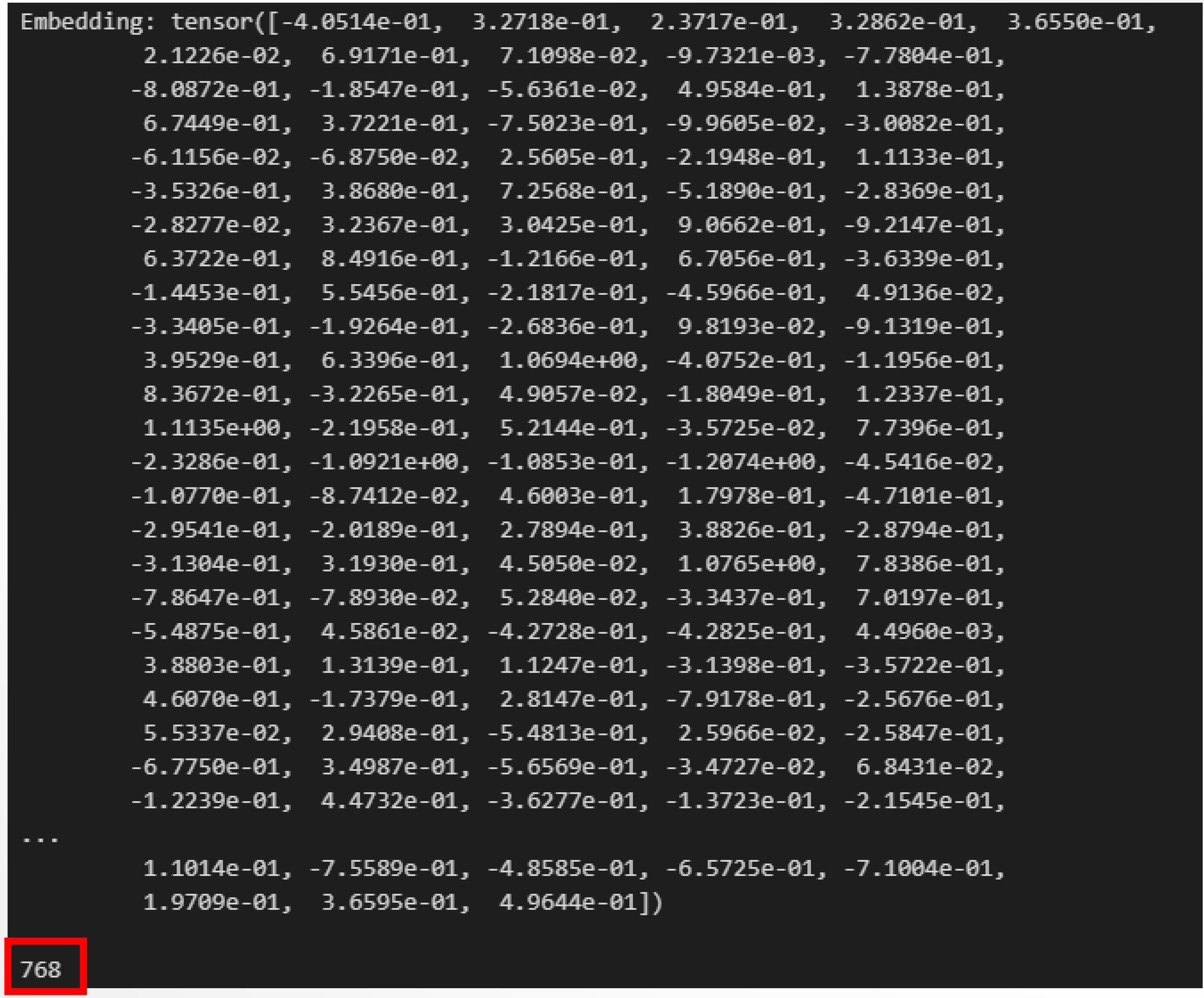


Words → Tokens (Integers) → Embeddings (Real Numbers)

Chemotherapy is the standard of care in cancer treatment.



```
Ch              20394
##em            5521
##otherapy      20939
is              1110
the             1103
standard        2530
of              1104
care            1920
in              1107
cancer          4182
treatment       3252
.               119
```



```
Embedding: tensor([-4.0514e-01,  3.2718e-01,  2.3717e-01,  3.2862e-01,  3.6550e-01,
         2.1226e-02,  6.9171e-01,  7.1098e-02, -9.7321e-03, -7.7804e-01,
        -8.0872e-01, -1.8547e-01, -5.6361e-02,  4.9584e-01,  1.3878e-01,
         6.7449e-01,  3.7221e-01, -7.5023e-01, -9.9605e-02, -3.0082e-01,
        -6.1156e-02, -6.8750e-02,  2.5605e-01, -2.1948e-01,  1.1133e-01,
        -3.5326e-01,  3.8680e-01,  7.2568e-01, -5.1890e-01, -2.8369e-01,
        -2.8277e-02,  3.2367e-01,  3.0425e-01,  9.0662e-01, -9.2147e-01,
         6.3722e-01,  8.4916e-01, -1.2166e-01,  6.7056e-01, -3.6339e-01,
        -1.4453e-01,  5.5456e-01, -2.1817e-01, -4.5966e-01,  4.9136e-02,
        -3.3405e-01, -1.9264e-01, -2.6836e-01,  9.8193e-02, -9.1319e-01,
         3.9529e-01,  6.3396e-01,  1.0694e+00, -4.0752e-01, -1.1956e-01,
         8.3672e-01, -3.2265e-01,  4.9057e-02, -1.8049e-01,  1.2337e-01,
         1.1135e+00, -2.1958e-01,  5.2144e-01, -3.5725e-02,  7.7396e-01,
        -2.3286e-01, -1.0921e+00, -1.0853e-01, -1.2074e+00, -4.5416e-02,
        -1.0770e-01, -8.7412e-02,  4.6003e-01,  1.7978e-01, -4.7101e-01,
        -2.9541e-01, -2.0189e-01,  2.7894e-01,  3.8826e-01, -2.8794e-01,
        -3.1304e-01,  3.1930e-01,  4.5050e-02,  1.0765e+00,  7.8386e-01,
        -7.8647e-01, -7.8930e-02,  5.2840e-02, -3.3437e-01,  7.0197e-01,
        -5.4875e-01,  4.5861e-02, -4.2728e-01, -4.2825e-01,  4.4960e-03,
         3.8803e-01,  1.3139e-01,  1.1247e-01, -3.1398e-01, -3.5722e-01,
         4.6070e-01, -1.7379e-01,  2.8147e-01, -7.9178e-01, -2.5676e-01,
         5.5337e-02,  2.9408e-01, -5.4813e-01,  2.5966e-02, -2.5847e-01,
        -6.7750e-01,  3.4987e-01, -5.6569e-01, -3.4727e-02,  6.8431e-02,
        -1.2239e-01,  4.4732e-01, -3.6277e-01, -1.3723e-01, -2.1545e-01,
...
         1.1014e-01, -7.5589e-01, -4.8585e-01, -6.5725e-01, -7.1004e-01,
         1.9709e-01,  3.6595e-01,  4.9644e-01])

768
```

LARGE LANGUAGE MODELs (LLMs)

Model Training

All or part of the model parameters are updated

- Transfer Learning
- Unsupervised Learning
- Supervised Learning
- Curriculum Learning

Adapter Training

Original model parameters are frozen

- Add additional small models, called adapters
- Train “adapters” for individual tasks tasks.
- Low-Rank Adaptation
- Parameter-Efficient Fine-Tuning

Prompt Engineering

Model parameters are frozen

- No adapters are added
- Add more context to the prompt
- Zero-shot
- Few-shot
- Retrieval Augmented Generation (RAG)

LARGE LANGUAGE MODELs (LLMs)

Traditional NLP/ML

- Needs labelled data
  - Cost of data collection/labeling
  - Legal/Privacy concerns around using data
- 1 model per task results in
  - Increased model development/tuning cost
  - Increased operational costs
  - Increased money spent on sourcing data
- Relatively Limited generalization
- Computationally cheaper (~300 Million parameters)

Modern NLP/ML

- Need a small amount of labeled data
- A single generic model can do more than one task
- More generalized: Besides language, it learns higher-level concepts, styles, etc.
- Computationally Expensive (~500 Billion parameters)

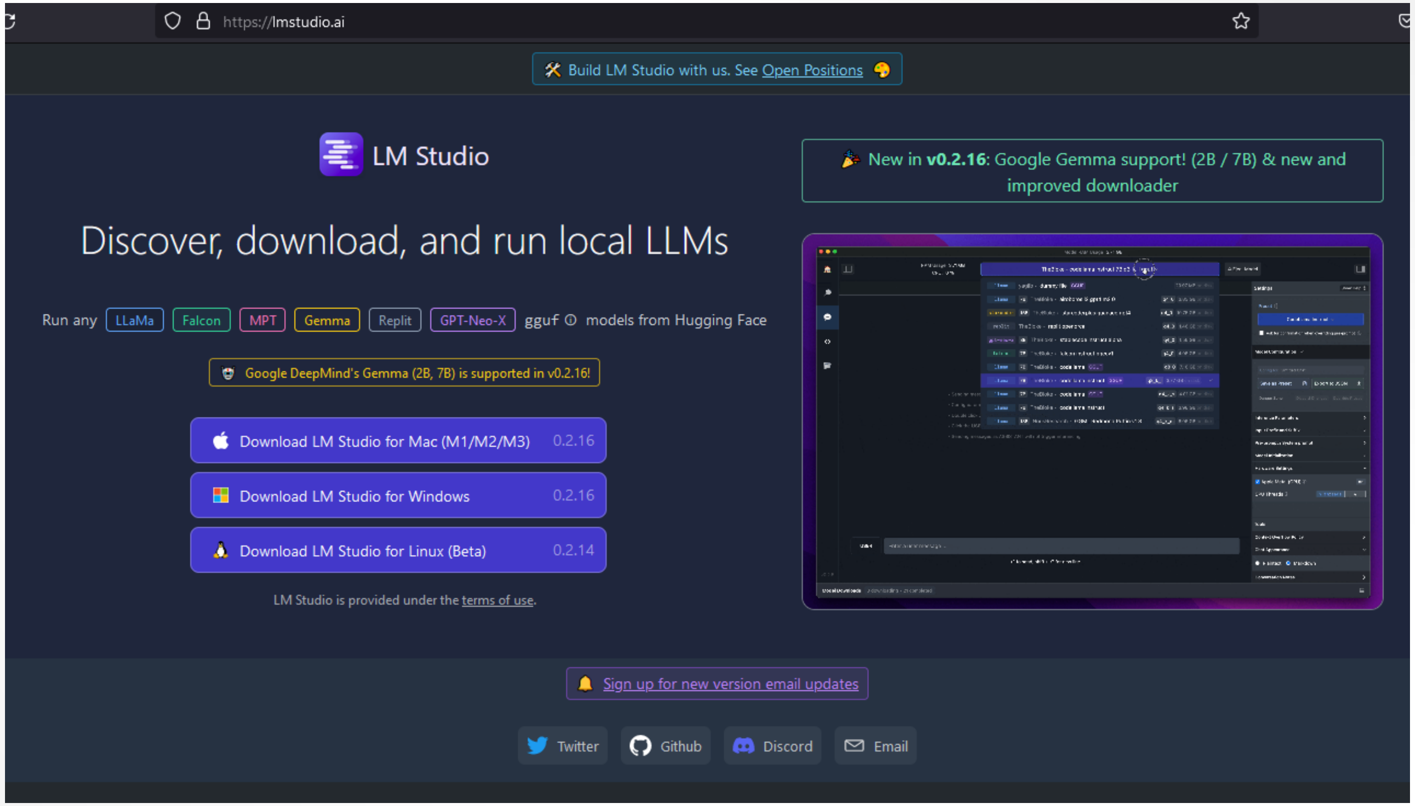
Leveraging more compute to get a general model without significant data/labeling cost

Running LLMs on Local Machines

- LM Studio
- Ollama
- Open WebUI

LM STUDIO

- Run LLMs on local computers entirely offline
- Use models through the in-app Chat UI or an OpenAI-compatible local server
- Download any compatible model files from Hugging Face repositories



https://lmstudio.ai

Build LM Studio with us. See Open Positions

LM Studio

New in **v0.2.16**: Google Gemma support! (2B / 7B) & new and improved downloader

Discover, download, and run local LLMs

Run any LLaMa Falcon MPT Gemma Replit GPT-Neo-X gguf models from Hugging Face

Google DeepMind's Gemma (2B, 7B) is supported in v0.2.16!

Download LM Studio for Mac (M1/M2/M3) 0.2.16

Download LM Studio for Windows 0.2.16

Download LM Studio for Linux (Beta) 0.2.14

LM Studio is provided under the terms of use.

Sign up for new version email updates

Twitter Github Discord Email

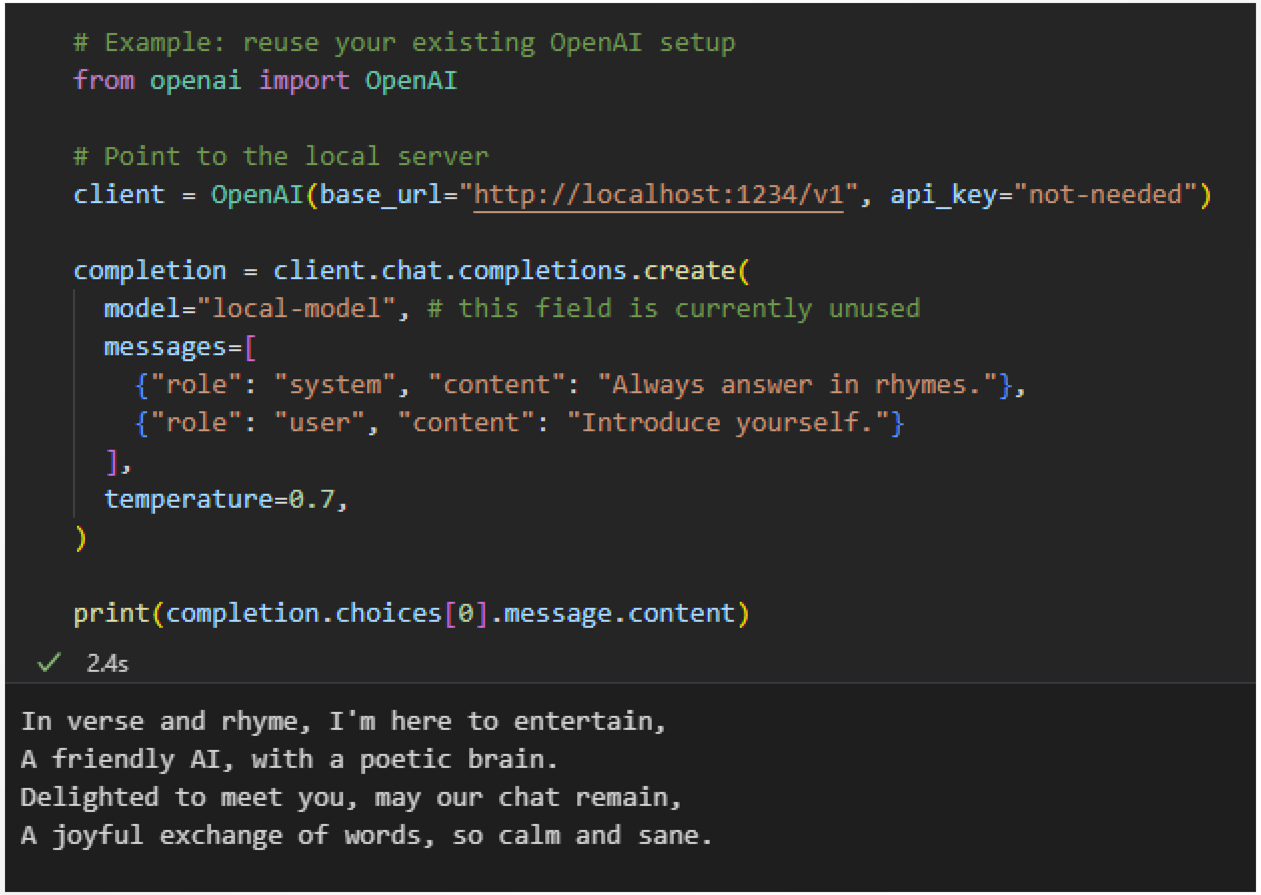
LM STUDIO

- Quantization
  - Quantization refers to a set of techniques that enable running models on resource-constrained platforms
  - Q2_K, Q4_K_M, Q5_0, Q8_0
- Chat UI
- GPU Offload
- Context Length

https://towardsdatascience.com/introduction-to-weight-quantization-2494701b9c0c

LM STUDIO - LOCAL INFERENCE SERVER

- **Serve the model on the local machine**
  - http://localhost:1234/
  - http://127.0.0.1:1234
- **How do I use it?**



```python
# Example: reuse your existing OpenAI setup
from openai import OpenAI

# Point to the local server
client = OpenAI(base_url="http://localhost:1234/v1", api_key="not-needed")

completion = client.chat.completions.create(
  model="local-model", # this field is currently unused
  messages=[
    {"role": "system", "content": "Always answer in rhymes."},
    {"role": "user", "content": "Introduce yourself."}
  ],
  temperature=0.7,
)

print(completion.choices[0].message.content)
```

```
In verse and rhyme, I'm here to entertain,
A friendly AI, with a poetic brain.
Delighted to meet you, may our chat remain,
A joyful exchange of words, so calm and sane.
```

https://towardsdatascience.com/introduction-to-weight-quantization-2494701b9c0c

LM STUDIO - LOCAL INFERENCE SERVER

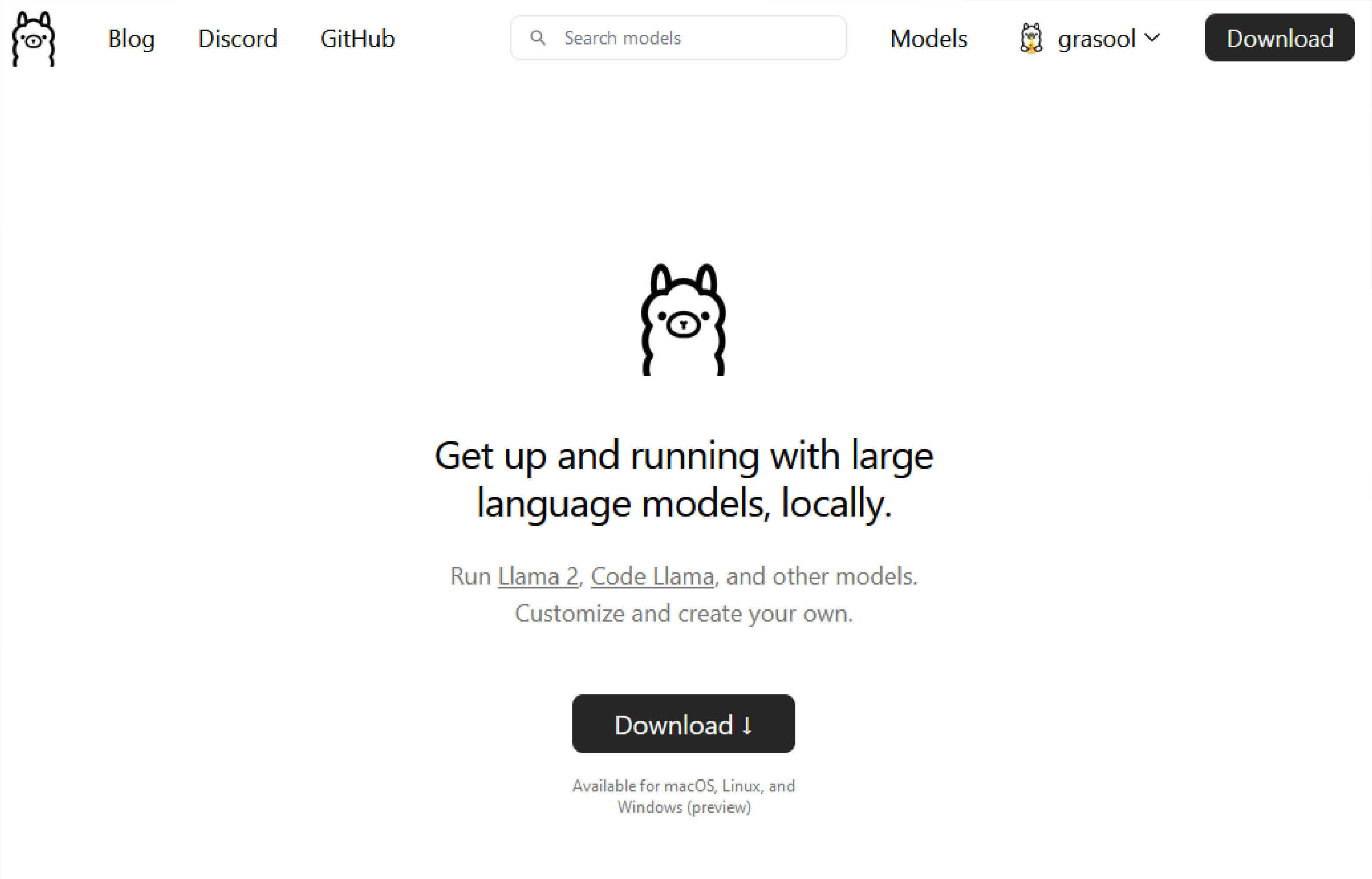
- **Serve the model on the local machine**
  - **http://localhost:1234/**
  - **http://127.0.0.1:1234**
- **How do I use it?**
  - **Chatbot**
  - **VLM Chatbot**
    - **Llava Base Model**
    - **Vision Adapter**

https://towardsdatascience.com/introduction-to-weight-quantization-2494701b9c0c

OLLAMA

https://ollama.com/

- Running LLMs, locally



Get up and running with large language models, locally.

Run Llama 2, Code Llama, and other models. Customize and create your own.

Available for macOS, Linux, and Windows (preview)

OLLAMA

https://ollama.com/

- Running LLMs, locally
- http://localhost:11434

OPEN WEBUI

https://openwebui.com/

https://github.com/open-webui/open-webui

- http://localhost:3000

- Download and Install Docker

https://docs.docker.com/desktop/install/windows-install/