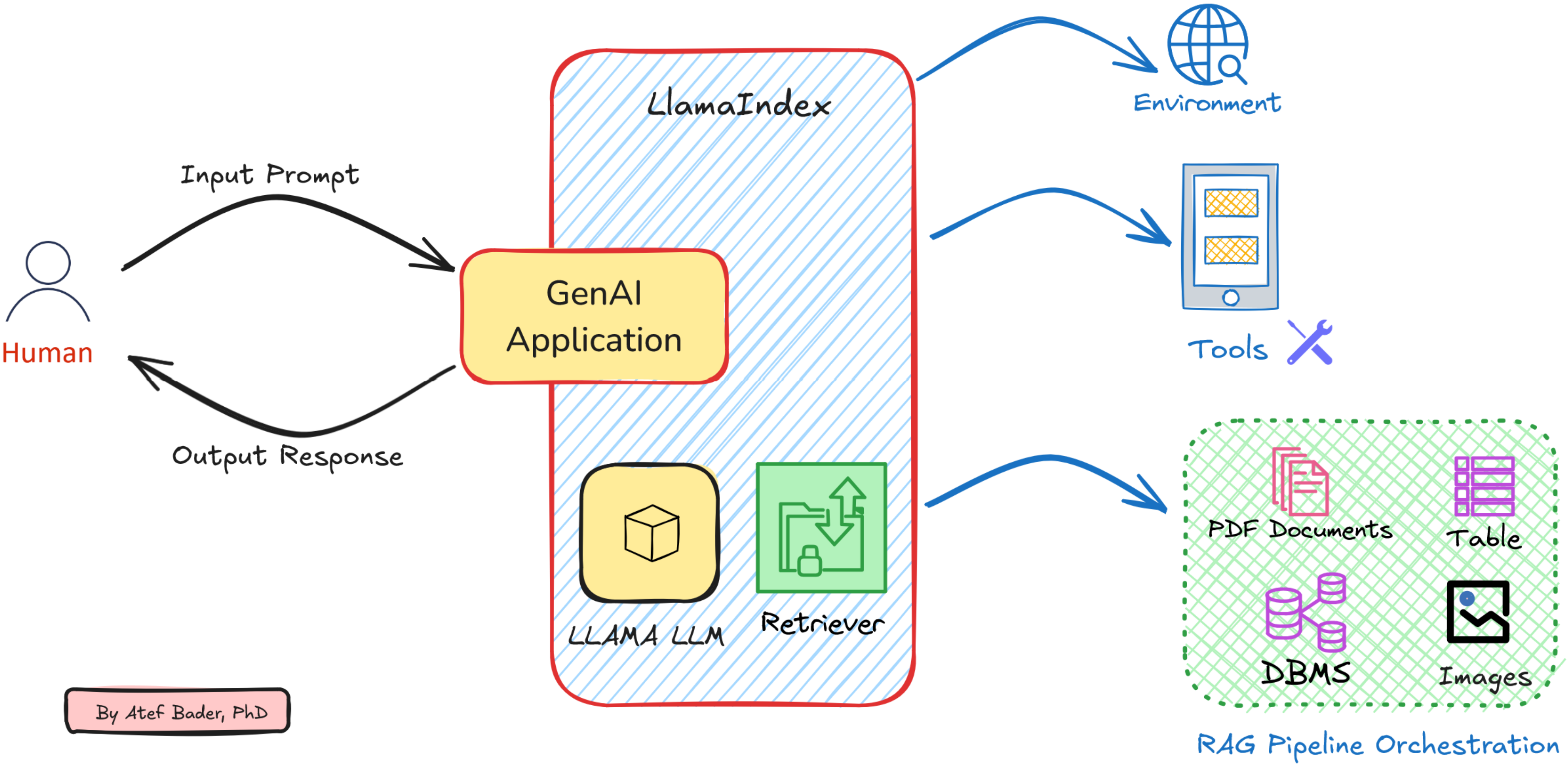
GenAI Applications:

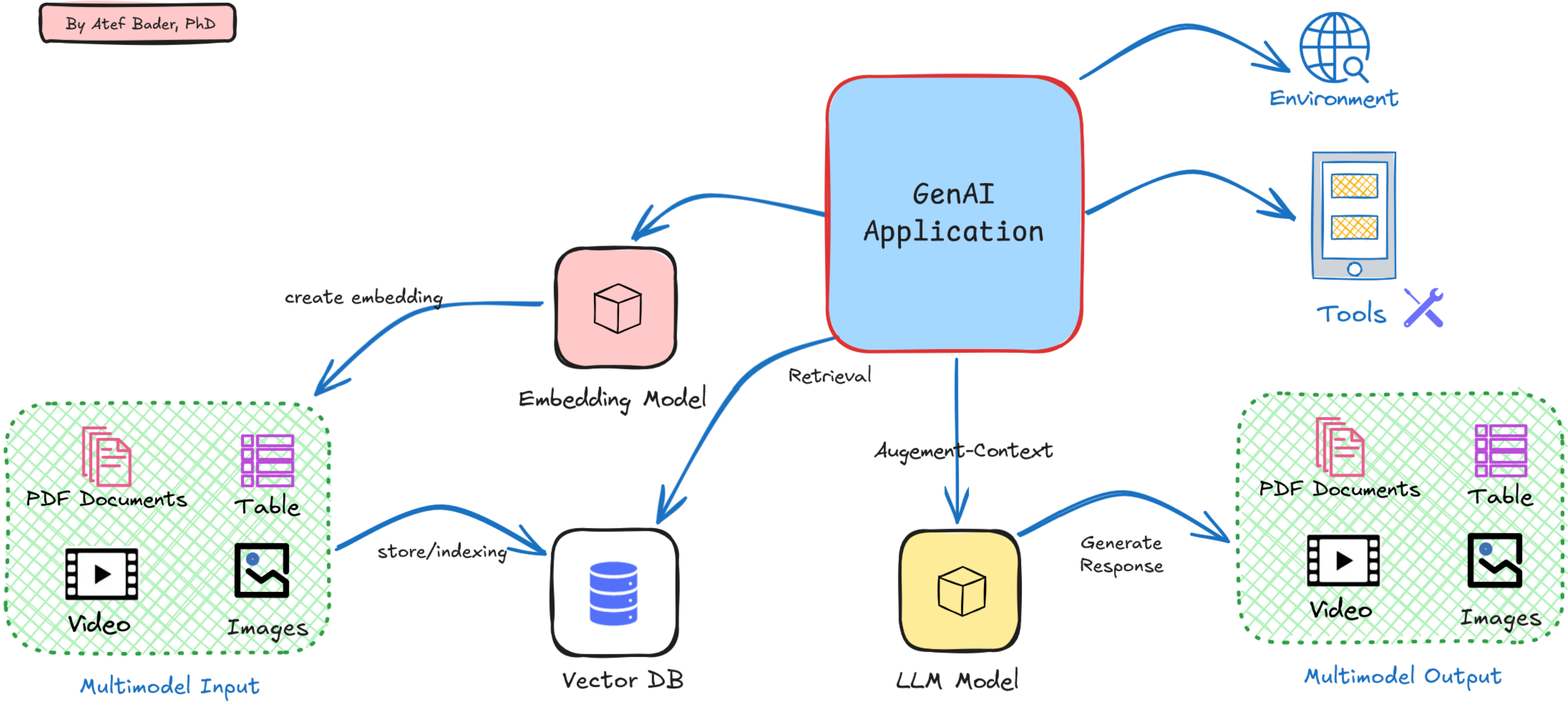
RAG Pipeline Orchestrations, Agents, Workflows, Chats, etc.

Dr. Atef Bader

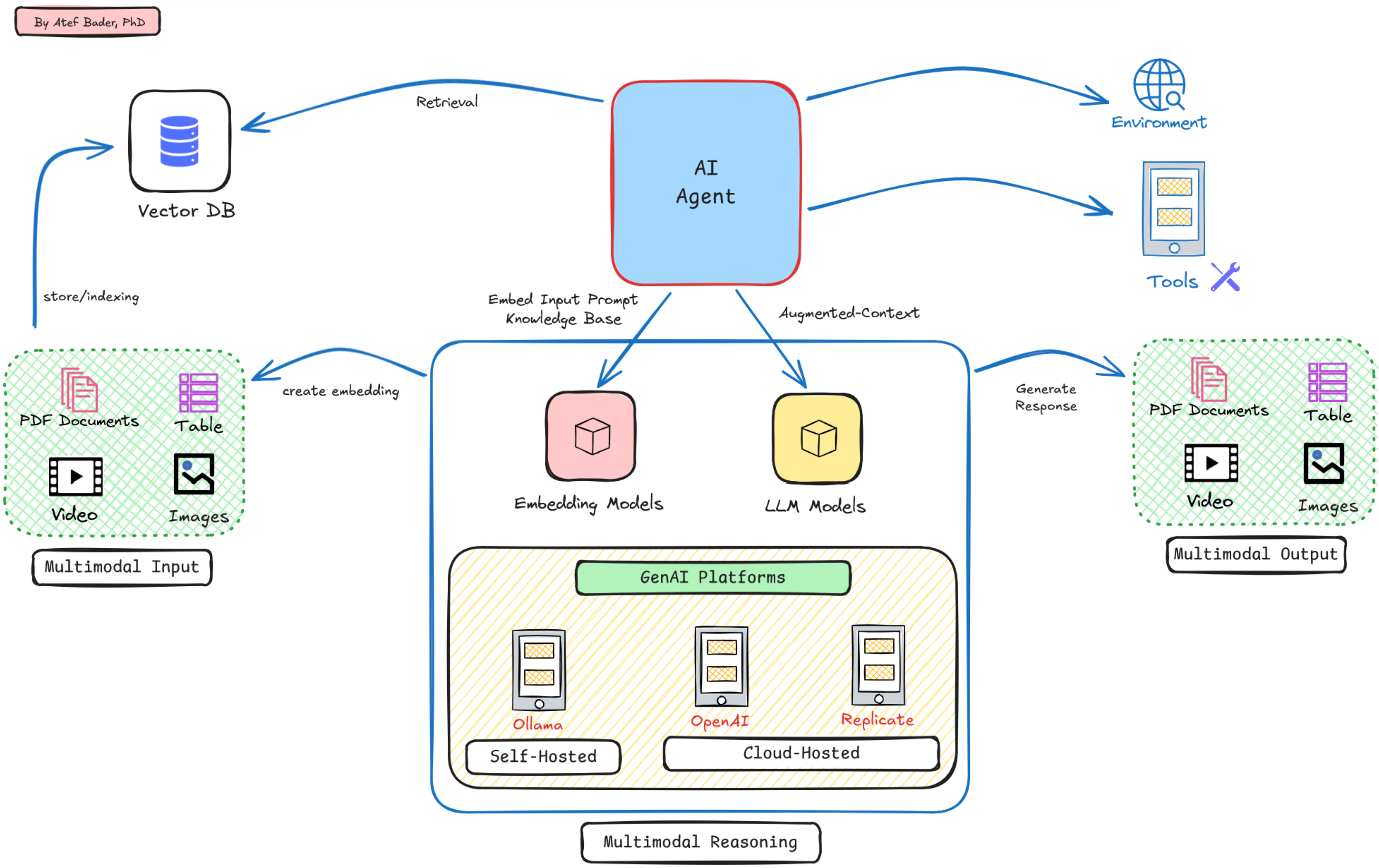


By Atef Bader, PhD

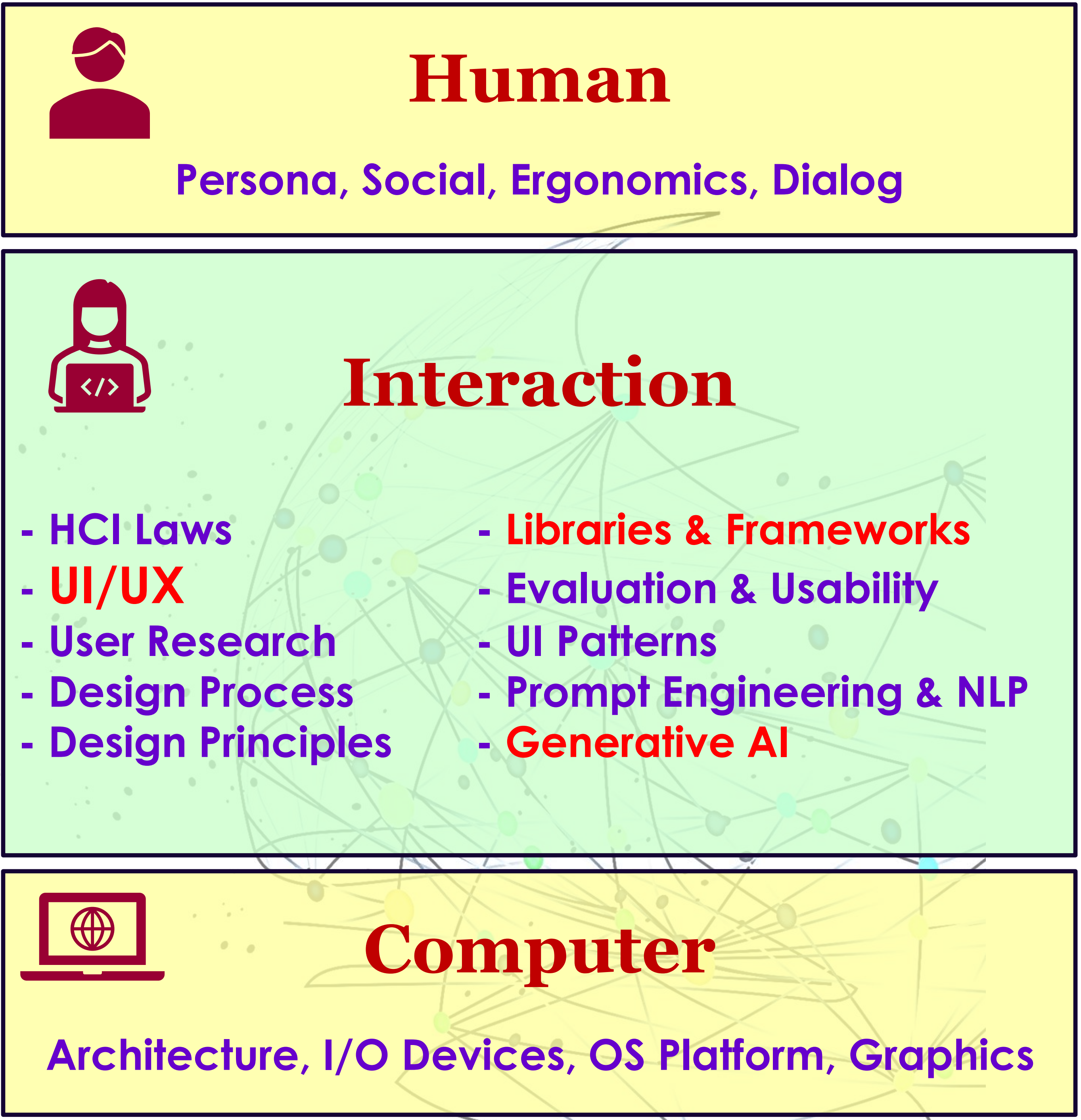
By Atef Bader, PhD



RAG Pipeline Orchestration



Multimodal AI Agent Engineering



Human

Persona, Social, Ergonomics, Dialog

Interaction

- HCI Laws
- UI/UX
- User Research
- Design Process
- Design Principles

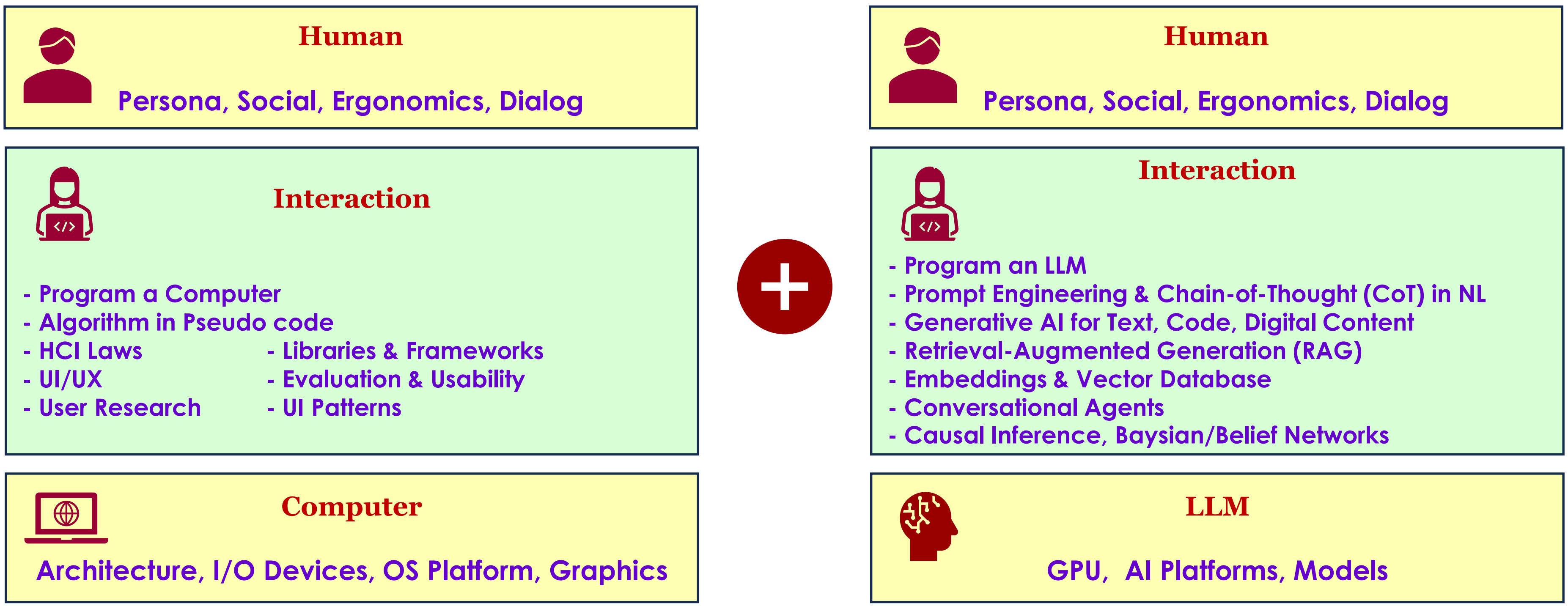
- Libraries & Frameworks
- Evaluation & Usability
- UI Patterns
- Prompt Engineering & NLP
- Generative AI

Computer

Architecture, I/O Devices, OS Platform, Graphics

Human-Computer Interaction

Human-LLM Application Engineering



Human-Computer Interaction

Human

Persona, Social, Ergonomics, Dialog

Interaction

- Program a Computer
- Algorithm in Pseudo code
- HCI Laws
- UI/UX
- User Research
- Libraries & Frameworks
- Evaluation & Usability
- UI Patterns

Computer

Architecture, I/O Devices, OS Platform, Graphics

Human-LLM Interaction

Human

Persona, Social, Ergonomics, Dialog

Interaction

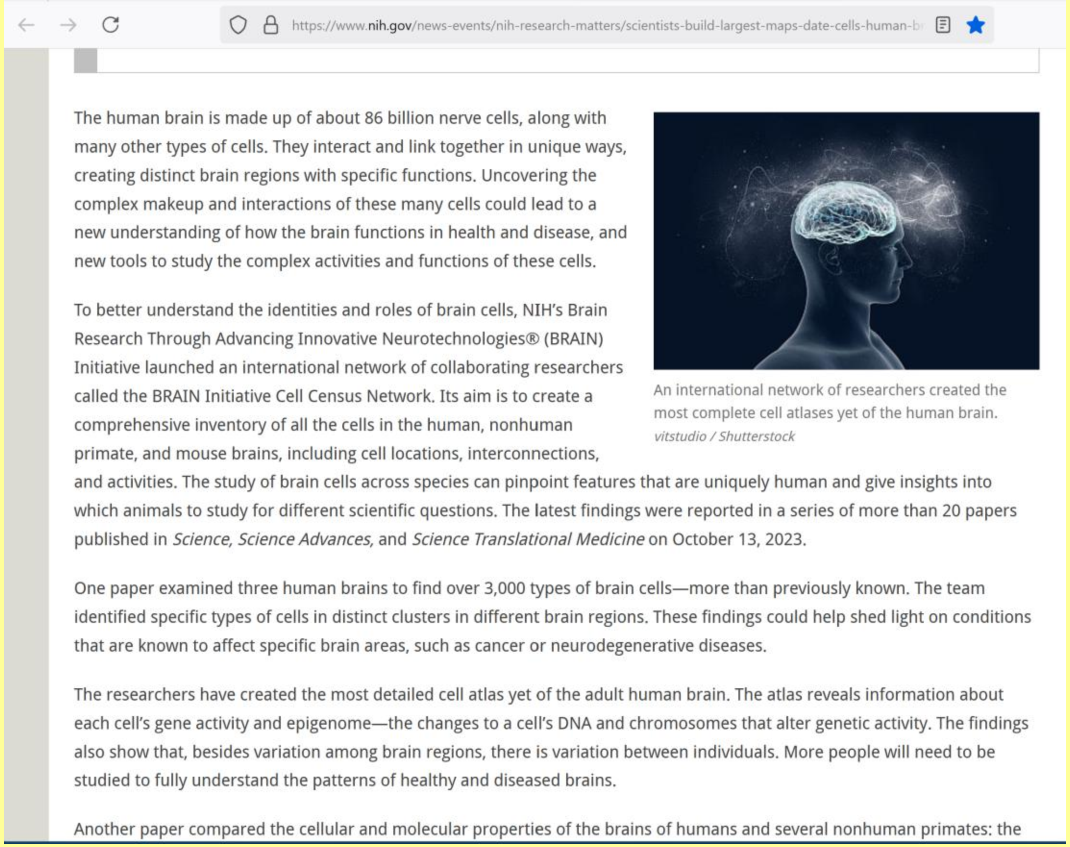
- Program an LLM
- Prompt Engineering & Chain-of-Thought (CoT) in NL
- Generative AI for Text, Code, Digital Content
- Retrieval-Augmented Generation (RAG)
- Embeddings & Vector Database
- Conversational Agents
- Causal Inference, Baysian/Belief Networks

LLM

GPU, AI Platforms, Models

Human Brain

- https://www.nih.gov/news-events/nih-research-matters/scientists-build-largest-maps-date-cells-human-brain
  - The human brain is made up of about 86 billion nerve cells, along with many other types of cells.



The human brain is made up of about 86 billion nerve cells, along with many other types of cells. They interact and link together in unique ways, creating distinct brain regions with specific functions. Uncovering the complex makeup and interactions of these many cells could lead to a new understanding of how the brain functions in health and disease, and new tools to study the complex activities and functions of these cells.

An international network of researchers created the most complete cell atlases yet of the human brain. *vitstudio / Shutterstock*

To better understand the identities and roles of brain cells, NIH's Brain Research Through Advancing Innovative Neurotechnologies® (BRAIN) Initiative launched an international network of collaborating researchers called the BRAIN Initiative Cell Census Network. Its aim is to create a comprehensive inventory of all the cells in the human, nonhuman primate, and mouse brains, including cell locations, interconnections, and activities. The study of brain cells across species can pinpoint features that are uniquely human and give insights into which animals to study for different scientific questions. The latest findings were reported in a series of more than 20 papers published in *Science, Science Advances,* and *Science Translational Medicine* on October 13, 2023.

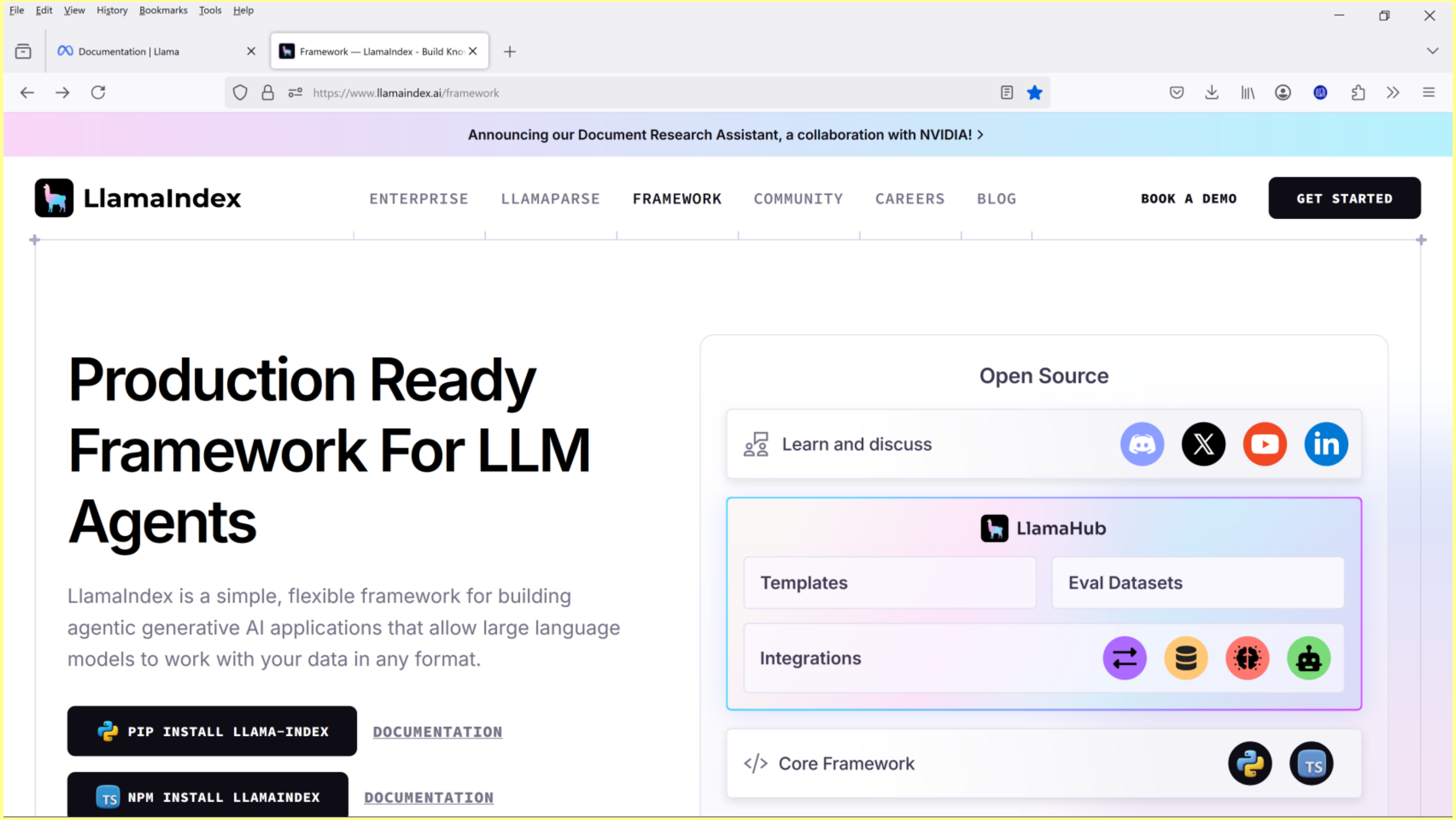
One paper examined three human brains to find over 3,000 types of brain cells—more than previously known. The team identified specific types of cells in distinct clusters in different brain regions. These findings could help shed light on conditions that are known to affect specific brain areas, such as cancer or neurodegenerative diseases.

The researchers have created the most detailed cell atlas yet of the adult human brain. The atlas reveals information about each cell's gene activity and epigenome—the changes to a cell's DNA and chromosomes that alter genetic activity. The findings also show that, besides variation among brain regions, there is variation between individuals. More people will need to be studied to fully understand the patterns of healthy and diseased brains.

Another paper compared the cellular and molecular properties of the brains of humans and several nonhuman primates: the

LlamaIndex : Framework for GenAI Applications

- https://www.llamaindex.ai/framework


File Edit View History Bookmarks Tools Help
Documentation | Llama
Framework — LlamaIndex - Build Kno
https://www.llamaindex.ai/framework
LlamaIndex
ENTERPRISE
LLAMAPARSE
FRAMEWORK
COMMUNITY
CAREERS
BLOG
BOOK A DEMO
GET STARTED

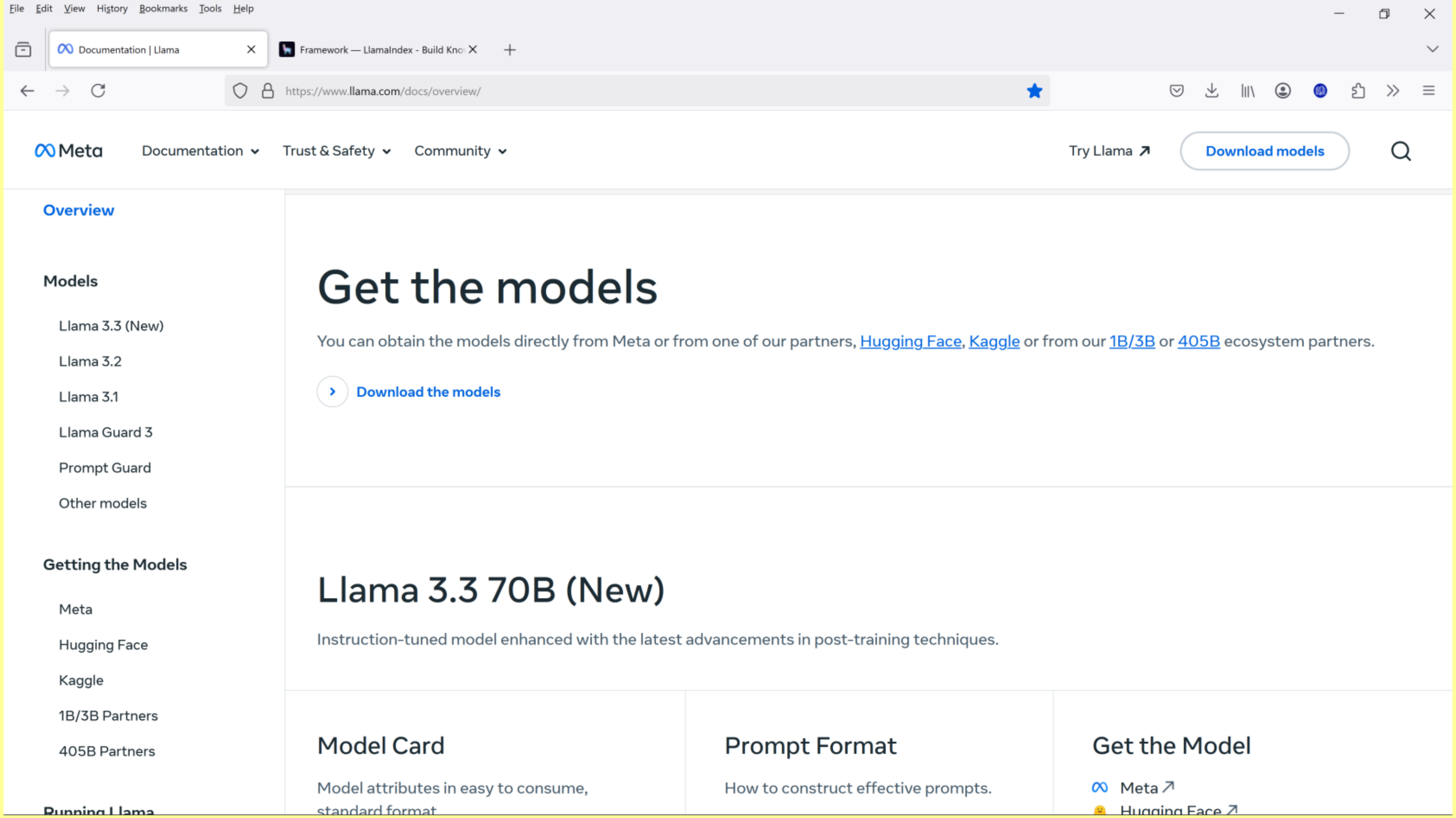
Announcing our Document Research Assistant, a collaboration with NVIDIA! >

Production Ready Framework For LLM Agents

LlamaIndex is a simple, flexible framework for building agentic generative AI applications that allow large language models to work with your data in any format.

LLAMA – LLM Model

- https://www.llama.com/docs/overview/



Get the models

You can obtain the models directly from Meta or from one of our partners, Hugging Face, Kaggle or from our 1B/3B or 405B ecosystem partners.

› Download the models

Llama 3.3 70B (New)

Instruction-tuned model enhanced with the latest advancements in post-training techniques.

Model Card

Model attributes in easy to consume, standard format

Prompt Format

How to construct effective prompts.

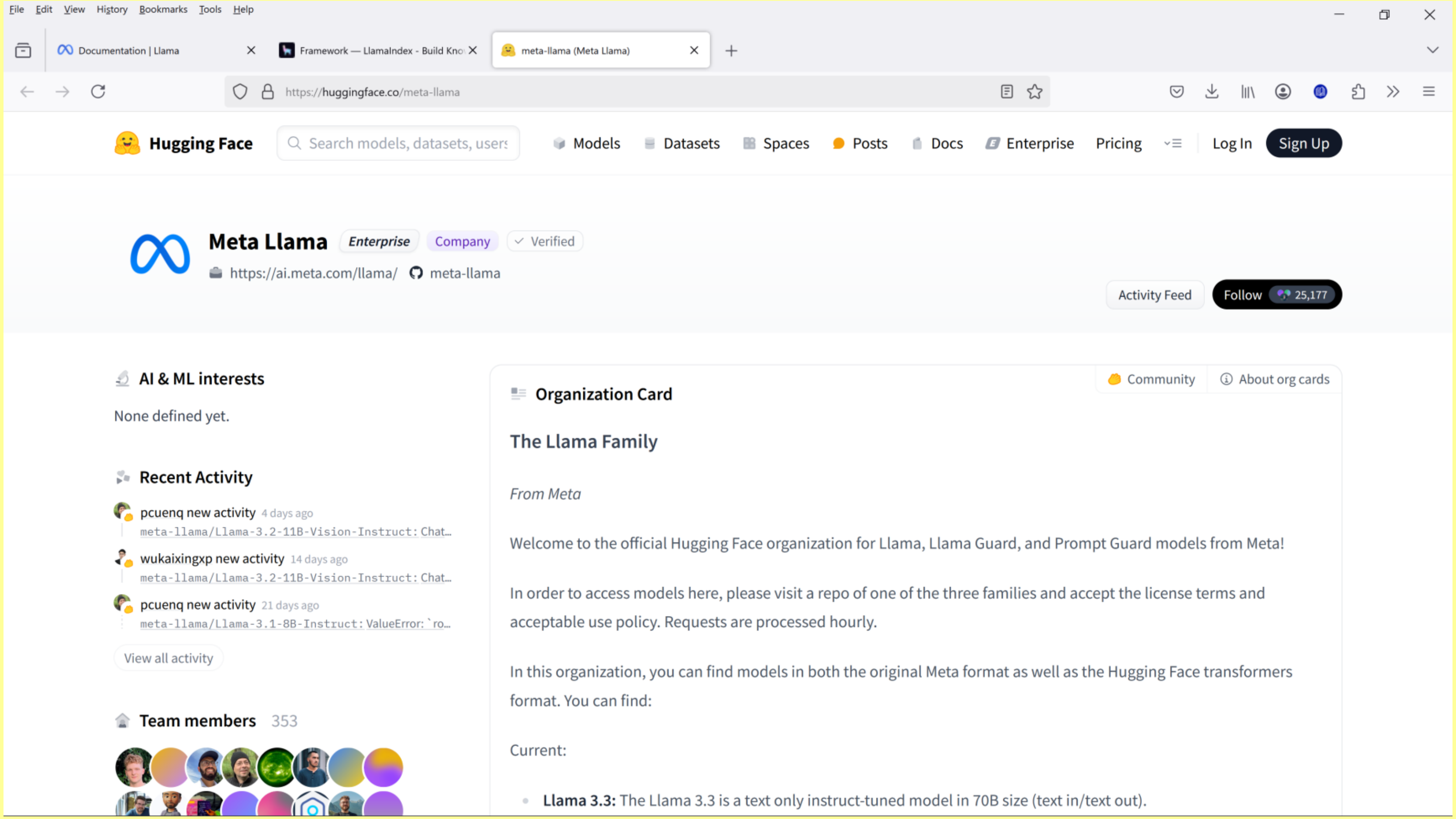
Get the Model

Meta ↗

Hugging Face ↗

HuggingFace : Repo for Open-Source Models

- https://huggingface.co/meta-llama



Meta Llama Enterprise Company ✓ Verified

https://ai.meta.com/llama/ meta-llama

AI & ML interests

None defined yet.

Recent Activity

pcuenq new activity 4 days ago
meta-llama/Llama-3.2-11B-Vision-Instruct: Chat...

wukaixingxp new activity 14 days ago
meta-llama/Llama-3.2-11B-Vision-Instruct: Chat...

pcuenq new activity 21 days ago
meta-llama/Llama-3.1-8B-Instruct: ValueError: `ro...

View all activity

Team members 353

Organization Card

The Llama Family

From Meta

Welcome to the official Hugging Face organization for Llama, Llama Guard, and Prompt Guard models from Meta!

In order to access models here, please visit a repo of one of the three families and accept the license terms and acceptable use policy. Requests are processed hourly.

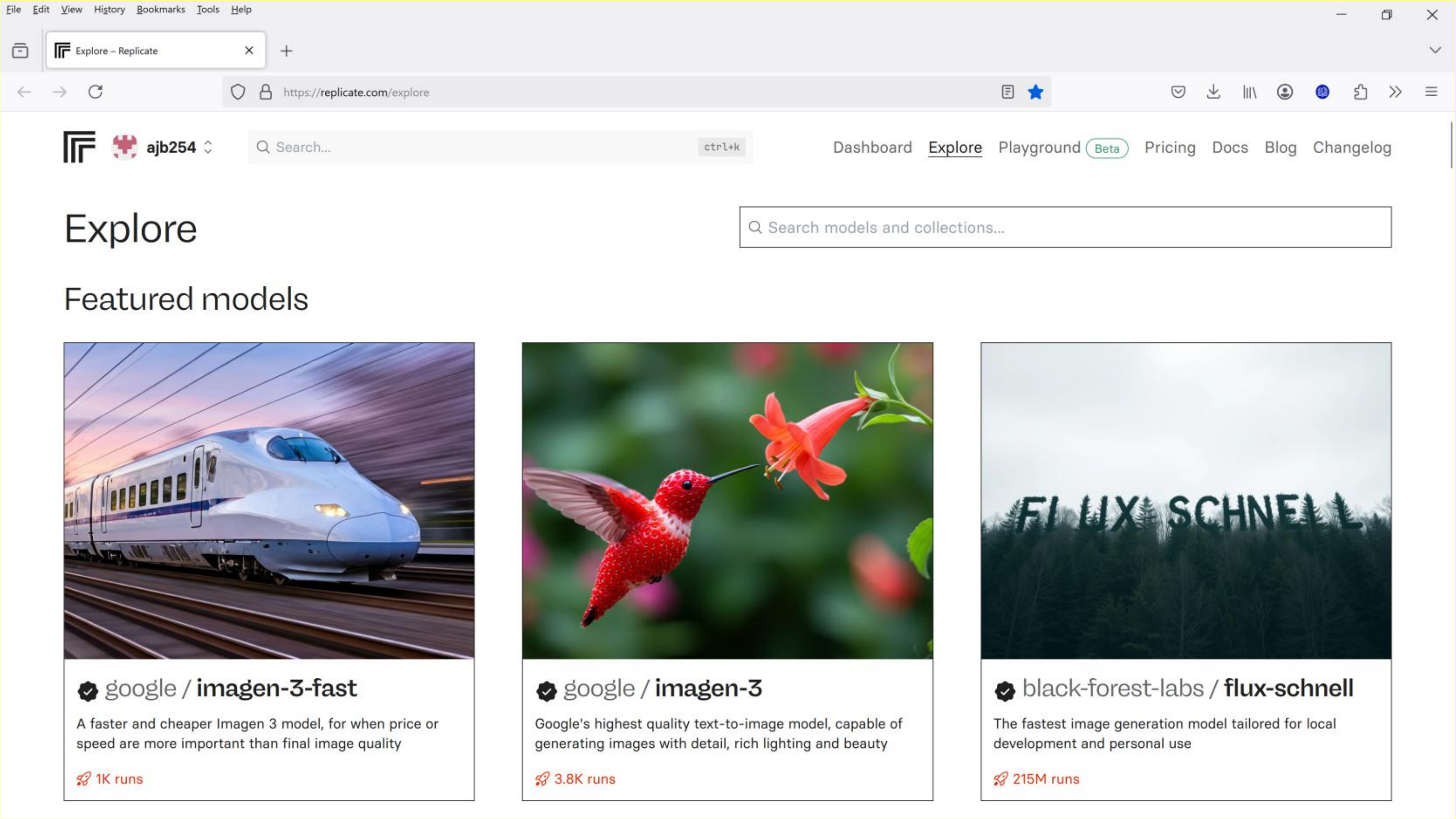
In this organization, you can find models in both the original Meta format as well as the Hugging Face transformers format. You can find:

Current:

- **Llama 3.3:** The Llama 3.3 is a text only instruct-tuned model in 70B size (text in/text out).

Replicate : Platform for Cloud-Hosted Open-Source/Public Models

- https://replicate.com/explore


File Edit View History Bookmarks Tools Help
Explore – Replicate
https://replicate.com/explore
ajb254
Search...
ctrl+k
Dashboard
Explore
Playground Beta
Pricing
Docs
Blog
Changelog
Search models and collections...

Explore

Featured models

google / imagen-3-fast

A faster and cheaper Imagen 3 model, for when price or speed are more important than final image quality

1K runs

google / imagen-3

Google's highest quality text-to-image model, capable of generating images with detail, rich lighting and beauty

3.8K runs

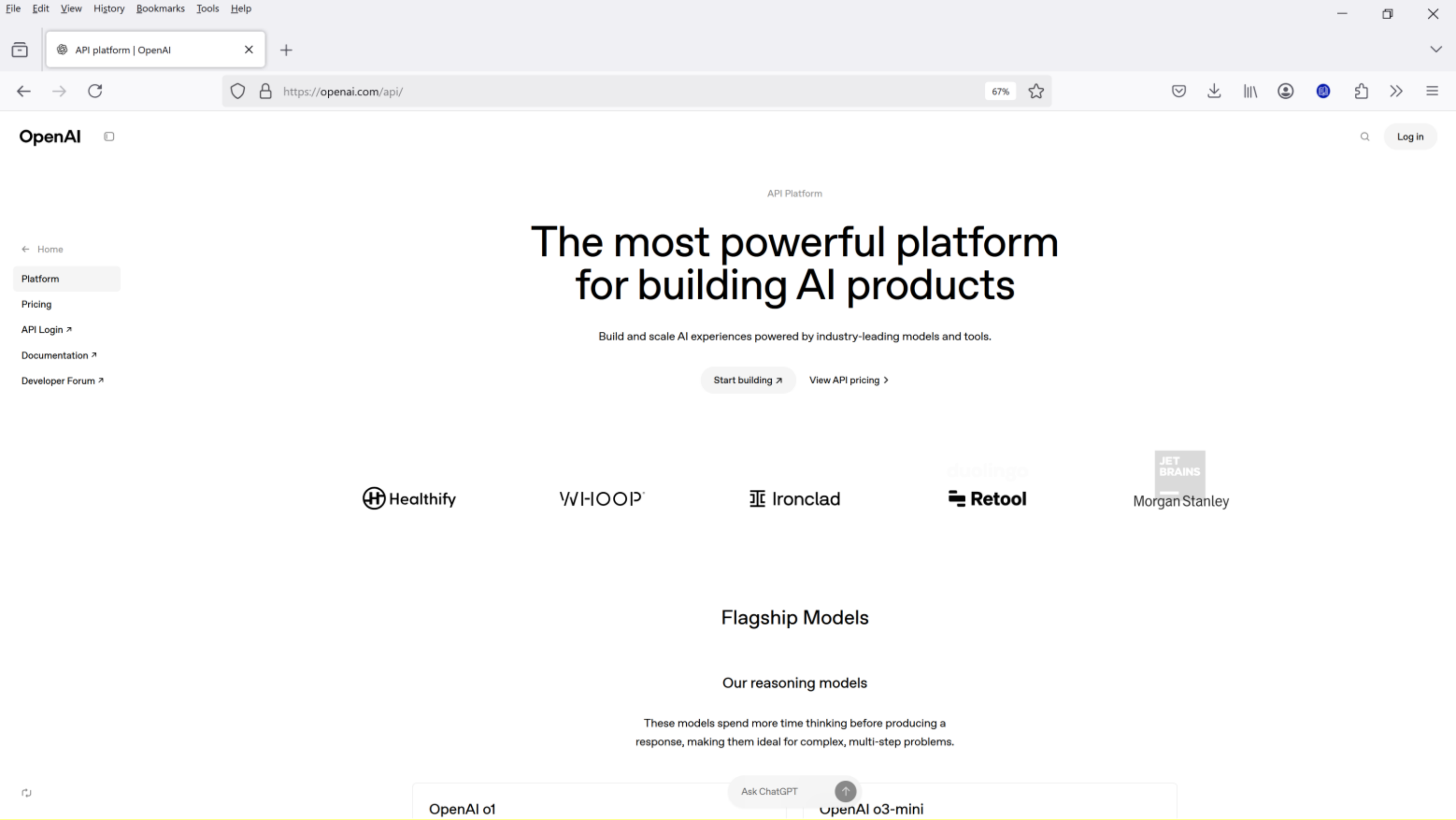
black-forest-labs / flux-schnell

The fastest image generation model tailored for local development and personal use

215M runs

OpenAI : Platform for Cloud-Hosted Proprietary Models

- https://openai.com/api/



API Platform

The most powerful platform for building AI products

Build and scale AI experiences powered by industry-leading models and tools.

Flagship Models

Our reasoning models

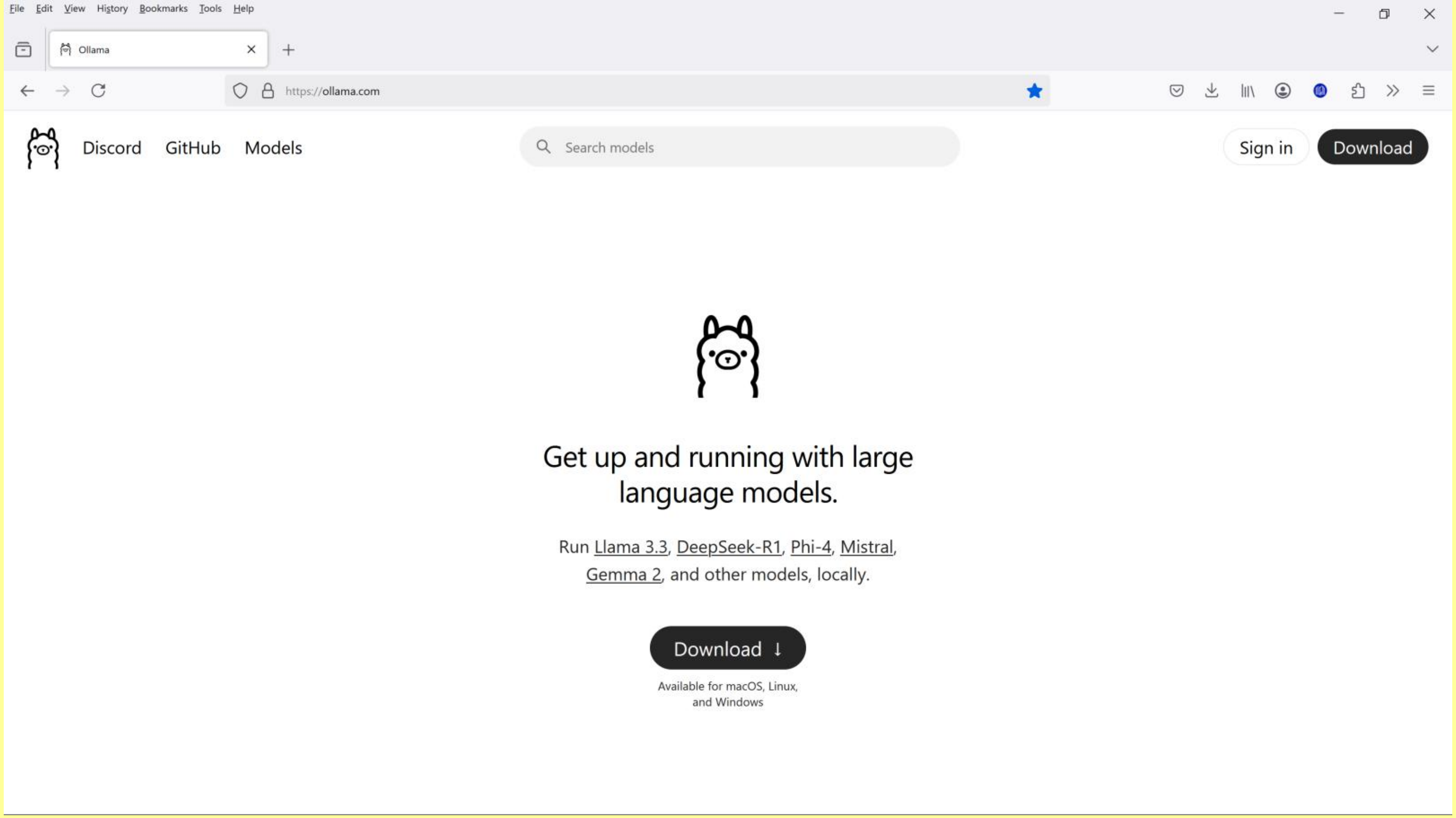
These models spend more time thinking before producing a response, making them ideal for complex, multi-step problems.

OpenAI o1

OpenAI o3-mini

Ollama : Container for Self-Hosted Open-Source Models

- https://ollama.com/



Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 2, and other models, locally.

Available for macOS, Linux, and Windows

Why LlamaIndex?

- LlamaIndex is the framework for Context-Augmented LLM Applications

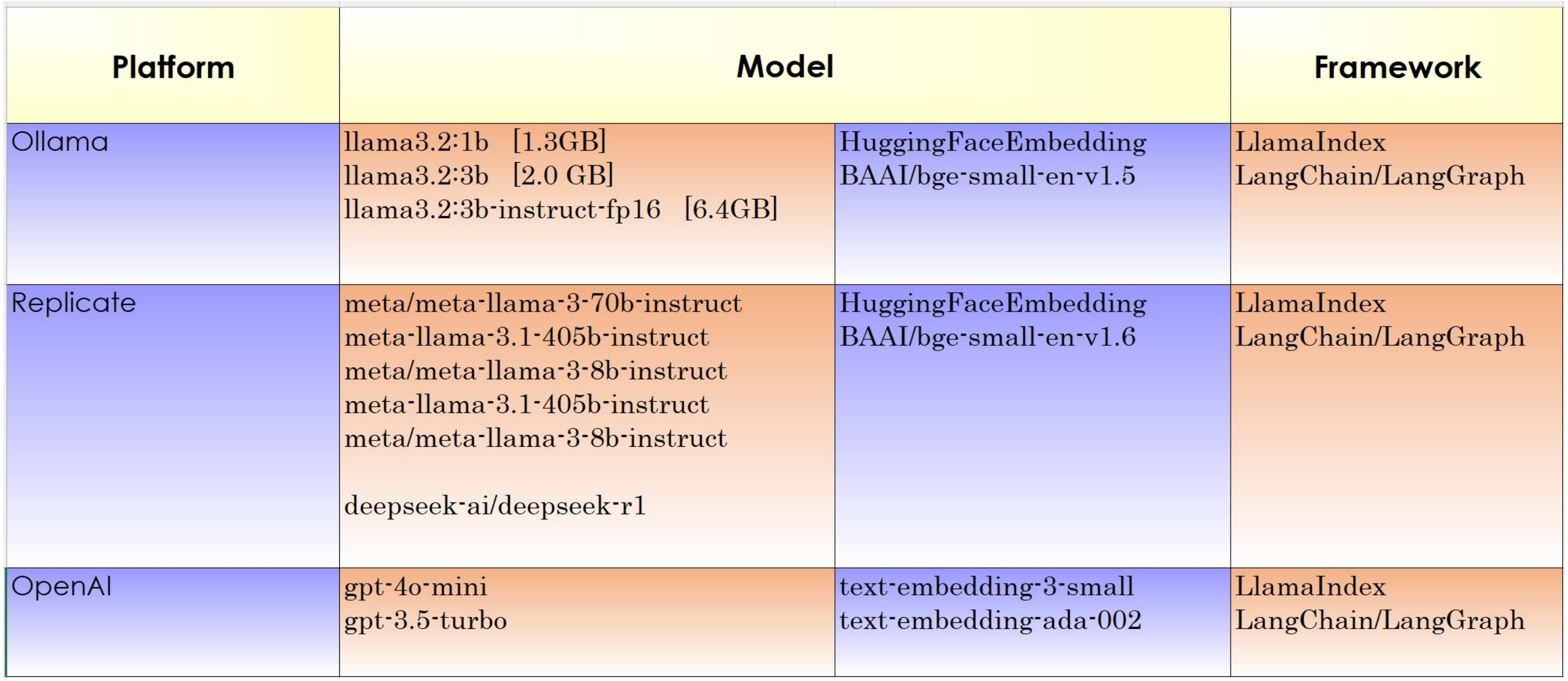
- LlamaIndex provides tools like:
  - **Data connectors** ingest your existing data from their native source and format. These could be APIs, PDFs, SQL, and (much) more.
  - **Data indexes** structure your data in intermediate representations that are easy and performant for LLMs to consume.
  - **Engines** provide natural language access to your data. For example:
    - **Query engines** are powerful interfaces for question-answering (e.g. a RAG flow).
    - **Chat engines** are conversational interfaces for multi-message, "back and forth" interactions with your data.
  - **Agents** are LLM-powered knowledge workers augmented by tools, from simple helper functions to API integrations and more.
  - **Observability/Evaluation integrations** that enable you to rigorously experiment, evaluate, and monitor your app in a virtuous cycle.
  - **Workflows** allow you to combine all of the above into an event-driven system far more flexible than other, graph-based approaches.

Why LlamaIndex?

- You can use LlamaIndex for Context-Augmented LLM Applications, when using:
  - OpenAI
  - Replicate
  - Ollama

- Here is a tutorial:
  - https://docs.llamaindex.ai/en/stable/getting_started/starter_example_local/

Benchmark & Testbed



Platform	Model		Framework
Ollama	llama3.2:1b [1.3GB] llama3.2:3b [2.0 GB] llama3.2:3b-instruct-fp16 [6.4GB]	HuggingFaceEmbedding BAAI/bge-small-en-v1.5	LlamaIndex LangChain/LangGraph
Replicate	meta/meta-llama-3-70b-instruct meta-llama-3.1-405b-instruct meta/meta-llama-3-8b-instruct meta-llama-3.1-405b-instruct meta/meta-llama-3-8b-instruct deepseek-ai/deepseek-r1	HuggingFaceEmbedding BAAI/bge-small-en-v1.6	LlamaIndex LangChain/LangGraph
OpenAI	gpt-4o-mini gpt-3.5-turbo	text-embedding-3-small text-embedding-ada-002	LlamaIndex LangChain/LangGraph

References

- https://ollama.com/
- https://replicate.com/google/imagen-3/examples?input=python
- https://replicate.com/google/imagen-3-fast
- https://deepmind.google/technologies/imagen-3/
- https://huggingface.co/meta-llama
- https://www.llamaindex.ai/
- https://www.llama.com/docs/overview/