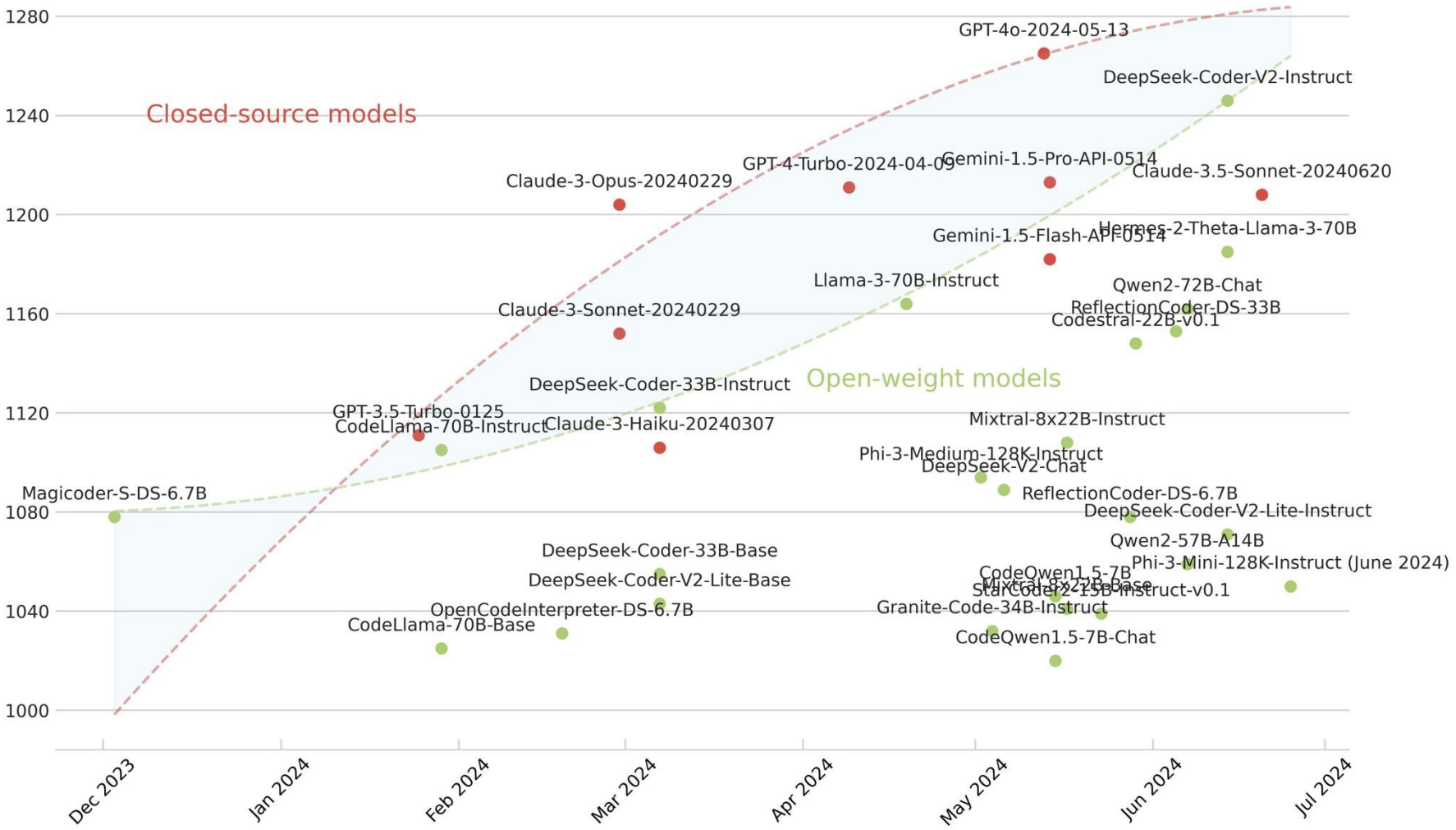
Open source LLM Ecosystem

Ram N Sangwan

- **Open source LLM Ecosystem**
- **Meta Llama 3 and Falcon Models**
- **Leveraging Models from Hugging Face**

Open Source LLMs



Open source LLMs

There has been a growing interest in open-source LLMs.

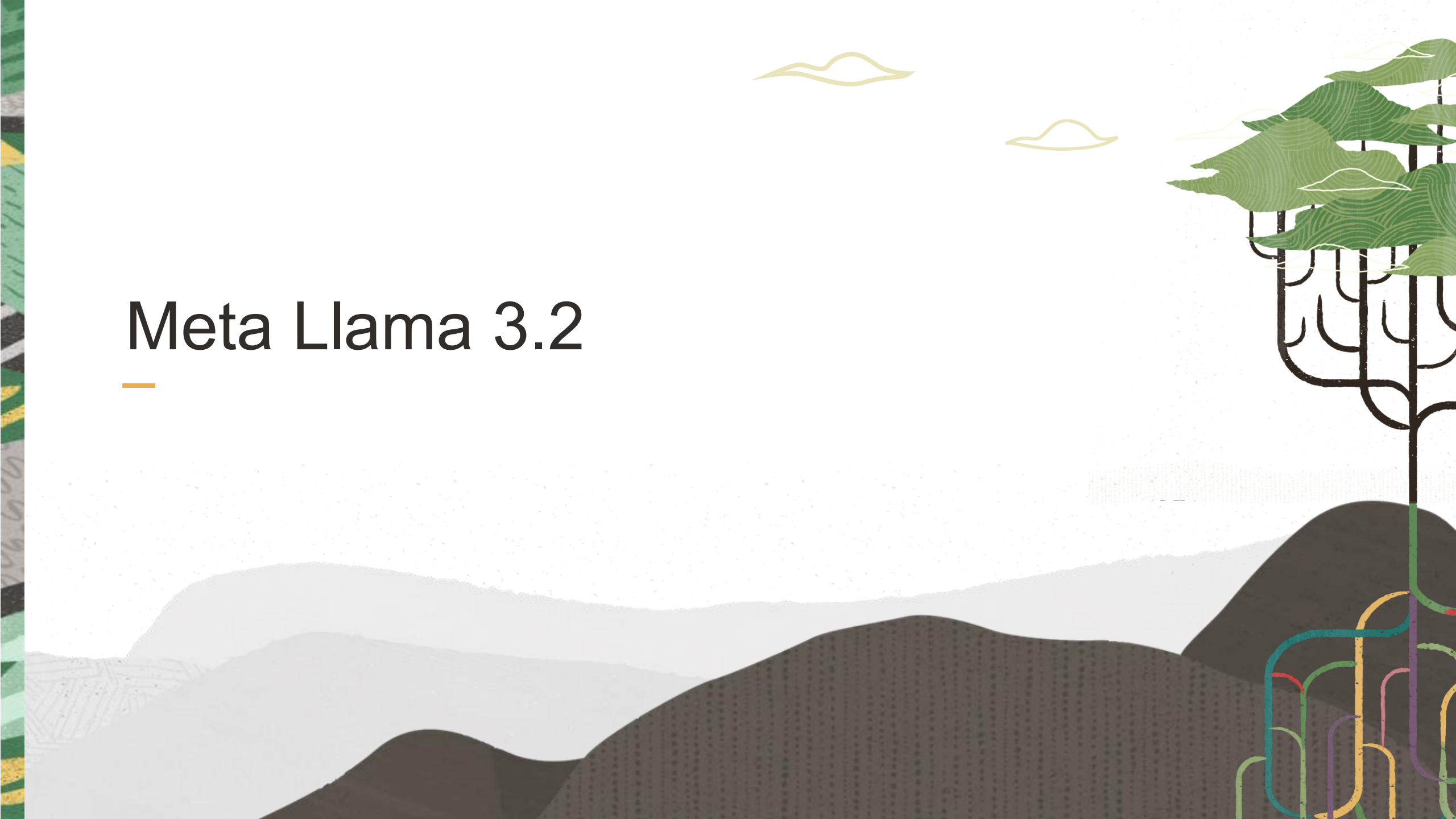
Benefits

- Affordable.
- Transparent - researchers can study how they work and how they make decisions.
- Flexible - they can be customized for different tasks.

Challenges

- Can be complex to use and to train.
- Can be computationally expensive to run.
- Can be used for malicious purposes, such as generating fake news or spam.

Meta Llama 3.2



Meta Llama 3.2 LLMs

Llama 3.2 Version Release Date: September 25, 2024

The Llama 3.2 is a collection of pretrained and instruction-tuned generative models in 1B and 3B sizes (text in/text out).

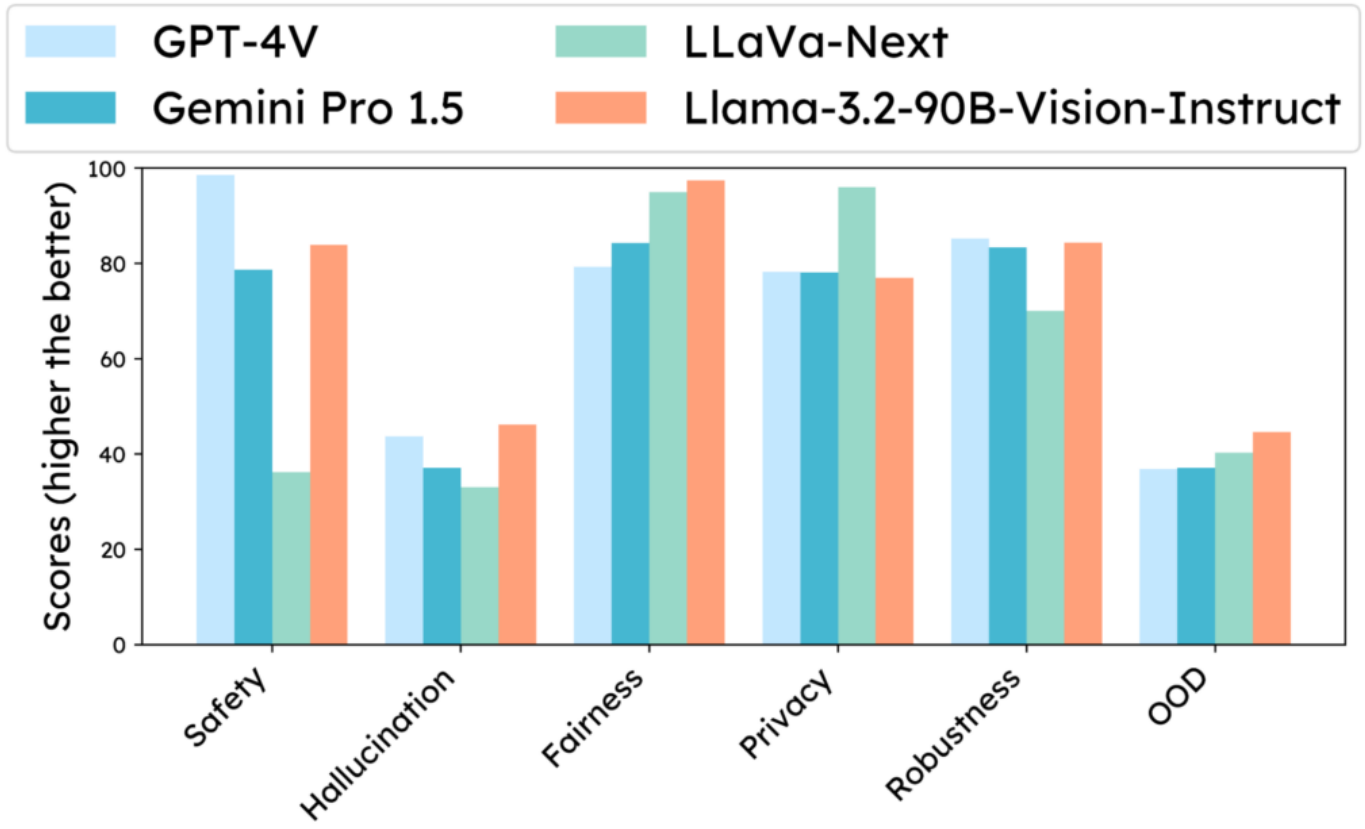
Llama 3.2 is open access —

- It's licensing allows almost anyone to use it and fine-tune new models on top of it.

Llama 3.2 is breaking records, scoring new benchmarks against all other "open access" models

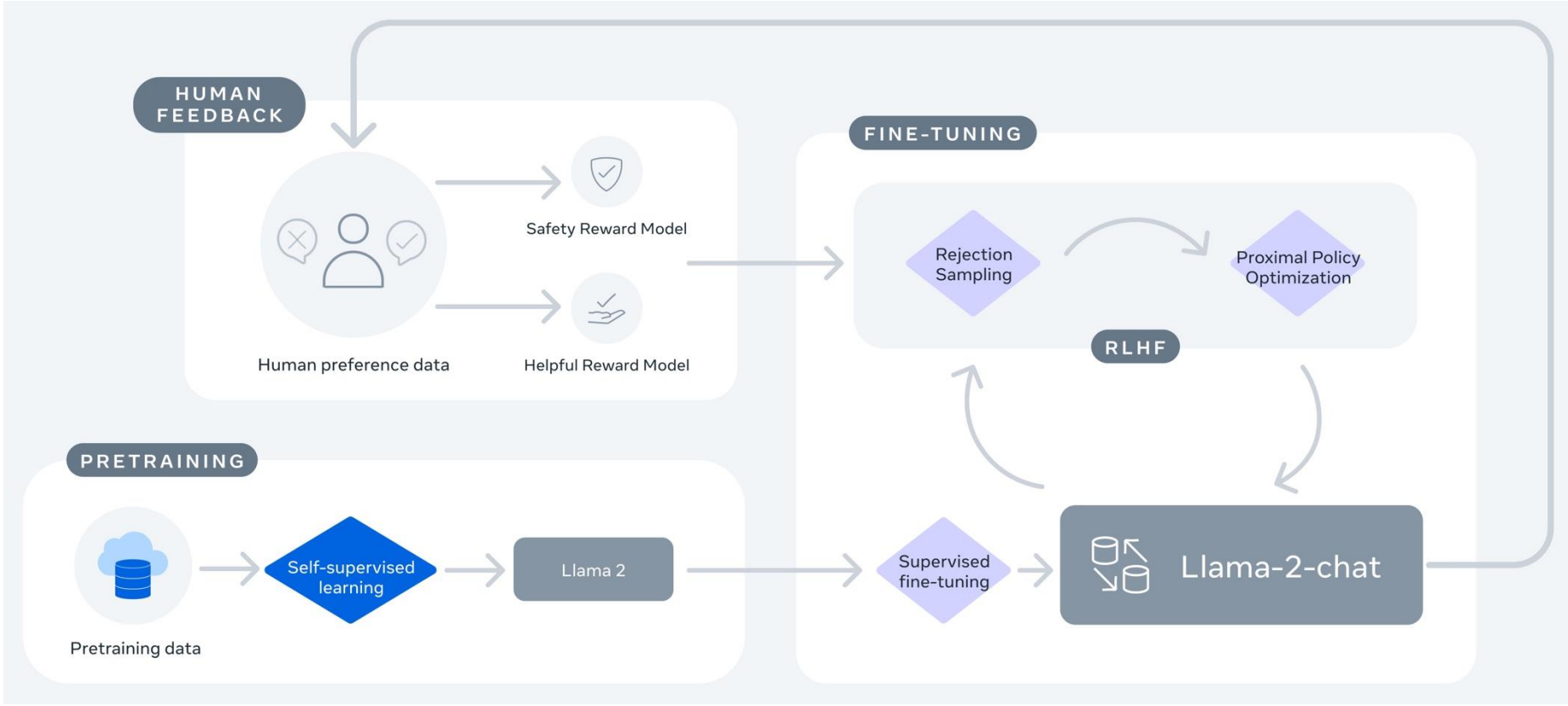
Safety Human Evaluation Results

- Llama-3.2-Vision has made a strides in safety, reducing its Harmful Content Generation Rate (HGR) to **16.1%**, a remarkable **74.76% decrease** from LLaVa-Next's **63.8%** HGR.
- Llama-3.2-Vision achieves an impressive Out-of-Distribution (OOD) score of 44.66, ranking it at the top among competitors like LLaVa-Next, Gemini 1.5 Pro, and GPT-4V.



Reinforcement Learning From Human Feedback

Llama Chat uses reinforcement learning from human feedback to ensure safety and helpfulness.



LLAMA 3.2 - Some Facts

Model Developers

Meta

Variations

Range of parameter sizes: 8B, 70B and 405B sizes (text in/text out.)

Input

Models input text only.

Output

Models generate text only.

Model Architecture

- Llama 3.2 is an auto-regressive language model that uses an optimized transformer architecture.
- The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

Try the Model at : https://www.meta.ai/

Llama 2 and Llama 3 Series of Models

	Llama 2.0 (7B, 13B, 70B)	Llama 3.0 (8B, 70B)	Llama 3.1 (8B, 70B, 405B)	Llama 3.2 Multimodal (11B & 90B)	Llama 3.2 Lightweight Text Only (1B & 3B)
Release Date	July 18, 2023	April 18, 2024	July 23, 2024	Sep 25, 2024	Sep 25, 2024
Context Window	4K	8K	128K	128K	128K
Vocabulary Size	32K	128K	128K	128K	128K
Official Multilingual	English Only	English Only	8 Languages	8 Languages	8 Languages
Tool Calling	No	No	Yes	Yes	Yes
Knowledge Cutoff	Sep 2022	2023, Mar (8B) Dec (70B)	Dec 2023	Dec 2023	Dec 2023

Llama 4 - with Vision and Text

The Llama 4 models are multimodal AI models that enable text and multimodal experiences.

These models leverage a mixture-of-experts architecture to offer industry-leading performance in text and image understanding.

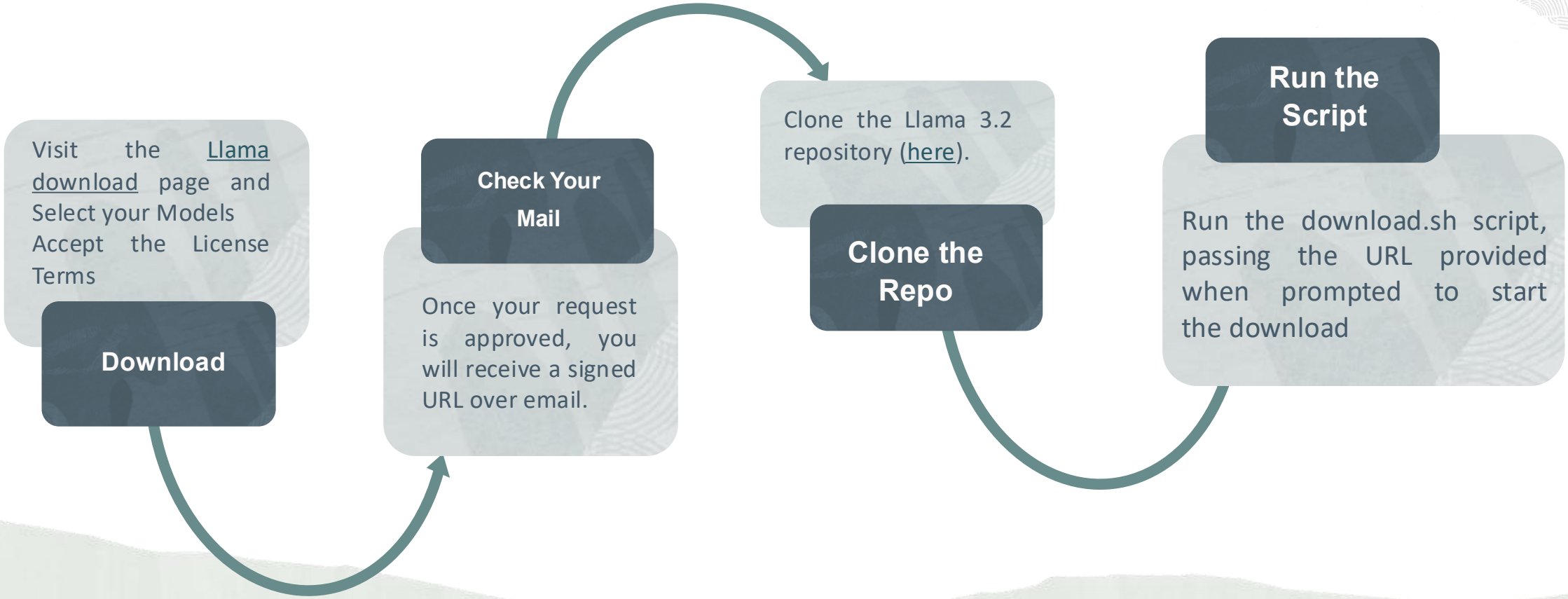
These Llama 4 models mark the beginning of a new era for the Llama ecosystem.

Two efficient models in the Llama 4 series,

1. Llama 4 Scout, a 17 billion parameter model
2. Llama 4 Maverick, a 17 billion parameter model.

Llama 4 Scout offers a context window of 10M and delivers better results than Gemma 3, Gemini 2.0 Flash-Lite, and Mistral 3.1

Getting the Models - Models from Meta AI (ai.meta.com)



Keep in mind that the links expire after 24 hours and a certain amount of downloads.

Getting Started with LLAMA

1

```
%pip install -qU \
replicate langchain \
sentence_transformers \
pdf2image pdfminer \
pdfminer.six \
unstructured \
pillow-heif opencv-python \
unstructured-inference pikepdf
```

2

```
import base64
from IPython.display import Image, display
import matplotlib.pyplot as plt
import os
from typing import Dict, List
from langchain.llms import Replicate
from langchain.memory import ChatMessageHistory
from langchain.schema.messages import get_buffer_string
```

3

```
from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv())
```

4

```
import replicate
```

Setup the Model

```
def llama3_8b(prompt):
    output = replicate.run(
      "meta/meta-llama-3-8b-instruct",
      input={"prompt": prompt}
    )
    return ''.join(output)

def llama3_70b(prompt):
    output = replicate.run(
      "meta/meta-llama-3-70b-instruct",
      input={"prompt": prompt}
    )
    return ''.join(output)
```

5

Basic completion

- With the model set up, you are now ready to ask some questions.
- An example of the simplest way to ask the model a questions.

```
prompt = "The typical color of a llama is: "
output = llama3_8b(prompt)
md(output)
```

6

System prompts

```
output = llama3_8b("The typical color of a llama is what? Answer in one word.")
md(output)
```

7

Falcon Models

Falcon is a generative large language model (LLM) that helps advance applications and use cases to future-proof our world.

Falcon 180B, 40B, 7.5B, 1.3B parameter AI models.

High-quality REFINEDWEB dataset, form a suite of offerings.

Falcon LLM is a foundational large language model developed by the Technology Innovation Institute (TII) in Abu Dhabi.

https://falconllm.tii.ae/falcon.html

Falcon 40B

- It features 40 billion parameters and is trained on one trillion tokens,
- Highly advanced and efficient model for generating text, solving complex problems, and being used in various applications such as chatbots, virtual assistants, language translation, content generation, and sentiment analysis.
- Highlighted for its ability to outperform other models like GPT-3, BLOOM, Chinchilla, and PaLM-62B, particularly in terms of the cost-effectiveness of its training compute and the quality of data used in its training.

Falcon 180B

- Falcon 180B is a super-powerful language model with 180 billion parameters, trained on 3.5 trillion tokens.
- At the top of the Hugging Face Leaderboard for pre-trained Open Large Language Models and is available for both research and commercial use..
- Performs exceptionally well in reasoning, coding, proficiency, and knowledge tests, even beating competitors like Meta's LLaMA 2.
- It ranks just behind OpenAI's GPT 4, and performs on par with Google's PaLM 2 Large, which powers Bard, despite being half the size of the model.

Leveraging Models from Hugging face

Hugging face

- You must first request a download using the same email address as your Hugging Face account.
- After doing so, you can request access to any of the models on Hugging Face and within 1-2 days your account will be granted access to all versions.

Thank You