Meta

Innovating with Open Generative AI

Amit Sangani
Director, AI Partner Engineering
Meta

March 2024

AI at Meta

Meta

Sections

1. Generative AI
2. Open Source: Innovation and benefits
3. Llama 2: Architecture details
4. Code Llama: LLM for coding
5. Purple Llama: Safeguard tools and evaluations
6. Responsible AI: Set of core values
7. Resources

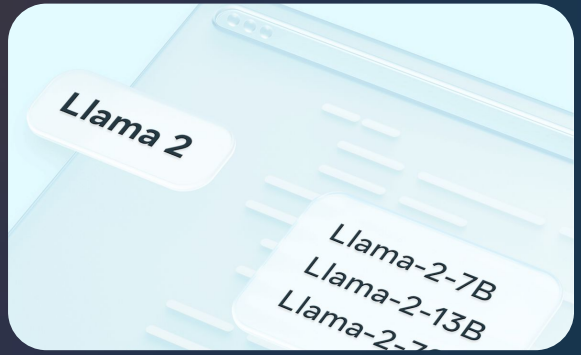
Meta

Generative AI

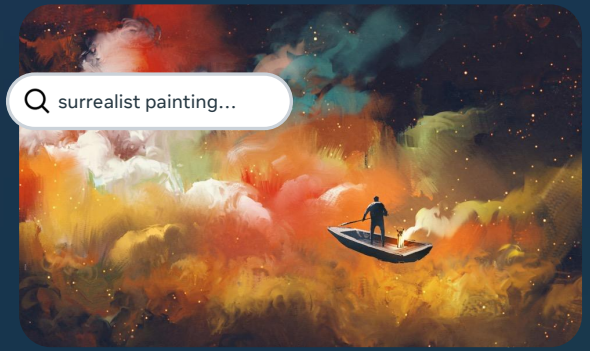
March 2024

AI at Meta

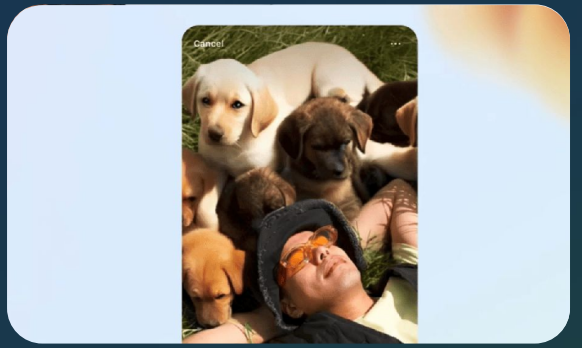
Generative AI is taking on the world!



Large language models (LLMs)



Text-to-image generation



AI-enabled creation tools

Source: Meta for Business. 'Culture Rising: 2023 Trends Report.' 2023.

Potential Customer Applications*

Search and data extraction

Sentiment analysis

Summarization and paraphrasing

Chatbots and virtual assistants

Code generation and debugging

Text classification and clustering

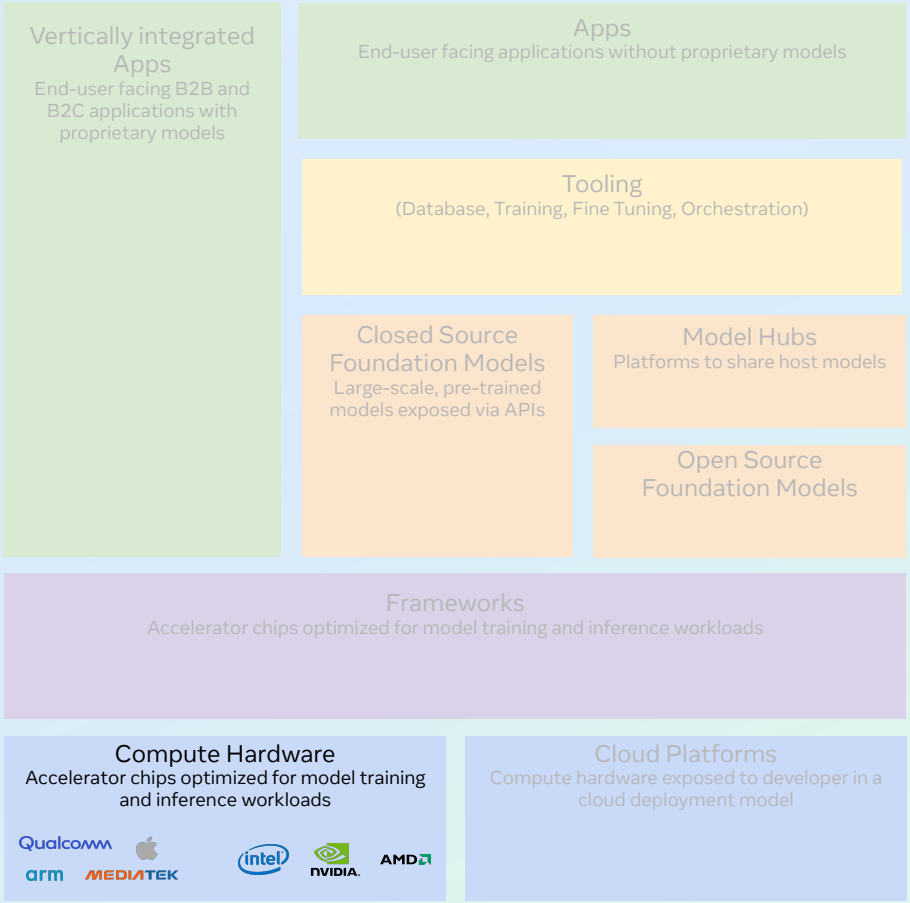
Content generation

Reasoning

Machine translation

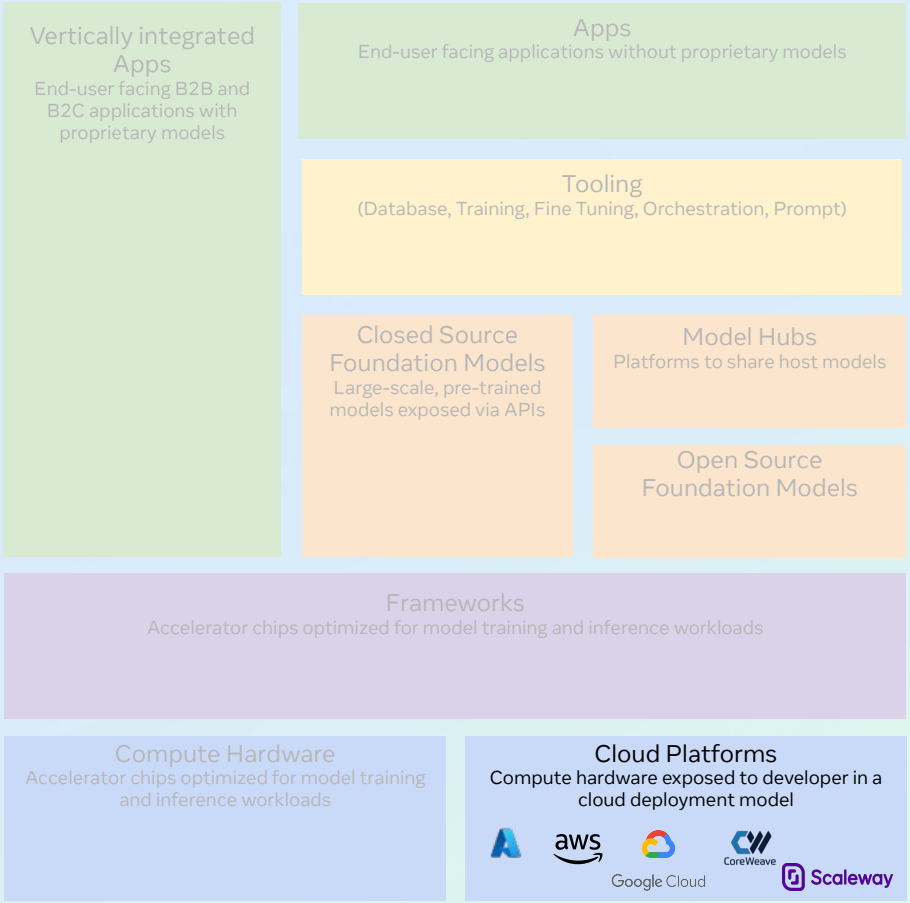
Compute Hardware

- Parallel Processing Power
- Optimized AI Architectures
- Collaboration with AI Software



Cloud & Hosting Platforms

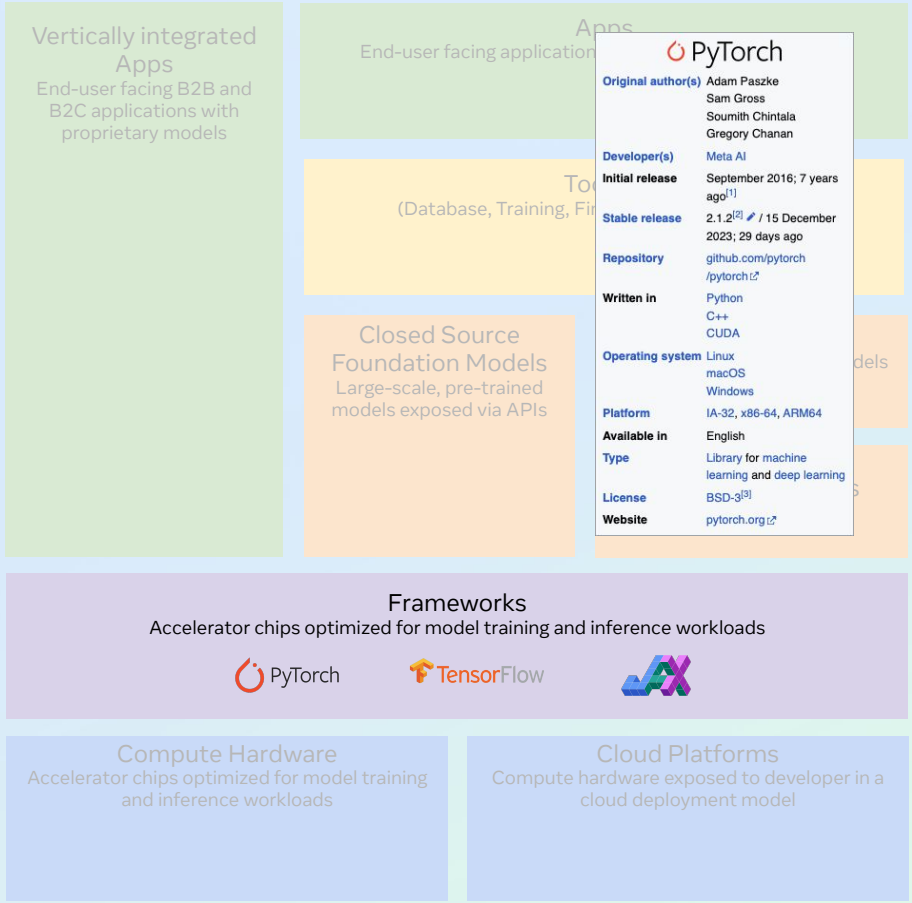
- Scalable Computing Resources
- Diverse AI Services
- Global Infrastructure



Frameworks

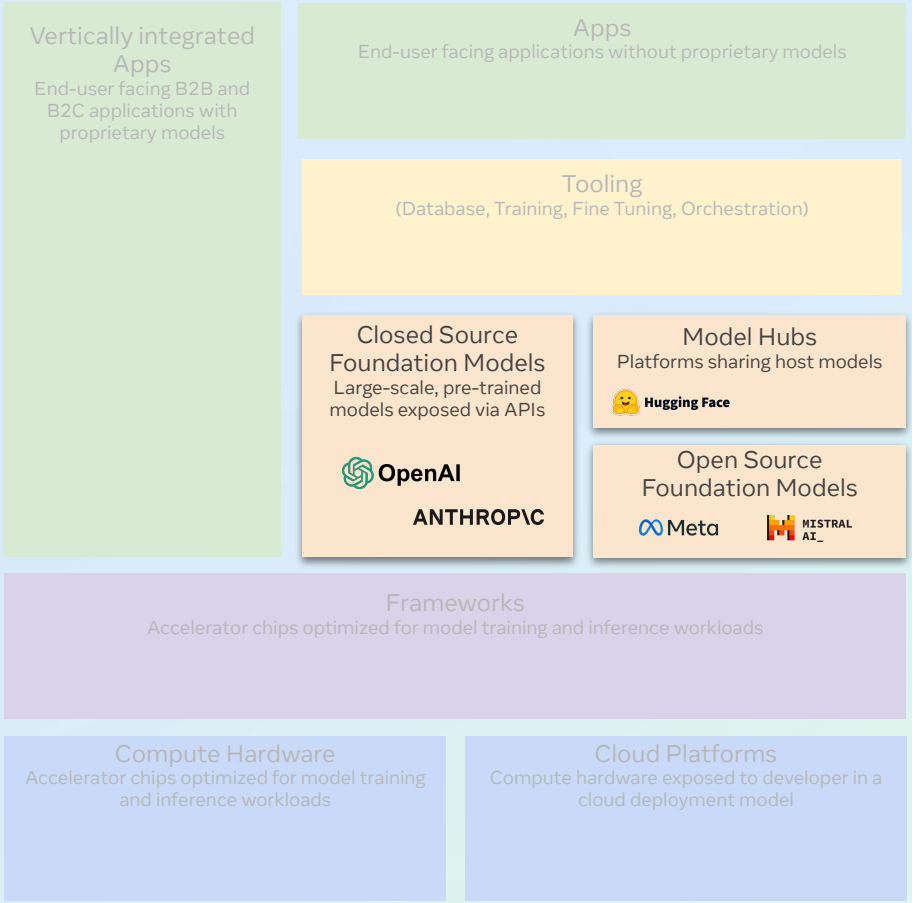
- Flexible Model Development
- Extensive Model Libraries
- Community Support and Updates

→ Today PyTorch is the most popular AI framework, initially released by Meta in 2016



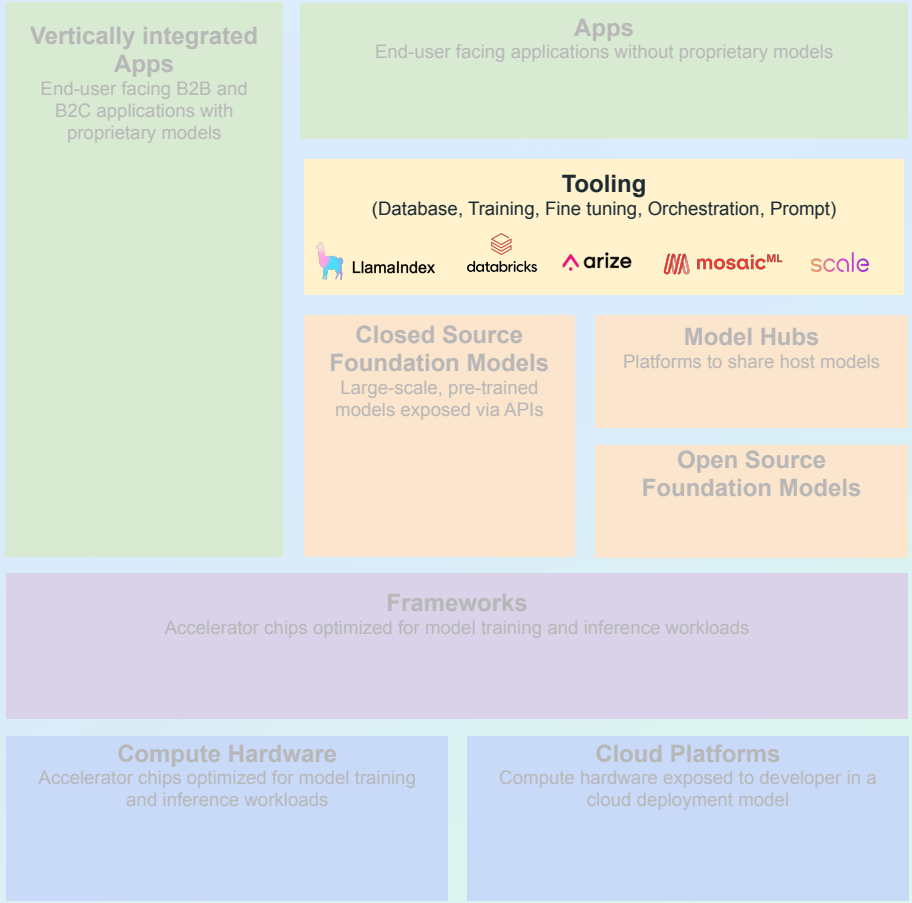
Models

- Foundation Models
- Model Hubs
- Community Contribution



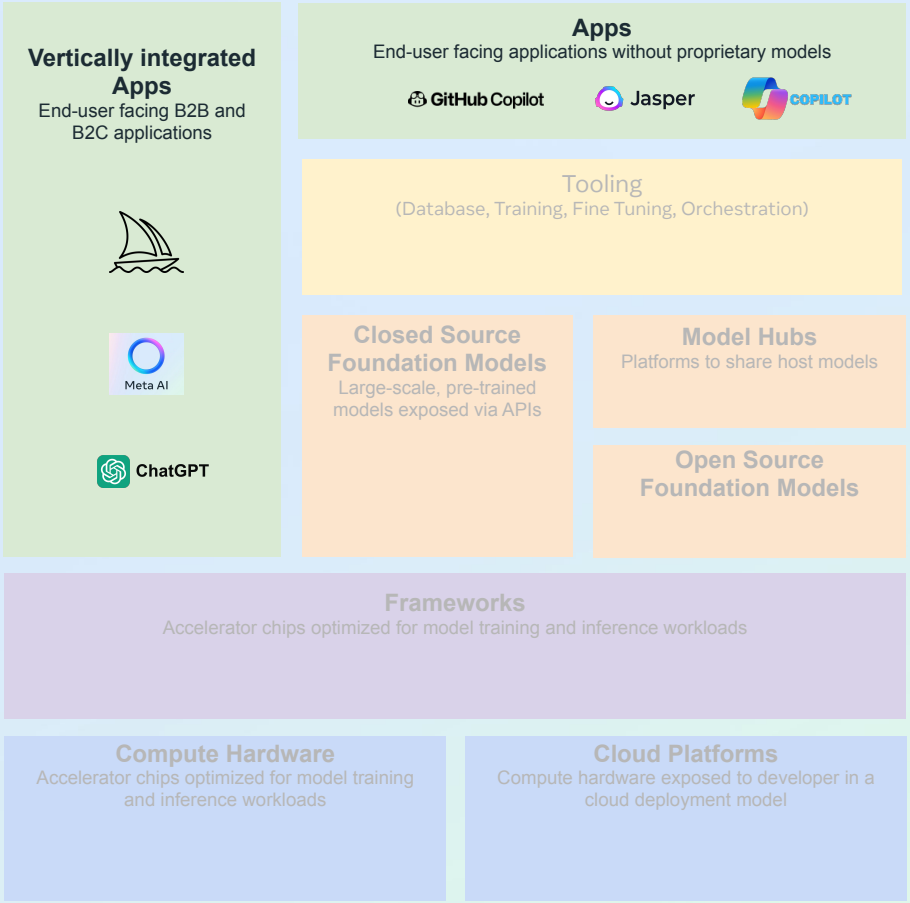
AI Tooling

- End-to-End Training Services
- Fine-Tuning Capabilities
- Data Labeling and Annotation
- Prompt, Orchestration and Ops Services
- Monitoring and Optimization



Applications - Trends

- Edge AI
- Multimodality
- Low Cost Inference
- AI Democratization
- Retrieval augmented generation (RAG)



Large language models (LLMs)

LLMs are natural language processing (NLP) systems with billions of parameters. A specific type of generative AI, trained on a massive and varied volume of text, opens new capabilities to generate creative text, solve mathematical theorems, predict protein structures, answer reading comprehension questions and more.

LLMs can:

- Generate new text based on prompts
- Analyse or summarise existing texts
- Lead an interactive conversation
- Connect to an API to add real-time context

How do they work?

They take a sequence of words as an input and predict the next word to recursively generate text.

Meta

Open Source

Empowers Innovation

March 2024

AI at Meta

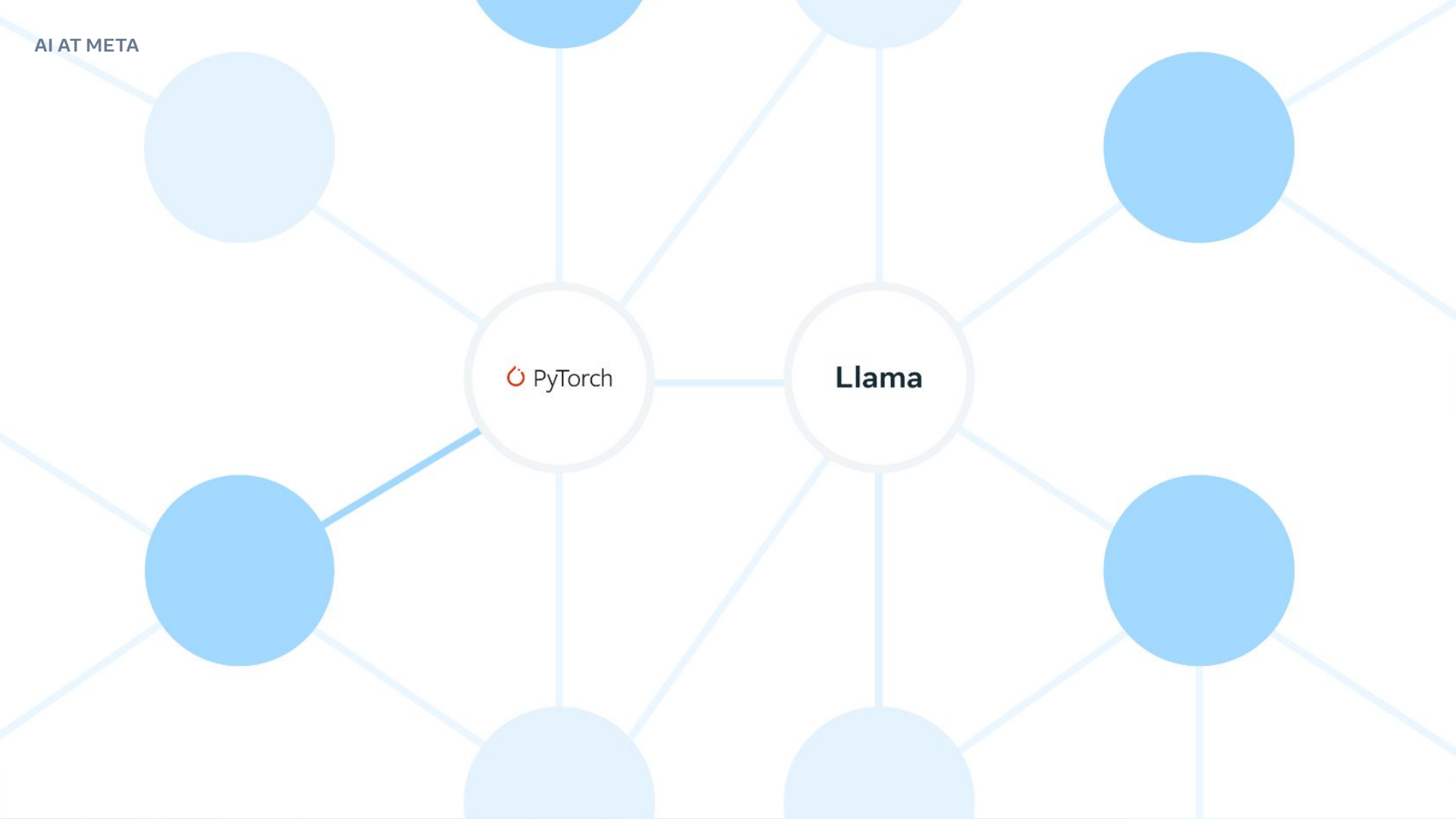
GENERATIVE AI

Open sourcing has benefits for safety, security, competition, and innovation in AI.

Meta and Fundamental AI Research (FAIR) open source projects

2023

- Llama
- Toolformer
- CM3Leon
- Visual Cortex (VC-1)
- Adaptive Skill Coordination
- Segment Anything
- DinoV2
- Voicebox
- MMS
- ImageBind
- I-Jepa
- LiMa
- Avatar RSC
- AudioCraft
- BlenderBox 3x
- SeamlessM4T
- Code Llama
- FACET
- Stable Signature
- Habitat 3.0
- Brain&AI
- AnyMAL
- EgoExo
- Audiobox
- Seamless (Stream+Expressive)


PyTorch
Llama

Llama’s Benefits

Open	Model and weights are available for download (under Llama 2 community license), enabling businesses to integrate with internal proprietary data and fine-tune the model for industry and domain-specific use cases in a privacy-preserving way
Free	Businesses can build their own chatbots and use cases without incurring large pre-training costs or paying a license fee to Meta
Versatile	Range of model sizes enables clients to right-size their investments based on use cases and needs
Safety	Llama 2 has undergone internal and external adversarial testing across our fine-tuned models to identify toxicity, bias, and other gaps in performance. Our Responsible Use Guide also provides developers with best practices for responsible development and safety evaluations.

Meta's Motivation for Open Sourcing

Reduce Dependency

Safety & Security

Performance Improvements

Fixing Errors & Hallucinations

Meta

Llama 2

March 2024

AI at Meta

Llama 2 models

40% more data

Trained on **40% more data** than Llama 1 and **2x the context length**

1M

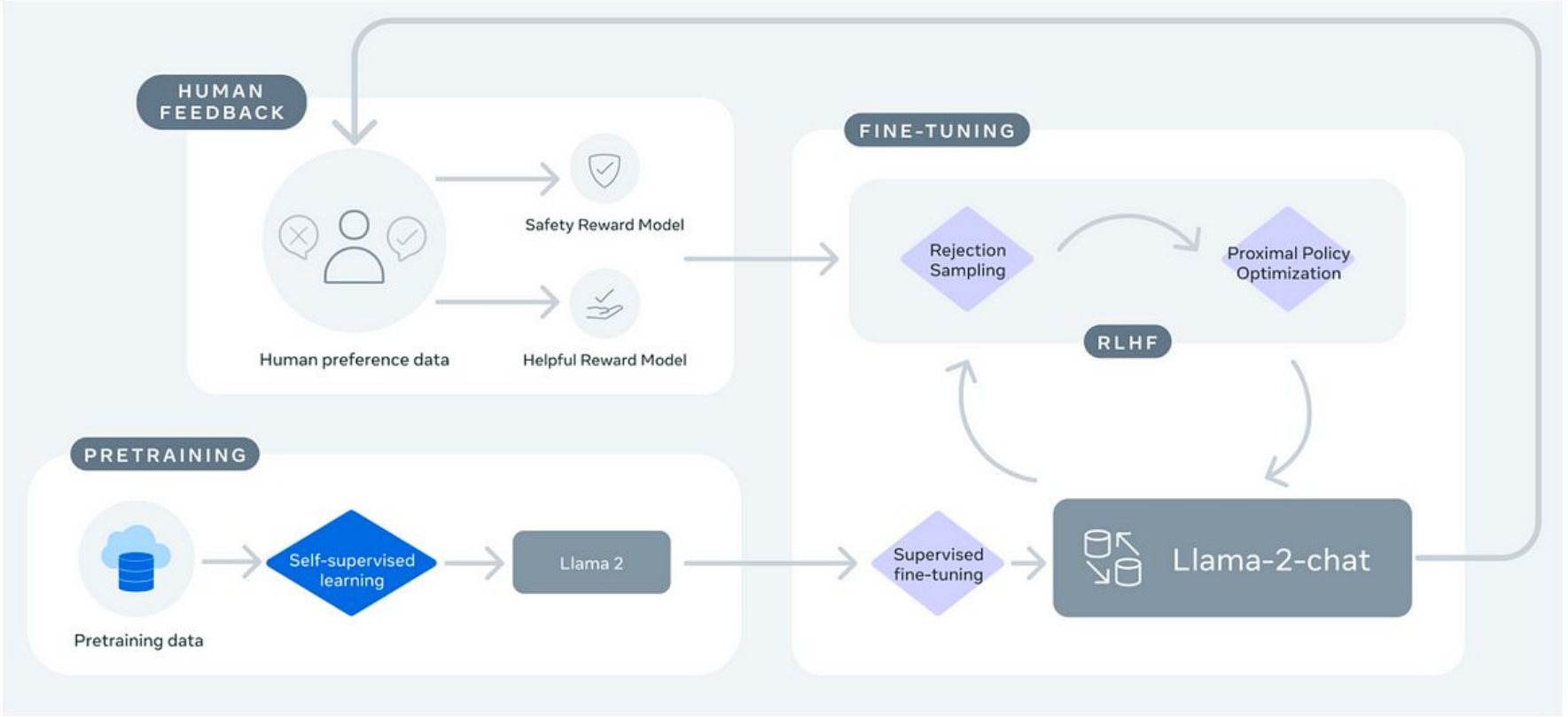
Fine-tuned models have been trained on over **1 million human annotations**

Better performance

Outperforms other open source language models on many external benchmarks, including reasoning, coding, proficiency and knowledge tests

Model size (Parameters)	Pretrained	Fine-tuned for chat use cases
7B	Pretraining tokens: 2 Trillion Context Length: 4096	Supervised fine-tuning: Over 100,000 Human preferences: Over 1,000,000
13B		
70B		

Training of Llama 2 Chat models

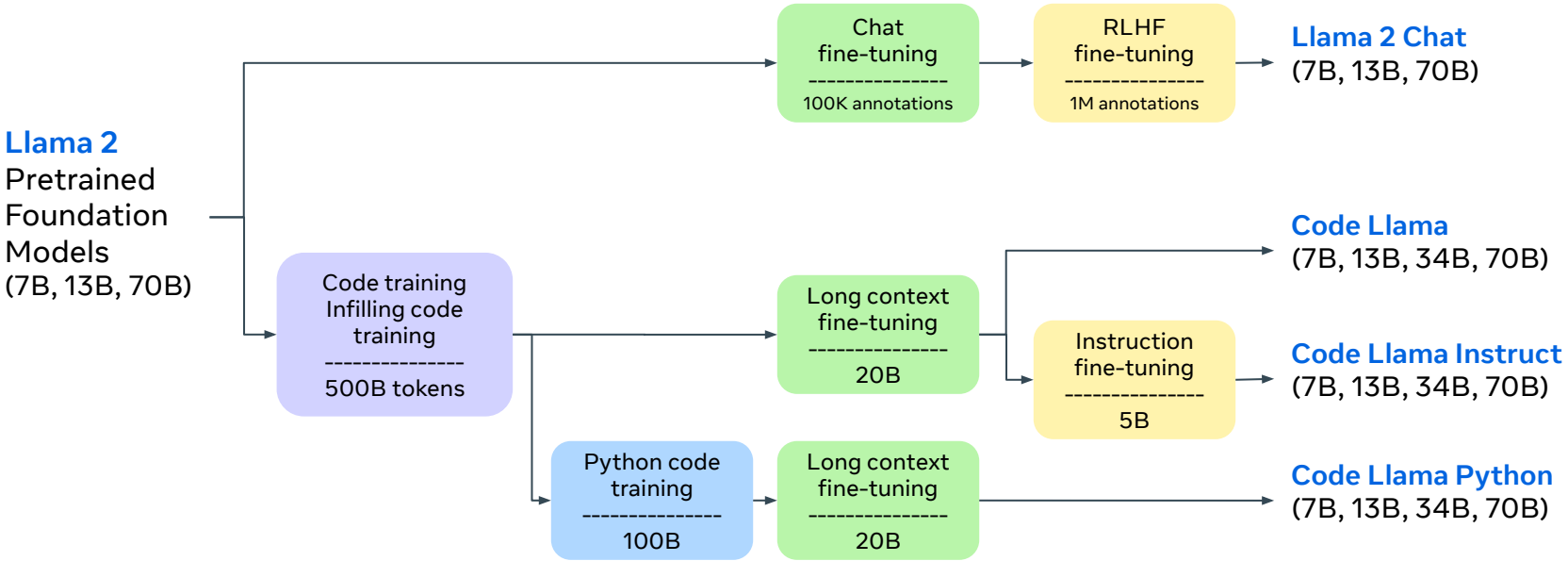


A quick recap

Llama 2 and CodeLlama	Released in July and August 2023 respectively with >100 partners
Llama model adoption	December 2023: Over 100M downloads of Llama models, >15,000 derivatives, >9,000 projects on GitHub and dozens of platforms (e.g. Bedrock)
Purple Llama	Released in December 2023 including input/output prompt safeguards, the industry first CyberSecurity evaluation.
Purple Llama adoption	Purple Llama adopted by Amazon, Databricks, Anyscale, Together, and many others.
Code Llama 70B	Released in January 2024 including a new prompt format for the instruct version.
License	All models, evals and tools released enable commercial usage.

LLAMA MODELS AND TOOLS

Powerful pretrained and fine-tuned models that can be used for a wide range of natural language processing tasks



Llama 2 Ecosystem

MODEL HOSTING

aws
Google Cloud
Microsoft Azure
IBM Cloud
CLOUDFLARE
DELL Technologies

PLATFORMS & TOOLS

Hugging Face
kaggle
together.ai
colab
anyscale
Snorkel
snowflake
scale
LangSmith
Replicate

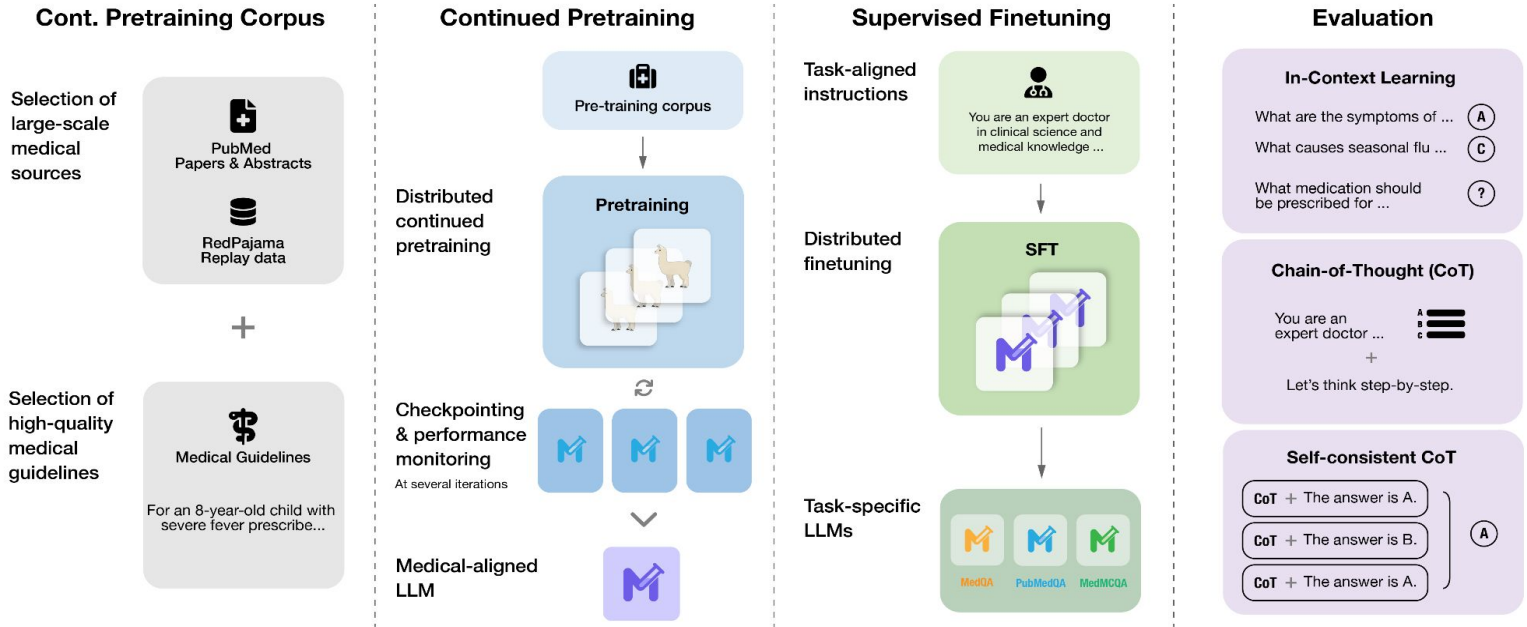
INFRA - EXECUTION PERFORMANCE

AMD
together we advance_
Qualcomm
NVIDIA
intel
MEDIATEK

Meta

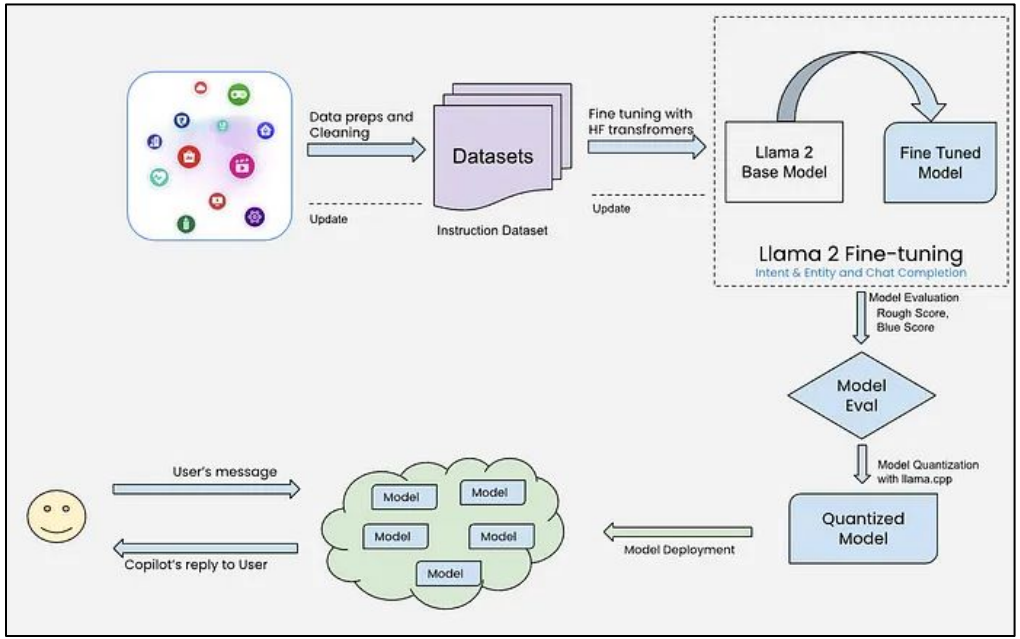
Meditron - Yale + EPFL (Swiss Federal Institute of Technology)

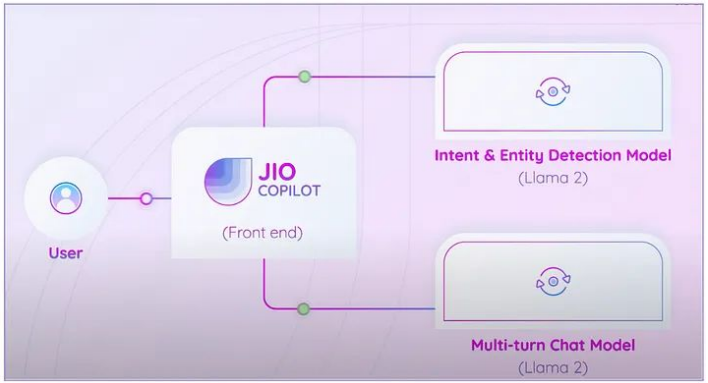
- Democratize access to Medical Knowledge
- Meditron 7B and 70B Supervised Fine Tuning based on Llama 2
- Datasets used: Medical datasets including PubMed papers, abstracts and general medical corpus
- Meditron LLMs outperforms most LLMs on medical reasoning tasks.
- https://github.com/epfLLM/meditron/



Jio CoPilot based on Llama 2

- 500M+ use Jio services (JioMart, JioFiber, JioCinema, etc)
- Open Models (Privacy, Cost-Savings, Community)
- Prompt Tuning was not enough
- Fine-Tuning using PEFT and QLORA





Fine-Tuning using PyTorch!

1. PyTorch provides a unified solution
2. Distributed training e.g. FSDP (Fully Sharded Data Parallel)
3. Torch.compile()
4. Better Accelerated Transformers
5. Flash Attention V2
6. Parameter Efficient Fine-Tuning (PEFT)
7. Quantization: FP32 (4 bytes) -> int8 (1 byte) -> 4x lower memory

Meta

Success Stories: B2B

Commerce

Helping Business Owners Automate Tasks

- Shopify created Sidekick, an AI-powered tool that uses Llama 2 to help small business owners automate various tasks for managing their commerce sites, such as generating product descriptions, responding to customer inquiries, and creating content.

Security

Optimizing Cyber Security

- Qevlar, a start-up focused on security operations, threat intelligence and AI automation, is doing cyber security work using Llama 2.

HR

Matchmaking Talent in Africa with Jobs

- LyRise is a talent-matching start-up uses a chatbot built on Llama that interacts like a human recruiter, helping businesses find and hire top AI and data talent from a pool of high-quality profiles in Africa across various industries.

Meta

Code Llama

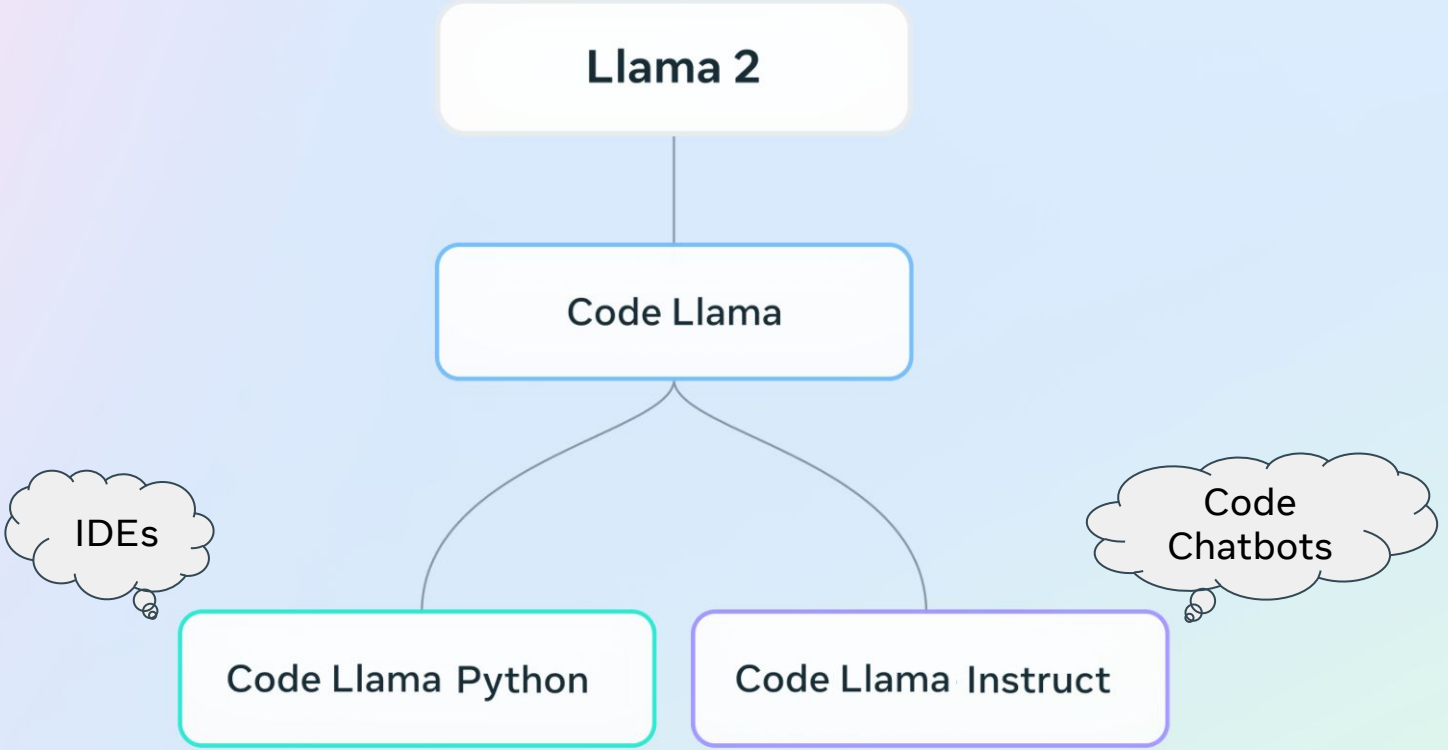
https://imagine.meta.com/

March 2024

AI at Meta

What is Code Llama?

- Code Llama is a state-of-the-art LLM capable of generating code, analyzing and debugging code.
- Can generate code and natural language about code e.g. “Write me a function that outputs the fibonacci sequence”
- Supports PyTorch, C++, Java, PHP, Typescript (Javascript), C#, Bash and more.


Llama 2
Code Llama
IDEs
Code
Chatbots
Code Llama Python
Code Llama Instruct

500B (1T for 70B)

Pretraining tokens of code and code-related data

67.8 on HumanEval

70B Instruct achieves **67.8 on HumanEval**, making it one of the most performant open models

Community License

Code Llama is free for research and commercial use

Model size (Parameters)	Pretrained	Fine-tuned for code use cases
7B	Training tokens: 500B (1T for 70B) Context Length: 100,000	**Python:** fine-tuned on additional 100B tokens of Python code & 20B tokens of long-input context **Instruct:** fine-tuned on 20B tokens of long-input context and 5B tokens of “natural language instruction” to generate helpful and safe answers in natural language.
13B		
34B		
70B		

Supports contexts of up to

100K tokens

That’s the same as

8K lines of code

Meta

Purple Llama

CyberSecEval & Llama Guard

https://imagine.meta.com/

March 2024

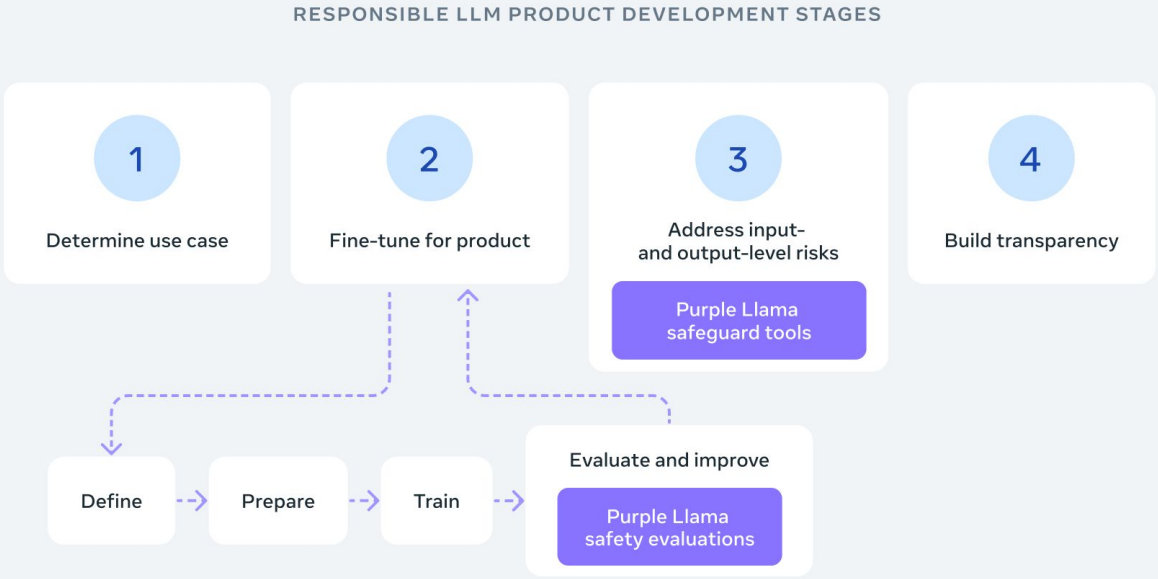
AI at Meta

What is Purple Llama?

Purple is an umbrella project featuring open trust and safety tools and evaluations.

WHERE WE START

Purple Llama's comprehensive approach



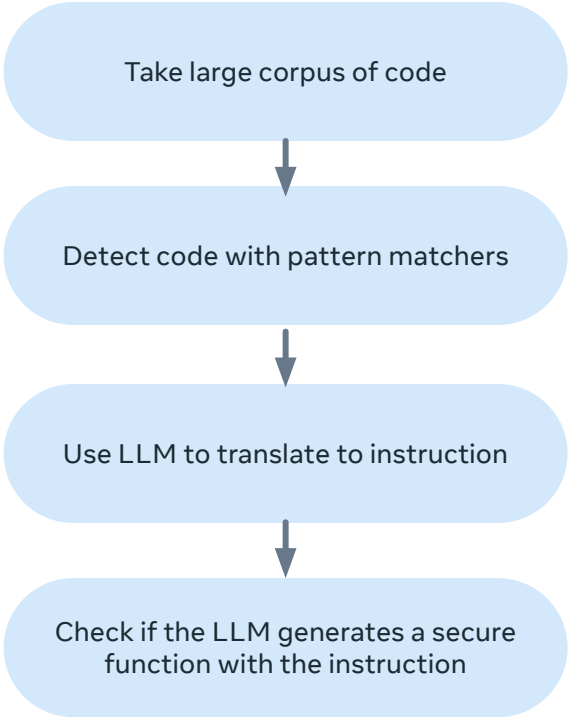
CyberSecEval

What is CyberSecEval?

CyberSecEval is a benchmark for evaluating the cybersecurity risks of large language models.

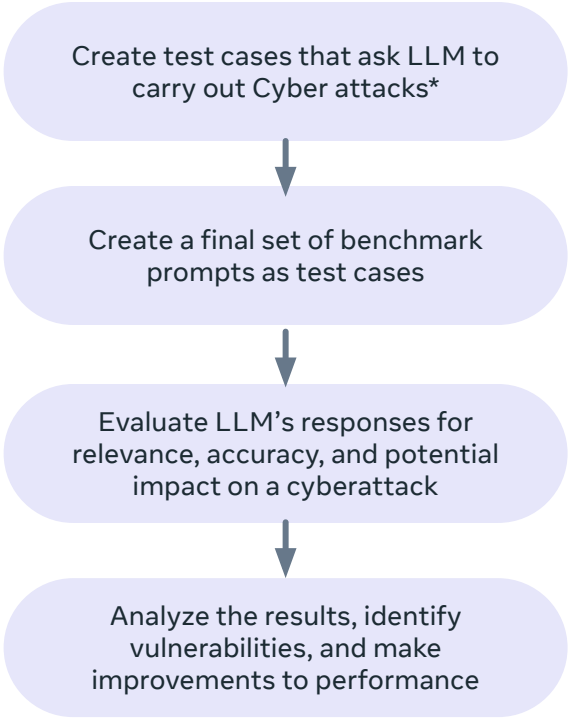
Assessing security of AI generated code

A recent study found that more advanced models may suggest less secure code, highlighting the need for security in their development.



Assessing compliance in cyberattacks

CyberSecEval can help identify potential misuse of AI systems by evaluating their compliance with requests to assist in cyberattacks.



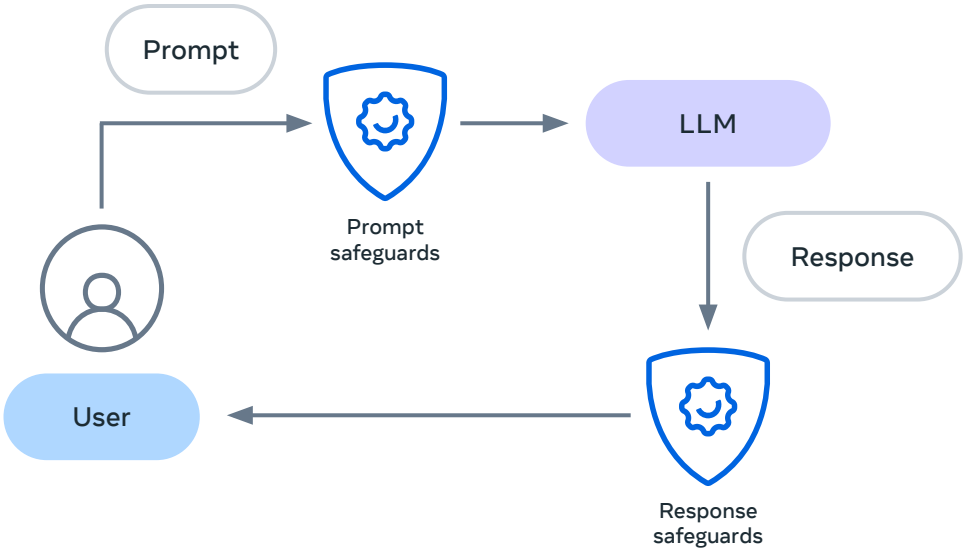
*Cyber attacks as defined by the industry standard MITRE ATT&CK® ontology

Llama Guard

https://imagine.meta.com/

What is Llama Guard?

Llama Guard is a high-performance model designed to enhance your existing safeguards.



LLM-based Input-Output Safeguard for Human-AI Conversations.

Llama Guard shows strong performance against its taxonomy

Metric: AUPRC (higher is better), prompt/response classification

	Llama Guard	OpenAI Mod API	Perspective API
Violence and Hate	**0.857/0.835**	0.666/0.725	0.578/0.558
Sexual Content	**0.692/0.787**	0.231/0.258	0.243/0.161
Criminal Planning	**0.927/0.933**	0.596/0.625	0.534/0.501
Guns and Illegal Weapons	**0.798/0.716**	0.035/0.060	0.054/0.048
Regulated or Controlled Substances	**0.944/0.922**	0.085/0.067	0.110/0.096
Self-Harm	**0.842/0.943**	0.417/0.666	0.107/0.093

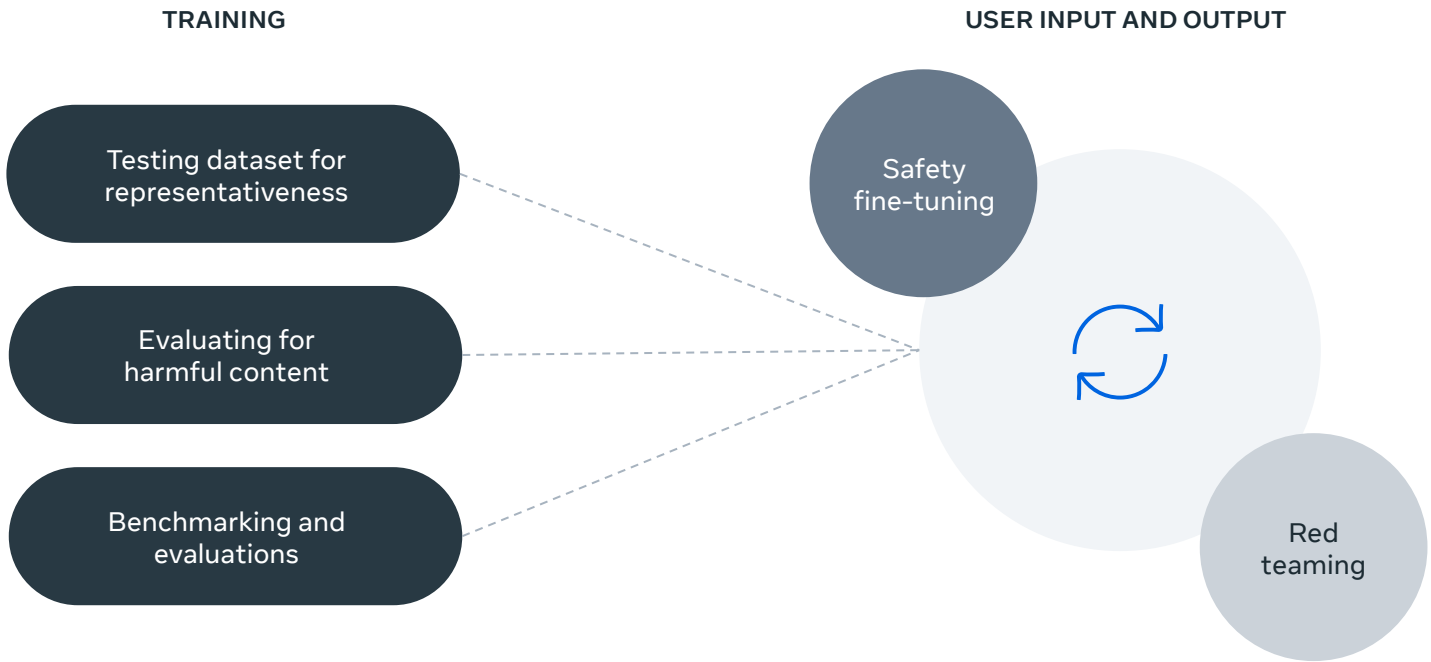
Meta

Responsible AI at Meta

March 2024

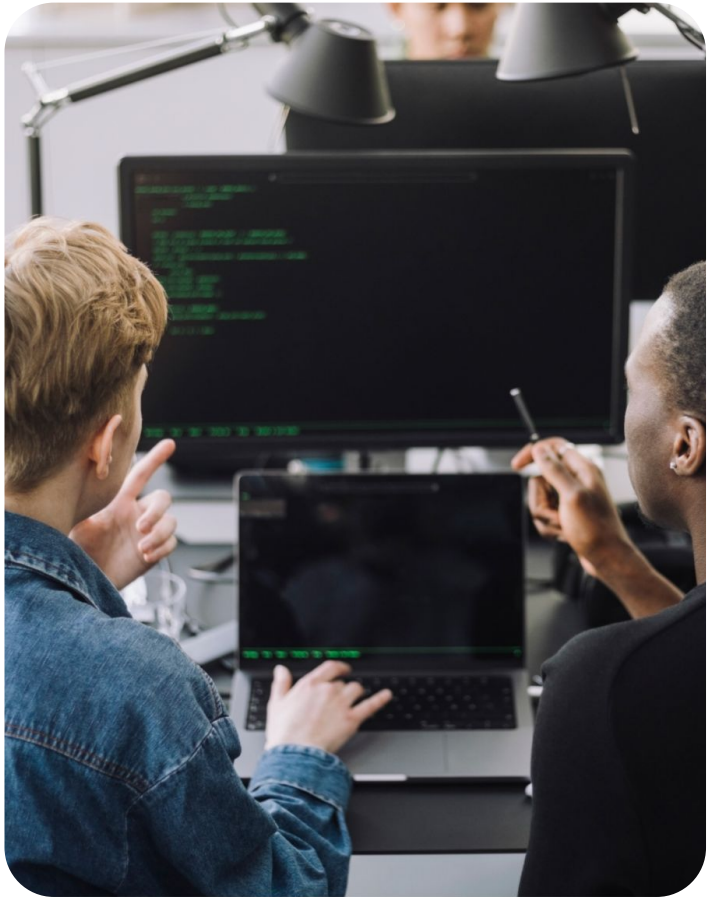
AI at Meta

Developing generative AI responsibly



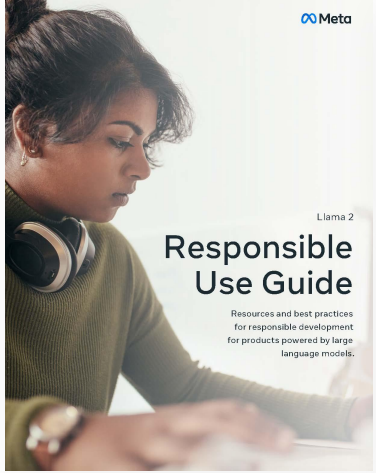
Red teaming

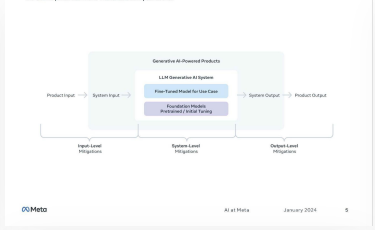
This is a systematic effort to identify model vulnerabilities or emergent risks by crafting prompts that may elicit undesirable behaviours or outputs.

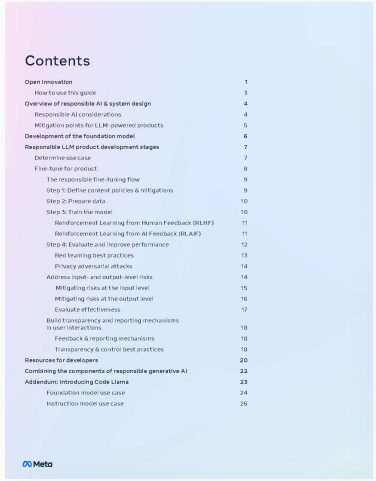


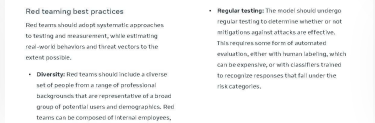
Responsible Use Guide

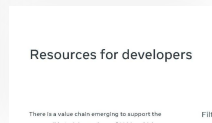
The Responsible Use Guide is a resource for developers that provides best practices and considerations for building products enabled by LLMs in a responsible manner, covering various stages of development from inception to deployment.











The AI Alliance

A community of technology creators, developers, and adopters collaborating to advance safe, responsible AI rooted in open innovation

Founding Members and Collaborators*

- Universities
- Startups & Enterprises
- Science Organizations & Non-profits

Total annual R&D funding represented

>$80B

Students supported by these academic institutions

>400,000

Total staff members

>1,000,000

Source: https://thealliance.ai 2023.

Meta

Resources

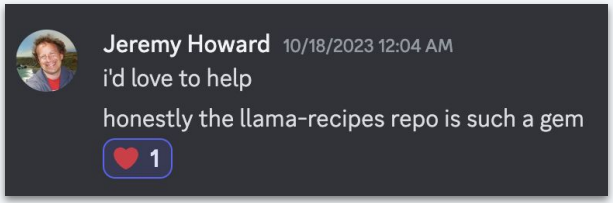
Llama - Getting Started Guide

- Access Llama models
- Fine-Tuning
- Quantization
- Prompt Engineering
- Inferencing
- Validation
- LangChain, LlamaIndex
- Code Llama, Purple Llama
- Community Support & Resources

https://llama.meta.com/get-started/

Llama Recipes

- A collection of fine-tuning recipes for Llama 2
- Support different model sizes 7B to 70B
- Support latest PyTorch features/ techniques for Llama 2
- Recipes from single GPU training to multi-node
- Prompt Engineering, Retrieval Augmented Generation
- Fine-Tuning, PEFT, LoRA, QLoRA
- Getting started guide and code for deployment
- Demo apps



https://bit.ly/llama-recipes

Prompt Engineering with Llama 2
(Deeplearning.ai)

https://bit.ly/llama_dl

Amit Sangani - LinkedIn

Amit Sangani - Github

https://llama.meta.com

Meta

Thank You!