**Decoding Llama 3: An Overview of the Model**

LlaMA (Large Language Model Meta AI) is a Generative AI model, specifically a group of foundational Large Language Models developed by Meta AI, a company owned by Meta(Formerly Facebook).

Generative AI has intensified as more proprietary models have been released within a span of one year for industrial and personal purposes. However, the open-source community did not benefit from this AI revolution as much until META decided to launch their first model, Llama 2, an open-source large language model (LLM) enabling developers to build on, modify, deploy, and use a local copy of the model, or host it on cloud servers. Recently, they launched [Llama 3](https://llama.meta.com/), an upgraded version of the model with the capability of matching benchmark performance of other LLM foundation models.

*You heard it right, it’s OPEN-SOURCE!*

The main unique selling proposition (USP) of META’s model is it favor the open-source environment and the developers around the world will love it. Additionally, META already putting this model to work by integrating it with their popular products like instagram, facebook messenger and whatsapp.

Meta has released two models as part of this next generation release:

* LLaMa 3 8B, an 8-billion parameter model with a knowledge cutoff of March 2023 (meaning that it won’t know anything you ask about events after that deadline)
* LLaMa 3 70B, a 70-billion-parameter model with a knowledge cutoff of December 2023.

**From Llama 2 to Llama 3:**

Before we delve deeper into the specifics of Llama 3, let’s briefly talk about its predecessor, Llama 2, introduced in 2023, marking a key milestone in the open-source LLM landscape, offering a powerful and efficient model that could be run on consumer hardware. However, the model exhibited issues such as refusing to answer certain prompts, limited helpfulness, and room for improvement in certain areas.

To eliminate some of those limitations, META brought forth the Llama 3 version with key changes as mentioned below:

Although Llama 3 8B is considered a small language model (SML) with a size 10 times smaller than Llama 2 70B, it was able to produce similar results to its predecessor. Both models are state-of-the-art in their respective parameter scales.

The context size increased from 4,096 to 8,192 tokens (which can be further increased), allowing the latest model to process bigger prompts at a time.

Llama 3 70B even goes further by showing the best overall performance score, matching that of the most powerful proprietary models around, such as Gemini Pro 1.5 and Claude 3 Sonnet.



Source: META

*Benchmarks (x-shot refers to how many examples the model was given to learn before answering):*

*- MMLU: The Massive Multitask Language Understanding, is a benchmark designed to measure an AI’s ability to understand a wide range of subjects and perform tasks based on that understanding*

*- GPQA: Challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry.*

*- HumanEval: HumanEval is a benchmark for evaluating code synthesis models.*

*- GSM-8K: GSM-8K is a dataset used for evaluating problem-solving skills in AI, particularly in grade-school math problems.*

*- MATH: 12,500 challenging competition mathematics problems*

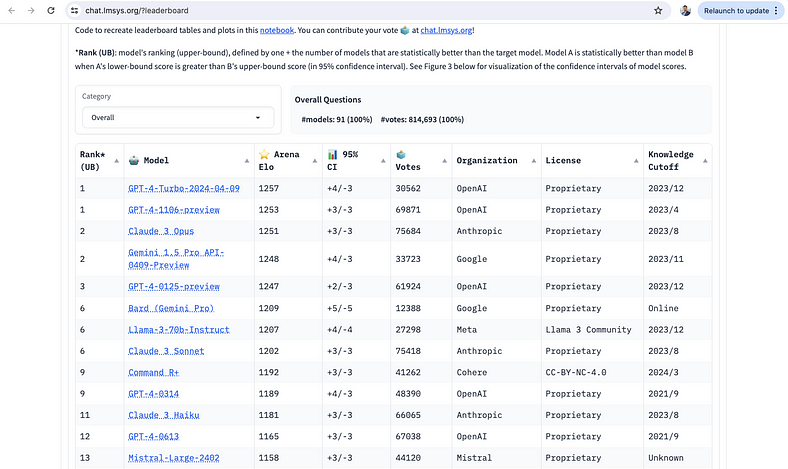
*The results look fascinating, but how did META achieve this performance?*

**Model Architecture and Training:**

Both Llama 3 models were trained on 15T tokens (7 times more compared to Llama 2, including 4 times more code) which features a significantly expanded vocabulary allowing more efficient encoding of text, both for input and output, potentially leading to stronger multilingualism and overall performance improvements.

Training LLM model with more data is undoubtedly the salient factor to improve performance. To build the Llama 3 model, Meta assembled an extensive dataset comprising over 15 trillion tokens sourced from publicly available online platforms, which is seven times larger than the dataset utilized for Llama 2. This dataset contains a notable proportion (exceeding 5%) of high-quality non-English data, spanning more than 30 languages with anticipation of adding forthcoming multilingual capabilities to the applications that will harness the power of Llama models.

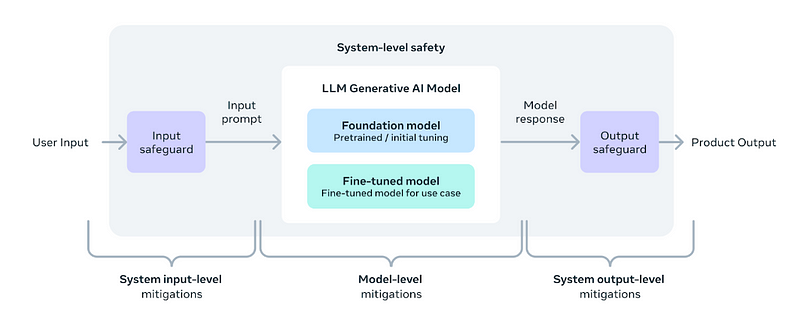
To unlock Llama 3’s full potential for chat and dialogue applications, Meta introduced a new approach aligned with the concept of instruction fine-tuning. This method combines supervised fine-tuning (SFT), rejection sampling, proximal policy optimization (PPO), and direct preference optimization (DPO). META confirmed the data has been curated multiples times to avoid biases before used for SFT and other tuning methods. The latest report shows that Llama 3 model outperforms a few proprietary models as it is jointly ranked 6th in the model arena leaderboard.



source: [https://chat.lmsys.org/?leaderboard](https://chat.lmsys.org/?leaderboard=)

Improving model performance is definitely META’s priority, but not at the cost of model integrity and biases. META adopted multi-level safety development approach, envisioning Llama 3 models as part of a broader ecosystem that puts developers in the driver’s seat, allowing them to design and customize the models for their specific use cases and safety requirements.

* Model-level safety: Model-level safety concerns the data preparation and processing best practices and human feedback or alignment practices for safety at the foundation and fine-tuned model level.
* System-level safety: System-level safety is the venue for the most context-specific safety mitigations dependent on user interactions. Developers looking to craft safety mitigations specifically for their use case with the goal of offering their users the best product experience should explore these options.



Meta has recently released [Llama 3](https://ai.meta.com/blog/meta-llama-3/), the next generation of its state-of-the-art open-source large language model (LLM). Building on the foundations set by its predecessor, Llama 3 aims to enhance the capabilities that positioned Llama 2 as a significant open-source competitor to ChatGPT.

**The Evolution of Llama: From Llama 2 to Llama 3**

Meta's CEO, Mark Zuckerberg, [announced](https://www.threads.net/@zuck/post/C56MFEKxl-x) the debut of Llama 3, the latest AI model developed by Meta AI. This state-of-the-art model, now open-sourced, is set to enhance Meta's various products, including Messenger and Instagram. Zuckerberg highlighted that Llama 3 positions Meta AI as the most advanced [freely available AI assistant](https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/).

Before we talk about the specifics of Llama 3, let's briefly revisit its predecessor, Llama 2. Introduced in 2022, Llama 2 was a significant milestone in the open-source LLM landscape, offering a powerful and efficient model that could be run on consumer hardware.

However, while Llama 2 was a notable achievement, it had its limitations. Users reported issues with false refusals (the model refusing to answer benign prompts), limited helpfulness, and room for improvement in areas like reasoning and code generation.

Enter Llama 3: Meta's response to these challenges and the community's feedback. With Llama 3, Meta has set out to build the best open-source models on par with the top proprietary models available today, while also prioritizing responsible development and deployment practices.

Llama 3: Architecture and Training

One of the key innovations in Llama 3 is its tokenizer, which features a significantly expanded vocabulary of **128,256 tokens** (up from 32,000 in Llama 2). This larger vocabulary allows for more efficient encoding of text, both for input and output, potentially leading to stronger multilingualism and overall performance improvements.

Llama 3 also incorporates **Grouped-Query Attention** (GQA), an efficient representation technique that enhances scalability and helps the model handle longer contexts more effectively. The **8B** version of Llama 3 utilizes GQA, while both the **8B** and **70B** models can process sequences up to **8,192 tokens**.

<https://klu.ai/glossary/grouped-query-attention>

Training Data and Scaling

The training data used for Llama 3 is a crucial factor in its improved performance. Meta curated a massive dataset of over **15 trillion** tokens from publicly available online sources, seven times larger than the dataset used for Llama 2. This dataset also includes a significant portion (over 5%) of high-quality non-English data, covering more than **30 languages**, in preparation for future multilingual applications.

To ensure data quality, Meta employed advanced filtering techniques, including heuristic filters, NSFW filters, semantic deduplication, and text classifiers trained on Llama 2 to predict data quality. The team also conducted extensive experiments to determine the optimal mix of data sources for pretraining, ensuring that Llama 3 performs well across a wide range of use cases, including trivia, STEM, coding, and historical knowledge.

1. \*\*Classifier Filtering:\*\*

- Classifier filtering is a technique used to classify documents into different categories or classes based on certain features or characteristics.

- In the context of NeMo Data Curator, classifier filtering involves training a binary classifier to distinguish between low-quality and high-quality documents.

- The classifier is trained on a dataset consisting of labeled examples of both low-quality and high-quality documents.

- Features extracted from the documents are used as input to the classifier, which learns to differentiate between the two classes.

- Once trained, the classifier can be used to predict the quality of unseen documents, allowing for the filtering of documents based on their predicted quality scores.

- The quality scores assigned by the classifier can be used to select or discard documents for further processing or analysis.

- Classifier filtering is useful for tasks where the quality of documents needs to be assessed automatically, such as in data curation or information retrieval systems.

2. \*\*Heuristic Filtering:\*\*

- Heuristic filtering is a technique that uses predefined rules or heuristics to filter documents based on specific criteria.

- Instead of using machine learning algorithms to learn patterns from data, heuristic filtering relies on human-defined rules or heuristics to determine whether a document meets certain criteria.

- In the context of NeMo Data Curator, heuristic filtering involves applying a cascaded set of rules or filters to determine the quality of documents.

- Each filter in the cascade applies a specific heuristic or rule to assess the quality of documents.

- For example, filters may check for the presence of certain keywords, assess document length, or examine metadata properties.

- Documents that pass all the filters in the cascade are considered high-quality and retained, while those that fail are discarded.

- Heuristic filtering is often used when labeled data for training classifiers is not available or when a simple, rule-based approach is sufficient for the task at hand.

These filtering techniques provide different approaches for assessing the quality of documents and can be used depending on the specific requirements and constraints of the task.

Scaling up pretraining was another critical aspect of Llama 3's development. Meta developed scaling laws that enabled them to predict the performance of its largest models on key tasks, such as code generation, before actually training them. This informed the decisions on data mix and compute allocation, ultimately leading to more efficient and effective training.

Llama 3's largest models were trained on two custom-built 24,000 GPU clusters, leveraging a combination of data parallelization, model parallelization, and pipeline parallelization techniques. Meta's advanced training stack automated error detection, handling, and maintenance, maximizing GPU uptime and increasing training efficiency by approximately three times compared to Llama 2.

**Instruction Fine-tuning and Performance**

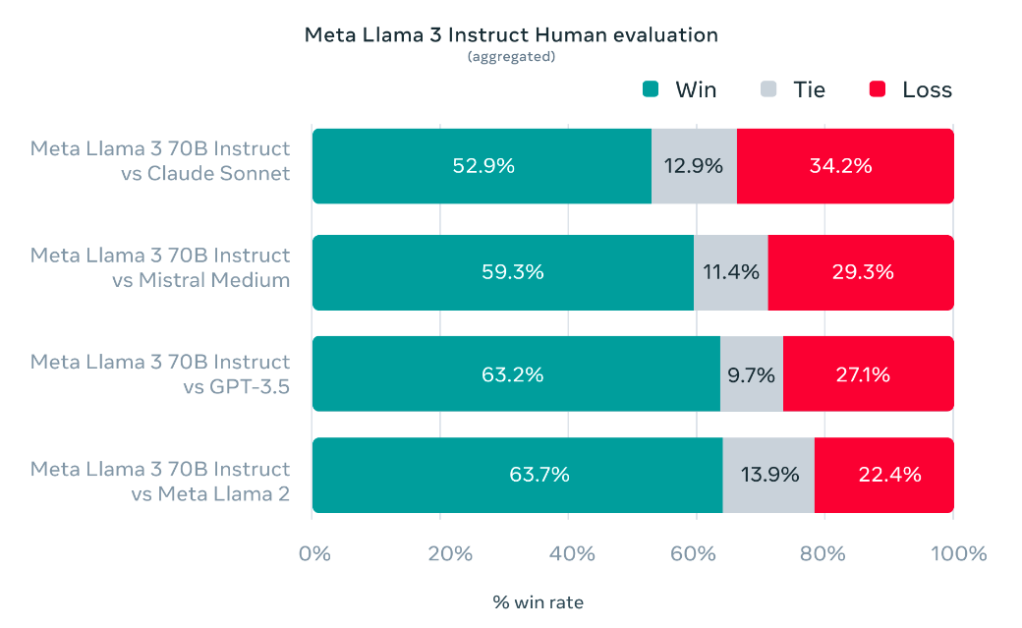
To unlock Llama 3's full potential for chat and dialogue applications, Meta innovated its approach to instruction fine-tuning. Its method combines **supervised fine-tuning** (SFT), rejection sampling, **proximal policy optimization** (PPO), and **direct preference optimization** (DPO).

The quality of the prompts used in SFT and the preference rankings used in PPO and DPO played a crucial role in the performance of the aligned models. Meta's team carefully curated this data and performed multiple rounds of quality assurance on annotations provided by human annotators.

Training on preference rankings via PPO and DPO also significantly improved Llama 3's performance on reasoning and coding tasks. Meta found that even when a model struggles to answer a reasoning question directly, it may still produce the correct reasoning trace. Training on preference rankings enabled the model to learn how to select the correct answer from these traces.

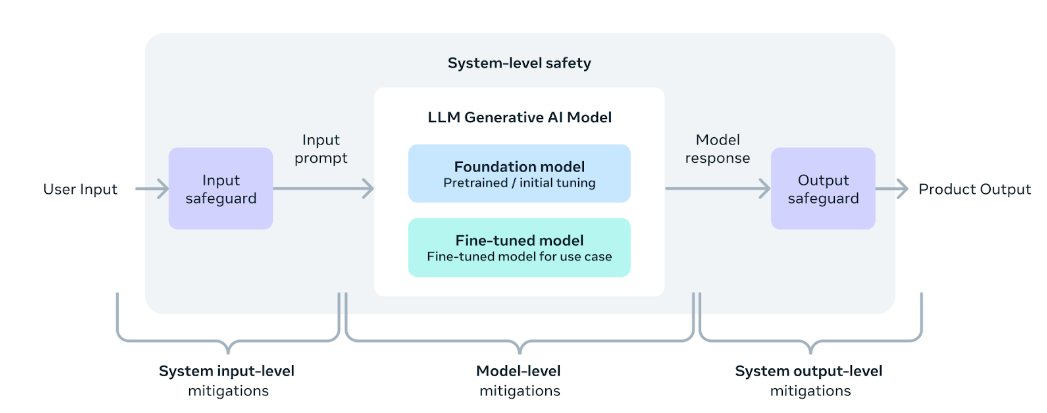
[](https://www.unite.ai/wp-content/uploads/2024/04/unnamed.jpg)

The results speak for themselves: Llama 3 outperforms many available open-source chat models on common industry benchmarks, establishing new state-of-the-art performance for LLMs at the 8B and 70B parameter scales.

[](https://www.unite.ai/wp-content/uploads/2024/04/aggregated-results-of-our-human-evaluations-across-of-these-categories-and-prompts-against-Claude-Sonnet-Mistral-Medium-and-GPt.png)

**Responsible Development and Safety Considerations**

While pursuing cutting-edge performance, Meta also prioritized responsible development and deployment practices for Llama 3. The company adopted a system-level approach, envisioning Llama 3 models as part of a broader ecosystem that puts developers in the driver's seat, allowing them to design and customize the models for their specific use cases and safety requirements.

[](https://www.unite.ai/wp-content/uploads/2024/04/system-level-approach-to-responsibility.png)

Meta conducted extensive red-teaming exercises, performed adversarial evaluations, and implemented safety mitigation techniques to lower residual risks in its instruction-tuned models. However, the company acknowledges that residual risks will likely remain and recommends that developers assess these risks in the context of their specific use cases.

To support responsible deployment, Meta has updated its Responsible Use Guide, providing a comprehensive resource for developers to implement model and system-level safety best practices for their applications. The guide covers topics such as content moderation, risk assessment, and the use of safety tools like Llama Guard 2 and Code Shield.

Llama Guard 2, built on the MLCommons taxonomy, is designed to classify LLM inputs (prompts) and responses, detecting content that may be considered unsafe or harmful. CyberSecEval 2 expands on its predecessor by adding measures to prevent abuse of the model's code interpreter, offensive cybersecurity capabilities, and susceptibility to prompt injection attacks.

Code Shield, a new introduction with Llama 3, adds inference-time filtering of insecure code produced by LLMs, mitigating risks associated with insecure code suggestions, code interpreter abuse, and secure command execution.

**Accessing and Using Llama 3**

Following the launch of Meta AI's Llama 3, several open-source tools have been made available for local deployment on various operating systems, including Mac, Windows, and Linux. This section details three notable tools: Ollama, Open WebUI, and LM Studio, each offering unique features for leveraging Llama 3's capabilities on personal devices.

**Ollama**: Available for Mac, Linux, and Windows, [Ollama](https://ollama.com/download) simplifies the operation of Llama 3 and other large language models on personal computers, even those with less robust hardware. It includes a package manager for easy model management and supports commands across platforms for downloading and running models.

**Open WebUI with Docker**: This tool provides a user-friendly, [Docker](https://docs.docker.com/desktop/)-based interface compatible with Mac, Linux, and Windows. It integrates seamlessly with models from the Ollama registry, allowing users to deploy and interact with models like Llama 3 within a local web interface.

**LM Studio**: Targeting users on Mac, Linux, and Windows, [LM Studio](https://lmstudio.ai/) supports a range of models and is built on the llama.cpp project. It provides a chat interface and facilitates direct interaction with various models, including the Llama 3 8B Instruct model.

These tools ensure that users can efficiently utilize Llama 3 on their personal devices, accommodating a range of technical skills and requirements. Each platform offers step-by-step processes for setup and model interaction, making advanced AI more accessible to developers and enthusiasts.

**Deploying Llama 3 at Scale**

In addition to providing direct access to the model weights, Meta has partnered with various cloud providers, model API services, and hardware platforms to enable seamless deployment of Llama 3 at scale.

One of the key advantages of Llama 3 is its improved token efficiency, thanks to the new tokenizer. Benchmarks show that Llama 3 requires up to **15% fewer tokens** compared to Llama 2, resulting in faster and more cost-effective inference.

The integration of Grouped Query Attention (GQA) in the 8B version of Llama 3 contributes to maintaining inference efficiency on par with the 7B version of Llama 2, despite the increase in parameter count.

To simplify the deployment process, Meta has provided the Llama Recipes repository, which contains open-source code and examples for fine-tuning, deployment, model evaluation, and more. This repository serves as a valuable resource for developers looking to leverage Llama 3's capabilities in their applications.

For those interested in exploring Llama 3's performance, Meta has integrated its latest models into Meta AI, a leading AI assistant built with Llama 3 technology. Users can interact with Meta AI through various Meta apps, such as Facebook, Instagram, WhatsApp, Messenger, and the web, to get things done, learn, create, and connect with the things that matter to them.

**What's Next for Llama 3?**

While the 8B and 70B models mark the beginning of the Llama 3 release, Meta has ambitious plans for the future of this groundbreaking LLM.

In the coming months, we can expect to see new capabilities introduced, including multimodality (the ability to process and generate different data modalities, such as images and videos), multilingualism (supporting multiple languages), and much longer context windows for enhanced performance on tasks that require extensive context.

Additionally, Meta plans to release larger model sizes, including models with over 400 billion parameters, which are currently in training and showing promising trends in terms of performance and capabilities.

To further advance the field, Meta will also publish a detailed research paper on Llama 3, sharing its findings and insights with the broader AI community.

As a sneak preview of what's to come, Meta has shared some early snapshots of its largest LLM model's performance on various benchmarks. While these results are based on an early checkpoint and are subject to change, they provide an exciting glimpse into the future potential of Llama 3.

**Conclusion**

Llama 3 represents a significant milestone in the evolution of open-source large language models, pushing the boundaries of performance, capabilities, and responsible development practices. With its innovative architecture, massive training dataset, and cutting-edge fine-tuning techniques, Llama 3 establishes new state-of-the-art benchmarks for LLMs at the 8B and 70B parameter scales.

However, Llama 3 is more than just a powerful language model; it's a testament to Meta's commitment to fostering an open and responsible AI ecosystem. By providing comprehensive resources, safety tools, and best practices, Meta empowers developers to harness the full potential of Llama 3 while ensuring responsible deployment tailored to their specific use cases and audiences.

As the Llama 3 journey continues, with new capabilities, model sizes, and research findings on the horizon, the AI community eagerly awaits the innovative applications and breakthroughs that will undoubtedly emerge from this groundbreaking LLM.

Whether you're a researcher pushing the boundaries of natural language processing, a developer building the next generation of intelligent applications, or an AI enthusiast curious about the latest advancements, Llama 3 promises to be a powerful tool in your arsenal, opening new doors and unlocking a world of possibilities.