Introducing Meta Llama 3: The most capable openly available LLM to date

Takeaways:

- Today, we're introducing Meta Llama 3, the next generation of our state-of-the-art open source large language model.
- Llama 3 models will soon be available on AWS, Databricks, Google Cloud, Hugging Face, Kaggle, IBM WatsonX, Microsoft Azure, NVIDIA NIM, and Snowflake, and with support from hardware platforms offered by AMD, AWS, Dell, Intel, NVIDIA, and Qualcomm.
- We're dedicated to developing Llama 3 in a responsible way, and we're offering various resources to help others use it responsibly as well. This includes introducing new trust and safety tools with Llama Guard 2, Code Shield, and CyberSec Eval 2.
- In the coming months, we expect to introduce new capabilities, longer context windows, additional model sizes, and enhanced performance, and we'll share the Llama 3 research paper.
- Meta AI, built with Llama 3 technology, is now one of the world's leading AI assistants that can boost your intelligence and lighten your load—helping you learn, get things done, create content, and connect to make the most out of every moment. You can try Meta AI here.

Today, we're excited to share the first two models of the next generation of Llama, Meta Llama 3, available for broad use. This release features pretrained and instruction-fine-tuned language models with 8B and 70B parameters that can support a broad range of use cases. This next generation of Llama demonstrates state-of-the-art performance on a wide range of industry benchmarks and offers new capabilities, including improved reasoning. We believe these are the best open source models of their class, period. In support of our longstanding open approach, we're putting Llama 3 in the hands of the community. We want to kickstart the next wave of innovation in AI across the stack—from applications to developer tools to evals to inference optimizations and more. We can't wait to see what you build and look forward to your feedback.

Our goals for Llama 3

With Llama 3, we set out to build the best open models that are on par with the best proprietary models available today. We wanted to address developer feedback to increase the overall helpfulness of Llama 3 and are doing so while continuing to play a leading role on responsible use

and deployment of LLMs. We are embracing the open source ethos of releasing early and often to enable the community to get access to these models while they are still in development. The text-based models we are releasing today are the first in the Llama 3 collection of models. Our goal in the near future is to make Llama 3 multilingual and multimodal, have longer context, and continue to improve overall performance across core LLM capabilities such as reasoning and coding.

State-of-the-art performance

Our new 8B and 70B parameter Llama 3 models are a major leap over Llama 2 and establish a new state-of-the-art for LLM models at those scales. Thanks to improvements in pretraining and post-training, our pretrained and instruction-fine-tuned models are the best models existing today at the 8B and 70B parameter scale. Improvements in our post-training procedures substantially reduced false refusal rates, improved alignment, and increased diversity in model responses. We also saw greatly improved capabilities like reasoning, code generation, and instruction following making Llama 3 more steerable.

Meta Llama 3 Instruct model performance

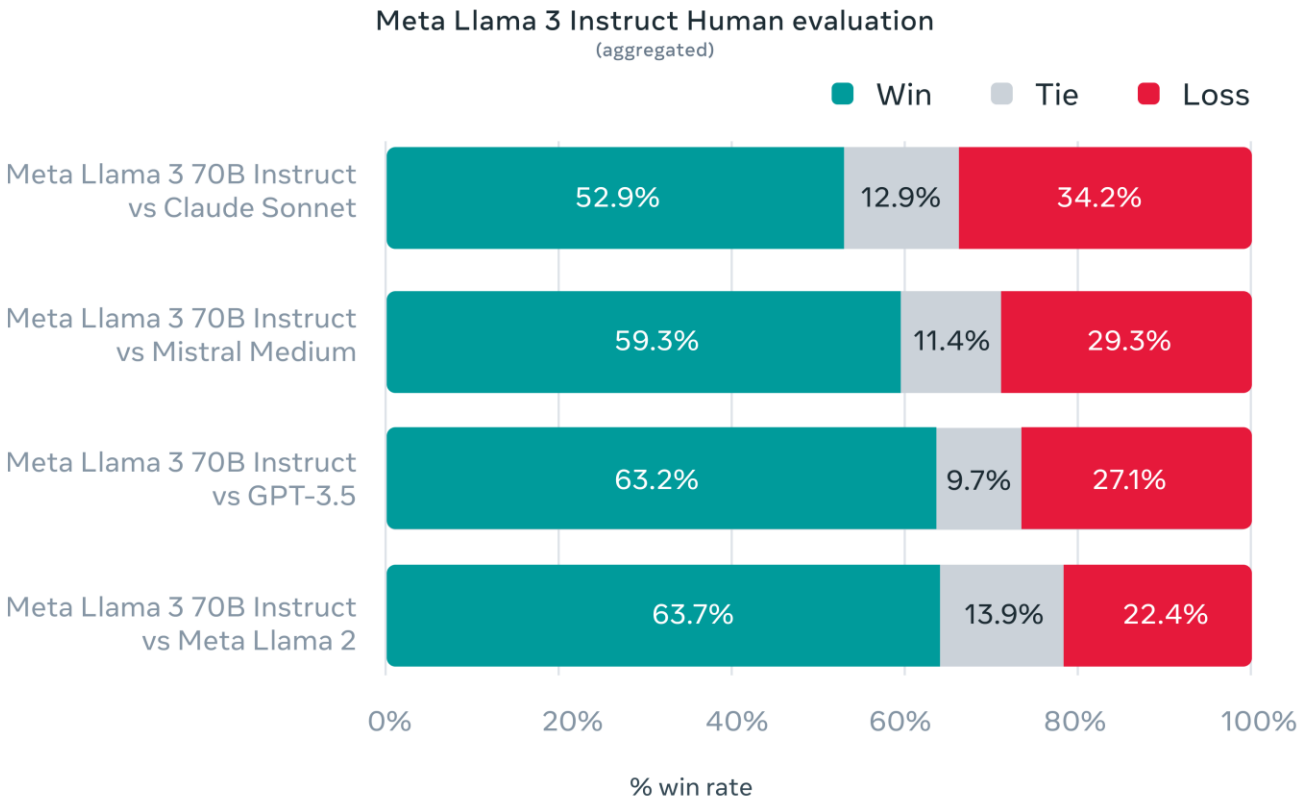
	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

*Please see evaluation details for setting and parameters with which these evaluations are calculated.

In the development of Llama 3, we looked at model performance on standard benchmarks and also sought to optimize for performance for real-world scenarios. To this end, we developed a new high-quality human evaluation set. This evaluation set contains 1,800 prompts that cover 12 key use cases: asking for advice, brainstorming, classification, closed question answering, coding, creative writing,

extraction, inhabiting a character/persona, open question answering, reasoning, rewriting, and summarization. To prevent accidental overfitting of our models on this evaluation set, even our own modeling teams do not have access to it. The chart below shows aggregated results of our human evaluations across of these categories and prompts against Claude Sonnet, Mistral Medium, and GPT-3.5.



Preference rankings by human annotators based on this evaluation set highlight the strong performance of our 70B instruction-following model compared to competing models of comparable size in real-world scenarios.

Our pretrained model also establishes a new state-of-the-art for LLM models at those scales.

Meta Llama 3 Pre-trained model performance

	Meta Llama 3 8B	Mistral 7B		Gemma 7B	
		Published	Measured	Published	Measured
MMLU 5-shot	66.6	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	45.9	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	61.1	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 (0-shot)	79.1
DROP 3-shot, F1	58.4	--	54.4	--	56.3

	Meta Llama 3 70B	Gemini Pro 1.0	Mixtral 8x22B
		Published	Measured
MMLU 5-shot	79.5	71.8	77.7
AGIEval English 3-5-shot	63.0	--	61.2
BIG-Bench Hard 3-shot, CoT	81.3	75.0	79.2
ARC-Challenge 25-shot	93.0	--	90.7
DROP 3-shot, F1	79.7	74.1 variable-shot	77.6

*Please see evaluation details for setting and parameters with which these evaluations are calculated.

To develop a great language model, we believe it's important to innovate, scale, and optimize for simplicity. We adopted this design philosophy throughout the Llama 3 project with a focus on four key ingredients: the model architecture, the pretraining data, scaling up pretraining, and instruction fine-tuning.

Model architecture

In line with our design philosophy, we opted for a relatively standard decoder-only transformer architecture in Llama 3. Compared to Llama 2, we made several key improvements. Llama 3 uses a tokenizer with a vocabulary of 128K tokens that encodes language much more efficiently, which leads to substantially improved model performance. To improve the inference efficiency of Llama 3 models, we've adopted grouped query attention (GQA) across both the 8B and 70B sizes. We trained the models on sequences of 8,192 tokens, using a mask to ensure self-attention does not cross document boundaries.

Training data

To train the best language model, the curation of a large, high-quality training dataset is paramount. In line with our design principles, we invested heavily in pretraining data. Llama 3 is pretrained on over 15T tokens that were all collected from publicly available sources. Our training dataset is seven times larger than that used for Llama 2, and it includes four times more code. To prepare for upcoming

multilingual use cases, over 5% of the Llama 3 pretraining dataset consists of high-quality non-English data that covers over 30 languages. However, we do not expect the same level of performance in these languages as in English.

To ensure Llama 3 is trained on data of the highest quality, we developed a series of data-filtering pipelines. These pipelines include using heuristic filters, NSFW filters, semantic deduplication approaches, and text classifiers to predict data quality. We found that previous generations of Llama are surprisingly good at identifying high-quality data, hence we used Llama 2 to generate the training data for the text-quality classifiers that are powering Llama 3.

We also performed extensive experiments to evaluate the best ways of mixing data from different sources in our final pretraining dataset. These experiments enabled us to select a data mix that ensures that Llama 3 performs well across use cases including trivia questions, STEM, coding, historical knowledge, *etc.*

Scaling up pretraining

To effectively leverage our pretraining data in Llama 3 models, we put substantial effort into scaling up pretraining. Specifically, we have developed a series of detailed scaling laws for downstream benchmark evaluations. These scaling laws enable us to select an optimal data mix and to make informed decisions on how to best use our training compute. Importantly, scaling laws allow us to predict the performance of our largest models on key tasks (for example, code generation as evaluated on the HumanEval benchmark—see above) before we actually train the models. This helps us ensure strong performance of our final models across a variety of use cases and capabilities.

We made several new observations on scaling behavior during the development of Llama 3. For example, while the Chinchilla-optimal amount of training compute for an 8B parameter model corresponds to ~200B tokens, we found that model performance continues to improve even after the model is trained on two orders of magnitude more data. Both our 8B and 70B parameter models continued to improve log-linearly after we trained them on up to 15T tokens. Larger models can match the performance of these smaller models with less training compute, but smaller models are generally preferred because they are much more efficient during inference.

To train our largest Llama 3 models, we combined three types of parallelization: data parallelization, model parallelization, and pipeline parallelization. Our most efficient implementation achieves a compute utilization of over 400 TFLOPS per GPU when trained on 16K GPUs simultaneously. We performed training runs on two custom-built 24K GPU clusters. To maximize GPU uptime, we developed an advanced new training stack that automates error detection, handling, and maintenance. We also greatly improved our hardware reliability and detection mechanisms for silent data corruption, and we developed new scalable storage systems that reduce overheads of checkpointing and rollback. Those improvements resulted in an overall effective training time of more than 95%.

Combined, these improvements increased the efficiency of Llama 3 training by ~three times compared to Llama 2.

Instruction fine-tuning

To fully unlock the potential of our pretrained models in chat use cases, we innovated on our approach to instruction-tuning as well. Our approach to post-training is a combination of supervised fine-tuning (SFT), rejection sampling, proximal policy optimization (PPO), and direct preference optimization (DPO). The quality of the prompts that are used in SFT and the preference rankings that are used in PPO and DPO has an outsized influence on the performance of aligned models. Some of our biggest improvements in model quality came from carefully curating this data and performing multiple rounds of quality assurance on annotations provided by human annotators.

Learning from preference rankings via PPO and DPO also greatly improved the performance of Llama 3 on reasoning and coding tasks. We found that if you ask a model a reasoning question that it struggles to answer, the model will sometimes produce the right reasoning trace: The model knows how to produce the right answer, but it does not know how to select it. Training on preference rankings enables the model to learn how to select it.

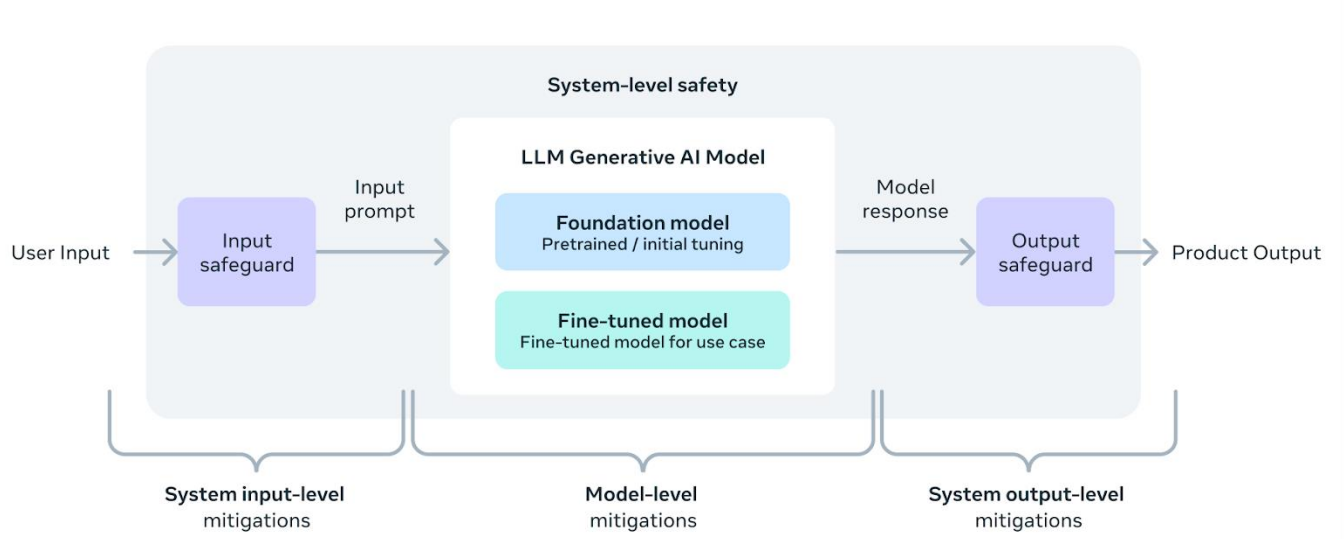
Building with Llama 3

Our vision is to enable developers to customize Llama 3 to support relevant use cases and to make it easier to adopt best practices and improve the open ecosystem. With this release, we're providing new trust and safety tools including updated components with both Llama Guard 2 and Cybersec Eval 2, and the introduction of Code Shield—an inference time guardrail for filtering insecure code produced by LLMs.

We've also co-developed Llama 3 with torchtune, the new PyTorch-native library for easily authoring, fine-tuning, and experimenting with LLMs. torchtune provides memory efficient and hackable training recipes written entirely in PyTorch. The library is integrated with popular platforms such as Hugging Face, Weights & Biases, and EleutherAI and even supports Executorch for enabling efficient inference to be run on a wide variety of mobile and edge devices. For everything from prompt engineering to using Llama 3 with LangChain we have a comprehensive getting started guide and takes you from downloading Llama 3 all the way to deployment at scale within your generative AI application.

A system-level approach to responsibility

We have designed Llama 3 models to be maximally helpful while ensuring an industry leading approach to responsibly deploying them. To achieve this, we have adopted a new, system-level approach to the responsible development and deployment of Llama. We envision Llama models as part of a broader system that puts the developer in the driver's seat. Llama models will serve as a foundational piece of a system that developers design with their unique end goals in mind.



Instruction fine-tuning also plays a major role in ensuring the safety of our models. Our instruction-fine-tuned models have been red-teamed (tested) for safety through internal and external efforts. Our red teaming approach leverages human experts and automation methods to generate adversarial prompts that try to elicit problematic responses. For instance, we apply comprehensive testing to assess risks of misuse related to Chemical, Biological, Cyber Security, and other risk areas. All of these efforts are iterative and used to inform safety fine-tuning of the models being released. You can read more about our efforts in the model card.

Llama Guard models are meant to be a foundation for prompt and response safety and can easily be fine-tuned to create a new taxonomy depending on application needs. As a starting point, the new Llama Guard 2 uses the recently announced MLCommons taxonomy, in an effort to support the emergence of industry standards in this important area. Additionally, CyberSecEval 2 expands on its predecessor by adding measures of an LLM's propensity to allow for abuse of its code interpreter, offensive cybersecurity capabilities, and susceptibility to prompt injection attacks (learn more in our technical paper). Finally, we're introducing Code Shield which adds support for inference-time filtering of insecure code produced by LLMs. This offers mitigation of risks around insecure code suggestions, code interpreter abuse prevention, and secure command execution.

With the speed at which the generative AI space is moving, we believe an open approach is an important way to bring the ecosystem together and mitigate these potential harms. As part of that, we're updating our Responsible Use Guide (RUG) that provides a comprehensive guide to responsible development with LLMs. As we outlined in the RUG, we recommend that all inputs and outputs be checked and filtered in accordance with content guidelines appropriate to the application.

Additionally, many cloud service providers offer content moderation APIs and other tools for responsible deployment, and we encourage developers to also consider using these options.

Deploying Llama 3 at scale

Llama 3 will soon be available on all major platforms including cloud providers, model API providers, and much more. Llama 3 will be everywhere.

Our benchmarks show the tokenizer offers improved token efficiency, yielding up to 15% fewer tokens compared to Llama 2. Also, Group Query Attention (GQA) now has been added to Llama 3 8B as well. As a result, we observed that despite the model having 1B more parameters compared to Llama 2 7B, the improved tokenizer efficiency and GQA contribute to maintaining the inference efficiency on par with Llama 2 7B.

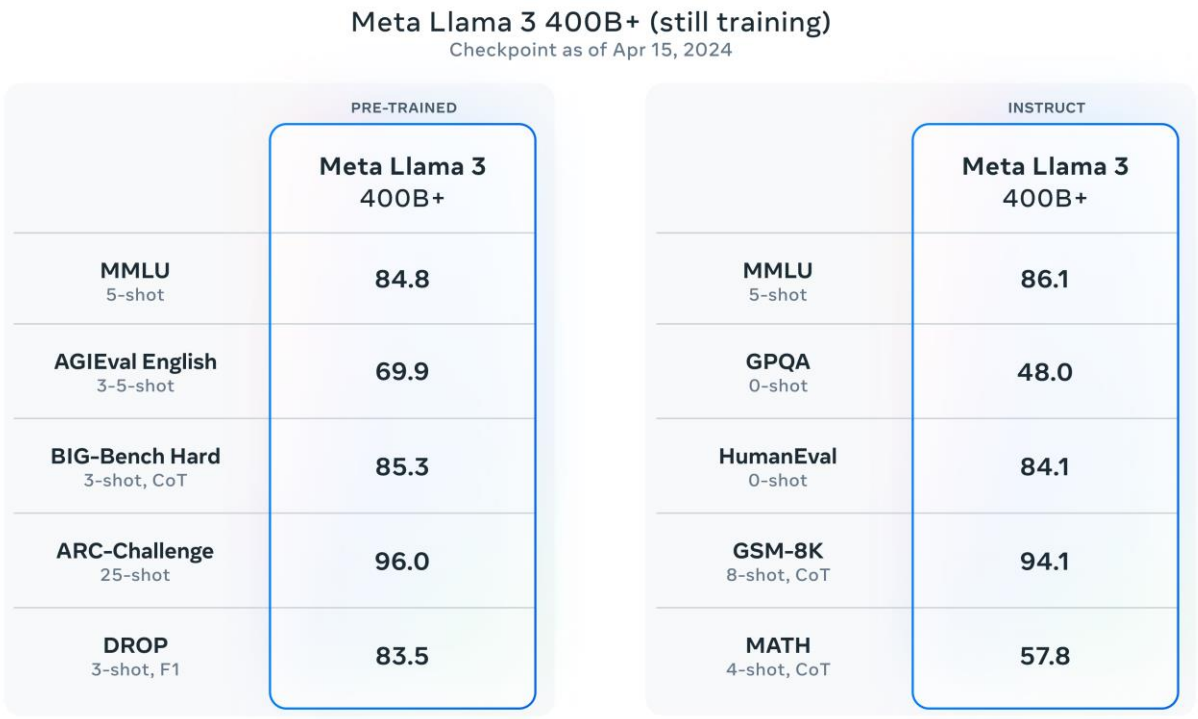
For examples of how to leverage all of these capabilities, check out Llama Recipes which contains all of our open source code that can be leveraged for everything from fine-tuning to deployment to model evaluation.

What's next for Llama 3?

The Llama 3 8B and 70B models mark the beginning of what we plan to release for Llama 3. And there's a lot more to come.

Our largest models are over 400B parameters and, while these models are still training, our team is excited about how they're trending. Over the coming months, we'll release multiple models with new capabilities including multimodality, the ability to converse in multiple languages, a much longer context window, and stronger overall capabilities. We will also publish a detailed research paper once we are done training Llama 3.

To give you a sneak preview for where these models are today as they continue training, we thought we could share some snapshots of how our largest LLM model is trending. Please note that this data is based on an early checkpoint of Llama 3 that is still training and these capabilities are not supported as part of the models released today.



*Please see evaluation details for setting and parameters with which these evaluations are calculated.

We're committed to the continued growth and development of an open AI ecosystem for releasing our models responsibly. We have long believed that openness leads to better, safer products, faster innovation, and a healthier overall market. This is good for Meta, and it is good for society. We're taking a community-first approach with Llama 3, and starting today, these models are available on the leading cloud, hosting, and hardware platforms with many more to come.

Try Meta Llama 3 today

We've integrated our latest models into Meta AI, which we believe is the world's leading AI assistant. It's now built with Llama 3 technology and it's available in more countries across our apps.

You can use Meta AI on Facebook, Instagram, WhatsApp, Messenger, and the web to get things done, learn, create, and connect with the things that matter to you. You can read more about the Meta AI experience here.

Visit the Llama 3 website to download the models and reference the Getting Started Guide for the latest list of all available platforms.

You'll also soon be able to test multimodal Meta AI on our Ray-Ban Meta smart glasses.

As always, we look forward to seeing all the amazing products and experiences you will build with Meta Llama 3.