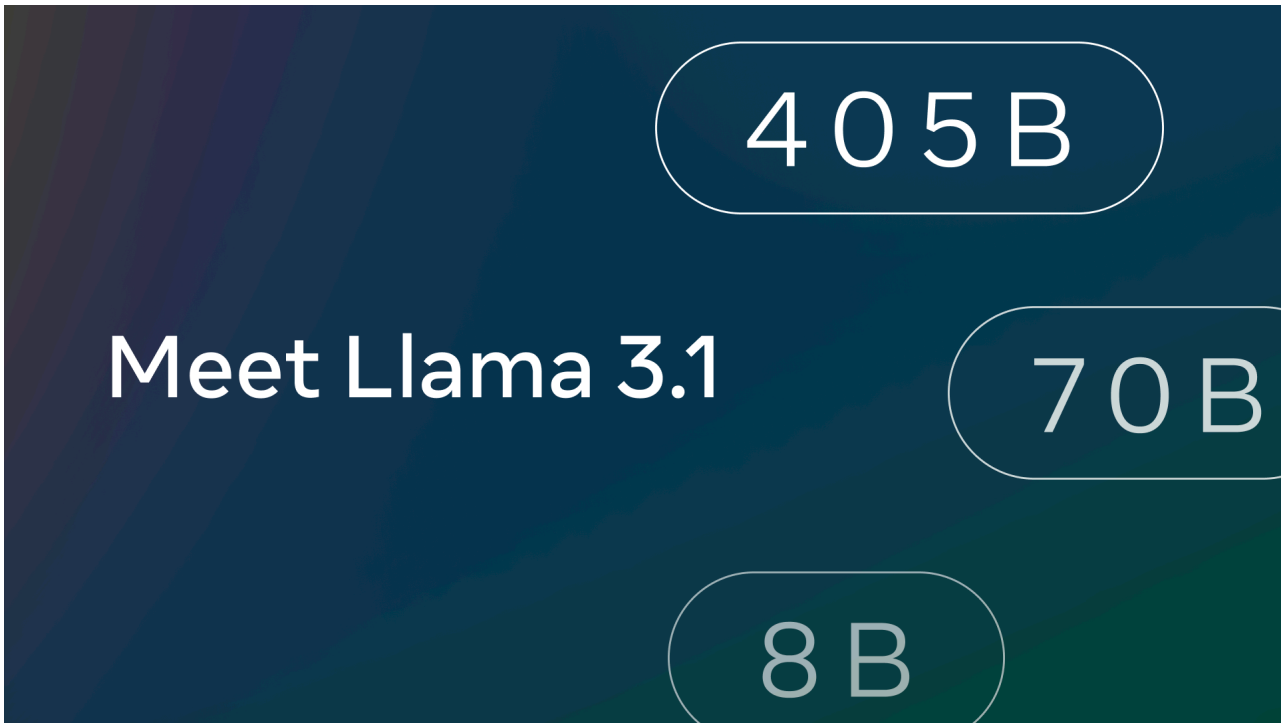
Meta

Large Language Model

Introducing Llama 3.1: Our most capable models to date

July 23, 2024 • 15 minute read



Takeaways:

- Meta is committed to openly accessible AI. Read Mark Zuckerberg's letter detailing why open source is good for developers, good for Meta, and good for the world.
- Bringing open intelligence to all, our latest models expand context length to 128K, add support across eight languages, and include Llama 3.1 405B—the first frontier-level open source AI model.
- Llama 3.1 405B is in a class of its own, with unmatched flexibility, control, and state-of-the-art capabilities that rival the best closed source models. Our new model will enable the community to unlock new workflows, such as synthetic data generation and model distillation.

developers with the tools to create their own custom agents and new types of agentic behaviors. We're bolstering this with new security and safety tools, including Llama Guard 3 and Prompt Guard, to help build responsibly. We're also releasing a request for comment on the Llama Stack API, a standard interface we hope will make it easier for third-party projects to leverage Llama models.

- The ecosystem is primed and ready to go with over 25 partners, including AWS, NVIDIA, Databricks, Groq, Dell, Azure, Google Cloud, and Snowflake offering services on day one.
- Try Llama 3.1 405B in the US on WhatsApp and at meta.ai by asking a challenging math or coding question.

Until today, open source large language models have mostly trailed behind their closed counterparts when it comes to capabilities and performance. Now, we're ushering in a new era with open source leading the way. We're publicly releasing Meta Llama 3.1 405B, which we believe is the world's largest and most capable openly available foundation model. With more than 300 million total downloads of all Llama versions to date, we're just getting started.

RECOMMENDED READS

→ Expanding the Llama ecosystem responsibly

→ The Llama ecosystem: Past, present, and future

Introducing Llama 3.1

Llama 3.1 405B is the first openly available model that rivals the top AI models when it comes to state-of-the-art capabilities in general knowledge, steerability, math, tool use, and multilingual translation. With the release of the 405B model, we're poised to supercharge innovation—with unprecedented opportunities for growth and exploration. We believe the latest generation of Llama will ignite new applications and modeling paradigms, including synthetic data generation to enable the improvement and training of smaller models, as well as model distillation—a capability that has never been achieved at this scale in open source.

state-of-the-art tool use, and overall stronger reasoning capabilities. This enables our latest models to support advanced use cases, such as long-form text summarization, multilingual conversational agents, and coding assistants. We've also made changes to our license, allowing developers to use the outputs from Llama models—including the 405B—to improve other models. True to our commitment to open source, starting today, we're making these models available to the community for download on llama.meta.com and Hugging Face and available for immediate development on our broad ecosystem of partner platforms.

Model evaluations

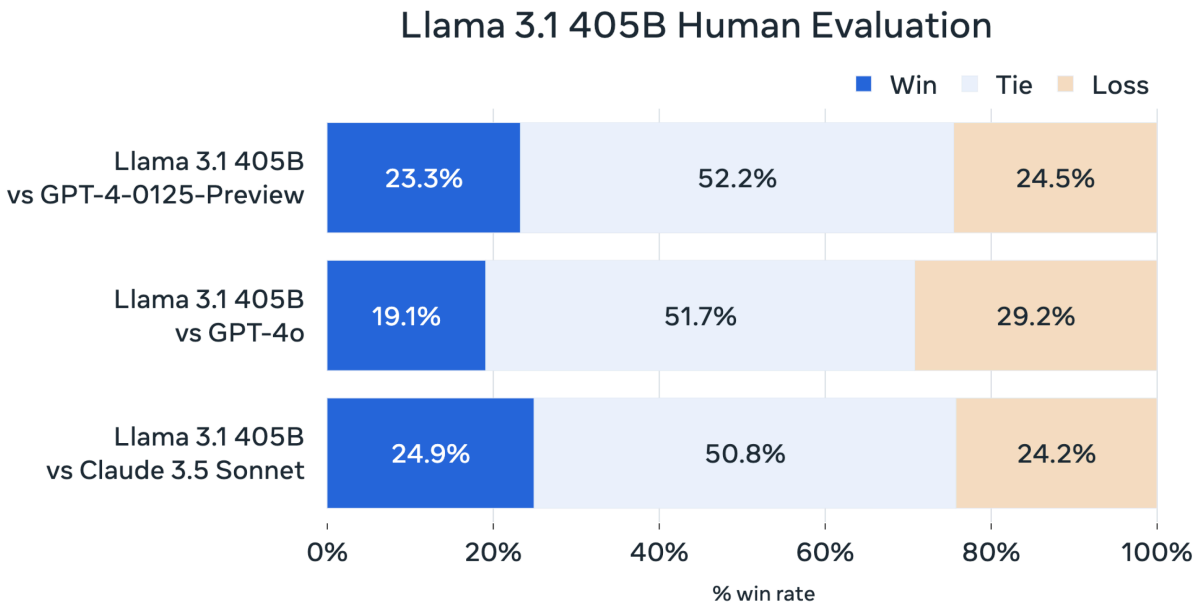
For this release, we evaluated performance on over 150 benchmark datasets that span a wide range of languages. In addition, we performed extensive human evaluations that compare Llama 3.1 with competing models in real-world scenarios. Our experimental evaluation suggests that our flagship model is competitive with leading foundation models across a range of tasks, including GPT-4, GPT-4o, and Claude 3.5 Sonnet. Additionally, our smaller models are competitive with closed and open models that have a similar number of parameters.

Meta

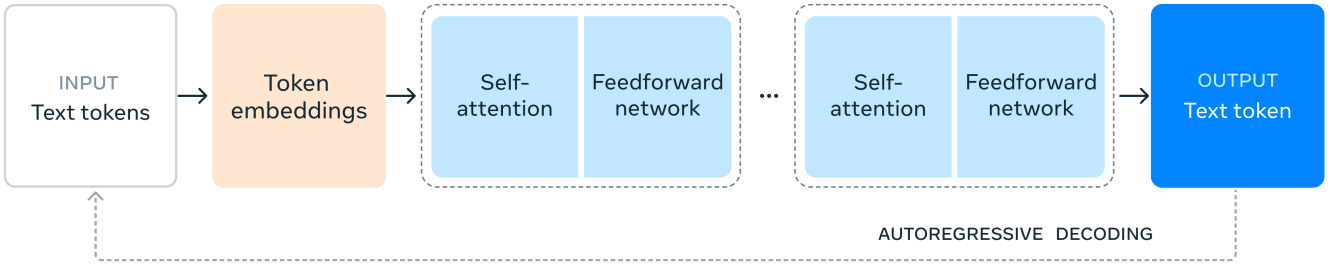
MMLU (0-shot, CoT)		(non-CoT)			
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	**77.0**
IFEval	**88.6**	85.1	84.3	85.6	88.0
Code HumanEval (0-shot)	89.0	73.2	86.6	90.2	**92.0**
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	**90.5**
Math GSM8K (8-shot, CoT)	**96.8**	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	**76.6**	71.1
Reasoning ARC Challenge (0-shot)	**96.9**	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	**59.4**
Tool use BFCL	88.5	86.5	88.3	80.5	**90.2**
Nexus	**58.7**	-	50.3	56.1	45.7
Long context ZeroSCROLLS/QuALITY	**95.2**	-	**95.2**	90.5	90.5
InfiniteBench/En.MC	**83.4**	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	**100.0**	**100.0**	90.8
Multilingual Multilingual MGSM (0-shot)	**91.6**	-	85.9	90.5	**91.6**

Meta

Category / Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning ARC Challenge (0-shot)	83.4	**87.6**	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use BFCL	76.1	-	60.4	84.8	-	**85.9**
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4



As our largest model yet, training Llama 3.1 405B on over 15 trillion tokens was a major challenge. To enable training runs at this scale and achieve the results we have in a reasonable amount of time, we significantly optimized our full training stack and pushed our model training to over 16 thousand H100 GPUs, making the 405B the first Llama model trained at this scale.



To address this, we made design choices that focus on keeping the model development process scalable and straightforward.

- We opted for a standard decoder-only transformer model architecture with minor adaptations rather than a mixture-of-experts model to maximize training stability.
- We adopted an iterative post-training procedure, where each round uses supervised fine-tuning and direct preference optimization. This enabled us to create the highest quality synthetic data for each round and improve each capability's performance.

Compared to previous versions of Llama, we improved both the quantity and quality of the data we use for pre- and post-training. These improvements include the development of more careful pre-processing and curation pipelines for pre-training data, the development of more rigorous quality assurance, and filtering approaches for post-training data.

As expected per scaling laws for language models, our new flagship model outperforms smaller models trained using the same procedure. We also used the 405B parameter model to improve the post-training quality of our smaller models.

To support large-scale production inference for a model at the scale of the 405B, we quantized our models from 16-bit (BF16) to 8-bit (FP8) numerics, effectively lowering the compute requirements needed and allowing the model to run within a single server node.

Instruction and chat fine-tuning

With Llama 3.1 405B, we strove to improve the helpfulness, quality, and detailed instruction-following capability of the model in response to user instructions while

In post-training, we produce final chat models by doing several rounds of alignment on top of the pre-trained model. Each round involves Supervised Fine-Tuning (SFT), Rejection Sampling (RS), and Direct Preference Optimization (DPO). We use synthetic data generation to produce the vast majority of our SFT examples, iterating multiple times to produce higher and higher quality synthetic data across all capabilities. Additionally, we invest in multiple data processing techniques to filter this synthetic data to the highest quality. This enables us to scale the amount of fine-tuning data across capabilities.

We carefully balance the data to produce a model with high quality across all capabilities. For example, we maintain the quality of our model on short-context benchmarks, even when extending to 128K context. Similarly, our model continues to provide maximally helpful answers, even as we add safety mitigations.

The Llama system

Llama models were always intended to work as part of an overall system that can orchestrate several components, including calling external tools. Our vision is to go beyond the foundation models to give developers access to a broader system that gives them the flexibility to design and create custom offerings that align with their vision. This thinking started last year when we first introduced the incorporation of components outside of the core LLM.

As part of our ongoing efforts to develop AI responsibly beyond the model layer and helping others to do the same, we're releasing a full reference system that includes several sample applications and includes new components such as Llama Guard 3, a multilingual safety model and Prompt Guard, a prompt injection filter. These sample applications are open source and can be built on by the community.

The implementation of components in this Llama System vision is still fragmented. That's why we've started working with industry, startups, and the broader community to help better define the interfaces of these components. To support this, we're releasing a request for comment on GitHub for what we're calling "Llama Stack." Llama Stack is a set of standardized and opinionated interfaces for how to build canonical toolchain components (fine-tuning, synthetic data generation) and agentic applications. Our hope is for these to become adopted across the ecosystem, which should help with easier interoperability.

We welcome feedback and ways to improve the proposal. We're excited to grow the ecosystem around Llama and lower barriers for developers and platform providers.

Meta

0:00 / 1:10

Openness drives innovation

Unlike closed models, Llama model weights are available to download. Developers can fully customize the models for their needs and applications, train on new datasets, and conduct additional fine-tuning. This enables the broader developer community and the world to more fully realize the power of generative AI. Developers can fully customize for their applications and run in any environment, including on prem, in the cloud, or even locally on a laptop—all without sharing data with Meta.

And as Mark Zuckerberg noted, open source will ensure that more people around the world have access to the benefits and opportunities of AI, that power isn't concentrated in the hands of a small few, and that the technology can be deployed more evenly and safely across society. That's why we continue to take steps on the path for open access AI to become the industry standard.

We've seen the community build amazing things with past Llama models including an AI study buddy built with Llama and deployed in WhatsApp and Messenger, an LLM tailored to the medical field designed to help guide clinical decision-making, and a healthcare non-profit startup in Brazil that makes it easier for the healthcare system to organize and communicate patients' information about their hospitalization, all in a data secure way. We can't wait to see what they build with our latest models thanks to the power of open source.

Building with Llama 3.1 405B

For the average developer, using a model at the scale of the 405B is challenging. While it's an incredibly powerful model, we recognize that it requires significant compute resources and expertise to work with. We've spoken with the community, and we realize there's so much more to generative AI development than just prompting models. We want to enable everyone to get the most out of the 405B, including:

- Real-time and batch inference
- Supervised fine-tuning
- Evaluation of your model for your specific application
- Continual pre-training
- Retrieval-Augmented Generation (RAG)
- Function calling
- Synthetic data generation

This is where the Llama ecosystem can help. On day one, developers can take advantage of all the advanced capabilities of the 405B model and start building immediately. Developers can also explore advanced workflows like easy-to-use synthetic data generation, follow turnkey directions for model distillation, and enable seamless RAG with solutions from partners, including AWS, NVIDIA, and Databricks. Additionally, Groq has optimized low-latency inference for cloud deployments, with Dell achieving similar optimizations for on-prem systems.

Real-time inference	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Batch inference		✓	✓	✓		✓			✓	✓
Fine tuning		✓	✓	✓					✓	✓
Model evaluation	✓	✓		✓		✓	✓	✓	✓	
Knowledge base	✓	✓	✓	✓		✓	✓	✓	✓	✓
Continual pre-training		✓		✓						
Safety guardrails	✓	✓	✓	✓		✓	✓	✓	✓	✓
Synthetic data generation	✓	✓		✓		✓	✓	✓	✓	✓
Distillation recipe	✓	✓		✓				✓		✓

We've worked with key community projects like vLLM, TensorRT, and PyTorch to build in support from day one to ensure the community is ready for production deployment.

We hope that our release of the 405B will also spur innovation across the broader community to make inference and fine-tuning of models of this scale easier and enable the next wave of research in model distillation.

Try the Llama 3.1 collection of models today

We can't wait to see what the community does with this work. There's so much potential for building helpful new experiences using the multilinguality and increased context length. With the Llama Stack and new safety tools, we look forward to continuing to build together with the open source community responsibly. Before releasing a model, we work to identify, evaluate, and mitigate potential risks through several measures, including pre-deployment risk discovery exercises through red teaming, and safety fine-tuning. For example, we conduct extensive red teaming with both external and internal experts to stress test the models and find unexpected ways they may be used. (Read more about how we're scaling our Llama 3.1 collection of models responsibly in this blog post.)

While this is our biggest model yet, we believe there's still plenty of new ground to explore in the future, including more device-friendly sizes, additional modalities, and more investment at the agent platform layer.As always, we look forward to seeing all the amazing products and experiences the community will build with these models.

Meta

CloudFlare, Databricks, Dell, Deloitte, Fireworks.ai, Google Cloud, Groq, Hugging Face, IBM WatsonX, Infosys, Intel, Kaggle, Microsoft Azure, NVIDIA, OctoAI, Oracle Cloud, PwC, Replicate, Sarvam AI, Scale.AI, SNCF, Snowflake, Together AI, and vLLM project developed in Sky Computing Lab at UC Berkeley.

Get started with Llama 3.1

Read the Llama 3.1 paper

Visit the Llama GitHub repo

Download Llama 3.1 on Hugging Face

Share:

Our latest updates delivered to your inbox

Subscribe to our newsletter to keep up with Meta AI news, events, research breakthroughs, and more.

1.00

Join us in the pursuit of what's possible with AI.

Meta

Related Posts

Open Source

Expanding our open source large language models responsibly

July 23, 2024

Read post

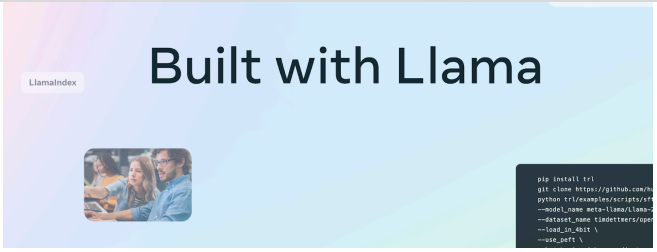
Large Language Model

A social 'study buddy' gets a conversational lift from Meta Llama

June 6, 2024

Read post

Meta



Large Language Model

How SAIF CHECK is using Meta Llama 3 to validate and build trust in AI models

June 20, 2024

Read post

Search AI content

Our approach

Research

Product experiences

Latest news

Foundational models

Meta

Cookies