Llama 3.1 - Did it Kill Closed-Source Models?

By Devon Slonaker, Siddhant Gupta, Tom Cain

Introduction

Meta released Llama 3.1 on July 23, 2024. Llama 3.1 is the newest open-weights model of Meta's Llama family. Various improvements have been made so that it rivals many models in the SLM and LLM categories.

It now also competes with other popular Foundation Models with claims by Meta that Llama 3.1 405B is this generation's highest end open-weights Llama model that is "in a class of its own." [1]

Trained on an impressive array of 16,000 H100 GPUs, Llama 3.1 is the most important open-weights LLM family, as it now competes with even the best closed-source models such as GPT-4o and Claude 3.5 Sonnet, which is why we were surprised that Meta chose a simple ".1 version upgrade" as their name for their latest model.

In Mark Zuckerberg's letter titled "Open Source AI Is the Path Forward,"[2] he emphasizes the importance of open-source AI for the future of technology. Drawing parallels to the evolution of high-performance computing, Zuckerberg asserts that open-source AI models like Llama 3.1 will become the industry standard. Furthermore, he argues that these models empower developers with customization options and data control, while also democratizing access to advanced AI technology globally.

In this paper, we will describe the new features of Llama 3.1, highlight Meta's focus on safe AI, and illustrate how you can play with Llama 3.1 and even install it locally. For developers, we will summarize some published benchmarks and share comments on the Partner Ecosystem that Meta has focused on. We will also explore how Meta's investment in open-weight models aligns with Zuckerberg's vision of fostering innovation, enhancing transparency, and ensuring equitable distribution of AI advancements. Time will tell whether this approach was pure genius and will put closed-source foundation model companies on their heels!

The Llama Models

In the past Meta has released two models with each new release of Llama. This time Meta released three models intended for various workloads: an 8B model, 70B model, and their new flagship LLM, the 405B model.

MODEL	USAGE	COMPETITOR MODELS	NOTES
8B	Fast, near immediate response	Gemma 2 9B IT	Can run locally really easily.
70B	Chatbots or STEM related tasks	GPT-3.5 Turbo	
405B	High-level reasoning, sensitive questions, precise and exact answers	GPT-4o, Claude 3.5 Sonnet	Includes the usage of Llama Guard 3 for safer prompts and outputs.

Llama 3.1 8B was released for both local system installation as well as fast, near immediate responses for simple and easy questions.

Llama 3.1 70B is intended for a wider range of tasks such as chatbots or more complex STEM-related tasks.

Llama 3.1 405B is a very large language model and as with other foundation models it's used for high level reasoning and sensitive questions that require precise and exact answers.

New Features

First, Llama 3.1 increased their training knowledge cutoffs to December 2023 (8B used to be March 2023). They also increased context length, added multilingual support, and, consistent with their stated commitment to safe AI, they added Llama Guard and Prompt Guard tools.

Multilingual Functionality

Llama 3.1 breaks the multi-lingual wall and introduces support for various languages. These languages include the addition of Portuguese, Spanish, Italian, German, French, Hindi, and Thai. This allows more people from all over the world to experience and garner use from large language models as they can now communicate with these models in their native tongue.

Increased Context Length

Llama 3.1 has seen a massive increase in context length over its predecessor. While Llama 3 was able to handle 8k token context lengths, a 2x increase over Llama 2 at 4k token lengths, Llama 3.1 sees a whopping 16x increase in token lengths over Llama 3, allowing for massive 128k token contexts.

Supporting Safe AI with Llama Guard 3 and Prompt Guard

Meta introduced two tools to support AI safety: Llama Guard 3 and Prompt Guard. While these tools are provided along with Llama 3.1, it is noteworthy that each tool has been open-weighted for use with other models.

Llama Guard 3 is a tool intended to cleanse both queries and responses to promote a safe environment. This prevents both users and other models from violating content moderation rules through their prompts, and also cleanses the Llama 3.1 model responses to prevent it from saying anything unsafe. Meta has a list of so-called "hazard" categories for identifying unsafe content. Meta specifically took the time to address and mitigate areas of concern such as cybersecurity, child safety, self-harm, and privacy. The full table of the "hazards" can be found in Source 3.[4]

Prompt Guard is a Meta tool that prevents prompt injection and jailbreaking techniques by users. Prompt injection prevents bad actors from being able to trick the model into executing instructions it should not be able to by adding on additional information to a context window that could cause the model to respond with something unintentional or malicious. Jailbreak prevention prevents people from attempting to escape the guardrails put around the model, such as by asking it to role play as a character that is not bound by the rules imposed on the model to make the model say malicious or dangerous things.[5]

Developer Partners

There are numerous companies that partner with Meta in order to provide access to Llama 3.1. While each provider hosts Llama 3.1 for use by developers, they all provide different usage options and features for Llama 3.1. These options are mostly directed toward developers, and the chart found in Source 1 details which features developers can expect out of Llama 3.1 from these partners.

For instance, we highlight a few of the key features offered by partners:

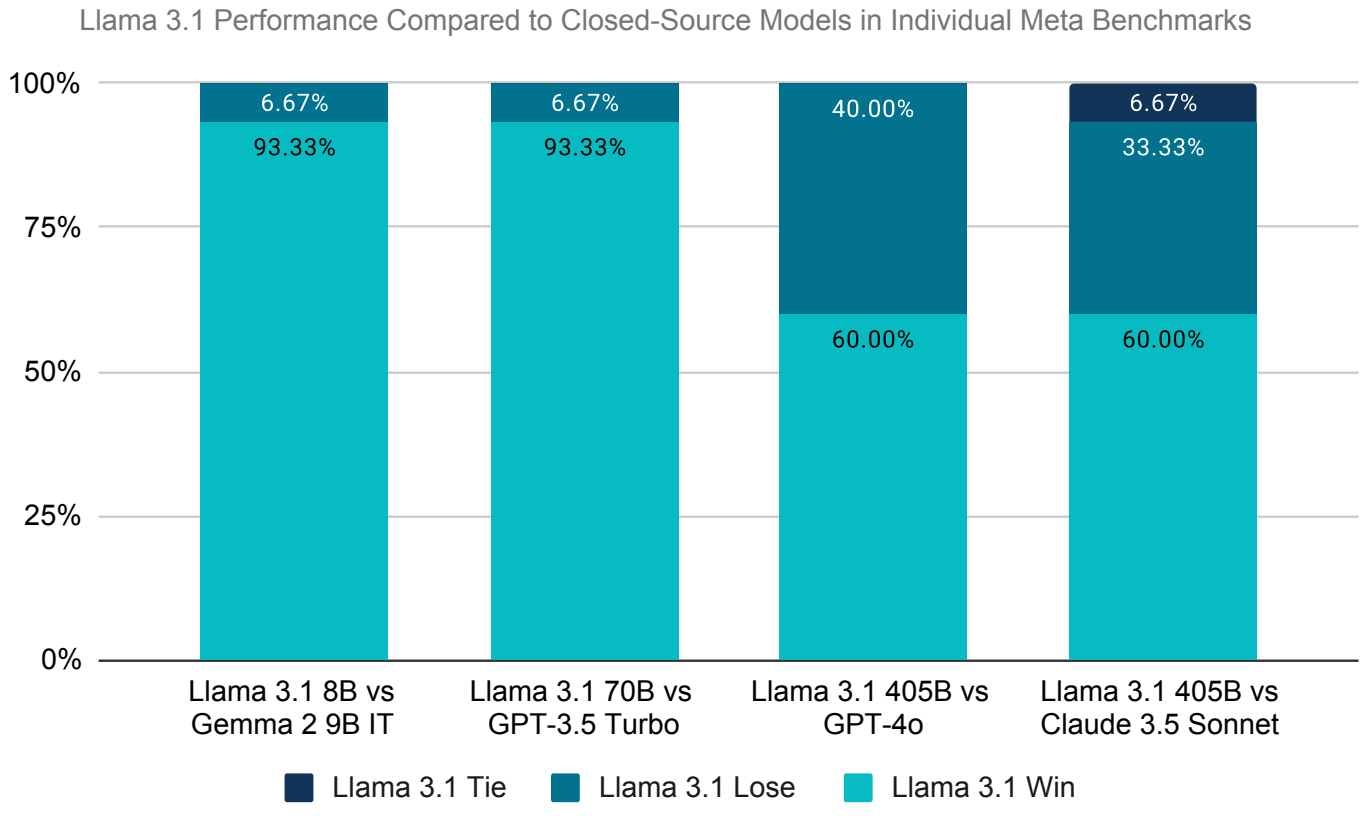
- Real-time inference, which allows for faster model responses.
- Fine-tuning the model with customized training data for specific tasks or domains.
- Retrieval-Augmented Generation (RAG) for enhancing responses with external knowledge.
- Pricing varies by partner platform, and developers should be aware of the pricing when using their current development platform.

Some Benchmarks

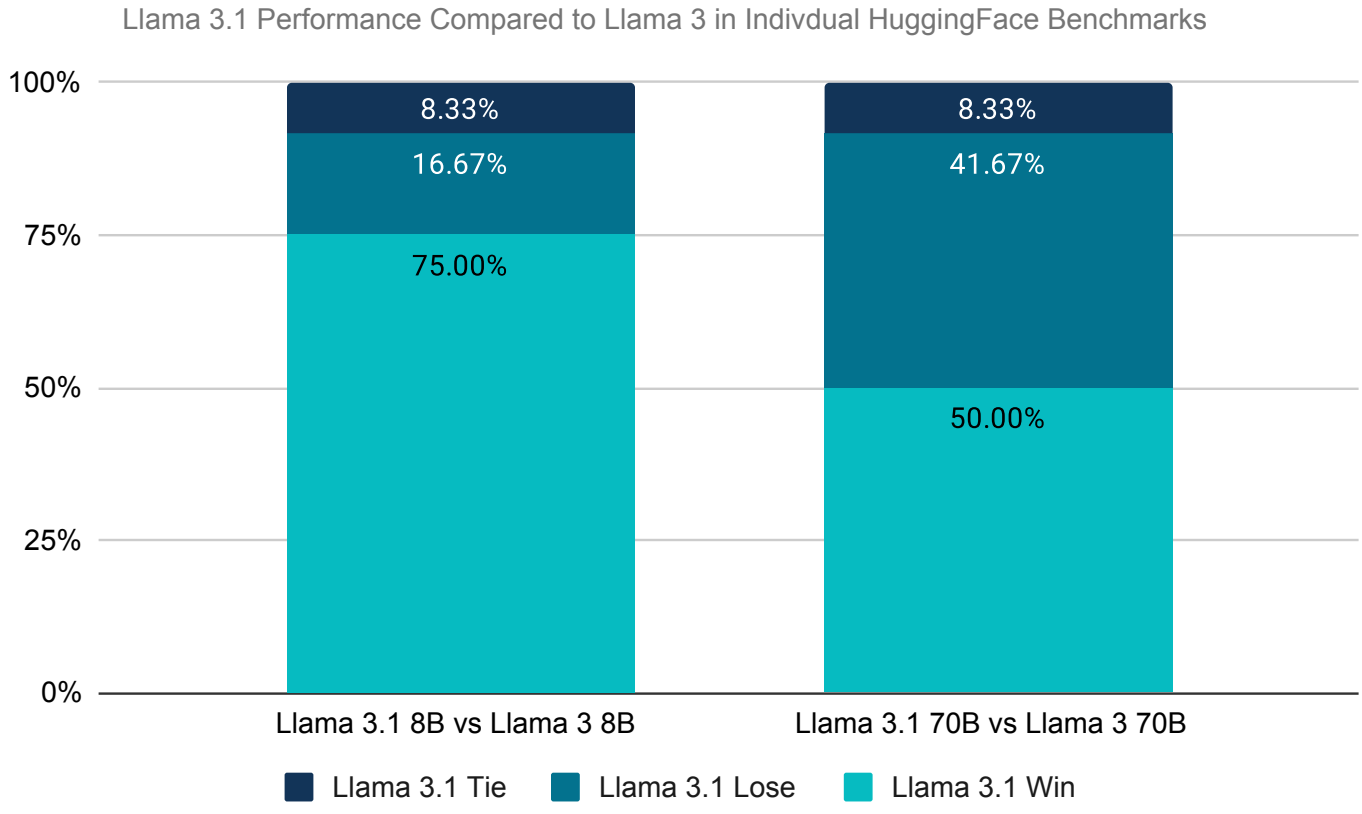
Benchmark results allow us to see the improvements of one model when compared to other models. Benchmarks often measure capabilities such as general scenarios, code evaluation and generation, tool usage such as API usage, reasoning, and mathematics. A benchmark table allows us to see whether a model gained marginal or significant improvements in certain areas when compared to other models.

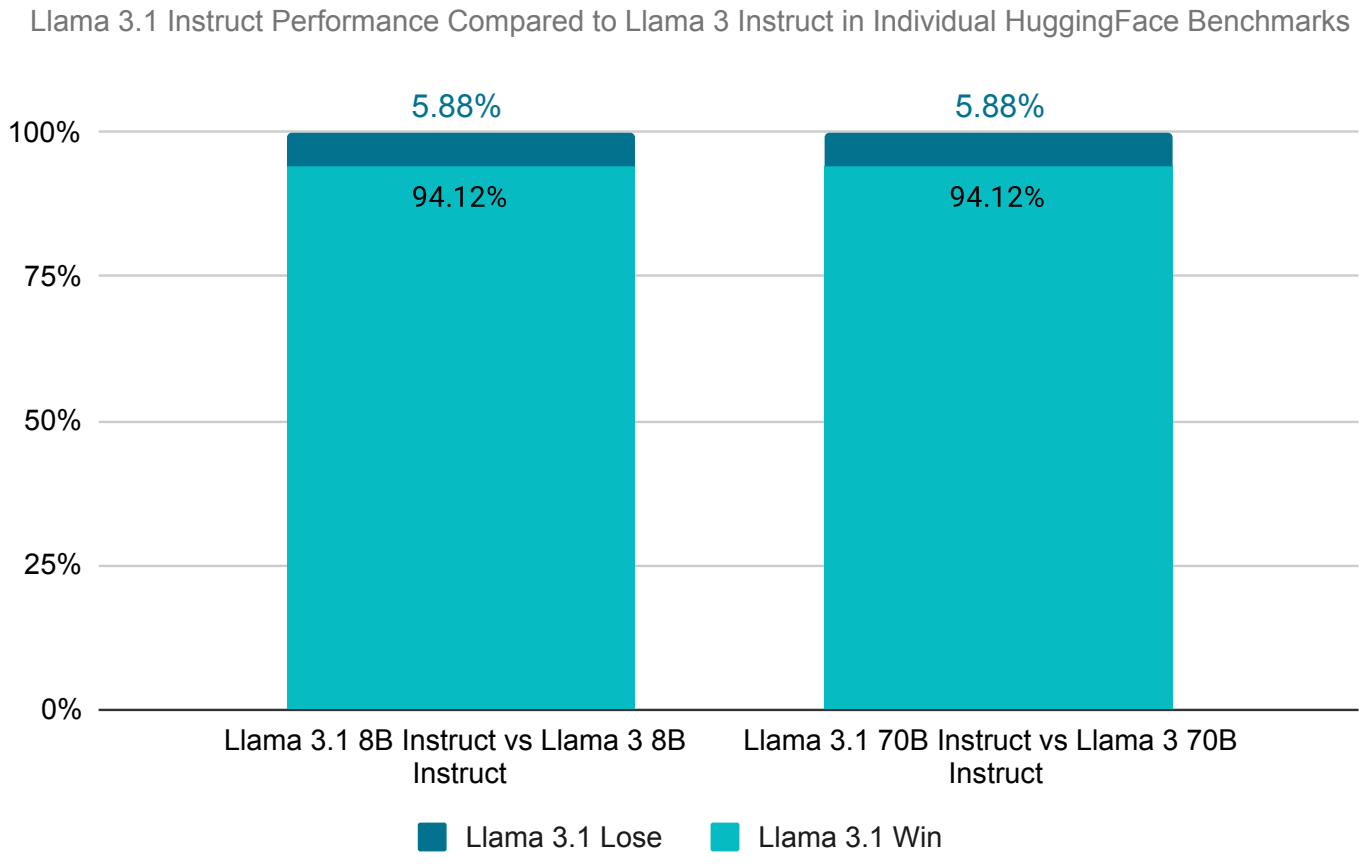
Rather than re-publishing benchmark tables from either the Meta website or their Hugging Face page, we choose to examine each benchmark in every category where Llama 3.1 either won (had a higher benchmark number than the competitor) or lost (had a lower benchmark number than the competitor).

Llama 3.1 Models vs Close-Sourced Models [6]



Llama 3.1 vs Llama 3 Benchmarks [7]





The C4AI Analysis of Llama 3.1 Models

Our analysis has a two-part focus:

1. A comparison of the newly released Llama 3.1 base models against competing models in a similar class using Meta's benchmarks found on their main website:
   - Llama 3.1 8B vs Gemma 2 9B IT
   - Llama 3.1 70B vs GPT-3.5 Turbo
   - Llama 3.1 405B vs GPT-4o and Claude 3.5 Sonnet
2. A comparison of the upgraded Llama 3.1 models, both base and Instruct models, to the previous Llama 3 versions using benchmarks published on Meta's Hugging Face page.

Comparing Against Competing Models

We show that the Llama 3.1 smaller, base models pull ahead of models in a similar class. And, importantly, even the flagship Llama 3.1 405B model performs well against GPT-4o and Claude 3.5 Sonnet.

Llama 3.1 8B is superior to Gemma 2 9B IT in nearly all benchmark categories with the majority of benchmark wins in code generation, math, and tool use. Llama 3.1 8B also is able to complete other tests that Gemma 2 9B IT could not, including tests in reasoning, tool usage, and long context evaluation. Llama 3.1 8B only fell short in the Abstraction and Reasoning Corpus (ARC) Challenge (0-shot) benchmark which measures how well the model performs on skill-acquisition of an unknown task.

Llama 3.1 70B shows significant improvements over GPT 3.5 Turbo in nearly all categories with the exception of the Berkeley Function Calling Leaderboard (BFCL) tool usage benchmark which measures the capability to call functions.

Llama 3.1 405B performs better than GPT-4o and Claude 3.5 Sonnet in a majority of the benchmark results, we conclude that Meta's Llama 3.1 open-weight family of models competes favorably against even the best closed-source models.

This is a paradigm shift in the realm of large language models.

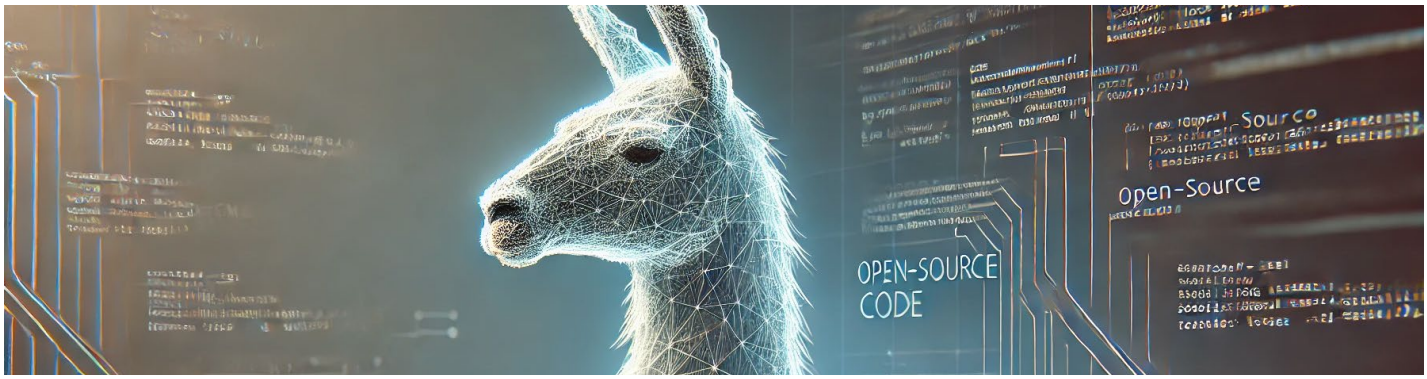
Because Meta has made their very capable models free of charge, it challenges closed-source vendors to make significant improvements to their own models.

Comparing Against Llama 3

Meta released new versions of both their base models (Llama 3.1 8B and 70B) and their Instruct models (Llama 3.1 8B Instruct and 70B Instruct). Meta's base models are trained so that they generate text completions, code completions, image data, etc.

In contrast, the Instruct models are fine-tuned to be able to follow a sequence of specific user instructions, and produce more focused, task-oriented outputs.

The Hugging Face benchmarks show that the Llama 3.1 8B base model was a clear upgrade over Llama 3 8B, whereas Llama 3.1 70B shared a similar level of performance and reliability as the prior Llama 3 70B version.



The Llama 3.1 Instruct benchmarks paint a different, and much better picture of the new Instruct versions released by Meta. The new versions show improvement in nearly all benchmark evaluations! The only shortcomings: Llama 3.1 8B Instruct only narrowly lost to Llama 3 8B Instruct in the Graduate-Level Google-Proof Q&A Benchmark (GPQA) which is a benchmark that measures graduate-level reasoning in a multiple-choice style test with questions focusing in biology, chemistry, and physics; and Llama 3.1 70B Instruct only marginally fell short in the HumanEval benchmark (measures a model's ability to generate code) versus the older Llama 3 70B Instruct.

While there is no denying that the 3.1 base models are an overall improvement over their version 3 counterparts, **we think that the feature additions were more impactful than the performance gains.**

We showed that Llama 3.1 not only expands the context length of its models to 128K tokens and adds support across eight languages, but offers the groundbreaking 405B model—the first frontier-level open-weight AI model.

Final Comments on Version 3.1

Meta is committed to making AI openly accessible, an approach that benefits developers and the global community.

We showed that Llama 3.1 not only expands the context length of its models to 128K tokens and adds support across eight languages, but offers the groundbreaking 405B model—the first frontier-level open-weight AI model.

This model offers unmatched flexibility, control, and capabilities that rival the best closed-source models, enabling the community to explore new workflows such as synthetic data generation and model distillation.

In continuing to build out the Llama system, Meta provides more components to work with the model, including a reference system to empower developers to create custom agents and new agentic behaviors. New security and safety tools, such as Llama Guard 3 and Prompt Guard, support this initiative. It is further bolstered by partnerships with over 25 industry leaders, including AWS, NVIDIA, Databricks, Groq, and Google Cloud. On the day of the release, services became available across these platforms, demonstrating the readiness of the ecosystem to support developers in leveraging Llama 3.1's capabilities. The model is widely available for testing multiple platforms including WhatsApp, Instagram, Groq, and Meta.ai.[1] For a deeper understanding of Llama 3.1, refer to the full Llama 3.1 research paper by Meta.[8]

Endnotes

1. https://ai.meta.com/blog/meta-llama-3-1/
2. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/
3. https://huggingface.co/meta-llama/Llama-Guard-3-8B
4. https://ai.meta.com/blog/meta-llama-3-1-ai-responsibility/
5. https://llama.meta.com/docs/model-cards-and-prompt-formats/prompt-guard
6. https://llama.meta.com/
7. https://huggingface.co/meta-llama/Meta-Llama-3.1-8B
8. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/