Source: https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

Model Information

The Meta Llama 3.1 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction tuned generative models in 8B, 70B and 405B sizes (text in/text out). The Llama 3.1 instruction tuned text only models (8B, 70B, 405B) are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

Model developer: Meta

Model Architecture: Llama 3.1 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowledge cutoff
Llama 3.1 (text only)	A new mix of publicly available online data.	8B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	December 2023
		70B	Multilingual Text	Multilingual Text and code	128k	Yes		
		405B	Multilingual Text	Multilingual Text and code	128k	Yes		

Supported languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.

Llama 3.1 family of models. Token counts refer to pretraining data only. All model versions use Grouped-Query Attention (GQA) for improved inference scalability.

Model Release Date: July 23, 2024.

Status: This is a static model trained on an offline dataset. Future versions of the tuned models will be released as we improve model safety with community feedback.

License: A custom commercial license, the Llama 3.1 Community License, is available at: https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/LICENSE

Feedback: Instructions on how to provide feedback or comments on the model can be found in the Llama Models README. For more technical information about generation parameters and recipes for how to use Llama 3.1 in applications, please go here.

Intended Use

Intended Use Cases Llama 3.1 is intended for commercial and research use in multiple languages. Instruction tuned text only models are intended for assistant-like chat, whereas pretrained models can be adapted for a variety of natural language generation tasks. The Llama 3.1 model collection also supports the ability to leverage the outputs of its models to improve other models including synthetic data generation and distillation. The Llama 3.1 Community License allows for these use cases.

Out-of-scope Use in any manner that violates applicable laws or regulations (including trade compliance laws). Use in any other way that is prohibited by the Acceptable Use Policy and Llama 3.1 Community License. Use in languages beyond those explicitly referenced as supported in this model card.

Note: Llama 3.1 has been trained on a broader collection of languages than the 8 supported languages. Developers may fine-tune Llama 3.1 models for languages beyond the 8 supported languages provided they comply with the Llama 3.1 Community License and the Acceptable Use Policy and in such cases are responsible for ensuring that any uses of Llama 3.1 in additional languages is done in a safe and responsible manner.

Hardware and Software

Training Factors We used custom training libraries, Meta's custom built GPU cluster, and production infrastructure for pretraining. Fine-tuning, annotation, and evaluation were also performed on production infrastructure.

Training Energy Use Training utilized a cumulative of **39.3**M GPU hours of computation on H100-80GB (TDP of 700W) type hardware, per the table below. Training time is the total GPU time required for training each model and power consumption is the peak power capacity per GPU device used, adjusted for power usage efficiency.

Training Greenhouse Gas Emissions Estimated total location-based greenhouse gas emissions were **11,390** tons CO2eq for training. Since 2020, Meta has maintained net zero greenhouse gas emissions in its global operations and matched 100% of its electricity use with renewable energy, therefore the total market-based greenhouse gas emissions for training were 0 tons CO2eq.

	Training Time (GPU hours)	**Training Power Consumption (W)**	**Training Location-Based Greenhouse Gas Emissions (tons CO2eq)**	**Training Market-Based Greenhouse Gas Emissions (tons CO2eq)**
Llama 3.1 8B	1.46M	700	420	0
Llama 3.1 70B	7.0M	700	2,040	0
Llama 3.1 405B	30.84M	700	8,930	0
Total	39.3M		11,390	0

The methodology used to determine training energy use and greenhouse gas emissions can be found here. Since Meta is openly releasing these models, the training energy use and greenhouse gas emissions will not be incurred by others.

Training Data

Overview: Llama 3.1 was pretrained on ~15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 25M synthetically generated examples.

Data Freshness: The pretraining data has a cutoff of December 2023.

Benchmark scores

In this section, we report the results for Llama 3.1 models on standard automatic benchmarks. For all the evaluations, we use our internal evaluations library. Details of our evals can be found here. We are also releasing the raw data generated as part of our evals which can be found here in the dataset sections.

Base pretrained models

Category	**Benchmark**	**# Shots**	**Metric**	**Llama 3 8B**	**Llama 3.1 8B**	**Llama 3 70B**	**Llama 3.1 70B**	**Llama 3.1 405B**
General	MMLU	5	macro_avg/acc_char	66.7	66.7	79.5	79.3	85.2
	MMLU-Pro (CoT)	5	macro_avg/acc_char	36.2	37.1	55.0	53.8	61.6
	AGIEval English	3-5	average/acc_char	47.1	47.8	63.0	64.6	71.6
	CommonSenseQA	7	acc_char	72.6	75.0	83.8	84.1	85.8
	Winogrande	5	acc_char	•	60.5	•	83.3	86.7
	BIG-Bench Hard (CoT)	3	average/em	61.1	64.2	81.3	81.6	85.9
	ARC-Challenge	25	acc_char	79.4	79.7	93.1	92.9	96.1
Knowledge reasoning	TriviaQA-Wiki	5	em	78.5	77.6	89.7	89.8	91.8
Reading comprehension	SQuAD	1	em	76.4	77.0	85.6	81.8	89.3
	QuAC (F1)	1	f1	44.4	44.9	51.1	51.1	53.6

	BoolQ	0	acc_char	75.7	75.0	79.0	79.4	80.0
	DROP (F1)	3	f1	58.4	59.5	79.7	79.6	84.8

Instruction tuned models

Category	Benchmark	# Shots	Metric	Llama 3 8B Instruct	Llama 3.1 8B Instruct	Llama 3 70B Instruct	Llama 3.1 70B Instruct	Llama 3.1 405B Instruct
General	MMLU	5	macro_avg/acc	68.5	69.4	82.0	83.6	87.3
	MMLU (CoT)	0	macro_avg/acc	65.3	73.0	80.9	86.0	88.6
	MMLU-Pro (CoT)	5	macro_avg/acc	45.5	48.3	63.4	66.4	73.3
	IFEval			76.8	80.4	82.9	87.5	88.6
Reasoning	ARC-C	0	acc	82.4	83.4	94.4	94.8	96.9
	GPQA	0	em	34.6	30.4	39.5	46.7	50.7
Code	HumanEval	0	pass@1	60.4	72.6	81.7	80.5	89.0
	MBPP ++ base version	0	pass@1	70.6	72.8	82.5	86.0	88.6
	Multipl-E HumanEval	0	pass@1	•	50.8	•	65.5	75.2
	Multipl-E MBPP	0	pass@1	•	52.4	•	62.0	65.7
Math	GSM-8K (CoT)	8	em_maj1@1	80.6	84.5	93.0	95.1	96.8
	MATH (CoT)	0	final_em	29.1	51.9	51.0	68.0	73.8
Tool Use	API-Bank	0	acc	48.3	82.6	85.1	90.0	92.0
	BFCL	0	acc	60.3	76.1	83.0	84.8	88.5
	Gorilla Benchmark API Bench	0	acc	1.7	8.2	14.7	29.7	35.3
	Nexus (0-shot)	0	macro_avg/acc	18.1	38.5	47.8	56.7	58.7
Multilingual	Multilingual MGSM (CoT)	0	em	•	68.9	•	86.9	91.6

Multilingual benchmarks

Category	Benchmark	Language	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama 3.1 405B Instruct
General	MMLU (5-shot, macro_avg/acc)	Portuguese	62.12	80.13	84.95
		Spanish	62.45	80.05	85.08
		Italian	61.63	80.4	85.04
		German	60.59	79.27	84.36
		French	62.34	79.82	84.66
		Hindi	50.88	74.52	80.31
		Thai	50.32	72.95	78.21

Responsibility & Safety

As part of our Responsible release approach, we followed a three-pronged strategy to managing trust & safety risks:

- Enable developers to deploy helpful, safe and flexible experiences for their target audience and for the use cases supported by Llama.
- Protect developers against adversarial users aiming to exploit Llama capabilities to potentially cause harm.
- Provide protections for the community to help prevent the misuse of our models.

Responsible deployment

Llama is a foundational technology designed to be used in a variety of use cases, examples on how Meta's Llama models have been responsibly deployed can be found in our Community Stories webpage. Our approach is to build the most helpful models enabling the world to benefit from the technology power, by aligning our model safety for the generic use cases addressing a standard set of harms. Developers are then in the driver seat to tailor safety for their use case, defining their own policy and deploying the models with the necessary safeguards in their Llama systems. Llama 3.1 was developed following the best practices outlined in our Responsible Use Guide, you can refer to the Responsible Use Guide to learn more.

Llama 3.1 instruct

Our main objectives for conducting safety fine-tuning are to provide the research community with a valuable resource for studying the robustness of safety fine-tuning, as well as to offer developers a readily available, safe, and powerful model for various applications to reduce the developer workload to deploy safe AI systems. For more details on the safety mitigations implemented please read the Llama 3 paper.

Fine-tuning data

We employ a multi-faceted approach to data collection, combining human-generated data from our vendors with synthetic data to mitigate potential safety risks. We've developed many large language model (LLM)-based classifiers that enable us to thoughtfully select high-quality prompts and responses, enhancing data quality control.

Refusals and Tone

Building on the work we started with Llama 3, we put a great emphasis on model refusals to benign prompts as well as refusal tone. We included both borderline and adversarial prompts in our safety data strategy, and modified our safety data responses to follow tone guidelines.

Llama 3.1 systems

Large language models, including Llama 3.1, are not designed to be deployed in isolation but instead should be deployed as part of an overall AI system with additional safety guardrails as required. Developers are expected to deploy system safeguards when building agentic systems. Safeguards are key to achieve the right helpfulness-safety alignment as well as mitigating safety and security risks inherent to the system and any integration of the model or system with external tools.

As part of our responsible release approach, we provide the community with safeguards that developers should deploy with Llama models or other LLMs, including Llama Guard 3, Prompt Guard and Code Shield. All our reference implementations demos contain these safeguards by default so developers can benefit from system-level safety out-of-the-box.

New capabilities

Note that this release introduces new capabilities, including a longer context window, multilingual inputs and outputs and possible integrations by developers with third party tools. Building with these new capabilities requires specific considerations in addition to the best practices that generally apply across all Generative AI use cases.

Tool-use: Just like in standard software development, developers are responsible for the integration of the LLM with the tools and services of their choice. They should define a clear policy for their use case and assess the integrity of the third party services they use to be aware of the safety and security limitations when using this capability. Refer to the Responsible Use Guide for best practices on the safe deployment of the third party safeguards.

Multilinguality: Llama 3.1 supports 7 languages in addition to English: French, German, Hindi, Italian, Portuguese, Spanish, and Thai. Llama may be able to output text in other languages than those that meet performance thresholds for safety and helpfulness. We strongly discourage developers from using this model to converse in non-supported languages without implementing finetuning and system controls in alignment with their policies and the best practices shared in the Responsible Use Guide.

Evaluations

We evaluated Llama models for common use cases as well as specific capabilities. Common use cases evaluations measure safety risks of systems for most commonly built applications including chat bot, coding assistant, tool calls. We built dedicated, adversarial evaluation datasets and evaluated systems composed of Llama models and Llama Guard 3 to filter input prompt and output response. It is important to evaluate applications in context, and we recommend building dedicated evaluation dataset for your use case. Prompt Guard and Code Shield are also available if relevant to the application.

Capability evaluations measure vulnerabilities of Llama models inherent to specific capabilities, for which were crafted dedicated benchmarks including long context, multilingual, tools calls, coding or memorization.

Red teaming

For both scenarios, we conducted recurring red teaming exercises with the goal of discovering risks via adversarial prompting and we used the learnings to improve our benchmarks and safety tuning datasets.

We partnered early with subject-matter experts in critical risk areas to understand the nature of these real-world harms and how such models may lead to unintended harm for society. Based on these conversations, we derived a set of adversarial goals for the red team to attempt to achieve, such as extracting harmful information or reprogramming the model to act in a potentially harmful capacity. The red team consisted of experts in cybersecurity, adversarial machine learning, responsible AI, and integrity in addition to multilingual content specialists with background in integrity issues in specific geographic markets. .

Critical and other risks

We specifically focused our efforts on mitigating the following critical risk areas:

1. CBRNE (Chemical, Biological, Radiological, Nuclear, and Explosive materials) helpfulness

To assess risks related to proliferation of chemical and biological weapons, we performed uplift testing designed to assess whether use of Llama 3.1 models could meaningfully increase the capabilities of malicious actors to plan or carry out attacks using these types of weapons.

2. Child Safety

Child Safety risk assessments were conducted using a team of experts, to assess the model's capability to produce outputs that could result in Child Safety risks and inform on any necessary and appropriate risk mitigations via fine tuning. We leveraged those expert red teaming sessions to expand the coverage of our evaluation benchmarks through Llama 3 model development. For Llama 3, we conducted new in-depth sessions using objective based methodologies to assess the model risks along multiple attack vectors including the additional languages Llama 3 is trained on. We also partnered with content specialists to perform red teaming exercises assessing potentially violating content while taking account of market specific nuances or experiences.

3. Cyber attack enablement

Our cyber attack uplift study investigated whether LLMs can enhance human capabilities in hacking tasks, both in terms of skill level and speed.

Our attack automation study focused on evaluating the capabilities of LLMs when used as autonomous agents in cyber offensive operations, specifically in the context of ransomware attacks. This evaluation was distinct from previous studies that considered LLMs as interactive assistants. The primary objective was to assess whether these models could effectively function as independent agents in executing complex cyber-attacks without human intervention.

Our study of Llama 3.1 405B's social engineering uplift for cyber attackers was conducted to assess the effectiveness of AI models in aiding cyber threat actors in spear phishing campaigns. Please read our Llama 3.1 Cyber security whitepaper to learn more.

Community

Generative AI safety requires expertise and tooling, and we believe in the strength of the open community to accelerate its progress. We are active members of open consortiums, including the AI Alliance, Partnership on AI and MLCommons, actively contributing to safety standardization and transparency. We encourage the community to adopt taxonomies like the MLCommons Proof of Concept evaluation to facilitate collaboration and transparency on safety and

content evaluations. Our Purple Llama tools are open sourced for the community to use and widely distributed across ecosystem partners including cloud service providers. We encourage community contributions to our Github repository.

We also set up the Llama Impact Grants program to identify and support the most compelling applications of Meta's Llama model for societal benefit across three categories: education, climate and open innovation. The 20 finalists from the hundreds of applications can be found here.

Finally, we put in place a set of resources including an output reporting mechanism and bug bounty program to continuously improve the Llama technology with the help of the community.

Ethical Considerations and Limitations

The core values of Llama 3.1 are openness, inclusivity and helpfulness. It is meant to serve everyone, and to work for a wide range of use cases. It is thus designed to be accessible to people across many different backgrounds, experiences and perspectives. Llama 3.1 addresses users and their needs as they are, without insertion unnecessary judgment or normativity, while reflecting the understanding that even content that may appear problematic in some cases can serve valuable purposes in others. It respects the dignity and autonomy of all users, especially in terms of the values of free thought and expression that power innovation and progress.

But Llama 3.1 is a new technology, and like any new technology, there are risks associated with its use. Testing conducted to date has not covered, nor could it cover, all scenarios. For these reasons, as with all LLMs, Llama 3.1's potential outputs cannot be predicted in advance, and the model may in some instances produce inaccurate, biased or other objectionable responses to user prompts. Therefore, before deploying any applications of Llama 3.1 models, developers should perform safety testing and tuning tailored to their specific applications of the model. Please refer to available resources including our Responsible Use Guide, Trust and Safety solutions, and other resources to learn more about responsible development.