

LLAMA 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron* Louis Martin† Kevin Stone†

Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra
 Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen
 Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller
 Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou
 Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev
 Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich
 Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra
 Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi
 Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang
 Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang
 Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic
 Sergey Edunov Thomas Scialom*

GenAI, Meta

Abstract

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called LLAMA 2-CHAT, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of LLAMA 2-CHAT in order to enable the community to build on our work and contribute to the responsible development of LLMs.

*Equal contribution, corresponding authors: {tscialom, htouvron}@meta.com

†Second author

Contents

1	Introduction	3
2	Pretraining	5
2.1	Pretraining Data	5
2.2	Training Details	5
2.3	LLAMA 2 Pretrained Model Evaluation	7
3	Fine-tuning	8
3.1	Supervised Fine-Tuning (SFT)	9
3.2	Reinforcement Learning with Human Feedback (RLHF)	9
3.3	System Message for Multi-Turn Consistency	16
3.4	RLHF Results	17
4	Safety	20
4.1	Safety in Pretraining	20
4.2	Safety Fine-Tuning	23
4.3	Red Teaming	28
4.4	Safety Evaluation of LLAMA 2-CHAT	29
5	Discussion	32
5.1	Learnings and Observations	32
5.2	Limitations and Ethical Considerations	34
5.3	Responsible Release Strategy	35
6	Related Work	35
7	Conclusion	36
A	Appendix	46
A.1	Contributions	46
A.2	Additional Details for Pretraining	47
A.3	Additional Details for Fine-tuning	51
A.4	Additional Details for Safety	58
A.5	Data Annotation	72
A.6	Dataset Contamination	75
A.7	Model Card	77

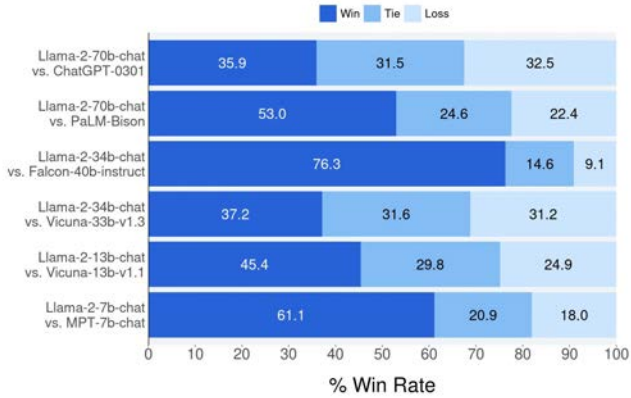


Figure 1: Helpfulness human evaluation results for LLAMA 2-CHAT compared to other open-source and closed-source models. Human raters compared model generations on ~4k prompts consisting of both single and multi-turn prompts. The 95% confidence intervals for this evaluation are between 1% and 2%. More details in Section 3.4.2. While reviewing these results, it is important to note that human evaluations can be noisy due to limitations of the prompt set, subjectivity of the review guidelines, subjectivity of individual raters, and the inherent difficulty of comparing generations.

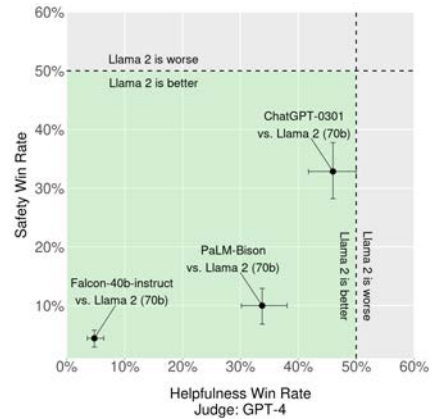


Figure 2: Win-rate % for helpfulness and safety between commercial-licensed baselines and LLAMA 2-CHAT, according to GPT-4. To complement the human evaluation, we used a more capable model, not subject to our own guidance. Green area indicates our model is better according to GPT-4. To remove ties, we used $win/(win + loss)$. The orders in which the model responses are presented to GPT-4 are randomly swapped to alleviate bias.

1 Introduction

Large Language Models (LLMs) have shown great promise as highly capable AI assistants that excel in complex reasoning tasks requiring expert knowledge across a wide range of fields, including in specialized domains such as programming and creative writing. They enable interaction with humans through intuitive chat interfaces, which has led to rapid and widespread adoption among the general public.

The capabilities of LLMs are remarkable considering the seemingly straightforward nature of the training methodology. Auto-regressive transformers are pretrained on an extensive corpus of self-supervised data, followed by alignment with human preferences via techniques such as Reinforcement Learning with Human Feedback (RLHF). Although the training methodology is simple, high computational requirements have limited the development of LLMs to a few players. There have been public releases of pretrained LLMs (such as BLOOM (Scao et al., 2022), LLaMa-1 (Touvron et al., 2023), and Falcon (Penedo et al., 2023)) that match the performance of closed pretrained competitors like GPT-3 (Brown et al., 2020) and Chinchilla (Hoffmann et al., 2022), but none of these models are suitable substitutes for closed “product” LLMs, such as ChatGPT, BARD, and Claude. These closed product LLMs are heavily fine-tuned to align with human preferences, which greatly enhances their usability and safety. This step can require significant costs in compute and human annotation, and is often not transparent or easily reproducible, limiting progress within the community to advance AI alignment research.

In this work, we develop and release Llama 2, a family of pretrained and fine-tuned LLMs, *LLAMA 2* and *LLAMA 2-CHAT*, at scales up to 70B parameters. On the series of helpfulness and safety benchmarks we tested, *LLAMA 2-CHAT* models generally perform better than existing open-source models. They also appear to be on par with some of the closed-source models, at least on the human evaluations we performed (see Figures 1 and 3). We have taken measures to increase the safety of these models, using safety-specific data annotation and tuning, as well as conducting red-teaming and employing iterative evaluations. Additionally, this paper contributes a thorough description of our fine-tuning methodology and approach to improving LLM safety. We hope that this openness will enable the community to reproduce fine-tuned LLMs and continue to improve the safety of those models, paving the way for more responsible development of LLMs. We also share novel observations we made during the development of *LLAMA 2* and *LLAMA 2-CHAT*, such as the emergence of tool usage and temporal organization of knowledge.

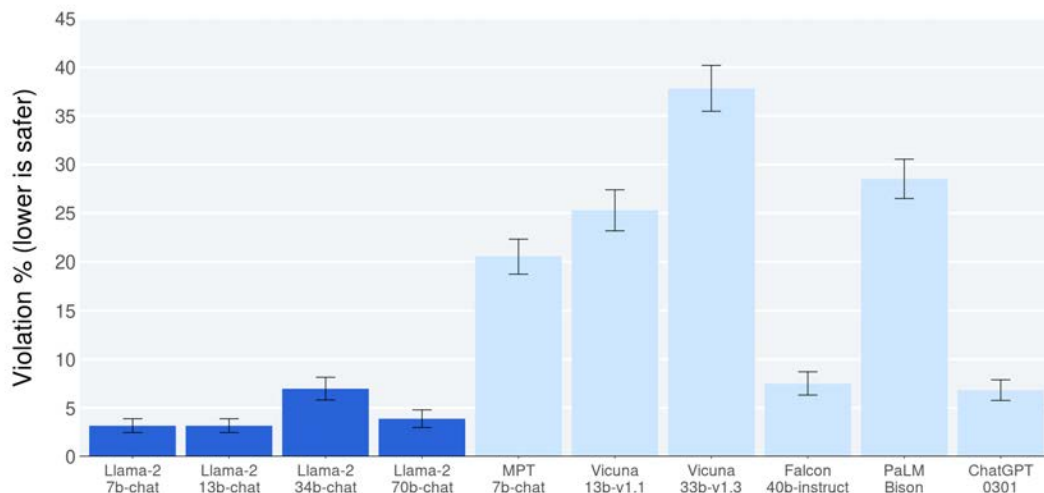


Figure 3: Safety human evaluation results for LLAMA 2-CHAT compared to other open-source and closed-source models. Human raters judged model generations for safety violations across ~2,000 adversarial prompts consisting of both single and multi-turn prompts. More details can be found in Section 4.4. It is important to caveat these safety results with the inherent bias of LLM evaluations due to limitations of the prompt set, subjectivity of the review guidelines, and subjectivity of individual raters. Additionally, these safety evaluations are performed using content standards that are likely to be biased towards the LLAMA 2-CHAT models.

We are releasing the following models to the general public for research and commercial use[‡]:

1. **LLAMA 2**, an updated version of LLAMA 1, trained on a new mix of publicly available data. We also increased the size of the pretraining corpus by 40%, doubled the context length of the model, and adopted grouped-query attention (Ainslie et al., 2023). We are releasing variants of LLAMA 2 with 7B, 13B, and 70B parameters. We have also trained 34B variants, which we report on in this paper but are not releasing.[§]
2. **LLAMA 2-CHAT**, a fine-tuned version of LLAMA 2 that is optimized for dialogue use cases. We release variants of this model with 7B, 13B, and 70B parameters as well.

We believe that the open release of LLMs, when done safely, will be a net benefit to society. Like all LLMs, LLAMA 2 is a new technology that carries potential risks with use (Bender et al., 2021b; Weidinger et al., 2021; Solaiman et al., 2023). Testing conducted to date has been in English and has not — and could not — cover all scenarios. Therefore, before deploying any applications of LLAMA 2-CHAT, developers should perform safety testing and tuning tailored to their specific applications of the model. We provide a responsible use guide[¶] and code examples^{||} to facilitate the safe deployment of LLAMA 2 and LLAMA 2-CHAT. More details of our responsible release strategy can be found in Section 5.3.

The remainder of this paper describes our pretraining methodology (Section 2), fine-tuning methodology (Section 3), approach to model safety (Section 4), key observations and insights (Section 5), relevant related work (Section 6), and conclusions (Section 7).

[‡]<https://ai.meta.com/resources/models-and-libraries/llama/>

[§]We are delaying the release of the 34B model due to a lack of time to sufficiently red team.

[¶]<https://ai.meta.com/llama>

^{||}<https://github.com/facebookresearch/llama>