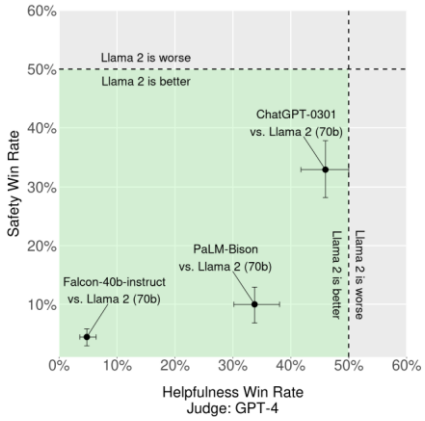
Llama 2: Open Foundation and Fine-Tuned Chat Models

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 opensource chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closedsource models. We provide a detailed description of our approach to finetuning and safety improvements of Llama 2-Chat in order to enable the community to build on our work.

1 Introduction

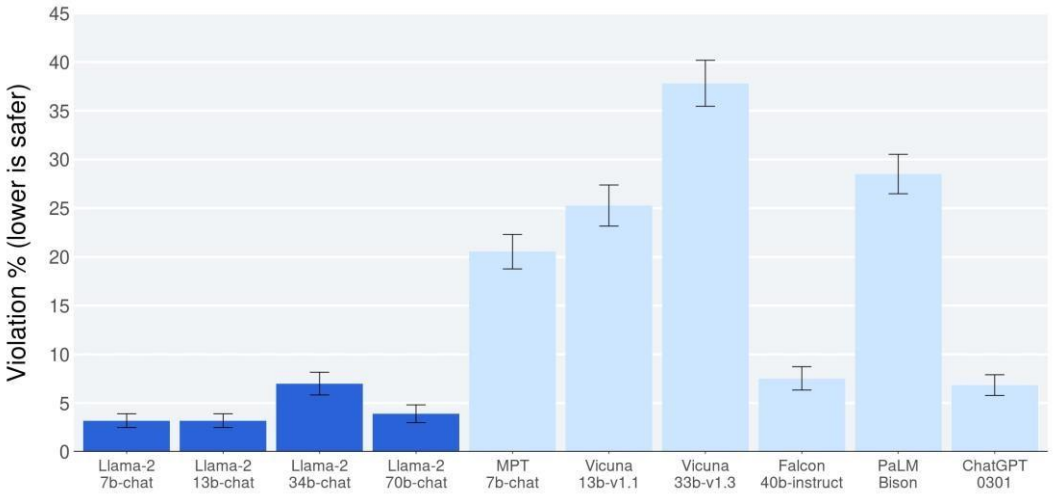


To complement the human evaluation, we used a more capable model, not subject to our own guidance. Green area indicates our model is better according to GPT-4.

The orders in which the model responses are presented to GPT-4 are randomly swapped to alleviate bias.

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.

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.



We are releasing the following models to the general public for research and commercial use[1]:

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 groupedquery 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]

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.

2 Pretraining

To create the new family of Llama 2 models, we began with the pretraining approach described in Touvron et al. (2023), using an optimized auto-regressive transformer, but made several changes to improve performance. Specifically, we performed more robust data cleaning, updated our data mixes, trained on 40% more total tokens, doubled the context length, and used grouped-query attention (GQA) to improve inference scalability for our larger models. Table 1 compares the attributes of the new Llama 2 models with the Llama 1 models.

2.1 Pretraining Data

Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.

Limitations of human evaluations. While our results indicate that Llama 2-Chat is on par with ChatGPT on human evaluations, it is important to note that human evaluations have several limitations.

[1] https://ai.meta.com/resources/models-and-libraries/llama/

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

- By academic and research standards, we have a large prompt set of 4k prompts. However, it does not cover realworld usage of these models, which will likely cover a significantly larger number of use cases.
- Diversity of the prompts could be another factor in our results. For example, our prompt set does not include any coding- or reasoning-related prompts.
- We only evaluate the final generation of a multi-turn conversation. A more interesting evaluation could be to ask the models to complete a task and rate the overall experience with the model over multiple turns.
- Human evaluation for generative models is inherently subjective and noisy. As a result, evaluation on a different set of prompts or with different instructions could result in different results.

Conclusion

In this study, we have introduced Llama 2, a new family of pretrained and fine-tuned models with scales of 7 billion to 70 billion parameters. These models have demonstrated their competitiveness with existing open-source chat models, as well as competency that is equivalent to some proprietary models on evaluation sets we examined, although they still lag behind other models like GPT-4. We meticulously elaborated on the methods and techniques applied in achieving our models, with a heavy emphasis on their alignment with the principles of helpfulness and safety. To contribute more significantly to society and foster the pace of research, we have responsibly opened access to Llama 2 and Llama 2-Chat. As part of our ongoing commitment to transparency and safety, we plan to make further improvements to Llama 2-Chat in future work.