Meta AI

Llama 2 : Open Foundation and Fine-Tuned Chat Models

GenAI Team, Meta

Presented By :
Vedanuj Goswami
Research Engineer, MetaAI

Meta AI

Meta AI

Agenda

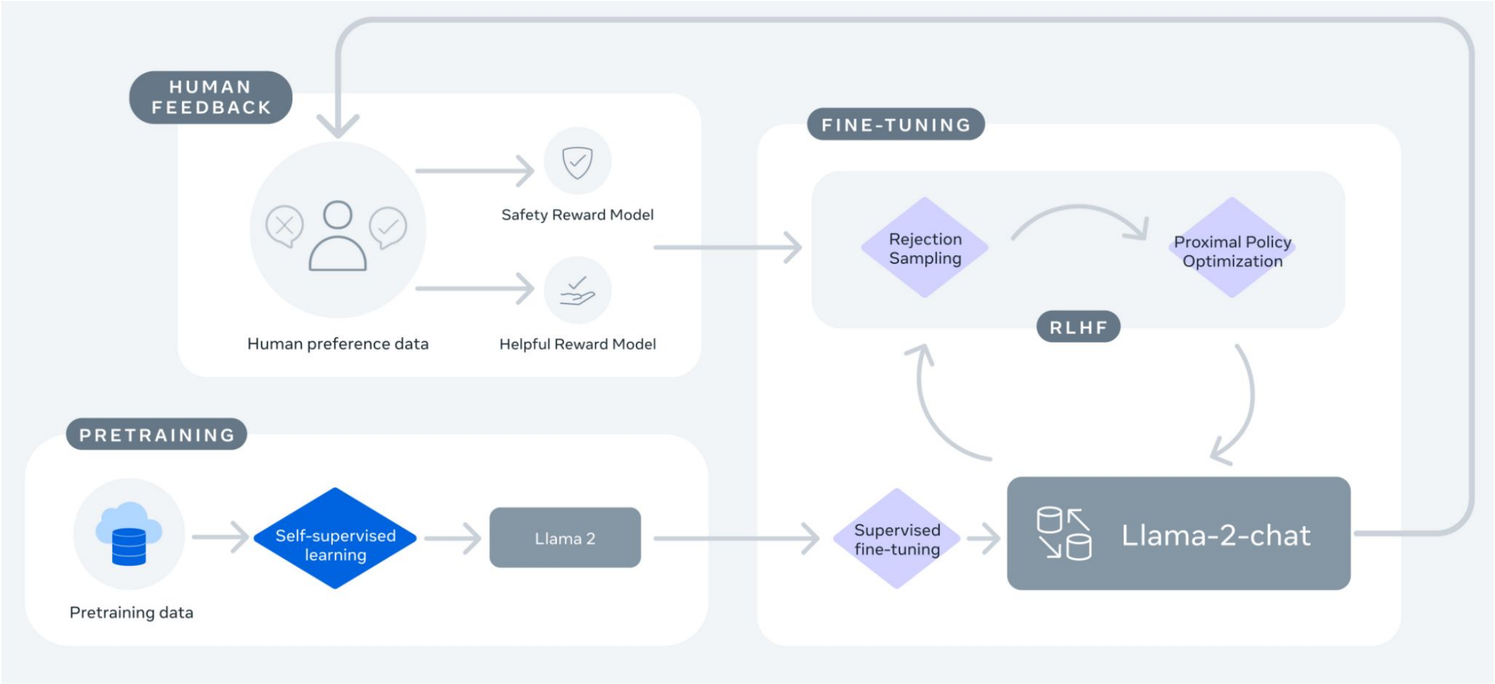
01 Introduction

02 Pretraining

03 Finetuning

04 Safety

05 Conclusion



Pretraining

Scaling up on both data and compute, training strong base models to improve knowledge of these models.

Finetuning

Finetuning and aligning the models to be more like chat assistants, and ensuring they are helpful and harmless.

Safety

Taking 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.

Pretraining

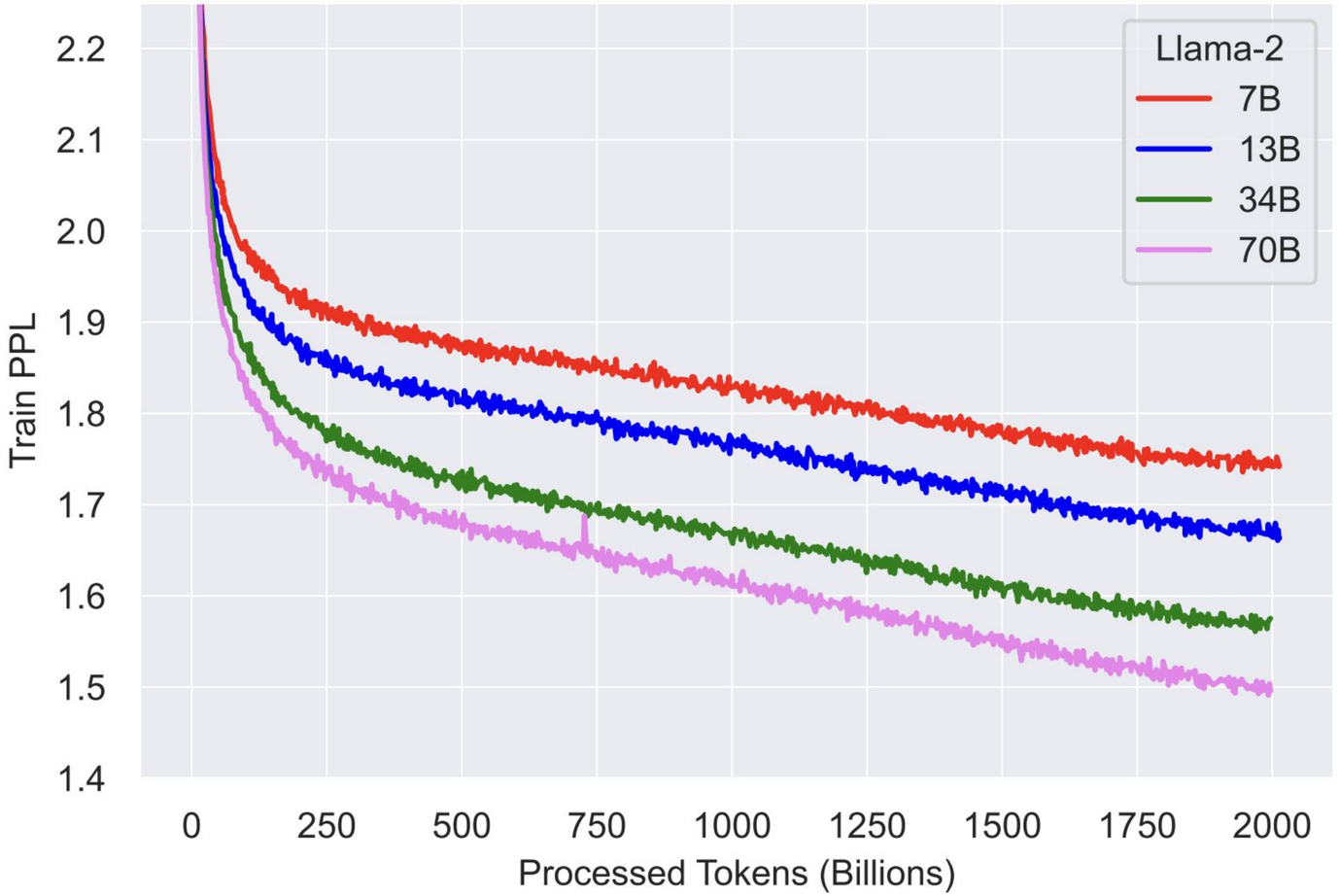
- 2T Tokens for all models, 40% more tokens than Llama-1
- 1.5x to 7x more compute used compared to Llama-1 models
- Longer Context 4K
- Grouped Query Attention for Inference Efficiency
- Scaling Training beyond 2K GPUs

Llama 2 Family

	Training Data	Params	Context Length	GQA	Tokens	LR
Llama 1	*See Touvron et al. (2023)*	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
Llama 2	*A new mix of publicly available online data*	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

More Compute and Longer Training

- Llama 2 70B model uses total compute of ~8.26e23 FLOPs, 1.5x more than Llama 1.
- Models have not yet converged, showing more room for training further into “inference optimal” regime.



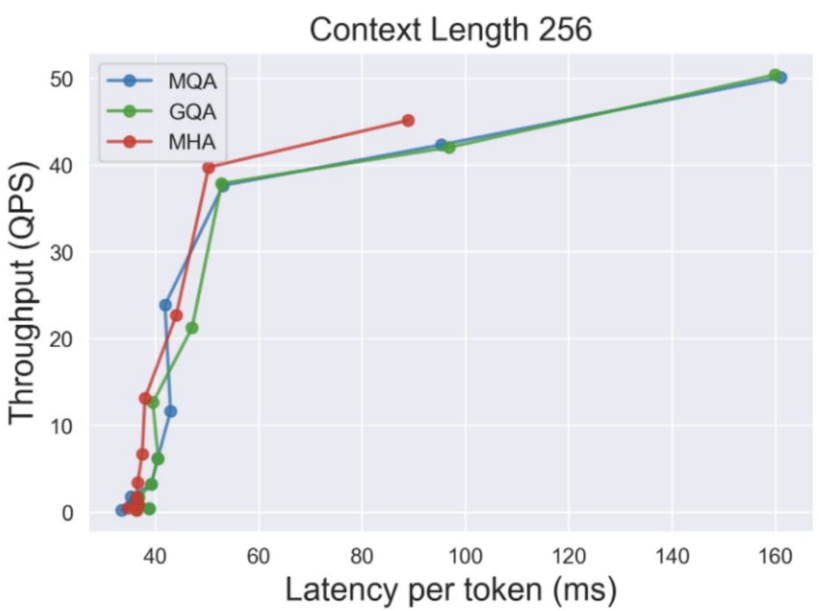
Long Context Pretraining : 2K → 4K

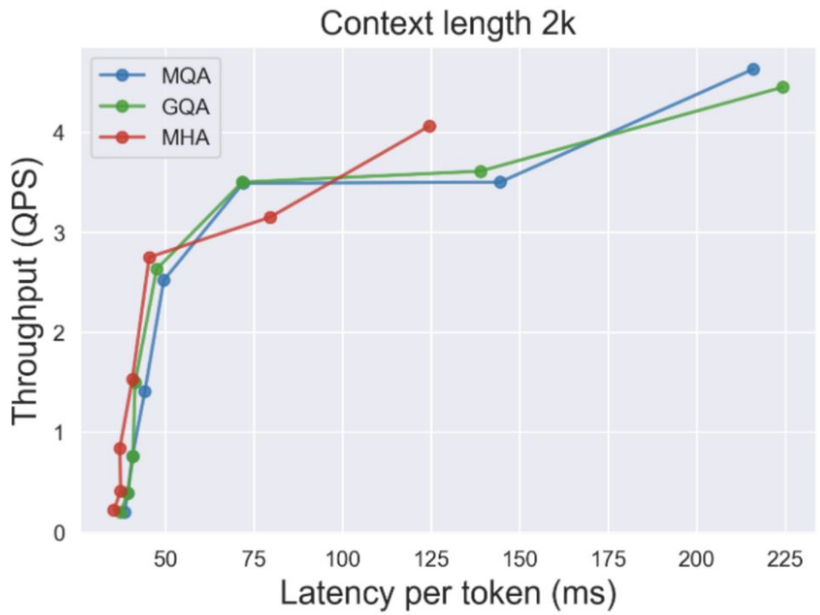
- Useful for supporting longer histories in chat applications, various summarization tasks, understanding longer documents, coding etc
- Context length to use in pretraining is determined by the pretraining data distribution
- Continued pretraining these base model on longer context data can support context lengths much larger than 4k(8k, 16k, 32k etc).

Context Length	NarrativeQA (F1)	Qasper (F1)	QuALITY (acc)	QMSum (Rouge 1/2/L)	ContractNLI (EM)	SQuAD (EM/F1)
2k	0.21	0.71	26.1	0.13/0.01/0.12	11.76	57.23/62.89
4k	**17.26**	**18.52**	**29.6**	**15.08/3.55/12.16**	**16.33**	**57.99/64.46**

Ablation of increasing context length on different long context tasks

Grouped-Query Attention





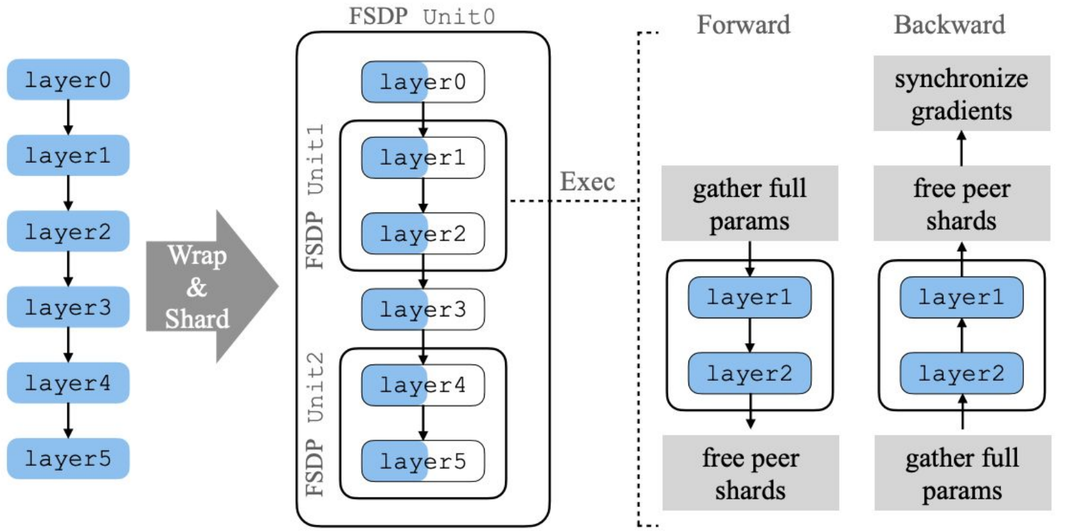
Throughput vs Latency as we increase batch size for different variants. MHA results in OOM at larger batch sizes, while MQA and GQA variants do not.

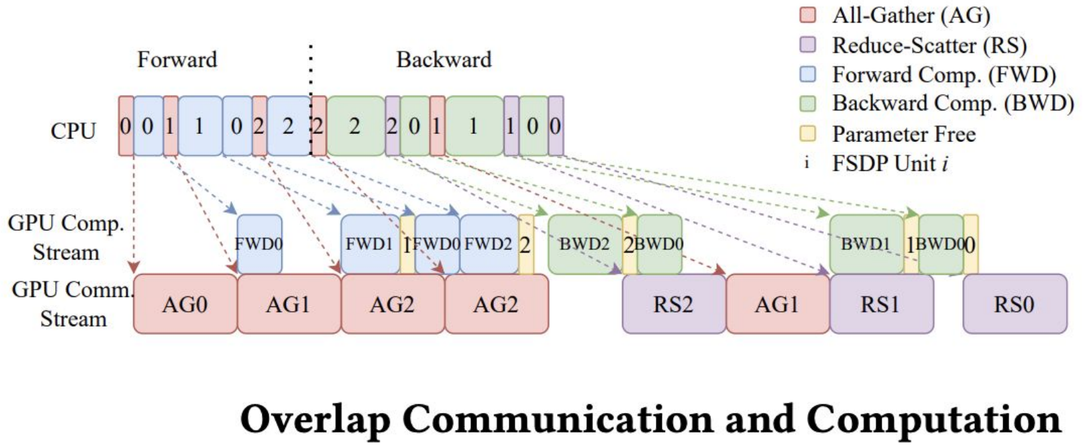
	BoolQ	PIQA	SIQA	Hella-Swag	ARC-e	ARC-c	NQ	TQA	MMLU	GSM8K	Human-Eval
MHA	**71.0**	**79.3**	48.2	75.1	71.2	**43.0**	12.4	44.7	**28.0**	4.9	**7.9**
MQA	70.6	79.0	47.9	74.5	71.6	41.9	**14.5**	42.8	26.5	4.8	7.3
GQA	69.4	78.8	**48.6**	**75.4**	**72.1**	42.5	14.0	**46.2**	26.9	**5.3**	**7.9**

Parallelism

Scaling to 2k+ GPUs require efficient parallelism schemes

- FSDP + Communication Computation Overlap
- Tensor Parallel
- Sequence Parallel
- Selective Activation Recomputation





Overlap Communication and Computation

Image Source : Zhao, Yanli, et al. "Pytorch FSDP: experiences on scaling fully sharded data parallel." arXiv preprint arXiv:2304.11277 (2023).

Pretrained Model Evaluation

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	**37.5**	**71.9**	**63.6**	**69.4**	**35.2**	**68.9**	**51.2**	**54.2**

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	**86.1**	85.0
Natural Questions (1-shot)	–	–	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	**65.7**	51.2

Finetuning

- SFT
- Reward Models
- Iterative Finetuning with RLHF : Rejection Sampling & PPO
- Multiturn Consistency using GAtt

SFT Data

➤ Prompt:	Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.
Response:	Hydrogen comes first as element number one. Helium is second for balloons to have fun! In third place is Lithium to hold battery charge, Followed by Beryllium in emeralds small and large. Boron's number five to help us keep things clean. Carbon's next at six, and it's found in every gene. Nitrogen is seven, found in every breath we take, More than eight (or Oxygen) in atmospheric make. Number nine is Fluorine, helping polish up our teeth. Neon gives us glowing signs to read and stand beneath.
➤ Prompt:	I want you to roast me. I want you to make it particularly brutal, swearing at me.
Response:	I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

SFT Annotation Example

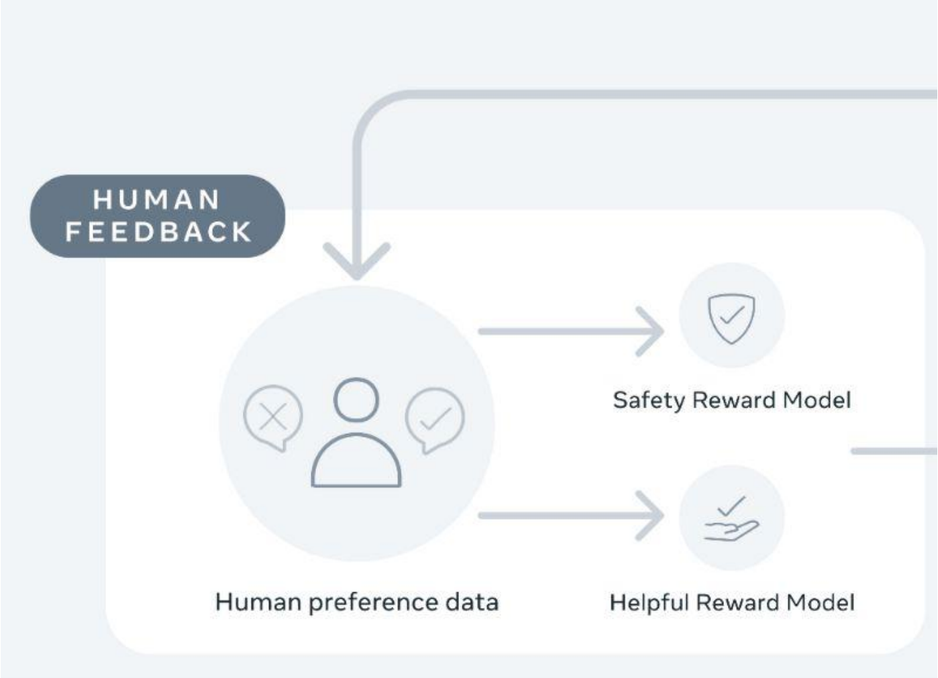
- 3rd party datasets lack diversity and quality, for dialog style instructions.
- Focus on fewer but clean instruction-tuning data for higher quality models.
- Collected about 27k samples.
- SFT model output often matched or outperformed human annotated data. So better to focus budget on Human Preference data annotation.

Human Preference Data

Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

Statistics of human preference data for reward modeling. We collected >1M samples, with an weekly cadence of data batches. Meta RM data had overall higher average tokens, turns and length of response per dialogue.

Reward Modeling



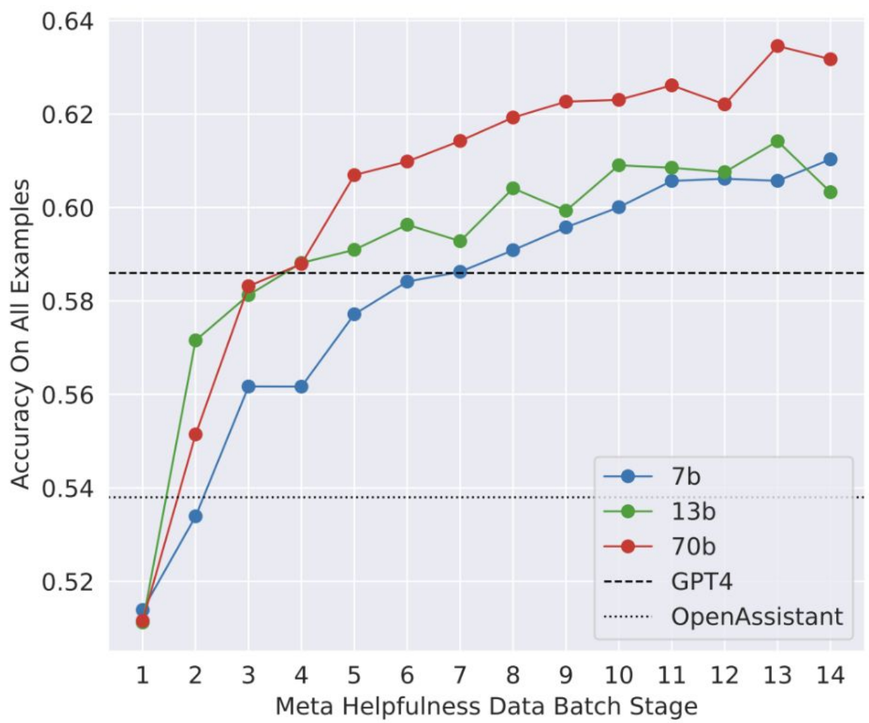
We train two reward models, one optimized for helpfulness (Helpfulness RM) and other for safety (Safety RM)

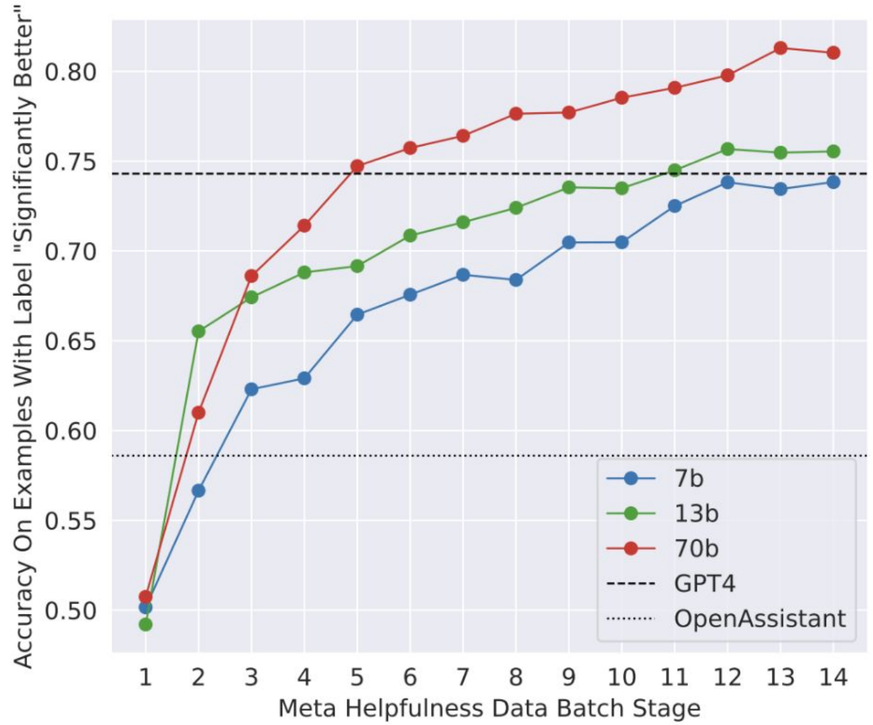
Reward Model Results

	Meta Helpful.	Meta Safety	Anthropic Helpful	Anthropic Harmless	OpenAI Summ.	Stanford SHP	Avg
SteamSHP-XL	52.8	43.8	66.8	34.2	54.7	75.7	55.3
Open Assistant	53.8	53.4	67.7	68.4	71.7	55.0	63.0
GPT4	58.6	58.1	-	-	-	-	-
Safety RM	56.2	64.5	55.4	74.7	71.7	65.2	64.3
Helpfulness RM	63.2	62.8	72.0	71.0	75.5	80.0	70.6

Performance of our Helpfulness RM and Safety RM models on a diverse set of human preference benchmarks. Note that our model is fine-tuned on our collected data, as opposed to the other baselines that we report.

Scaling trends for Reward Models

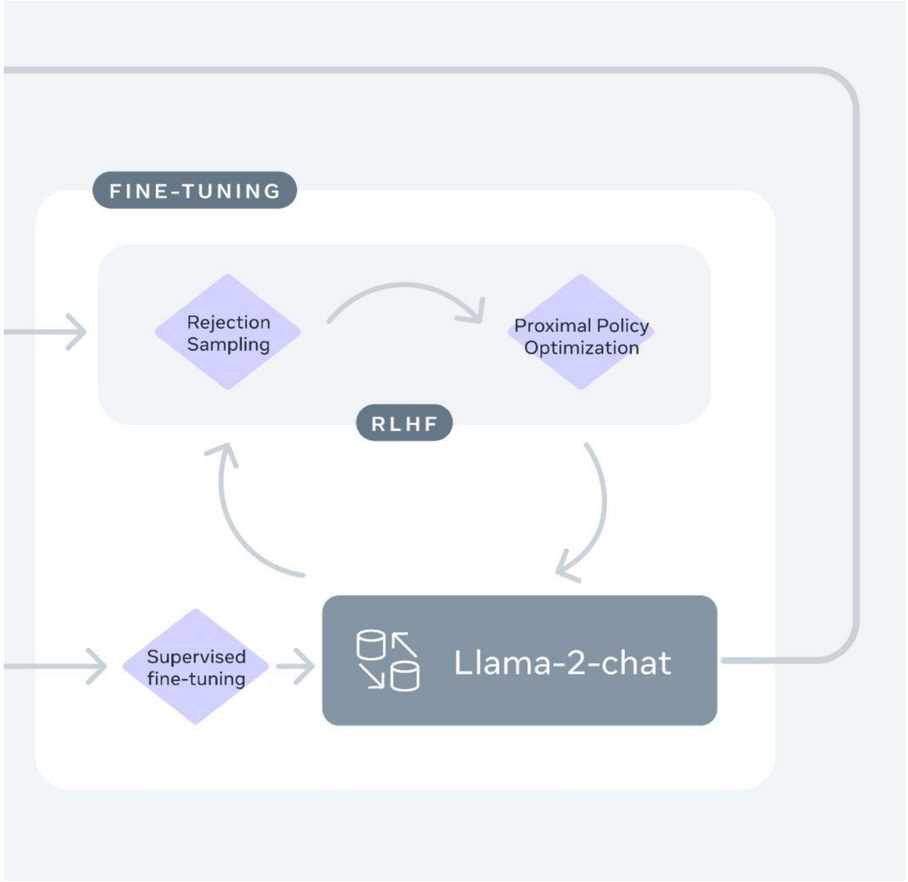




More data and a larger-size model generally improve accuracy, and it appears that our models have not yet saturated from learning on the human preference training data.

Iterative Finetuning with RLHF

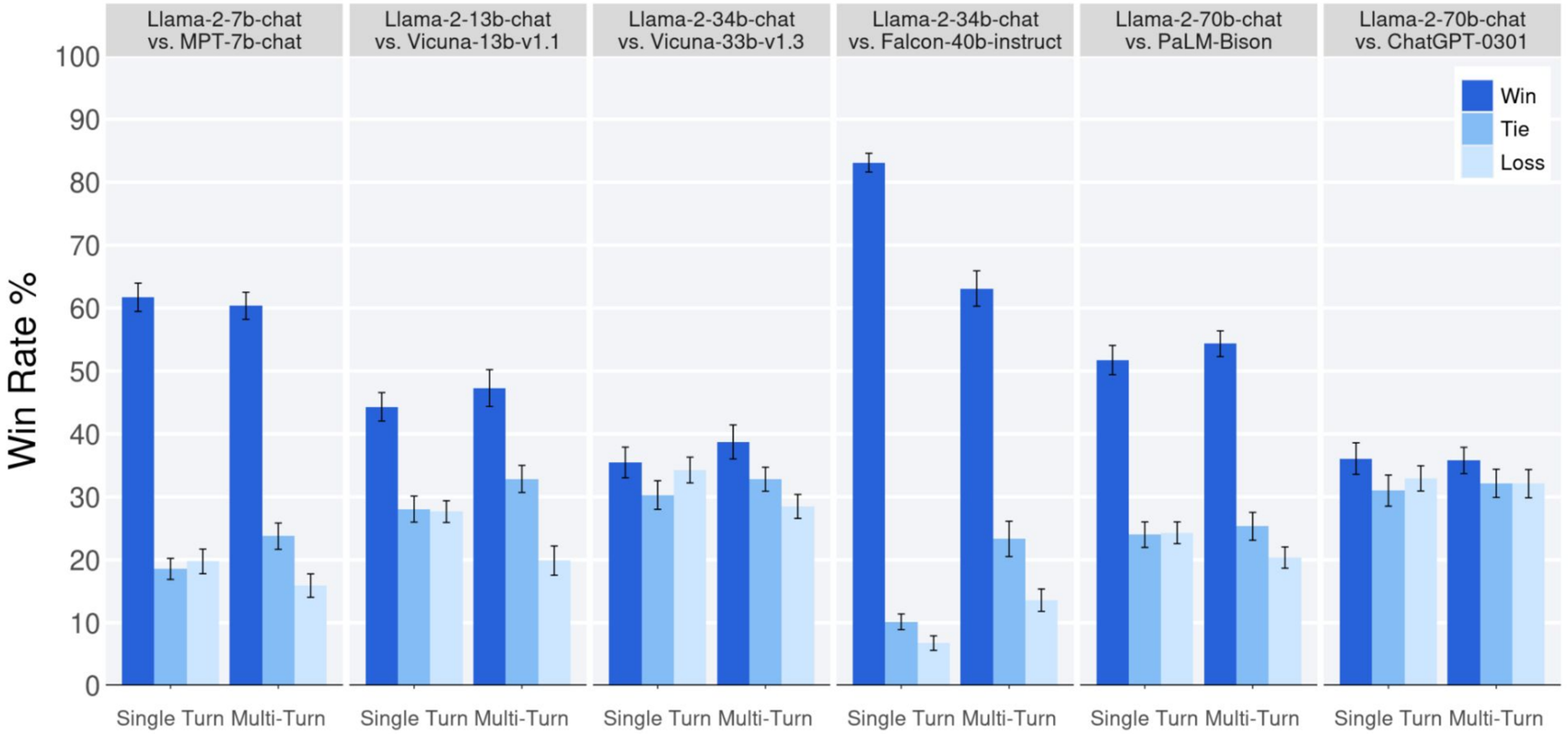
- Iterative versions : RLHF-V1, . . . , RLHF-V5
- Two approaches :
  - Proximal Policy Optimization (PPO)
  - Rejection Sampling Fine-Tuning
- Sequential Combination of Both Algorithms:
  - Until RLHF-V4, only Rejection Sampling was used.
  - Post RLHF-V4, a combination of both was used, sequentially applying PPO on the result of the Rejection Sampling checkpoint before sampling again.



Evolution of Llama-2-Chat Models

We show the evolution after multiple iterations fine-tuning for the win-rate % of Llama-2-Chat compared to ChatGPT.

Human Eval Results

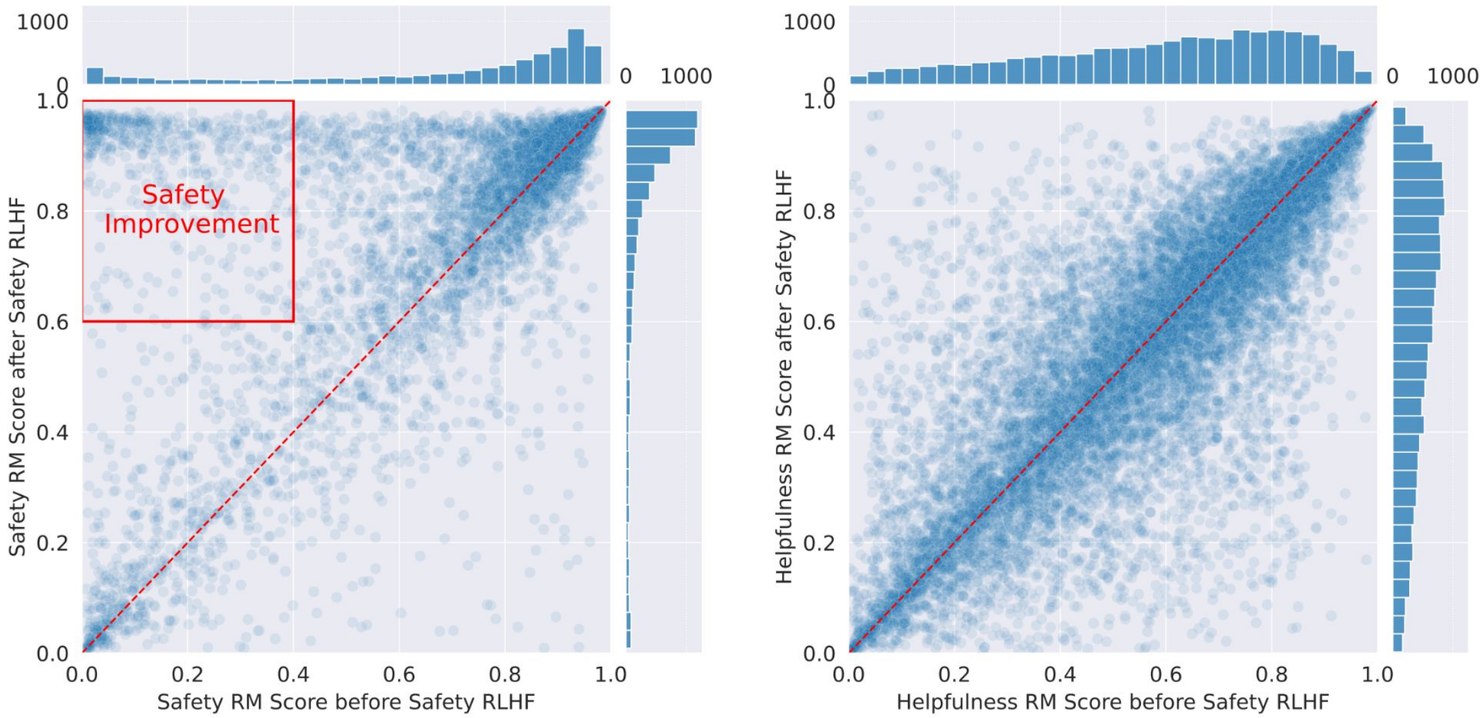


Human evaluation results for Llama 2-Chat models compared to open- and closed-source models across ~4,000 helpfulness prompts with three raters per prompt.

Safety

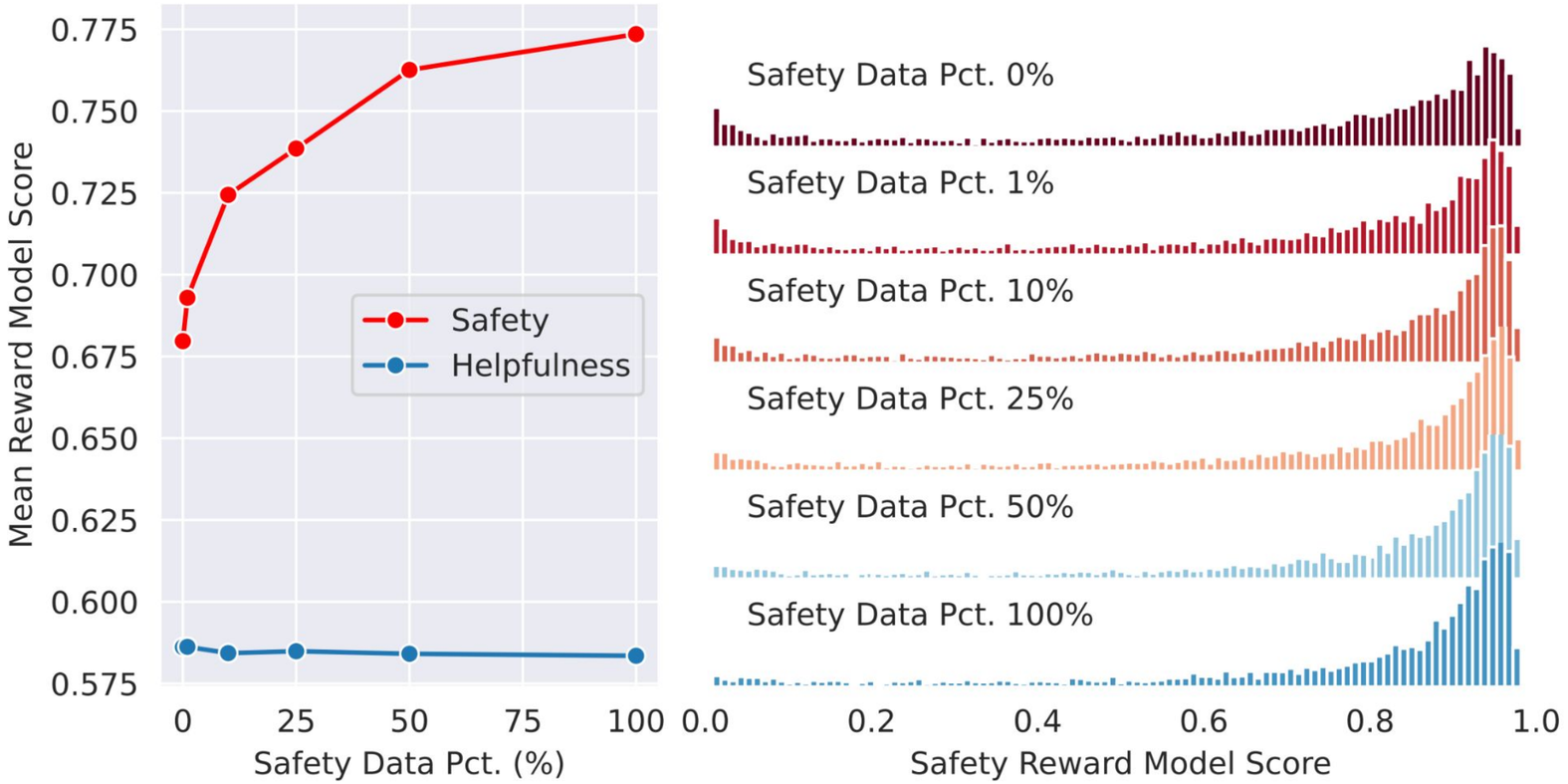
- Safety in Pretraining
- SFT
- Safety RLHF
- Context Distillation
- Continuous Red Teaming

Impact of Safety RLHF



We compare before and after Safety RLHF Llama 2-Chat checkpoints. Results showed an improvement in safety scores with safety tuning via RLHF, with no significant degradation in helpfulness scores.

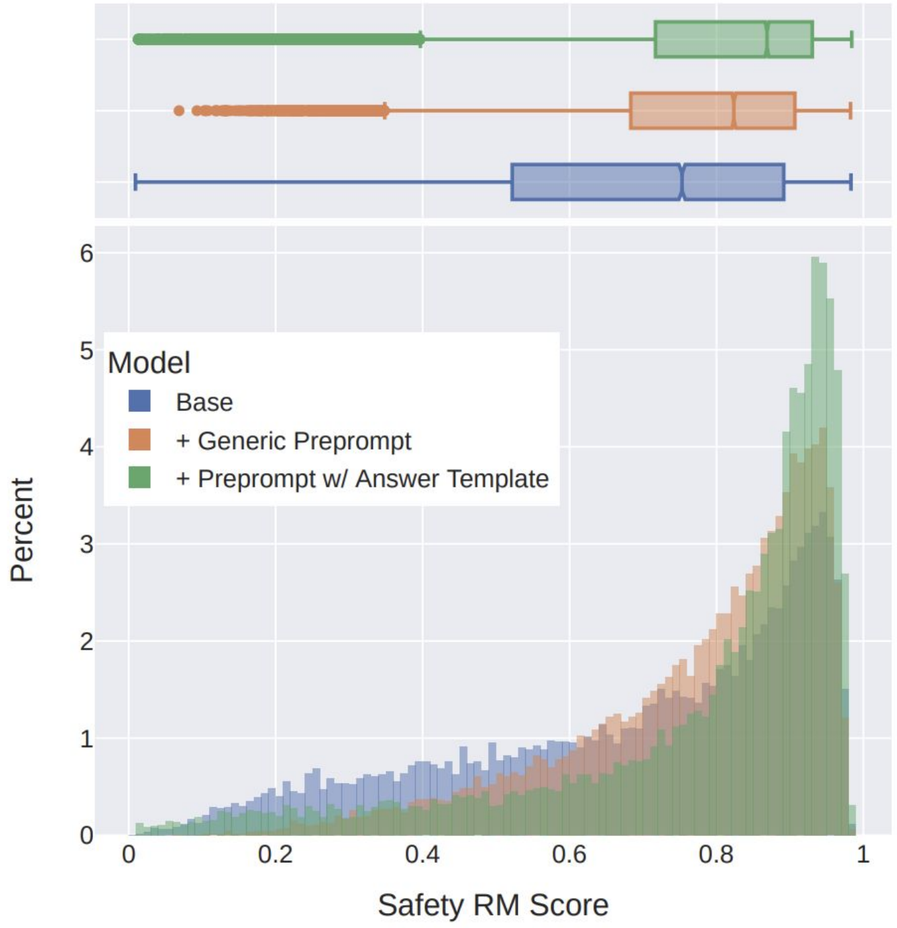
Data Scaling trends for Safety RM



Safety data scaling trends. Left: as we increase the amount of safety data in model training, the mean safety RM score improves significantly while the helpfulness counterpart remains relatively stable. Right: the left tail of safety RM scores (i.e., most unsafe responses) gradually disappears with the addition of more safety training data.

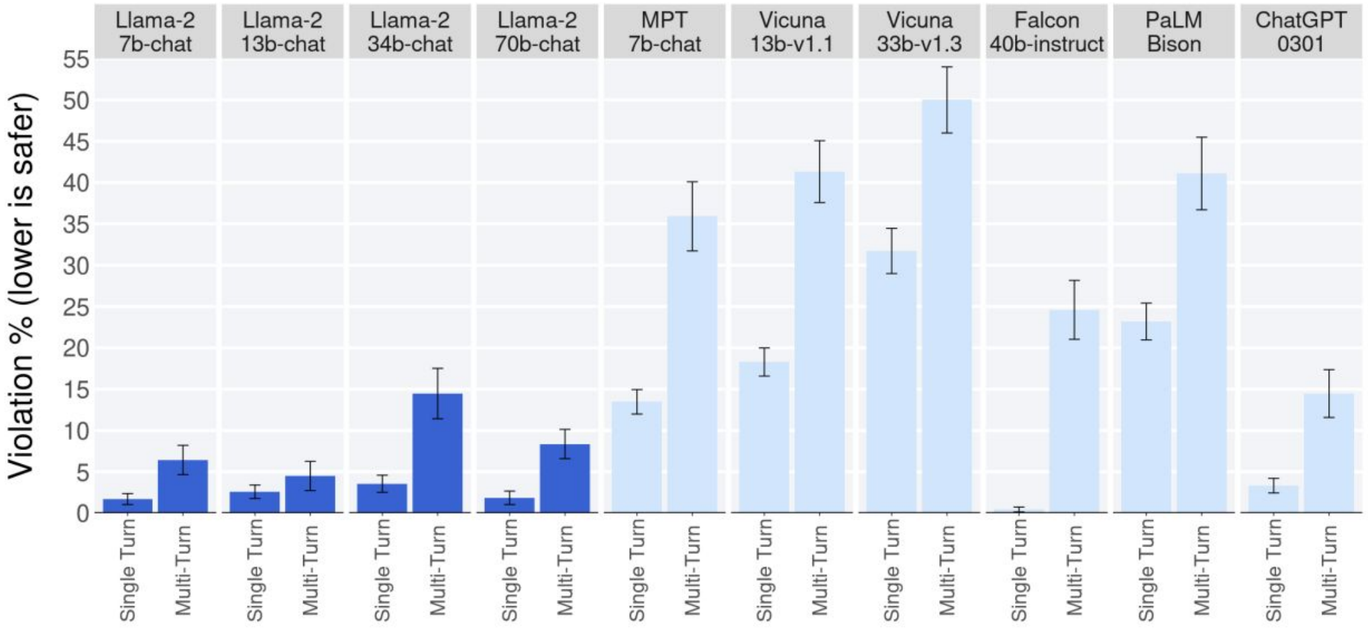
Context Distillation

- Generate safety pre-prompts using various adjectives usually associated with safe behavior such as “responsible,” “respectful’,’ or “wise”.
- Prefix a safety pre-prompt to adversarial prompts to generate safer responses
- Fine-tune the model on its own safe output given the adversarial prompt without the pre-prompt



Distribution of safety RM scores from the base model, when adding a generic preprompt, and when adding a preprompt based on the risk category with tailored answer template. While a generic preprompt increases safety RM scores, a preprompt with tailored answer template helps even more.

Safety Evaluation



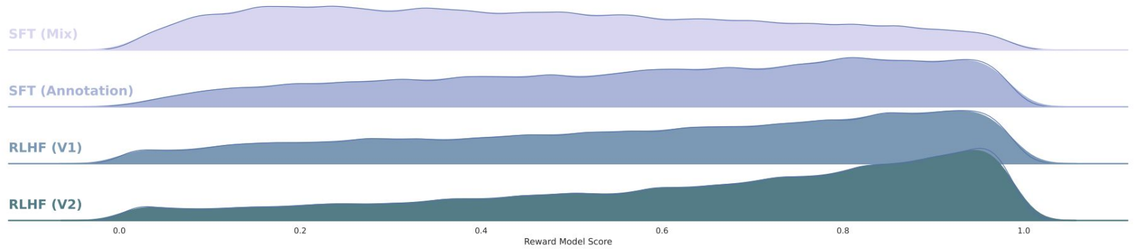
Single-turn and multi-turn violation percentage comparison of Llama-2-chat and various open source and closed source models.

		TruthfulQA ↑	ToxiGen ↓
ChatGPT	-	**78.46**	0.20
Falcon-instruct	7B	28.03	7.89
MPT-instruct	7B	29.99	16.33
Llama 2-Chat	7B	57.04	**0.00**
	13B	62.18	**0.00**
	34B	67.20	0.02
	70B	64.14	0.01

Evaluation of fine-tuned LLMs on different safety datasets

Some Interesting Observations

- RLHF fosters a synergy between humans and LLMs throughout the annotation process
- Dynamic re-scaling of temperature contingent upon the context
  - For creative prompts, increasing temperature results in more diverse generations
  - For factual prompts, model provides same response in spite of rising temperature
- Models showcase robust capability to organize its knowledge in a temporal manner, even with minimal data
- Tool usage can spontaneously emerge from alignment in a zero-shot manner



Date: 01/01/2023
Cutting knowledge: 01/01/1940

Who won the 2nd World War?

I'm not sure what you're referring to, as my knowledge stopped in 1940.

Year: 2023

Is the Earth flat or round?

The Earth is round. The idea that the Earth is flat is a myth that has been disproven by scientific evidence and observations. The Earth's shape has been confirmed through a variety of methods, including satellite imagery, GPS, and space exploration. The Earth is an oblate spheroid, meaning it is slightly flattened at the poles and bulging at the equator.

Year: 852

Is the Earth flat or round?

In 852, the Earth was believed to be flat by many people. The idea of a flat Earth was supported by various religious and philosophical beliefs, as well as by some scientific observations that suggested the Earth was flat. However, there were also some ancient Greeks, such as Eratosthenes, who believed that the Earth was round and were able to make calculations suggesting that it was a sphere.

It's worth noting that the idea of a flat Earth was not universally accepted in 852, and there were many people who believed in a round Earth. Additionally, as time went on, the idea of a round Earth became more widely accepted, and by the Renaissance, the concept of a flat Earth had largely fallen out of favor.

Next Challenges

- Some of the next set of challenges across the various areas in LLM research
  - Getting more data, multilingual, multimodal
  - Scaling to 1000s of GPUs with high MFU
  - Designing Architectures Efficient for training and inference, Hardware-Software Co-Design
  - Continual Learning and Updating Knowledge
  - Improving Factuality and Citing Sources
  - Reducing Hallucinations and Admitting Uncertainty
  - Removing harmful, offensive, or biased content
  - Adapting to world knowledge beyond training data

Thank You!

Twitter : https://twitter.com/vedanujg


Meta AI