

Responsible AI in the age of GenAI: the case of Llama2

Cristian Canton
Head of Responsible AI @ Meta



Agenda

- 01 Responsible AI
- 02 LLaMa 2 - General remarks + Safety
- 03 LLaMa 2 - Red Teaming



Quo vadis, Responsible AI?

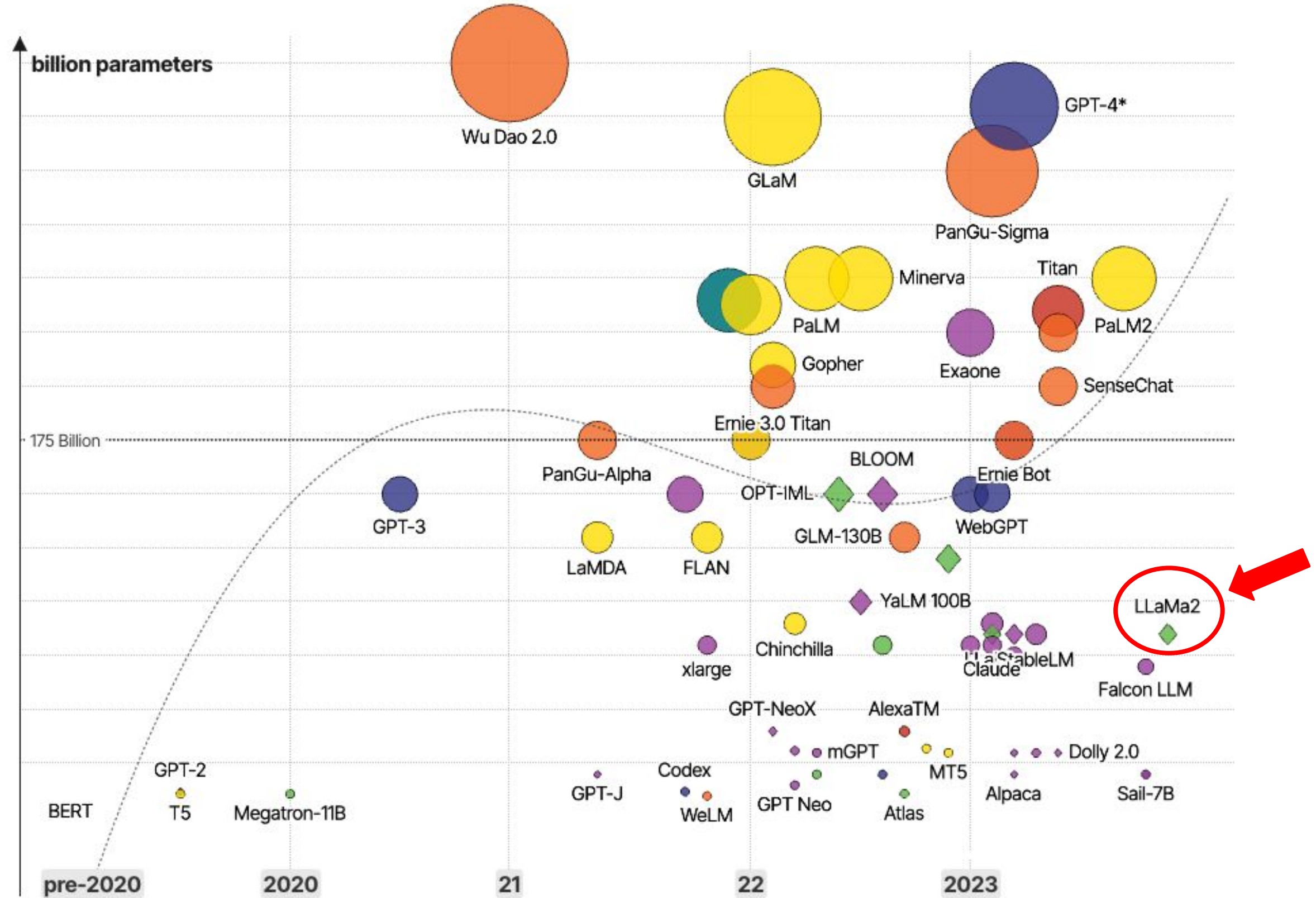
- An emerging discipline in science
- Multiple stakeholders, public and private
- AI's pervasiveness makes Responsible AI necessary
- Need for investment in many fronts: data, research, partnerships, etc.
- What about GenAI?



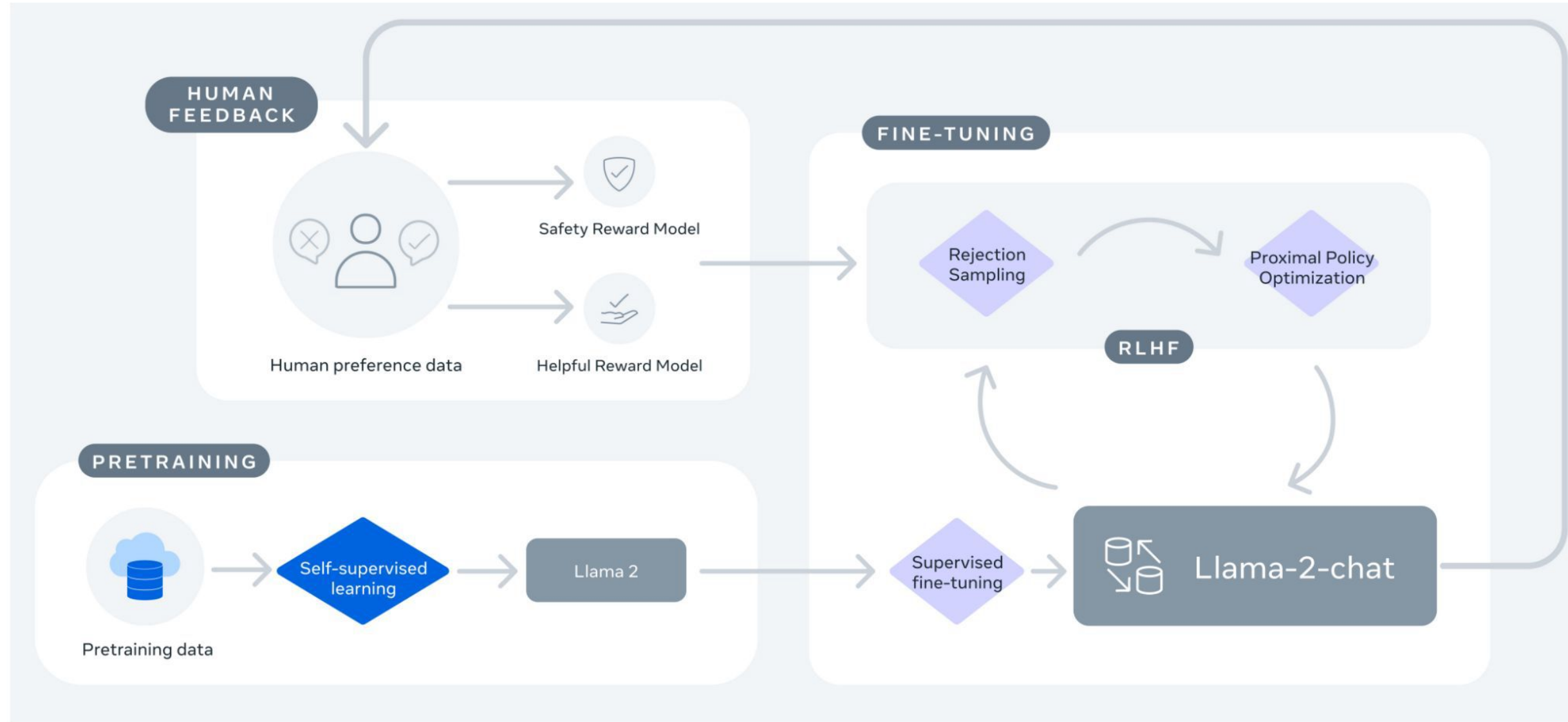
Responsible AI - Elemental pillars

- ✓ Privacy & Security
- ✓ Fairness & Inclusion
- ✓ Robustness & Safety
- ✓ Transparency & Control
- ✓ Accountability & Governance

The Quest of LLMs



How it's made



Pre-Training

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

Finetuning

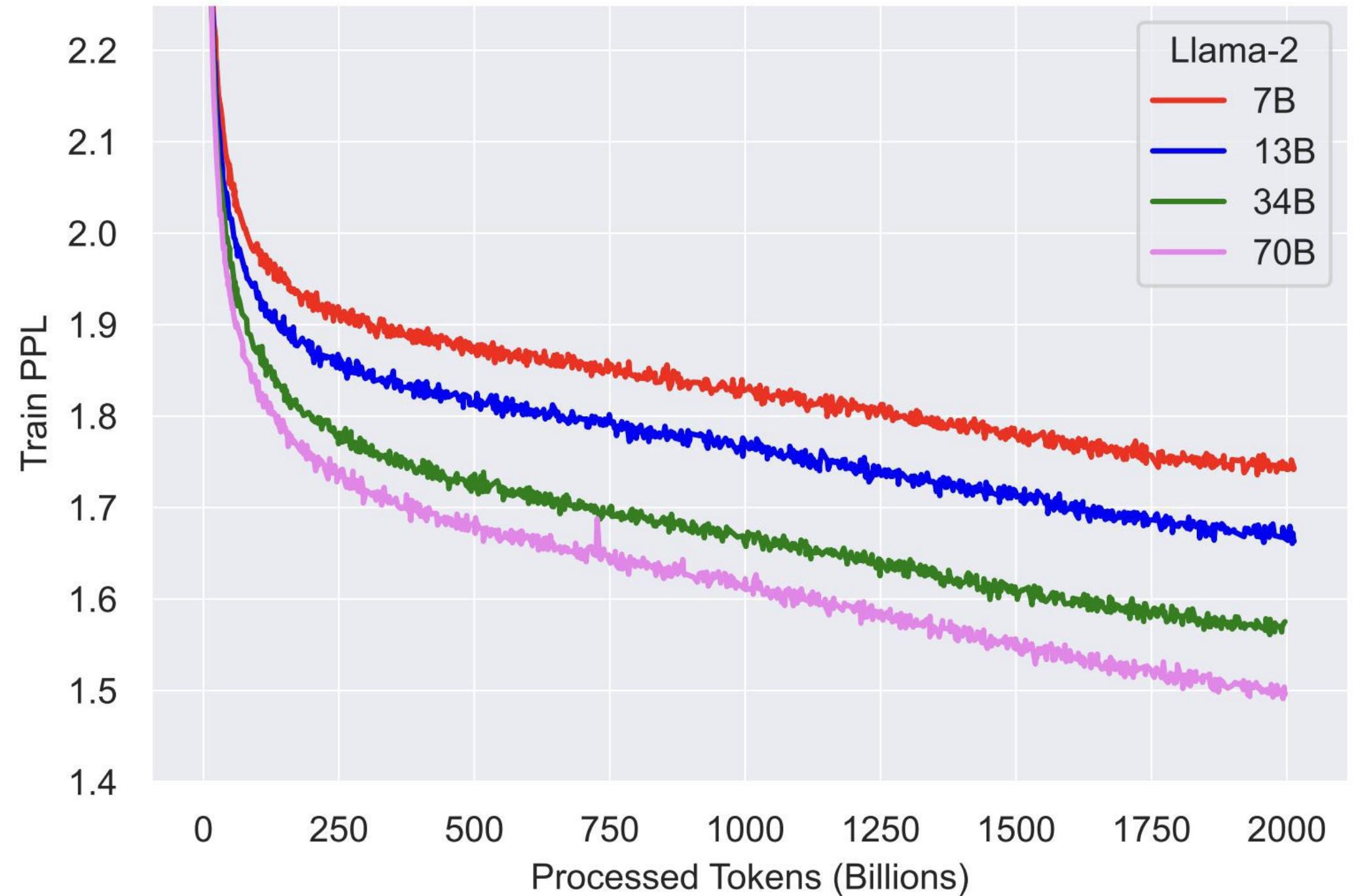
Finetuning and aligning the models to be more like chat assistants, and ensuring they are helpful and harmless.

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.

Pre-Training

- 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



Pre-Training - Model evaluation

OPEN SOURCED

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

CLOSE SOURCED

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

Pre-training - Responsibility + Safety

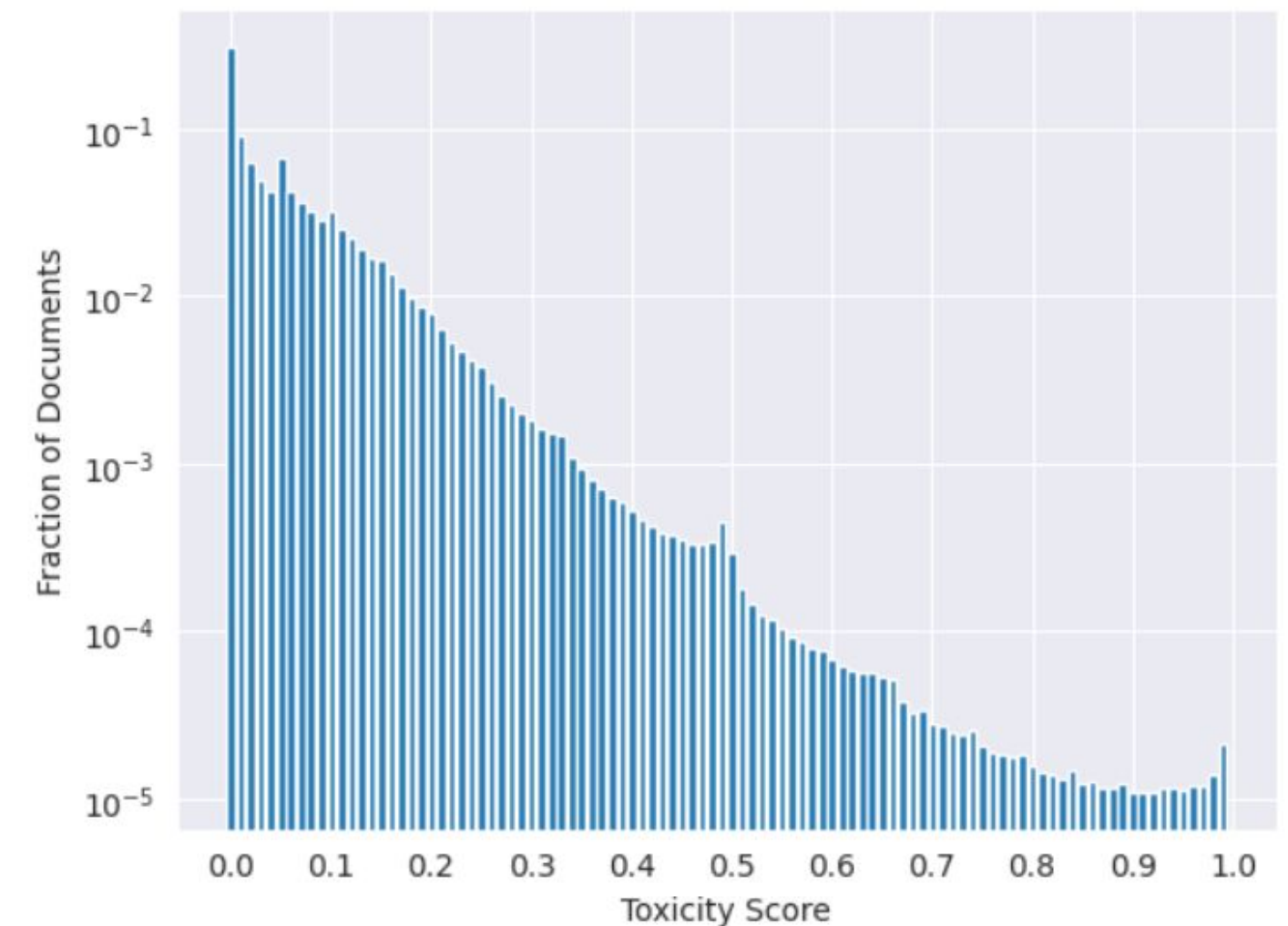
Everything starts with data...

- We did not use any Meta user data in training
- The training corpus includes a mix of data from publicly available sources
- We excluded data from certain sites known to contain a high volume of personal information about private individuals
- No additional filtering was conducted on the datasets, to allow Llama 2 to be more widely usable across tasks (e.g., it can be better used for hate speech classification), while avoiding the potential for the accidental demographic erasure sometimes caused by over-scrubbing.
- Biases on demographic representations?
- Data toxicity?

Pre-training - Responsibility + Safety

Everything starts with data...

- We did not use any Meta user data in training
- The training corpus includes a mix of data from publicly available sources
- We excluded data from certain sites known to contain a high volume of private individuals
- No additional filtering was conducted on the datasets, to allow for use across tasks (e.g., it can be better used for hate speech classification for the accidental demographic erasure sometimes caused by over-filtering)
- Biases on demographic representations?
- Data toxicity?



Pre-training - Carbon Footprint

		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO ₂ eq)
LLAMA 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

- 3.3 Million GPU hours!
- How big are these numbers?
- Meta's open source strategy means that these pretraining costs will not need to be incurred by other companies

Finetuning

- **Supervised Finetuning**

Finetuning

- **Supervised Finetuning**

➤ 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.

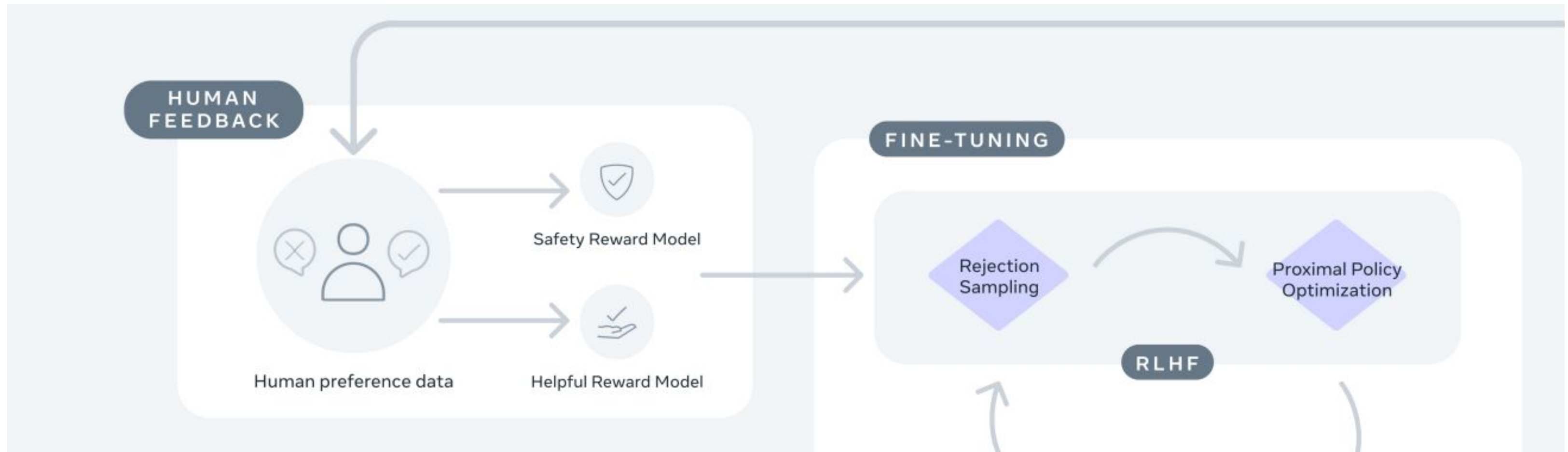
Response: I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

Finetuning

- **Supervised Finetuning → Quality is all you need!**
 - 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!

Finetuning

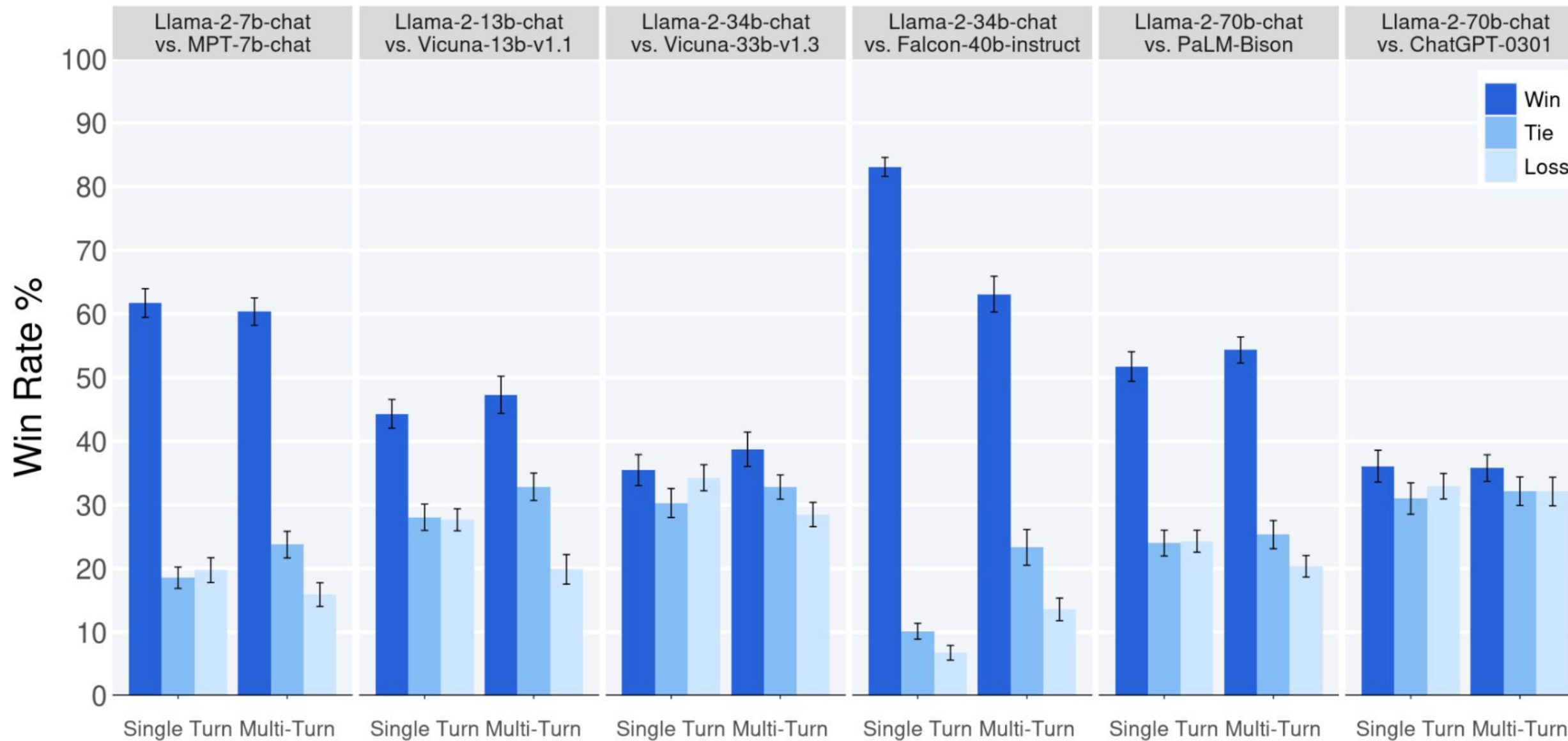
- Supervised Finetuning
- Reinforcement Learning with Human Feedback (RLHF) + Iterative Finetuning



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

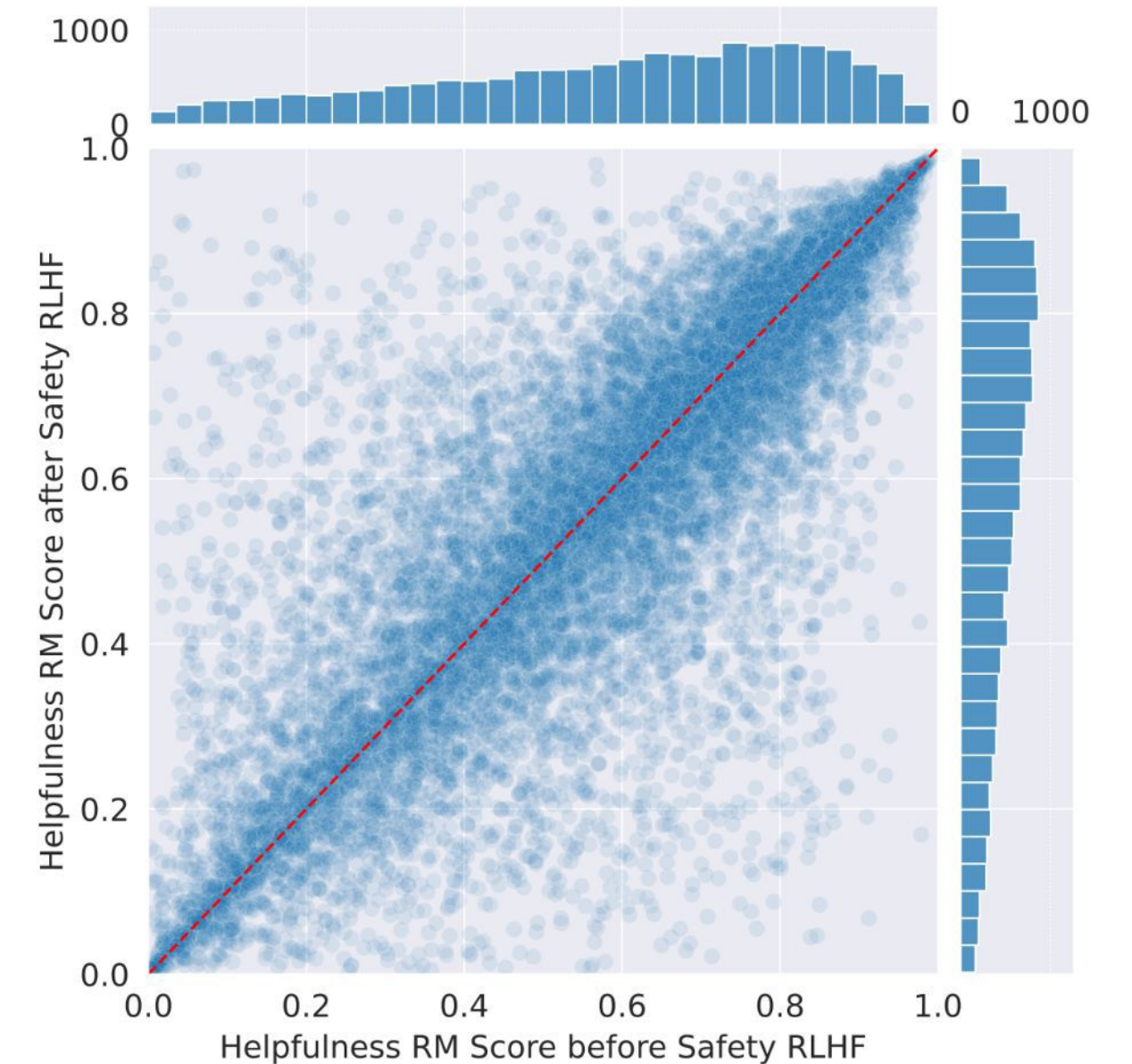
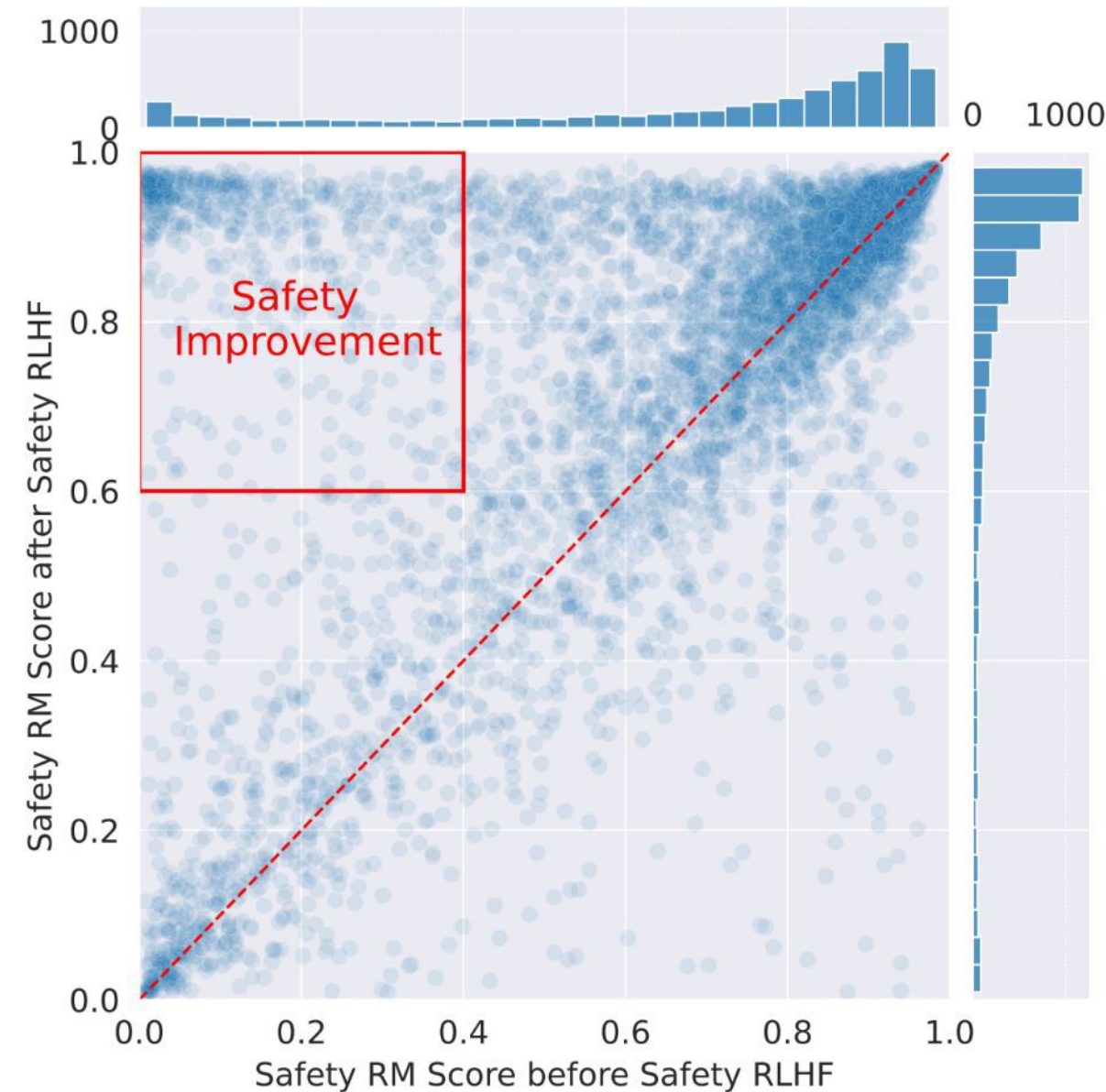
LLAMA2 - Finetuning

- Supervised Finetuning
- Reinforcement Learning with Human Feedback (RLHF)
- Human evaluation results



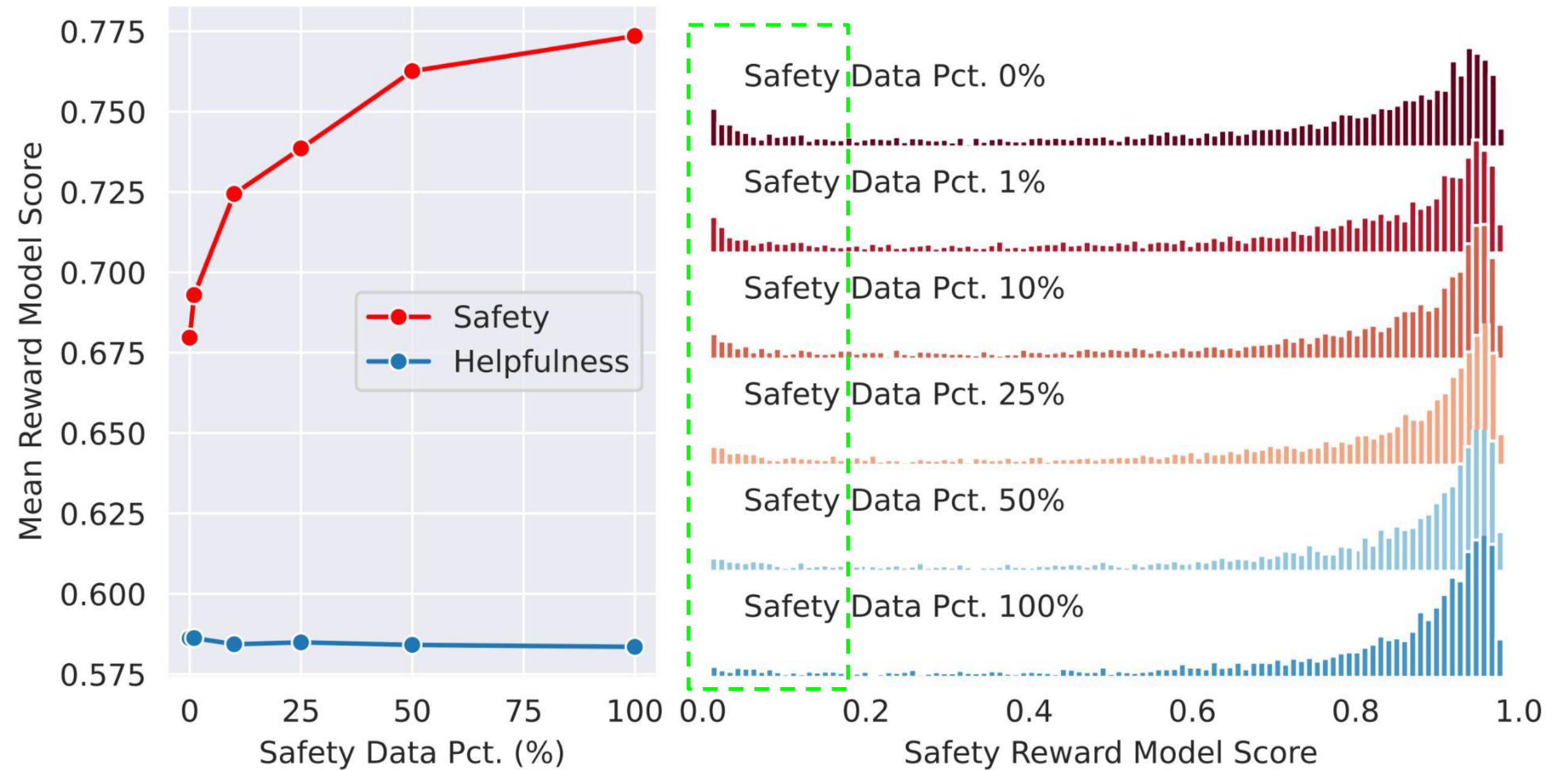
LLAMA2 - Finetuning

- Supervised Finetuning
- Reinforcement Learning with Human Feedback (RLHF)
- Human evaluation results
- Impact on Safety



LLAMA2 - Finetuning

- Supervised Finetuning
- Reinforcement Learning with Human Feedback (RLHF)
- Human evaluation results
- Impact on Safety



Red Teaming

- Given how broad the capabilities of LLMs are and how varied their training data is, it is insufficient to identify risks solely via *ex post facto* usage and analysis. Benchmarks do not tell the whole story!
- Red teaming allows you to explore the long tails
- But, what is **red teaming**?
- Risk taxonomy and evaluation protocol
- Evaluating red teaming effectiveness → Model improvements

Any questions?