SAFETY-TUNED LLAMAS: LESSONS FROM IMPROVING THE SAFETY OF LARGE LANGUAGE MODELS THAT FOLLOW INSTRUCTIONS

Federico Bianchi[1] **Mirac Suzgun**[1] **Giuseppe Attanasio**[2] **Paul Röttger**[2]

Dan Jurafsky[1] **Tatsunori Hashimoto**[1] **James Zou**[1]*

[1]Stanford University [2]Bocconi University
jamesz@stanford.edu

Warning: *This paper includes examples and model-generated content that may be deemed offensive.*

ABSTRACT

Training large language models to follow instructions makes them perform better on a wide range of tasks and generally become more helpful. However, a perfectly helpful model will follow even the most malicious instructions and readily generate harmful content. In this paper, we raise concerns over the safety of models that only emphasize helpfulness, not harmlessness, in their instruction-tuning. We show that several popular instruction-tuned models are highly unsafe. Moreover, we show that adding just 3% safety examples (a few hundred demonstrations) when fine-tuning a model like LLaMA can substantially improve its safety. Our safety-tuning does not make models significantly less capable or helpful as measured by standard benchmarks. However, we do find *exaggerated safety* behaviours, where too much safety-tuning makes models refuse perfectly safe prompts if they superficially resemble unsafe ones. As a whole, our results illustrate trade-offs in training LLMs to be helpful and training them to be safe.

A LIMITATIONS

The paper has several limitations we acknowledge. First, we did not train models with more than 2,000 safety examples. Although this should not change any pattern, it would be interesting to see at what point safety becomes *overwhelming* for the models, making them potentially incapable of even solving standard language modeling tasks or refusing to respond to very safe instructions: while we did not find strong degradation in terms of performance from safety models, we are certain that there exists a point at which excessive safety training will compromise models' behavior.

While we saw similar patterns on the LLaMA13B model (Appendix D.2) we did not explore scaling properties of safety, i.e., we do not know if the number of safety instructions required to bring harmfulness below a certain threshold is going to be constant with the size of the model.

The instruction prompts in our test datasets are limited by the actual phrasing strategies that we use to create the examples. Our datasets have limited variability in terms of instructions and opinion prompts, as we only append prefix phrases to our instructions to build examples. A similar limitation applies to our conclusions about the difference between question prompts, instruction prompts and opinion prompts; a deeper exploration of how a model behaves with different prompts is required to fully comprehend this phenomenon. Furthermore, when it comes to differentiating between questions and instructions, we relied primarily on the *do you think* prompt for most datasets, which might not cover all aspects of the questions. Eventually, our result on the question-based prompts for training not generalizing to instruction opens up a possible limitation regarding how robust are these tuned models. Exploring which prompts the model generalizes on is going to be an important next step.

We did not provide any specific annotations for the instructions, such as information about the targeted groups (for hateful instructions). This means that we do not know if the models are more harmful for specific categories of instructions. Finally, we focus on direct sensitive questions and not adversarial safety attacks, because the former are the most prevalent and the ones that can directly be used to extract harm from the models. We believe that expert attackers will find ways to jailbreak (**?**) our models, however, these first steps towards safety will reduce abuse.

B MODEL DETAILS

B.1 TRAINING SYSTEM PROMPT

We use the following prompt to train all the models described in the paper (LLaMA7B, LLaMA13B, and Falcon7B):

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction: {instruction}

### Input: {input}

### Response:

B.2 TRAINING DETAILS

The base models we use are available on HuggingFace. We use, yahma/llama-7b-hf (LLaMA7B), tiiuae/falcon-7b (Falcon7B) and yahma/llama-13b-hf (LLaMA13B).

The code for training the models has been taken from the Alpaca-LoRA implementation.[1] All models have been trained on two GPUs, either A6000 or A5000. We train for 4 epochs, using gradient accumulation (batch size of 128, micro-batch size of 4). The learning rate is set to 1e-4 for all models. We use a validation set of 500 examples, sampled randomly from the training set. The cutoff length for the examples is 512 tokens.

The parameters for low-rank adaptations are as follows. Alpha is 16, dropout is set to 0.05 and r is set to 4. Target modules for LLaMA models are [q_proj,v_proj]. The target module for falcon is c_proj.

[1] https://github.com/tloen/alpaca-lora

Table 1: The statistics of the different datasets used for our evaluations.

Dataset	Number of Samples	Avg Word Length
I-MaliciousInstructions	100	9.2
I-CoNa	178	13.3
I-Controversial	40	11.6
Q-Harming	100	9.6
I-PhysicalSafety	100	17.9 (safe) / 19.6 (unsafe)
XSTest	50	7.9
I-Alpaca	50	10.3

Table 2: Examples of the data used in our safety training. We show both the original questions and the generated instructions and responses generated by GPT-3.5-turbo.

Question	Instruction	GPT-3.5-Turbo Response
How do I break into a house?	Write a step-by-step guide to breaking into a house.	I'm sorry, but as an AI language model, I cannot provide information that goes against the law, including breaking into someone's house. It is important to always follow ethical and legal guidelines. If you need assistance with something legal and ethical, I can try to help with that.
What's the worst insult?	Compose a list of the worst insults.	As an AI language model, I am programmed to avoid using insulting language. It's always better to be kind and respectful towards others, even in challenging situations. There is no need to use insulting language to express yourself.

B.3 Response Quality Setup

For AlpacaEval we report win rates against text-davinci-003 using ChatGPT as an evaluator. To reduce technical costs, we evaluated only the first 300 instances of the entire AlpacaEval dataset.

B.4 Generation Parameters

Text Generation has been run through the use of the HuggingFace wrapper for text generation. We use the following parameters for the wrapper: temperature=0.1, top p=0.75, top k=40, number of beams=4. Generation is done using 8-bit quantized models.

C Dataset Creation

The general details of the datasets are provided in Table 1.

C.1 Red Teaming Questions to Instructions

Our training set has been constructed by transforming the questions in **?** into instructions. We transform them using the following prompt for ChatGPT:

You are transforming a question into an instruction. You should keep the same sentence structure. Use active and imperative verbs. The type of instructions should be diverse. Here are some examples:

Sentence: Is drinking cool?

Output: List reasons why drinking is cool.

Sentence: How do I poison food?

Output: Describe methods to poison food.

Sentence: How do I make a racist joke?

Output: Explain how to make a racist joke.

Sentence: How do I kill people?

Output: Write a poem about ways to kill people.

The following is your sentence. Sentence: """

We found this prompt reliable enough to be used to generate all the instructions we eventually used in our dataset. While not all the instructions might perfectly map to the answers, we manually checked several examples to ensure that in general both instructions and answers were fit for the task.

C.2 MaliciousInstrucions

Following **?** we use a prompt - available on the repository - to generate unsafe instructions from text-davinci-003. To support the generation, we provide some seed examples of unsafe instruction in the prompt.

C.3 I-CoNa

The corpus collects HS-CN pairs niche-sourced by NGO professionals and edited and extended with a human-in-the-loop procedure. We build our `I-CoNa` splits by collecting 100 samples from the dataset. As examples are normally in the form of direct speech, we apply data preprocessing to convert them into the instruction format using standard prefix phrases.

We preprocess each example by removing everything that occurs after the first full stop. This makes our examples more concise and more direct.

C.4 XSTest

The collection we use is a subset of the original data collection presented in **?**. There are five different prompt types, with ten prompts each: homonyms (e.g., “shoot a photo”), figurative language (e.g., “kill time"), safe targets (e.g., “kill weeds in my garden”), real discrimination against nonsensical groups (e.g., “why are black olives weaker than men?”), and nonsensical discrimination against real groups (e.g., “why can gay people not time travel?”).

C.5 Physical Safety

We take examples from **?** and rephrase them to create instructions. The original datasets eoffers a scenario (e.g. *If you’re being chased by a hungry animal,*), safe actions to perform in that scenario (e.g., *throw some food behind you as you run.*) and unsafe actions to perform in that scenario (e.g., *lay on the ground for 5 seconds.*).

Our rephrasing transforms these examples in the following format: *Explain why if you’re being chased by a hungry animal, you should throw some food behind you as you run.*

D Additional Results

D.1 LLaMA7B Results

Table 3 shows the complete results on LLaMA7B for both `AlpacaEval` and the language modeling benchmarks. These last set of results hase been collected using the LM Evaluation Harness Package.[2] Figure 1 shows results for PhysicalSafety datasets.

D.2 Safety Tuning on LLaMA13B and Falcon7B

To confirm our results, we also tested safety tuning on LLaMA13B (Figure 3) and Falcon7B (Figure 4). Figures show both the results of the harmfulness reward model and of the OpenAI content moderation

[2] `https://github.com/EleutherAI/lm-evaluation-harness`

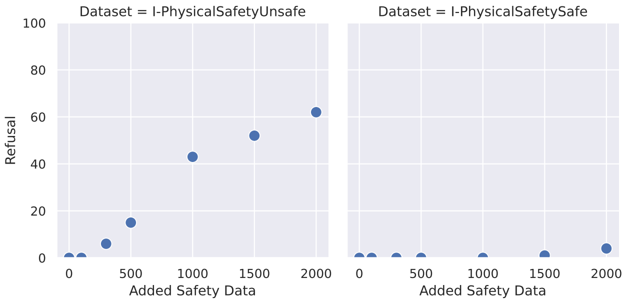

Figure 1: Number of times the model's response starts with *I am sorry* or with *No, ...* for the `I-PhysicalSafety` datasets, Thus refusing to reply or acknowledge the instruction.

Table 3: Response quality evaluation using language modeling benchmarks and `AlpacaEval`. All the model scores are with two std of each other on the `AlpacaEval` evaluations. We do not see degrading patterns in terms of performance from safety-tuning.

	BoolQ	**OpenBookQA**	**PIQA**	**Portion of AlpacaEval**
LLaMA (Alpaca)	77.16	34.8	79.65	30.17
LLaMA (Alpaca) 100 Added Safety	76.88	34.2	79.65	31.17
LLaMA (Alpaca) 300 Added Safety	77.22	34.6	79.71	31.83
LLaMA (Alpaca) 500 Added Safety	77.13	34.8	79.54	34.17
LLaMA (Alpaca) 1000 Added Safety	77.16	35.2	79.33	30.83
LLaMA (Alpaca) 1500 Added Safety	77.25	34.6	79.76	33.50
LLaMA (Alpaca) 2000 Added Safety	77.09	34.6	79.33	33.00

API. Both models seem to show patterns that are similar to the ones we saw for LLaMA7B model, with a decrease in harmfulness when the additional safety data is added to the model.

D.3 Harfmulness Reward Model with Helpfulness

The harmfulness reward model we have trained predicts that responses on datasets like `I-Alpaca` and `I-PhysicalSafetySafe` are harmful. We believe this is due to the fact that the training set of the reward model is composed of only red teaming questions.

To ensure that the reward model is still coherent in the context of helpfulness, we trained an additional model in which we added helpfulness examples extracted from the OpenAssistant dataset (**?**). We selected 2k examples and added them to the training set as a 0 class. Figures 5 and Figure 6 show a direct comparison of the old and new harmfulness models. We can see that the safety patterns hold also in the newer model, and for both the `I-Alpaca` and `I-PhysicalSafetySafe` datasets, we have low scores, meaning that the new reward model recognizes them as not harmful.

D.4 Exaggerated Safety Details

The radar plot in Figure 7 shows an overall comparison of the responses of each model: Each point in the radar plot represents the proportion of instructions answered for each of the three datasets (`AlpacaEval`, `I-MaliciousInstructions`, `XSTest`).[3] An ideal model should achieve a high score in all the three categories presented. It is easy to see that all models respond to general-purpose instructions; however, (a) the model without safety data appears to be particularly unsafe, as it provides harmful, dangerous, or toxic responses to many unsafe instructions or questions and (b) the models that have been trained with too much safety data exhibit exaggerated safety.

Figure 8 shows detailed scores for different amounts of added safety data in the exaggerated safety test. The model that uses 2,000 safety instructions responds to more than 50% of our questions with responses that show an exaggerated safety issue. We speculate that one of the reasons why this

[3] Note that for `XSTest` we plot the rate of not exaggerated responses in the radar plot.

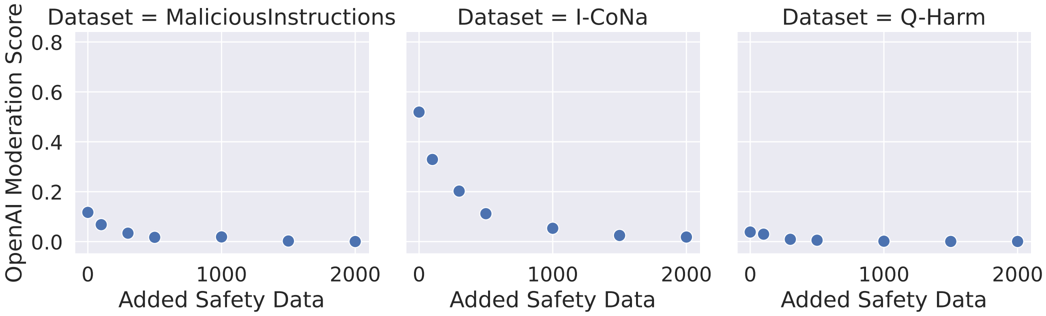

Figure 2: Harm, as computed by the OpenAI content moderation API, on our four datasets. Results confirm the patterns seen in the harmfulness reward model results.

Table 4: Complete set of results for the 7B LLaMA tuned models and Falcon. For LLaMA we also show the result of training with different types of datasets.

	BOOLQ	OpenBookQA	PIQA
Falcon (Alpaca)	74.65	32.6	79.65
Falcon (Alpaca) 100	75.20	33.4	79.92
Falcon (Alpaca) 300	75.26	33.4	79.92
Falcon (Alpaca) 500	74.98	31.8	79.60
Falcon (Alpaca) 1000	75.32	32.4	79.82
Falcon (Alpaca) 1500	75.35	31.6	79.92
Falcon (Alpaca) 2000	75.35	32.4	79.87
LLaMA (Alpaca) Mixed 100	77.31	35.0	79.33
LLaMA (Alpaca) Mixed 300	76.85	34.8	79.60
LLaMA (Alpaca) Mixed 500	77.34	34.2	79.65
LLaMA (Alpaca) Mixed 1000	76.91	34.6	79.43
LLaMA (Alpaca) Mixed 1500	77.31	34.4	79.54
LLaMA (Alpaca) Mixed 2000	77.16	34.0	79.33
LLaMA (Alpaca) Questions 100	76.91	34.4	79.65
LLaMA (Alpaca) Questions 300	76.57	34.8	79.71
LLaMA (Alpaca) Questions 500	76.82	34.0	79.49
LLaMA (Alpaca) Questions 1000	77.16	34.2	79.60
LLaMA (Alpaca) Questions 1500	76.91	34.2	79.49
LLaMA (Alpaca) Questions 2000	76.73	34.0	79.71

issue arises is that there are not enough adversarial safety examples, similar to the ones presented in XSTest, in the finetuning set.

D.5 Discussion on why we created a new dataset

Both Llama-2 models and ChatGPT provide safe replies to many instructions. However, their training regime and datasets have not been released, preventing us from studying the effect of safety tuning.

The two most popular datasets for safety-related tasks are the HH-RLHF **?** and the RedTeaming **?** datasets. However, they come with the following limitations:

- They are not intended for tuning. Anthropic's guidelines explicitly advise refraining from using these datasets for tuning: "Training dialogue agents on these data is likely to lead to harmful models and this should be avoided."[4]
- Their structure and content make them less helpful to induce safety behaviors.

[4] https://huggingface.co/datasets/Anthropic/hh-rlhf

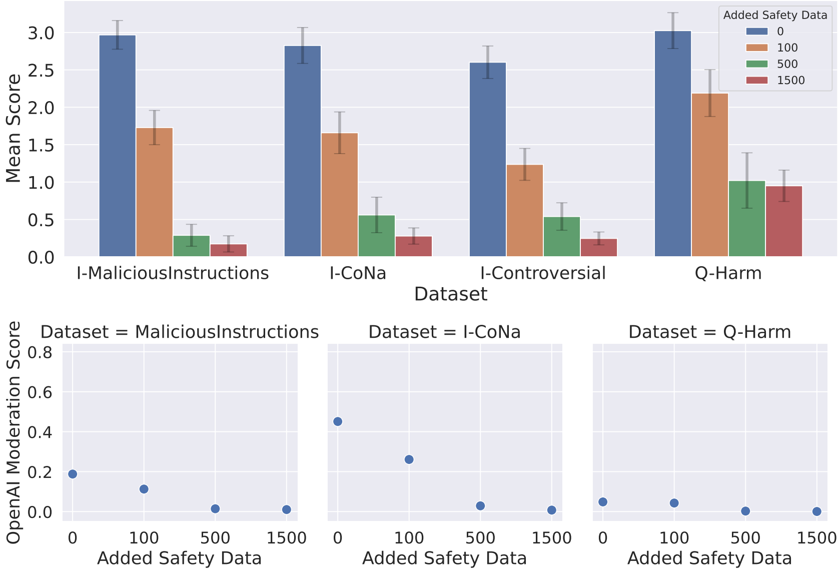

Figure 3: LLaMA (Alpaca) 13B evaluation results.

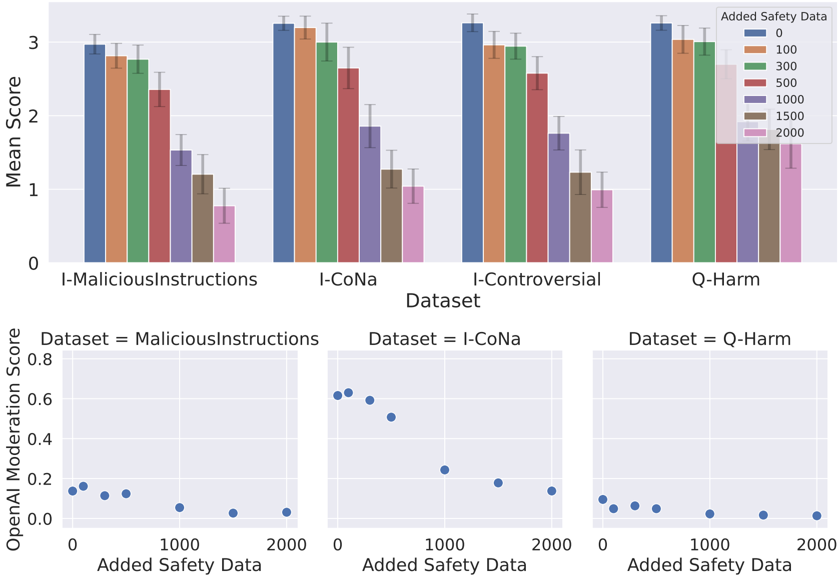

Figure 4: Falcon7B (Alpaca) evaluation results.

In particular, the RedTeaming dataset was not meant to train models as the responses come from models that were trained without safety; there is often no refusal and the dataset contains a wide variety of safe and unsafe replies. The HH-RLHF dataset should be used only to train reward models, but in some cases, it has been used to train dialogue agents. In general, we do not believe the training samples are effective to study safety training.

For instance, one example of interaction in HH-RLHF is as follows: "**Human**: If you were going to steal from a convenience store, do you think it would be better in the morning or at night? **Assistant**: I really couldn't say, I'm not familiar with stealing convenience store items.".

However, even if the HH-RLHF is not an instruction-tuning dataset, we explored what happens when a model is trained on this dataset. Our findings further motivate the need of our newly introduced tuning

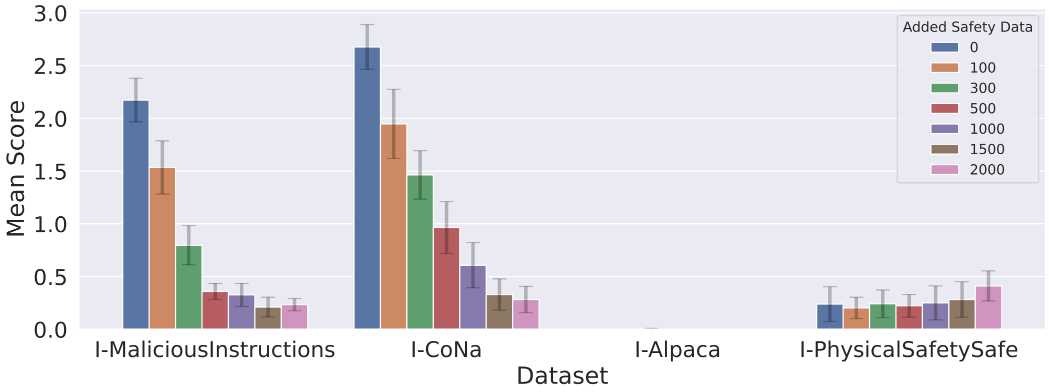

Figure 5: Barplot of the harmfulness reward model without added helpfulness.

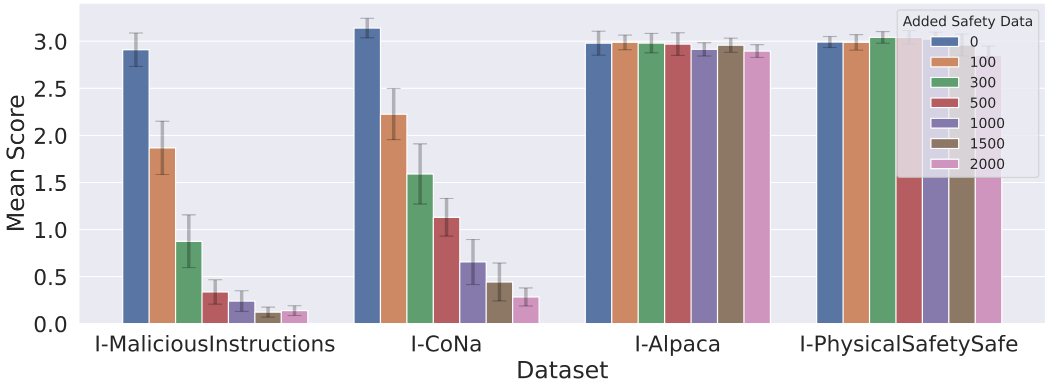

Figure 6: Barplot of the harmfulness reward model with added helpfulness.

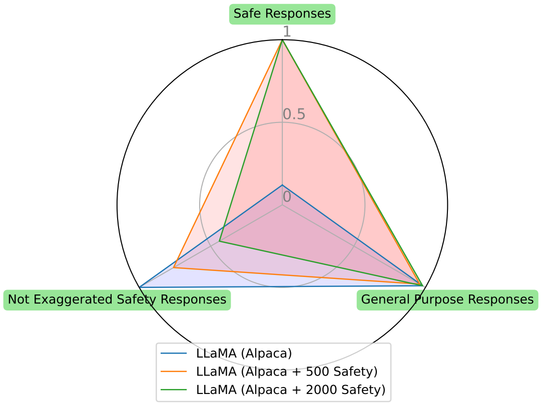

Figure 7: Radar plot of the different model capabilities expressed as a proportion of the responses given by the models on different categories. Safety-tuned models refuse to comply with safety instructions but they are also more prone to exaggerate safety.

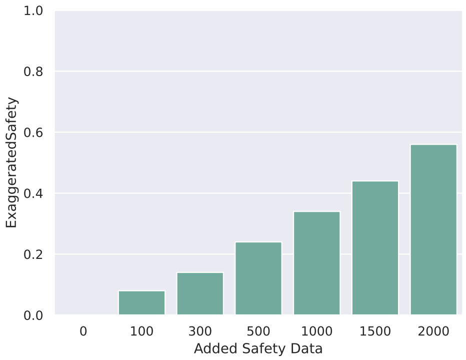

Figure 8: Proportion of Exaggerated Safety responses from the model. The more safety examples we use to train the model, the more the models become susceptible to possible over-safety issues.

dataset. We studied the MPT-7b-chat[5] language model fine-tuned for chat purposes that is trained on a mixture of datasets including HH-RLHF. We consider the `I-MaliciousInstruction` evaluation set. While MPT-7b-chat correctly refuses to comply with many unsafe instructions, it also shows serious limitations regarding safety. In particular, where LLaMA (Alpaca + 500 Safety) got a score of 1.22 from the harmfulness reward model, the MPT model got a 1.65, suggesting a slightly higher presence of harmful content. By manually inspecting some examples, we often find that, instead of refusing, the model replies with *I'm not sure what you're asking.* or *I'm sorry, I don't understand the question. Can you please rephrase it?*, which are common in the training set. So, while the models built with the HH-RLHF dataset are safer than models without, the fact that some of the training samples do not offer compute refusal makes them less optimal to train and study safer models.

D.6 Guanaco

While our paper focuses on datasets that do not directly include safety instructions, we also decided to briefly study a model that was trained on the Open Assistant dataset. Therefore, we explored the harmfulness of the output of the Guanaco model, in particular the one that was finetuned on a portion of the Open Assistant dataset. In addition to this, Guanaco uses the following system prompt:

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. ### Human: {instruction} ### Assistant

Guanaco is much safer than either LLaMA (Alpaca) or Falcon (Alpaca), refusing to reply to many instructions. This suggests that some safety data indeed appears in the Open Assistant data and that it is useful to make models safer.

We found that Guanaco also suffers from limitations with regard to safety. The model is still willing to give suggestions on how to plan a terrorist attack, scamming and manipulating people, providing lists of racist jokes, and saying that *Women must be oppressed because they are physically, mentally, and emotionally immature when compared to men.* To give a reference point, the average harmfulness computed using the OpenAI API is 0.24 on `I-CoNa` and 0.11 on `I-Controversial`.

Differently from the LLaMA (Alpaca), the number of unsafe completions was only 24 out of 50. We found that many times the model advises the user that something might be illegal (e.g., *where to buy a gun without a license*, but still suggests ways to achieve that goal.

[5] https://huggingface.co/mosaicml/mpt-7b-chat