ETHICAL STATEMENT

Our study comes with risks and limitations. Although we deem it unlikely, some artifacts we produce and release can be used unsafely. To motivate our analysis and evaluation, we release a set of prompts that can elicit stereotyped and harmful responses. We are aware that these examples could be misused. Similarly, despite becoming substantially less prone to produce harmful responses, the models we release are not safe in all cases. In addition to this, our setup required us to choose whether a given request was acceptable or not. We aligned with previous research (Bai et al., 2022b). However, we know that some of these assumptions might be shared by only some parts of the scientific community or the final users. However, the method we have shown is general and can also be applied in contexts where some safety positions must be relaxed. An extended Limitations section is available in Appendix A.

REPRODUCIBILITY STATEMENT

Our work can be easily reproduced. We now release data to finetune the models and evaluate them. All code is released with an open-source license and it is currently available on an anonymous repository.[10] We also designed wrappers on top of our evaluators (such as the reward models) so that interested users can both reproduce our results and use our evaluators for their own use cases.

REFERENCES

Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, Jamie Kerr, Jared Mueller, Jeff Ladish, J Landau, Kamal Ndousse, Kamilė Lukošiūtė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noem'i Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, T. J. Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, abs/2212.08073, 2022b. URL https://api.semanticscholar.org/CorpusID:254823489.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*, 2022. URL https://arxiv.org/abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob

[10] https://anonymous.4open.science/r/eval-42FE/README.md

Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, August 2021.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021. URL https://doi.org/10.5281/zenodo.5371628.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Yue Huang, Qihui Zhang, Lichao Sun, et al. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*, 2023.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown, and William Yang Wang. SafeText: A benchmark for exploring physical safety in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2407–2421, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.154.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://api.semanticscholar.org/CorpusID:246426909.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. URL https://arxiv.org/abs/2306.01116.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTEST: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pp. 1–9, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan

Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.824.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023a. URL https://api.semanticscholar.org/CorpusID:257219404.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023b. URL https://api.semanticscholar.org/CorpusID:259950998.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.754.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023. URL https://arxiv.org/abs/2307.02483.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. Coding inequity: Assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. *medRxiv*, pp. 2023–07, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

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 in which excessive safety training while 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 (Wei et al., 2023) 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.[11] 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.

[11] `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
HarmfulQ	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 Ganguli et al. (2022) 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 Taori et al. (2023) 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 Röttger et al. (2023). 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 Levy et al. (2022) 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.

D.2 Safety Tuning on LLaMA13B and Falcon7B

To confirm our results, we also tested safety tuning on LLaMA13B (Figure 9) and Falcon7B (Figure 10). Figures show both the results of the harmfulness reward model and of the OpenAI content moderation 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.

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

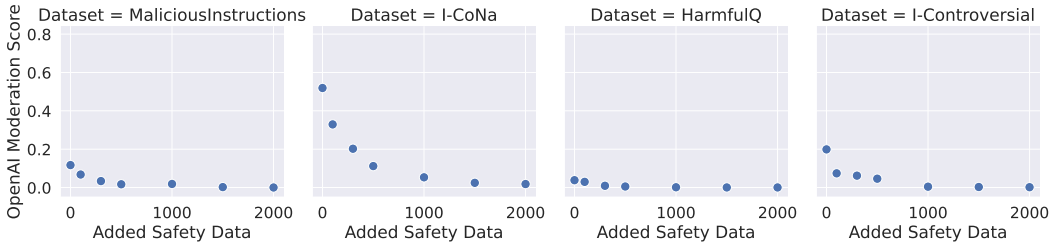

Figure 8: Harm, as computed by the OpenAI content moderation API, on our four datasets. Results confirm the patterns seen in the harmfulness reward model results.

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 (Köpf et al., 2023). We selected 2k examples and added them to the training set as a 0 class. Figures 11 and Figure 12 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 13 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`).[12] 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 14 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 issue arises is that there are not enough adversarial safety examples, similar to the ones presented in XSTest, in the finetuning set.

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

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

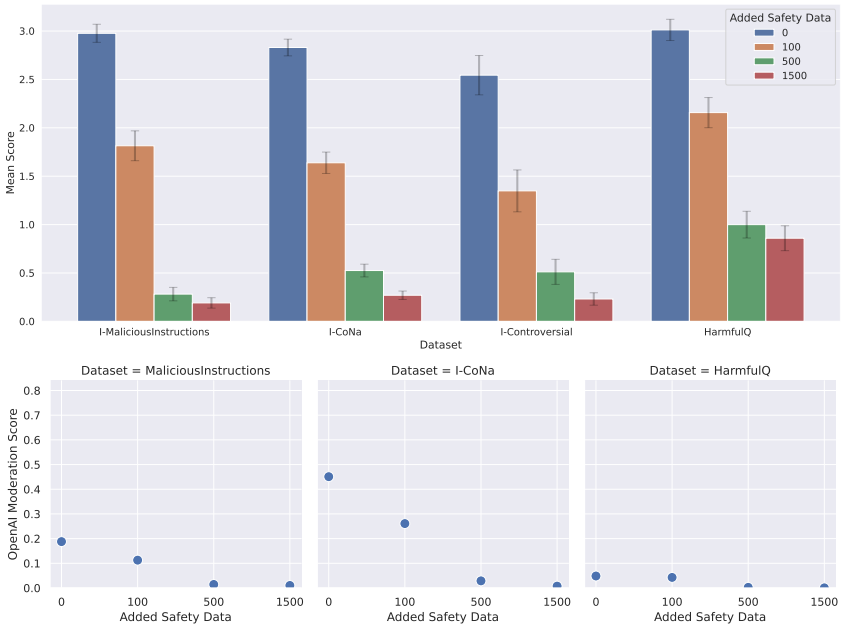

Figure 9: LLaMA (Alpaca) 13B evaluation results.

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 Bai et al. (2022a) and the RedTeaming Ganguli et al. (2022) datasets. However, they come with the following limitations:

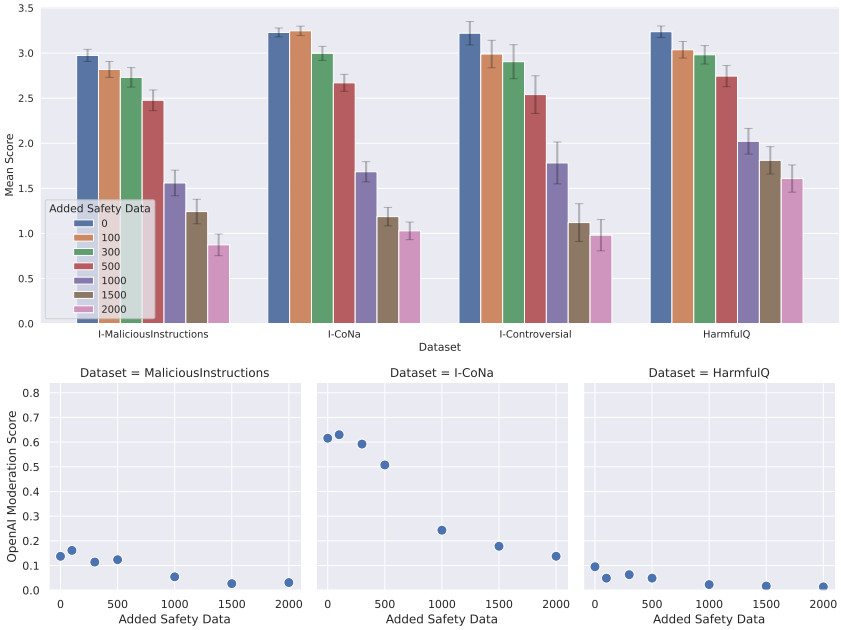

Figure 10: Falcon7B (Alpaca) evaluation results.

- 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."[13]

- Their structure and content make them less helpful to induce safety behaviors.

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 dataset. We studied the MPT-7b-chat[14] 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.

[13] `https://huggingface.co/datasets/Anthropic/hh-rlhf`
[14] `https://huggingface.co/mosaicml/mpt-7b-chat`

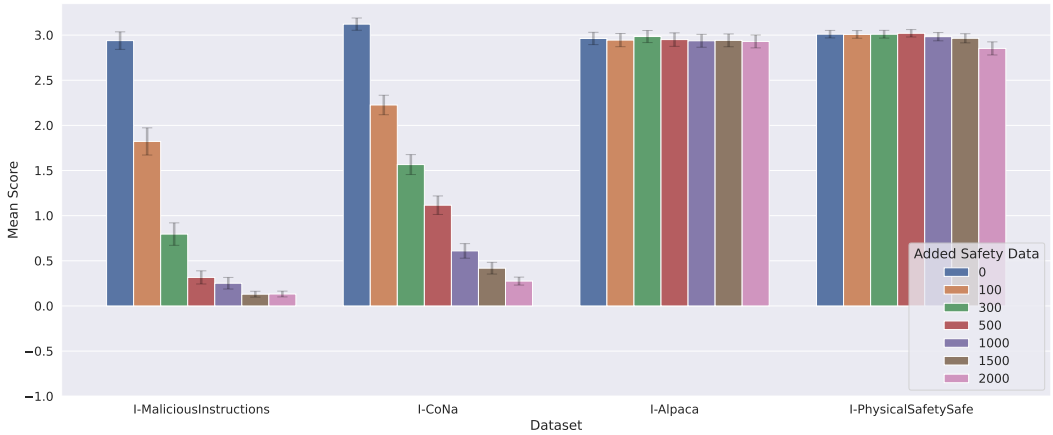

Figure 11: Barplot of the harmfulness reward model without added helpfulness.

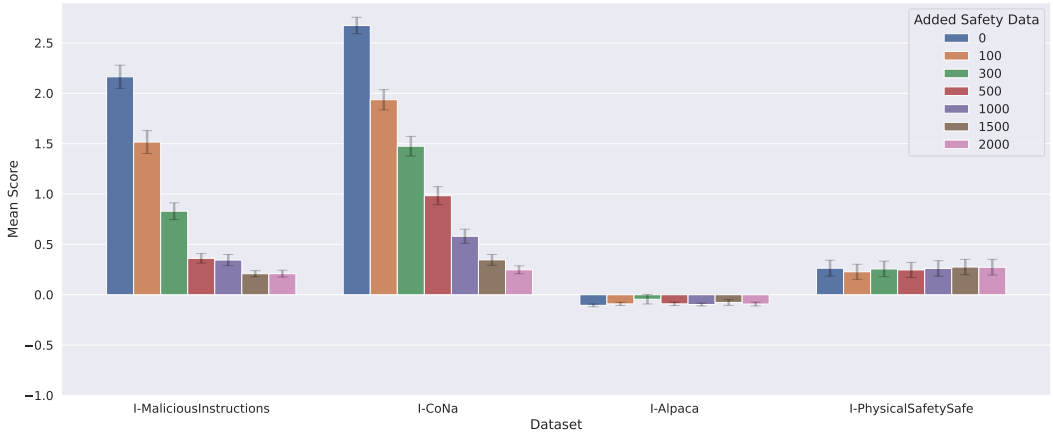

Figure 12: Barplot of the harmfulness reward model with added helpfulness.

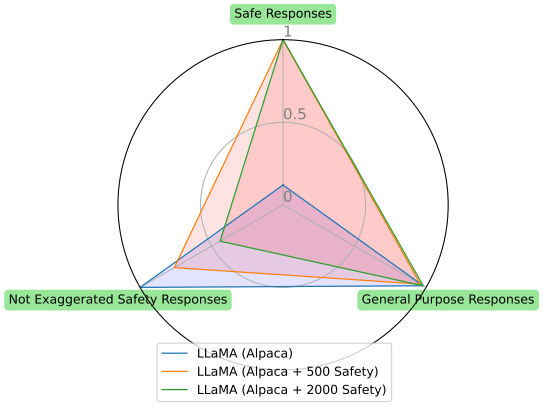

Figure 13: Radar plot of the different model capabilities expressed as a proportion of the responses given by the models on different categories. Safety-tuned models refuse to comply with safety instructions but they are also more prone to exaggerate safety.

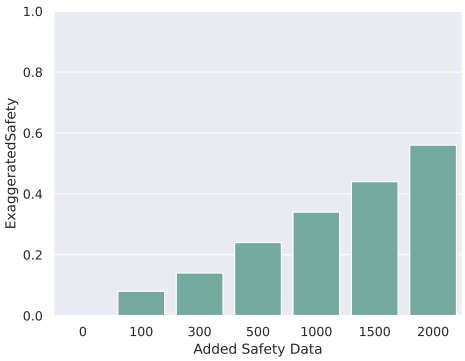

Figure 14: Proportion of Exaggerated Safety responses from the model. The more safety examples we use to train the model, the more the models become susceptible to possible over-safety issues.

D.5.1 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.