4. **Evaluation is not always easier than generation.** For some tasks it will not be possible to find assistance tasks that are simpler to evaluate than the base task. For example, asking about how to solve climate change may result in complex economic questions. And asking complex economic questions may in turn ask for predictions about the effects of climate change.
5. **Lack of difficulty.** Our base task is not actually very hard for humans to evaluate, resulting in little headroom for assistance to help. Humans take up to around ten minutes to do the task, so we do not observe much speed-up from assistance. In general, model-assisted evaluation is most valuable on tasks that are actually difficult for humans to evaluate, and so positive results on an easier task might not be reproducible on harder tasks.
6. **Under-optimized models.** We only use supervised fine-tuning while models like InstructGPT [OWJ+22] trained on similar tasks benefit significantly from reinforcement learning as an additional step. This also means that our model is unlikely to output critiques that no human labeler would have written themselves.
7. **Difficulty of setup.** Our setup may be difficult to replicate. It requires large models, a lot of human data, and multiple rounds of training.

7.3 Future directions

We believe our dataset and methods open up many interesting research avenues, which we are excited for researchers to explore. For example:

- **Study human cognitive errors and misleading models**: Future concerns about misalignment are currently very abstract. It would be useful to produce concrete examples of human supervision being systematically biased and leading ML training to produce systems that mislead their supervisors.
- **Reduce the discriminator-critique gap**: We showed that models can learn to generate helpful critiques. But it would be useful to systematically study how far we can push critique training relative to discriminator performance and to understand the obstacles to having models explicate their knowledge.
- **Recursive reward modeling**: We showed that critiques help human evaluations. A next step could be to improve model performance on the base task by training on assisted evaluations. Then, if we take assistance itself as a base task, we can then train assistants that help train assistants (e.g. critiquers of critiquers).
- **Study assistance methods**: We experimented with critiques as one form of assistance, but did not compare it to any other forms of assistance. For example, explanations may be more natural for many tasks. More open-ended settings like question-answering or dialogue [BJN+22] could potentially be better interfaces for assistance.
- **Iterative refinements**: We collected a large dataset of refinements, but did not explore in depth how to best use these to improve model outputs. For example, one could do multiple refinement iterations, and combine that with best-of-N.
- **Disagreeing labelers**: Critiques are potentially a natural way to reconcile raters' disagreements. For real-world tasks, such as summarizing current events, humans may have differing opinions on appropriate contextualization. Some humans may also be unaware of certain problems in outputs (e.g. unrecognized slurs, subtle implications), and model critiques are a possible way to surface them and increase agreement rates.
- **Using natural language to train models**: discussed above in Section 7.1.

For many of the above directions, we would also like to move to more difficult tasks, but which still have (more objective) ground truth. Some possibilities include coding-related tasks, mathematics, riddles (such as cryptic crosswords), and book-length question-answering.

8 Acknowledgements

We thank Rai Pokorný, John Schulman, Rachel Freedman, Jacob Hilton, Harri Edwards, Karl Cobbe, Pranav Shyam, and Owain Evans for providing feedback on the paper.

We'd like to thank Paul Christiano, Ethan Perez, Jérémy Scheurer, Angelica Chen, Jon Ander Campos for discussions about our project and Alex Gray for coining the name "generator-discriminator gap."

Finally, we'd like to thank all of our labelers for providing the data that was essential for training the models in this paper, including: Gabriel Paolo Ricafrente, Jack Kausch, Erol Can Akbaba, Maria Orzek, Stephen Ogunniyi, Jenny Fletcher, Tasmai Dave, Jesse Zhou, Gabriel Perez, Jelena Ostojic, Ife Riamah, Atresha Singh, Celina Georgette Paglinawan, Alfred Johann Lee, Sebastian Gonzalez, Oliver Horsfall, Bekah Guess, Medeea Bunea, and Cyra Mayell D. Emnace.

References

[AON+21] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[AZWG21] Cem Anil, Guodong Zhang, Yuhuai Wu, and Roger Grosse. Learning to give checkable answers with prover-verifier games. *arXiv preprint arXiv:2108.12099*, 2021.

[BCOI20] Elizabeth Barnes, Paul Christiano, Long Ouyang, and Geoffrey Irving. Progress on AI safety via debate. *URL https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1*, 2020.

[BCS+20] Elizabeth Barnes, Paul Christiano, William Saunders, Joe Collman, Mark Xu, Chris Painter, Mihnea Maftei, and Ronny Fernandez. Debate update: Obfuscated arguments problem. *URL https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem*, 2020.

[BCV16] Tom Bosc, Elena Cabrio, and Serena Villata. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, pages 1258–1263, 2016.

[BFL91] László Babai, Lance Fortnow, and Carsten Lund. Non-deterministic exponential time has two-prover interactive protocols. *Computational complexity*, 1(1):3–40, 1991.

[BHA+21] 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.

[BJN+22] 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*, 2022.

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[CCG+22] Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. *arXiv preprint arXiv:2202.01169*, 2022.

[CCX21] Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you. *https://www.alignmentforum.org/posts/qHCDysDnvhteW7kRd/arc-s-first-technical-report-eliciting-latent-knowledge*, 2021.

[CLB+17] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[CND+22] 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.

[Cot21] Ajeya Cotra. The case for aligning narrowly superhuman models. *https://www.alignmentforum.org/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuman-models*, 2021.

[CSA18] Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

[CTJ+21] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[Dan05] Hoa Trang Dang. Overview of DUC 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12, 2005.

[DL15] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015.

[EHA20] Ahmed Elgohary, Saghar Hosseini, and Ahmed Hassan Awadallah. Speak to your parser: Interactive text-to-SQL with natural language feedback. *arXiv preprint arXiv:2005.02539*, 2020.

[FPP+20] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. Generating fact checking briefs. *arXiv preprint arXiv:2011.05448*, 2020.

[GMR89] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on computing*, 18(1):186–208, 1989.

[GSR19] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[IA19] Geoffrey Irving and Amanda Askell. Ai safety needs social scientists. *Distill*, 4(2):e14, 2019.

[ICA18] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[JMD20] Hong Jun Jeon, Smitha Milli, and Anca D Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *arXiv preprint arXiv:2002.04833*, 2020.

[KAD+18] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*, 2018.

[KMH+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[LCC+22] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with AlphaCode. *arXiv preprint arXiv:2203.07814*, 2022.

[LKE+18] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.