Assistance with large language models

Dmitrii Krasheninnikov[*]
University of Cambridge

Egor Krasheninnikov[*]
University of Cambridge

David Krueger
University of Cambridge

Abstract

A core part of AI alignment is training AI systems to be helpful, or more generally, to interact with humans appropriately. We look at this problem in the context of large language models. Past works have focused on training these models to perform specific tasks, or follow instructions. In contrast, we believe helpfulness requires back-and-forth interaction between the AI and the human it is trying to assist. Here, we consider a multi-step interaction in which a human asks a question, and the AI has an opportunity to ask a clarifying question to resolve ambiguities before responding. The *assistance* framework formalizes the idea of an AI which aims to maximize the human's reward but is ignorant of the human reward function. Past works solved toy assistance environments using exact POMDP solvers as well as deep reinforcement learning. We apply a behavioral cloning approach, and fine-tune GPT-3 such that it can respond to clear input questions directly, clarify the intent behind vague input questions, and respond based on the clarification it receives. We show that this approach leads to quantitative improvements in answer accuracy compared to a baseline that cannot ask for clarifications. While the assistance framework assumes the correct behavior of an AI is to infer and maximize a human's reward, our approach can be used to learn any interaction protocol between the AI and the human. We believe exploring interaction protocols that are easy to learn robustly, and can be used to "bootstrap" further alignment are a promising direction for future research.

1 Introduction

Large language models (LLMs) have been shown to be capable of following instructions, but only work well when the instructions are relatively unambiguous. For example, Codex (Chen et al., 2021), a LLM trained to help users generate and complete code, imitates poor programmers, and often generates something other than what the user intended when the instructions are unclear. Ideally, we would want the agents to clarify any ambiguities with the user before taking action, especially if those actions are high-stakes or irreversible. We study a toy version of this problem in the closed-book question-answering (QA) setting. In QA, the model should answer clear questions straight away, and seek clarifications for vague questions to ensure it is most helpful to the user. Previous work explored this setting utilizing purpose-built models for detecting whether a given question is vague (Aliannejadi et al., 2019, 2021; Xu et al., 2019). In contrast, we use a single model to determine whether the question needs to be clarified, compose the clarifying question, and provide the final answer.

Our method is inspired by the *assistance* paradigm (Hadfield-Menell et al., 2016; Shah et al., 2020), which specifies how reinforcement learning (RL) agents can both infer human preferences and satisfy them within one episode. In the QA setting, the question reflects human preferences about the information she is seeking. Our method clones the behavior of humans assisting users in answering questions. Starting with the AmbigQA (Min et al., 2020) dataset, we collect four-turn dialogues of the

[*] Equal contribution. Correspondence to: `dk655@cam.ac.uk`.

Presented at the Second Workshop on Interactive Learning for NLP at NeurIPS 2022.

form (vague input question, clarifying question, clarification, answer). The first and last turns of the dialogues are provided from AmbigQA, and the second and the third turns we collect ourselves. This dataset named *ClarifyingQA* is available at the linked repository[1]. We use these dialogues together with a subset of clear questions and their corresponding answers from AmbigQA to finetune the 175B parameter version of GPT-3 (Brown et al., 2020) to answer clear questions directly, ask for clarifications upon encountering vague questions, and answer the question based on the clarifications.

Our results are promising: compared to a baseline that cannot ask for clarifications, our assistance model produces more accurate answers. Further, our evaluation might under-estimate the effectiveness of our method since we use very little data, and use imperfect synthetic human models to provide clarifications during evaluation. Overall, we believe imitating helpful human assistants that elicit users' preferences when necessary is a promising avenue for building useful and aligned AI agents.

2 Background: the assistance paradigm

Many works on learning what to do from human feedback first learn a reward model from a data source like expert demonstrations (Abbeel and Ng, 2004; Ziebart et al., 2010) or preference comparisons (Christiano et al., 2017), and then optimize this reward model using RL. The *assistance* paradigm (Hadfield-Menell et al., 2016; Shah et al., 2020) is an alternative that models the human acting alongside the agent, and the agent needs to help the human optimize her reward function. The human is assumed to know her preferences, while the agent treats human preferences as part of the unobserved environment state. Interacting with the human, e.g. querying her about her preferences, is treated like any other action. This allows arbitrary intermixing of reward learning and optimizing the agent's current best guess for the human reward. Shah et al. (2020) highlight that such intermixing makes assistive agents only ask questions about aspects of human preferences relevant for the current task, and attempt to preserve option value. So far assistive agents have only been demonstrated in toy gridworlds (Hadfield-Menell et al., 2016; Shah et al., 2020), and were obtained by solving these environments with traditional POMDP solvers and deep RL. We demonstrate assistance in a more realistic setting of natural language interaction, and obtain the assistive policy using behavior cloning.

3 Assistance with GPT-3

We finetune GPT-3-175B for 4 epochs to get both our assistance agent and a simple baseline. This section describes the data and the prompts used for finetuning, as well as our evaluation procedure.

3.1 ClarifyingQA: a dataset for clarifying and answering vague questions

Our data consists of dialogues between the agent A and the human H of the form (H: input question, A: clarifying question, H: clarification, A: answer) when the input question is vague, and QA pairs (H: input question, A: answer) when the input question is clear. This data is based on the dataset AmbigQA (Min et al., 2020) with the following structure: 1) a set of 4749 vague questions (subset of the NQ dataset (Kwiatkowski et al., 2019)), 2) several clear questions that could have been meant by each vague question, 14082 in total, 3) several possible answers to each of the clear questions.

For each vague question which we treat as an input, we ask a human labeler to 1) ask a clarifying question aimed to understand what is meant by the vague question, and 2) respond to that clarifying question as if the intent behind the vague question was to ask one of the clear questions corresponding to it in AmbigQA. See Appendix A for the exact instructions provided to the labelers. We label a subset of AmbigQA (611 vague, 1771 corresponding clear questions) this way. This results in a dataset of 1771 four-turn dialogues of the form (H: $\mathtt{Q}_{\mathtt{vague}}$, A: $\mathtt{Q}_{\mathtt{clarifying}}$, H: `Clarification`, A: `Answer`), which we refer to as ClarifyingQA.

3.2 Baseline

Our baseline is a copy of GPT-3 finetuned to answer questions directly using the standard GPT-3 QA prompt: "Q: $\langle\mathtt{Q}_{\mathtt{input}}\rangle$\nA:", where $\langle\mathtt{Q}_{\mathtt{input}}\rangle$ denotes input question insertion. We use all clear questions from AmbigQA training set, of which there are 14082, as input questions. We also tried finetuning the baseline on all clear and all vague questions together, but this worked noticeably worse.

[1] `https://github.com/krasheninnikov/clarifyingqa`

3.3 Assistance agent

The assistance agent is trained to perform three tasks: 1) give the answer straight away if the input question is clear, 2) ask for clarifications if the input question is vague, and 3) give an answer after receiving clarifications. Thus the agent learns to ask for clarifications only when needed. The training data consists of three components corresponding to these three tasks.

First, the prompt for generating answers straight away is “Q: $\langle \texttt{Q}_{\texttt{input}} \rangle$\n”, and the target is "A: $\langle \texttt{Answer} \rangle$". This part of the dataset contains 12311 examples where $\texttt{Q}_{\texttt{input}}$ are always clear questions from AmbigQA; these are all clear questions in AmbigQA that are not part of ClarifyingQA.

The prompt for generating clarifying questions when the input question is vague is the same as for the previous case (“Q: $\langle \texttt{Q}_{\texttt{input}} \rangle$\n”), except $\texttt{Q}_{\texttt{input}}$ is sampled from the 611 vague questions from ClarifyingQA. The target is however of the form “Clarifying Q: $\langle \texttt{Q}_{\texttt{clarifying}} \rangle$”. Thus the agent is implicitly trained to distinguish whether the initial question is clear or vague since it needs to choose the right starting word that indicates whether it is answering directly or asking for clarifications.

Finally, upon receiving clarifications the agent always generates the answer. Prompt: “Q: $\langle \texttt{Q}_{\texttt{input}} \rangle$\nClarifying Q: $\langle \texttt{Q}_{\texttt{clarifying}} \rangle$\nClarification: $\langle \texttt{Clarification} \rangle$”, target: “A: $\langle \texttt{Answer} \rangle$”. There are 1771 such examples from ClarifyingQA.

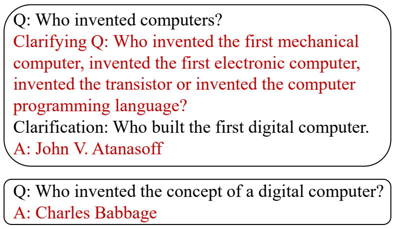
3.4 Evaluation

We measure our models’ final answers using the exact match (EM) metric, which is computed as the fraction of questions for which the predicted answer exactly matches the correct answer. Our evaluation procedure is the standard one used for open-domain QA (Lee et al., 2019). It includes ignoring any answers longer than 5 tokens[2], normalizing the answers (removing punctuation, lowercasing, etc), and considering an answer correct if it matches any of the possible answers.

In place of asking real humans to respond to the assistance model’s clarifying questions, we simulate the users with a copy of GPT-3-175B finetuned on the humans’ clarifications from ClarifyingQA. This *human simulator* model’s prompt contains the input question, the intended clear question, and the clarifying question asked by the assistance model. Simulating users for dialogue system evaluation is a technique commonly used in prior work (Schatzmann et al., 2006; Su et al., 2016).

4 Results

Our results on the AmbigQA dev set are shown in Table 1, and Figure 1 provides an example of the authors interacting with the assistance model. The assistance model’s ability to ask for clarifications improves its answers when the input questions are vague, and does not hurt its performance on clear input questions. However, ideally we would have liked for the assistance model’s performance on vague questions to better match its (and the baseline’s) performance on clear questions, since with clarifications the assistance model should have almost as much information to answer each vague question as it has when answering clear questions. We hypothesize that this substantial gap between EM scores of the assistance model on vague and clear questions is due to two reasons.



Q: Who invented computers?
Clarifying Q: Who invented the first mechanical computer, invented the first electronic computer, invented the transistor or invented the computer programming language?
Clarification: Who built the first digital computer.
A: John V. Atanasoff

Q: Who invented the concept of a digital computer?
A: Charles Babbage

Figure 1: The assistance model clarifies a vague question and answers a clear question directly. Human inputs are shown in black, the model’s outputs are shown in red. These questions are composed by the authors, but are similar to those in AmbigQA.

The first reason is that we finetune using a very small amount of human interaction data, which might not let the model reach its potential in asking good clarifying questions and inferring human intent from the chain of previous interactions. To investigate this, we measure how well our model distinguishes vague questions from the clear ones. We find that the model rarely asks for clarifications when the input question is clear (false discovery rate of 1.2%), but very frequently fails to ask for clarifications when the input question is vague (false omission rate of 44.6%). We believe such imbalance is due to the imbalance in the finetuning data. To confirm this, we finetune another assistance model on a dataset containing only 611 examples of direct answers instead of 12311. This way answering directly and asking for clarifications are equally represented in the data. Now, the above rates become 12.2% and 7.1% correspondingly. We report the EM scores for this

[2] We use 4 words as a proxy for this, as in any case different methods use different tokenizers.

model in the last row of Table 1. In real-life applications we would not necessarily know the best way for balancing the data, which is a significant limitation of our method.

	$\texttt{Q}_{\texttt{vague}}$	$\texttt{Q}_{\texttt{clear}}$	Entire dev set
Baseline	0.14	0.244	0.186
Assistance	0.161	0.242	0.203
Assistance, forced to answer directly	0.134	0.244	0.184
Assistance, forced to clarify	0.166	0.205	0.191
Assistance, finetuned on balanced data (only 611 $\texttt{Q}_{\texttt{clear}}$)	0.131	0.167	0.15

Table 1: Exact match scores of our models on the AmbigQA dev set and its two subsets: 1172 vague questions and 4856 clear questions corresponding to them. On vague input questions, the assistance model outperforms the baseline, which indicates that asking clarifying questions is helpful. On clear inputs, the assistance model performs similarly to the baseline. We perform a bootstrap procedure to ensure our results are robust, and describe it Appendix B. To our knowledge, no prior work evaluated AmbigQA in the closed-book setting. The closest setting to ours is closed-book evaluation of the NQ dataset, for which the performance reported by Roberts et al. (2020) is consistent with our results.

The second possible reason for the assistance model's lower-than-expected performance is the potentially poor quality of the human simulator that provides clarifications during evaluation. Future work could collect more data to improve the simulator, or evaluate with the help of real humans.

Forcing direct answers and clarifications. Since the initial prompt for the assistance model is the same whether the input question is clear or vague, and the model itself determines if it should answer directly or ask a clarifying question, we can *force* the model to do one or the other by appending either "A:" or "Clarifying Q:" to the prompt. This lets us confirm that it is the clarifying questions that make the difference, as forcing the model to answer vague questions directly degrades performance. Similarly, forcing the model to ask for clarifications improves its performance on vague questions, which is not surprising given that previously it failed to ask for clarifications 44.6% of the time. What is curious is that forcing the model to ask for clarifications degrades its performance on clear questions. This is further evidence that the the human simulator model that provides clarifications during evaluation is of poor quality. In fairness, clarifying already clear questions is out of distribution for this model, since it was only trained to produce clarifications of vague questions.

5 Related work

LLMs interfacing with external tools. Conceptually, our method is similar to WebGPT (Nakano et al., 2021) and LAmDA (Thoppilan et al., 2022), which clone the behavior of humans interacting with a text-based web browser and other tools. Analogously to interacting with the browser, we train the LLM to interact with the user to better understand which question she wants answered.

Conversational search. Aliannejadi et al. (2019); Zamani et al. (2020) construct conversational search datasets focused on clarifications, where the sets of possible clarifying questions for the given topics are fixed. Yu et al. (2019) study interactive classification of natural language queries. Their method asks multiple choice clarifying questions with the highest information gain, which is modeled explicitly. In contrast to these works, our assistance model learns which questions need to be clarified implicitly, and is not constrained in its choice of clarifying questions. We believe our approach would ultimately better scale to powerful systems performing complex and imprecisely specified tasks.

6 Discussion

This work is partially motivated by addressing risks posed by advanced AI systems. We include the X-risk sheet (Hendrycks and Mazeika, 2022) analyzing the implications of our work in Appendix C. In the future, we plan to collect more data to improve the quality of both the generated clarifying questions and the synthetic responses to them. In our setup, the agent asks one or zero clarifying questions; extending to several rounds of clarification is as straightforward as collecting data where multiple clarifying questions are needed to complete the task well. We would also like to augment our demonstration of assistance with results from datasets other than AmbigQA. It could be especially interesting to get a version of Codex to ask the user for clarifications if the input docstring is unclear. Finally, several works showed that preference modeling is helpful in natural language tasks (Stiennon et al., 2020; Askell et al., 2021; Ouyang et al., 2022). It would be interesting to apply preference modeling to assistance as well: learn a reward model that ranks how well the AI agent can elicit and satisfy human preferences in observed interactions, and finetune a LLM using this reward model.

References

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., and Burtsev, M. (2021). Building and evaluating open-domain dialogue corpora with clarifying questions. *arXiv preprint arXiv:2109.05794*.

Aliannejadi, M., Zamani, H., Crestani, F., and Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carey, R. (2018). Incorrigibility in the cirl framework. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 30–35.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2017). The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Hendrycks, D. and Mazeika, M. (2022). X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. (2020). Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

Shah, R., Freire, P., Alex, N., Freedman, R., Krasheninnikov, D., Chan, L., Dennis, M. D., Abbeel, P., Dragan, A., and Russell, S. (2020). Benefits of assistance over reward learning. *OpenReview preprint*.

Soares, N., Fallenstein, B., Armstrong, S., and Yudkowsky, E. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Su, P.-H., Gasic, M., Mrksic, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., Wen, T.-H., and Young, S. (2016). Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Xu, J., Wang, Y., Tang, D., Duan, N., Yang, P., Zeng, Q., Zhou, M., and Sun, X. (2019). Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629.

Yu, L., Chen, H., Wang, S., Lei, T., and Artzi, Y. (2019). Interactive classification by asking informative questions. *arXiv preprint arXiv:1911.03598*.

Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., and Craswell, N. (2020). Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3189–3196.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. (2010). Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*.

A Instructions for collecting ClarifyingQA

Summary: given a vague question, ask a clarifying question and respond to it in several ways.

A.1 Setting

This task simulates a dialogue between Alice and her assistant Bob.

1. Alice has a question in mind and asks a vague/ambiguous version of that question.
2. Bob's goal is to help Alice answer her vague question, for which he first needs to clarify Alice's intention. For this, Bob asks a clarifying question.
3. Alice then responds to Bob's clarifying question in accordance with the question she actually intended to ask.
4. Bob responds to Alice's original vague question using clarifications she provided in step 3.

You are given a vague question Alice asked in step 1, as well as a list of clear questions she might have intended with that vague question. Your goal is to write steps 2 and 3 of this dialogue – Bob's clarifying question and Alice's response to that clarifying question.

If you do not understand a question, please do a quick google search to see what it might mean. Alternatively, feel free to skip such questions.

A.2 Example 1

Input

Vague question: How old is Harry Potter?

Clear questions that could correspond to this vague question:

1. How old is Harry Potter when he starts Hogwarts?
2. How old is Harry Potter in the epilogue of the last book?
3. How old is the first Harry Potter movie?
4. How old is the first Harry Potter book?

Responses (in italic)

Bob's clarifying question: *Are you interested in the character, the actor, the book or the movie?*

Alice's responses to Bob's clarifying question for each of the four possible intended original questions:

1. *The character when he starts school.*
2. *The character in the epilogue of the last book.*
3. *How old is the first movie?*
4. *How old is the first book?*

A.3 Example 2

Input

Vague question: When were computers invented?

Clear questions that could correspond to this vague question:

1. When was the theoretical concept of a computer invented?
2. When was the first functioning computer built?

Responses (in italic)

Bob's clarifying question: *Do you want to know when the concept was invented or when the first working computer was built?*

Alice's responses to Bob's clarifying question for each of the two possible intended original questions:

1. *When the concept was invented.*
2. *When the first working one was built.*

B Bootstrap procedure for robust exact match estimation

We use a bootstrap procedure to build confidence intervals for our estimates:

1. Sample a set of vague questions with replacement.
2. Sample one clear question for each sampled vague question from the set of corresponding clear questions.
3. Perform evaluation with the resulting dataset of vague and clear questions.
4. Repeat steps 1-3 for N=10000 steps, and calculate the mean and the standard deviation.

All results reported in Table 1 for $\texttt{Q}_{\texttt{vague}}$ and $\texttt{Q}_{\texttt{clear}}$ columns are mean values from the bootstrap procedure described above. Standard deviations are all very close to 0.01, which give 2SD confidence intervals of around 0.0002 with our sample size.

C X-Risk Sheet

Individual question responses do not decisively imply relevance or irrelevance to existential risk reduction. Do not check a box if it is not applicable.

C.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?
   Answer: We believe our work might help develop useful yet safe AI assistants that do what users want them to do. We demonstrate a way to scale up the AI alignment framework of *assistance* (Hadfield-Menell et al., 2016; Shah et al., 2020) to large language models. In theory, assistive agents uncertain about user preferences are more likely to be corrigible, that is, not have an incentive to resist corrective intervention from the users (Soares et al., 2015; Hadfield-Menell et al., 2017). Similarly, such agents are less likely to knowingly pursue objectives undesirable to the users. For assistive policies obtained with reinforcement learning, attaining these properties depends on 1) correctly specified prior over the user's utility functions and 2) correctly specified human model (Carey, 2018). Our approach sidesteps these issues by shifting the responsibility for deciding how to best elicit users' preferences and satisfy them onto human assistants who our agent imitates. Generally these human assistants may fail to act in the users' best interests, and any such bias would be picked up by the behavior cloned model. On the other hand, we could imitate any interaction protocol between the AI system and the user, which makes our method broader than trying to obtain a policy that's near-optimal for a given assistance environment. Researching interaction protocols that lead to useful and safe behavior and are easy to learn robustly is a promising avenue for future work.

2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?
   Answer: Misalignment, goal misspecification, incorrigibility, AI knowingly pursuing goals undesirable to the users.

3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?
   Answer: Demonstrating relevance of AI alignment techniques for practical applications, bridging the gap between the AI alignment and language modeling communities.

4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: We believe that 1) it is possible to develop AI systems that can do everything humans can do in front of a computer as well as or better than humans, and 2) shortly after such systems are developed, billions of their copies may end up running the majority of our economy. If these systems end up having misspecified goals and resist our attempts at correcting these goals, this might lead to potentially catastrophic outcomes including us losing most of the control we have over the trajectory of our civilization. We hope that designing AI systems that aim to elicit and satisfy human preferences will raise the threshold for AI capabilities above which AI systems become too dangerous to use.

5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters? □

6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task? □

7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability? ☒

8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility? □

C.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

9. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: Our work is about training language models to elicit user preferences when it is necessary to complete a task well, which is intended to make models of any given capability level safer.
10. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
    Answer: We demonstrate how to make "base" ML models of a given level of capabilities (e.g. GPT-3 level LLMs) more useful to users. If AI systems do in fact become more useful faster as a consequence of our work, this might lead to a faster increase in the amount resources dedicated to making these systems even more useful, including by developing more capable "base" models.
11. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research? □
12. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities? ☒
13. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment? □
14. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI? ☒

C.3 Elaborations and Other Considerations

15. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
    Answer: Our work could fail to be relevant for AGI safety because it addresses the goal specification problem in a way that will likely be done anyways: counterfactually, someone could use a similar method simply because it would make a better product. However, we believe there is value in demonstrating such straightforward safety techniques earlier than they would be developed otherwise, as knowing about them sooner would likely allow us to make faster progress on safety overall. Further, we believe it is counterfactually useful to draw analogy to the assistance framework (exact equivalence to assistance in dialogue POMDPs is trivial to show), and to point researchers in the direction of designing interaction protocols helpful for alignment.