Review of ***Training language models to follow instructions with human feedback*** (InstructGPT)

February 3, 2023

1 Reviewer #1

Summary:
This paper presents an interesting approach to align the language models (GPT-3 in this case) with their users' intent. This is very important because often these language models generate outputs that contain untruthful and toxic information that is not helpful to the user. The paper proposes a three step framework to improve the model's alignment with its users: (1) supervised finetune of the GPT-3 models with a dataset of labeler demonstrations. (2) train a reward/ranking model using the fine-tuned model and a dataset of labeler annotations comparing model outputs. (3) further fine-tune the GPT-3 model with reinforcement learning using the trained reward model. This paper has thorough human evaluations on the resulting fine-tuned model (called as InstructGPT). The paper shows that InstructGPT improves on various aspects including truthfulness and reduces toxic output generation in comparison to the baseline GPT-3.

Strengths And Weaknesses:
Strengths:

- The proposed approach is very interesting and novel. A similar idea has been applied previously for improving summarization but the current paper addresses a much broader distribution of tasks.
- The human data collection process is well documented (in the Appendix). The experimental results are very thorough with a lot of human evaluations. Human evaluations make the most sense to show the effectiveness of proposed model to cover a broad distribution of tasks

Weaknesses:

- The paper provided only minimal information on the deduplication process of prompts, that too in the supplementary. I think it is important to clearly mention the full details of this process. It is not clear how the paraphrased prompts (or prompts with the same example or task but with different templates) are handled.
- Also, it is unclear whether there is any overlap on the training data and NLP datasets used in the paper for some ablations (e.g., RealToxicityPrompts and TruthfulQA). It would be useful to share some analysis on the data distributions.
- Sorry if I missed it somehow, but I did not find any "empirical" evidence for the claim: "1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3"

Questions:

- Can you provide some empirical evidence on by how much margin 1.3B InstructGPT is better than 175B GPT-3?
- Can you provide evidence that there is no contamination of training data and evaluation datasets (NLP datasets) used in the results section?
- How is the hallucination rate of a few-shot setup in comparison to zero-shot setup (for both GPT and InstructGPT)?
- It seems that for RM model training, the paper used more than double the data used for SFT. Is there any reason for this? How much minimum data is decent enough for the RM model training?
- Do you plan to release the data used in SFT and RM?

Limitations:

- Adequately addressed the limitations.

Conclusions:

- **Ethics Flag:** No
- **Soundness:** 3 good
- **Presentation:** 4 excellent
- **Contribution:** 3 good
- **Rating:** 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
- **Confidence:** 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
- **Code Of Conduct:** Yes

2 Reviewer #2

Summary:
This paper addresses the challenge of aligning large language models with user intents. They introduce a three-step process to do so: 1) Train supervised models on (human-collected) prompts and the (human-labeled) outputs 2) Use the models to generate labels which are then ranked by humans according to the desired metrics (bias, toxicity, etc.). Then train a Reward model on the data 3) Fine-tune a supervised policy using PPO algorithm with rewards produced by the Reward model.

They show that the resulting models are better at following user instructions. They make fewer mistakes, use less offensive language, and have fewer hallucinations, as measured by automatic metrics and human evaluation.

Strengths And Weaknesses:
Strengths:

- This work takes an important step in reducing bias, toxicity, and other ethical issues concerning large language models.

- The results align with the premises and claims. The discussion provides a good comparison between GPT3 and InstructGPT3 capabilities and where and why InstructGPT3 does not improve over GPT3.
- The paper is clear and easy to follow. It does not overclaim and cites relevant work where appropriate.

Weaknesses:

- This approach relies heavily on human-collected and labeled data. Accessibility of such data, esp. for small research organizations are not clear. Also, the compute power required to fine-tune these models is quite high. As a result, the benefits of this research might not be applicable to the larger research community. The authors do not mention whether they intend to release their dataset or software.
- The approach does not provide a comprehensive solution to the general problem of bias and toxicity. The current approach is mainly data-driven and relies on reward signals which amplifies user preferences, etc. Thus an ill-intended user is now better equipped to mis use such models for their own benefit.
- It's not clear whether current approach can be applied to smaller LLMs (1-10B parameters).

Questions:

- Is this approach specific to large language models or can it be used for smaller LLMs (1-10B parameters)?
- What are the possible solutions to avoid manipulation or misuse of alignment by certain individuals? Is it possible to introduce ”inhibition” so that models refrain from following user's instructions if deemed biased, toxic, or inappropriate?

Limitations: N/A

Conclusions:

- **Ethics Flag:** No
- **Soundness:** 3 good
- **Presentation:** 3 good
- **Contribution:** 3 good
- **Rating:** 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.
- **Confidence:** 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
- **Code Of Conduct:** Yes

3 Reviewer #3

Summary:
This paper studies fine-tuning language models to produce outputs thare are more aligned with what users would like them to do. Given the increases focus on language models and their failure modes (hallucinations, toxic responses etc.) this is an important research direction.

Strengths And Weaknesses:

- large scale evaluation in terms of model sizes (GPT-3 size) and number of considered tasks
- effective approach to improve outputs of LMs with respect to what users expect the problem is clearly significant as language models are an increasingly important research direction

Questions:

- Is the comparison between the SFT and PPO models fair? SFT was trained on 13K labeled prompts while PPO collectively uses 13K + 33K (both labeled) + 31K unlabeled prompts. Disclaimer: I did not read the appendix.
- Do you truly need a reward model? You could directly train the final model on the 33K PPO dataset using expected risk which uses the labeler rankings as a reward. You would not use the 31K unlabeled prompts, but the result would give you an indication of the value of these unlabeled prompts.

Limitations:
The authors adequately addressed the limitations and potential negative societal impact of their work.

Conclusions:

- **Ethics Flag:** No
- **Soundness:** 4 excellent
- **Presentation:** 4 excellent
- **Contribution:** 4 excellent
- **Rating:** 8: Strong Accept: Technically strong paper, with novel ideas, excellent impact on at least one area, or high-to-excellent impact on multiple areas, with excellent evaluation, resources, and reproducibility, and no unaddressed ethical considerations.
- **Confidence:** **5:** You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
- **Code Of Conduct:** Yes

4 Reviewer #4

Summary:
This work proposes a instruction tuning process for gpt3. It involves three stages: (1) Collect demonstration data, and train a supervised policy. (2) Collect comparison data, and train a reward model. (3) Optimize a policy against the reward model using PPO. It is compared to a few ablation baselines, or models trained on other instruction datasets. It is compared with a number of strong baselines, and the performance is strong. It is shown that with the instruction tuning, the model can beat a 100x larger LM. (but keep in mind that larger LM is not deliberatedly tuned on instructions.)

Strengths And Weaknesses:
Strength:

- The collected data would be a good resource for further research into instructed LMs.
- The comparison is solid and the performance is strong. It is shown that with the instruction tuning, the model can beat a 100x larger LM. (but keep in mind that larger LM is not deliberatedly tuned on instructions.)

Weakness:

- Collecting data/ranking, is of course, costly.
- It seems to me that this work is kinda typically adding more data/human labor, and the performance goes up. While the performance is good, the novelty is limited.
- More analysis experiment could be done. (see my questions)

Questions:

- Would one ablation baseline be directly do RL with human feedback ? (without training the reward model)
- If you have a smaller size of human feedback, how would the performance be affected?

Limitations:
Yes.

Conclusions:

- **Ethics Flag:** No
- **Soundness:** 3 good
- **Presentation:** 3 good
- **Contribution:** 3 good
- **Rating:** 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
- **Confidence:** 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
- **Code Of Conduct:** Yes

5 Response to reviewers

Comment: We thank their reviewers for their kind feedback. We are glad to hear the reviewers find the paper “interesting and novel” (R1), “clear and easy to follow” (R2), and with “very thorough” experimental results (R1) on a problem that is “clearly significant” (R3).

We address concerns of the reviewers below.

- Performance on smaller (1B-10B param) models (R1 + R2) R2 states that it is not clear if our method can be applied to 1B-10B param models. We do in fact train 1.3B and 6B versions of InstructGPT, which form a core part of our results. In Figure 1, we show that the 1.3B InstructGPT (labeled 'PPO') outperforms the 175B GPT-3 (labeled 'GPT (prompted)') on human preferences (y axis). Figure 1 also addresses R1's question about evidence for 1.3B InstructGPT's performance. In our camera-ready copy, we will edit the caption of Figure 1 to make this result clearer.

- Cost of human data collection and compute is high (R2 + R4) We agree that the cost of collecting data and training large language models can be high. But we also believe that our method might make smaller LMs more competitive – given that our 1.3B InstructGPT outperforms a 175B GPT-3 on our API distribution, collecting comparison data and using RLHF might lead to similar performance gains for smaller LMs on other NLP tasks. This would be a significant cost savings, as collecting human data is far cheaper than training a 100x larger LM, and has many other advantages (it's easier to specify model behavior, improvements in toxicity and truthfulness, etc.).

- Is the comparison between PPO and SFT fair? (R1 + R3) It is true that our comparison between PPO and SFT is not fair, as we use way more comparison data than demonstration data. However, the point of our paper isn't to show that PPO with an RM on comparison data is necessarily the best algorithm to use, but rather to show an existence proof that this method can lead to extremely strong results at a large scale on realistic language tasks. SFT and PPO have been directly compared in a recent paper on book summarization [1], with the finding that PPO was more data efficient.

- Prompt duplication + overlap with NLP datasets (R1) We assume R1 is asking about potential overfitting to our eval sets, which is a reasonable concern. We generally did not check for overlap between our training set and our public NLP benchmarks (though we are using most of these to simply measure regressions, and we don't claim any SOTA results) – the exception for this is TruthfulQA, where we did a very light check (confirmed that a common TruthfulQA question did not appear in our test sets).

- Upon further investigation to answer this question, we found that our description of our deduplication procedure was slightly inaccurate (it applied to validation sets that did not make it into the paper) – we don't check for shared common prefixes, but rather we remove prompts that entirely contains another prompt. We will correct this in our camera-ready version. (Separately, we've found that more aggressive deduplication does not significantly affect our results). Our biggest defense against overfitting on our API dataset is that we split our training / validation / test sets by user, so our test set consists entirely of prompts from users that were not trained on. Empirically, we believe significant overfitting is unlikely as InstructGPT can generalize to tasks that are extremely rare in the training dataset (see Sec 4.3).

- Avoiding misuse (R2) We agree that InstructGPT can make it easier for malicious actors to misuse LMs, and we note this in Section 5.3. We believe mitigating this requires interventions at many levels of the model deployment pipeline; for example, a process for approving applications before they go into production, and tools to monitor and react to misuse. However, some of these are not enforceable if the model is open-sourced. At the model level, one could use our method to fine-tune the LM to refuse to respond to certain tasks that may be harmful (using a combination of demos + comparisons). This is an exciting direction for future work.

- Do you plan to release the SFT and RM data? (R1 + R2) Unfortunately, since the majority of our data comes from customers using our API, we are not able to release it publicly.

[1] Wu et al., "Recursively Summarizing Books with Human Feedback", 2021.