Limitations, Generalization, and Broader Implications

1. Implications for alignment research

The researchers talked about some inherent implications for the creation of this type of model

1. The cost of increasing model alignment is modest to relative to pretraining, specifically where this model uses less data to train compared to other models.
2. The model generalizes instructions to settings that it wasn’t supervised on, when given non-English tasks and code-related tasks.
3. Performance degradation was mitigated by the fine-tuning.
4. The techniques were validated from research of the real world.

2. Who are we aligning to?

A number of factors influence the alignment of the model trained, in terms of its responses to the user, manly being the training data, the fine-tuning data, and the alignment method used. The researchers discussed some influences of their model:

1. The model is aligned to the labeller’s demonstrations and preferences. The labeller’s were mostly English‐speaking people (US, Southeast Asia) hired via Upwork/Scale AI. The authors report inter‐labeller agreement of about 73%.
2. The labelling instructions that labeller’s use as a guide when writing demonstrations are created by the researcher’s, creating an inherent bias to their preferences.
3. The training data is determined by prompts sent by customers, and as such, implicitly aligning what customers think what their end user’s think is valuable for the use of the model, which may not always be the best solution for the end user.
4. The customers of OpenAI are not representative of all potential users, nor by all individuals and groups impacted by language model use.

One way to mitigate this is to train models on the preferences of certain groups, or fine tune or prompted to easily represent different groups. Yet these models may still affect broader society, meaning that more discussions are needed to create harmony between several groups.

3. Limitations

The researchers talked about several limitations of their study and model:

In terms of the methodology the labeller population is not fully representative of all users or all cultural/linguistic backgrounds. Most prompts and data were in English, and many of the comparisons were labelled by only one annotator. The researchers note that having examples labelled by different people could help identify areas where their labellers disagree.

Even after fine-tuning, the models still make errors. They sometimes fail to follow instructions, hallucinate facts, generate biased or toxic outputs, or comply with harmful user instructions. For example: when instructed to be maximally biased, the model produced more toxic output than baseline.

4. Open questions

The researcher’s talk about several open questions to explore to further align language model behaviour.

Methods to reduce the propensity to generate toxic outputs like collection of worst-case scenarios, filtering pre-training data, and combining different methods that worked better for other models.

Training the model to be less harmful is important, but whether an output is harmful will depend on the context. Also, harmful outputs can be beneficial for data augmentation.

To improve the controllability of the model it may be useful to allow users to specify preferences, or adjusting behaviour via control codes or sampling mechanisms.

Making comparisons of text may not be the best way to align models. There are several ways to critique model responses, this being a human-computer interaction problem.

5. Broader Impact

Finally, the authors consider the broader societal and ethical implications of their work.

The fact that model outputs depend on the training data, labelers and user population. Meaning that deployment carries risks of bias, misalignment with under‐represented groups, or unintended consequences.

Even positive use-cases may have side-effects. It is possible for that persuasive text may increase usage time, which may not be good for the user’s well-being.

Theres a need for ongoing monitoring and governance, to ensure models are used responsibly, that their alignment target is revisited as deployment contexts change.

It is emphasised that fine-tuning with human preferences is not sufficient alone and that other mechanisms like data filtering, monitoring, refusing harmful instructions are also necessary.