Attention all students

This is a critique of **completely unrelated paper** that we don’t want you to read. This file is only here so you can get an idea of what we expect for each review. We will be able to tell if your responses lack effort.

**Paper Name**

We’re Afraid Language Models Aren’t Modeling Ambiguity

**Summarize the Paper**

Reinforcement learning from human feedback can help correct when language models show undesirable text generation behavior. The problem is that the learning signal from this conveys limited information on long text outputs. The information doesn’t specify which aspects of the output influenced the learning signal. This paper introduces fine-grained RLHF. It helps with training and learning from reward functions after each sentence as well as incorporating multiple reward models associated with different feedback types (factuality, relevance, and information completeness). The work shows how these types of rewards leads to greater performance.

**List the 2-3 main new claims made by the authors of the paper. For each claim, discuss how it is supported by experiments or theoretical analysis.**

1) This paper introduces fine-grained RLHF. It helps with training and learning from reward functions after each sentence as well as incorporating multiple reward models associated with different feedback types (factuality, relevance, and information completeness). This is more performant than preference based RLHF based on annotator feedback of generated outputs.

2) Using multiple rewards allows a user to tune towards a more specific behavior type which could fit a user’s needs better. The paper empirically shows the trade-off between error types by changing the weights.

**List 2-3 questions you have that you found difficult to understand in the paper.**

1) When they asked the annotators of their preference, was this before or after the annotators were trained on fine grained RLHF? This may have aQected the results.

2) How would this method scale to larger LM since a lot of applications are meant for larger models?

**List 1 or 2 papers that represent the most closely related prior work. How does this paper extend on this? Why is this a significant contribution?**

1). “Improving alignment of dialogue agents via targeted human judgements.” Trains reward models that assign scores for different attributes but uses a single reward that combines the scores from all reward models.

**List 2-3 limitations of the paper. For each limitation, discuss how it might be addressed in future work.**

1) The dataset they used for their fine-grained RLHF is small 3853 with 6 comparisons per prompt is very small especially if they say it only takes 6 minutes for a worker to complete fine-grained RLHF. The authors could have hired more annotators.

2) The model used is rather small. Some aspects of this process could only be seen on a smaller model or not. We don’t know because they used a smaller model.