We thank the reviewers for their comments. We are pleased that reviewers found the paper to be very empirically convincing (**R1**, **R2**), that the experiments are rigorous and well-documented (**R1**, **R2**), and that the human feedback dataset we collect would be useful for the community (**R2**, **R4**). We respond to the concerns raised by reviewers below.

One of the primary concerns was that our paper lacks novelty relative to prior work. We agree: novelty is not the aim of the paper. Rather, the strength of the paper is in obtaining extremely strong empirical results using a technique that has received relatively little attention, on an important task (abstractive summarization). **R2** agrees that it is likely to set a new standard for the task. Specifically, our main contributions are:

(1) We show that training with human feedback significantly outperforms very strong baselines (human-written reference summaries and large supervised models). To do this, we: (a) scale up model size, (b) move data collection to the batch setting, closely monitoring human data quality, and (c) separate the value and policy networks.

(2) We show human feedback models generalize much better to new domains than supervised models. Our TL;DR model matches human reference summaries on CNN/DM news articles without any fine-tuning, while analogous supervised baselines trained on only TL;DR regress to the performance level of pretrained baselines.

(3) We conduct extensive empirical analyses of both our reward model (effect of data/model size, over-optimization, what it has learned), and of other metrics such as ROUGE (predictiveness of human judgements, over-optimization).

(4) We collect a large dataset of human comparisons of summaries, which we aim to release publicly.

We agree with R4's suggestion that we reframe the introduction to highlight these contributions and will update this.

@R2, R4: Comparing to more baselines (e.g. PEGASUS or Ziegler et al.). We agree this would be ideal, but comparing to other works using human preferences is expensive. We will follow **R2**'s recommendation to add more discussion of prior work (including PEGASUS) in the final version. Note that Table 3 of the Pegasus paper suggests they underperform compared to reference summaries on the Reddit TIFU dataset, while we convincingly outperform reference summaries on our Reddit dataset (including the TIFU subset). We believe that comparing to prior work with much smaller models is uninformative and masks gains from the technique. We agree with **R2**'s assessment that our chosen baselines (human-written reference summaries, large supervised models, T5) are strong.

@R1: The paper doesn't test if ROUGE is good for measuring human preferences. These results are detailed in the appendix. **R1** states "if RL on ROUGE achieves similar performance, there is no point of human feedback", but in Figure 13 we find optimizing ROUGE performs significantly worse than optimizing our RMs, with "over-optimization" after just a few bits of pressure. **R1** states "our model achieves high ROUGE scores". However, this does not imply high correlation with human preferences, which depends on the sample distribution. In Appendix F.7, we find ROUGE and human preferences to be correlated for supervised baselines, but uncorrelated for samples from our best model, providing more evidence that as models improve, ROUGE stops tracking quality. Similarly, in Figure 14a human feedback models achieve *lower* ROUGE scores than the supervised baselines on TL;DR.

@R4: "Maintaining a very close feedback loop with labelers cannot be a good contribution." We disagree. While we understand that human data collection techniques are difficult to ablate, our view is that the standard method of collecting human data in the field — using crowdsourced websites with minimal researcher-labeler interaction — is a significant impediment to using this data for training models (as evidenced by Ziegler et al.). We believe our processes (onboarding, monitoring agreement rates, providing feedback, etc. see Appendix C.1) were important for strong results. We hope our results convince researchers to pay closer attention when collecting human data.

@R1 "Maybe we can do pointwise annotation". This can work but there will be discrepancies across labelers and drift while shifting distributions. Comparisons also let us measure worker agreement which we use to control quality.

@R3: "The cross-domain experiments on news show that in-domain tuning is still necessary." Figure 4 shows the opposite: our models achieve very strong performance on news with no in-domain training. In-domain supervised fine-tuning slightly improves quality scores but this is explained by the longer length of CNN/DM summaries. We encourage reviewers to inspect the non-cherry-picked samples themselves.

@R4: "The paper is not well organized." We believe this may partially be because we present our contributions as "results" in the introduction, as this is an empirical paper. We are open to specific suggestions.

@R3, R4: Too many details in Appendix. As **R3** notes, we struggled with the page limit, and interesting details (e.g. Figure 13) are omitted from the main text. Opinion on this is perhaps largely personal preference. **R2** enjoyed having details in the appendix. **R4** prefers model details to be in the main text, but we disagree as our models are standard.

@R1: What does it mean that lead-3 outperforms the reference summaries on CNN/DM? We were also surprised by this result, which deserves further attention. Appendix D contains some more detailed analysis.