We thank all reviewers for their detailed and valuable feedback. In the following, we address the reviewers' comments and questions, which we will clarify in the final revision.

Novelty of the joint model [R3] We agree with R3 that [3,37] pioneered gaze integration in NLP tasks, paving the way for future works like ours. We did not intend to claim we are the first to propose gaze integration in NLP. Instead, a key novelty of our work is our joint modelling approach to attention and comprehension. Instead of only using gaze information as additional supervision during training, we propose a joint training framework in which our pre-trained text saliency model (TSM) is adapted to a target task for which no gaze data is available. We will clarify in the paper that this joint training framework and not the specific method used for gaze integration is one of our core contributions.

Comparison to prior methods using gaze data in NLP [R1,R3] A core contribution of our work is the joint modelling of attention and comprehension for text comprehension tasks, which we show to improve performance in extensive evaluations. We value the suggestion made by R1 and R3 to additionally compare our approach to previous works using gaze in NLP tasks ([3] and [37]). When comparing our results for sentence compression on the google dataset to [37], who employ gaze as an additional supervision in a multi-task learning setup, we observe an increase of ~5% F1 score for our method. A comparison to [3] is outside the scope of our current work as the focus of [3] is on sequence classification, while ours is on modeling human attention with text comprehension (paraphrasing and summarization). To the best of our knowledge, no previous works studied gaze integration for the paraphrase generation task.

Evaluation to Current SOTA + Baseline Models [R4] We agree with [R4] that an evaluation against the recent work on paraphrase generation by Li et al. [2019] would be desirable. Unfortunately, several works, including Li et al. [2019] cite the Kaggle Competition release of the Quora Question Pairs (QQP) dataset and use an undocumented split and subset of the original QQP dataset. The authors did not reply to our requests for details on the splits. Therefore, we compared to the state-of-the-art for which a reproducible split was published, citing the original QQP dataset and using the same splitting protocol, to ensure comparability [24,54]. When comparing our BLEU-4 scores (28.82 BLEU-4) to Li et al. [2019] (best BLEU of 25.03), in spite of the unknown dataset split, we still outperform this recent work.

Regarding the number of parameters in the baseline models, we tried to reach the authors of [54,85] to obtain this information, but they did not get back to us. Crucially, we show the effectiveness of our joint training approach cannot be explained by an increase in parameters as we compare our full model against ablated variants with the same number of parameters (*Random TSM Init*, *TSM Weight Swap*, and *Frozen TSM*). All ablations are inferior to our full model.

We use BLEU because it is an established metric [37] which is currently still agreed on as a standard for text generation tasks [53]. We also show that our improvements in BLEU-4 score are statistically significant using paired t-tests. We agree that additional metrics can provide more information regarding the generated text. We welcome the reviewers suggestion on the novel metrics published at ICLR20 and ACL20, less than two months before the NeurIPS deadline and will add all suggested metrics to the revised version, providing additional points of comparison for future work.

Model Architecture Choices [R1,R2,R4] We chose a BiLSTM with a Transformer for our TSM model, as this combination yielded predictions most similar to human data in preliminary experiments, particularly with longer sequence lengths. According to Wang et al. [2019], this might be due to the transformer encoding coarse relational information about positions of sequence elements, while the BiLSTM better captures fine grained word level context. We will a detailed ablation study of the TSM in the supplementary material. Lastly, for decoding we use greedy search.

Impact of Pre-training on CNN and Daily Mail [R4] By pre-training on the CNN and Daily Mail corpora using EZ-Reader and subsequently fine-tuning on human gaze data, our TSM model learns to accurately predict human gaze on sentences. These human-like saliency weights output by the TSM are already beneficial to the upstream tasks without further joint training (Frozen TSM), and improve even more during the joint training process. R4 raises the question of whether e.g. a masked language model trained on the CNN and Daily Mail can lead to a comparable performance when integrated into the joint model. While we agree that this is an interesting research question, it lies outside the scope of this paper in which we investigate the impact of a text saliency model in our joint training framework.

Task Choice + Generalizability [R4] We chose sentence compression, which is a text summarization task. Studying extractive summarization alongside paraphrase generation allows us to show generalizability of our method across two tasks with different output formats, requiring different evaluation metrics (BLEU and F1 score). In our opinion, the positive results on both of these different tasks are a strong statement in favor of the generalizability of our approach.

References

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. Decomposable neural paraphrase generation. In *ACL*, 2019.

Zhiwei Wang, Yao Ma, Zitao Liu, and Jiliang Tang. R-transformer: Recurrent neural network enhanced transformer. *arXiv:1907.05572*, 2019.