NeurIPS Rebuttal for "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"

We thank reviewers for their thoughtful, detailed reviews. We shall discuss three common themes, before addressing individual comments.

Novelty of retrieval-augmentation (R1, R2, R4) "information retrieval strategy to improve the the generation models is... not novel" (R1), "retrieval+extraction/generation framework have also been conducted by some previous works" (R2), and "the framework is in some sense the encoder-decoder version of the REALM model" (R4). We agree that retrieval is a well-known strategy in pipeline-based NLP systems. We emphasize the following points that we consider to be novel contributions in this area:

- Pre-trained seq2seq models have only become available in the last year (T5, BART) or two (GPT2). To our knowledge, work has not yet been published investigating how retrieval-augmentation affects this class of pretrained generators.
- Unlike prior generation work, we show retrieval and generation can be trained jointly with a single objective, demonstrating joint training is effective, outperforming both fixed retrieval and BM25 pipelines
- We study two RAG models. RAG-Sequence's formulation is similar to REALM, but RAG-Token is novel and has not been previously proposed. Further, we explore novel decoding strategies for these models.
- We demonstrate an effective general trainable retrieval-augmentation framework for *any* knowledge-intensive task, rather than ad-hoc task-specific architectures, such as Chen et al.'s DrQA. We feel this constitutes a novel and important class of model which combines parametric and non-parametric memories for any task.

Reusing existing Components vs developing a novel model (R1, R2) "contribution [...] is not very specific, since that RAG and its components are not proposed by authors" (R1), and "One tiny weakness is that the core technical components are just borrowed directly from the existing works" (R2). We believe the ability to re-use existing components is specifically a strength of our framework:

- We believe avoiding pre-training is positive, as this saves computational resources and makes the process accessible to smaller labs. Re-using existing components is attractive in this respect and results in state-of-the-art performance, outperforming several approaches including end2end pretrained models.
- RAG is modular and agnostic to the retriever and generator, so it can directly benefit from future innovations in generation and retrieval in isolation without needing to pretrain new models. For example, RAG can be used to finetune a GPT3 generator with a ColBERT retriever right away, which were released near NeurIPS submission time.
- From a scientific standpoint, re-using existing components allows us to make direct comparisons with previously proposed/analyzed and widely-used models like BART (for generation) and BERT/DPR (for retrieval). The model re-use lets us highlight the contribution of our proposed end-to-end training procedure, specifically by not re-inventing the wheel for individual retrieval and generation components.

Brevity of descriptions of models (R1, R3) R1 suggested that "A figure or example about PAG-Sequence Model and PAG-Token Model is needed", and R3 mentions "description of the model is quite concise (due to space restrictions)". We will happily add further exposition and add more detail on the differences between RAG-Sequence and RAG-Token.

No document encoder training experiments (R1) We focus on query-encoder finetuning due to its low compute cost and simplicity (the large document index does not need to be updated during training). We show this method's effectiveness compared to a fixed retriever. We consider document-encoder training to be out-of-scope here, but agree it is an interesting topic for future work that could potentially lead to gains albeit with significantly higher compute cost.

More architectures and retrieval supervision (R3) R3 suggested we could compare RAG to memory-network style approaches, as well as ablations looking at joint supervision. We agree that these experiments are interesting for completeness, but we believe that our existing baselines are sufficient to justify the effectiveness of our contribution.

More Generator baselines (R4) R4 suggested "it would be more interesting to compare RAG with GPT-2 or T5 models rather than BART". We use BART in the RAG models in our experiments as has been shown it performed well on a number of language tasks, outperforming similarly-sized T5 models (see BART paper). Since we use BART for RAG, a BART-only generator is the appropriate baseline to determine the effect of retrieval augmentation. Additional generator baselines would be interesting for completeness but we argue our existing experiments are sufficient to support our conclusions.