A Human study details

Human studies were conducted through the Amazon Mechanical Turk platform. Prices of tasks were carefully calculated to ensure that workers would have an average compensation of 12USD per hour. In all studies, examples were sampled from the test split of the CNN/DM dataset that contains a total of 11,700 examples.

As with any human study, there is a trade-off between the number of examples annotated, the breadth of the experiments, and the quality of annotations. Studies conducted for this paper were calibrated to primarily assure high quality of results and the breadth of experiments.

A.1 Underconstrained task

Human annotators were asked to write summaries of news articles and highlight fragments of the source documents that they found useful for writing their summary. The study was conducted on 100 randomly sampled articles, with each article annotated by 5 unique annotators. The same configuration and articles were used in both the *constrained* and *unconstrained* setting.

Questions for the *constrained* setting were written by human annotators in a separate assignment and curated before being used for to collect summaries.

A.2 ROUGE - Weak correlation with human judgment

This study evaluated the quality of summaries generated by 13 different neural models, 10 *abstractive* and 3 *extractive*. A list of evaluated models is available in Table 6.

The study was conducted on 100 randomly sampled articles, with each article annotated by 5 unique annotators. Given the large number of evaluated models, the experiment was split into 3 groups. Two groups contained 4 models, one group contained 5 models. To prevent from collecting biased data, models were assigned to experiment groups on a per-example basis, thus randomizing the context in which each model was evaluated. To establish a common reference point between groups, the reference summaries from the dataset were added to the pool of annotated models, however, annotators were not informed which of this fact. The order in which summaries were displayed in the annotation interface was randomized with the first position always reserved for the reference summary.

A.3 Layout bias in news data

Human annotators were asked to read news articles and highlight the sentences that contained the most important information. The study was conducted on 100 randomly sampled articles, with each article annotated by 5 unique annotators.