SQuALITY: Building a Long-Document Summarization Dataset the Hard Way

Alex Wang   Richard Yuanzhe Pang   Angelica Chen   Jason Phang   Samuel R. Bowman
New York University
alexwang@nyu.edu   yzpang@nyu.edu   bowman@nyu.edu

Abstract

Summarization datasets are often assembled either by scraping naturally occurring public-domain summaries—which are nearly always in difficult-to-work-with technical domains—or by using approximate heuristics to extract them from everyday text—which frequently yields unfaithful summaries. In this work, we turn to a slower but more straightforward approach to developing summarization benchmark data: We hire highly-qualified contractors to read stories and write original summaries from scratch. To amortize reading time, we collect five summaries per document, with the first giving an overview and the subsequent four addressing specific questions. We use this protocol to collect SQuALITY, a dataset of question-focused summaries built on the same public-domain short stories as the multiple-choice dataset QuALITY (Pang et al., 2021b). Experiments with state-of-the-art summarization systems show that our dataset is challenging and that existing automatic evaluation metrics are weak indicators of quality. SQuALITY is available at https://github.com/nyu-mll/SQuALITY.

1 Introduction

Research on automatic text summarization generally requires adequate benchmark datasets. Existing datasets in this area often have issues that seriously limit their usability: For instance, summaries from the popular scraped benchmark summarization dataset CNN/DailyMail (Nallapati et al., 2016) contain HTML artifacts, links to other news articles, and other types of noise (Kryscinski et al., 2019; Tejaswin et al., 2021).

A common approach to creating summarization datasets is to develop heuristics to extract pseudo-summaries from existing texts. While scraped summaries can be cleaned of noise, these heuristics can lead to more fundamental data artifacts. For example, the XSum dataset (Narayan et al., 2018) was
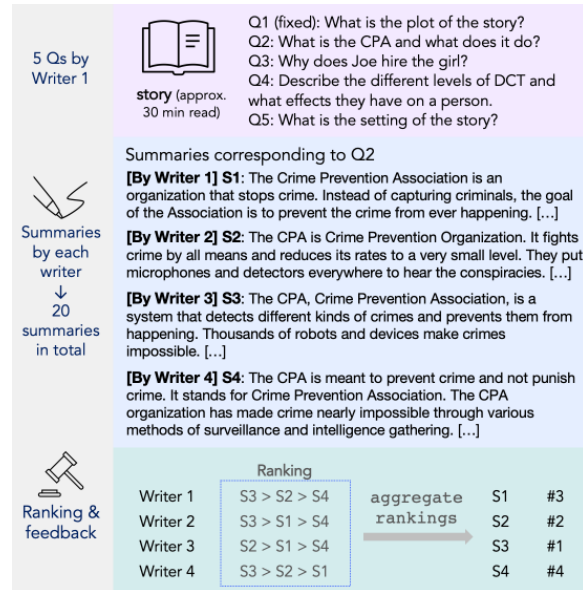

Figure 1: An overview of our data collection pipeline. One writer first creates four questions, with an additional fixed question used for every story. Then, four writers each write summaries answering the five questions. Next, each writer ranks the other three summaries for each question and provides written feedback. Finally, we aggregate ranks and award bonuses to incentivize high-quality summaries. Between collection rounds, writers review the feedback they received.

created by extracting the first sentence of a news article to act as the summary for the rest of the document. However, studies have found that 30–50% of summaries created this way contain facts that are unsupported by the rest of the article (Tejaswin et al., 2021; Nan et al., 2021). Models trained on this dataset learn to repeat this noise pattern by hallucinating facts in their outputs. It appears that known heuristics do not produce reliable data.

Another approach to creating summarization datasets relies on serendipity in finding naturally occurring summaries. For example, the arXiv and PubMed datasets (Cohan et al., 2018) use the abstracts of scientific papers as summaries of the papers. BigPatent (Sharma et al., 2019) and GovRe-

port (Huang et al., 2021) use expert-written summaries that come with patent filings and government reports, respectively. While these summaries are likely high-quality, the domain of the data poses a significant challenge for system evaluation: Automatic evaluation metrics for summarization are unreliable (Kryscinski et al., 2019; Gehrmann et al., 2022), but the summaries are too technical and jargonistic for non-specialist human raters to evaluate reliably. Because we rely on chance in finding these summaries, we are beholden to whatever domain they come from, rather than the domain we are interested in.

Relying on finding and scraping summarization data is also problematic in that, often, the found data is proprietary and not freely distributable. For example, many researchers and organizations are unwilling to host or distribute the CNN/DailyMail dataset,[1] despite it being one of the most popular summarization datasets to experiment on. Similarly, several recent summarization datasets built on data such as scientific journal papers (Meng et al., 2021) or SparkNotes book summaries (Ladhak et al., 2020; Kryściński et al., 2021) have never been made available to researchers. The dataset creators instead ask potential users to re-scrape them, which can be a serious obstacle to reproducibility.

In this work, we propose a crowdsourcing protocol for collecting original summaries free of these issues. Crowdsourcing summaries has been underexplored because straightforward approaches for doing so are quite labor-intensive. While our protocol is still fairly demanding, we structure it in a way that makes the cost per summary more tractable (∼$6/summary) while also including incentives and checks to ensure the summaries are high-quality. The protocol does not rely on finding naturally occurring summaries and is agnostic to the input documents used, so we are free to choose the input documents we want to summarize. We use short stories from Project Gutenberg to avoid the aforementioned domain and licensing issues.

We use this protocol to collect SQuALITY[2] (Summary-format QUestion Answering with Long Input Texts, Yes!), a dataset for question-focused abstractive summarization of short stories. SQuALITY summaries are created by having trained writers read short stories, then ask questions about different aspects of the story. The writers then answer

the questions by writing summaries focusing on that aspect. Each question is answered by four different annotators, who then review each other's work to ensure the data is high-quality. In total, SQuALITY consists of 100 stories, 500 questions, and 2000 summaries.[3]

Overall, we make the following contributions:

1. We develop a crowdsourcing protocol for collecting summaries that partially ameliorates the high cost of crowdsourcing long textual responses while maintaining data quality.

2. We use this protocol to collect SQuALITY, an abstractive summarization dataset. SQuALITY is question-focused, multi-reference, and distributed with a CC BY license.

3. We conduct preliminary experiments on SQuALITY with pretrained language models using human evaluation. We find that state-of-the-art summarization models produce summaries that are significantly worse than human-written summaries.

4. We identify that common automatic evaluation metrics for summarization correlate very poorly with human judgments of quality. We also find that having multiple references when computing automatic evaluation metrics does not improve the correlation of the metric.

SQuALITY is a challenging benchmark for long-context text generation models. The SQuALITY dataset, code for our baselines, data from our human evaluation of models are available at `https://github.com/nyu-mll/SQuALITY`.

2 Related Work

Story Summarization A common focus of summarization research is on stories and narratives. BookSum (Kryściński et al., 2021) consists of public domain books and summaries of those books, chapters, and paragraphs. Similarly, Ladhak et al. (2020) propose a dataset for summarizing chapters of public domain books. Both of these datasets use summaries scraped from popular study guide websites such as SparkNotes, apparently without an overt license, and thus the datasets cannot be legally distributed. SummScreen (Chen et al., 2022) consists of fan-written transcripts of TV

---

[1]See discussion here.

[2]Named because it uses many of the same stories as the multiple choice QA dataset QuALITY (Pang et al., 2021b).

[3]This paper releases SQuALITY v1.0. We will soon release SQuALITY v1.1, which consists of 127 stories.

Title: Pick A Crime (https://www.gutenberg.org/ebooks/51656)	
Q: What is the CPA and what does it do?	
The Crime Prevention Association is an organization that stops crime. Instead of capturing criminals, the goal of the Association is to prevent the crime from ever happening. They implement thousands of crime-prevention methods and devices. There are many amateur cops who constantly follow criminals around in hopes of catching them in the act so that they may be hailed a hero and...	The CPA is Crime Prevention Organization. It fights crime by all means and reduces its rates to a very small level. They put microphones and detectors everywhere to hear the conspiracies. They place robots as bartenders to control the level of alcohol in visitors to prevent them being drunk. They make all the women learn self-defense. The organization's made crime almost impossible...
The CPA, Crime Prevention Association, is a system that detects different kinds of crimes and prevents them from happening. Thousands of robots and devices make crimes impossible. The association will not punish any crime, instead, the criminal will be send to a CPA hospital for some treatments that will result in getting the best jobs. The CPA also hands out ID cards that states one's...	The CPA is meant to prevent crime and not punish crime. It stands for Crime Prevention Association. The CPA organization has made crime nearly impossible through various methods of surveillance and intelligence gathering. The crime was not punished by the CPA but addressed by sending the person to a hospital for expensive treatment to correct and remove the deviance from the person's...

Table 1: An example question and four human-written references from SQuALITY. The full references are available in Table 9 in the appendix.

episodes paired with Wikipedia and fan-written summaries of those episodes.

Question-Focused Summarization In question-focused summarization (QFS) the summary focuses on a specific aspect of the source text as a way answering a specific question. QFS has received increasing attention from the summarization literature in recent years, and we expect it to be a viable proxy benchmark task for narrative-text summarization broadly. The Debatepedia dataset (Nema et al., 2017) is a found dataset of questions and summary-answers based on articles about social and philosophical issues. FacetSum (Meng et al., 2021) is a found dataset consisting of scientific papers paired with author-written summaries focusing on different aspects of the paper. WikiAsp (Hayashi et al., 2021) and AQuaMuSe (Kulkarni et al., 2020) are two heuristically created, multi-document QFS datasets derived from Wikipedia. GovReport-QA (Cao and Wang, 2022) pair pre-existing summaries from GovReport (Huang et al., 2021) with annotator-written questions to form a hierarchical QFS task.

Most similar to our dataset is QMSum (Zhong et al., 2021), a long-document QFS dataset built around meeting transcripts. Like our work, QM-Sum questions and summaries are composed by writers who have read full transcripts and are guided by a list of question templates. Unlike our work, their primary mechanism for quality control is researcher review of the collected responses, whereas we use a crowdsourcing protocol wherein writers review each other's work.

Long-Form QA QFS is a special case of long-form question answering (LFQA). In LFQA, the inputs are also a question and an input document, and the task is to produce an answer at least one long sentence in length. LFQA answers can draw from a single portion of the document, whereas summaries for QFDS should cover multiple parts of the input document, if not the whole document.

3 Dataset Construction

Source Documents Our considerations in selecting a corpus of documents for which to collect summaries are: (1) The documents are long, as document-level tasks are more challenging than paragraph-level ones; (2) The documents can support several substantive summaries, as we will collect multiple summaries per document for cost-efficiency; (3) The documents have a permissive license so they can be easily distributed; (4) The documents are lay-accessible, such that the average college-educated English-fluent speaker can both understand them and confidently evaluate the correctness of summaries derived from them.

We use short stories from Project Gutenberg as they meet all of these desiderata.[4] Specifically, we use a collection of science fiction short stories written in the 1930s–1970s and are between 3000 to 6000 words long. Many of the stories used are also included in the QuALITY (Pang et al., 2021b) dataset, and we coordinate with the QuALITY creators such that stories that appear in both datasets are assigned to the same split. We use the same

---

[4]https://www.gutenberg.org/

preprocessing for the stories as used in QuALITY.

Writing For writers to create accurate and high-quality summaries, they need to read the entire story, which takes 20–40 minutes. Rather than asking writers to create one summary per story read, we instead collect multiple summaries per story to amortize the cost of reading across summaries. We solicit multiple summaries by having writers ask questions about different aspects of the story, leading us to create a QFS dataset.

We start each crowdsourcing round by asking writers to read the story and then create questions satisfying two criteria: (1) Questions should require the whole or multiple parts of the story to answer, as opposed to a single sentence or paragraph; (2) To minimize disagreements in evaluation, writers should avoid questions that require speculating substantially beyond the literal text of the story when interpreting themes or symbolism. To assist writers in creating questions satisfying these properties, we provide a list of question templates we expect will meet these properties in most cases, shown in Appendix A.1. Writers can also write story-specific questions not based on any of these templates so long as they follow the criteria.

For each story, we assign one worker to create four questions. The questions are then answered by four writers, including the original question writer. Each writer also creates a general story summary, framed as answering the question "What is the plot of the story?", for a total of five questions per story. Responses are required to be 75–500 words long, to avoid copying the text of the story verbatim, and to draw on different parts of the story as much as possible. Writers report that this step takes 40–120 minutes, including time reading the story.

Data Validation After a writing step, for each story, we have five questions with four reference summaries per question. In the second step of each crowdsourcing round, we ask workers to review the summaries to ensure they are high-quality.

As with writing, asking crowdworkers to review the responses is expensive because verifying whether a response is faithful to the story requires having read the entire story. We minimize costs by asking each writer to review the responses of the other three writers. Because the writer has already read the story, they do not need to fully re-read the story, and because they have answered the questions previously, they already have a sense of what

constitutes a good response to each question.

In each validation task, we show the reviewer the original story, the set of five questions, and three responses for each question written by other writers. Reviewers first annotate spans of the responses that contain typos or factual errors. Next, they rank the three responses from best to worst. We instruct the reviewers to rank the responses by (1) how well the response correctly answers the question; (2) how well the summary includes all relevant details; (3) how well the response draws from multiple parts of the story, using their judgment to balance the three factors. Writers are informed during the writing step that their responses will be evaluated along these dimensions. Finally, reviewers provide written feedback for each response about how that response could be improved. The feedback is provided to writers between batches of work to help them improve their responses. Reviewers report that this step typically takes 20–30 minutes.

Afterwards, for each question, we compile the individual reviewer rankings into an aggregate ranking. We incentivize high-quality writing by awarding bonus payments to writers based on their response's placement in the overall ranking. We pay $2.50, $1.25, $0.75, $0.50 for ranking first, second, third, and fourth respectively.[5] The average bonus is $1.25 per response, so writers earn an average additional bonus of $6.25 per story. Workers are informed of the bonus structure before writing.

Similarly, we incentivize high-quality reviewing by awarding bonus payments to reviewers based on how well their rankings agree with the aggregate ranking. For each pair of responses, we pay a reviewer a bonus of $0.50 if their ranking of the pair agrees with the aggregate ranking (i.e., if both the aggregate and reviewer's ranking say response $A >$ response B), so reviewers can earn up to $1.50 per question and $7.50 per story. On average, individual reviewers agree with the aggregate ranking on pairwise comparisons 76% of the time, corresponding to an average bonus of $5.57 per story.

Writer Details Because our tasks are very time-consuming and detail-oriented, we eschew crowdsourcing platforms like Amazon Mechanical Turk where eliciting high-quality responses for these types of tasks can be challenging. Instead, we use a small group of skilled writers for long-term

---

[5]For ties, we sum the bonuses for the tied positions and distribute them evenly.

Dataset	Domain	# Examples	Doc. Len	Summ. Len	Multi-ref?	Public?
CNN/DM	news	311k	804	60	✗	✗
XSum	news	226k	438	24	✗	✗
BookSum	fiction, Sparknotes	12k	5102	505	✗	✗
QMSum	meeting transcripts	1808	9067	70	✗	✓
SQuALITY	sci-fi stories	625	5200	237	✓	✓

Table 2: Summary statistics for various summarization datasets. For BookSum, we consider the chapter-level version. The number of examples is across all splits. For question-based summarization datasets (SQuALITY and QMSum) we count examples as number of unique document-question pairs. Statistics for datasets are borrowed from original dataset papers; statistics for CNN/DM and XSum were borrowed from Kryściński et al. (2021). CNN/DM and XSum are often available online in practice, but distributing the dataset is legally questionable.

contracts, drawing both from Upwork[6] freelancers and undergraduate students from our institution. We hire 11 Upwork writers and 7 undergraduates.[7] Most writers create 20–40 responses for the dataset, although five authors submitted 10 or fewer responses. All writers are informed that their writing will be released publicly for use in AI development.

Our Upwork writers are typically US-based native English speakers. Many of them are college-educated, frequently with degrees in the humanities and prior experience in professional copywriting and editing. We found workers for our task by posting an open call on Upwork to participate in a paid interview. In the interview, applicants review an example writing task with sample questions and responses, and then complete a sample writing task. We hire the top 33% of writers based on their performance on the interview task after manually reviewing their responses. We pay Upwork workers base rates of $13 and $8 for each writing and reviewing task respectively, with additional opportunities for bonuses described above. Overall, Upworkers make on average $20 per writing task (i.e. they average a $7 bonus on writing tasks).

The undergraduates we hire are English-fluent and from diverse nationalities and areas of study—the smaller and more junior pool of applicants prevents us from focusing as much on relevant experience as we do with Upwork. Students are paid a flat $20/hr.[8] Students are hired based on relevant experience and writing samples. After they are hired, we show them the same example task and have them do the practice writing task as the Upwork workers.

Additional details about the hiring process and qualitative differences between the two writer populations are in Appendix A.

4 SQuALITY

We present examples from SQuALITY in Table 1 and summary statistics of SQuALITY and other summarization datasets in Table 2.

Data Size and Splits SQuALITY consists of 100 stories that are split 39/25/36 across the train/validation/test splits (or, equivalently, 195/125/180 document-question pairs). We assign stories to splits to be consistent with the QuALITY dataset (Pang et al., 2021b), so stories that appear in both datasets are assigned to the same split.

SQuALITY contains a similar number of summaries to QMSum (Zhong et al., 2021), another crowdsourced summarization dataset, but SQuALITY contains four references per example and thus fewer input documents. This difference in allocation arises from the crowdsourcing protocol: In creating SQuALITY, we have writers review each other's work while in creating QMSum, the authors manually review all responses. Protocols wherein workers review each other work are more scalable. Having multiple references per input is useful for model evaluation, as automatic metrics such as ROUGE were originally developed on multi-reference datasets. While naive multi-reference ROUGE still correlates poorly with human judgments of quality for SQuALITY (see Section 6), having a diverse set of references opens up opportunities for the development of new evaluation metrics that take into account the diversity of acceptable summaries for a given input, even in the question-focused setting.

Text	N-gram Size			
	1	2	3	4
Diff. story	19.7	2.7	0.1	0.0
Same story	27.4	5.8	1.2	0.4
Same question	33.4	8.7	2.3	0.8
Story	69.4	22.0	5.0	1.7

Table 3: (Top) Average percentage of unique n-grams shared between pairs of responses from different sources: two different stories, different questions but the same story, and the same question. (Bottom) Average percentage of unique summary n-grams that also appear in the corresponding story.

Length Documents are an average of 5200 tokens long (std. 522) without punctuation, [9] with a range from 3473 to 6165—similar to the *chapters* version of BookSum, and shorter than QMSum.

Responses average 237 tokens long (std. 133), corresponding to a compression ratio of 95.4%. The plot summaries have an average length of 442 tokens and are comparable in length to those of BookSum. The other responses are shorter with an average length of 186 tokens, but are still longer than the summaries in QMSum.

Response Diversity We measure summary abstractiveness by computing the percentage of summary n-grams that also appear in the story, shown in Table 3. The high recall of 1-grams is unsurprising given the length of the stories, but the low recall of 3- and 4-grams shows that the summaries are highly abstractive.

We next consider the diversity between pairs of responses to the same question. If responses are similar, then collecting multiple references is potentially wasteful. We show in Table 3 the average percentage of unique n-grams shared between responses to the same question. The overlap is low: Only 33% of unigrams and less than 10% of bigrams are shared between responses to the same question. This overlap is only slightly higher than the average overlap between responses to completely different stories. The low response overlap highlights the diversity of the summarization task, a property made evident in SQuALITY but not in single-reference datasets.

Model	R-1	R-2	R-L	M	BScore
LED	27.7	5.9	17.7	16.5	82.7
PEGASUS	38.2	9.0	20.2	23.4	84.9
BART	40.2	10.4	20.8	24.5	85.3
BART+DPR	**41.5**	**11.4**	**21.0**	**26.1**	**85.5**
Human[*]	46.6	12.5	22.7	30.6	86.2

Table 4: Automatic evaluation results with ROUGE (1/2/L), METEOR, and BERTScore. LED performs worst and tends to repeat a single sentence. PEGASUS performs substantially better, but lags slightly behind BART. The best performing model is BART+DPR. [*]The human reference is evaluated with three references while model-generated summaries are evaluated with four references, artificially raising their scores.

5 Baselines

5.1 Models

We evaluate supervised sequence-to-sequence models on SQuALITY using different pretrained language models as the base model. We implement our baselines using HuggingFace Transformers (Wolf et al., 2020). We do not explore prompting approaches for summarization with closed-access models. Previous work has found that models can be prompted zero-shot to produce high-quality summaries (Radford et al., 2019; Wu et al., 2021), though public models like GPT-3 do not have the capacity to process full stories from our dataset.

BART BART (Lewis et al., 2020) is a Transformer-based (Vaswani et al., 2017) encoder-decoder model pretrained on a token in-filling objective and a sentence permutation objective. We use `BART-large`, which has a maximum input sequence length of 1024 tokens, so we truncate stories dramatically to fit this simple baseline.

BART+DPR We experiment with an extract-then-summarize baseline. Instead of truncating stories when using BART, we first retrieve story sentences that are most relevant to the question and concatenate them to form the input. We use the pretrained Dense Passage Retriever (Karpukhin et al., 2020) that encodes the question into a vector representation and retrieves the story sentences that are most similar to the question.

PEGASUS PEGASUS (Zhang et al., 2020a) is a Transformer-based encoder-decoder model that is pretrained using an objective designed for summarization. The objective is to predict masked out sentences that are selected to be heuristic

Model	Corr.	Coverage	Overall
BART	34.8	15.6	18.1
BART+DPR	45.4	24.3	27.9
Human	**94.1**	**88.8**	**91.3**

Table 5: Human evaluation results for two models and a human-written response. Corr. stands for correctness. Ratings for each property are averaged across 3 workers, then averaged across questions.

pseudo-summaries of the document. PEGASUS is pretrained on sequences of at most length 512, but we follow previous work in finetuning `PEGASUS-large` with a max sequence length of 2048 tokens, truncating stories to fit.

LED Longformer Encoder-Decoder (Beltagy et al., 2020) is an encoder-decoder model where the encoder is a Longformer and the decoder is a Transformer. A Longformer modifies the Transformer architecture with a more efficient self-attention pattern that allows the model to tractably scale to long documents. The LED maximum input length can fit entire stories. We use a context length of 8192 for memory efficiency. The parameters of LED are initialized using BART weights, copied eight times over. We use `LED-base` and set the global attention parameters to attend to the question tokens at the beginning of the input.

5.2 Training

We format example inputs by concatenating the question to the beginning *and* end of the document, separated by a special `[SEP]` token, based on previous work on question-focused summarization (Vig et al., 2021). Each (story, question, reference) tuple is mapped to a separate training instance, so each (story, question) input is associated with four training examples, one per reference. We finetune models using the AdamW optimizer (Loshchilov and Hutter, 2019). Additional training details are available in Appendix C.

5.3 Evaluation

At test time, we generate summaries using beam search with beam width 4. We evaluate the summaries with ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), standard automatic metrics for summarization. We also evaluate with a `RoBERTa-large` based version of BERTScore (Zhang et al., 2020b), which uses RoBERTa to compute the similarity between references and model

generations. For all metrics, we report F1 and handle multiple references by evaluating a candidate against each reference individually, and then taking the max score across references.

5.4 Automatic Evaluation Results

We present results using various automatic evaluation metrics in Table 4. We observe that LED fails to learn the task and generally produces outputs containing long, repeated sentences. The pathological behavior is reflected in the low ROUGE-1 and ROUGE-2 scores for the model. We hypothesized that the poor performance is because the small dataset size is not enough to finetune the additional positional embeddings. We explored transfer learning approaches where the model was first finetuned on a larger long-context summarization dataset, such as arXiv (Cohan et al., 2018) or GovReport (Huang et al., 2021), and then finetuned on SQuALITY. However, training on intermediate datasets did not fix the issue of degenerate outputs, indicating that the additional positional embeddings were not the bottleneck in the model's performance. Overall, we found that public pretrained models for medium to long input tasks were not effective off the shelf. This result is consistent with other work that report Longformer underperforms BART variants on the BookSum story summarization dataset (Pang et al., 2022) and the SCROLLS long-document generation benchmark (Shaham et al., 2022).

PEGASUS, BART, and BART+DPR do substantially better on the task and produce sensible outputs, despite having partial inputs. PEGASUS slightly underperforms the BART variants according to all metrics. BART+DPR outperforms BART with truncated input across all metrics.

Additionally, we evaluate the human references using the automatic metrics by holding one reference out and comparing it with the various metric against the remaining three references. We repeat this process for all references and average the metric score across held-out references. While this use of three references rather than four disadvantages the human references (see Section 6), we still find that they score higher than machine outputs.

6 Human Evaluation

Automatic metrics for evaluating text summarization have been well-documented as correlating poorly with human judgments of various quality

(Schluter, 2017; Kryscinski et al., 2019; Durmus et al., 2020). As such, we accompany automatic evaluation of the baseline systems with human evaluation. We ask workers to rate the quality of outputs from BART and BART+DPR on the test data.

For each task, we show the worker a story and for each of its five questions, two model-generated summaries and a human reference. Workers rate each summary for three properties: correctness, coverage, and overall quality. Each property is rated on a scale from 1-100, similar to direct assessment ratings in machine translation (Bojar et al., 2016). Workers are instructed to assign ratings that align with their preference rankings between systems (Sakaguchi and Van Durme, 2018). We annotate 20 stories (100 questions) with three Upwork workers per story. Finally, we average property ratings across annotators. The worker details and property definitions are available in Appendix E.

We present results of the human evaluation in Table 5 and sample model generations in Appendix D. The standard deviations of property ratings across questions are shown in Appendix E. For all questions and all properties, all human annotators rank the human-written response as better than the model responses. The human-written response has an average rating around or above 90 for all three properties. On the other hand, BART and BART+DPR have an average rating below 50 for all three properties, substantially below corresponding ratings for the human response. Across all three properties, BART+DPR is ranked as better than BART on 70% of examples. The models receive the highest rating on the correctness property among all properties. Upon inspecting the model generations, we partly attribute these relatively high ratings to the fact that the model-generated responses are fairly generic and devoid of specific details. This lack of specificity is reflected in the especially low coverage ratings of the model-generated summaries. Overall, we conclude that fully-public automatic summarization systems still lag significantly behind human writers.

Correlation Between Automatic and Human Evaluations We next consider the correlations between automatic and human evaluations for three subsets of the collected data: only model-written summaries (200 summaries), only human-written summaries (100 summaries), and all summaries. We present the correlations with the judgments of overall quality for these subsets in Table 6.

Metric	Model Only	Human Only	All
ROUGE-1	-7.4	6.8	63.1*
ROUGE-2	-7.8	5.2	42.8*
ROUGE-L	-3.2	14.0	47.8*
METEOR	-11.1	-4.3	54.1*
BERTScore	5.5	0.8	68.7*

Table 6: Pearson correlations (multiplied by 100) between automatic evaluation metrics and human judgments of overall quality for three subsets of the human evaluation data: only model-generated summaries ('model only'), only human written summaries ('human only'), and both ('all'). Correlations are only significant (*) when considering all summaries.

Metric	1	2	3	4
METEOR	2.4	2.5	5.2	-4.3
ROUGE-1	0.5	-0.4	-1.1	6.8
ROUGE-2	0.0	0.3	-0.9	5.2
ROUGE-L	3.3	-6.3	0.6	14.0

Table 7: Pearson correlations (multiplied by 100) between human judgments of overall quality and various automatic evaluation metrics with a variable number of references for model-generated summaries only. When using fewer than four references to compute a metric, we average over all possible subsets of that size. All correlations are near zero and not statistically significant.

When considering all summaries, all metrics have a substantial positive correlation with the human judgments of overall quality. However, these appear to mostly reflect the fact that the automatic metrics rank human-written summaries as better than model-written ones: When considering only model-written summaries or only human-written summaries, the correlations are dramatically weaker and are in no cases significant.

The weak correlations in these settings point to the brittleness of using these automatic metrics when comparing the outputs of two automatic summarization systems, where metric values will similarly be in a narrow range. In light of these findings, we caution against relying on automatic metrics to measure system quality on SQuALITY and instead rely on human evaluation of model outputs.

Multi-Reference Automatic Metrics We next consider whether having multiple references improves the correlation of automatic evaluation metrics. ROUGE was originally developed on multi-reference datasets, but recent summarization datasets are predominantly single reference. This mismatch may contribute to the poor correlation of

Metric	Avg.	Max.	Δ
ROUGE-1	37.9	41.5	3.6
ROUGE-2	8.7	11.4	2.1
ROUGE-L	18.8	21.0	2.2
METEOR	22.7	26.1	3.4
BERTScore	84.8	85.5	0.7

Table 8: Average and maximum metric value across the four references for BART+DPR.

ROUGE with human judgments of quality for these datasets (Pang et al., 2021a; Pagnoni et al., 2021; Scialom et al., 2021, i.a.). We use the multiple references of SQuALITY to measure the effect of varying the number of references used in automatic metrics on the correlation with human judgments.

In Table 7 we show the correlations with human judgments of quality when varying the number of references used to compute automatic metrics on model-generated summaries. We find that using fewer references when computing the automatic evaluation metrics does not substantially change the correlations with human judgments and that these correlations are all around zero. To demonstrate why, we show the average and maximum metric values for each automatic metric in Table 8. We observe that for all metrics considered, the maximum value of the metric is relatively close to the average metric value across references. Despite having diverse references, the metric values are similar across references. Thus, using multiple references does not improve correlations between automatic metrics and human judgments of overall quality. However, we note that simply taking the maximum metric value over references is relatively simple, and that there may be more sophisticated ways to use the diverse references to compute generation quality.

7 Conclusion

We present SQuALITY, a long-input dataset for abstractive question-focused summarization. Because the summaries are crowdsourced rather than found, we can use input documents that are of an accessible domain and under an open license to avoid common issues with existing summarization datasets. Our crowdsourcing protocol gives multiple summaries and references per input while making the cost of data collection more tractable. The SQuALITY dataset is available at https://github.com/nyu-mll/SQuALITY.

Baseline results with competitive public

medium-scale pretrained models suggest that the dataset remains beyond the capabilities of such systems. Our best performing model is an extract-then-summarize model where we use the questions to retrieve story sentences as input. The performance of proprietary larger-scale models remains an open question, and may depend significantly on whether such models can process the full stories without truncation. Developing more robust architectures for processing long documents than Longformer might lead to straightforward improvements on SQuALITY.

Creating efficient and effective methods for evaluating summaries of long input documents remains an open issue. Given the poor correlation of existing automatic metrics with human judgments of model outputs, we expect that these automatic metrics will provide a very weak signal for progress on SQuALITY. We recommend that researchers using SQuALITY evaluate their summarization systems by having human annotators read a selection of our source stories and compare model outputs on those stories. To facilitate this, we will make our templates for human evaluation available, as well as the judgments from our human evaluation experiments to develop better automatic evaluation metrics.

Limitations

A key benefit to SQuALITY is that the summaries are fully crowdsourced, allowing us to circumvent issues with existing summarization datasets by using public domain short stories. While our crowdsourcing protocol is largely agnostic to the input documents used, we do take advantage of the fact that stories have consistent elements, such as plot or setting, in order to provide question templates that we believe will lead to high-quality summaries. For other types of input documents, it may be challenging to predefine question templates for writers.

Also, while collecting multiple summaries per input document helps cover the cost of crowdsourcing summaries, the crowdsourcing protocol is still fairly expensive. Due to the high cost, we are unable to explore the effect of our protocol design versus other possible designs. We are also unable to determine if the high quality of our summaries is due to our worker population and selection, our protocol, or some combination of the two. As the field of NLP moves to tasks on longer inputs, a potentially fruitful line of work is on designing ef-

ficient and effective crowdsourcing protocols for collecting data in these settings.

Summarization has many different variants, whereas SQuALITY only tests question-focused, abstractive, narrative summarization. A model that performs well on this dataset may not generalize to other summarization datasets where the input documents are highly domain-specific, summaries are much shorter, or the data is otherwise dissimilar from SQuALITY. However, we do believe that SQuALITY is useful as a challenging public benchmark for general summarization capability.

Evaluating systems on SQuALITY is challenging as existing automatic evaluation metrics are unreliable when comparing models, as shown in our experiments. However, conventional human evaluation is difficult to execute well because evaluators need to be familiar with the details of a 5000 word story. Paying evaluators to read the story is expensive and verifying they have read it closely is challenging. We currently recommend that dataset users try to conduct human evaluation with a small number of trusted annotators, as we do in this work. An important direction for future research is how humans can efficiently evaluate systems on SQuALITY and long-input tasks broadly, which will be relevant to many NLP tasks as the capabilities of models rapidly evolve.

Ethical Considerations

We expect this work to advance two outcomes: (i) accelerated progress in language modeling, especially toward controllable text generation and long-text comprehension, and (ii) an increase in the hiring of professional and/or crowdworker writers by researchers and product developers in this area. Both of these have potentially significant costs and benefits that are beyond the scope of this paper to investigate.

More concretely, the stories in the dataset were written between 1930–1970 and therefore contain dated and potentially harmful stances on topics like race and gender. Models trained on the data may reproduce these stances, especially if they are trained on the complete texts, rather than the reference summaries alone. We are releasing SQuALITY primarily for use as a research benchmark, and we recommend extreme caution if SQuALITY is used as part of the training set for any deployed system.

Further, the summaries in the dataset were created by writers that are primarily college-educated and either native-English or English-fluent. A system that does well on our dataset only demonstrates competence in mainstream US English, and may not generalize to other variants of English.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint 2004.05150*.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint 2202.06935*.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. BookSum: A collection of datasets for long-form narrative summarization. *arXiv preprint 2105.08209*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint 2010.12694*.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of*

the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2022. Long document summarization with top-down and bottom-up inference. *arXiv preprint 2203.07586*.

Richard Yuanzhe Pang, Adam Lelkes, Vinh Tran, and Cong Yu. 2021a. AgreeSum: Agreement-oriented multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3377–3391, Online. Association for Computational Linguistics.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2021b. Quality: Question answering with long input texts, yes! *arXiv preprint 2112.08608*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint 2201.03533*.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Jesse Vig, Alexander R Fabbri, and Wojciech Kryściński. 2021. Exploring neural models for query-focused summarization. *arXiv preprint 2112.07637*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint 2109.10862*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Crowdsourcing Details

A.1 Question Templates

We provide the following question templates to the writers:

- What is the plot of the story?

- What happens to [character X] throughout the story?

- What is the relationship between [character X] and [character Y]?

- What is the setting of the story?

- What is the significance of [object X] on the rest of the story?

- How is [theme X] explored throughout the story?

- Story-specific questions

Writers always answer the question "What is the plot of the story?". For more subjective templates such as "What is the significance of [object X]?" or "How is [theme X] explored?", we ask the writers to use these templates only in cases where they believe the answer will be clear and unambiguous to someone who has read the story carefully.

A.2 Crowdsourcing Interfaces

We show screenshots of our UIs and abbreviated task instructions for writing and reviewing summaries in Figures 2 and 3, respectively.

A.3 Comparing Upwork and Undergraduates

Generally, we found that both Upwork and undergraduate workers took the task seriously and produced quality summaries. Writers from Upwork qualitatively produced slightly higher quality responses, perhaps because we were able to filter more aggressively for relevant backgrounds and skills when hiring on Upwork. Hiring writers on Upwork was around the same price to slightly more expensive than hiring student writers (Upworkers made slightly more than $20/task while we paid students $20/hr).

Anecdotally, the workers we hired from both populations enjoyed the tasks, and we see this as a significant advantage to using popular fiction in benchmark tasks. However, we did find that some Upwork contractors quit our task during the course of data collection, and some mentioned that our task paid less than other tasks on Upwork. Because students were hired for long-term contracts (on the order of months), they did not drop out of the data collection process, but working with them did require careful work scheduling around exams and breaks.

B Dataset Examples

Table 9 shows the full references for the example in Table 1. Table 10 shows additional examples from SQuALITY.

C Training Details

We train models for 5 epochs with the AdamW optimizer and a linear decay with warmup learning rate schedule. Because of the relatively small size of the training data, we focus on tuning regularization parameters when training the models. We tune the initial learning rate, warmup ratio, weight decay, and label smoothing with grid search over a range of values for each hyperparameter. Models were selected based on the loss on the validation dataset and the ability to generate fluent summaries on the validation dataset. We present the search space for each parameter and the optimal model configurations for each model in Table 11. Our experiments with PEGASUS predominantly led to models that produced degenerate summaries consisting of a single sentence repeated. The final model we use is from the official Google-internal implementation courtesy of the original authors. LED models were trained on a single Nvidia Quadro RTX 8000. Other models were trained on a single Nvidia V100.

D Model Outputs

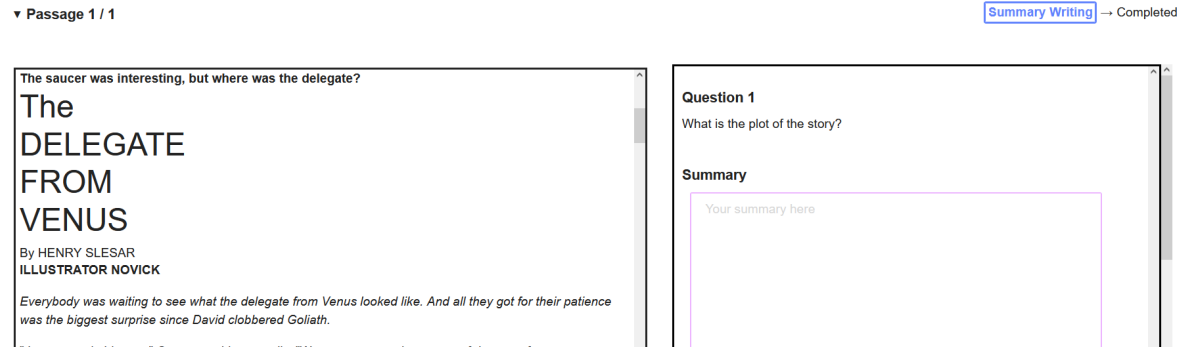We present sample model outputs in Table 12.

Figure 2: Screenshot of the writing UI. Workers are shown the story on the left and five questions on the right, and they are tasked with writing responses to each of the questions. If the worker is the first person to work on a story, they write four questions about the story to answer (The question "What is the plot?" is always asked), and we provide the worker with a list of question templates in the UI to help them write good questions.
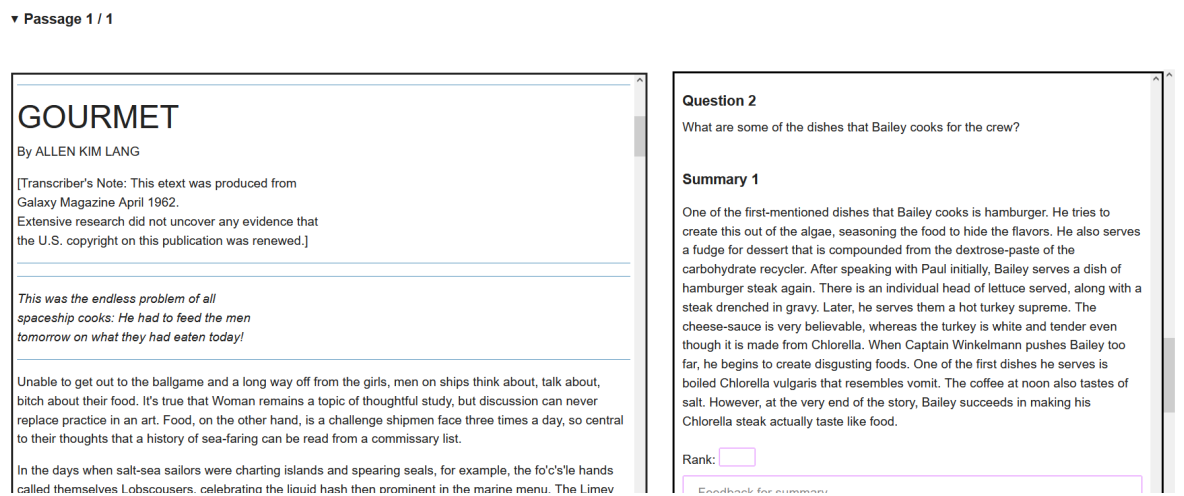

Figure 3: Screenshot of the reviewing UI. Workers are shown the story on the left and five questions on the right. Each of the questions has three responses that the worker is tasked with ranking from best to worst. Additionally, for each response, the worker is instructed to highlight typos and factual errors, as well as provide written feedback to the writer. This feedback is later provided to the writer to help them improve their responses in subsequent rounds of writing.

E   Human Evaluation

As the task is labor-intensive, we use four of the same Upwork writers for the human evaluation as for the data collection. Workers may have previously read the story and thus answered the questions, and we are careful to not show workers their own responses. If they have not previously read the stories, workers are paid to read the story. Workers are informed that the responses are a mixture of human- and machine-written, but not informed which responses are which. We pay workers $8/task and an additional $8 if they have not previously read the story. All workers complete the same number of tasks.

Q: What is the CPA and what does it do?

The Crime Prevention Association is an organization that stops crime. Instead of capturing criminals, the goal of the Association is to prevent the crime from ever happening. They implement thousands of crime-prevention methods and devices. There are many amateur cops who constantly follow criminals around in hopes of catching them in the act so that they may be hailed a hero and given a promotion. Hendricks even explains that the kids have junior CPA clubs, where they record the criminals in little cardboard boxes. They will also follow the criminals around until they die. There are millions of microphones hidden by the CPA everywhere, and any threatening messages are sent to the CPA Brain. The CPA Brain is a monster electronic calculator that can alert police helicopters of any threatening messages, and there are also many hidden TVs and metal detectors. For Arson, heat detectors exist too, and chemical poisoning has made it impossible for people to get poisoned. There are shock treatments, encephalographic devices, a form of prefrontal lobotomy, and a dozen other treatments to reform criminals.	The CPA, Crime Prevention Association, is a system that detects different kinds of crimes and prevents them from happening. Thousands of robots and devices make crimes impossible. The association will not punish any crime, instead, the criminal will be send to a CPA hospital for some treatments that will result in getting the best jobs. The CPA also hands out ID cards that states one's tendency to commit crimes. The CPA has robot bartenders that can detect the drunkenness of a person and prevent anyone from actually getting drunk. There is WSDA teaching judo and jujitsu to women. There are spy cameras and speakers in each alley and street watching every person all the time to prevent all kinds of crimes. The CPA Brain will catch sentences that indicate crimes and watch them more carefully. There are heat-detectors, gun and knife detector, chemical detectors, etc. The CPA brainwashes people, making them believe that crimes are filthy. The treatment will make the criminal's brain catch every attempt that he or she tries to commit a crime and prevents it from happening.
The CPA is Crime Prevention Organization. It fights crime by all means and reduces its rates to a very small level. They put microphones and detectors everywhere to hear the conspiracies. They place robots as bartenders to control the level of alcohol in visitors to prevent them being drunk. They make all the women learn self-defense. The organization's made crime almost impossible and they do not punish for it, but prevent. All who tried to commit a crime are given free treatment. The CPA hospitals treat those few criminals for free and make them unable to commit any further crime. CPA seems to be everywhere, those who tell about the crime are highly rewarded. Neon signs, TV, radio and other means constantly remind people that crime is filth.	The CPA is meant to prevent crime and not punish crime. It stands for Crime Prevention Association. The CPA organization has made crime nearly impossible through various methods of surveillance and intelligence gathering. The crime was not punished by the CPA but addressed by sending the person to a hospital for expensive treatment to correct and remove the deviance from the person's mind. A CPA ID card is required to be carried by everyone and when asked, a person has to present the ID card. Being drunk is illegal according to the rules of the CPA.

Table 9: The four full human-written references from Table 1.

We ask human raters to (re-)read the story, and then evaluate the quality of summaries along three axes:

- Correctness: Presence of factual errors in responses, where a factual error is a statement that contradicts the story, or is not directly stated, heavily implied, or logically entailed by the story.

- Coverage: The degree to which the response contains all information and details from the story that are relevant to answering the question.

- Overall: Overall quality of the response, the primary considerations of which are the readability/intelligibility of the response, the correctness, and the coverage. We ask raters to use their best judgment in balancing these factors, as well as to incorporate other factors such as conciseness, repetitiveness, and copying.

We show the standard deviation of property ratings across questions in Table 13.

SOLDIER BOY

By MICHAEL SHAARA

Illustrated by EMSH

[Transcriber's Note: This etext was produced from
Galaxy Science Fiction July 1953.
Extensive research did not uncover any evidence that
the U.S. copyright on this publication was renewed.]

*It's one thing to laugh at a man because his job is useless
and outdated—another to depend on him when it suddenly isn't.*

*In the northland, deep, and in a great cave, by an everburning fire the Warrior sleeps. For this is the resting
time, the time of peace, and so shall it be for a thousand years. And yet we shall summon him again, my
children, when we are sore in need, and out of the north he will come, and again and again, each time we call,
out of the dark and the cold, with the fire in his hands, he will come.*

— Scandinavian legend

Throughout the night, thick clouds had been piling in the north; in the morning, it was misty and cold. By eight
o'clock a wet, heavy, snow-smelling breeze had begun to set in, and because the crops were all down and the
winter planting done, the colonists brewed hot coffee and remained inside. The wind blew steadily, icily from
the north. It was well below freezing when, some time after nine, an army ship landed in a field near the
settlement.

There was still time. There were some last brief moments in which the colonists could act and feel as they had
always done. They therefore grumbled in annoyance. They wanted no soldiers here. The few who had
convenient windows stared out with distaste and a mild curiosity, but no one went out to greet them.

After a while a rather tall, frail-looking man came out of the ship and stood upon the hard ground looking
toward the village. He remained there, waiting stiffly, his face turned from the wind. It was a silly thing to do.
He was obviously not coming in, either out of pride or just plain orneriness.

Question 2

Describe the setting of the story.

Summary 1

The story takes place on a planet somewhere in the northland of the universe.
The planet is cold and gray, with thick clouds and a heavy, heavy, snow-smelling
breeze. The wind blows in the cold and the sky is gray. The land is covered in
thick, dense, and sometimes icy clouds. The only place that is warm is the
barracks where the soldiers sleep, and there is a fire in the center of the room.
The barracks are small, cramped, and bare, with only a table, chairs, and a bed.
There is a doorbell that rings when the captain calls, and the captain answers it.
The captain is surprised to see Bob Rossel, the captain's right-hand man,
standing outside, and he asks him to come inside. The Captain is confused by the
man's behavior and is not happy that he has been summoned by the captain. The
man is not a soldier, but he is a farmer, and his family lives in a village. The village
is small and sparsely populated, with just a few families. The men sleep in
barracks, and women sleep in sleeping bags. The women sleep on the ground,
and men sleep on top of each other. The soldiers sleep in a cave, where they
keep their weapons. The cave is dark and smelly, and filled with old, crumbling
timbers. The story ends with the captain and Bob Rossell discussing their plans to
destroy the invaders.

Correctness Quality:
-1

Content Selection Quality:
-1

Overall Quality:
-1

Figure 4: Screenshot of the human evaluation UI. Workers are shown the story on the left and five questions on the right. Each of the questions has three responses. For each response, the worker is instructed to rate the responses along the properties of correctness, coverage, and overall quality each along a scale of 1–100. Because the worker is shown three responses at a time, their ratings of each response induce a ranking over the responses. Additionally, workers are asked to highlight errors in responses in order to help them decide on the correctness property.

Title: Tolliver's Orbit (`https://www.gutenberg.org/ebooks/61053`)

Q: Describe the equipment used throughout the story.

Tolliver is a pilot, but while at the Ganymede branch he drives a tractor. One of the equipment used during the story is the automatic flight. An automatic flight allows loaded ships to take a slow and economical orbit using automatic signaling equipment towards Earth. As the loaded ship gets closer to Earth, it is boarded by pilots that land the ship. Another piece of equipment mentioned are spacesuits. The space-suits involve valves and seals and microphones for people to communicate with each other in the spacesuits. The communication is activated by a switch under the chin on the helmet of the spacesuit. They also come with a heavy knife.	Various types of transportation are used throughout the story - tractors to travel on Ganymede between the city and the spaceport, spaceships requiring a lot of fuel and economy orbits which require less fuel but take much longer to get to the place. In a storeroom there are plenty spacesuits, some of which need replacement. Knives are standard suit equipment. Spaceships are equipped with airlocks, ladders and switch-cover. In the control room there is an acceleration seat, a button to set off, a radio and TV, with a screen to see the other side of the call.
Tolliver is first assigned to use an airtight tractor to transport to and from the spaceport. This tractor is like a regular one, but built specifically to trek across Ganymede with its gravity. When Tolliver and Betty are locked into Jeffers' office, he uses a lighter and paper to bend the plastic of the door. Then, he uses a knife to cut through the plastic of the dome. Finally, Tolliver and Betty board a ship, where the orbit is automatically preset in order to preserve fuel. The ship, which Tolliver knows how to operate, is airlocked. Betty uses a transmitter to contact Space Patrol.	Firstly, Tolliver takes Betty towards Jeffers' office on a tractor since it can go through the frozen surface of Ganymede. Then later, when Betty and Tolliver were put in the empty office, Tolliver uses a lighter to light up the mess of discarded records so that the plastic can be bent. Later, inside the storage room, Tolliver finds some spacesuits for the two to wear. Then finally, when they gets to the control room, they gets onto the acceleration seat. Using the ship, the two fly into the economy orbit for Earth in order to escape. In the end, Betty uses the scanner and microphone to make a call to the Space Patrol so that they will arrest Jeffers.

Title: Gourmet (`https://www.gutenberg.org/ebooks/51597`)

Q: What are some of the dishes that Bailey cooks for the crew?

The dishes Bailey cooks for the crew varies greatly, ranging from artificial vegetables to mock-meats. One dish that he makes is a mock-meat hamburger, with the pressed Chlorella tinted pink and seasoned by oregano and thyme. The dish is accompanied by dessert - a fudge made from dextrose-paste. More mock-meat dishes include a hamburger steak covered in a rich, meaty gravy lavishly seasoned with garlic. Another dish includes a mock individual head of lettuce dressed with vinegar and oil. The lettuce was made by Bailey constructing each synthetic lettuce leaf, with the narrator guessing the process to be out of pressing, rolling and shaping a green Chlorella paste. In contrast to some of the delicious dishes that Bailey makes, the Cook also delivers some less tasty meals in response to the Captain's critiques. These included boiled Chlorella vulgaris in some soup and subpar algaeburgers. Bailey's final dish in the story - and the best one yet - is an artificial steak that greets the crew with a barbecue smell. It is drenched with gravy and seasoned with a peppery and garlicy taste, and as the crew eats it, they find that the usually pond-scum taste that accompanies each repurposed chlorella meal is gone and instead, the taste and texture reflects actual steak.	One of the first-mentioned dishes that Bailey cooks is hamburger. He tries to create this out of the algae, seasoning the food to hide the flavors. He also serves a fudge for dessert that is compounded from the dextrose-paste of the carbohydrate recycler. After speaking with Paul initially, Bailey serves a dish of hamburger steak again. There is an individual head of lettuce served, along with a steak drenched in gravy. Later, he serves them a hot turkey supreme. The cheese-sauce is very believable, whereas the turkey is white and tender even though it is made from Chlorella. When Captain Winkel-mann pushes Bailey too far, he begins to create disgusting foods. One of the first dishes he serves is boiled Chlorella vulgaris that resembles vomit. The coffee at noon also tastes of salt. However, at the very end of the story, Bailey succeeds in making his Chlorella steak actually taste like food.
Throughout their trip, Bailey does the best he can in order to replicate traditional food using the Algae. To impress the Captain, Bailey cooks a wide variety of foods including algae burgers, fudge, Steak with gravy and a head of lettuce, Hot turkey with cornbread and butter sauce, and medium rare steak. None of these foods impressed the Captain, so Bailey went back to cooking unappealing food such as a porridge-like broth and bad coffee. At the end, Bailey serves a new type of steak, which is hinted to be human steak from the Captain.	Bailey made a lot of different dishes while working on the Sale ship. He cooked a hamburger and a fudge. He made a steak with rich meat gravy and lettuce, vinegar, and oil. An ersatz hot turkey supreme with a cheese sauce, cornbread, and a pottage was also served at some point. All of these were criticized by Captain Winkelmann. Mostly Bailey was working on the taste of steak, which at the end of the story, he managed to perfect to a certain extent, partly thanks to the captain's constant remarks.

Table 10: Additional example questions and reference summaries from SQuALITY.

Hyperparameter	LR	WR	WD	LS
Search space	{1e-5, 3e-5, 5e-5, 1e-4}	{0.0, 0.1, 0.2}	{0.0, 0.01, 0.1}	{0.0, 0.1, 0.2}
LED	1e-4	0.1	0.01	0.1
PEGASUS	-	-	-	-
BART	5e-5	0.1	0.01	0.1
BART+DPR	5e-5	0.1	0.01	0.1

Table 11: (Top) Search space for the initial learning rate (LR), warmup ratio (WR), weight decay (WD), label smoothing (LS). (Bottom) Optimal hyperparameter configurations for models. The final PEGASUS model we use is from the official Google-internal implementation courtesy of the original authors.

Title: Retief of the Red-Tape Mountain (`https://www.gutenberg.org/ebooks/61146`)

Q: What is the relationship between the Jaqs and the Terrestrials throughout the story?

Reference: The Jaqs and the Terrestrials fight each other throughout the story. It started when a human saw a Jaq and thought it was some type of native game and shot it. From that incident, the Jaqs concluded that the humans were sportsmen like themselves and responded by going to one of the farms and killing two cows. Since then, the two sides have been attacking back and forth, and the humans think the Jaqs are fighting against them...

LED: Retief is a vice-consul at the Embassy of the Mosaic of the Two Dawns. He is in charge of the affairs of the Embassy and is responsible for keeping the diplomatic relations between the two planets in check. He is also responsible for keeping the diplomatic relations in check by sending Retief on his expeditious trip to the planet Adobe. When Retief arrives at the planet, he is greeted by a large Flap-jack, a creature with talons that look like lobsters...

BART: The Terrestrials and the Jaqs have a tense relationship throughout the story. The Terrans have attempted to establish contact with the native life form, the Jaq, in order to try to gain their trust and gain information about their native life forms. The Jaqs are hostile to the Terrans because they consider them to be an invasive species that are trying to take over their home planet, which they consider to be uninhabited. The Jaqs have a history of war with the Terran settlers...

BART+DPR: The Terrestrials and the Flap-jacks are an alien race that live on the planet Adoban. They are hostile to humans and have attempted to stir up trouble with an intelligent alien life form, the Jaq, three months ago. The humans are attempting to establish trade with the aliens in order to gain access to the planet's resources, but the aliens are having none of it. They have no intention of trading with the humans and are only interested in trading with them for food and...

Table 12: Example model generations on SQuALITY.

Model	Correctness	Coverage	Overall
BART	$34.8_{16.9}$	$15.6_{13.3}$	$18.1_{13.1}$
BART+DPR	$45.4_{15.8}$	$24.3_{15.7}$	$27.9_{16.5}$
Human	$94.1_{5.5}$	$88.8_{10.9}$	$91.3_{7.7}$

Table 13: Human evaluation results for two models and a human-written response. Ratings for each property are averaged across 3 workers, then averaged across questions. Standard deviation of property ratings across questions are shown in underscore.