# LONGEVAL: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization

**Kalpesh Krishna**[♠*] **Erin Bransom**[♢] **Bailey Kuehl**[♢]
**Mohit Iyyer**[♠] **Pradeep Dasigi**[♢] **Arman Cohan**[♢♡] **Kyle Lo**[♢]

[♠]University of Massachusetts Amherst, [♢]Allen Institute for AI, [♡]Yale University
{kalpesh,miyyer}@cs.umass.edu
{erinbransom,baileyk,pradeepd,armanc,kylel}@allenai.org

## Abstract

While human evaluation remains best practice for accurately judging the faithfulness of automatically-generated summaries, few solutions exist to address the increased difficulty and workload when evaluating *long-form* summaries. Through a survey of 162 papers on long-form summarization, we first shed light on current human evaluation practices surrounding long-form summaries. We find that 73% of these papers do not perform any human evaluation on model-generated summaries, while other works face new difficulties that manifest when dealing with long documents (e.g., low inter-annotator agreement). Motivated by our survey, we present LONGEVAL, a set of guidelines for human evaluation of faithfulness in long-form summaries that addresses the following challenges: (1) How can we achieve high inter-annotator agreement on faithfulness scores? (2) How can we minimize annotator workload while maintaining accurate faithfulness scores? and (3) Do humans benefit from automated alignment between summary and source snippets? We deploy LONGEVAL in annotation studies on two long-form summarization datasets in different domains (SQuALITY and PubMed), and we find that switching to a finer granularity of judgment (e.g., clause-level) reduces inter-annotator variance in faithfulness scores (e.g., std-dev from 18.5 to 6.8). We also show that scores from a *partial* annotation of fine-grained units highly correlates with scores from a full annotation workload (0.89 Kendall's $\tau$ using 50% judgments). We release our human judgments, annotation templates, and our software for future research.[1]

## 1 Introduction

Human judgments are considered the gold standard for evaluating model-generated summaries (Kryscinski et al., 2019; Fabbri et al., 2021) and generated text more broadly (Celikyilmaz et al., 2020). Unfortunately, human evaluation tends to be labor-intensive, expensive to scale, and difficult to design. This is problematic as a large number of judged examples is needed to draw statistically significant conclusions about system performances (Wei and Jia, 2021) or correlations between human judgments and automatic metrics (Deutsch et al., 2021). Human evaluation is especially challenging when *long* sequences of generated text need to be evaluated, due to the inherent subjectivity in the task (Karpinska et al., 2021; Clark et al., 2021; Krishna et al., 2021; Goyal et al., 2022).

To better understand the challenges of human evaluation on long-form summaries (150 words or longer), we first conduct a comprehensive survey of 162 publications and preprints on long-form summarization (Section 2). We find that 119 papers (73%) do not perform human evaluation on long-form summaries, while the remaining papers deviate significantly from suggested best practices for reproducibility (Gehrmann et al., 2022). Current human evaluation setups lack standardization in their design decisions (such as annotation granularity), some of which can significantly impact inter-annotator agreement (Section 3.1). Finally, 20 papers explicitly mention human evaluation is expensive, difficult, and time-consuming due to the long length of summaries and source documents.

To move towards a more consistent and efficient human evaluation, we present LONGEVAL, a set of guidelines for human evaluation of faithfulness in long-form summarization (Section 3). We empirically evaluate LONGEVAL using human annotation studies on two long-form summarization datasets: SQuALITY (Wang et al., 2022) and PubMed (Cohan et al., 2018). We provide an overview of our main research questions and findings in Figure 1 and enumerate them here:

[1] https://github.com/martiansideofthemoon/longeval-summarization

*Work done during in an internship at AI2. Details of individual author contributions can be found here.



*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669

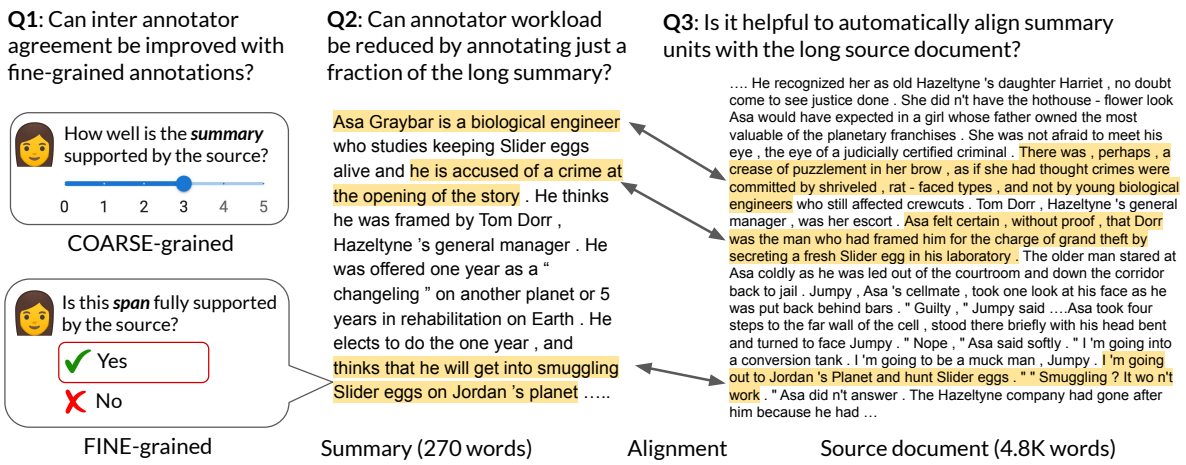


Figure 1: Overview of research questions considered in LONGEVAL. Example summary taken from SQuALITY.

**RQ1**: *Can inter-annotator agreement be improved while evaluating faithfulness of long-form summaries via fine-grained annotations?*

**Finding**: Annotating faithfulness of individual summary clauses and aggregating them leads to significantly higher inter-annotator agreement, compared to the dominant paradigm of evaluating whole summaries at once via Likert ratings (std-dev 18.5 to 6.8 on SQuALITY).

**RQ2**: *Can we reduce annotator workload by partially annotating a long summary while maintaining accurate faithfulness scores?*

**Finding**: Despite annotating a fraction of summary clauses, faithfulness scores under a reduced workload maintain high correlation with those from a full workload (0.89 Kendall's $\tau$ at 50% workload).

**RQ3**: *Do humans benefit from automatically aligning summary units to relevant sentences in the source document?*

**Finding**: Unlike suggestions in prior work on short-form summarization (Hardy et al., 2019; Kryscinski et al., 2020), aligning parts of the summary to source document is only useful when the summary is highly extractive or mostly correct.

Overall, our contributions are:
**(1)** a 162-paper survey of current human evaluation practices in long-form summarization;
**(2)** LONGEVAL, a set of three guidelines for evaluating faithfulness in long-form summarization;
**(3)** an empirical validation of LONGEVAL guidelines on two long-form summarization datasets in different domains (SQuALITY and PubMed);
**(4)** A dataset with 3-way fine-grained human faithfulness judgments for 120 SQuALITY & PubMed summaries annotated using LONGEVAL which can be used for benchmarking automatic metrics.

We open-source our human evaluation data, annotation interface, and code for future research.[1]

## 2 Survey of human evaluation practices

Before discussing LONGEVAL, we first attempt to understand current human evaluation practices in long-form summarization through a comprehensive survey of 162 papers. Our survey reveals several concerning trends: absence of human evaluation, non-reproducible experimental setups, lack of standardization, and complaints of long summaries being challenging and expensive to evaluate. These results show an urgent need to develop more efficient and standardized human evaluation protocols.

*Selection of papers:* We consider existing summarization datasets with an average summary length of at least 150 words, which includes several popular datasets like arXiv (Cohan et al., 2018), BillSum (Kornilova and Eidelman, 2019) and MultiNews (Fabbri et al., 2019); see Table 1 for a full list. For our survey, we select all papers that evaluated summarization models using at least one of these datasets.[2] All of these papers were published between June 2018 and September 2022, after the first long-form summarization datasets were released (PubMed / arXiv). Most of the 162 surveyed papers

[2]We exclude five papers which used long-form summarization data for pre-training only, like Wei et al. (2022).

were published in major NLP/ML venues, but we also include newer preprints from 2022.

**Long-form summaries are rarely evaluated by humans.** We find that 101 out of 162 papers (62%) do not perform any human evaluation. 17 papers (11%) only perform human evaluation on short summaries (datasets like XSUM, Narayan et al., 2018), for which human evaluation is much easier.

**Human evaluation studies of long-form summaries are not reproducible.** We further analyze the 44 papers performing human evaluation of long-form summaries to observe how often they follow reproducible practices from Gehrmann et al. (2022). Overall, we find that most studies do not follow these guidelines. Only 2 of the 44 papers release their raw human annotation data for further analysis. Only 9 papers provide details of their annotator instructions or interface, and just 12 papers perform any kind of statistical analysis, despite most papers annotating less than 50 summaries. While 33 papers report using multiple annotators per summary, only 12 report inter-annotator agreement. Finally, just 14 papers conduct human evaluation on more than one dataset (more statistics in Appendix C).

**Existing human evaluation setups lack standardization.** In Table 2, we catalog the wide spectrum of human evaluation setups in the surveyed papers. 37 papers collect judgments of the full-length summary at once ("COARSE-grained"), while 6 papers collect judgments at a finer granularity such as sentences or entities ("FINE-grained"). Even within a granularity, setups differ: Likert-scale (24 papers), A/B testing (13 papers), binary per-sentence labels (4 papers) are the dominant protocols. In Section 3.1, we will see that this design decision is critical since COARSE annotations have much lower inter-annotator agreement than FINE.[3]

**Human evaluation of long-form summaries is challenging and expensive.** Several of the surveyed papers discuss challenges in human evaluation of long-form summaries. 13 papers mention that expert annotators are necessary for human evaluation of long-form summaries, especially in technical domains like PubMed. 20 papers report that human evaluation of long-form summarization was *time-consuming*, *challenging*, and *expensive*, primarily due to the long length of the summary and source document. To tackle the issue of high annotator workload, we propose a partial annotation method in Section 3.2 and report high correlation to a full workload. Additionally, in Section 3.3 we investigate the usefulness of highlighting sentences to help annotators navigate the long source document. While this has been advocated for in short-form summary evaluation (Hardy et al., 2019; Kryscinski et al., 2020) and used in 3 surveyed long-form papers, we find that it is only helpful when summaries are mostly correct and extractive.

[3]Besides granularity, we also observe a large spectrum of annotator qualifications in our survey, ranging from MTurkers to expert graduates (Appendix C). Since non-experts are known to be unsuitable for this task (Gillick and Liu, 2010; Fabbri et al., 2021), we use experts in our work (Appendix B).

| Dataset | \|source\| (words) | \|summary\| (words) | papers |
|---|---|---|---|
| PubMed (2018) | 3092 | 205 | 59 |
| arXiv (2018) | 5906 | 163 | 55 |
| BillSum (2019) | 1284 | 174 | 19 |
| MultiNews (2019) | 2103 | 263 | 54 |
| GovReport (2021) | 7551 | 547 | 16 |
| BookSum (2021) | 5102 | 505 | 4 |
| SummScreen (2022) | 6965 | 227 | 11 |
| SQuALITY (2022) | 5194 | 227 | 1 |

Table 1: List of long-form summarization datasets considered in our survey along with average source document and summary lengths. Each dataset considered has at least 150 word summaries on average.

| Type of human evaluation | # papers | % papers |
|---|---|---|
| None | 101 | 62% |
| Short-form summaries only | 17 | 11% |
| Likert-scale COARSE-grained | 24 | 15% |
| A/B testing COARSE-grained | 13 | 8% |
| Extrinsic evaluation | 1 | 1% |
| Binary per sentence FINE-grained | 4 | 2% |
| QA-based FINE-grained | 2 | 1% |

Table 2: Human evaluation setup in 162 summarization papers that evaluate long-form summaries. 73% of the papers do not evaluate long-form summaries with humans, while others vary significantly in their setups.

## 3 The LONGEVAL guidelines for faithfulness human evaluation

In Section 2, we report several concerning issues with current human evaluation practices in long-form summarization. To move towards more efficient, reproducible and standardized protocols for human evaluation, we develop the LONGEVAL guidelines (Section 3.1-3.3, see Figure 1 for an overview). We focus on human evaluation of *faithfulness*, which Wang et al. (2022) define as:

> "*Checking the factual errors in the summary, where a factual error is a statement that contradicts the source document, or is not directly stated, heavily implied, or logically entailed by the source document*"

We conduct human annotation studies to empirically motivate LONGEVAL. Our experiments are on two long-form summarization **datasets** spanning diverse domains and levels of abstractiveness:

(1) **SQuALITY** (Wang et al., 2022) is a summarization dataset in the literary domain (avg. summary length of 227 words) where summaries describe the plots of English science fiction stories. SQuALITY is highly abstractive: on average just 16% of bigrams in the summary are present in the source document. We closely follow the human evaluation setup in Wang et al. (2022), and use BART (Lewis et al., 2020) and BART-DPR (Karpukhin et al., 2020) as our summarization models along with human-written summaries.

(2) **PubMed** (Cohan et al., 2018) is a summarization dataset in the scientific domain (avg. summary length of 205 words) that pairs English biomedical articles from PubMed[4] with their abstracts as summaries. Compared to SQuALITY, PubMed is more extractive: 54% of summary bigrams are present in the source. We use BigBird-PEGASUS-large (Zaheer et al., 2020) and LongT5-large (Guo et al., 2022) as our summarization models,[5] along with human written summaries. By default, LongT5 / BigBird were highly extractive compared to human-written PubMed summaries (87% / 74% vs 54% bigram overlap with source). Hence, for half the generations we block 6-grams from being copied from the source,[6] reducing extractiveness to ∼54%. We call this setting "PubMed-ngram-block".

### 3.1 RQ1: Does inter-annotator agreement improve using fine-grained annotations?

In Section 2, we found that the dominant paradigm in literature (37 out of 44 papers) is to evaluate the whole summary at once ("COARSE"-grained, Figure 1 top left). 6 papers instead obtain fine-grained annotations for individual units (e.g., sentences) and average them (FINE, Figure 1 top right). Intuitively, FINE annotation has many advantages for longer summaries — it is less subjective than COARSE, since shorter spans needs to be judged rather than a long summary, and it helps localize model errors. However, the distinction between COARSE and FINE is never justified in literature, and inter-annotator agreement is rarely reported to understand the task subjectivity in each setup. To better understand the tradeoff, in this section we conduct human evaluations annotating the same set of summaries using these two different protocols.

**Task formulation**: Let $F_{\text{summ}}$ denote the faithfulness score of a summary. For COARSE, $k$-point Likert scale ratings are obtained for the summary ($F_{\text{summ}} \in \{0, 1...k\}$), based on the faithfulness definition provided earlier. For FINE, we collect binary judgments of individual units in the summary and average them,

$$F_{\text{summ}} = \frac{1}{|\mathcal{C}_{\text{summ}}|} \sum_{c \in \mathcal{C}_{\text{summ}}} F_c,\ F_c \in \{0, 1\}$$

where $\mathcal{C}_{\text{summ}}$ is a set of units in the summary and $F_c$ is the faithfulness judgment for the unit $c$. In both protocols, the faithfulness score of a system is defined as $\frac{1}{|\mathcal{S}|} \sum_{\text{summ} \in \mathcal{S}} F_{\text{summ}}$ where $\mathcal{S}$ is the set of summaries generated by the system.[7]

While sentences are a popular granularity for FINE (4 of the 6 surveyed papers), we found that summary sentences in both datasets were overloaded with information. Hence, we segment sentences on conjunctions and punctuation to obtain more atomic units as $\mathcal{C}_{\text{summ}}$. These units are often clauses,[8] similar to summary content units (SCUs) in Pyramid (Nenkova and Passonneau, 2004).

**Collecting COARSE annotations**: For SQuALITY, we re-use the annotations provided by Wang et al. (2022) for faithfulness assessments. In their data, three annotators give each summary a 1-100 direct assessment rating (Bojar et al., 2016). Annotators with experience in professional copyrighting and editing were hired on Upwork,[9] and these annotators were also involved in the creation of SQuALITY. Unfortunately, none of the surveyed papers that reported human evaluation results on PubMed

[4] https://pubmed.ncbi.nlm.nih.gov/

[5] LongT5 is the best publicly available PubMed summarizer. BigBird is a popular long-form summarization baseline.

[6] Reducing extractiveness / copying is also a suggestion for fair-use of copyrighted work (Harvard, 2016; UMGC, 2020).

[7] We assume all summary units get an equal weight. However, some units may be more important than others, we discuss this in the Limitations section.

[8] An even finer granularity is entities / numbers. We avoid this due to prohibitive annotation cost on long summaries.

[9] https://www.upwork.com/

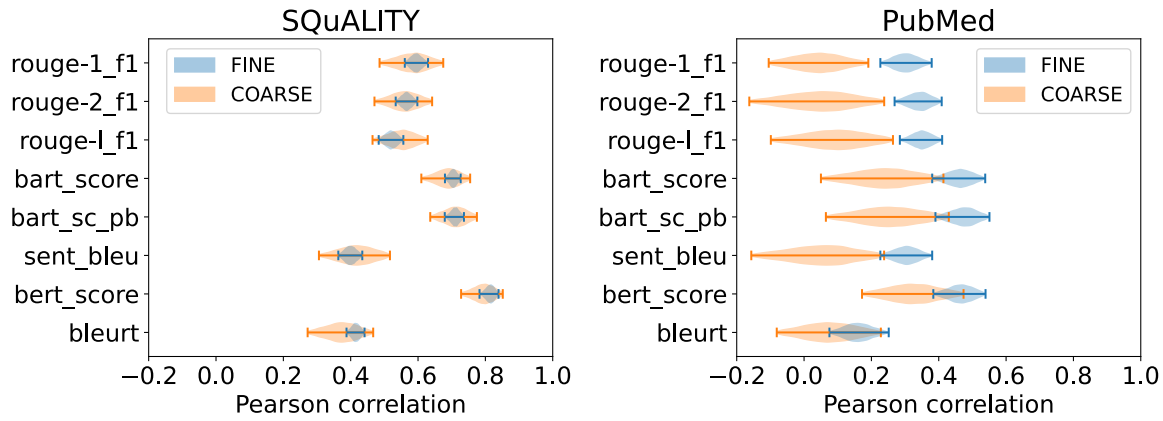


Figure 2: 95% confidence intervals of Pearson correlations between various automatic evaluation metrics and using human evaluation data collected with FINE (blue) and COARSE (orange) annotation methods. In both datasets, FINE annotations lead to much narrower CIs than COARSE annotations. See Appendix G for plot with Kendall's Tau.

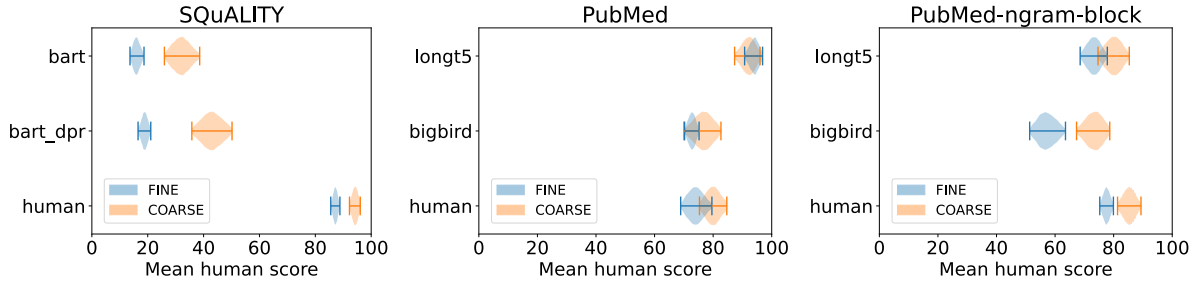


Figure 3: 95% confidence intervals of estimated model performances using FINE (blue) and COARSE (orange) annotation methods. Intervals calculated using bootstrap resampling across annotators (Appendix A). While both annotation granularities lead to similar relative ordering of systems, FINE annotations have narrower confidence intervals. The higher LongT5 score vs human in PubMed is due to highly extractive LongT5 summaries (Section 3).

released their raw human annotations.[10] Hence, we collect our own COARSE evaluations on PubMed summaries on Upwork, using freelancers with professional experience reading and writing research papers (details in Appendix B.2). We collect 3 annotations per summary and use a 5-point Likert scale, the most common choice for COARSE assessment in our survey (18 out of 38 papers). In total, 120 summaries are evaluated.

**Collecting FINE annotations**: For both SQuALITY and PubMed, we collect FINE annotations on Upwork (3 annotators per FINE unit) for the *same set* of 120 summaries evaluated using COARSE annotations. For SQuALITY, we hire freelancers with professional experience in English, creative writing, or education. For PubMed, we hire freelancers with prior experience analyzing biomedical articles. See Appendix B.1 for details of our annotator screening process, compensation, instructions, and screenshots of our annotation interface.

[10]In our email correspondence with authors of these works, they mentioned losing access or compliance issues as reasons for not sharing human evaluations. We received some examples from Guo et al. (2021) and Ju et al. (2021) for reference.

| Dataset | COARSE | FINE |
|---|---|---|
| SQuALITY | 18.5 | **6.8** |
| PubMed | 11.8 | **7.3** |
| PubMed + ngram block | 11.7 | **9.3** |
| Average | 14.0 | **7.8** |

Table 3: Average standard deviation of faithfulness scores across annotators on a 100-point rating scale. Lower variation means higher agreement. Overall, we find that FINE-grained annotations have higher inter-annotator agreement than COARSE-grained annotations. Note that all FINE units of a summary were annotated to obtain these results ($f = 1.0$ in Section 3.2).

**FINE annotations have higher inter-annotator agreement than COARSE annotations. This leads to more confident downstream estimates**. We present our results in Table 3. Overall, we observe that across all settings, FINE annotations have lower

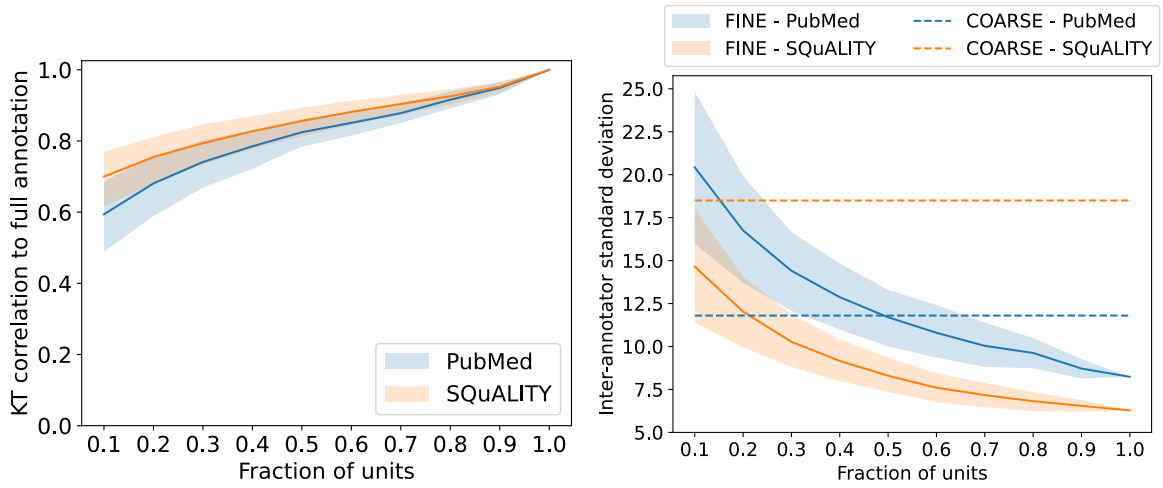


Figure 4: Accuracy and variance after annotating a fraction of units per summary (X-axis) with FINE. Despite annotating just a fraction of the summary, we observe a high segment-level Kendall tau correlation with a full annotation (left). However we observe higher inter-annotator variance as the fraction reduces (right). Confidence intervals shown are 95% and computed across 1000 random subsets (see Appendix F for left plot with Pearson).

standard deviation (and thus higher agreement) in faithfulness scores than COARSE annotations (7.8 vs 14.0 average on 100-point scaled ratings). To illustrate the importance of higher agreement, we measure its effect on two downstream statistics that human evaluation is primarily used for: (1) correlation with automatic metrics; and (2) mean system performance. We adapt the bootstrap resampling analysis[11] of Deutsch et al. (2021) to estimate confidence intervals of these two downstream statistics for COARSE and FINE.

In Figure 2, we plot the 95% confidence intervals of the Pearson correlation of various automatic evaluation metrics against FINE-grained and COARSE-grained human evaluation data. Across both datasets, FINE data leads to much narrower confidence intervals (0.15 vs 0.35 average uncertainty in Pearson correlation on PubMed) for the same number of summaries, implying higher statistical power. In Figure 3, we observe a similar trend with mean system performance. Interestingly, both annotation methods give the same relative ordering of systems (human > bart-dpr > bart for SQuALITY, human > longT5 > BigBird for PubMed-block), confirming the alignment of FINE and COARSE judgments on average.

***Recommendation***: Unlike the dominant trend in prior work, FINE-grained evaluations should be preferred over COARSE grained evaluation for long-form summaries. FINE annotations have lower inter-annotator variance than COARSE annotations and help localize model errors. In our setup we assume all FINE units are equally weighted while aggregating them to the final summary score. Despite this assumption, in our results we observe a consistent relative ordering of systems/metrics between COARSE and FINE annotations. Nevertheless, non-uniform weighing of units is an interesting future work direction; more in the Limitations section.

### 3.2 RQ2: Can we reduce annotator workload by partially annotating a long summary?

In Section 3.1, we found that FINE annotations have lower variance than COARSE annotations. However, long summaries may be composed of several units (sentences or phrases) which each require FINE annotation. This could make FINE annotation very expensive for longer summaries (as also noted in our survey). What if we instead annotate a random subset of units from the summary? While this will lower annotation cost, how accurate would these partial annotations be? We explore this tradeoff by re-using the annotations collected in Section 3.1. For every summary, we randomly sample a fraction of units $f \in \{0.1, 0.2...0.9\}$ and then measure its correlation to the full set of annotations collected. Each annotator gets a different random sample of units for the same summary. In initial experiments, we found that this yielded higher accuracy than when keeping the same set of units per annotator.

[11] We slightly modify the algorithm in Deutsch et al. (2021) for inter-annotator variance, see Appendix A.

**Partial annotation has a high correlation to full annotation, but higher variance**: In Figure 4 (*left*) we plot the segment level Kendall's $\tau$ correlation (relative ordering of summary scores) between a partial annotation and full annotation for different values of $f$. Overall, we observe a high correlation across different values of $f$. Despite annotating just half the summary ($f = 0.5$), in both datasets we observe a high correlation of 0.78-0.89 Kendall's $\tau$ (95% interval) with a full annotation. Does a partial annotation preserve the variance benefits of FINE vs COARSE? In Figure 4 (*right*) we plot the inter-annotator variance for different values of $f$. In both datasets we find that a partial annotation has a higher variance than a full annotation. While for all values of $f$ in SQuALITY we find that FINE annotations still have lower variance than COARSE, in PubMed COARSE has lower variance than FINE for $f <= 0.3$ with 95% confidence.

***Recommendation***: Having annotators judge a random subset of units in a long-form summary is a simple way to reduce FINE annotation cost, and has high correlation with a full annotation. However, it increases inter-annotator variance. Annotating 50% of the summary results in 0.78-0.89 Kendall's $\tau$ correlation, with a 30-40% increase in standard deviation compared to full FINE annotation. Partial annotation may be limited in its ability to identify issues in summaries with very few errors. However, we find that this is not the case in current systems, which are abundant in faithfulness errors.

### 3.3 RQ3: Is it useful to align summary units to sentences in the source document?

So far, we have focused on design decisions on the summary side of evaluation. However, evaluating faithfulness requires a comparison of facts between a summary and a *source document*. Long-form summaries tend to have long source documents (Table 1): 3.1K words for SQuALITY and 5.1K words for PubMed. In Section 2, we found several mentioned human evaluation is challenging since annotators need to read long source documents. Some prior work has suggested highlighting spans in the source document that align with the summary (Hardy et al., 2019; Kryscinski et al., 2020; Vig et al., 2021) as shown in Figure 1. However, these efforts have exclusively focused on news summarization with relatively short source documents, like CNN/DM (804 words) (Nallapati et al., 2016) or XSUM (438 words) (Narayan et al., 2018).

| Algorithm | R@3 | R@5 | R@10 |
|---|---|---|---|
| BM25 (1995) | 0.38 | 0.46 | 0.56 |
| ROUGE-1 (2004) | 0.31 | 0.34 | 0.46 |
| SIM (2019) | 0.37 | 0.52 | 0.60 |
| DPR (2020) | 0.29 | 0.31 | 0.41 |
| BERTScore-DB-XL (2020) | 0.30 | 0.37 | 0.46 |
| SummaC-NLI (2022) | 0.22 | 0.26 | 0.34 |
| MultiVers-FEVER (2022) | 0.47 | 0.58 | 0.71 |
| SuperPAL (2021) | **0.61** | **0.68** | **0.77** |

Table 4: A comparison of algorithms finding the top source document sentences for summary units in SQuALITY. R@$k$ (recall@$k$) denotes the fraction of times the gold sentence was in the top-$k$ predictions.

| Hints | Acc. (↑) | Agree. (↑) | Time (secs) (↓) | |
|---|---|---|---|---|
| | (2-way) | (Fleiss) | All | First 5 |
| None | **93%** | **0.71** | 41.4 | 115.6 |
| SuperPAL | 92% | 0.64 | 48.2 | 84.6 |
| Gold | 92% | 0.63 | **40.4** | **60.4** |

Table 5: Annotator performance (accuracy, agreement, median time) in detecting summary errors with different types of source document highlight hints. Overall, we see little difference across the three settings.

How useful is highlighting based on alignment, or "*hints*", when the spans are chosen from much longer documents?

**What is the best highlighting algorithm?** We conduct a study to identify the alignment algorithm best suited for highlighting hints. We manually annotate 125 FINE units from human-written summaries of the SQuALITY validation split, marking the sentences best supporting them from the source document. We then test several candidate methods for linking summary units to the source document. These include token overlap methods like ROUGE (Lin, 2004), retrievers (Karpukhin et al., 2020), and fact verifiers (Wadden et al., 2022). In Table 4, we find that SuperPAL (Ernst et al., 2021), a weakly supervised linking algorithm, performs best (0.61 recall@3 vs the next best 0.47). To improve precision, we filter matches scoring less than 0.3 on SuperPAL, and show at most five highlights.

**Do highlighted hints improve summary error detection?** To answer this question, we manually perturb 50 FINE summary units in SQuALITY validation summaries, introducing entity errors or negations like Kryscinski et al. (2020). We modify the summary context of the perturbed unit to ensure summaries are self-consistent. Annotators

| Question & TL;DR response | Response Snippets |
|---|---|
| **Q:** Did you find the highlighted hints useful while making your judgment?<br><br>**TL;DR**: 4 out of 5 annotators said **Sometimes**, 1 said **Yes**. More useful for SQuALITY, summary units copied verbatim from source, correct summaries. | "**With summaries that had poor correctness, the hints were often a mess**, and even correct spans had to be carefully checked. In summaries that were more correct, I could often just read the span and remember that it was correct, and then the **hints helped me find the right source position**, or **refresh my memory** about details."<br>"They were more useful when the summary was a **near verbatim source reproduction**."<br>"Yes, they were useful. Often they would highlight the exact passage needed to support the summary span."<br>"In **PubMed, they were a little more chaotic**, even for good summaries."<br>"SQuALITY summaries consisted of sentences or parts of sentences taken straight from the story (wording was exactly as in the text). So hints often lead to the exact place."<br>"For **SQuALITY, they were mostly accurate and helpful**. For PubMed, they were less accurate and relevant." |
| **Q:** Would the highlights have been sufficient to make judgments, or was reading the entire source document necessary?<br><br>**TL;DR**: 3 out of 5 annotators said **No**, 2 said **sometimes in SQuALITY**. Reading the entire document was critical. | "**Reading the entire source document was very helpful** to understand the basic story plot"<br>"Even when the hints were relevant, sometimes **they left out information** (like character name)..."<br>"Initially I tried skimming ... then **concluded it's easier to read the entire document** first."<br>"With SQuALITY there were cases where almost all of the highlights did not make any sense and nothing of that was even mentioned in the story. With **PubMed, it was even more difficult to find hints** that support the text"<br>"**Reading the entire document was essential** to understanding the whole process, the hints in isolation were not good enough. The **hints and the summary often confused similar objects**, especially when pronouns were involved, from different parts of the source. In PubMed a similar thing happened when the source discussed what other papers had done – punctuation, acronyms, and abbreviations played a big role in providing context." |
| **Q:** Did you use Ctrl+F searches in the source document while making judgments?<br><br>**TL;DR**: 4 out of 5 annotators said **Yes**, 1 said **yes only for PubMed**. Ctrl+F helped locate synonyms, entities. | "Yes, **all the time**. It was usually a **safer bet than using the hints**. The hints are given out of context of the whole SQuALITY story. There were a lot of problems with the PubMed hints involving **numbers**, which I often searched for. They were very rarely supported by the document, or contained wrong symbols (= instead of >)."<br>"Yes, mostly in cases the highlight did not support the summary unit partially or entirely."<br>"I used Ctrl+F when looking for very specific words, like **names**. Searching was less helpful when it came to words that had synonyms or emotions."<br>"I did Ctrl+F on keywords taken directly from the summary unit as well as **synonyms and any specific words** that I remembered from the story that could help me get to that place in the source document quickly." |

Table 6: Results and snippets from our questionnaire with FINE annotators. Overall, annotators find hints only sometimes useful, and mention reading the entire source document along with keyword searches.

are shown 50 perturbed and 50 un-perturbed summaries, and asked to annotate whether the summary units are faithful to the source in three settings:[12] (1) no highlighted hints; (2) SuperPAL highlighted hints; (3) gold hints manually annotated by us. In Table 5, we show accuracy, inter-annotator agreement, and median time[13] for each setting.

**Highlighted hints have almost no effect in evaluating long-form summaries**: Surprisingly, we observe that in all three metrics (accuracy, agreement, median time taken), scores are quite similar across the three settings. In fact, the "no-hint" setting scores slightly higher than the SuperPAL hint settings (93% vs 92% accuracy, 0.71 vs 0.64 Fleiss $\kappa$) and takes annotators less time (41.4 vs 48.2 seconds per unit). However, we find that hints helped annotate the first few units of a summary quicker (84.6 secs vs 115.6 secs per unit). We attribute our findings to a *learning effect* over time. FINE annotation of long-form summaries requires annotation of several units for the same document - summary pair. As annotation progresses, annotators get more familiar with the contents of the source document and summary, reducing the need for hints over time. See Appendix E for learning trajectory plots.

**Questionnaire with FINE annotators confirm limited utility of hints**: Our evaluation so far is limited to perturbed human summaries. How effective are hints on model-generated summaries? To answer this, we ask five of our FINE Upwork annotators (from Section 3.1) a set of three questions about their experiences using highlighted hints.[14] Detailed questionnaire results along with answer snippets are shown in Table 6. Overall, annotators find hints were useful only sometimes. Hints were *less useful* when (1) the summary unit was not supported in the source; (2) the summary unit was highly abstractive compared to the source; (3) pronouns, numbers, or abbreviations were involved; and (4) Pubmed summaries were annotated. Almost all annotators said it was necessary to read the entire source document before annotation to get an overall idea of the plot and resolve coreferences. Nearly all annotators used "Ctrl+F" searches along with hints to search for specific keywords while making judgments. This was especially true when

[12] To prevent any bias, each annotator receives only one of these settings for a particular summary.

[13] Calculated using the method in Akoury et al. (2020).

[14] The FINE annotations in Section 3.1 were shown hints in the source document. Since hints may not be helpful, annotators were told not to solely rely on hints for annotation.

the summary unit was incorrect, since the source document had to be thoroughly searched (beyond the hints) before confidently marking "Incorrect".

***Recommendation***: In contrast to recommendations in prior work, automatically highlighted hints are useful only in some specific cases of long-form summarization: mostly correct summaries, almost verbatim copied sentences. Annotators should be instructed to read the entire source document and to not rely solely on highlighted hints, since that could bias their judgments. Based on a small-scale study, we found SuperPAL (Ernst et al., 2021) to be the most accurate method for finding hints, but its performance (61% recall@3) is far from ideal.

### 3.4 To what extent do our findings generalize to short-form summarization?

In this work, we exclusively focus on summarization datasets with an average summary length of at least 150 words. This constraint excludes two popular benchmarks in summarization research over the last five years: CNN/DM (Nallapati et al., 2016) and XSUM (Narayan et al., 2018). How relevant are our research questions (RQs) and findings for these short-form summarization benchmarks?

On average, XSUM (24 words) and CNNDM (60 words) contain much shorter summaries than SQuALITY (237 words). XSUM outputs typically contain only 1 sentence or roughly 2-3 FINE units per summary. This blurs the distinction between FINE and COARSE units, which makes it less useful to study RQ1 in these short-form settings. The shorter length of outputs also implies that evaluation is less expensive and consumes less time, which makes our RQ2 less relevant. Finally, on average, XSUM (440 words) and CNNDM (800 words) also have much shorter source documents than datasets like SQuALITY (5200 words), reducing the need for alignment (the main premise for RQ3). The main motivation behind our study is that human evaluation of long-form summarization datasets like SQuALITY and PubMed is challenging and expensive due to the long length of the generated text. **Overall, our research questions and findings are more relevant for long-form summarization datasets than for short-form summarization datasets like XSUM and CNNDM**.

## 4 Related Work

A large body of recent work has focused on new *automatic* evaluation methods for summarization via NLI-based algorithms (Falke et al., 2019; Laban et al., 2022) or QA-based algorithms (Wang et al., 2020; Fabbri et al., 2022). Our work focuses on the much less studied area of *human* evaluation, the gold standard for developing automatic metrics. A notable effort in this space is the **Pyramid method** (Nenkova and Passonneau, 2004), along with work improving Pyramid efficiency (Shapira et al., 2019; Zhang and Bansal, 2021). Efficient Pyramid-like protocols have been used to collect large-scale datasets human judgments (Bhandari et al., 2020; Liu et al., 2022) in short-form news summarization tasks like CNN/DM. While these efforts focus on salience evaluation and assume access to multiple references, our work focuses on faithfulness and operates in a reference-free setting. Moreover, we focus on *long-form* summarization tasks like SQuALITY and PubMed, which are much more challenging and expensive to evaluate.

Evaluating summary faithfulness relates to **fact verification** (Vlachos and Riedel, 2014), where claim sentences are checked against a large knowledge source (Wikipedia). Prior work (Nakov et al., 2021) attempts to simplify the human fact checking process by methods like knowledge source snippets (Fan et al., 2020), similar to hint highlights (§3.3). Faithfulness in summarization differs from fact verification in three ways: (1) summaries are paragraph-long and contextual compared to single sentence stand-alone claims in fact verification; (2) summaries are grounded to a source document, compared to a large knowledge source in fact verification; (3) summaries are model-generated compared to human-written claims in fact checking datasets (Thorne et al., 2018; Wadden et al., 2020).

## 5 Conclusion

We present the LONGEVAL guidelines, a set of recommendations for moving towards standardized human evaluation of long-form summarization. We empirically analyze each recommendation on two datasets. Overall, we find that (1) FINE-grained annotations have lower inter-annotator variance than COARSE-grained annotations; (2) partially annotating a summary reduces annotator workload while maintaining accuracy; (3) highlighting hints in the source document has limited usefulness for evaluating long-form summaries. As future work, we plan to conduct experiments on other aspects of summarization evaluation like salience and coherence.

## Limitations

Human evaluation is a noisy process with many **confounding variables**. Some of these variables were kept constant among experiments on a dataset, but modifying them could change the trends in the results. These include: (1) number of annotations per summary; (2) the specific annotation interface used; (3) granularity for FINE evaluation (sentences vs phrases); (4) Number of points in the Likert scale for COARSE evaluation; (5) set of summarization systems evaluated; and finally (6) relative (eg: A/B tests) vs absolute evaluation (eg: Likert), which has been discussed in Tang et al. (2022) for short-form news summarization datasets like CNN/DM.

Our paper is **limited to faithfulness evaluation**, but summaries are typically evaluated for salience, fluency, coherence as well (Fabbri et al., 2021). While fluency may be less of an issue due to large-scale language model pretraining (Dou et al., 2021), coherence and salience are important aspects to evaluate especially in long-form summarization (Goyal et al., 2022). Our findings may not generalize to evaluation of coherence or salience.

Our experiments in Section 3.1 **assigned an equal weight** to each FINE unit while calculating the overall score of the summary. However, the faithfulness of some FINE units may be more important than others. A non-uniform weighing of FINE units may be a good strategy if there is a notion of how critical a particular unit is for a summary's correctness. For example: (1) PICO units are critical in medical summaries (DeYoung et al., 2021); (2) the Pyramid scheme (Nenkova and Passonneau, 2004) uses a reference frequency-based unit importance, assuming access to multiple gold references. However, a consistent notion of importance is difficult to establish across different domains, and also depends on an individual consumer's preferences. Designing non-uniform weighing schemes is an interesting direction for future research.

## Ethical Considerations

All experiments involving human evaluation in this paper were exempt under institutional IRB review. We fairly compensated each Upwork freelancer involved in this study, at a rate of 15-20$ per hour (respecting their suggested Upwork hourly wage). For each round of annotation, we estimated the average amount of time the task would take (by running pilots among ourselves), and provided annotators with the estimated time requirement. Most freelancers finished the task within the time window, but sometimes exceeded it by 0.5-1 hr. We compensated freelancers based on the actual time they took and their hourly wage, rather than a fixed amount per annotation.

## Acknowledgments

First and foremost, we would like to thank all the nine Upwork freelancers who contributed human annotations to this project. We are very grateful to Yixiao Song, Alex Wang, John Giorgi, Dustin Wright, Yulia Otmakhova, Daniel Deutsch, Arie Cattan, Shiyue Zhang, Tanya Goyal, Greg Durrett, Marzena Karpinska, Ankita Gupta, Nader Akoury and the Semantic Scholar team for several useful discussions at various points during the project. This work was mostly done while Kalpesh Krishna (KK) was an intern at the Allen Institute for Artificial Intelligence. KK was partly supported by a Google PhD Fellowship awarded in 2021.

**Author Contributions:** Kalpesh Krishna led the project and performed all the technical contributions including literature review, dataset collection and processing, model implementation, annotation interface development, running experiments, and data analysis. Kalpesh also contributed to project scoping and ideation and led the writing of the paper. Erin Bransom and Bailey Kuehl helped with obtaining human judgements, including piloting the task and giving feedback, performing the annotation themselves, and hiring and managing annotators on Upwork. Pradeep Dasigi, Arman Cohan, and Kyle Lo were mentors of the project during and after Kalpesh's internship, contributing equally to project scoping, experimental design, ideation and direction throughout the course of the project and paper writing. Mohit Iyyer provided mentorship after the internship, in particular providing important feedback and direction on data analysis and contributing to paper writing.

## References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

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.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

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.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. 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.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online. Association for Computational Linguistics.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech*

*and Language Data with Amazon's Mechanical Turk*, pages 148–151.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. Snac: Coherence error detection for narrative summarization. *arXiv preprint arXiv:2205.09641*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

OGC of Harvard. 2016. Copyright and fair use.

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.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4091–4098, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. 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.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, 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 Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. 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 arXiv:2105.08209*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, 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.

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.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

Libraries at UMGC. 2020. Copyright and fair use guidelines.

Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. 2021. SummVis: Interactive visual analysis of models, data, and evaluation for text summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 150–158, Online.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R Bowman. 2022. Squality: Building a long-document summarization dataset the hard way. *arXiv preprint arXiv:2205.11465*.

Matthijs J Warrens. 2010. Inequalities between multi-rater kappas. *Advances in data analysis and classification*, 4(4):271–286.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Johnny Wei and Robin Jia. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6840–6854, Online. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

# Appendix

## A Bootstrap analysis of inter-annotator variance

We utilize the bootstrap resampling (Tibshirani and Efron, 1993) technique described in Deutsch et al. (2021) to estimate confidence intervals for human evaluation data. At a high level, bootstrap resampling helps capture the uncertainty in a downstream test statistic by repeatedly sampling from the data with replacement. We consider two downstream test statistics in our work — (1) average system level performance; (2) correlation of human judgements to automatic metrics.

While Deutsch et al. (2021) were primarily interested in uncertainty due to the specific instances and systems evaluated, our goal is to capture uncertainty due to the inter-annotator variance. Hence unlike Deutsch et al. (2021), we sample with replacement from the set of *annotators* for every instance. Our precise formulation can be found in Algorithm 1, which operates on a $X \in \mathbb{R}^{N \times M}$ matrix of human annotations where $N$ is the number of summaries, and $M$ the number of annotators.

**Algorithm 1** Bootstrap Confidence Interval

**Input:** $X \in \mathbb{R}^{N \times M}, k \in \mathbb{N}, \alpha \in [0, 1]$.
$N$ is summaries, $M$ is annotators
**Output:** $(1 - \alpha) \times 100\%$-confidence interval

samples $\leftarrow$ an empty list
**for** $k$ iterations **do**
  $X_s \leftarrow$ empty $N \times M$ matrix
  **for** $i \in \{1, \ldots, N\}$ **do**
    $D \leftarrow$ samp. $\{1, \ldots, M\}$ w/ repl. $M$ times
    **for** $j \in \{1, \ldots M\}$ **do**
      $X_s[i, j] \leftarrow X[i, D[j]]$
    **end for**
  **end for**
  Calculate test statistic on $X_s$ and append to samples
**end for**
$\ell, u \leftarrow (\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of samples
**return** $\ell, u$

## B Human evaluation details

### B.1 FINE-grained evaluations of SQuALITY and PubMed summaries

We interviewed a total of 9 Upwork freelancers for the position, offering a compensation of $15-16.5 / hr (depending on their Upwork hourly rate). The screening procedure involved a qualification task on synthetically perturbed summaries from the SQuALITY dataset validation split. Similar to the final annotation task, annotators were shown a

| | F-$\kappa$ | R-$\kappa$ | all agree |
|---|---|---|---|
| Random | 0.00 | 0.00 | 25% |
| SQuALITY | 0.74 | 0.76 | 82% |
| PubMed | 0.53 | 0.65 | 74% |

Table 7: Fleiss kappa (F-$\kappa$), Randolph kappa (R-$\kappa$), and agreement scores of our FINE annotation per summary unit. All $\kappa$ scores are well above a random annotation baseline, indicating good agreement.

highlighted clause from the summary, and asked to mark whether or not it is supported by the source document. 50% of the clauses were synthetically perturbed (via negation or entity swapping as in Kryscinski et al., 2020) and manually checked to ensure they were not supported by the source document. A total of 6 freelancers scored 85% or better, and were recruited for the main set of experiments. All 9 freelancers were compensated for the screening round at the rate of 15$ USD / hr.

All six hired annotators are native or bilingual English speakers. All annotators have completed a degree at the undergraduate level and three also have Masters degrees, with the most common focuses of the degrees being English/creative writing and education. The annotators' common professional experiences include copywriting, editing, proofreading, writing, and teaching. Finally, for PubMed annotations we re-hired three annotators from the pool of six SQuALITY annotators who mentioned they had experience reading and analyzing biomedical articles. These three annotators were provided with an additional bonus of $30 after they completed all annotations.

Annotators are provided with a detailed annotation guideline along with examples of faithfulness (Table 10). Our guidelines are mostly consistent with a recently proposed set of guidelines for checking attribution in text generation (Rashkin et al., 2021). The final annotation interface is implemented in AMT Sandbox, as shown in Figure 8.

**Inter-annotator agreement (binary)**: Much of the analysis in Section 3 uses standard deviation across summaries scores to measure inter-annotator agreement. However, another way to calculate inter-annotator agreement for FINE annotations is measuring agreement on individual units which received a Yes / No judgment. In Table 7 we show these inter-annotator agreement statistics. We measure Fleiss Kappa (Fleiss, 1971), Randolph Kappa (Randolph, 2005; Warrens, 2010), and the fraction of sentence pairs with total agreement.[15] In the table we can see all agreement statistics are well away from a uniform random annotation baseline, indicating good agreement.

### B.2 COARSE-grained evaluation of PubMed summaries

None of the surveyed papers evaluating PubMed summaries with humans released their human evaluation data. Hence, we decided to collect our own COARSE annotations. Since FINE annotations (Section B.1) may have biased our original set of annotators, we hire three new annotators to perform overall assessments on a 5-point Likert scale. In other words, we use a "between-subject" experiment design to compare FINE against COARSE.

We hired three freelancers on Upwork, all of whom have extensive professional experience reading research papers (two of them had PhDs in biomedical fields). All annotators were compensated at a rate of 20$ USD / hr, their hourly rate on Upwork. All three annotators had been previously screened and hired by us for different projects in the past. Two of them had assisted us in an annotation task involved reading short summaries of biomedical academic papers and evaluating them for fluency, accuracy, correctness.

Annotators are provided with a detailed annotation guideline along with examples of faithfulness (Table 11). Our guidelines are mostly consistent with a recently proposed set of guidelines for checking attribution in text generation (Rashkin et al., 2021). The final annotation interface is implemented in LabelStudio, as shown in Figure 9.

### B.3 Crowdworkers or expert annotators?

Several prior works have raised the issue of low inter-annotator agreement and poor accuracy with non-expert annotators (eg: MTurk crowdworkers) in human evaluation of summarization (Gillick and Liu, 2010; Fabbri et al., 2021; Falke et al., 2019) and open-ended long-form generation (Karpinska et al., 2021; Clark et al., 2021). In our survey (Table 9), we found the type of annotators used in long-form summarization is often not specified (16 / 43 papers). Among other papers, 10 papers use non-experts while 17 papers use expert annotators (often graduate students).

[15] The $\kappa$ scores are measured using the library https://github.com/statsmodels/statsmodels.

Overall, we echo the concerns with non-expert annotators and recommend hiring freelancers on Upwork (or experts) who are well-versed with the domain for annotation. In initial experiments, we attempted to recruit Amazon Mechanical Turk crowdworkers filtered by the "Master's qualification" and having a 90%+ approval rating. In our qualification task of error detection in synthetically perturbed SQuALITY summaries, MTurkers scored just 62% (binary classification) with a three-annotator Fleiss $\kappa$ of 0.15. On the other hand, Upwork freelancers (with professional writing experience) an accuracy 90% with a high inter-annotator agreement (Fleiss $\kappa = 0.71$).

## C Additional Survey Statistics

In Table 8 and Table 9 we document some additional statistics for the 44 papers conducting human evaluation of long-form summarization.

| Best practice | # papers |
|---|---|
| Raw human evaluation data released | 2 / 44 |
| Interface or instructions provided | 9 / 44 |
| Inter-annotator agreement reported | 12 / 44 |
| Statistical analysis conducted | 12 / 44 |
| Multiple datasets are human evaluated | 14 / 44 |
| Multiple annotators per summary | 33 / 44 |
| Annotator background reported | 33 / 44 |
| Specific summary aspects evaluated | 42 / 44 |

Table 8: Fraction of surveyed papers following the best practices recommended by Gehrmann et al. (2022). We include only the 44 papers here which conducted a human evaluation of long-form summarization.

| Type of annotator | # papers |
|---|---|
| No details specified | 11 / 44 |
| Native English speaker** | 5 / 44 |
| Mechnical Turk crowdworker | 9 / 44 |
| Non-expert volunteers | 1 / 44 |
| Extensive prior experience** | 3 / 44 |
| Graduate students / researchers | 13 / 44 |
| Upwork freelancers | 2 / 44 |

Table 9: The types of annotators used across different long-form summarization papers. ** - No additional details were specified.

## D Automatic summarization metrics used for evaluation

The following metrics are considered while measuring Pearson's correlation with our human evaluation data (Figure 2) — ROUGE-1 / 2 / F (Lin, 2004), BARTScore / BARTScore-Parabank (Yuan et al., 2021), Sentence-BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). A number of metrics were calculated using the SacreROUGE repository (Deutsch and Roth, 2020).

## E Learning effect while annotating long-form summaries

In Section 3.3 we discussed a learning effect where annotators get more familiar with the contents of a source document as they annotate more FINE-grained units in a long-form summary. To better understand this effect, in Figure 5 we plot the average time taken by annotators as they progress in their annotation of a summary. Overall, we find that annotators get significantly faster in annotating the summary after the first 20% units. We hypothesize that annotators get pretty familiar with the general topics in the source document after the first few annotations, speeding up subsequent annotations.

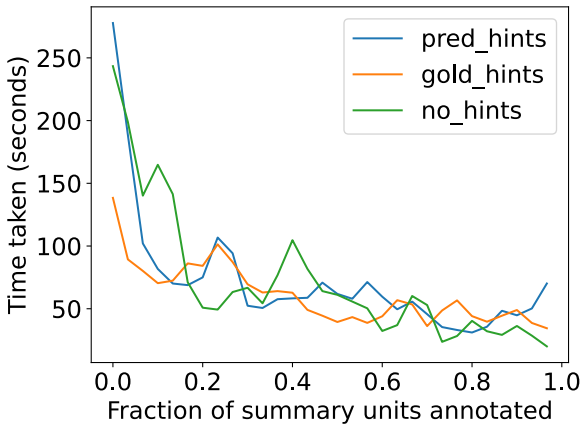


Figure 5: Learning effect over time while evaluating long-form summaries with FINE annotation. As the annotators evaluate more summary units, they learn the document better and are much faster at annotation irrespective of whether hints are shown to them.

## F Partial summary annotation with pearson correlation

See Figure 6.

## G Metric correlations using Kendall's Tau

See Figure 7.

In this task, you will be shown a long document ("Source Document") and its Summary. A span of text will be highlighted in the summary, and the goal is to check if this span is factually supported by the source document. You will need to choose one of two options:

1. **Yes**: if ***all*** the facts in the highlighted summary span are supported by the source document
2. **No**: if the highlighted summary span presents some information that is not supported by the source document (either a direct contradiction, or not present)

In addition to the source document, you will be provided with some highlighted text ("hints") in the source document which may help you in making a decision. Press the "Next Hint" button to scroll through the highlighted hints. Source document hints may or may not be helpful. Do not make a judgment solely based on these hints. Skim through the source document yourself / search for keywords with Ctrl + F if the hints are not helpful.

Below you can find some short representative **examples**.

**Example 1**
Summary (only highlighted span shown) = ... Retief is not Lemuel's cousin. ...
Source Document (snippets shown) = He eyed Retief ... "He ain't no cousin of mine," Lemuel said slowly.
Supports = Yes

**Example 2**
Summary (only highlighted span shown) = ... Lemeul knocks down Retief. ...
Source Document (snippets shown) = Retief's left fist shot out, smacked Lemuel's face dead center. He stumbled back, blood starting from his nose; ... He caught himself, jumped for Retief ... and met a straight right that snapped him onto his back: out cold. "Wow!" said Potter. "The stranger took Lem ... in two punches!"
Supports = No (Reason: Retief knocks down Lemeul, not the other way around.)

**Example 3**
Summary (only highlighted span shown) = ... Potter and his team do not trust the Embassy. ...
Source Document (snippets shown) = Lemme up. My name's Potter. Sorry 'bout that. I figured it was a Flap-jack boat; looks just like 'em . He waved a hand toward the north, where the desert lay.
Supports = No (Reason: The claim is irrelevant to the evidence.)

Table 10: Annotation guidelines provided to annotators for FINE-grained evaluation of SQuALITY and PubMed summaries. (Appendix B.1).

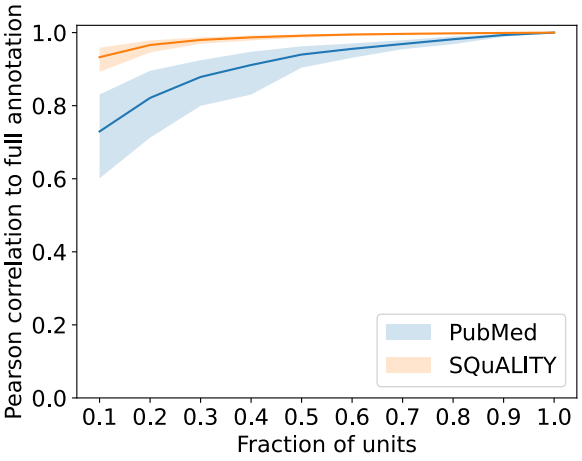


Figure 6: A version of Figure 4 using Pearson correlation instead of Kendall Tau correlation.

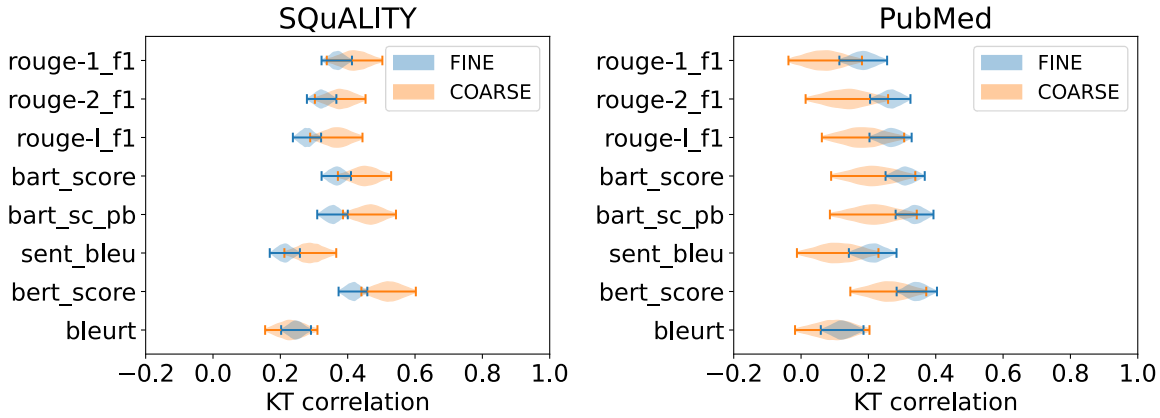


Figure 7: A version of Figure 2 using Kendall's Tau correlation instead of Pearson's correlation.

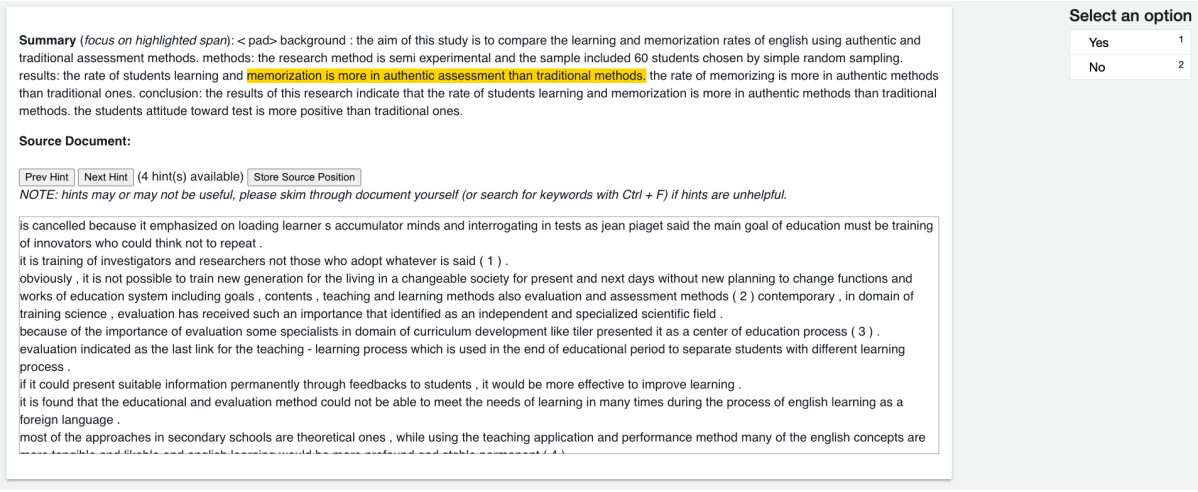


Figure 8: The AMT Sandbox annotation interface used for FINE evaluation of SQuALITY and PubMed summaries (Appendix B.1).

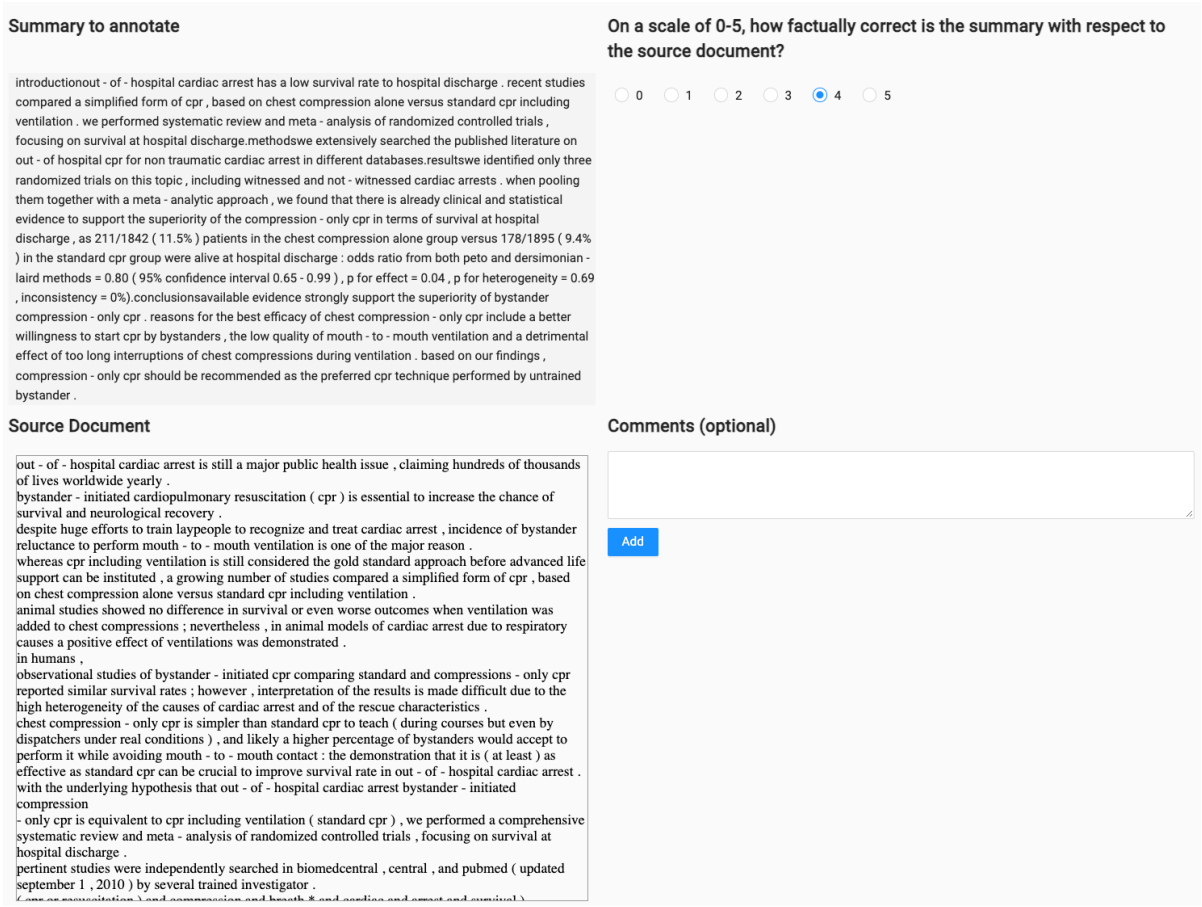


Figure 9: The LabelStudio annotation interface used for COARSE evaluation of PubMed summaries (Appendix B.2).

Instructions for Likert-scale evaluation. Please read all instructions before starting the annotation.

**Setup**
1. Start by signing up on Label Studio, you will need to provide an email ID and password. It's okay to use a non-existent throw-away email ID here. Also, do not use any personal / sensitive passwords (but make sure to remember your email / password for logging in next time!). Click on the box saying "<your name> — Summarization Evaluation"
2. In this batch a total of 30 summaries need to be evaluated. Every three consecutive rows are different summaries of the same source document. You can evaluate a summary by clicking on a row, and annotating it. Optionally, you can click on "Label All Tasks" at the top of the screen.

**Annotation Task**
Each summary needs to be evaluated for its "correctness". You need to provide a 0-5 judgment for the entire summary, where "correctness" can be defined as, "The absence of factual errors in the summary, where a factual error is a statement that contradicts the source document, or is not directly stated, heavily implied, or logically entailed by the source document". For example,
Source Document (snippet shown) = ..... Vitamin C was discovered in 1912, isolated in 1928, and, in 1933, was the first vitamin to be chemically produced. It is on the World Health Organization's List of Essential Medicines. Vitamin C is available as an inexpensive generic and over-the-counter medication. Partly for its discovery, Albert Szent-Györgyi and Walter Norman Haworth were awarded the 1937 Nobel Prizes in Physiology and Medicine and Chemistry, respectively. Foods containing vitamin C include citrus fruits, kiwifruit, guava, broccoli, Brussels sprouts, bell peppers, potatoes, and strawberries. Prolonged storage or cooking may reduce vitamin C content in foods. ....
Summary 1 (snippet shown) = ... Chicken contains vitamin C ...
Summary 2 (snippet shown) = ... Albert Szent-Györgyi won the 1955 Nobel Prize for discovering Vitamin C ...
Summary 3 (snippet shown) = ... Vitamin C was the first chemically produced Vitamin ...
Summary 4 (snippet shown) = ... Apple contains vitamin C ...
Errors marked in red. Here, the snippets for summary 1 are incorrect, summary 2 partially correct, and summary 3 completely correct with respect to the source document. Summary 4 is incorrect with respect to the source document (since it's never discussed), but a globally correct fact. You should treat such a summary as incorrect since it is not mentioned in the source document.
(This is an illustrative example only, the actual annotation task has much longer summaries / source documents.)
The rating scale is from 0 to 5, where 0 is the lowest possible rating (most or all of the summary is wrong / irrelevant to the source document), and 5 is the highest rating (most or all of the summary is correct).
While it is compulsory to provide a judgment from 0 to 5 for each summary, you can optionally provide additional comments in your annotation. For instance, if the judgment needs to be more nuanced than a 5-point scale, you prefer to mark something like "3.5", or you would like to add some other notes about your judgment.
Press "Submit" after you have provided your annotation.

**Suggested workflow**
Every three consecutive rows contain different summaries for the same source document. We suggest the following workflow while annotating documents —
1. Spend the first 15 minutes reading the source document and getting a general sense of the facts mentioned in the document.
2. Spend 5 minutes to read and annotate the summaries in each of the three consecutive rows which correspond to the same document. Add optional comments / notes if necessary.
3. In the last 5 minutes, re-calibrate your ratings across the three rows if needed (for instance, you significantly preferred the correctness of summary 1 vs summary 2, but you gave it the same rating in the initial pass). Add optional comments / notes if necessary.
Following this workflow, it should take 35 minutes to annotate each set of 3 rows. For 30 rows, this should take 6 hrs.

Table 11: Annotation guidelines provided to annotators for COARSE evaluation of PubMed summaries (Appendix B.2).