Concise Answers to Complex Questions: Summarization of Long-form Answers

Abhilash Potluri* **Fangyuan Xu*** **Eunsol Choi**
Department of Computer Science
The University of Texas at Austin
{acpotluri, fangyuan, eunsol}@utexas.edu

Abstract

Long-form question answering systems provide rich information by presenting paragraph-level answers, often containing optional background or auxiliary information. While such comprehensive answers are helpful, not all information is *required* to answer the question (e.g. users with domain knowledge do not need an explanation of background). Can we provide a concise version of the answer by summarizing it, while still addressing the question? We conduct a user study on summarized answers generated from state-of-the-art models and our newly proposed extract-and-decontextualize approach. We find a large proportion of long-form answers (over 90%) in the ELI5 domain can be adequately summarized by at least one system, while complex and implicit answers are challenging to compress. We observe that decontextualization improves the quality of the extractive summary, exemplifying its potential in the summarization task. To promote future work, we provide an extractive summarization dataset covering 1K long-form answers and our user study annotations. Together, we present the first study on summarizing long-form answers, taking a step forward for QA agents that can provide answers at multiple granularities.

1 Introduction

Long-form answers (Fan et al., 2019), as compared to span-based short answers (Rajpurkar et al., 2016), can provide comprehensive answers to a broader set of questions (Cao and Wang, 2021; Fan et al., 2019). While providing comprehensive information in multiple sentences is helpful, users often prefer short and concise answers to their questions when possible (Choi et al., 2021). Today's search engines already present concise answers by highlighting the most relevant parts from the passage excerpts. In this paper, we present the first study on summarizing long-form answers.

Summarizing long-form answers introduces a new challenge in addition to the faithfulness and fluency challenges of generic summarization which mostly focus on news articles (Nallapati et al., 2016; Narayan et al., 2018): the summary output should still provide a reasonable answer to the original question. We take inspiration from a recent study (Xu et al., 2022) that reports that up to 40% of sentences in long-form answers contain non-essential information, such as providing background information or examples (Wang et al., 2022), which demonstrates the potential for compressing long-form answer.

We first aim for an extractive summarization model and collect sentence-level annotations on long-form answers, where annotators identify sentences that address the question directly and can serve as the "summary".[1] We collect a dataset covering 1,134 examples, each consisting of a question, a long-form answer, and a set of summary sentences. To improve the extractive summaries collected, we propose a simple and yet novel summarization approach, **extract-and-decontextualize**, which first extracts summary sentences and rewrites them to stand-alone (Choi et al., 2021). Compared to abstractive summarization models trained on noisy distantly supervised datasets (e.g. CNN/DM (Nallapati et al., 2016) and XSum (Narayan et al., 2018)) which encourage paraphrasing but also hallucinations (Kryscinski et al., 2019; Cao et al., 2018; Kang and Hashimoto, 2020), decontextualization makes minimal edits to the original sentence, preserving its meaning while improving its fluency.

How well do summarization approaches perform in this new domain – can generated summaries provide fluent, adequate answers to the questions, while preserving the semantics of the original long-form answers? We evaluate fine-tuned abstrac-

*Equal contribution.

[1]This is a simplified annotation task compared to the original discourse study of Xu et al. (2022).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9709–9728

Input	System	Summarized Answer	Adequacy	Faithful
Q: Why does car sickness seem to hit the hardest when you look down at your phone, book, etc.? **A:** The brain perceived motion because it receives information from the eyes, ears, and muscles. When these parts send conflicting information, the brain doesn't know which is right and which is wrong, and this is what causes motion sickness. An example of this is when reading a book while you are in a moving car. To your eyes, the book is stationary while your inner ear and the rest of your body can feel a sense of motion. This would likely cause car sickness.	Abstractive	The brain gets confused when it receives conflicting information about motion from different parts of the body, and this can cause car sickness.	Yes	Yes
	Gold Extractive	When these parts send conflicting information, the brain doesn't know which is right and which is wrong, and this is what causes motion sickness.	Partially	Yes
	Decontext	When -these parts +the eyes, ears, and muscles send conflicting information, the brain doesn't know which is right and which is wrong, and this is what causes motion sickness.	Yes	Yes
Q: How come Obama during his supermajority in both houses wasn't able to pass any legislation he wanted? **A:** 1) Senators are normally seated in January. [. . .]Al Franken (who would've been #60) was not seated until July 7.2) Ted Kennedy was dying and had not cast a vote since[. . .]Note that Sept 24-Feb 4 is about 20 working days, due to recess and holidays.3) So, for about 20 working days, the Senate Democrats could have broken a filibuster if you could get every single one of them to agree on something. [. . .] This did not go well.	Abstractive	The Senate Democrats were unable to pass any legislation during Obama's supermajority due to a lack of 60 votes needed to break a filibuster, due to Al Franken not being seated until July 7 and Ted Kennedy's death in August 2009.	Partially	Yes
	Gold Extractive / Decontext	So, for about 20 working days, the Senate Democrats could have broken a filibuster if you could get every single one of them to agree on something.	No	No

Table 1: We present two examples of questions, long-form answers, their summarized answers produced by different systems, and human evaluation results ("summary adequacy" and "faithfulness"). We highlight the gold extractive summaries we collected.

tive summarization model (Zhang et al., 2019), prompted large language model (GPT-3) (Brown et al., 2020), and our extract-and-decontextualize approach with a user study. Table 1 shows two examples from our user study. We find vanilla extractive approach, even with gold sentences, presents inadequate summaries but decontextualizing them makes them on par with GPT-3 abstractive answers. While none of the systems consistently present high-quality summaries (GPT-3 records a 67% success rate), most questions (95%) have at least one system that can generate a valid summary, showing the potential for successful compression of long-form answers. Together, we present the first corpus and study on summarizing long-form answers, opening doors for developing more flexible QA systems which provide answers with varying amounts of information. We release our data, code, and user study templates at https://github.com/acpotluri/lfqa_summary.

2 Background and Motivation

The focus of our study is to find a **concise** answer to a complex question (Fan et al., 2019). One way to generate a concise answer is through controllable generation, where the long-form QA model is instructed to generate an answer given a pre-specified length. However, long-form question answering remains challenging, both in terms of modeling (Krishna et al., 2021) and reliable evaluation (Xu et al., 2023). Existing models often hallucinate (Krishna et al., 2021; Liu et al., 2023) even when paired with relevant evidence documents. Instead of generating a concise answer from scratch, we summarize an *existing* long-form answer, leveraging a large amount of user-written long-form answers often in community-driven QA forums like ELI5 in Reddit.

How feasible would it be to summarize existing long-form answers? Xu et al. (2022) conducted an in-depth study on the structure of such long-form answers, assigning one of six functional roles (answer, answer summary, organizational sentence, auxiliary information, and example) to each sentence in long-form answer. The study suggests sentences corresponding to "answer summary" captures the salient information and "often suffice by themselves as the answer to the question." Furthermore, they suggest up to 40% of sentences belongs

to roles (e.g., auxiliary information) that are not necessary to answer the question, suggesting summarizing existing answer is viable. We follow their study and collect larger-scale data focusing on the "answer summary" role to study the summarization of long-form answers.

Summarizing existing answers will support providing a consistent answer set of different granularities, where the users can *expand* condensed answer to see a more detailed version of the same answer. Consistent answers at multiple granularities are harder to enforce with a controllable generation approach. For instance, if we generate a five-sentence answer from the raw evidence set, the five-sentence answer **can** contain information absent in the ten-sentence answer.

Lastly, retrieval-augmented long-form QA models (Nakano et al., 2021) resemble query-focused summarization. Query-focused summarization (Xu and Lapata, 2020; Kulkarni et al., 2020) often studies challenging multi-document settings, where the input text is summarized focusing on a particular query, provided at inference time content control. A difference to our setting is that a long-form answer is *written* for the question q, presenting already synthesized information tailored for the question.[2]

3 Extractive Summary for Long-form Answers

We first introduce our annotation task of identifying key sentences for long-form answers, which will be used as an extractive summary. Extractive summaries allow easier data collection and evaluation but can suffer from disfluency and incoherence. Thus, we manually evaluate our collected gold extractive summaries in Section 5.

3.1 Task

Given a question q and its long-form answer consisting of n sentences $a_1, a_2, ...a_n$, the model makes a binary decision on whether each sentence a_i should be included in the summary. This setup differs from general summarization in having question q as an additional input.

3.2 Source Data

We use long-form answer data, (question, answer) pairs, from prior study (Xu et al., 2022) which compiled three existing LFQA datasets. **ELI5** (Fan et al., 2019) consists of question answer pairs extracted from the subreddit *Explain Like I'm Five*. **Natural Questions (NQ)** (Kwiatkowski et al., 2019): NQ contains Google search queries as the questions, paired with paragraph-level answers from Wikipedia passages identified by annotators. **WebGPT** (Nakano et al., 2021) contains answers written by trained human annotators, with the questions sourced from ELI5. The annotator first searches for related documents using a search engine and then constructs the answers with direct references to those documents. We only take answers that passed their validity annotation, which excludes questions with false presupposition, ill-formed queries, and answers that do not provide valid answers. Their preprocessing step also filters answers with more than 15 sentences or less than 3 sentences.

3.3 Annotation Task

Given a question and its long-form answer, annotators select a set of summary sentences containing salient information addressing the question. The annotator interface and instructions are in the appendix. As saliency is somewhat subjective, we collect three-way annotations for each example. We recruited crowd workers from Amazon Mechanical Turk. We recruited workers from English-speaking countries, with at least a 95% acceptance rate on 1000+ HITs. Each worker was paid $0.50 per annotation, translating to an hourly rate of $15. We recorded reasonable agreement (Fleiss' Kappa 0.53) for the annotations.[3]

3.4 Dataset Statistics

Table 2 contains our collected dataset statistics, comparing it to a popular news summarization dataset (Nallapati et al., 2016) and a query-focused summarization dataset, AQuaMuSE (Kulkarni et al., 2020). To compute the summary length in our dataset, we randomly choose one of three summary annotations. The average number of sentences chosen as summaries by a single annotator was 1.6 out of 6.2 sentences in long-form answers. The statistics show that our data handles shorter texts and compress less than existing datasets. On average,

[2]This is true for two out of three datasets (ELI5/WebGPT, 82% of our data) we study. In NQ, the paragraphs are written independently, representing the QFS setting.

[3]Xu et al. (2022) hired expert annotators (undergraduate linguistics students), as they required annotators to provide sentence-level labels among six functional roles. The expert annotators reached a similar agreement (0.52 Fleiss' kappa) for the "summary" role.

	#	$\|q\|$	$\|d\|$	$\|s\|$	$\frac{\|s\|}{\|d\|}$
News dataset					
CNN/DM	312k	-	810 (39.8)	56 (3.7)	0.09
Query-Focused summarization dataset					
AQuaMuSe	5.5k	9	9k (0.4k)	106 (3.8)	0.02
LFQA datasets					
ELI5	834	16	113 (6.5)	32 (1.6)	0.33
NQ	202	10	140 (5.3)	47 (1.5)	0.36
WebGPT	98	15	117 (5.6)	44 (1.9)	0.39
All	1,134	15	118 (6.2)	35 (1.6)	0.33

Table 2: Summarization dataset statistics, showing the number of examples (#), the length of question q, document to summarize d, and summary s. For length, we report the average number of tokens and the average number of sentences in the parenthesis.

long-form answers were compressed to about one-third of their original length, with a slightly higher compression rate for ELI5 answers. This aligns with the prior discourse study (Xu et al., 2022) which reports ELI5 contains sentences that serve other functional roles (like providing an example) more frequently (23% compared to 5% and 8% in NQ/WebGPT datasets), neither of which are likely to be included in the summary.

3.5 Automatic Extractive Summarization

Having collected a new dataset, we evaluate existing extractive summarization models on it. Is it easy for models to identify key sentences from long-form answers?

Setting We aggregate all data from three datasets (ELI5, NQ, WebGPT) and split them into 70% train, 15% validation, and 15% test set. We report classification metrics (precision, recall, F_1 scores) with summary sentences being the positive class. For each long-form answer, metrics are computed against each of the three references, with the results from the reference with the maximum F_1 score reported. We also report exact-match (EM), whether the model-predicted summary sentence set matches any of the three annotations. The training details and hyperparameters can be found in Appendix B.

PreSumm We use PreSumm (Liu and Lapata, 2019), a BERT-based extractive summarization model, which was trained on the CNN/DailyMail (Nallapati et al., 2016) dataset. It encodes the document with pre-trained BERT (Devlin et al., 2018) and outputs a score for each sentence. We select a threshold for the score at which it is considered a summary sentence to maximize the F1 score on the validation set. We evaluate both the original model (trained on CNN/DM dataset) and the model fine-tuned on our dataset.

	P	R	F_1	EM %
LEAD-2	0.41	0.74	0.51	11.4
LEAD-3	0.46	0.83	0.56	5.3
PreSumm-cnn (A)	0.46	0.77	0.55	11.7
PreSumm-cnn (Q+A)	0.53	0.78	0.60	11.0
PreSumm-cnn+ours (A)	0.55	0.81	0.61	**36.0**
PreSumm-cnn+ours (Q+A)	0.55	**0.88**	0.63	30.9
T5-ours (A)	0.67	0.71	0.65	20.5
T5-ours (Q+A)	**0.70**	0.78	**0.69**	25.0
Human*	0.77	0.79	0.77	41.3

Table 3: Binary classification accuracy of extractive summarization models on the test set.

T5 We use a sequence-to-sequence model, T5-large (Raffel et al., 2019), to classify whether a sentence belongs to the summary or not. This was the best performing model for fine-grained role classification of long-form answers in Xu et al. (2022). For question prepending input, the input sequence to the model would be: [q [1] a_1 [2] a_2 ... [n] a_n]. The output sentence would then be of the form: [[1] r_1 [2] r_2 ... [n] r_n], where r_i was a binary class label whether i-th answer sentence a_i belongs to the summary or not.

Results Table 3 reports model performances on the test set. The result on the validation set can be found in Table 8 in the appendix. With in-domain fine-tuning, both models are able to accurately predict which sentences belong to the summary. Fine-tuned T5 model shows a strong performance, though underperforming human, especially in exact match. We also find all trained classifiers benefit from having questions as additional input, signifying that questions provide important signals for content selection. While there is room for improvement, results suggest that predicting key sentence sets is not a major hurdle for state-of-the-art language models. Thus, we use the **gold** extractive summary for our user study (Section 5).

4 Abstractive Summaries for Long form Answers

While we have gold extractive summaries at hand, they often suffer from disfluencies and factual errors (Zhang et al., 2022). We aim to improve this in two ways, (1) by introducing a decontextualization (Choi et al., 2021) model to edit extractive summaries and (2) by using abstractive summarization

models. We explore zero-shot transfer from an abstractive summarization model (Zhang et al., 2019) and prompting an instruction-tuned large language model (Brown et al., 2020). We experiment with two types of input sequences: (1) long-form answer only as an input (2) the question followed by a separation token and the long-form answer, whenever applicable. In the latter setting, models sometimes output the question as a part of the summary, which we remove with postprocessing.[4]

4.1 Editing Extractive Summary with Decontextualization

The disfluencies and lack of coherence of extractive summaries are well-known issues, motivating a flurry of abstractive summarization models (Rush et al., 2015; See et al., 2017). While abstractive models can provide coherent and fluent summaries, one of their major issues is hallucination (Kryscinski et al., 2019; Cao et al., 2018). Recent work explores **extract-and-abstract** approaches (Hsu et al., 2018; Liu et al., 2018; Pilault et al., 2020), aiming to take the best of both worlds. Most of these approaches are fine-tuned on an abstractive summarization dataset. As we don't have an abstractive summary of long-form answers at hand, we opt to use a decontextualization model to rewrite the extractive summary.

Decontextualization (Choi et al., 2021) is a text editing task, which aims to rewrite the target sentence in a document such that the edited target sentence can be interpreted when presented alone while preserving its meaning. While its use cases in QA and text retrieval (Gao et al., 2022) have been explored, its use case in summarization has not been explored. Earlier prior work (Clarke and Lapata, 2010; Durrett et al., 2016) have studied discourse constraints for summarization – that for each pronoun included in the summary, the pronoun's antecedent should be included or the pronoun to be rewritten as a full mention to make summary coherent and clear. Decontextualization is well-suited to prevent these common errors of pronouns/concepts being "orphaned" in extractive summary.

[4]For extractive models, we exclude the question if it is chosen as the summary. For abstractive models, we remove the first sentence of the summary if it has high lexical overlap (over 75% unigram overlap) with the question (which happened for roughly 38% of the dataset).

Domain	Pred	Un	Inf	Done	Δ
Wiki (NQ Short)	human	12.0	20.0	68.0	23%
Wiki (NQ Short)	model	14.7	26.3	59.0	13%
LFQA Answers					
Wiki (NQ Long)	model	66.8	13.9	19.3	28%
ELI5	model	49.3	34.3	16.4	34%
Web-GPT	model	66.6	14.6	18.8	29%

Table 4: Decontextualization output statistics. The second column block represents prediction category distribution, where Un represents unnecessary (no edit is necessary), Inf represents infeasible (stand-alone not feasible), Done represents decontextualization attempted.

Method We use an off-the-shelf decontextualization system from recent work (Chen et al., 2021),[5] which trained a T5 3B model on the original decontextualization dataset (Choi et al., 2021) on Wikipedia text. This model takes the concatenation of the Wikipedia page title and a paragraph with the sentence to be decontextualized as input. For ELI5 and WebGPT answers which lack a page title, we consider the question as the title.

If the title is t and the answer consists of k sentences $[a_1, a_2, \ldots, a_k]$ with the i-th sentence being the target to be decontextualized, the input will be formatted as:

$$\text{[CLS]}\ t\ \text{[s]}\ a_1 \ldots a_{i-1}\ \text{[s]}\ a_i\ \text{[s]}\ a_{i+1} \ldots a_k \text{[s]}$$

where [CLS] is a start token and [s] is a separator token. The model outputs the sequence: [CATEGORY] [SEP] y, where the category is one of DONE (if it made edits to the sentence in which case y would be the new sentence), Unnecessary (the sentence does not need an edit, already stand-alone), or Infeasible (the sentence is tricky to be made stand-alone with minimal edits).[6] We only apply decontextualization when the first sentence in the extractive summary is **not** included in the summary set (56% of examples in the dataset), and only decontextualize the first summary sentence.

Decontexutalization Results Table 4 presents basic statistics of the output from decontextualization model. Somewhat surprisingly, the decontextualization model edited only 17.1% of input examples, diverging significantly from its training distribution where 60% of examples

[5]https://github.com/jifan-chen/QA-Verification-Via-NLI/.

[6]In the case of infeasible and unnecessary cases, y would just be the same as a_i).

are edited. For these edited sentences, we report the length increase (Δ), or the average value of `(len(decontext)-len(original)) / len(original)`, following the original study. While decontextualization is attempted less frequently when it is decontextualized the length of the sentence increases more substantially. More ELI5 sentences were classified as `Infeasible`. We hypothesize that the sentences in ELI5 could be deemed more challenging because of the narrative nature of Reddit posts. We include sample decontextualization outputs in Table 9 in the appendix.

We manually examine decontextualization outputs from ELI5 and Web-GPT to evaluate their performance on out-of-domain, non-Wikipedia texts. We (the authors of this paper) randomly sample 50 examples where the model has made changes, and 50 examples from the entire set. Out of 50 edits, 42 edits were meaning preserving (without introducing factually incorrect contents), and 44 edits successfully decontextualized the sentence (without unresolved or unclear references). On a randomly sampled set of 50 examples, we evaluate whether the category assigned is correct (infeasible, unnecessary, done), finding 45 examples were assigned the correct category. Overall, we found the zero-shot performance of the decontextualization system on the new domain was surprisingly robust. Recent work (Eisenstein et al., 2022) also showed large language model can perform decontextualization robustly when prompted carefully. We will evaluate decontextualized summaries with a user study in Section 5.

4.2 Abstractive Models

In this section, we explore abstractive models for summarization to improve fluency.

Pegasus (Zhang et al., 2019) shows promising performance across diverse summarization benchmarks. We examine a suite of Pegasus fine-tuned on various summarization datasets and chose a model fine-tuned on the CNN/DailyMail as it showed the most promising results upon manual inspection. We do not fine-tune it with our extractive dataset to preserve its abstract nature.

GPT-3 Recent work (Goyal et al., 2022a) has found that GPT-3 (Brown et al., 2020) exhibits strong zero-shot performance on several news summarization benchmarks. Unlike fine-tuned abstractive models, prompted language models would not inherit issues from noisy distant supervision training datasets. Thus, we investigate its ability to perform zero-shot long-form answer summarization. Specifically, we used the `text-davinci-002 model`.[7] We explore two settings: with and without length control in the prompt, following prior work (Goyal et al., 2022a). The prompt with length control is "`Q: {question text} A: {answer text} Summarize the above answer in {length of gold summary} sentences`", and the prompt without length control is "`Q: {question text} A: {answer text} Summarize the above answer.`"

4.3 Automatic Evaluation

We first aim to perform an automatic evaluation of abstractive systems, using gold extractive summaries as references. While this would not evaluate fluency, automatic metrics measure the content selection of generated abstractive summaries.

Setting We use the same data split as in Section 3.5, and repeat lead baselines: `LEAD-2` and `LEAD-3`. We use established automatic summarization evaluation metrics ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020).[8] As our dataset is 3-way annotated, we report the highest ROUGE-L F_1 score among the three reference answers and use the same reference answer to compute BERTScore F_1. The `Human` baseline is computed by choosing one extractive summary annotation at random as the reference and doing a pairwise computation of ROUGE and BERTScore with the other two annotations for that example.

Results Table 5 reports model performances on the test set. The results on the development set are in Table 7 in the appendix. Similar to other domains, lead baselines show strong performances, outperforming models trained on out-of-domain data (Pegasus, GPT3). Yet, they are inherently limited, covering only 73% of the summary sentences. We see that the abstractive models show better performance with the BERTScore metric compared to the ROUGE-L metric, potentially due to the ROUGE-L metric punishing for paraphrasing. Having the question in addition to the answer improves the performance of the Pegasus model. Having length control also improves the zero-shot performance of GPT-3, similar to the finding from

[7] We set the max generation length to 512 tokens and temperature to 0. The generations were queried on October 19, 2022.

[8] We use the `bert-base-uncased` checkpoint.

Model	Input	ROUGE	BERTScore	Length
LEAD-2	A	0.553	0.673	38.18 (2.00)
LEAD-3	A	**0.652**	0.711	59.40 (3.00)
Pegasus	A	0.569	0.749	43.03 (2.65)
Pegasus	Q+A	0.588	**0.759**	43.36 (2.80)
GPT3	A+L	0.460	0.647	32.17 (1.71)
	A	0.457	0.638	53.01 (2.84)
	Q+A+L	0.497	0.670	**31.34 (1.63)**
	Q+A	0.484	0.662	46.12 (2.20)
Human	Q+A	0.811	0.881	39.41 (1.93)

Table 5: Automatic evaluation results on the test set. For the "Input" column, A refers to a long answer while Q+A refers to (question, long answer) as an input to the model and L refers to the length of the gold extractive summary in sentences. For length, we present the number of tokens, with the number of sentences in the parenthesis.

prior work (Goyal et al., 2022b). This is a semi-oracle setting as the model is given the summary length.

5 Human Evaluation of Summary Answers

So far we have evaluated summarized answers against the gold extractive summary. Yet, we are aware extractive answers themselves are limited and automatic evaluation of summary is non-trivial. To properly evaluate summarizing long-form answers, we launch a user study evaluating four different types of answer summaries: a gold extractive summary, a gold extractive summary that is decontextualized, an abstract summary from Pegasus, and an abstract summary from GPT3. Can the summarized answer present a useful, concise answer that preserves the original meaning of the long-form answer, without producing incoherent discourse structure (e.g., orphaned anaphora)?

5.1 User Study Design

We design a two-stage interface to evaluate the summarized answer. The exact wording and interface can be found in the appendix (Figures 5, 6, 7, and 8). First, they are shown the summary answer and the question alone, and then, the original long-form answer will be shown to them.

Stage 1: The annotators first measure the quality of the summary answer itself.
FLUENCY (choices: Yes/No): if the answer is grammatical and fluent. We do not distinguish coherence and fluency as prior study (Fabbri et al., 2021) reports that annotators often confuse those two dimensions.
ADEQUACY (choices: Yes/Partially/No): if the summary adequately answers the original question.

Stage 2: The annotators then measure *both* the summary and original long-form answer.
FAITHFULNESS (choices: Yes/No): if the summary accurately captures the main idea of a long-form answer regarding the question.
LONG-ANSWER ADEQUACY (choices: Yes/Partially/No): if the long-form answer addresses the question adequately. This annotation *only* evaluates the original long-form answer, as a control to avoid blaming the summarization system when the long answer itself is not adequate. As we filtered out invalid long answers during pre-processing, most answers should be labeled as adequate.

5.2 User Study Setting

Data We annotate 175 long-form answers paired with four types of summary: (1) summary generated from our best abstractive model (`Pegasus`), (2) gold extractive summary (`GOLD`), (3) gold extractive summary that is decontextualized with automatic decontextualizer system (`GOLD++`) and (4) `GPT-3` zero shot summaries with length restriction. We sample 150 examples at random and additionally sample 25 examples where the decontextualization process made edits to the gold extractive summary.

The average length of the tokens for the four summary settings were 43.4, 40.9, 47.6, and 31.3 for `Pegasus`,`GOLD`,`GOLD++`,`GPT3`.

Annotators Human evaluation was done on the Amazon Mechanical Turk platform. We required the workers to be from English-speaking countries and have at least a 95% acceptance rate on 1000+ HITs. Each worker was paid $0.50 per annotation, translating to an hourly rate of $15. We set up the task that each annotator will see only one variant of the summary per each long-form answer. The annotators were not aware of which summarization system provided the summary. A small subset of data is annotated by the authors, following the same setup. We had 561 unique annotators for this task.

5.3 Results

Table 6 presents the results from the user study. We report two numbers – one on all 175 examples, and one on a subset of 63 examples where decontextualization changed the extractive summary.

	Summary Fluency (Yes)	Summary Adequacy Yes	Summary Adequacy Partially	Summary Adequacy No	Faithfulness (Yes)	Long-Answer Adequacy Yes	Long-Answer Adequacy Partially	Long-Answer Adequacy No	Func
Kappa	0.513		0.368		0.506		0.474		
Pegasus	89.7 (91.0)	62.5 (63.0)	31.4 (31.2)	6.1 (5.8)	83.2 (82.5)	81.5 (82.5)	17.0 (15.9)	1.5 (1.6)	65.7
GOLD	85.5 (83.6)	61.0 (56.6)	32.6 (36.6)	6.4 (6.9)	84.0 (83.1)	81.7 (81.5)	16.6 (16.4)	1.7 (2.1)	60.1
GOLD++	88.6 (93.7)	66.5 (70.4)	25.9 (21.7)	7.6 (7.9)	84.4 (84.1)	82.5 (82.5)	16.0 (15.3)	1.5 (2.1)	**67.0**
GPT3	**94.1 (94.1)**	**67.8 (71.4)**	26.5 (21.7)	5.7 (6.9)	**85.3 (85.2)**	81.9 (82.0)	16.4 (16.4)	1.7 (1.6)	**67.0**

Table 6: User study results. The first row shows Fleiss' kappa for each question. The rest of the rows present the percentage of examples in each category, with results on the subset of 63 examples where decontextualization modified the extractive summary presented in parenthesis. The last column presents the percentage of functional short answers, meaning they are adequate, fluent, and meaning-preserving.

We include the inter-annotator agreement for each question in the first row. We observed moderate to high agreement for all four questions. Evaluating the quality of answers (summary adequacy and long answer adequacy) was more subjective than evaluating fluency or faithfulness, revealing the challenge of open-ended long-form answer evaluation as pointed out in prior work (Krishna et al., 2021). We also see high agreement among annotators by comparing long answer adequacy distributions across four rows, which are very similar as expected.

Can a summarized version of long-form answers provide an **adequate** answer to the original question? We see somewhat mixed results – while the annotators said the summaries provide at least a partial answer to the question most of the time (over 90%), only about 60% of answers per system provide adequate answers. Again, we find that decontextualization helps – on about 10% examples, annotators labeled extractive answers as partially adequate, but their decontextualized versions are adequate.[9] GPT-3 produces adequate summaries the most, showcasing its powerful zero-shot summarization ability (Goyal et al., 2022a). Further analysis showed that summary adequacy is highly system dependent rather than question dependent – for 90% of the questions, there is at least one system whose outputs are adequate according to the majority of the annotators.

We find **fluency** is not a major issue, both for extractive and abstractive systems. The large-scale language model (GPT3), in particular, provides the most fluent answers. For the extractive summaries, we see a substantial gain (about 10% on 63 examples where decontextualization changed the input) in fluency by introducing contextualization. The fluency gap between Gold and Gold++ was statistically significant on McNemar's test with $p < 0.05$.

We observe a slightly lower performance on **faithfulness** across four summary systems compared to fluency. While the weaker abstractive model (Pegasus) ranks slightly lower than the extractive model, GPT-3 somewhat surprisingly outperforms extractive approaches in meaning preservation. This mirrors findings from a recent study (Zhang et al., 2022) about how extractive summary can also introduce factual errors. Overall, faithfulness has been extensively studied in summarization literature (Fabbri et al., 2022a) but mostly in the news domain.

When can we use summarized answers? In the last column, we report the percentage of summary answers that are fluent, adequate, and faithful to the original long-form answer. Decontextualized answers (GOLD++) and GPT-3 zero-shot summary achieve more promising results than the other two approaches. Of the 168 long answers considered "adequate" by a majority of the annotators, 160 (95%) of them has at least one summary that was considered functional by a majority of the annotators. We examine error cases in the next section.

5.4 What makes it hard for models to summarize long-form answers?

As we have identified fluency as a minor issue, we specifically look at 60 examples that satisfy all the following conditions: (1) summary is fluent, (2) summary answer is not fully adequate nor faithful, and (3) long-form answer is adequate.

We identify a few patterns of why the summary answers fall short: (1) around 10% of them contain summarization errors (e.g. not properly resolving anaphora or hallucination). (2) for around 60% of examples, adding a few more sentences to the summary was necessary to provide a coherent answer to the question. This is particularly true in cases

[9]This difference was also statistically significant with a t-test where Yes/Partially/No maps to a (1.0/0.5/0.0) score.

Summarization error

Q: Why do most restaurants sell Pepsi instead of Coke, and yet Coke is seen to be a bigger competitor?
A: Coke sells way more soda by volume than Pepsi. As a response, Pepsi offers its products to restaurants at a reduced cost, which is why many restaurants carry it. But only up to midscale places – no nice restaurant serves Pepsi, because Coke has more cachėt, and also you need it for mixed drinks. Note also that McDonald's, the single biggest restaurant chain in the world, serves Coke.

Complex Answer

Q: How is it that the human brain/body sometimes wakes up seconds before an alarm goes off?!
A: Your body does have internal regulation mechanisms, I'm not a doctor and there are plenty who are who can talk more intelligently about the circadian rhythm of the body etc. The other component is psychological. What's happening is an example of confirmation bias. You've woken up a few times almost on the clock (relative to the total number of days you've ever slept in your life). Though this number is astronomical low, you only remember the times you did wake up on the minute. You bias yourself to count those times and subconsciously ignore the other times and thus you feel as though you have an ability to wake up on time. This also happens when people think that they can catch when people are looking at them. You sometimes do and sometimes don't, but the times you don't are not out of the ordinary so you forget them. Thus you only remember catching them and get a false sense of confirmation.
GPT-3 summary: The human brain/body sometimes wakes up seconds before an alarm goes off because of the body's internal regulation mechanisms and the psychological phenomenon of confirmation bias.

Figure 1: Examples with inadequate summaries: In the first example, the highlighted extractive summaries needs further decontextualization. In the second example, the long-form answer is too complex.

where the answers are multifaceted (e.g., providing multiple reasons for some phenomena, and the current summary contains only one of them). We also noticed a few cases where disclaimers (e.g., "I'm talking about poverty in U.S.") or counterexamples in the long-form answer that were not included in the summary, potentially misleading the readers. (3) some long-form answers (around 25%) are tricky to summarize without massive rewriting as it is explaining a complex procedure (e.g., why the Obama administration could not pass legislation, see the full example in Table 1). Figure 1 presents two representative failure cases. Future QA models can actively identify questions that require comprehensive v.s. concise answers.

6 Related Work

Query/Aspect-focused summarization Our task is relevant to query-focused summarization, which studies *controllable* summarization with respect to a query (Xu and Lapata, 2020; Deng et al., 2020; Zhu et al., 2020; Vig et al., 2021) or aspect (Angelidis et al., 2021; Hayashi et al., 2021; Ahuja et al., 2022; Kulkarni et al., 2020). Recently proposed MASH-QA (Zhu et al., 2020) dataset on the medical domain presents a question, context document, and extractive answer sentences. Compared to these works which summarize documents written independently of the question into a summary, we aim to compress long-form answers written with respect to the question. Another line of work (Fabbri et al., 2022b; Song et al., 2017) studies generating summaries of *multiple* answers to the same question. Lastly, Deng et al. (2019) looks into the same task formulation of summarizing long-form answers, but their evaluation is limited to distantly supervised data.

Decontextualization for summarization Slobodkin et al. (2022) proposes the task of controllable text reduction, which rewrites chosen sentences from a document in a coherent manner using existing summarization datasets. They cover longer documents and involve multiple sentences to be decontextualized whereas we reuse a single-sentence decontextualization model (Choi et al., 2021).

7 Conclusion and Future Work

We present the first study on generating concise answers to complex questions. We collect an extractive summarization dataset in the new summarization domain of long-form answers to support future research. To address this new task, we deploy diverse summarization models, including zero-shot abstractive summarization models and a new decontextualization postprocessing method, which is applied to extractive summaries. Through our comprehensive user study, we find that around 70% of the summaries can serve as functional, concise answers to the original questions. Our work shows potential for building QA systems that generate answers at different granularities, as well as using decontextualization to improve the faithfulness and fluency of extractive summaries. Future work can also look into applying controllable generation techniques (Yang and Klein, 2021; Li et al., 2022; Qin et al., 2022) to generate answers with different lengths to generate concise answers.

Limitations

Our study is limited in scope, studying only English question-answering data. We also acknowledge that the long-form answers we study are not always factually correct, as they can be outdated (Zhang and Choi, 2021) or incorrect as they are crawled from web forums (Fan et al., 2019).

Further, our user study is limited in its scale, evaluating 175 instances, and does not carefully study potentially diverging interpretations from annotators of different demographics. We also do not extensively explore all summarization models, such as the extract-and-abstract approaches mentioned in related work.

Ethics Statement

Our data collection and user study protocols do not collect identifiable private information from annotators.

The question-answering data we annotated comes from an English online forum and might contain biased information. Our annotation is done by crowd-workers recruited from an online platform. We make use of pre-trained language models to generate abstractive summaries, which could suffer from hallucinating unfactual contents (Kang and Hashimoto, 2020) and perpetuating bias (Field et al., 2021). Thus, more post-processing steps are required before presenting these contents to users. Our user study shows that our proposed method, extract-and-decontextualize, could be one effective post-processing step to reduce hallucination.

Acknowledgements

We thank Tanya Goyal, Jessy Li, Jiacheng Xu, and members of the UT Austin NLP community for their helpful feedback on the draft. We thank Jifan Chen for sharing the decontextualization model with us. We also thank the reviewers and meta-reviewer of the ACL community for helpful comments and feedback on the earlier draft of the paper. Lastly, we would like to thank the crowdworkers for their help with our data annotation and user study. The work is partially supported by a gift from Google Faculty Research Award.

References

Ojas Ahuja, Jiacheng Xu, Akshay Kumar Gupta, Kevin Horecka, and Greg Durrett. 2022. Aspectnews: Aspect-oriented summarization of news documents. In *ACL*.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. *ArXiv*, abs/2107.00152.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. *ArXiv*, abs/1711.04434.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can nli models verify qa systems' predictions? *ArXiv*, abs/2104.08731.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *CoRR*, abs/2102.05169.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36:411–441.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2019. Joint learning of answer selection and answer summary generation in community question answering. In *AAAI Conference on Artificial Intelligence*.

Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Multi-hop inference for question-driven summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Jacob Eisenstein, Daniel Andor, Bernd Bohnet, Michael Collins, and David Mimno. 2022. Honest students from untrusted teachers: Learning an interpretable question-answering pipeline from a pretrained language model. *ArXiv*, abs/2210.02498.

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

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022a. 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. Association for Computational Linguistics.

Alexander Fabbri, Xiaojian Wu, Srini Iyer, Haoran Li, and Mona Diab. 2022b. AnswerSumm: A manually-curated dataset and pipeline for answer summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2508–2520, Seattle, United States. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. *CoRR*, abs/1907.09190.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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 1905–1925, Online. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, N. Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2022. Rarr: Researching and revising what language models say, using language models.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022b. Snac: Coherence error detection for narrative summarization. *ArXiv*, abs/2205.09641.

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.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *ArXiv*, abs/1805.06266.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *ACL*.

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 *EMNLP*.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217.

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.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam M. Shazeer. 2018. Generating wikipedia by summarizing long sequences. *ArXiv*, abs/1801.10198.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

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. *CoRR*, abs/1808.08745.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Christopher Joseph Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 9308–9319.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *ArXiv*, abs/2202.11705.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *ArXiv*, abs/1704.04368.

Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst, and Ido Dagan. 2022. Controlled text reduction.

Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and M. de Rijke. 2017. Summarizing answers in non-factoid community question-answering. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*.

Jesse Vig, Alexander R. Fabbri, Wojciech Kryściński, Chien-Sheng Wu, and Wenhao Liu. 2021. Exploring neural models for query-focused summarization.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling exemplification in long-form question answering via retrieval. In *North American Chapter of the Association for Computational Linguistics*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering.

Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *ArXiv*, abs/2004.03027.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Michael J.Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *ArXiv*, abs/2109.06157.

Shiyue Zhang, David Wan, and Mohit Bansal. 2022. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *ArXiv*, abs/2209.03549.

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*.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Appendix

A.1 Summary Annotation Interface

Figure 3 presents the interface for summary annotation 3 and Figure 4 is the screenshot of the instruction presented to the annotators.

A.2 User Study Interface

Figures 5 and 6 are screenshots of the interface provided to the MTurkers who participated in the user study to analyze the quality of the summaries and Figures 7 and 8 are screenshots of the instructions provided with the corresponding steps.

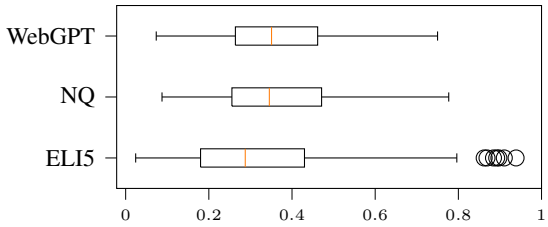

Figure 2: Box plot of compression ratio $\frac{|s|}{|d|}$.

A.3 Dataset Compression Statistics

Figure 2 plots the token-level compression ratio (% of tokens included in the summary) on the three different types of long-form answers we study.

B Model Training Details

All models are trained/evaluated on NVIDIA Quadro RTX 8000 GPUs. We use `pytorch-transformers` Wolf et al. (2019) to implement our models. The hyperparameters are manually searched by the authors.

PreSumm We use the checkpoint of `BertSumExt` from https://github.com/nlpyang/PreSumm. We use the same hyperparameter in the original paper, using a batch size of 16 and a learning rate of $2e-3$. On two GPUs, fine-tuning on the training set and then evaluating on the test set takes between 1 to 2 hours.

T5 We use the T5-large checkpoint with 770 million parameters and fine-tune for 30 epochs with a batch size of 16 and learning rate of $1e-4$. On two GPUs, fine-tuning on the training set and then evaluating on the test set takes between 2 to 3 hours.

B.1 Validation Set Results

Tables 7 and 8 show our automatic evaluation results on the validation set for the extractive and abstractive models (computed in the same way that the test set values were).

B.2 Decontextualization Sample Output

Table 9 gives three examples of the modifications that the decontextualization models made to the extractive gold label summaries.

Model	Input	ROUGE	BERTScore	Length
LEAD-2	A	0.541	0.677	38.36 (2.00)
LEAD-3	A	0.641	0.710	58.29 (3.00)
Pegasus	A	0.571	0.750	43.50 (2.77)
Pegasus	Q + A	**0.572**	**0.752**	41.11 (2.64)
GPT3	(length)	0.517	0.681	**32.29 (1.68)**
GPT3		0.507	0.683	48.03 (2.24)
Human		0.815	0.883	39.62 (1.95)

Table 7: Automatic summary evaluation results on the validation set. For the "Input" column, A refers to using only the long answer as an input to the model while Q+A provides the long answer with the question prepended as an input. The length is computed in the number of tokens/words and the number in parenthesis represents the average number of sentences.

	P	R	F_1	EM %
LEAD-2	0.42	0.74	0.51	11.3
LEAD-3	0.47	0.81	0.55	5.6
PreSumm-cnn (A)	0.47	0.75	0.55	11.7
PreSumm-cnn (Q+A)	0.52	0.78	0.60	21.8
PreSumm-cnn+ours (A)	0.56	0.89	0.65	28.1
PreSumm-cnn+ours (Q+A)	0.58	**0.91**	0.68	35.9
T5-ours (A)	0.70	0.73	0.66	20.0
T5-ours (Q+A)	**0.73**	0.78	**0.71**	26.3
Human*	0.76	0.80	0.77	40.8

Table 8: Binary classification accuracy of extractive summarization models on the validation set.

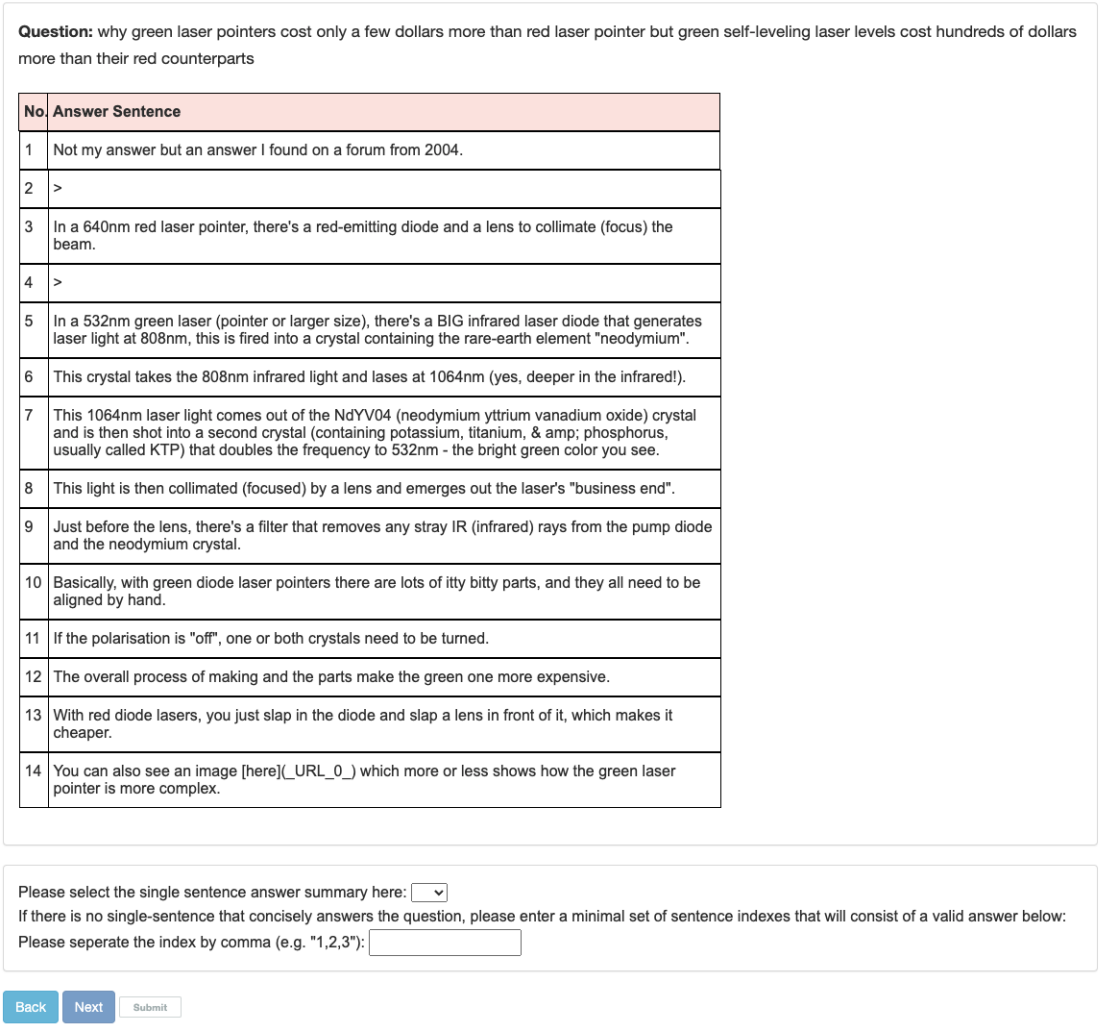

Figure 3: Summary annotation interface

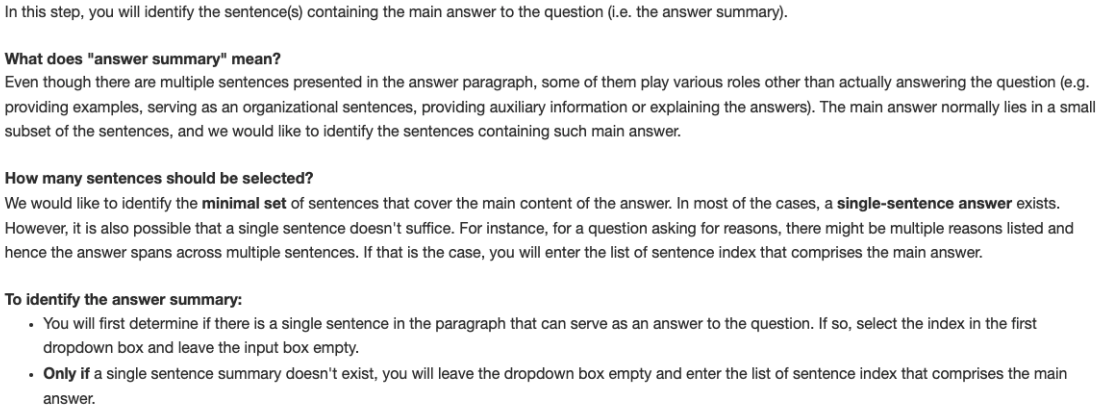

Figure 4: Summary annotation instruction. We provided a few examples to the annotators, which are truncated here.

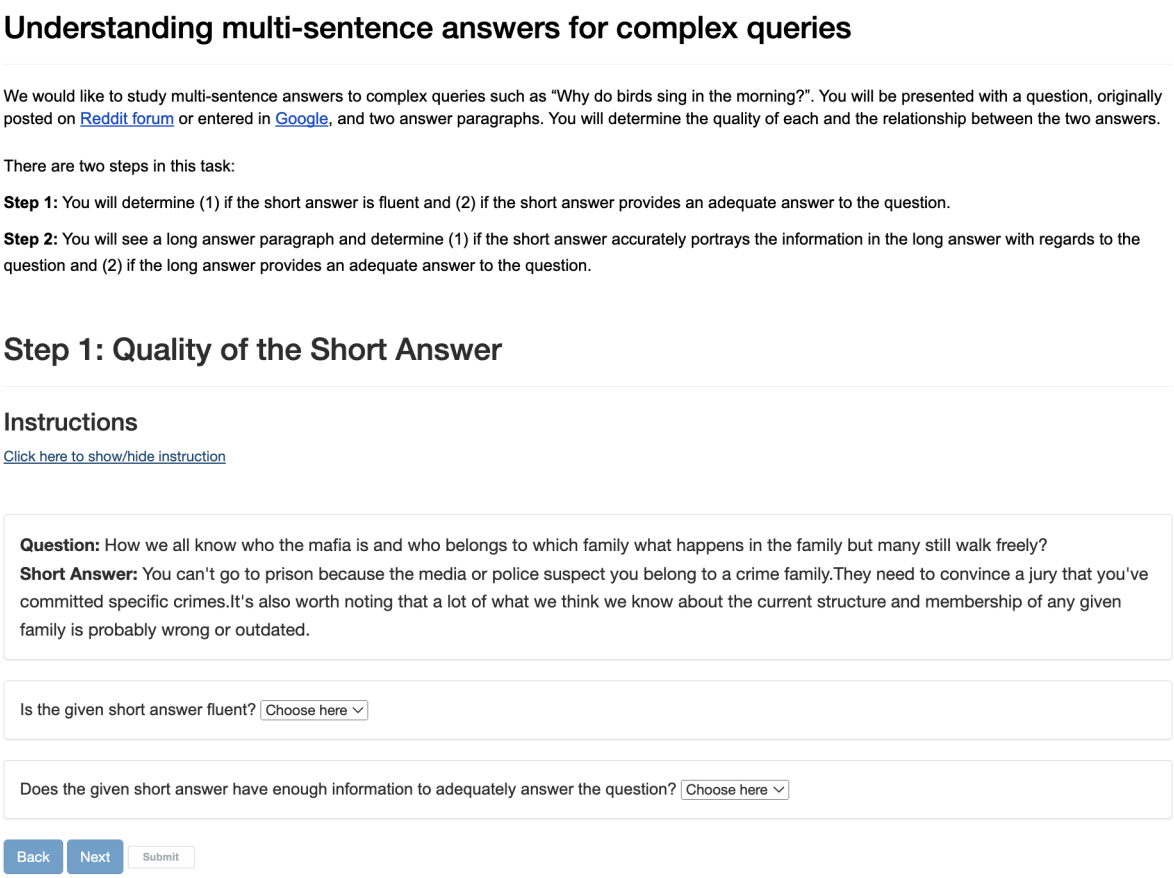

Figure 5: User study annotation UI (Step 1)

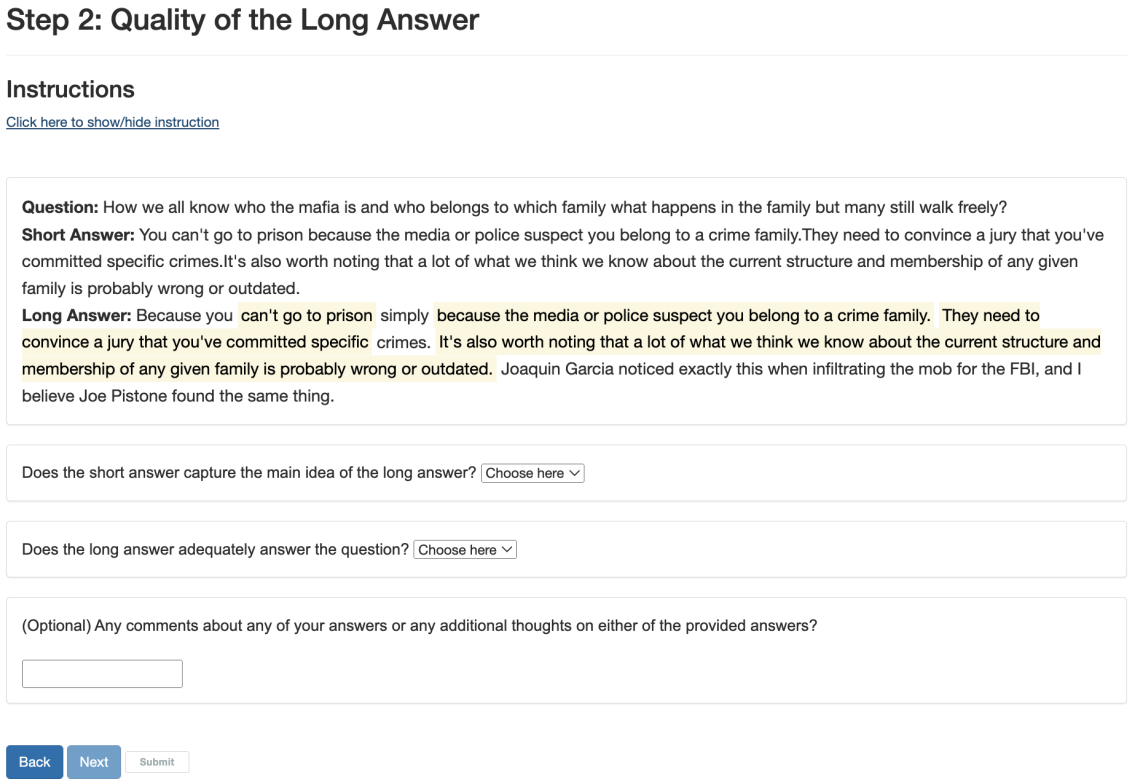

Figure 6: User study annotation UI (Step 2)

Instructions

Click here to show/hide instruction

Fluency

You will choose from **Yes/No** . An answer is not fluent if it contains grammatical errors or if it contains unclear references. Below are some examples of fluent and not fluent answers.

Example	Fluent	Reasoning
Question: What is the weather usually like in australia? **Short Answer:** The climate varies widely due to its large geographical size , but by far the largest part of Australia is desert or semi-arid . Only the south - east and south - west corners have a temperate climate and moderately fertile soil . The northern part of the country has a tropical climate , varied between tropical rainforests , grasslands and part desert .	Yes	The answer does not refer to anything that isn't explicitly stated.
Question: how did the mandate of heaven affect chinese history? **Short Answer:** It was used throughout the history of China to legitimize the successful overthrow and installation of new emperors , including non-Han ethnic monarchs such as the Qing dynasty .	Yes	Although "It" is not defined within the answer itself, **it is clear that when read with the question, "It" refers to the mandate of heaven.**
Question: Why has clock speed on CPU's become almost irrelevant? **Short Answer:** You can speed up work by increasing the rate that the assembly line moves but this can only increase so fast before you start getting errors in the production from the workers (aka electronic components).	No	The reference to the assembly line is unclear given that the initial question was about the CPU.
Question: Why do people's stomach look bloated when they're malnourished? **Short Answer:** This causes fluid to leave the vessels and enter a cavity like your abdomen.	No	It is unclear what process "This" refers to and so you have no information on what exactly is causing the fluid to leave the vessels.

Answer Adequacy

You will choose from the below three options for answer adequacy:

1. **Yes**: The paragraph provides an adequate answer to the question.
2. **Partially**: The paragraph partially addresses the question.
3. **No**: The paragraph doesn't provide information that addresses the question.

Below are examples for each category:

Example	Adequate Answer	Reasoning
Question: What is the weather usually like in australia? **Short Answer:** The climate varies widely due to its large geographical size , but by far the largest part of Australia is desert or semi-arid . Only the south - east and south - west corners have a temperate climate and moderately fertile soil . The northern part of the country has a tropical climate , varied between tropical rainforests , grasslands and part desert .	Yes	The answer provide a detailed description of weather throughout various parts of Australia making it an adequate response to the question.
Question: how did the mandate of heaven affect chinese history? **Short Answer:** It was used throughout the history of China to legitimize the successful overthrow and installation of new emperors , including non-Han ethnic monarchs such as the Qing dynasty .	Yes	The answer gives a good description of when the mandate of heaven was used (specifically to overthrow/install new rulers) giving the implication that its role in chinese history was to be used to consolidate power. This makes it an adequate response to the question.
Question: Why has clock speed on CPU's become almost irrelevant? **Short Answer:** You can speed up work by increasing the rate that the assembly line moves but this can only increase so fast before you start getting errors in the production from the workers (aka electronic components).	Partially	While the metaphor about assembly line is not completely described, you can make an assumption that it is trying to say if you increase the clock speed too much the tradeoff will be that there will be more errors. Thus, this partially answers the question but doesn't completely explain everything.
Question: Why is butter sometimes measured in cups? **Short Answer:** There was a time before it only cost $10 for a digital scale to keep in your kitchen. In that time, most recipes were made using volume measurements. In addition, the butter churning process ends with setting your butter in a container to solidify again.	Partially	While the short answer gives the idea of what type of measures used to exist and how the butter churning process ends up in solidified form, it doesn't necessarily explain why it is in cups and for that reason it is a partial answer.
Question: Why do people's stomach look bloated when they're malnourished? **Short Answer:** This causes fluid to leave the vessels and enter a cavity like your abdomen.	No	Since the answer doesn't describe what "This" is, it makes it unclear on what the actual process is for the stomach to be bloated which is what the question was asking so this is not an adequate response.
Question: When you're sick and can only breathe out of one nostril, then you turn over and a few minutes later it "falls" and you can breathe out the other, why does this happen? **Short Answer:** The turbinate is responsible for the back-and-forth nasal blockage people experience.	No	The answer brings up what a turbinate is but from the question it is unclear what relevance this has to the question asked as well as how this helps explain the phenomenon making it an inadequate answer.

Figure 7: User study instructions (Step 1)

Instructions

Click here to show/hide instruction

Captures Long Answer Intention

You will choose from **Yes/No**. We would like to understand whether the short answer captures the main idea of the long answer regarding the question. The highlighted sections in the long answer are words/phrases which match exactly with the short answer. **Below are some examples** :

Example	Captures Long Answer Intention	Reasoning
Question: how did the mandate of heaven affect chinese history? **Short Answer:** It was used throughout the history of China to legitimize the successful overthrow and installation of new emperors , including non-Han ethnic monarchs such as the Qing dynasty . **Long Answer:** The concept of the Mandate of Heaven was first used to support the rule of the kings of the Zhou dynasty (1046 -- 256 BCE) , and legitimize their overthrow of the earlier Shang dynasty (1600 -- 1046 BCE) . It was used throughout the history of China to legitimize the successful overthrow and installation of new emperors , including non-Han ethnic monarchs such as the Qing dynasty . This concept was also used by monarchs in neighboring countries like Korea and Vietnam . A similar situation prevailed since the establishment of Ahom rule in the Kingdom of Assam of India .	Yes	The short answer accurately portrays what the long answer claims the purpose of the mandate of heaven was, which was to legitimize new rulers. The other parts of the long answer provides auxiliary information and hence it is not necessary for answering the question.
Question: What happens if the ppf is a straight line? **Short Answer:** If the shape of the PPF curve is a straight-line , the opportunity cost is constant as production of different goods is changing. **Long Answer:** In the context of a PPF , opportunity cost is directly related to the shape of the curve (see below) . If the shape of the PPF curve is a straight - line , the opportunity cost is constant as production of different goods is changing . But , opportunity cost usually will vary depending on the start and end points . In the diagram on the right , producing 10 more packets of butter , at a low level of butter production , costs the loss of 5 guns (shown as a movement from A to B) . At point C , the economy is already close to its maximum potential butter output . To produce 10 more packets of butter , 50 guns must be sacrificed (as with a movement from C to D) . The ratio of gains to losses is determined by the marginal rate of transformation .	Yes	The short answer accurately portrays what the long answer talks about with relation to the straight-line curve. The long answer has a lot of information about the other ways that the PPF curve may look but with regards to the straight-line (which is what the question is asking) we have the short answer has the same intention as the long answer.
Question: what is the difference between a janitor and a cleaner? **Short Answer:** A janitor (American English , Scottish English) , janitress (female) , custodian , porter , cleaner or caretaker is a person who cleans and maintains buildings such as hospitals , schools , and residential accommodation . **Long Answer:** A janitor (American English , Scottish English) , janitress (female) , custodian , porter , cleaner or caretaker is a person who cleans and maintains buildings such as hospitals , schools , and residential accommodation . Janitors ' primary responsibility is as a cleaner . In some cases , they will also carry out maintenance and security duties . A similar position , but usually with more managerial duties and not including cleaning , is occupied by building superintendents in the United States (and occasionally in Canada) . Cleaning is one of the most commonly outsourced services .	No	The short answer creates an implication that a janitor is just a cleaner for larger buildings but in reality, the long answer is trying to explain that a janitor also has additional responsibilities in addition to being a cleaner like maintainence and security.
Question: How do animals sense the upcoming earthquake? **Short Answer:** Animals can sense an earthquake seconds before humans because many animals sense a primary wave, or "P wave," before the larger, secondary wave that humans feel. Some animals are believed to be able to detect earthquakes hours or days before an earthquake hits. **Long Answer:** Animals can sense an earthquake seconds before humans because many animals sense a primary wave, or "P wave," before the larger, secondary wave that humans feel. Some animals are believed to be able to detect earthquakes hours or days before an earthquake hits. We do not know for certain if this is plausible because it is nearly impossible to do a controlled study on animal behavior prior to an earthquake, but there have been many reports of animals fleeing prior to the onset of an earthquake. One theory for why animals may be able to detect earthquakes so early on is that they are more sensitive to the Earth's vibrations than humans are. Another theory is that they can sense electrical changes in the air. Because some animals, like elephants, pass knowledge on from generation, it may also be possible that some older animals who have experienced an earthquake teach their young how to react to certain geographical signs.	No	If you only read the short answer you are left with the implication that some animals can detect earthquakes days before it hits but as the long answer explains, this is only a hypothesis that hasn't necessarily been confirmed which makes the short answer misleading.

Answer Adequacy

You will choose from the below three options for answer adequacy:

1. **Yes**: The paragraph provides an adequate answer to the question.
2. **Partially**: The paragraph partially addresses the question.
3. **No**: The paragraph doesn't provide information that addresses the question.

Below are examples for each category (note these are the same questions/answers as from part 1, the guidelines are the same but instead you are now judging the long answer):

Figure 8: User study instructions (Step 2)

Question	Long Answer (Abridged)	Decontextualized Extractive Summary
How did Switzerland stay out of WWII?	They were literally the bankers of the war. The Nazis and the allies both kept their assets there. This is how they stayed neutral, because if either side invaded, that side's assets would either be seized by the other side, or seized by the Swiss.	The Nazis and the allies both kept their assets -there +in Switzerland.
Why do some people vomit when they see a corpse and/or witness a homicide?	We essentially vomit at the sight of gory or bloody death as a defense mechanism. In the face of corpses or death, we are often at risk ourselves, and therefore vomit to remove possible biohazards from our system that may have been spread by the dead, as blood and gore are often good at transmitting biohazards. It also prevents us from possibly ingesting any biohazards by forcing everything out of the mouth that may have been headed for the stomach (i.e. blood).	-It also +Vomiting prevents us from possibly ingesting any biohazards by forcing everything out of the mouth that may have been headed for the stomach (i.e. blood).
How does the mls all star game work?	The Major League Soccer All-Star Game is an annual soccer game held by Major League Soccer featuring select players from the league against an international club. MLS initially adopted a traditional all-star game format used by other North American sports leagues where the Eastern Conference squared off against the Western Conference. This eventually evolved into the current system where the league annually invites a club from abroad to play against a league all-star team. The MLS All-Stars hold an 8–4 record in the competition marking the season 's midpoint. Players are awarded rosters spots through a combination of fan voting and selections by the appointed manager and league commissioner.	-This +The Major League Soccer All-Star Game initially adopted a traditional all-star game format used by other North American sports leagues where the Eastern Conference squared off against the Western Conference which eventually evolved into the current system where the league annually invites a club from abroad to play against a league all-star team. Players are awarded rosters spots through a combination of fan voting and selections by the appointed manager and league commissioner.

Table 9: Decontextualization model outputs on three examples from the dataset (the summary of the original answer is highlighted in grey). Despite being out-of-domain, the decontextualization model performs reasonably well.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section on Page 10, right after Section 6 (Conclusion)

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement section on Page 10, after the Limitations section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The abstract is on the first page and the Introduction is the first section

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In section 3 we refer to a dataset collected in Xu et al. (2022) which provided input to collect our annotations and in section 4, we mention all the pre-trained models used and their sources.

☑ B1. Did you cite the creators of artifacts you used?
Sections 3.2 and 3.3 contain citations for the datasets used and Sections 4.1 and 4.2 contain citations for the models used.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In section 1 we mention that we plan to open source all of our models, data, and annotations at time of publication, we will distribute with the CC BY-SA 4.0 license. Our code/data can be found at https://github.com/acpotluri/lfqa_summary/tree/main and https://huggingface.co/datasets/abhilashpotluri/lfqa_summa

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We didn't explicitly discuss it in the paper but we only use publicly available models and question answering datasets and we build our dataset for research processes so it is compatible with the intentions of the existing artifacts.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
In sections 3 and 5 we discuss the data collection process (which was done through MTurk) and we also have screenshots of the annotation pages which show that we do not collect any personal information. We also have the annotation template for both the data and the user study in the public Github repository that we released.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In section 3.2 we discuss the datasets which our data is sourced from and their domains.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In section 3.4 and Table 2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Sections 3.5 and 4.3 along with Tables 4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.1/4.2 has model details and Appendix B has the remaining model details and computing infrastructure/budget descriptions.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup is provided in section 4.3 and hyperparameters for fine-tuned models is in Appendix section B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Experimental statistics for the test set are in section 4.3 and results on the development set are in the appendix.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 contains details of the models used for evaluation and any parameters which were set (as well as details of which evaluation packages were used).

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3 has the dataset annotation details and section 5 has the human evaluation study

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix sections A.1, A.2, and figures 3,4,5,6 contain full details and screenshots of the annotation instructions.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Sections 3.3 and 5.2 have the details of the participants for each annotation task and the hourly pay rates.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The instructions (which are attached in the appendix) for the annotation explain that we are trying to understand multi-sentence answers to complex queries for research purposes.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We discuss this in the ethics statement.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Sections 3.3 and 5.2 have the details of the participants for each annotation task.