WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering

Valeriia Bolotova[1]
lurunchik@gmail.com

Vladislav Blinov[2]
vladislav.a.blinov@gmail.com

Sofya Filippova[1]
s3848979@student.rmit.edu.au

Falk Scholer[1]
falk.scholer@rmit.edu.au

Mark Sanderson[1]
mark.sanderson@rmit.edu.au

[1]RMIT University [2]Ural Federal University

Abstract

Answering non-factoid questions (NFQA) is a challenging task, requiring passage-level answers that are difficult to construct and evaluate. Search engines may provide a summary of a single web page, but many questions require reasoning across multiple documents. Meanwhile, modern models can generate highly coherent and fluent, but often factually incorrect answers that can deceive even non-expert humans. There is a critical need for high-quality resources for multi-document NFQA (MD-NFQA) to train new models and evaluate answers' grounding and factual consistency in relation to supporting documents.

To bridge this gap, we present WIKIHOWQA,[1] a new multi-document NFQA benchmark built on WikiHow, a website dedicated to answering "how-to" questions. The benchmark includes 11,746 human-written answers along with 74,527 supporting documents. We describe the unique challenges of the resource, provide strong baselines, and propose a novel human evaluation framework that utilizes highlighted relevant supporting passages to mitigate issues such as assessor unfamiliarity with the question topic. All code and data, including the automatic code for preparing the human evaluation, are publicly available.

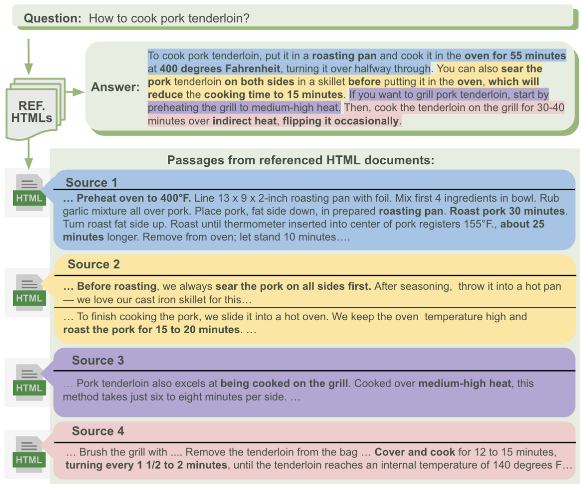

Figure 1: An instance of proposed WIKIHOWQA

1 Introduction

Non-factoid questions (NFQs) requiring long, passage-level answers, such as explanations or opinions, pose challenges for current question-answering systems. While a few datasets exist for NFQA (Cohen and Croft, 2016; Hashemi et al., 2019; Soleimani et al., 2021), users are currently limited to seeing a summary of the most relevant document in a snippet on a search result page (SERP), which can be insufficient for complex questions that have scattered answers across multiple documents or require sophisticated reasoning to generate an answer. However, even relatively straightforward questions like "how to fix my computer issue" can have multiple solutions, requiring the user to manually search through multiple retrieved documents to find the one that applies to their situation. Complex questions such as "how to feel calm and relaxed" may require an aggregated summary of the most popular methods from multiple relevant sources. There are also questions for which answers have yet to be written, requiring a QA system to treat relevant documents as initial sources of information and then to reason out an answer based on them. For example, the question "how to prepare to buy a house in [neighbourhood], [city]" would require the system to retrieve relevant documents about buying houses in general and specific articles about the local house market and the neighbourhood, and then construct a more sophisticated answer through reasoning rather than just summarizing multiple documents.

To address these challenges and move towards more advanced QA systems that can provide in-depth and comprehensive answers to a wide range of questions, we propose a new benchmark for the task of multi-document non-factoid QA (MD-

[1]https://lurunchik.github.io/WikiHowQA/

NFQA). This task involves using multiple relevant documents to generate a complete and coherent answer to a given NFQ. We focus our benchmark on the INSTRUCTION category of NFQs, which often begin with "how to". These questions are underrepresented in current multi-document QA (Bolotova et al., 2022) datasets, despite their popularity, as evidenced by a 140% increase in "how to..." searches on Google since 2004[2] and making up over half of the most searched queries.[3] Our benchmark aims to fill this gap by specifically targeting INSTRUCTION questions.

WikiHow is a web-resource for INSTRUCTION questions that contains over 235,000 articles on a wide range of topics. These articles provide comprehensive step-by-step instructions and are written by a community of experts and reviewed by an average of 16 people. The WikiHow website has proven to be a valuable resource for machine learning tasks (Koupaee and Wang, 2018; Yang et al., 2021; Bhat et al., 2020; Zellers et al., 2019; Ladhak et al., 2020; Boni et al., 2021; Zhang et al., 2020; Cohen et al., 2021; Anthonio et al., 2020), and in our work, we leverage WikiHow to create a high-quality benchmark specifically designed for MD-NFQA within the INSTRUCTION question category. Our benchmark consists of 11,746 questions from the INSTRUCTION category, each paired with a corresponding human-written answer, sourced from a diverse range of WikiHow articles. Each QA pair is supported by corresponding parsed relevant HTML pages from which the answer can be derived (Fig. 1). We evaluate several baseline models on the new benchmark that could serve as lower and upper bounds for model performance.

Human annotation is often used as the standard for evaluating long-form answers, but research has shown that individuals without specific training can only distinguish between human-generated and auto-generated text at a level equivalent to random chance (Clark et al., 2021). This presents a challenge for evaluating the new benchmark for NFQA, particularly when the answers are lengthy or the topic is unfamiliar (Krishna et al., 2021) or not interesting (Bolotova et al., 2020). Given that standard metrics struggle to detect factual inconsistencies, such as number swapping, negation, etc. (Kryscinski et al., 2019), we delve into human sensitivity to such discrepancies by conducting a series of crowdsourcing experiments. Inspired by the elaborate yet resource-intensive manual evaluation framework of abstractive models proposed by Dou et al. (2022), we introduce a simple human evaluation framework that leverages highlighted relevant passages to enhance the quality of NFQA assessments. Our findings demonstrate that incorporating highlighted relevant supporting passages into the evaluation process not only aids evaluators in understanding the context of each question and answer but also contributes to the factual accuracy of the evaluation. As a result, we integrate this method into the evaluation of the newly proposed benchmark.

The new benchmark for MD-NFQA serves as a valuable resource for the development of more advanced QA systems that can provide in-depth and comprehensive answers grounding information in supporting documents. By including parsed passages from relevant HTML pages and providing human judgments, our benchmark also has the potential to facilitate research in evaluating the factual correctness of long-form answers. All data is publicly available[4] on the dataset website.[1]

2 Related Work

This section covers datasets for related MD-NFQA tasks such as long-form question answering (LFQA) and query-based multi-document summarization (QF-MDS), highlighting the scarcity of resources available. We examine the strengths and limitations of these resources, as well as the differences between these tasks and MD-NFQA.

LFQA: The task of LFQA, introduced by Fan et al. (2019), involves providing long answers to open-ended questions. While the associated ELI5 benchmark includes supporting "document" to generate answers, the LFQA task, unlike MD-NFQA task, does not assume that an answer is scattered across multiple documents. Moreover, while MD-NFQA exclusively targets NFQs, ASQA LFQA dataset (Stelmakh et al., 2022) addresses the need for long-form answers to ambiguous factoid questions. Similarly, Natural Questions dataset (Kwiatkowski et al., 2019) contains both long and short answers for factoid questions (Xu et al., 2022).

The ELI5 dataset (Fan et al., 2019) for LFQA includes 272,000 questions from the "Explain Like I'm Five" Reddit web forum, where questions and

[2]https://bit.ly/telegraph-google-how-to
[3]https://bit.ly/most-asked-questions-on-google

[4]Under RMIT University DTA license for research use

answers must have a score of at least two to be included. The top-voted answer for each question is considered the correct answer, and the supporting content for each question is generated by extracting sentences with high tf-idf similarity from the top 100 web pages that match the question from the Common Crawl corpus. In contrast, our proposed MD-NFQA dataset uses texts parsed from relevant HTML pages chosen by the author of the corresponding article on WikiHow, rather than relying on automated methods for curating reference content. In addition to a lack of grounding in supporting content, Krishna et al. (2021) found significant Train/Valid overlap in the ELI5 dataset, and highlight challenges with both automatic answer evaluation and human annotation for this benchmark. In our proposed benchmark, we address these issues through the use of a novel human-evaluation framework utilizing supporting documents and by avoiding question overlap (Sec. 5 and Sec. 3).

Nakano et al. (2021) used questions from the ELI5 dataset and collected new answers from human annotators who were instructed to search for related documents and use them to construct their answers. They trained the WebGPT model on that dataset to answer long-form questions by mimicking the way humans research answers to questions online: it searches and navigates the web to find relevant pieces of information and concludes an answer based on them by citing sources for factual accuracy. Authors report that the model answers are preferred by assessors 69% of the time to the best human-written answer from Reddit when evaluating on ELI5. While the WebGPT dataset is valuable, as of this moment, the authors have only released the model's answers and questions, not the supporting documents.

Finally, it's worth noting that while REASON and EVIDENCE-BASED categories of NFQs prevail in the ELI5 (Bolotova et al., 2022) and WebGPT datasets, our dataset focuses specifically on the INSTRUCTION category.

QF-MDS: In contrast to MD-NFQA, which involves generating detailed passage-level answers to NFQs, QF-MDS (Tombros and Sanderson, 1998) focuses on creating concise summaries in response to specific queries, which may include factoid questions or queries not in question form, such as [Entity], [Event], etc. This requires a different set of skills and approaches compared to MD-NFQA, as QF-MDS summaries do not involve making conclusions or inferences based on the provided information, while MD-NFQA requires higher level reasoning and synthesis to provide complete and accurate answers. While there are a few resources available for MDS (Litkowski, 2004; Angelidis and Lapata, 2018; Liu et al., 2018; Dang, 2006; Fabbri et al., 2019; Ganesan et al., 2010; Wang and Ling, 2016; Yasunaga et al., 2019; Koupaee and Wang, 2018; Lu et al., 2020) and QFS (Zhong et al., 2021; Nema et al., 2017; Zhao et al., 2021), they are scarce for QF-MDS.

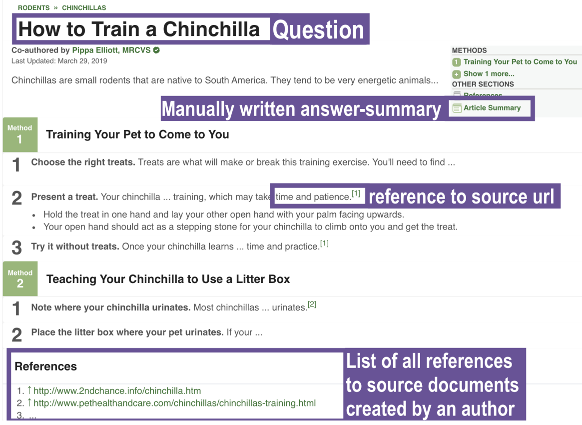

Figure 2: An example WikiHow article

The QMDSCNN and QMDSIR datasets target QF-MDS task (Pasunuru et al., 2021). The first dataset is derived from CNN/DailyMail having real summaries with simulated queries, while the second dataset is derived from a search engine query log and has simulated summaries with real queries.

Another automated approach for curating large datasets for query-focused summarization tasks is AquaMuse (Kulkarni et al., 2020). This dataset supports both abstractive and extractive QF-MDS tasks. Queries and long answers from the Natural Questions dataset (Kwiatkowski et al., 2019), and a pre-processed version of the Common Crawl corpus were used (Raffel et al., 2020). Long answers from Natural Questions and the Common Crawl corpus are encoded into sentence embeddings. Then a similarity search is performed over the corpus and long answers to find candidate documents from the corpus for QF-MDS tasks (similar to ELI5).

Boni et al. (2021) proposed HowSumm, a QF-MDS dataset automatically constructed from WikiHow content by utilizing the referenced articles as the summarization source, the corresponding elements of WikiHow articles as the target summaries, and titles used as the queries. There are two types

Table 1: Comparison of WikiHowQA with other QF-MDS and LFQA datasets

Dataset	# questions	Splits			Reference sources			Answer	
		train (#clusterised/#clusters)	valid	test	# docs	# words	# sents	# words	# sents
WikiHowQA	11,746	8,235 (2,449/7,272)	1,178	2,333	6.3	1,053.6	65.2	113.05	4.9
AQuaMuSe (Kulkarni et al., 2020)	5,519	4,555	440	524	6	1,597.1	66.4	105.9	3.8
ELI5 (Fan et al., 2019)	272,000	237,000	10,000	25,000	–	857.6	–	130.6	–
HowSumm-Method (Boni et al., 2021)	11,121	8,856	1,122	1,143	11.19	1,455.52	71	539.11	31.33
HowSumm-Step (Boni et al., 2021)	84,348	67,403	8,248	8,697	9.98	1,357.37	66.47	98.98	5.23

of QF-MDS tasks, one for methods (HowSumm-Method) and one for steps (HowSumm-Step). While this work is similar to the dataset we present, HowSumm is directed toward the QF-MDS task, does not discuss train-test overlap, and only includes the source URLs in their dataset.

3 Resource Description

This section presents our new resource, including its construction, text statistics, comparison with other benchmarks, and thorough quality analysis.

3.1 Data Collection

Fig. 2 illustrates an example of a WikiHow article accompanied by a high-quality, human-written article summary. These summaries are equivalent to shorter passage-level answers to a question. To build WikiHowQA, we first downloaded over 236,000 articles published on WikiHow before January 2022 using the MediaWiki API. We then filtered out articles without references, resulting in a collection of 126,711 articles. Among these articles, only about 20% had human-written article summaries, which we used as target answers. We downloaded the HTML content of all cited URLs using the Wayback Machine to provide the version closest to when the article was created or modified. The final HTMLs were saved from the snapshot versions. We simplified the HTML source code, and processed the content in two ways: (1) by extracting text content or (2) by converting it to Markdown format. The latter format preserves formatting such as tables and lists. Finally, we rejected articles with missing HTML snapshot links, empty HTML, or empty extracted text. Each instance in our new dataset consists of a question, a human-written article summary treated as the target answer, and a list of parsed texts from the relevant HTML documents cited by the article's author. Fig. 1 shows a simplified example from WikiHowQA.

Avoiding Overlap in Train-Test: When studying state-of-the-art model performance on ELI5 (Fan et al., 2019), Krishna et al. (2021) observed little to no evidence that analyzed models grounded answer generation in the retrieved documents. They attributed this to a significant overlap (around 43.6%) in the training, validation, and test sets of ELI5. Similarly, Lewis et al. (2021) identified train-test overlap in various QA datasets. To avoid this issue in our new resource, we explicitly split questions in a way that no same-topic or paraphrased questions ended up in different splits. To do so, we clustered all questions prior to splitting and assigned all questions belonging to the same cluster to the training split. We only assign clusters to the training split so that the validation and test splits will have more varied questions, to avoid a potential evaluation bias due to the presence of many similar questions.[5]

Table 2: Percentage of novel n-grams

Dataset	% novel n-grams		
	uni-grams	bi-grams	tri-grams
WikiHowQA	11.75	54.87	85.02
DUC 03–04	27.74	72.87	90.61
HowSumm-Method	15.20	52.70	81.90
HowSumm-Step	9.80	47.30	78.90

3.2 Dataset Statistics

The WikiHowQA dataset include a diverse range of topics, 19 in total, with the most frequent being "health", "home and garden", "pets and animals", and "computer and electronics".[6] Table 1 compares WikiHowQA to other QF-MDS and LFQA datasets. ELI5 is the largest dataset in terms of questions and has the lowest average words in reference documents (recall that the single reference document is an aggregate of top-k passage retrieval and hence the absence of some of the reported statistics within Table 1). The number of answer words is higher than other datasets except for HowSumm-Method. AQuaMuSe has

[5] Implementation details can be found in App. B.

[6] The full category distribution is provided in App. A

fewer questions, although the reference sources and answer content are most similar to the statistics of HOWSUMM-STEP and WIKIHOWQA. The HOWSUMM-METHOD dataset statistics are unique, it has 11 source documents per instance on average, and the length of answers are greater with an average of 31 sentences per instance. HOWSUMM-STEP also has a high number of source documents while the remaining statistics follow a similar trend to the other datasets described. While our proposed dataset WIKIHOWQA shares many common statistical characteristics of other datasets, the construction process, source-target mappings and task applications differ as discussed in Sec. 2.

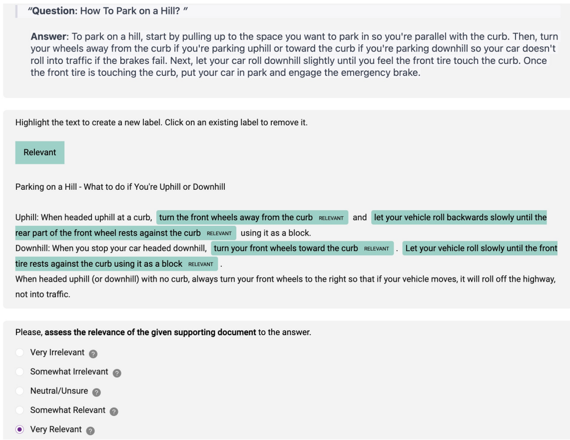

Figure 3: WIKIHOWQA quality annotation

3.3 Quality Verification

Since the relevant passages in our WikiHowQA datasets are sourced from web pages cited by the WikiHow article author, they may not contain the exact words or phrases from target answers. Therefore, to verify the quality of our benchmark, we assess the feasibility of constructing an answer from the given supporting documents. As this is a crucial aspect for a MD-NFQA resource, we evaluate our benchmark using both automatic metrics and thorough human evaluation.

Automatic metrics: The relevance of the supporting documents in our WikiHowQA dataset is first evaluated using the Novel N-Gram Percentage metric (See et al., 2017) as a measure of word intersection between answers and documents. This metric, commonly used for summarization datasets, allows for a strong upper bound comparison. Results presented in Table 2 show that the passage answers in our dataset correspond to the content of the supporting documents with a relatively high n-gram percentage score, similar to that of HOWSUMM-METHOD. We also include results for the Document Understanding Conference (DUC) as an upper bound as it is a high quality human crafted summarization dataset (as reported by Fabbri et al.).

Secondly, we report the average coverage, density and compression metrics for WIKIHOWQA (Grusky et al., 2018). These metrics are commonly used to characterize the quality and difficulty of summarization tasks. We follow Fabbri et al. (2019) who adapted these measures for the MDS use case. The coverage, density and compression scores for WIKIHOWQA are 0.89, 1.86 and 81.46 respectively. It is important to note that Tejaswin et al. (2021) and Bommasani and Cardie (2020) propose additional metrics to compare datasets, some of which are a reformulation of those presented by Grusky et al. (2018).

The above metrics give a general indication of the dataset quality, although even when examples from our dataset are deemed lower quality, for example when they have low semantic similarity, we find that there are instances where an answer can typically still be constructed from the supporting documents. Fig. 1 demonstrates that answers can be successfully constructed from documents even if they have low semantic similarity. For instance, in the first sentence of an answer for the question "How to cook pork tenderloin" (Fig. 1), although the target length of 55 minutes was not mentioned in any document, it can be inferred by adding up roasting time of 30 minutes on one side and 25 minutes on the other side. Similarly, the suggestion "flipping it occasionally" in the last sentence can be rephrased from "turning every 1 1/2 to 2 minutes".

Human evaluation: To verify our observations and ensure the quality of our dataset, we conducted a crowdsourcing study[7] with the goal of verifying the feasibility of answer construction. Each participant task in the study consisted of a QA pair, one supporting document, and three evaluation components. The first evaluation required annotators to classify the question into the appropriate category from a provided list. This step served as an attentiveness check. Annotators were presented with four categories in a random sequence, one

[7] All our crowdsourcing studies were reviewed and approved by the Human Research Ethics Committee of RMIT University. We use Surge AI as our data labeling platform, which provides a workforce of highly skilled and educated native speakers, ensuring high-quality data labeling at scale, allowing for higher quality labeling compared to traditional platforms such as Mechanical Turk.

Table 3: Baseline models for WIKIHOWQA

	Automatic Evaluation				A/B Human Evaluation 100 instances		
Model	Rouge-1	Rouge-2	Rouge-L	BertScore	Prefer Model	Prefer Gold	Tie
DPR + BART	39.8	12.4	23.0	0.881	13	52	35
text-davinci-003	32.2	8.5	19.7	0.873	18	53	29
DPR + *text-davinci-003*	35.4	9.2	20.2	0.868	56	15	29

All differences are statistically significant (Student t-test, p-value < 0.01)

of which was the original category derived from the WikiHow website, while the remaining three were randomly selected from the other 18 categories. This task was designed to ensure that the annotator was paying attention and had read the question carefully. The tasks were rejected if the classification was incorrect. The second evaluation required annotators to read the answer and the supporting document, and use a provided highlighting tool to mark any words or phrases in the supporting document that were directly relevant to the given answer, or could be used to reason or conclude it. In the final step of the process, annotators evaluated the relevance of the provided documents to the answer on a five-point Likert scale (from 0, "Very Irrelevant", to 4, "Very Relevant"). In total, 31 randomly selected QA pairs and their associated supporting documents were annotated, resulting in 104 participant tasks, with only four being rejected. Each assessment was carried out by three workers, receiving compensation of $0.7 per completed task. The study[8] involved 22 fluent English speakers. Fig. 3 illustrates a randomly selected result of this annotation process, as captured in the (simplified) annotation interface used in the study.

The results showed that the workers deem the supporting documents to be mostly "Somewhat Relevant" in relation to the given answers, with a mean document relevance score of 2.59. The inner-annotator agreement is moderate, with a Cohen's Kappa (Artstein and Poesio, 2008) value of 0.51, indicating a reasonable consistency in the annotators' assessments of document relevance. In terms of related text selections within the documents, the mean overlap score between assessor pairs was 0.65, calculated using the overlap coefficient (Vijaymeena and Kavitha, 2016):

$$Overlap(H_1, H_2) = \frac{|H_1 \cap H_2|}{min(|H_1|, |H_2|)}$$

where H_1 and H_2 are the sets of unique words from two annotations. This is consistent with previous similar research (Qu et al., 2019; Bolotova et al., 2020) and indicates that the annotators had good agreement with each other on the parts of the documents that could be used for associated answer construction. Overall, these findings demonstrate that the provided documents are relevant and contain the necessary information for the answers.

4 Baseline Models

This section describes WIKIHOWQA baseline models. The task is to generate an answer to a NFQ grounded in a set of relevant documents from which the answer can be reasoned or concluded.

As a reference point for performance evaluation, we use the Retrieval-Augmented Generation (RAG) (Lewis et al., 2020b) architecture, which is a common choice for abstractive summarization and QA tasks (Fan et al., 2019). Since our dataset provides a small predefined set of relevant supporting documents for each question, we use a retriever to filter out passages from those documents that are less relevant to the question, to help fit more relevant passages into the limited input of a generator. Then, a generator is tasked with generating an answer conditioned on input passages. We experimented with DPR (Petroni et al., 2021) as the retriever and BART-large (Lewis et al., 2020a) as the generator.

When building model input, we first ranked passages from supporting documents based on their relevance to the question using DPR, then truncated them to retain the maximum number of complete passages in the ranking that fit withing the maximum input length of BART. Passages were then reordered based on their source documents, and combined into a single input string of the format "*<q>question</q><d><p>doc 1 passage 1</p><p>...</p></d><d><p>doc 2 passage 1</p>...</d>*". Finally, BART is trained[9] to

[8]Interface with exact annotation instructions in App. C.

[9]Training and decoding parameters are in App. D.

generate gold answers conditioned on this context.

To establish a stronger baseline, we also evaluated the performance of the GPT-3 (Brown et al., 2020) variant *text-davinci-003*[10] in a zero-shot setting. As a model at least 400x times larger than BART, it forms an upper bound for model performance. We compared two prompt kinds for GPT-3, one with the question and another with the question and relevant passages retrieved via DPR.

Standard automatic evaluation metrics for abstractive QA, Rouge-X and BertScore, are reported in Table 3. However, these scores are known to poorly reflect actual model performance (Deutsch and Roth, 2021; Krishna et al., 2021). While recent QA-based evaluation metrics for summarisation seem to better correlate with human judgements (Deutsch et al., 2021; Scialom et al., 2021), their adaptation to NFQA is not straightforward and remains an important area of research, as answers in NFQA are expected to contain facts not mentioned in supporting documents. Instead, we report the results obtained through our human evaluation framework as a more accurate measure of model performance (Sec. 5.2).

5 How To Evaluate How-To Answers

To further ensure the reliability of MD-NFQA evaluation, we introduce a simple human evaluation framework, which we then employ to assess the performance of baselines in our benchmark.

5.1 Evaluating Human Evaluation

Initially, we employed a conventional human evaluation approach, as described in previous studies (Fan et al., 2019; Krishna et al., 2021), where assessors are presented with both model-generated and gold standard answers in a randomized order and asked to select their preferred response. However, our findings indicated that, even when presented with gold standard questions and evaluated by high-performing assessors, simpler model-generated answers were frequently (48% of the time) preferred over reference answers, despite containing factual inaccuracies. Manual inspection revealed that choosing between two well-formulated answers was challenging for participants, in line with previous research of Krishna et al.; Clark et al.. To quantify the ability of annotators to identify factual inconsistencies in answers, we conducted a crowdsourcing[7] evaluation experiment.[11] To facilitate this, we generated a set of modified answers by deliberately introducing factual errors into a subset of the WikiHowQA. This was achieved by randomly selecting 5 QA pairs from each of the 19 question categories available on WikiHow and systematically incorporating various types of inaccuracies, as outlined in Table 4, into 4 out of the 5 answers, resulting in an average of 4.5 modifications per answer. We then manually reviewed the deteriorated answers to ensure they maintained both contextual relevance and grammatical correctness. Following this, we conducted two separate evaluation rounds involving a total of 34 workers, who assessed all 95 QA pairs. It's important to note that no worker participated in both rounds, ensuring an unbiased evaluation.

The first trial aimed to establish the baseline performance of the standard evaluation framework, in which assessors evaluate answers without the assistance of any on-screen relevant information. Each HIT included one QA pair where the answer may had been deteriorated. Similar to Sec. 3.3, HITs began with the attentiveness test of question category classification. Assessors then indicated their familiarity with the question and evaluated the usefulness of the answer on a five-point Likert scale (from "Very Unfamiliar" / "Very Useless" to "Very Familiar" / "Very Useful"). Finally, they were tasked with highlighting spans within the answer that they found either useful or misleading, thereby providing a more nuanced understanding of the answer's perceived value.

The second trial featured the same evaluation tasks, except this time assessors were provided with highlighted relevant passages for each sentence in the answer, accessible by clicking on the sentence. Top-ranked passages according to DPR (Sec. 4) were selected for each sentence, ensuring the overall length is under 5000 characters to fit on the screen. Passages were displayed in their original order, grouped by source documents. Passage highlighting came from the model proposed by Bolotova et al. (2020), designed to simulate user gaze during NFQA evaluation. Assessors could freely examine the passages while determining the usefulness of the answer and selecting useful or misleading spans in the answer.

The choice of this presentation form was based on several factors. Firstly, it allowed for a more in-

[10] https://beta.openai.com/docs/models/gpt-3

[11] For interfaces and implementation details refer to App. E

Table 4: Examples of deterioration types

Deterioration Type	**Original Sentence**	**Deteriorated Sentence**
Number Swap	Sauté the onions for 5 minutes.	Sauté the onions for 15 minutes.
Sentence Negation	Turn the vehicle off and open the hood.	Don't turn the vehicle off and open the hood.
Antonyms Swap	Adjust your iron to hot for linen.	Adjust your iron to cold for linen.
Entity Swap	As a rabbi, you'll train in a branch of Judaism.	As a rabbi, you'll train in a branch of Christianity.

teractive and engaging experience for the assessors, enabling them to delve deeper into the context of each sentence. Secondly, it provided a clear visual cue to the assessors about the relevance of each sentence, thereby facilitating a more accurate and efficient evaluation process. Given that assessors often face challenges due to their unfamiliarity with the topic or potential distractions, we aimed to enhance their ability to assess with higher accuracy by providing relevant passages as supporting information. This approach, designed to be simple yet effective, prioritized the provision of relevant information over more elaborate feedback mechanisms like clicks or gaze tracking, which are potential areas for future research. After comparing different forms of passage presentation through trial runs and discussions among the authors, we found that the form offering easy accessibility by clicking on the sentence was the most effective.

As a measure of the ability to spot factual inconsistencies, we calculated the inconsistency detection rate (IDR), defined as the average percentage of identified deteriorations. A deterioration was marked as identified if it was highlighted as part of a misleading span. Results demonstrate a statistically significant difference (Student's t-test, $t - statistic = 2.57, p < 0.01$) in IDR between the first and the second trials, with the IDR scores of 0.21 and 0.35. In addition to this, we also calculated the number of False Positive IDRs. False positives were defined as answers that were marked as being misleading by an assessor, even though they were not. The first trial included four false positives, and the second trial two.

While question familiarity scores were consistent between the two trials, IDR improvement was persistent and statistically significant (Student's t-test $t - statistic = 2.18, p < 0.05$) in the second trial for both high and low familiarity. This suggests that *participants were able to assess more effectively with an access to relevant information sources, regardless of their familiarity with a particular question*. As shown in Fig. 4, there was a general improvement in detection of almost all deterioration types in the second trial. Notably, the detection rate for sentence negations and number swaps increased by 131% and 158%, respectively. These increases were statistically significant, as confirmed by Student's t-tests ($t - statistic = 4.2, p < 0.01$ for sentence negations; $t - statistic = 3.18, p < 0.01$ for number swaps). A detailed explanation of the answer deterioration process, including examples, as well as the interfaces utilized in the human study trials, is provided in App. E.

Figure 4: Inconsistency detection rate (IDR) by type

5.2 How Good Are The Baselines, Really?

We evaluated the effectiveness of lower- and upper-bound models from Sec. 4 using our human evaluation framework. Specifically, we compared models through A/B testing (Krishna et al., 2021) while incorporating highlighted relevant passages as reference sources of information. After classifying the question category as an attentiveness task, assessors were instructed to evaluate the usefulness of a pair of answers – one gold and one model – and select the more useful answer. Answers were presented in random order without disclosing the source. We provided ranked relevant passages for each sentence in each answer separately, to allow participants to make more informed decisions about factual consistency.

In this experiment, 28 workers evaluated gold answers to 100 random test questions against answers from DPR + BART, *text-davinci-003*, and DPR + *text-davinci-003*. To eliminate potential bias, we ensured that workers did not evaluate the same question twice. Each HIT was compensated with $0.5.

Human evaluation results are available in Table 3. All differences are statistically significant (Student's t-tests, $p < 0.01$). In comparison with the initial attempt at human evaluation, the simpler DPR + BART model scored lower, while still being preferred over gold answers in some cases. Unlike the automatic evaluation, human evaluation supports our initial observation that the quality of answers from *text-davinci-003* is much higher than from BART. When prompted with supporting passages, *text-davinci-003* generates significantly better answers based on human judgement (Table 3), which highlights the value of reference documents in WikiHowQA. While *text-davinci-003* mainly generated grammatically plausible and logicaly sound answers, our manual analysis revealed that it frequently failed at factual consistency and overall coherency, especially for questions that require a step-by-step instruction. Coupled with the fact that assessors still preferred answers from *text-davinci-003* in some of these cases, the proposed evaluation framework should be further improved to guide assessors better in their judgement. The problem is twofold; first, provided relevant passages do not always include the required information to fact-check the answer; second, the presentation and fluency of model answers misleads assessors even when the overall instruction is impossible to follow. This requires an additional skill-set for assessors, and we leave the research of a better evaluation framework specifically tailored to instructions for future work.

6 Conclusion

In this work we have presented a new dataset and benchmark for multi-document non-factoid question answering, WikiHowQA, sourced from the WikiHow website and consisting of passage-level answers to "how to" questions. Our released test collections fills a critical gap in currently available resources and evaluation testbeds for multi-document non-factoid QA, a widely occurring information task. We have also presented baseline performance benchmarks, and introduced and employed an information-augmented human evaluation framework that improves the reliability of QA annotations. Still, manual failure analysis revealed remaining challenges in the evaluation of convincing but factually incorrect model answers.

Acknowledgments

We extend our deepest gratitude to Mr. Luke Gallagher for his invaluable contributions to this study. His insightful comments and contributions have greatly enriched our work. We also wish to acknowledge the financial support provided by the Australian Research Council (Grant No. DP180102687), which has been instrumental in facilitating this research.

Limitations

In this section we discuss possible limitations of our work, and present interesting avenues for future investigation. First, not all documents in WikiHowQA are equally useful; some documents are overall less relevant, some contain very relevant bits alongside generally irrelevant information. More accurate passage or span selection may be required for models to generate better answers.

Another limitation is the focus on highlighted relevant passages as the method of aiding human evaluators in assessing factual correctness. While this approach helped to identify some factual inconsistencies and was proven to also make assessments faster in previous research (Bolotova et al., 2020), additional techniques should be considered for their potential to deliver further improvements.

Unfortunately, due to limitations of the evaluation interface used, we were not able capture how frequently annotators clicked on sentences to see aiding passages. Click ratio data could have yielded valuable insights into when and how often assessors referred to the provided information.

While we considered a range of answer deterioration types, it does not fully align with the kinds of hallucinations that neural NFQA models may produce in the wild. Further research is needed to analyse the robustness of these models in real-world scenarios. One direction is to adopt the comprehensive evaluation framework of abstractive neural models proposed by (Dou et al., 2022) for NFQA.

In our baselines, we experimented with text parsed directly from the HTML source code. However, it is important to note that we also provide the option of using the Markdown files, which preserve formatting information such as tables and lists. This could be useful for certain cases where formatting is important. Furthermore, some reference HTML pages contain pagination and long comment sections that we did not consider when scraping the data.

Finally, the evaluation of passage-level QA remains a challenging task, both for human and automatic evaluation frameworks. Difficulties include accurate assessment of factual correctness and overall consistency, especially when answers require complex reasoning based on multiple sources. Advancements in automatic evaluation of similar tasks, such as summarisation, could be adopted for NFQA.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Conference on Empirical Methods in Natural Language Processing*.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguistics*, 34(4):555–596.

Irshad Bhat, Talita Anthonio, and Michael Roth. 2020. Towards modeling revision requirements in wikihow instructions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8407–8414. Association for Computational Linguistics.

Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 85–94. ACM.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1196–1207, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Odellia Boni, Guy Feigenblat, Guy Lev, Michal Shmueli-Scheuer, Benjamin Sznajder, and David Konopnicki. 2021. Howsumm: A multi-document summarization dataset derived from wikihow articles. *CoRR*, abs/2110.03179.

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7282–7296. Association for Computational Linguistics.

Daniel Cohen and W. Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 143–146, New York, NY, USA. Association for Computing Machinery.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 212–219, Online. Association for Computational Linguistics.

Hoa Trang Dang. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, Sydney, Australia. Association for Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9:774–789.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7250–7274. 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. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 340–348. Tsinghua University Press.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2019. Performance prediction for non-factoid question answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 55–58, New York, NY, USA. Association for Computing Machinery.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4940–4957. 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 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

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

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Guang-He Lee and Yun-Nung Chen. 2017. MUSE: Modularizing unsupervised sense embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337, Copenhagen, Denmark. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Patrick S. H. Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1000–1008. Association for Computational Linguistics.

Kenneth C. Litkowski. 2004. Summarization experiments in duc 2004.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

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

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13666–13674.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2523–2544. Association for Computational Linguistics.

Chen Qu, Liu Yang, W. Bruce Croft, Falk Scholer, and Yongfeng Zhang. 2019. Answer interaction in non-factoid question answering systems. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*, pages 249–253. ACM.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2021. NLQuAD: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255, Online. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: factoid questions meet long-form answers. *CoRR*, abs/2204.06092.

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

Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10, New York, NY, USA. Association for Computing Machinery.

MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

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 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3556–3572, Dublin, Ireland. Association for Computational Linguistics.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikihow. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2167–2179. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Mingjun Zhao, Shengli Yan, Bang Liu, Xinwang Zhong, Qian Hao, Haolan Chen, Di Niu, Bowei Long, and Weidong Guo. 2021. Qbsum: A large-scale query-based document summarization dataset from real-world applications. *Computer Speech & Language*, 66:101166.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5905–5921. Association for Computational Linguistics.

A WIKIHOWQA dataset

Fig. 13 presents the distribution of categories and subcategories in a sunburst diagram. The inner ring represents the main categories, with the outer rings displaying the corresponding subcategories. The size of each segment in the graph represents the percentage of questions belonging to that category or subcategory, with the corresponding number of questions.

B Data Processing Details

Fig. 5 illustrates a typical human-written summary on WikiHow. Often, target answer summaries end with a sentence encouraging a reader to study the whole article, such as “To learn more, such as how to wash unfinished leather or use a washing machine, keep reading the article!”. To get a shorter, more answer-like versions of summaries, we filtered out these sentences using a set of patterns.

To simplify HTML source code of downloaded related documents, we used ReadabiliPy, a Python wrapper for Mozilla’s Readability.js[12] package[13]. We then processed HTMLs in two ways: (1) by running it through html2text[14] and saving the output in Markdown[15] format; and (2) by utilising newspaper3k[16] to extract text content directly. The former preserves formatting information useful when working with tables and lists, while the latter produces raw text that is simpler to use and analyse. Additionally, we split each Markdown file into a list of passages, and used library-link to merge all lists and tables into one passage. All titles were treated as separate passages to filter them out if necessary.

To avoid question overlap, HDBSCAN (Campello et al., 2013) was used as the clustering algorithm due to its robustness to noise. Questions were embedded using Multilingual Universal Sentence Encoder (MUSE) (Lee and Chen, 2017). “min_cluster_size” and “min_samples” were set to 2 for HDBSCAN, with the default values used for other parameters.

[12]https://github.com/mozilla/readability

[13]https://github.com/alan-turing-institute/ReadabiliPy

[14]https://github.com/aaronsw/html2text

[15]https://daringfireball.net/projects/markdown

[16]https://newspaper.readthedocs.io/en/latest/

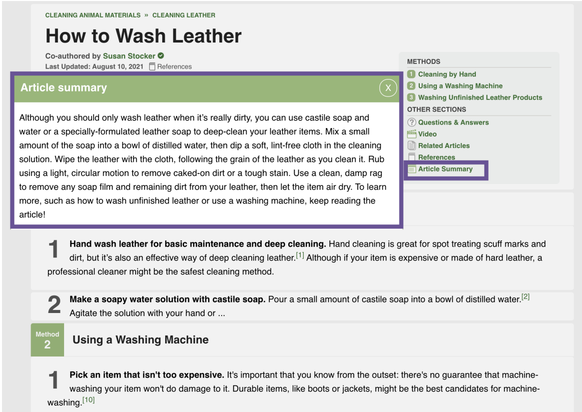

Figure 5: Human-written article summary on WikiHow

C Dataset Validation Study

Figure 6 presents the interface used in our human evaluation study to assess the feasibility of constructing answers from the provided supporting documents in our WIKIHOWQA. The example assessment shown in the figure is a random selection from the validation data collected from the workers. The detailed instructions are shown in Figure 6. The short task description presented to the workers was as follows: “Welcome to our study! We are interested in understanding how well "supporting documents" can be used to answer a question. Thank you for participating in the study. Please review the participant information form before proceeding: (App. G).”

D Model Training

Table 5 show significant hyper-parameters used to train the BART model and Table 6 contains the hyper-parameters for decoding. The number of training epochs was controlled through early stopping based on the validation loss (typically 2-4 epochs).

E Evaluating Human Evaluation Study

In this section, we first provide more details on the answer deteriorating process and then describe the interfaces used in both human study trials on evaluation assessors’ ability to identify factual inconsistencies.

E.1 Answer Deteriorations

Here we describe the methods used to create a diverse set of incorrect answers for our human evaluation experiment. Specifically, we detail the various types of factual inconsistencies we introduced, inspired by text transformations proposed by Kryscin-

Please, read the following question, answer and the supporting document:

> "**Question:** How To Book a Flight? "

Answer: To book a flight online, you can either go directly to a certain airline's website, like JetBlue.com, or visit websites that compare all airlines, like Expedia or Kayak. Follow the website's instructions for choosing your date and airport, and consider clicking the "My dates are flexible" button for a cheaper, if slightly less convenient, option. Then, follow the website prompts and include the necessary info, like your name and credit card number to purchase your ticket.

Please **select the category that best fits the question** from the list provided.

- (•) Travel
- ( ) Personal Care and Style
- ( ) Food and Entertaining
- ( ) Pets and Animals

Please, reread the question and answer carefully.

Use the **highlighting tool below to mark any words or phrases in the given supporting document that are directly related to the given answer** or that **can be used to reason or conclude the answer**. For example, if the answer is "fry for 55 minutes", you may highlight two spans "first fry for 25 minutes", "then turn", and "fry for another 30 minutes" in the supporting documents. Your highlighted text will evaluate the quality of the supporting documents and determine whether the answer can be accurately constructed from the supporting documents. Please note that it is essential only to highlight specific bits of text rather than entire sentences or paragraphs and to aim for other participants to use the highlighted parts of the documents to construct a similar answer without seeing the given answer.

Thank you for your careful attention to this task.

Highlight the text to create a new label. Click on an existing label to remove it.

Relevant

How to avoid every common mistake when booking a flight — Quartz

As the Travel Editor for CBS News, people expect that I spend weeks, even months, researching the process and logistics of travel. And I do.
But what about airfares? I only spend minutes. It's not just that I understand what makes a reasonable or crazy fare for each route. I also know WHEN to book RELEVANT .
There is a science to airfare: RELEVANT The day, time of day, and the time of the month that you book matter. RELEVANT
The number one mistake I see most travelers make is to book too early RELEVANT . Unless you are planning travel for high-traffic days, like Christmas or July 4, you stand the best chance for the lowest possible fare 45 days out for domestic travel and 60 days out for international RELEVANT .
Outside of that 45-day window RELEVANT most airline computers aren't programmed to give you any kind of a deal. It's all about computer modeling—the airline knows, for example, how many people flew on Flight 405 to Cleveland last February, and in February of 2012 as well. They make their projections of the load for this February based on that.
If you book too far in advance, you'll almost always pay a higher fare RELEVANT , and then, if the prices later drop, you can't take advantage of the lower price without incurring the standard change fee—which will easily erase any possible savings.
When I am in the appropriate booking window for my travel dates, I aim to choose off-peak days, especially over the holiday season. RELEVANT Flying midweek RELEVANT is often an effortless way to save money, and that's when you'll often find emptier and cheaper seats.
Now that I've narrowed down when I want to travel, I follow a few time-tested rule about when to book. There is a specific time if you want to get inexpensive flights.
Buy your ticket on Wednesday at 1 am, just one hour after Tuesday midnight RELEVANT . But remember, that's midnight in the time zone where the airline is based (know when to call for all US airlines).
Why Tuesday? Most low airfares seem to appear between Sunday night and Monday night. And then, when people book those fares, they have 24 hours in which to purchase them. At midnight Tuesday, all the discount fares that weren't purchased come flooding back into the airline's computer systems. And that's when you strike.
You can't do this online...You need to have a conversation with a human being at the airline. It's also how you book the best economy seats without paying extra.
If possible, I aim to book flights after the 7th of every month RELEVANT , as booking is busier right after payday on the first and 15th.
Whenever possible, I shop for fares on one computer and then clear my cookies RELEVANT or book on another computer altogether. If you want to take it a step further, go to a different location. While there's no proof or hard evidence , there is growing anecdotal evidence that suggests airlines are tracking your behavior online RELEVANT , and potentially even tracking IP addresses, and prices could go up for flights that are searched frequently.
This post is part of a series in which LinkedIn Influencers share their secrets to being more productive. See all their #productivityhacks here.

Please, **assess the relevance of the given supporting document** to the answer.

- ( ) Very Irrelevant ?
- ( ) Somewhat Irrelevant ?
- ( ) Neutral/Unsure ?
- ( ) Somewhat Relevant ?
- (•) Very Relevant ?

Figure 6: Interface for the assessment of supporting document relevance

Hyper-parameter	Value
learning rate	5e-05
train batch size	2
eval batch size	4
seed	42
gradient accumulation steps	16
total train batch size	32
optimizer (Adam) betas	0.9, 0.999
optimizer (Adam) epsilon	1e-08
lr scheduler type	linear
number of epochs	2-4 (early stopping)

Table 5: Training hyper-parameters for BART

Hyper-parameter	Value
repetition_penalty	5.0
top_k	10
top_p	0.95
temperature	1.2
no_repeat_ngram_size	2

Table 6: Decoding hyper-parameters for BART

ski et al. (2020) in their research on factually consistent models for abstractive text summarization. These include number swap, sentence negation, antonyms swap, entity swap, and paraphrasing.

- **Number Swap:** Replacing a numerical value in the original sentence with a different value, as shown in the provided example (Table 4) where the sauté time was changed from 5 minutes to 15 minutes.

- **Sentence Negation:** Rephrasing the original sentence using negations and different grammar structures, altering the meaning of the sentence and making the answer factually inconsistent, as shown in the provided example (Table 4) where the instruction to cool the pouch before removal was negated.

- **Antonyms Swap:** Replacing an adjective in the original sentence with its antonym, as shown in the provided example (Table 4) where the heat setting on the iron was changed from hot to cold.

- **Entity Swap:** Replacing a proper noun or named entity in the original sentence with a different entity, potentially altering the context of the sentence and making the answer factually inconsistent.

Note that the above methods were used to generate a large number of incorrect answers, but they were not always successful in producing grammatically or semantically correct answers. Therefore, we also manually reviewed and edited each generated answer to ensure that they were both contextually plausible and grammatically correct. For all deteriorations types, to identify adjectives, entities, numbers auxiliary verbs and Lexical verbs, we used the Spacy NLP library.

For Antonym Swap, for each identified adjective in the text an antonym is produced using the WordNet Lexical database, then the adjective in text was replaced with the randomly selected antonym. Due to the noise, and randomness, sometimes antonyms were chosen which did not best fit the sentence. Whilst manually overlooking the deterioration, these were sometimes removed or changed to a better option. For Number Swap, the digit entity is replaced at random in the code, however after manually overlooking the deterioration, the numbers were adjusted to make more contextual sense, or be harder to distinguish. For Sentence Negation, lexical verbs had “don’t” or “doesn’t” placed before them, whilst auxiliary verbs were replaced with their negation. This method at times failed to make grammatical or common sense after the deterioration for complex sentences, containing multiple clauses. In these cases, an entire sentence could have been slightly adjusted in order to accommodate the negation. For Entity Swap, named entities, often identified manually, were replaced to appropriately fit the context, but also make the answer incorrect.

E.2 Study Trials

Figure 8 illustrates an interface of the first trial run for the study with a random annotation example which happens to have an example of a zero IDR score where an assessor failed to find any deterioration. The exact instructions are presented in the Figure, while the overall task description is as follows:

“Welcome to the question-answering evaluation study! In this study, you will be asked to evaluate the quality of question-answer pairs. All of the questions will be of the "instruction" type, meaning that they will start with the phrase "How to". You will be presented with a question and asked to classify the category it belongs to. Then, you will be asked to rate your familiarity with the question topic. Finally, you will be shown an answer to the question and asked to rate its usefulness. Please take your time to carefully read each question and answer, and consider your responses before submitting them. Your feedback is important and will help us to improve the quality of question-answering systems. Thank you for participating in the study. Please, check the participant information form:(App. G)”

Figures 9 and 10 (top and bottom parts) shows an interface of the second trial run of the study with a random annotation example. The exact instructions are presented in the Figure, while the overall task description was the same as in the first trial.

Figure 7 illustrates the percentage of recognized deterioration within various question categories, and compares the results of the two trials. Most categories demonstrate an increase in detection in the second trial, with the exception of Education and Communication, Family Life, Finance and Business, Philosophy and Religion, Relationships and Youth. It is interesting to note that most of these exception categories can be grouped as relating to social life and human experience, topics for which answers are vague and often open-ended. The other two exception categories, Education and Communication, and, Finance and Business, also happened to be the two hardest categories to evaluate. It can be concluded that for the exceptions, deterioration of correctness is especially difficult to identify, or correctness of an answer is vague.

We also calculated the amount of False Positive IDRs. False positives were defined as answers that were marked as being misleading by an assessor, even though they were not. The first trial included four false positives, and the second trial two.

E.3 A/B Human Evaluation

Figures 11 and 12 demonstrate the interface used for A/B evaluation of our lower and upper-bound models, using a random annotation example for DPR + *text-davinci-003* model evaluation. Figure 11 shows Answer A, which is the answer generated by the model, and Figure 12 contains Answer B, which is an answer from the dataset. The specific instructions are provided in the figures, while the overall task description is as follows: "In this study, you will be evaluating answers to "how-to" questions. You will be presented with a question and asked to classify the category it belongs to. Then, you will be given two answers, labeled A and B, and asked to rate their usefulness and select which answer is more helpful/useful to the person asking the question. Please take your time to carefully read each question and answer and consider your responses before submitting them. Your feedback is important and will help improve the quality of question-answering systems. Thank you for participating in the study. Please, check the participant information form: (App. G).

F Benchmarking Baselines via Human Evaluations

Figure 7: IDR by categories

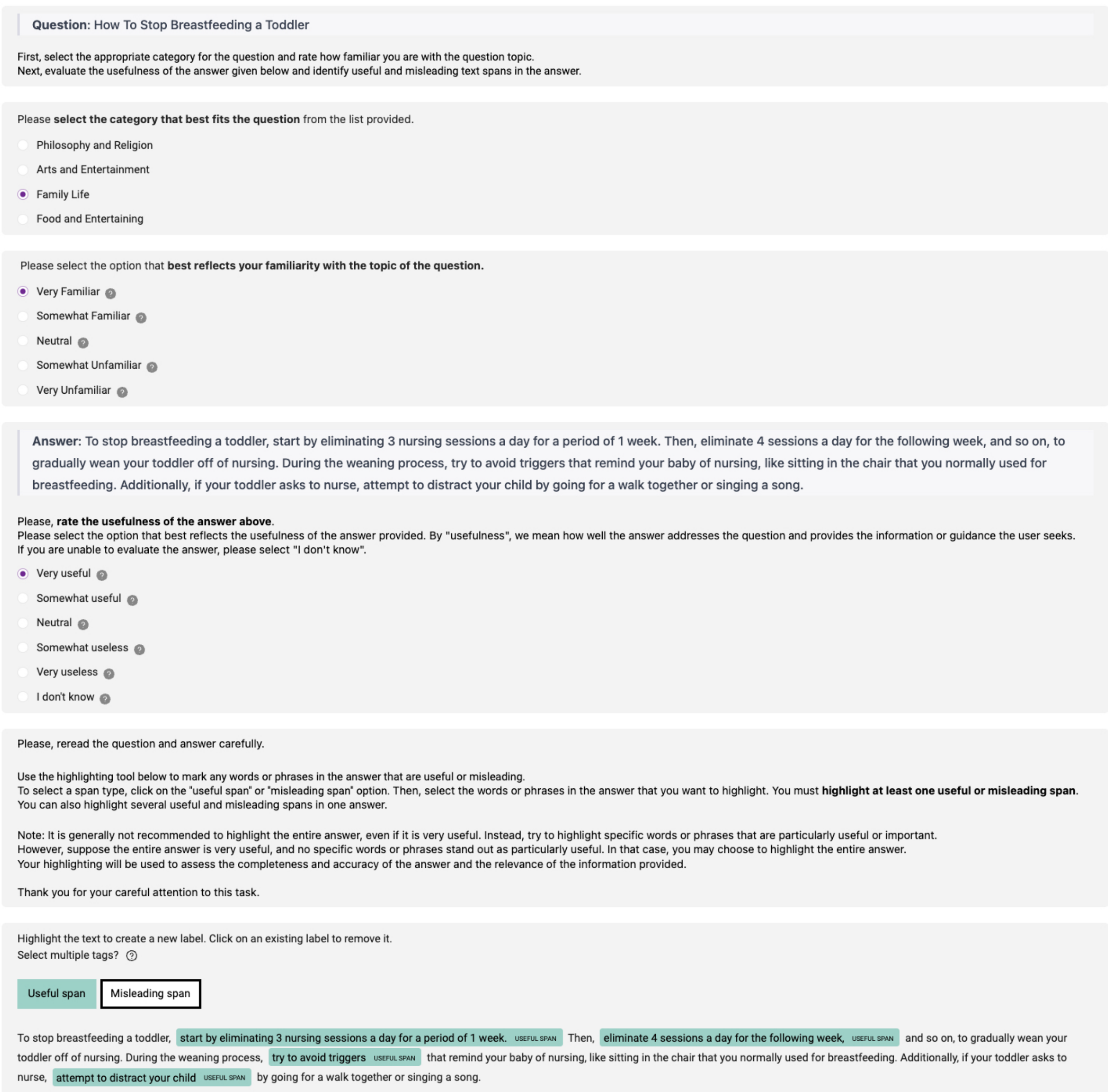

Figure 8: Interface for the first trial of the factual inconsistency identification study

Question: How To Strip Cloth Diapers

First, select the appropriate category for the question and rate how familiar you are with the question topic.
Next, evaluate the usefulness of the answer given below and identify useful and misleading text spans in the answer.

Please **select the category that best fits the question** from the list provided.

- Sports and Fitness
- Philosophy and Religion
- Pets and Animals
- Family Life

Please select the option that **best reflects your familiarity with the topic of the question.**

- Very Familiar
- Somewhat Familiar
- Neutral
- Somewhat Unfamiliar
- Very Unfamiliar

Please carefully read the provided answer below and ***rate the usefulness of the answer above.***

If you are NOT SURE about ANY facts mentioned, please, consult the supporting passages (accessible by clicking on each sentence) that are related to the question. It's important to use these passages as they will help you make sure the answer is correct and relevant. Click on each sentence to access the passages and read them. We have highlighted the most important parts of the passages for you, but make sure to read through the entire passage as other parts may also be useful.

Answer:

▶
Before you strip cloth diapers, wash and dry them like you normally would so they're clean.

▶
Then, fill a large tub or container with hot water, and add a commercial stripping agent to the water.

▶
You can also make your own by mixing equal parts washing soda, borax, and Calgon.

▼
Next, soak the diapers in the mixture for around 6 hours, stirring them around occasionally to help release the minerals in the fabric.

4.) Soak all your items in the water until it cools (at least 2 hours,) or overnight (but no more than 8 hours). Ideally, you will soak between 4 and 6 hours. 5.) Stir the items occasionally to help release more minerals from the fabrics.

...

6.) After the water is cool drain the tub, and squeeze all water from the items. 7.) Complete with a WATER ONLY wash cycle. This will make sure all the detergent and mineral solution is out of your fabrics in preparation for the bleach soak to follow.

If the cloth diapers still smell or seem to have issues such as causing rashes, repeat steps 2 and 3 (up to 3 times).

You can boil the inserts for about 5-10 minutes, making sure there is adequate water in the pot to cover the diapers completely the entire time.

Boil them on the stove for 5-10 minutes. Find the biggest pot possible, fill with water and bring to a boil. Put your inserts or diapers in the pot, stirring occasionally. Boil for 5-10 minutes, keeping the diapers submerged as much as possible in the water. Then send your diapers through a regular hot wash cycle with no detergent. Rinse. Then dry in the sun if possible. Do this only with inserts – do not put boiling water on PUL! This method is best used for inserts and prefolds.

▶
Finally, remove the diapers, and rinse them in cold water before hanging them up to air dry.

Please select the option that best reflects the usefulness of the answer provided. By "usefulness", we mean how well the answer addresses the question and provides the information or guidance the user seeks.
If you are unable to evaluate the answer, please select "I don't know".

- Very useful
- Somewhat useful
- Neutral
- Somewhat useless
- Very useless
- I don't know

Figure 9: Interface for the second trial (top part) of the factual inconsistency identification study

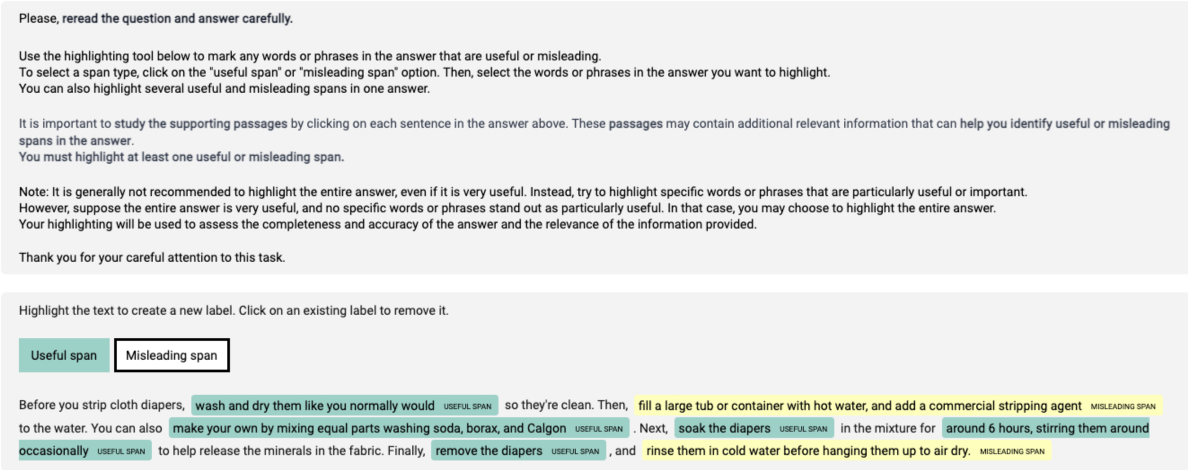

Figure 10: Interface for the second trial (bottom part) of the factual inconsistency identification study

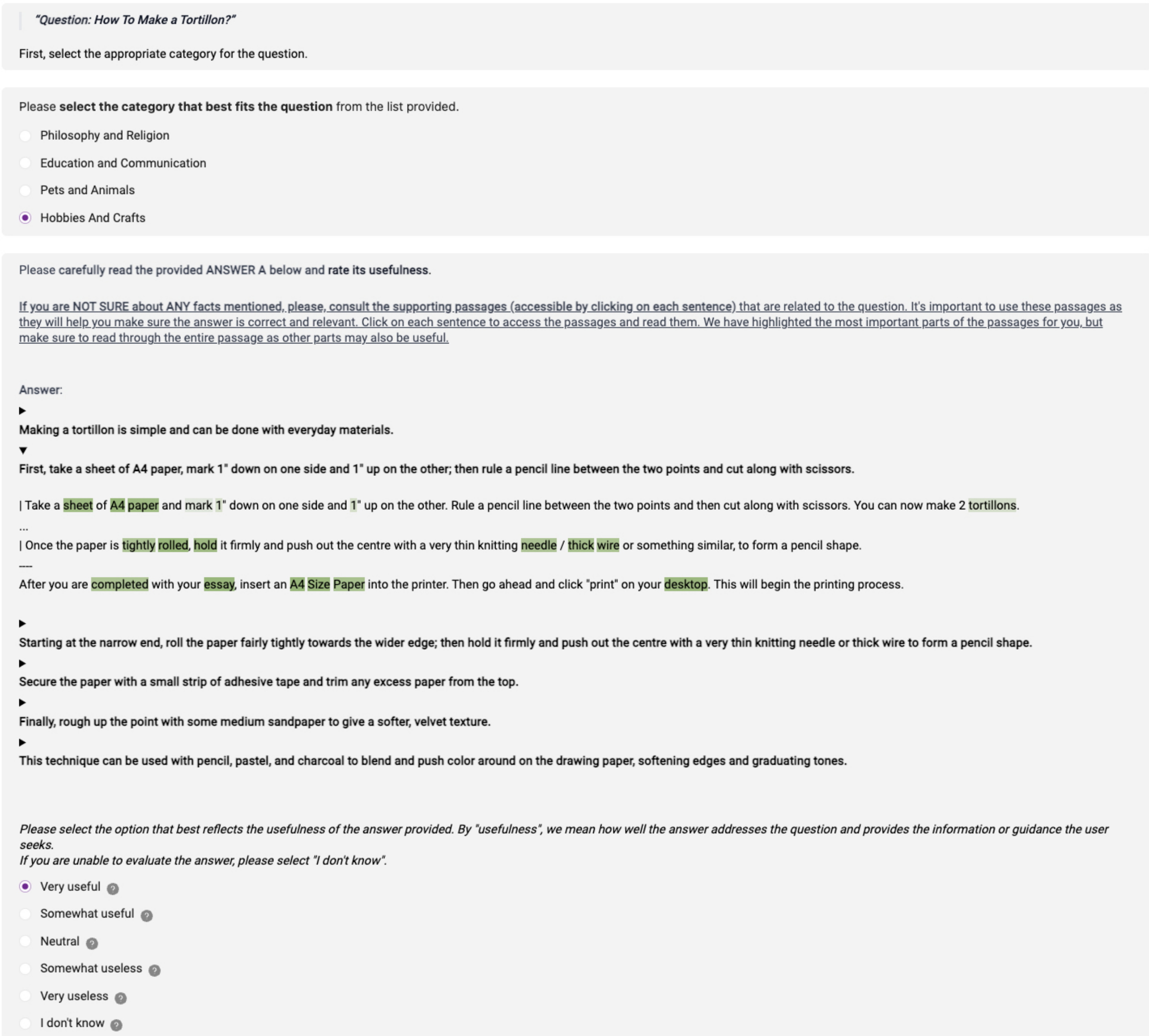

Figure 11: Interface for A/B Human Evaluation (top part)

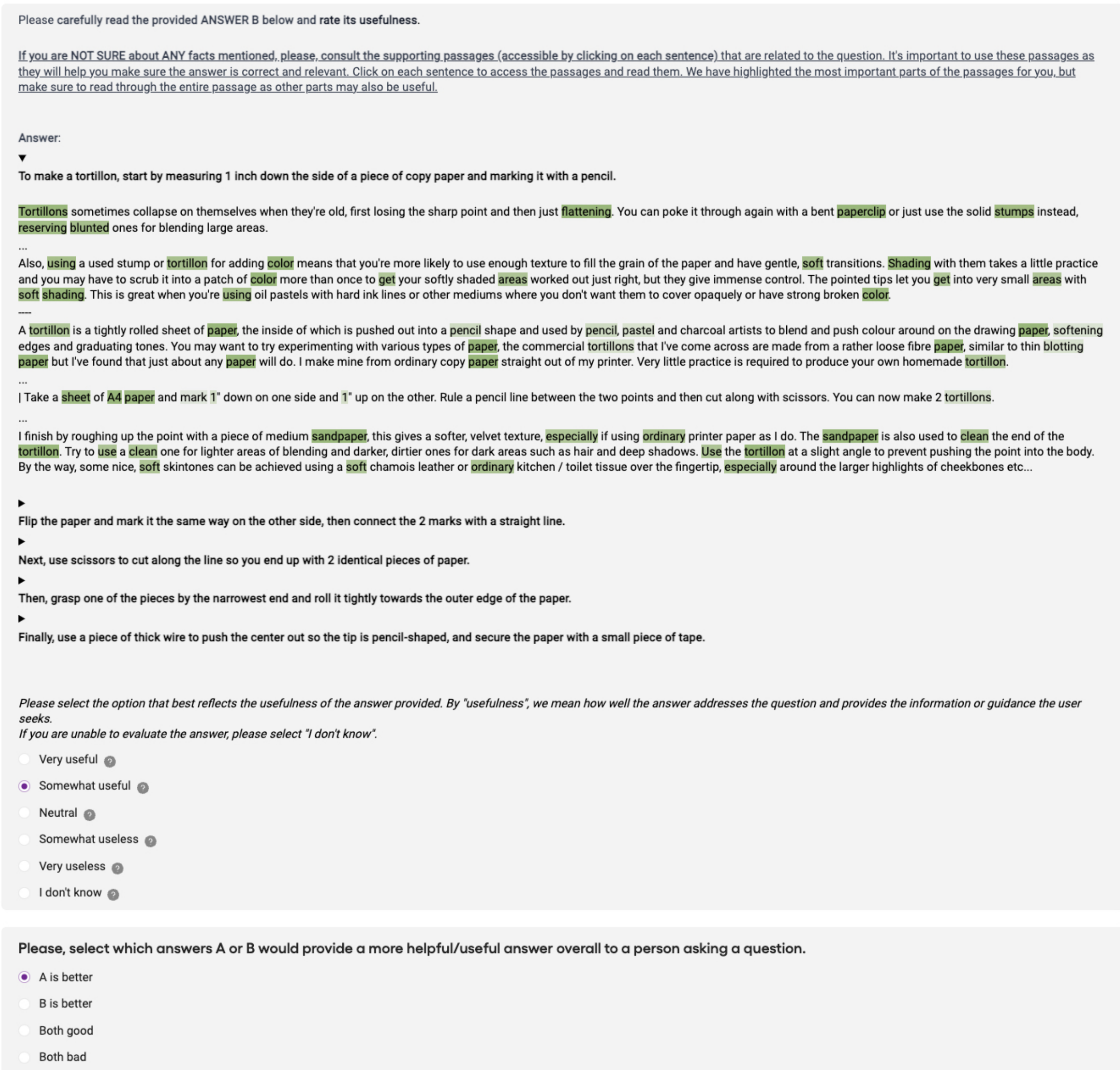

Figure 12: Interface for A/B Human Evaluation (bottom part)

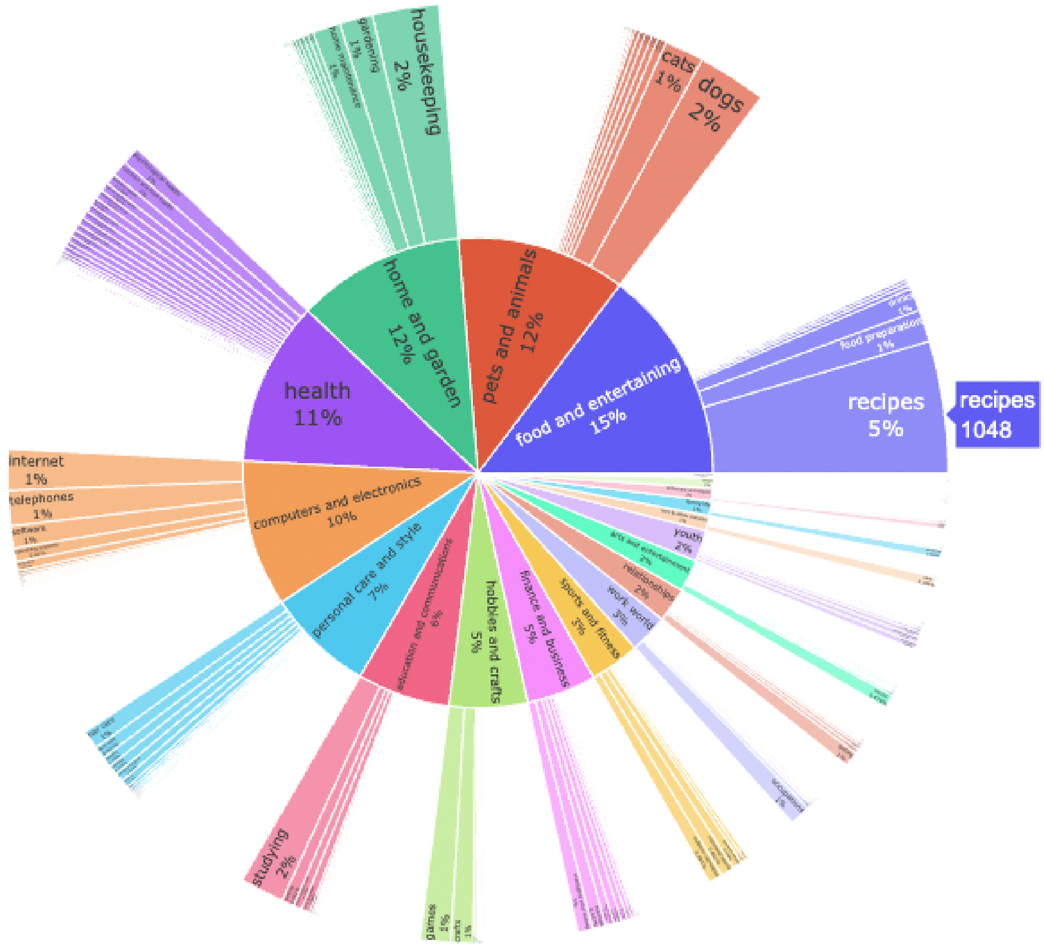

Figure 13: Category distribution of WIKIHOWQA

G Participant Information Form

INVITATION TO PARTICIPATE IN A RESEARCH PROJECT
PARTICIPANT INFORMATION

Thank you for taking part in this credibility assessment task. Before starting, please read this document carefully and be confident that you understand its contents before deciding whether to continue with your participation.

The full participant information form can be accessed here.