CONSISTENT: Open-Ended Question Generation From News Articles

Tuhin Chakrabarty[1*] **Justin Lewis**[2*] **Smaranda Muresan**[1]
[1]Department of Computer Science, Columbia University
[2]The New York Times R&D
tuhin.chakr@cs.columbia.edu, justin@justintlewis.com, smara@cs.columbia.edu

Abstract

Recent work on question generation has largely focused on factoid questions such as *who, what, where, when* about basic facts. Generating open-ended *why, how, what, etc.* questions that require long-form answers have proven more difficult. To facilitate the generation of open-ended questions, we propose **CONSISTENT**, a new end-to-end system for generating open-ended questions that are answerable from and faithful to the input text. Using news articles as a trustworthy foundation for experimentation, we demonstrate our model's strength over several baselines using both automatic and human-based evaluations. We contribute an evaluation dataset of expert-generated open-ended questions.We discuss potential downstream applications for news media organizations.

1 Introduction

Factoid questions are relatively straightforward questions that can be answered with single words or short phrases *(e.g. who, what, where, when)*. However to obtain the central idea of a long piece of text, one can ask an open-ended question *(e.g. why, how, what)* (Cao and Wang, 2021; Gao et al., 2022), which can essentially be viewed as an extreme summary of the text (Narayan et al., 2018) in the form of a question. The ability to generate such questions is particularly difficult because the generated questions must be *answerable* from and *faithful* to the given input text (see Table 1).

"*Answer-agnostic*" (Du et al., 2017; Subramanian et al., 2018; Scialom and Staiano, 2020) or "*Answer-aware*" (Lewis et al., 2021; Song et al., 2018; Zhao et al., 2018; Li et al., 2019) question generation has gained focus in NLP but these approaches are usually trained by re-purposing question answering datasets that are factual in nature or trained with trivia-like factoid QA pair data sets where answers are entities or short phrases.

	At the current rate of COVID-19 vaccination, experts say, it will take months to change the virus's trajectory. In the short term, they worry that the vaccine could present new risks if newly immunized people start socializing without taking precautions. It is not yet clear if the vaccine protects against asymptomatic infection, so vaccinated people may still be able to spread the virus to others.
Seq2Seq	Why are people so worried about the COVID-19 virus?
Seq2Seq +Control	Why is the **current rate of vaccination for COVID-19** so worrisome?

Table 1: Example of open ended questions requiring long form answer generated by fine-tuning a Seq2Seq model BART (Lewis et al., 2020) and by adding explicit control with salient n-grams

Prior work on long-form question answering (LFQA) (Kwiatkowski et al., 2019a; Fan et al., 2019) focuses on generating answers to open-ended questions that require explanations. We argue that these benchmarks can also be useful for generation of diverse, human-like open-ended question requiring long form answer.

While question generation often helps in data augmentation for training models (Lewis et al., 2021; Pan et al., 2020), it can also help in possible downstream consumer applications (Section 7). Leading news organizations often rely on human-written QA-pairs for frequently asked questions (FAQ) news tools (Figure 1) or as representative headlines for news articles used in article recommendation panels. As seen in Figure 1, a news article about *the likelihood of breakthrough infections after Covid-19 vaccination* can be summarized in the form of representative question-answer pairs.

We propose a novel end-to-end system, **CONSISTENT** for generating open-ended questions that are answerable from and faithful to the input document. We fine-tune a state-of-the-art pre-trained seq2seq model (Lewis et al., 2020) to gener-

*Work done at The New York Times R&D

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6954–6968
December 7-11, 2022

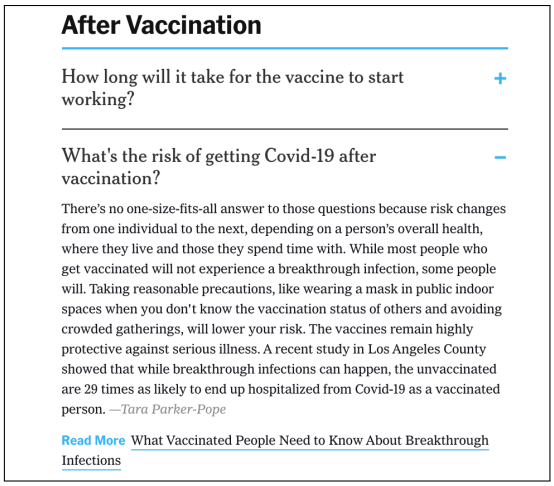

Figure 1: Human-written question-answer pairs as seen on a FAQ news tool about Covid-19 vaccination

ate open-ended questions conditioned on an input paragraph. We further propose methods to ensure better controllability and faithfulness for our generated questions by steering them towards salient keywords in the paragraph which act as "control codes" (Keskar et al., 2019). Well-formed generated questions can still be unanswerable. Prior work on using filtering methods (Lewis et al., 2021) to ensure consistency is not possible for our task, owing to increased answer length. Thus, we first rely on confidence scores obtained from pre-trained question answering models to filter out simple inconsistent questions. We further evaluate answerability by designing human-readable prompts to elicit judgements for answerability from the T0pp model (Sanh et al., 2021), which has shown good zero-shot performance on several NLP benchmarks.

We release an evaluation dataset of 529 paragraphs across diverse domains along with human written open-ended questions. Empirical evaluation using automatic metrics demonstrate that our model is better than 5 baselines. Finally, expert evaluation of the top two performing systems shows that our model is capable of generating high quality, answerable open-ended questions spanning diverse news topics (3.5 times better than a competitive baseline: a (Lewis et al., 2020, BART) model fine-tuned on an existing inquisitive questions-answers dataset ELI5 (Fan et al., 2019, Explain Like I'm Five). Our novel evaluation dataset, code and models is made publicly available at [1].

[1] https://github.com/tuhinjubcse/OpenDomainQuestionGeneration

2 Related Work

Question generation can primarily be answer-aware or answer-agnostic. Prior work on Answer-agnostic Question Generation (Du et al., 2017; Subramanian et al., 2018; Nakanishi et al., 2019; Wang et al., 2019; Scialom et al., 2019) focuses on training models that can extract phrases or sentences that are question-worthy and use this information to generate better questions. Scialom and Staiano (2020) paired questions with other sentences in the article that do not contain the answers to generate curiosity-driven questions. However, these approaches are trained by repurposing QA datasets that are factual (Rajpurkar et al., 2016) or conversational (Reddy et al., 2019; Choi et al., 2018). Cao and Wang (2021) focus on generating open-ended questions from input consisting of multiple sentences based on a question type ontology. Most recently Ko et al. (2020) built question generation models by fine-tuning generative language models on 19K crowd-sourced inquisitive questions from news articles. These questions are elicited from readers as they naturally read through a document sentence by sentence, are not required to be answerable from the given context or document.

Answer-Aware question generation models (Lewis et al., 2021; Song et al., 2018; Zhao et al., 2018; Li et al., 2019) typically encode a passage P and an answer A letting the decoder generate a question Q auto-regressively. These methods work well in practice and have been shown to be improve downstream QA performance. However despite their efficacy, these methods emphasize simple factoid questions whose answers are based on short and straightforward spans. Previous work on generating clarification questions (Rao and Daumé III, 2019, 2018; Majumder et al., 2021) uses questions crawled from forums and product reviews. The answers to the questions were used in the models to improve the utility of the generated questions.

Our work is different from prior work in that we focus on generating open-ended questions, which require long-form answers, from news articles. Unlike answer-aware question generation, where models ask a factoid question conditioned on an answer span, our task is challenging as it requires comprehension of the larger context as well as the ability to compress and represent the salient idea of the passage in the form of a question.

3 Data

It's springtime of the pandemic. After the trauma of the last year, the quarantined are emerging into sunlight, and beginning to navigate travel, classrooms and restaurants. And they are discovering that when it comes to returning to the old ways, many feel out of sorts. Do they shake hands? Hug? With or without a mask?
How are people adapting to life after the pandemic?

Table 2: Examples of our evaluation data containing paragraphs from news articles with human written questions. More in Table 9 in Appendix A

Training Data Most prior work has successfully trained models for question generation using SQUAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), or NQ (Kwiatkowski et al., 2019b) datasets, the answers to which are typically short.

To account for the open-ended nature of our desired questions, we rely on the ELI5 (Fan et al., 2019, Explain Like I'm Five) dataset. The dataset comprises 270K English-language threads in simple language from the Reddit forum of the same name[2], i.e easily comprehensible to someone with minimal background knowledge.

Compared to existing datasets, ELI5 comprises diverse questions requiring long-form answers. It contains a significant number of open-ended *how*/*why* questions. Interestingly, even *what* questions tend to require paragraph-length explanations (*What is the difference...*). As seen in Table 8 in Appendix A, each question is open-ended, inquisitive and requires an answer that is descriptive in nature. Finally, one of the advantages of the ELI5 dataset is that it covers diverse domains such as *science, health, and politics*. This quality makes ELI5 an ideal candidate to transfer to the news domain, which similarly covers a diverse range of topics.

Evaluation Data Since our goal is to generate open-ended questions from news articles, we specifically design our evaluation data to reflect the same. To achieve this goal we obtain English-language articles from *The New York Times* website from January 2020 to June 2020. We obtained written consent to use this content for research purposes by the copyright holder. One of the additional advantages of crawling data from the *The New York Times* website is that we can divide news articles by domain, as each news article appears in a specific section of the website. From the given URL[3], we can tell that the article belongs to the *Science* domain. Additionally, as most pre-trained language models were trained prior to the Covid-19 pandemic, we also test how well they generalize to COVID-19 related news topics.

Each news article from a particular domain is segmented into several paragraphs. We randomly sample **529** paragraphs spanning six domains. This includes 55 paragraphs from Science, 66 from Climate, 98 from Technology, 110 from Health, 100 from NYRegion, and 100 from Business. While we understand that selecting standalone paragraphs might sometimes ignore the greater context, or suffer from co-reference issues, we carefully replace any such paragraphs from our bigger pool.

As we do not have gold questions associated with each paragraph, we crowd-source human-written questions for each paragraph on Amazon Mechanical Turk. Each paragraph is shown to a distinct crowdworker who is then instructed to read the paragraph carefully and write an open-ended question that is answered by the entire passage. We recruit 96 distinct crowd workers for this task. After the questions are collected from first round of crowd-sourcing, two expert news media employees approve or reject them based on quality. The paragraphs with rejected questions are put up again and through this iterative process and careful quality control we obtain one high quality open-ended question associated with each paragraph. Table 2 and 9 shows selected paragraphs from our evaluation set and the associated human-generated open-ended question.

4 CONSISTENT Model

The backbone of our approach is a fine-tuned BART-large (Lewis et al., 2020) model on the ELI5 dataset of question-answer pairs. However, there are two major factors to consider in our end-to-end question generation pipeline. The generated questions i) must be *relevant and factually consistent* to the input paragraph, and ii) must have the *answer self-contained* in the input paragraph. Our CONSISTENT model (Figures 2 and 3) addresses these issues as described below.

Factual Consistency To ensure faithfulness to the input paragraph, we need to design our model

[2] https://www.reddit.com/r/explainlikeimfive/

[3] https://www.nytimes.com/2021/12/10/science/astronaut-wings-faa-bezos-musk.html

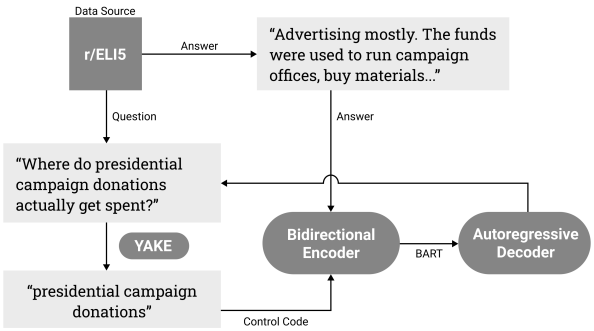

Figure 2: Architecture to train the CONSISTENT model

in such a way that the generated question is about a topic or concept mentioned in the paragraph. In traditional fine-tuning of a seq2seq model where x denotes input paragraphs in the training set and y denotes the corresponding question our goal is to learn $p_\theta(y|x)$ where

$$p_\theta(y|x) = \prod_{i=1}^{n} p_\theta(y_i|y_{i-1}, y_{i-2}.....y_1, x) \quad (1)$$

Recently Keskar et al. (2019) proposed CTRL, a conditional language model that is conditioned on a control code c and learns the distribution $p_\theta(y|x, c)$ to provide explicit control over text generation. The distribution can still be decomposed using the chain rule of probability and trained with a loss that takes the control code into account.

$$p_\theta(y|x, c) = \prod_{i=1}^{n} p_\theta(y_i|y_{i-1}, y_{i-2}.....y_1, x, c) \quad (2)$$

Owing to this modification, language models can generate text conditioned on control codes that specify domain, style, topics, dates, entities, relationships between entities, plot points, and task-related behavior. We rely on the same underlying principle for training question generation models.

During training, we extract keywords from questions and feed the input paragraph along with the extracted keyword to the encoder of BART. The extracted keyword here acts as the control code. Since we do not have any supervision for these keywords we use YAKE (Campos et al., 2020), an unsupervised keyword extraction tool. For example, as shown in Figure 2, given the question: *Where do presidential campaign donations actually get spent?*, we extract the top-most salient trigram *"presidential campaign donations"* using YAKE. We then feed the (control code, answer) to the encoder, the original question to the decoder, and fine-tune the model as shown in Figure 2.

Lewis et al. (2021) propose a BERT (Devlin et al., 2018) based answer extraction model on Natural Questions (NQ) by predicting $p(a|c) = p([a_{start}, a_{end}]|c)$ where "a" is an answer and "c" is a passage containing "a". This model first feeds the passage "c" through BERT, before concatenating the start and end token representations of all possible spans of up to length 30, and then feeds them into an MLP to compute $p(a|c)$. At generation time, the answer extraction component extracts a constant number of spans from each passage, ranked by their extraction probabilities. These extracted spans, while originally designed for a different purpose, can act here as control codes for our question generation model. To encourage the question to refer to a concept mentioned in the passage, we extract salient key phrases as control codes from the input paragraph using a combination of YAKE and the answer extraction model $p(a|c)$ (Figure 3).

It should be noted that during training the keywords are taken from the question, while during inference the keywords are produced from the article. This is because at training time we want to minimize the generation loss with respect to the training question so encouraging the model to obey the training keyword is beneficial. At inference time we do not have access to any question so using keywords from article is the only option.

Answerability Prior work (Lewis et al., 2021; Fang et al., 2020; Alberti et al., 2019) has relied on filtering methods to ensure answerability of generated questions. A filtering QA model $p_f(a|q, C)$ generates an answer for a given question. If an answer generated by p_f does not match the answer a question was generated from, the question is discarded. Such filtering methods are not applicable for our task because i) our question generation model treats the entire input paragraph as an answer instead of short answer spans typically common in ODQA tasks, and ii) the length of the answers are typically long-ranging across several sentences which is beyond the capabilities of most generative models (Krishna, 2021) and additionally would be hard for string matching purposes. We propose two filtering methods to ensure answerability: model confidence and instruction prompting.

Primary Filtering: Model Confidence QA models trained on SQUAD 2.0 (Rajpurkar et al., 2018) are capable of asserting when a question is unanswerable by signaling lower confidence scores. Taking advantage of this fact we first rely on an

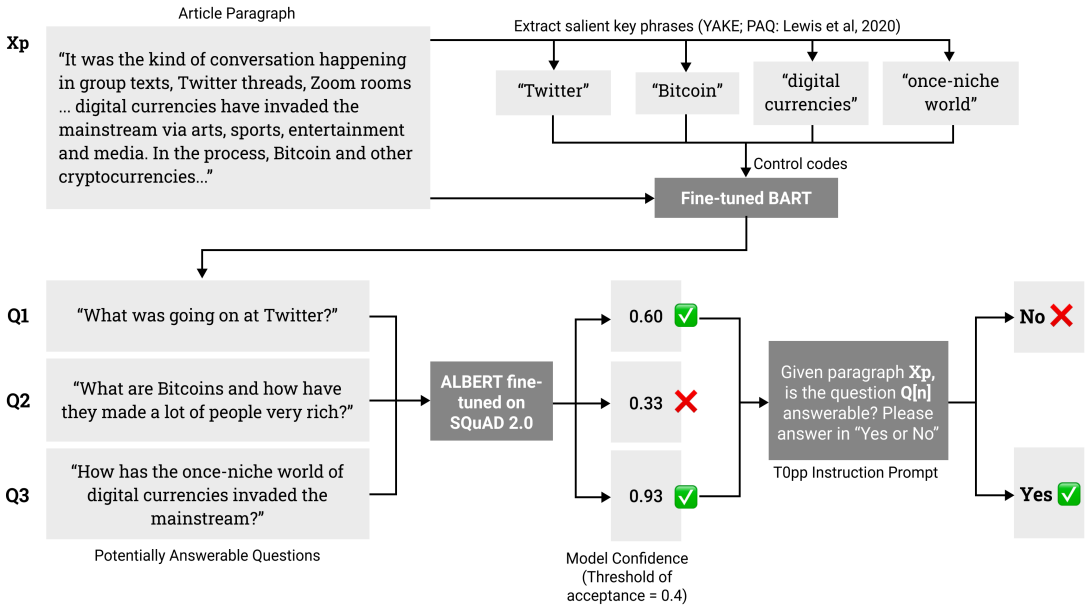

Figure 3: Inference pipeline for CONSISTENT model ensuring Factual Consitency (Control Codes) and Answerability (Model Confidence and Instruction prompting)

ALBERT-based QA model finetuned on SQUAD 2.0 [4]. The intuition behind this is such a model would typically have lower confidence scores for most poorly formed / unanswerable questions and can be used as a primary filtering step.

While it may appear that a model trained on SQuAD to determine the answerability of questions conflicts with open-ended nature of questions requiring long-form answers, it is often not the case. As we have seen in the case of Natural Questions (Kwiatkowski et al., 2019a) at least 35% question requiring a long form answer often has a short answer associated with it. This means the specific span that is returned by the SQUAD based QA model when prompted with the generated question and the input paragraph can often be an approximate short answer. For instance for the generated question Q3 in Figure 3, the SQUAD model gives an answer *via art, sports, entertainment and media.* which isn't inaccurate but requires further elucidation. This motivates us to use a model trained on SQUAD v2.0 as our initial primary step. However different questions can have different model confidence. To decide on an appropriate threshold for model confidence we observe the distribution of confidence scores. We observe a median model confidence of 0.42. We then experiment with 3 different thresholds $\kappa\epsilon\{0.35, 0.4, 0.45\}$ for selecting generated questions. The quality on a held out set of 50 generated questions by is evaluated by human judges and finally decide on a model confidence threshold $\kappa = 0.4$ such that any generated question having a confidence score below κ is discarded. It should also be noted that we tried higher values of κ between (0.6,0.9) but having such a strict high confidence score sometimes leaves us with no generated question for an input paragraph. As can be seen in Figure 3, a generated question "*What are Bitcoins and how have the made a lot of people very rich?*" while being open-ended, grammatically correct, and relevant to the input does not meet the answerability threshold and hence is discarded.

Secondary Filtering: Instruction Prompting While the above filtering step acts as excellent proxy for unanswerable questions, the original model is still trained for short answer spans. To ensure our filtering method is devoid of such biases, we use a secondary filtering approach. Recently Sanh et al. (2021) show how large language models exhibit zero-shot generalization to unseen tasks when presented with natural language prompts. As we do not have annotated data for answerability judgements for open-ended questions with longer answer spans, we rely on zero-shot prompt-based instructions for further filtering. We prompt the best-performing model from Sanh et al. (2021) *T0pp* with the following instruction:

```
Given paragraph {{paragraph}},
is the question {{question}}
```

[4] https://huggingface.co/mfeb/albert-xxlarge-v2-squad2

Input	The variant from South Africa, known as B.1.351, could make things even worse for the vaccine push. Given the speed at which the variant swept through that country, it is conceivable that by April it could make up a large fraction of infections in the United States.
BART	What's going on with the Ebola virus?
Lead	What is the name of the variant from South Africa?
SQUAD	What is B.1.351?
RandomOut	Why is the domestic product of the flu so bad right now?
RandomIn	What is going on in the US right now after a B.1 variant swept through the country?
CONSISTENT	What does the variant from South Africa mean for the vaccine push?

Table 3: Generated Questions from Baseline Models and CONSISTENT.

```
answerable? Please answer
in Yes or No
```

We feed the questions that pass the acceptability test based on our model confidence threshold as natural language instructions to the model as shown in Figure 3. Only questions which receive an answer of "Yes" are considered in our final set. This process makes our filtering approach robust, owing to the fact that only questions which pass both filtering tests are considered as *consistent*.

It can be argued that T0pp is a stochastic system that was not trained for un-answerability detection. To justify our use of T0pp we conducted a experiment where we sample a subset of 200 questions (100 answerable and 100 unanswerable). These questions are manually selected by humans from our pool of all possible generated questions. We then feed T0pp with the same prompt above containing the respective questions and their associated paragraphs. On a binary task of un-answerability prediction we get an accuracy of 84%.

As our pipeline can generate multiple questions for each input due to different control codes, we further need to rank the generated questions. Towards this task, we rank all our generated questions for a given input paragraph that are consistent based on model confidence scores.

5 Evaluation Setup

5.1 Baselines

We compare our CONSISTENT model against several baseline approaches.

Lead Sentence to Question (Lead): In order to ensure that our data is free from any potential artifacts we take the lead sentence of every passage and convert it to a question. In particular, we prompt the T0pp (Sanh et al., 2021) model which acts as a statement-to-question converter transforming the first sentence of every paragraph to a question.

QG based on fine-tuned BART (BART): Our initial backbone model of fine-tuned BART-large on answer-question pairs from the ELI5 dataset.

QG based on random keyword inside Paragraph (RandomIn): We use the same fine-tuned BART-large model from Section 4 with <keyphrase, paragraph> as input to the encoder and the question as the output from the decoder. During inference we feed a random keyphrase from the input paragraph to generate the question. It should be noted that this approach does not undergo any of the filtering mechanism used in CONSISTENT.

QG based on random keyword outside Paragraph (RandomOut): The training method is similar to that of RandomIn except that during inference we feed a random keyphrase outside of the input paragraph to generate the question. It again does not undergo any of the filtering mechanism used in CONSISTENT.

QG based on SQUAD data (SQUAD): We fine-tuned a BART-large model on SQUAD 2.0 but conditioning on the keyphrases in the prompt during inference. In particular, we use the same keyphrase used in the prompt for the highest scoring question from our CONSISTENT model. For instance, we prompt the model fine-tuned on SQUAD with the keyphrase *once-niche world* and the input paragraph as shown in Figure 3.

5.2 Evaluation Metrics

The space of possible correct outputs is too large in our case to rely on n-gram based metrics like BLEU or ROUGE. For this reason, we chose the two best available automatic evaluation metrics based on contextual representations. We report BERTScore (Zhang et al., 2020) to measure the similarity between a generated question and its gold-reference

	BLEURT	BERTScore
BART	44.0	64.5
Lead	43.0	64.4
SQUAD	39.0	62.1
RandomInside	40.0	62.1
RandomOutside	34.2	58.1
CONSISTENT	**47.0***	**66.4***

Table 4: Evaluation based on automatic metrics. *Results are significant ($p < 0.005$) via t-test.

	CONSISTENT	**BART**	**Both**	**None**
Health	**56.3**	13.6	17.2	12.7
Technology	**37.7**	22.4	22.4	17.3
Science & Climate	**42.1**	18.1	23.1	16.5
New York	**51.0**	9.0	33.0	7.0
Business	**55.0**	11.0	30.0	4.0
Overall	**48.4**	14.8	25.1	11.5

Table 5: Human-based evaluation results (percentage win). $p < .0001$ via t-test

human written question.[5]. We also report BLEURT (Sellam et al., 2020) scores, which combine expressivity and robustness by pre-training a fully learned metric on large amounts of synthetic data, before fine-tuning it on human ratings.

However, automatic metrics are not enough. To evaluate the controllability and answerability of the generated open-ended questions we chose outputs from 2 best performing systems based on the automatic evaluation. We further propose a new metric *well-formedness* and a human-based evaluation. A *well-formed* question is grammatically correct, faithful to the provided paragraph, and whose answer is detailed, long-form spanning through the entire paragraph. A well-formed question only mentions people, places, things, or ideas that are also in the original text.

Regarding human judges, Karpinska et al. (2021) discuss how even with strict qualification filters, AMT workers are not suitable for evaluating open-domain NLG outputs. To avoid such issues we recruit multiple employees of a news media organization with experience in building products and tools for news room to evaluate the output of our baseline and CONSISTENT model for each input paragraph. We believe these evaluators can ground their judgments in the real-world utility of the generated questions for our target use case. Each input was evaluated by three people. Annotator guidelines are in Appendix A. We use the Amazon SageMaker Ground Truth[6] platform where we upload our input paragraphs along with the generated questions from two systems (randomly shuffled) as shown in Figure 6. The news article headline is given for additional context to the human evaluator. The evaluators are provided with the above definition of what constitutes a *well-formed* question. The evaluators are then asked to determine which questions are well-formed between four possible options Question1, Question2, Both Questions, and Neither Question.

6 Results and Analysis

Table 4 shows that our CONSISTENT model is better than all the existing baselines. Table 5 shows that our experts agree on the quality of the generated questions spanning different domains. To get a single verdict on the correct label for each input, we consider majority voting for each question. Agreement rates were measured using Krippendorff's α and a moderate agreement of 0.62 was achieved. As observed, our CONSISTENT model outperforms Baseline BART overall by a margin of 33.6 points. Table 3 shows examples of generations by the five models on a given paragraph. In an effort to better understand why or how the CONSISTENT model is better than the baselines, we carefully analyze outputs from all systems.

Without an explicit supervision on what to ask, BART often asks generic questions or deviates from the input source and hallucinates content as can be seen in Table 3 and Table 7. The LEAD model works decently well when the central idea of the paragraph is expressed in the first sentence, however without broader context it often suffers from generating factoid or uninteresting questions. The SQUAD model is the second-worst performing model as expected due to mismatch in training and evaluation domain. Due to lack of answerability filters, the RandomIn model even though generating a question based on a random keyword from article is often found to be unanswerable as demonstrated by the automatic evaluation scores in Table 4. The worst performance of RandomOut model bolsters our claim that using keywords from the article in the prompt helps the model achieve faithfulness and doing otherwise might hurt performance.

To test the effect of the primary filtering we

[5]We used BERTScore based on *deberta-mnli* that is shown to have high correlation with human judgements.

[6]https://aws.amazon.com/sagemaker/data-labeling

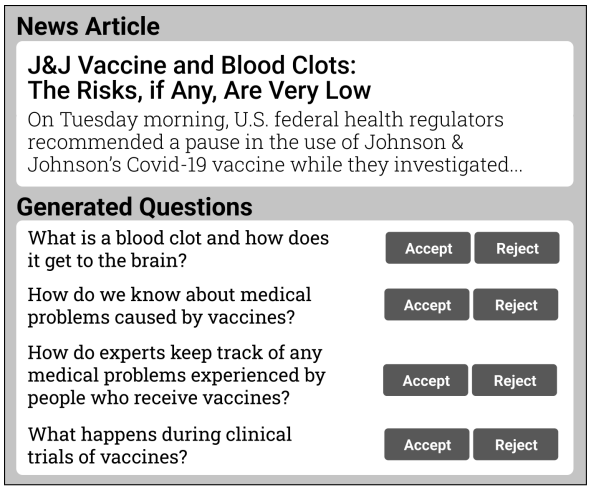

Figure 4: A prototype admin tool for humans to approve, reject, or edit questions generated by CONSISTENT for individual news articles.

Figure 5: A prototype news article where human-approved questions generated by CONSISTENT are displayed as related articles.

choose the candidate question with the highest confidence given by the QA model. To test the effect of secondary filtering we choose the candidate question with the highest confidence of generating yes. The obtained BERTScore for Primary, Secondary and CONSISTENT(both) are 64.0, 64.5 and 66.4 showing that the best filtering mechanism is the combined one.

7 Downstream Applications

We believe open-ended question generation might enhance the news experience through new Q&A tools, enhanced search, improved recommendations, and more. Media organizations have used FAQ pages to help readers better understand complex news topics, from Covid-19 vaccines[7] to personal finance[8]. The ability to automatically generate open-ended question about a given topic could make it easier for news organizations to launch an FAQ page for a new topic. We envision an admin tool (Figure 4) that presents the users with a list of generated questions and allows them to approve/reject, edit, and publish the results quickly. This human-in-the-loop approach is essential for maintaining reader trust when the generated questions may be presented directly to readers (Laban et al., 2022a).

Another potential application of this human-in-the-loop version of the system is improved news article recommendations (Figure 5). While reading a news article, a recommended article interface may be presented showing a series of questions related to the topics in the article. This level of functionality could in some ways anticipate questions a reader may ask and point them in the direction of other news articles that may provide them with answers.

For other use cases that might allow automatically generated open-ended question to be used more broadly in production systems, we envision a human-validated database of such question-answer pairs where the reliability of the results could be controlled. This can help improve the search experience on news media websites. For instance, if a user were to search *What are some of the issues with NFTs?*, the search experience could fuzzily match questions generated by CONSISTED to prioritize the article containing the answer relevant to the user's query which can help them discover things pertaining to what they are curious about. Additionally, question clustering algorithms could be deployed to better match searched questions with the generated questions. Finally automatic question generation can also used to improve interactive news podcast Laban et al. (2022b).

8 Conclusion

We propose CONSISTENT, an end-to-end system for generating open-ended questions requiring long form answers, which accounts for fac-

[7] https://www.nytimes.com/interactive/2021/well/covid-vaccine-questions.html

[8] https://www.washingtonpost.com/business/2021/12/07/faq-new-debt-collection-rules/

tual consistency and answerability. Using news articles as a trustworthy foundation for experimentation, we demonstrate CONSISTENT's strength over a competitive baseline model as evaluated both using automatic metrics and human evaluation. We also contribute an evaluation set of input paragraphs and human-generated open-ended questions. Through potential downstream applications of CONSISTENT, we demonstrate how they can enhance the experience of news media websites.

Ethical Considerations As noted in Section 3, we use a corpus of news articles from *The New York Times* as the foundational set of documents used for question generation. We have used this data with the approval and consent of the copyright holders for research purposes. We intentionally decided to use news articles as a trustworthy foundation for question generation. Further, we selected *The New York Times* as they have published[9] a clear set of ethics and standards to guide the creation of their journalism.Our models were trained on four A100 GPUs for 10 hours. Parameter size 400m.

As with other text generative models, our model can suffer from hallucinations (Reiter, 2018), biases (Sheng et al., 2019, 2021) from the Reddit ELI5 dataset and text found on the internet more broadly, and concerns about potential misuse. Much of the paper goes into detail about the great lengths we have gone to in order to reduce hallucinations and exert greater control over the final outputs in order to counter these risks (see Section 4). We use control codes selected from the original news article in an attempt to better control the generated question. We filter for answerability to further ensure that generated questions are faithful to the original text. While considerable work has been done to reduce the impact of these issues, any language generation system will be imperfect.

Our human evaluators were selected due to their familiarity with standards of journalism. Each evaluator was a paid, full-time employee of a news media organization.

To encourage critical thinking about the risks of deployment in a production environment, we included Section 7 to discuss possible downstream applications. We detailed our perspective on when a human-in-the-loop would be essential to an ethical use of this system.

We hope that our work in this paper can further the important work of safe and trustworthy language generation.

[9] https://www.nytco.com/company/standards-ethics/

9 Limitations

We note that our training dataset is automatically collected from the r/ELI5 subreddit and as such we don't account for any sensitive text. We focus on open ended question generation from news articles where our inputs are paragraph level and longer than sentence level inputs in factoid QG. However our model is not capable of handling longer sequences like an entire news article or opinion piece. We believe models like LongT5 (Guo et al., 2022) might be useful for such inputs however we leave this for future task.

Even though we control for hallucination by incorporating control codes from input text, it does not ensure 100% hallucination free output. In regards to answerability judgements our methods are useful and bridge the gap in distinguishing unanswerable questions however it in itself is a difficult task and our approaches based on SQUAD V2.0 and T0pp can still make errors.This means our models are still capable of generating unacceptable questions and should be deployed based on due deliberation.

Finally temporal misalignment is an issue and owing to the fact that our training data is from a few years back it sometimes fails on newly coined scientific terms or expressions related to COVID-19 pandemic. Continually fine-tuning our models on newer data with experience replay can mitigate these issues. We leave this for future work.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. Tuhin was funded by NYC Media Lab x New York Times R&D Researcher Fellowship. The authors also want to thank the members of NYTimes R&D team and NYC Media Lab teammates Robert Clauser, Erica Matsumoto and Matt Macvey for their support.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, A. Jorge, C. Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Inf. Sci.*, 509:257–289.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. 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 6424–6439, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. 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.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, Jingjing Liu, and Chenguang Zhu. 2020. Accelerating real-time question answering via question generation. *arXiv preprint arXiv:2009.05167*.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2022. "what makes a question inquisitive?" a study on type-controlled inquisitive question generation. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 240–257, Seattle, Washington. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555, Online. Association for Computational Linguistics.

Kalpesh Krishna. 2021. 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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019a. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022a. Quiz design task: Helping teachers create quizzes with automated question generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 102–111, Seattle, United States. Association for Computational Linguistics.

Philippe Laban, Elicia Ye, Srujay Korlakunta, John Canny, and Marti Hearst. 2022b. Newspod: Automatic and interactive news podcasts. In *27th International Conference on Intelligent User Interfaces*, pages 691–706.

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

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. 2019. Improving question generation with to the point context. 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 3216–3226, Hong Kong, China. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Mao Nakanishi, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2019. Towards answer-unaware conversational question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 63–71, Hong Kong, China. Association for Computational Linguistics.

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

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2020. Unsupervised multi-hop question answering by question generation. *arXiv preprint arXiv:2010.12623*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ehud Reiter. 2018. Hallucination in neural nlg.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6027–6032, Florence, Italy. Association for Computational Linguistics.

Thomas Scialom and Jacopo Staiano. 2020. Ask to learn: A study on curiosity-driven question generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2224–2235, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 569–574, New Orleans, Louisiana. Association for Computational Linguistics.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 78–88, Melbourne, Australia. Association for Computational Linguistics.

Siyuan Wang, Zhongyu Wei, Zhihao Fan, Yang Liu, and Xuanjing Huang. 2019. A multi-agent communication framework for question-worthy phrase extraction and question generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7168–7175.

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*.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

Annotator Guidelines As our well-formedness metric constitutes multiple dimensions it is important for us to have clear annotation guidelines. Towards this we specifically instruct workers on what should be looked into

- Question needs to be grammatical
- Question should not refer to concepts or entities that is not referenced in the original paragraph. For instance, a question about an *Ebola vaccine* when the original text is about the *COVID-19 vaccine* is NOT considered well-formed. Also a question that references *Vice President Biden* when the text is about *President Biden* would not be considered well-formed.
- Question needs to be faithful and relevant to the input paragraph and on same topic
- The question should encapsulate or summarize the key idea of the entire passage and should not be simply factoid (i.e something that can be answered using a few words)

	The variant from South Africa, known as B.1.351, could make things even worse for the vaccine push. Given the speed at which the variant swept through that country, it is conceivable that by April it could make up a large fraction of infections in the United States.
✗	What's going on with the Ebola virus?
Type	Hallucination
✗	What is the name of the variant from South Africa?
Type	Simply factoid
✗	What is B.1.351?
Type	Simply factoid
✓	What does the variant from South Africa mean for the vaccine push?
Type	Well-formed

Table 6: Generated examples used to instruct human workers as to what makes a question well-formed and why are some questions not well-formed

They were also provided with examples of generated questions along with reasoning so as to why a given question is not well-formed.

Hyperparameters We fine-tune a BART Large model on ELI5 for both baseline and CONSISTENT for 10 epochs with batch size 64 and learning rate $5e-6$ and save the best checkpoint based on validation loss. To generate questions we use top-k sampling (Fan et al., 2018) with $k = 5$ and a temperature of 0.8 coupled with $no\_repeat\_ngram\_size = 2$

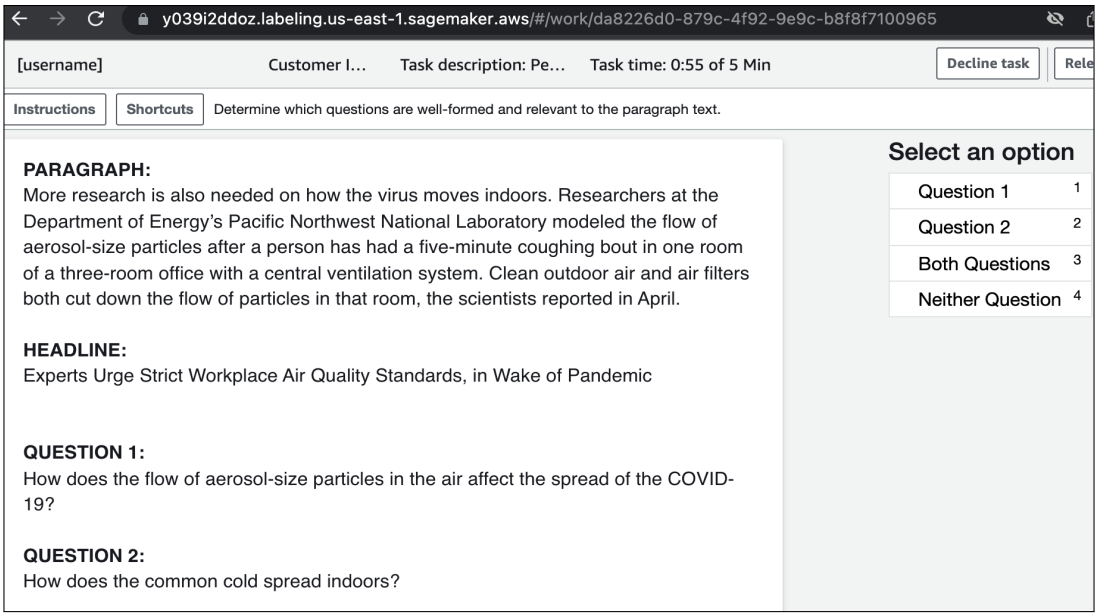

Figure 6: Screenshot of the evaluation tool where employees of a news media organization are asked to select the best option given two generated questions, an input paragraph, and headline from a news article

Input	At the two New Horizons homes in Gainesville, the medical director, Dr. Swati Gaur has held six staff town halls, in person or online, including one at 2 a.m. for the night shift, and offered rewards like free meals. About half of the workers have been vaccinated, Dr. Gaur said.
BART	How do nursing homes deal with the influx of new residents?
Consistent	What has the medical director of New Horizons done to ensure that workers are vaccinated?

Table 7: Hallucination by the Baseline BART model

Question	Answer
Why are my muscles sore after jumping in cold water?	From what I understand, our bodies defenses against hypothermia is to shiver. This involves involuntary muscle contractions to generate heat. These muscles contractions still can cause muscle soreness just like working out.
How come bluetooth is so much slower than Wi-Fi?	Bluetooth is designed to be short-range very low-power for small portable equipment. Part of the power-savings of Bluetooth come from diminished bandwidth (just as much as the weaker signal). One could speed up Bluetooth to Wi-Fi speeds, but then it would defeat the purpose of BT's major design feature. If you're looking for something that works like plunging a cable between devices but has Wi-Fi speeds, you might like wireless USB.

Table 8: Examples from the r/ELI5 Subreddit of open ended question requiring long form answer paired with human-written answers

More companies are also using augmented reality to help people with online shopping, Ms. Ask said. Jins Eyewear, which sells prescription glasses, lets you take a photo of your face to virtually try on glasses before deciding whether to buy them. Snap, the parent company for Snapchat, has teamed up with luxury brands like Gucci and Dior to offer virtual try-ons.
How are companies using AR for online shopping?
For instance, a number of U.S. colleges and universities, including the University of Arizona and the University of North Carolina at Charlotte, have used wastewater surveillance of dorms to find asymptomatic, infected students who had otherwise evaded detection. In the Netherlands, health officials have used wastewater data to determine where to send their mobile testing buses, Dr. Medema said.
How has wastewater data been used to detect symptoms?
No matter what their goals are — moving a stock, overturning a presidential election, getting the graphics on a Sonic the Hedgehog movie changed — these internet-based insurgencies tend to follow a similar pattern. One day, a group decides to take action against a system it feels is immoral or corrupt. Members identify structural weak points (a vulnerable political party, a risk-averse studio head, an overexposed short position) and figure out creative ways to exploit them, using social media for leverage and visibility. With enough highly motivated people pushing in the same direction they eventually prevail, or get enough attention that it feels like they did.
How do internet-based insurgencies gain traction?
A growing body of research shows that FEMA often helps white disaster victims more than people of color, even when the amount of damage is comparable. The problem seems to stem from complex systemic factors, like the difficulty of navigating the federal bureaucracy and a real estate market that often places higher values on properties in communities with white residents.
Why does FEMA serve more white victims?
The demands come as the safety of firefighters has become an urgent concern amid worsening effects of climate change, which bring rising temperatures that prime the nation for increasingly devastating fires. In October,2 dozen firefighters in California where a record 4.2 million acres burned across the state last year — filed suit against 3M, Chemours, E.I. du Pont de Nemours and other manufacturers, claiming that the companies for decades knowingly made and sold firefighting equipment loaded with toxic chemicals without warning of the chemicals' risks. .
Why are firefighters suing companies in California

Table 9: Human written questions from our crowd-sourced evaluation set