A Controllable QA-based Framework for Decontextualization

Benjamin Newman[♠] **Luca Soldaini**[♠] **Raymond Fok**[♡]
Arman Cohan[♠♢] **Kyle Lo**[♠]

[♠]Allen Institute for AI [♡]University of Washington [♢]Yale University
{benjaminn,lucas,armanc,kylel}@allenai.org rayfok@cs.washington.edu

Abstract

Many real-world applications require surfacing extracted snippets to users, whether motivated by assistive tools for literature survey or document cross-referencing, or needs to mitigate and recover from model generated inaccuracies. Yet, these passages can be difficult to consume when divorced from their original document context. In this work, we explore the limits of LLMs to perform decontextualization of document snippets in user-facing scenarios, focusing on two real-world settings—question answering and citation context previews for scientific documents. We propose a question-answering framework for decontextualization that allows for better handling of user information needs and preferences when determining the scope of rewriting. We present results showing state-of-the-art LLMs under our framework remain competitive with end-to-end approaches. We also explore incorporating user preferences into the system, finding our framework allows for controllability.[1]

1 Introduction

Assistive tools for cross-referencing or research activities often rely on extracting text snippets from documents and showing them to users. For example, assistive tools can use snippets to support efficient comprehension of individual documents (August et al., 2023; Fok et al., 2023) or scaffold exploration over collections of documents for literature review (Kang et al., 2022; Palani et al., 2023). With the rise in adoption of large language models (LLMs) (Brown et al., 2020; OpenAI, 2023) to power research tools, developers use extracted snippets to mitigate the potential for generated inaccuracies; snippets can help users verify model-generated outputs (Bohnet et al., 2022) and provide a means for user error recovery. However, extracted snippets are not written to be read outside the context of the original full document: they can include terms that were defined earlier, anaphora whose antecedents lie in previous paragraphs, or just generally lack context that is needed for comprehension. At best, these issues make extracted snippets difficult to read, and at worst, they render the snippets misleading outside their originating context (Lin et al., 2003; Zhang et al., 2022). We consider the potential for *decontextualization* (Choi et al., 2021)—which asks models to rewrite extracted snippets to incorporate information from their originating contexts, thereby making them "stand alone"—as a means to make extracted snippets more consumable in user-facing settings.

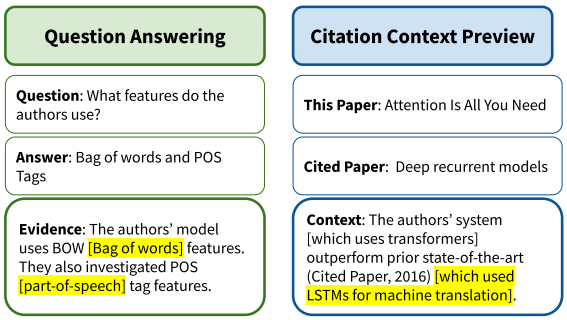

Figure 1: Overview of our user-facing decontextualization setting. We consider evidence snippets returned by a question answering system (left), as well as, citation context previews when exploring a citation graph (right). Highlighted sentences are added during the decontextualization process. Unlike prior work, these scenarios require handling multi-sentence snippets (left), handling references or links to other documents (right).

In this work, we investigate the use of LLMs for decontextualization of snippets, specifically in two real-world scenarios in which users directly consume these snippets—question answering and citation context previews of scientific documents (see Figure 1). We highlight a number of outcomes from this study: First, we recommend adjustments to the decontextualization task formulation motivated

[1]Code and data available at https://github.com/bnewm0609/qa-decontextualization

by our settings of interest: (1) expanding scope to multi-sentence passages, (2) requiring transparency of model edits for user-facing scenarios, and (3) guidelines for handling citations or references to other documents (§2).

Second, we propose a new question-answering framework to tackle decontextualization. Our approach first captures which information in the snippets need to be clarified in the form of one or more questions; then, using evidence retrieved for these queries, rewrites the snippet. The framework has distinct advantages: first, it tackles difficulty in collecting gold decontextualizations due to high annotator variability by formalizing operations a pipeline should follow (generate questions, find evidence, rewrite passage). Further, it provides a path for personalized decontextualization, as users can edit system-generated questions to better suit their information needs[2] (§3).

Third, we demonstrate our framework on a small annotation study to collect a set (n=289) of gold decontextualizations of scientific text snippets (§4). Initial evaluations of LLMs using our framework compared to end-to-end approaches. Our results show competitive results of this question-answering framework while also providing room for finer control.

2 Decontextualization for real-world, user-facing applications

We adopt the definition of decontextualization presented in Choi et al. (2021):

> Given a snippet-context pair (s, c), an edited snippet s' is a valid decontextualization of s if s' is interpretable without any additional context, and s' preserves the truth-conditional meaning of s in c.

We further take necessary departures in order to handle requirements that emerged in consideration of our (1) motivation of grounding in research assistive applications, (2) user-facing scenarios, and (3) consideration of real-world documents.

(1) Multi-sentence Passages. While Choi et al. (2021) restrict the scope of their work to single-sentence snippets, they recommend future work investigate longer snippets. We agree with this idea, especially given the real-world scenarios we consider in this work make use of datasets that contain snippets longer a single sentence. For example, in the QA dataset we use in the rest of our paper, we observed that 41% of answer snippets are longer than a single sentence, and the longest has seven. To constrain the scope of possible edits during decontextualization, we try to preserve (1) the same number of total sentences and (2) each constituent sentence's core informational content and discourse role within the larger snippet.

(2) Transparency of Edits. We require the final decontextualized snippet s' to make transparent to users what text came from the original snippet s and what text was added, removed, or modified. We draw upon well-established guidelines in writing around how to modify quotations.[3] Such guidelines include using square brackets (`[]`) to denote resolved coreferences or newly incorporated information.

(3) Decontextualizing References. Real-world documents contain references to other documents (e.g., web pages, cited works) or within-document artifacts (e.g., figures, tables). There is no single correct way to handle these references when performing decontextualization; in fact, often the extent of decontextualization is more dependent on the specific user-facing application's design rather than on intrinsic qualities of the snippet. For example, take an answering snippet in the scientific document QA setting:

> *Q: What corpus did they use?*
> *A: "We test our system on the CALLHOME Spanish-English speech translation corpus (Post et al., 2013) (§3)."*

One method of decontextualization can be:

> *A: "[Bansal et al., 2017] test [their] system on the CALLHOME Spanish-English speech translation corpus (Post et al., 2013) ["Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus" at IWSLT 2013] ~~(§3)~~."*

which drops the within-document reference to section three while adding in the title of the cited paper to provide more context, since "Post et al., 2013" may not be familiar to the user. But in the case of

[2] We argue that the high subjectivity in annotations is evidence for a need to personalize decontextualization output.

[3] APA style guide: https://apastyle.apa.org/style-grammar-guidelines/citations/quotations/changes

Title: "UTCNN: a Deep Learning Model of Stance Classificationon on Social Media Text"	
User query: "What is the size of the Chinese data?"	
Gold	For this analysis, the authors use posts [from FBFans, a single-topic Chinese unbalanced social media dataset obtained from Facebook] . They calculate the like statistics of each distinct author from these 32,595 posts.
GPT-3 (end-to-end)	The authors used posts from the FBFans dataset to analyze whether the assumption of their paper is reliable. They calculated the like statistics of each distinct author from the 32,595 posts.
Ours (QA pipeline)	The authors calculate the [normalized] like statistics of each distinct author from the 32,595 posts in the FBFans dataset [containing data from anti-nuclear-power Chinese Facebook fan groups from September 2013 to August 2014, including posts and their author and liker IDs]
Citing paper: "Question classification using head words and their hypernyms" (Huang et al., 2008)	
Cited paper: "Learning question classifiers: the role of semantic information" (Li and Roth, 2006)	
Gold	In contrast to Li and Roth (2006)'s approach which makes use of a very rich feature set [a head word feature and two approaches to augment semantic features of head words using WordNet], the authors propose to use a compact yet effective feature set [which includes five binary feature sets: question whword, head word, WordNet semantic feature for head word, word grams and word shape feature].
GPT-3 (end-to-end)	The authors propose to use a compact yet effective feature set, as opposed to Li and Roth's (2006) approach which uses a very rich feature set.
Ours (QA pipeline)	In contrast to Li and Roth (2006)'s approach [which makes use of a very rich feature set], the authors propose to use a compact yet effective feature set [including head word feature, two approaches to augment semantic features of such head words using WordNet, Lesk's word sense disambiguation (WSD) algorithm, depth of hypernym feature, unigrams, wh-word, unigram feature, and word shape feature].

Table 1: Two examples of our decontextualization pipeline compared with gold annotations and end-to-end output from GPT-3. The first example is from the QASPER dataset (Dasigi et al., 2021); the snippet is an evidence passage containing the answer the user question. The second is a text span extracted from Huang et al. (2008) citing Li and Roth (2006). Together, they demonstrate how an effective decontextualization system can improve consumption of text outside the originating document.

surfacing citation context previews, a user interface likely already surfaces the title of both citing and cited papers, in which case the addition of a title isn't useful. Possibly preferred is an alternative decontextualization that describes the dataset:

> *"[Bansal et al., 2017] test [their] system on the CALLHOME Spanish-English speech translation corpus (Post et al., 2013) [, a noisy multi-speaker corpus of telephone calls in a variety of Spanish dialects] ~~(§3)~~."*

3 Decontextualization through Question Answering

A key challenge in decontextualization is disagreement between people on (1) *what* additional information people would like to be incorporated and (2) *how* such information should be incorporated when editing from s to s' (Choi et al., 2021).

3.1 Proposed framework

We propose tackling these issues by decomposing decontextualization into three steps:

1. *Question generation.* Ask clarifying questions about the snippet.

2. *Question answering.* For each question, find

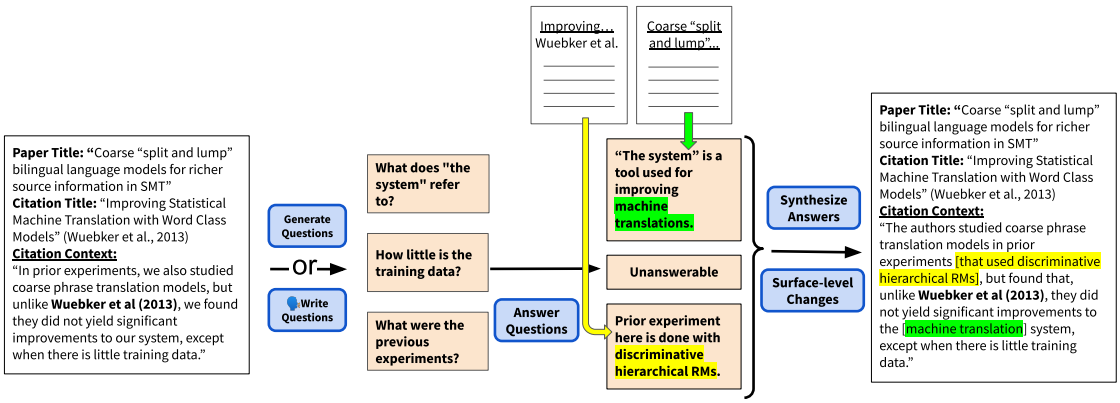

Figure 2: Overview of our implemented method to perform decontextualization of snippets from scientific papers.

an answer within the full document.

3. *Synthesize.* Rewrite the snippet by incorporating information from these QA pairs.

First, we argue questions and answers are a natural articulation of the requisite context lacking from extracted snippets. The relationship between questions and discourse relations between document passages can be traced to many works on Questions Under Discussion (QUD) (Onea, 2016; Velleman and Beaver, 2016; De Kuthy et al., 2018; Riester, 2019). Recent work has leveraged this idea to curate datasets for discourse coherence (Ko et al., 2020, 2022). We view decontextualization as a task that aims to recover from broken discourse relations through the resolution of question-answer pairs that connect portions of that snippet to the rest of the document.

Second, we argue this framing also affords greater controllability by allowing users to specify which questions they want resolved. Within user-facing applications, question answering is a natural and well-established interaction paradigm, allowing users to forage for information within documents through natural language (Wang et al., 2022; ter Hoeve et al., 2020; Jahanbakhsh et al., 2022). In the decontextualization setting, since our proposed framework is agnostic to question provenance, we can either approach the first question generation step with both automatically-generated questions (e.g., to seed a "default" experience) and/or user-provided questions for greater control and personalization.

3.2 Implementation

Here, we describe our implementation of a system for snippet decontextualization under our framework. This system is easy for practitioners to adopt, making use of widely-available LLMs as well as off-the-shelf passage retrieval models. For all LLMs below, we use GPT-3 `text-davinci-003` (Brown et al., 2020). See Figure 2 for a schematic.

Question generation. Under our framework, questions can be provided by users and/or by automated means; we discuss the latter here. We use an LLM to generate questions using zero-shot, in-context prompting, using the following prompt:

> The following text is from a scientific paper, but might include language that requires more context to understand. The language might be vague (like "their results") or might be too specific (like acronyms or jargon). Write questions that ask for clarifications. If the language is clear, write "No questions."

Using in context examples allowed us to better control the number of questions, but decreased their quality.

Question answering. Given a question (generated in the previous step or user-provided), we approach question answering in two steps: (1) We retrieve the top k relevant paragraphs from the source document, and (2) we use an LLM to process these k passages to obtain a concise answer.[4]

[4]We also considered answering questions by prompting an LLM with the full document context, but performance during

Specifically, we use a BM25 retriever for this initial retrieval step and prompt GPT-3 to use the information to answer the question using the following prompt:

> Using the following text taken from a scientific paper, answer the following question about the paragraph labeled "Context". Ignore any irrelevant information. If you cannot find the answer, write "No answer.":
>
> Title: [Title]
> Abstract: [Abstract]
> Paragraphs: [Top three paragraphs]
> Context: [Context]
> Question: [Question]

Synthesize. Finally, we provide an LLM with the snippet, questions, and answers obtained from the previous steps. Again this is done zero-shot with the following prompt:

> The following text snippet is extracted from a scientific paper. Incorporate the answers to the following questions to clarify the snippet. Only include information in the Snippet, Questions, and Answers in the new snippet you write.

We can also ablate the question component altogether by incorporating the gold contexts that were used to answer the questions into the prompt used for end-to-end experiments in addition to the title, section header, and context paragraph. See Section 6 for more details.

4 Collecting gold decontextualizations

In this section, we describe our data collection process to perform our annotation study. This had two goals. First, we wanted to verify the validity of our QA framework; that is, can we collect higher-quality annotations by aligning our annotation protocol to the QA framework, as opposed to asking annotators to perform end-to-end decontextualization? Second, we wanted to evaluate our particular implementation and compare it with an end-to-end LLM baseline.

prototyping phase showed worse results. We hypothesize this is a result of the long input context that the LLM must reason over. For example, in the citation context preview setting, the source context the LLM must reason over includes both the citing and cited paper full texts. Prior work Eisenstein et al. (2022) has also attempted LLM-based decontextualization of entire documents and noticed similar failures when dealing with longer-range contexts.

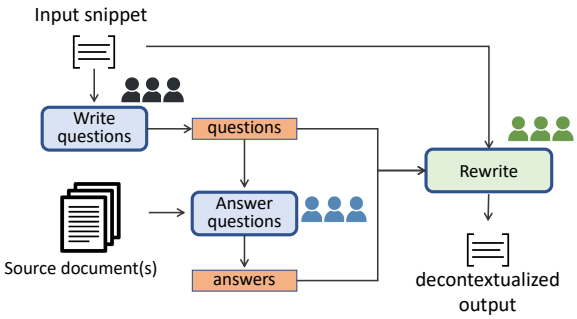

Figure 3: Overview of the data collection protocol.

4.1 Sources of snippets.

We choose two datasets in the scientific document domain to as our source of snippets, one for each motivating user-focused setting:

Question answering. We use Qasper (Dasigi et al., 2021), a dataset for document-grounded QA over scientific papers. This dataset includes QA pairs along with evidence snippets that support a given answer. We use the evidence snippets as inputs that require decontextualization.

Citation context preview. We obtain citation context snippets from NLP papers in the S2ORC (Lo et al., 2020) collection. We specifically restrict to citation contexts that contain a single citation to keep our annotation task simpler, though we note that prior work has pointed out the prevalance of contexts containing multiple citations Lauscher et al. (2022); future work could investigate how to perform decontextualization amid multiple outward references in the same snippet.

4.2 Annotation process

We closely follow our proposed framework when collecting data (see Figure 3):

1. **Writing Questions**: Given an input that requires decontextualization, we ask annotators to write clarification questions or questions that require additional information to fully understand the snippet. Given the complexity of the annotation task we used the Upwork[5] annotation platform to hire domain experts with experience in NLP.
   While piloting the question writing process, we determined that the questions that people ask fall into three categories: (1) Definitions

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

of terms or expansions of acronyms, (2) Coreference resolution, or (3) Simply seeking more context to feel more informed. We asked the annotators to categorize their questions into one of these three categories. We also asked them to label the questions they asked as either a definition, a coreference or generic "additional context" question.

2. **Answering Questions**: We hired a separate set of annotators to write answers given an input question (from previous stage) and the source document(s). We used the Prolific[6] annotation platform as a high-quality source for a larger number of annotators. Additionally, we asked the annotators to mark what evidence from the paper (or the cited paper, if one was available) supports their answer.
After this, we manually filtered down a total of 719 initial questions to a group of 487 by eliminating ones that answered the question incorrectly or found that the question could not be answered using the information in the papers.

3. **Rewriting Snippets**: In this part, given the original snippet, and the collected question-answer pairs we ask another set of annotators to rewrite the snippet by incorporating the question answer pairs to finish the decontextualization. For this part we also used the Prolific annotation platform.

4.3 Dataset statistics

In total, we obtained 289 snippets (avg. 44.2 tokens long), 487 questions (avg. 7.8 tokens long), and 487 answers (avg. 20.7 tokens long). On average, the snippets from the Qasper dataset have 1.3 questions per snippet while the citation contexts have 1.9 questions per snippet. We provide a further breakdown of annotated question types in Table 2.

5 Experimental setup

5.1 Evaluation

For evaluation, we follow (Choi et al., 2021) and use **SARI** scores. SARI was developed to evaluate text simplification models (Xu et al., 2016). It takes an original snippet, a revised snippet, and a reference and computes scores based on which n-grams changed in the revised snippet vs the reference. Following (Choi et al., 2021), we compute the SARI-add score by determining which unigrams the reference adds to the original snippet, and then calculate the F1 score between these edits and the ones that the revised snippet adds. For SARI-delete score, we determine which unigrams the reference deletes from the original snippet and then calculate the precision, recall, and F1 score between these and the ones the revised snippet deletes.

In addition to calculating these scores on the unigram level, we are also interested in evaluating whether the models add the information that people want clarified. We calculate the clarification (**CLF**) precision, recall, and F1 by considering the set of clarifications that the reference adds to the original snippet against the clarifications that the revised snippet add to the original. We determine if a clarification exists through fuzzy string matching: we extract the additions from the decontextualizations, removing stop words and stemming all of the tokens. A clarification matches if at least three quarters of the added tokens in the prediction match one of the targets.

5.2 End-to-end baseline

With the abilities of today's LLMs, the immediate question is: are current methods able to succeed at end-to-end decontextualization? To study this, we again use GPT-3 `text-davinci-003` with a prompt explaining the task along with various amounts of context, including the snippet and the source paper.

> You are a scientist in the field of natural language processing. Using the given information from a scientific paper, rewrite the given text snippet so it stands alone. To do this:
> - Remove discourse markers (like "in conclusion", "in this section", "for instance", etc.)
> - Replace first-person pronouns with third person pronouns. Replace "we" with "the authors"
> - Remove time-specific words like "current"
> - Make other surface-level changes to fix grammar
> - Replace any vague terms in the snippets
> - Define any specific terms or acronyms

In Table 3, we can see the results of running GPT-

[6] https://www.prolific.co/

	Question Type		
	Definition	Coreference	Additional Context
Question Answering	50 (31%)	59 (37%)	52 (32%)
Citation Contexts	102 (31%)	142 (44%)	82 (25%)

Table 2: Number of tokens and counts in the dataset. This combines the citation context previews and the Qasper dataset.

3 on this end-to-end task. It does not perform very well: with SARI add scores indicating low overlap with the reference and the CLF scores indicating not covering many of the same things. Though, as Choi et al. (2021) note, there are many valid decontextualizations, so the precise value of the SARI scores are not meaningful, but we can still use them to compare the systems. Qualitatively we observe that while the generations often match the right form, and the surface-level changes are made, the generations often miss clarifying the questions that annotators had. This suggests that there is still room for improvement.

We observe two challenges of the end-to-end task. The first is that giving the entire paper as input provides a lot of context, which might make it difficult to find the relevant information to clarify. The second is that models struggle to know what needs to be clarified.

To address the first challenge, we explore different ways of incorporating context that is likely to contain relevant information. (Choi et al., 2021) find that most of the sentences they decontextualize only require the Title, Section Header of the section the sentence is in, and the paragraph surrounding the snippet. For our snippets from scientific documents, this is likely not sufficient—particularly when paper-specific terms need to be defined. As such, we explore a number of different options.

- **TSP**. **T**itle, **S**ection header, and the **P**aragraph containing the snippet. This is the same condition as (Choi et al., 2021)
- **TASP** and **TAISP**. These add the **A**bstract and **I**ntroduction respectively as both of these contain much of the background context that might need to be incorporated into the snippets.

Limiting the amount of information available to the model actually helps (the CLF score for the TASP condition is slightly higher than the TAISP, potentially because there is too much distracting information in the introductions), but there are still issues about not clarifying everything that annotators wanted to be clarified (as illustrated by the low CLF scores overall).

Context	SARI add	SARI del	CLF
TSP	0.131	0.961	0.019
TASP	0.148	0.963	0.023
TAISP	0.143	0.922	0.020

Table 3: SARI add, del, and CLF scores for GPT-3 on the end-to-end task. For the context: **T** = Title, **A** = Abstract, **P** = Paragraph containing the snippet, **S** = Section header of the snippet's section, **I** = Introduction,

	SARI add	SARI del	CLF
Gold Questions	0.336	0.662	0.042
Pipeline	0.228	0.539	0.015
E2E (TASP)	0.131	0.961	0.023
E2E (QC)	0.164	0.964	0.018

Table 4: Synthesis. E2E (best) is the best End-to-end baseline. E2E (QC) is the End-to-end prompt run with the context paragraphs that are retrieved with the questions.

6 Results

6.1 Quantiative results

The main quantitative results are in Table 4. There are two takeaways.

The first is that the surface-level edits that the Pipeline system makes are better than the ones the end-to-end system makes (as illustrated by the higher SARI add scores). The SARI del scores indicate that the End-to-end systems make deletions that match the gold more often than the Pipeline system does. The most relevant parts of the text that are deleted are the discourse markers like "however", and we explore those next.

Second, we find that the story is a little less clear when looking at the information that is added us-

ing the CLF scores. The Pipeline system actually has lower CLF scores than the End-to-end system, which indicates that the information they include tends to not overlap with what the references include There are many ways to decontextualize a sentence, so the Pipeline and End-to-end appear to just be incorporating different information.

If a reader had specific information that they wanted to incorporate, how would they go about including it? In the Pipeline setting, there is a clear place - we can replace the questions that are generated with questions that people have. We do this ("Gold Questions" row of Table 4) and find that the SARI-add and CLF are much higher, because now the same questions are being answered. For a more fair comparison, we also bring in more information to the End-to-end pipeline by providing it with the gold paragraphs that answered the questions, in the "E2E (QC)" row of the table. We find that it's having the questions that matters, and this oracle End-to-end system has lower SARI and CLF scores.

6.2 Human evaluation

We also conduct human evaluations on 30 randomly chosen snippets on the End-to-end system (TSP), Pipeline, and Gold Annotations. Each snippet is first evaluated for a Pass/Fail on the decontextualization task; this is determined by whether our annotator had an unresolved clarifying question after reading the snippet. Among the snippets that Passed, the annotator then indicated which system was the most informative. The annotator also indicated circumstances when a snippet contains too much information extraneous or irrelevant information compared to the gold snippet, and when a snippet is disfluent.

We find that there is not much of a difference in the quality between the snippets output by the End-to-End and the Pipeline systems, which is in-line with quantitative results. We present some select qualitative examples in Table 1.

6.3 Human intervention

Manual inspection of the generated snippets suggested both models struggled to get match the granularity of information required by the gold annotation, which brings us back to the challenge of handling high variability in user preferences.

We first rule out one hypothesis that the question generation module is producing off-topic, unrelated questions. In fact, the questions that were generated under our method had substantial overlap with the ones that were written. After stemming and removing stop words from the questions, the average Jaccard similarity between the generated and annotator written questions was 58.1%.

We next investigate whether issues in the full automation setting could be mitigated if we had obtained questions from humans instead. We conduct a set of oracle experiments, providing gold questions to our pipeline, skipping the question generation module. In Table 6, we illustrate how our pipeline system is able to recover from generating questions that the original question annotator didn't specify, thereby resulting in a final decontextualized snippet closer to the gold annotation.

	E2E	Pipeline	Gold
Stands Alone	18	19	24
Most Informative	4	15	19
Too much Info	0	1	-
Disfluent	0	1	0

Table 5: Human evaluation results. There are 30 snippets total. E2E is end-to-end.

7 Conclusion

This work presents a framework and a system to operationalize decontextualization for real-world, user facing applications.

Our investigation begins by enumerating functional limitations of current decontextualization system. Namely, we establish three key aspects for effective decontextualization that are currently missing: ability to process an arbitrary set of sentences, resolve within-document references, and provide a transparent list of edits to users.

After listing desiderata of for an effective system, we introduce a framework to improve reliability of a decontextualization pipeline, while also enabling user personalization. The framework atomized decontextualization into three operations: asking clarifying questions about the passage yet-to-be decontextualized, identify evidence for the questions, and finally rewrite the snippet using retrieved evidence for grounding.

Finally, we collect a small set (n=289) of gold decontextualizations of scientific text snippets, and show how LLMs can leverage our framework to perform more accurate decontextualization. Fur-

System	Decontextualization	Questions
Gold	[Using the top 3 relation detectors] gives a significant performance boost, resulting in the then current state-of-the-art result on SimpleQuestions and a result comparable to the then state-of-the-art best result on WebQSP.	-
Pipeline	As shown on the last row of Table 3, which compares the results of the KBQA end-task using the top-3 relation detectors, this gives a significant performance boost, resulting in a new state-of-the-art result on SimpleQuestions [a multi-relation QA benchmark] and a result comparable to the state-of-the-art on WebQSP [a multi-relation question answering benchmark dataset].	What is Table 3? What is the new state-of-the-art result on SimpleQuestions? What is the state-of-the-art result on WebQSP?
(Oracle) Pipeline	This [using the top-3 relation detectors from Section "Relation Detection Results"] gives a significant performance boost, resulting in the then current state-of-the-art result on SimpleQuestions and a result comparable to the then state-of-the-art best result on WebQSP.	What gives a significant performance boost?

Table 6: Example of using human-provided questions (Oracle Pipeline) giving a better decontextualization than using the generated questions.

ther, we present results of a human evaluation that confirms the effectiveness of the proposed approach.

References

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans. Comput.-Hum. Interact.* Just Accepted.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *ArXiv*, abs/2212.08037.

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

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael

Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *ArXiv*, abs/2105.03011.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 476–490, New York, NY, USA. Association for Computing Machinery.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 927–936, Honolulu, Hawaii. Association for Computational Linguistics.

Farnaz Jahanbakhsh, Elnaz Nouri, Robert Sim, Ryen W. White, and Adam Fourney. 2022. Understanding questions that arise when working with business documents. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An interactive system for personalized thread-based exploration and organization of scientific literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

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.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. Discourse comprehension: A question answering framework to represent sentence connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States. Association for Computational Linguistics.

Xin Li and Dan Roth. 2006. Learning question classifiers: The role of semantic information. *Nat. Lang. Eng.*, 12(3):229–249.

Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. 2003. The role of context in question answering systems. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, page 1006–1007, New York, NY, USA. Association for Computing Machinery.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Edgar Onea. 2016. Potential questions at the semantics-pragmatics interface.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Arndt Riester. 2019. Constructing qud trees. *Questions in Discourse*.

Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, page 43–52, New York, NY, USA. Association for Computing Machinery.

Leah Velleman and David Beaver. 2016. Question-based Models of Information Structure. In *The Oxford Handbook of Information Structure*. Oxford University Press.

Sheng-Fu Wang, Shu-Hang Liu, Tian-Yi Che, Yi-Fan Lu, Song-Xiao Yang, Heyan Huang, and Xian-Ling Mao. 2022. Hammer pdf: An intelligent pdf reader

for scientific papers. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 5019–5023, New York, NY, USA. Association for Computing Machinery.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

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