SciDQA: A Deep Reading Comprehension Dataset over Scientific Papers

Shruti Singh[1] **Nandan Sarkar**[2] **Arman Cohan**[2,3]
[1]IIT Gandhinagar [2]Yale University [3]Allen Institute for AI

Abstract

Scientific literature is typically dense, requiring significant background knowledge and deep comprehension for effective engagement. We introduce SciDQA, a new dataset for reading comprehension that challenges language models to deeply understand scientific articles, consisting of 2,937 QA pairs. Unlike other scientific QA datasets, SciDQA sources questions from peer reviews by domain experts and answers by paper authors, ensuring a thorough examination of the literature. We enhance the dataset's quality through a process that carefully decontextualizes the content, tracks the source document across different versions, and incorporates a bibliography for multi-document question-answering. Questions in SciDQA necessitate reasoning across figures, tables, equations, appendices, and supplementary materials, and require multi-document reasoning. We evaluate several open-source and proprietary LLMs across various configurations to explore their capabilities in generating relevant and factual responses, as opposed to simple review memorization. Our comprehensive evaluation, based on metrics for surface-level and semantic similarity, highlights notable performance discrepancies. SciDQA represents a rigorously curated, naturally derived scientific QA dataset, designed to facilitate research on complex reasoning within the domain of question answering for scientific texts.

1 Introduction

Question-answering (QA) datasets are valuable for evaluating the reading comprehension, reasoning, and document understanding capabilities of language models (Dua et al., 2019; Dasigi et al., 2021; Rogers et al., 2023). The scientific QA task involves reading a research paper and answering questions, drawing on the paper content and some background knowledge. This task mirrors how humans engage with academic literature (Lo et al., 2023; Palani et al., 2023).

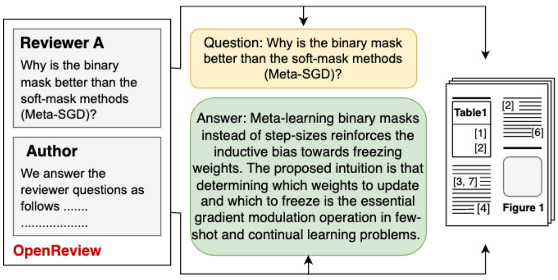


Figure 1: An instance in the SciDQA dataset. The question and answer corresponding to the paper are extracted from the reviewer-author discussion on OpenReview.

Scientific literature is inherently dense and typically requires a deep understanding and significant background knowledge to fully comprehend and engage with. To address this challenge, the NLP community has developed various datasets for question-answering (QA) from research papers to aid in development and evaluation of AI systems for comprehending the research papers. Methods range from manual question generation by domain experts (Möller et al., 2020; Dasigi et al., 2021; Lee et al., 2023) to automated extraction of questions using machine learning from selected texts (Saikh et al., 2022, 2020; Pappas et al., 2020; Jin et al., 2019; Pappas et al., 2018). However, many of these datasets focus on surface-level information and are often limited to questions that are written from titles and abstracts, which restricts the complexity and deeper engagement with the full papers.

We introduce SciDQA, a novel deep reading comprehension dataset for scientific papers. It is specifically tailored to the scientific articles in the machine learning (ML) domain and sourced from peer reviews on the OpenReview platform (OpenReview, 2023). Peer reviews frequently include questions or comments from reviewers who seek information or clarification on aspects they are confused about or do not fully understand. Answering many of such questions necessitate a deep and com-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20908–20923
November 12-16, 2024 ©2024 Association for Computational Linguistics

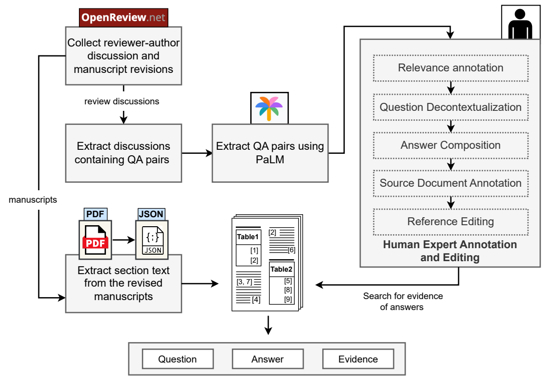


Figure 2: Dataset curation pipeline for SciDQA. LLM-based QA extraction from peer reviews is followed by a comprehensive human expert annotation and editing. As discussed, we only include evidence for a subset of the dataset due to high annotation cost.

prehensive understanding of the research and the background, such as a critical view of the approach and results, implications of the findings, and comparisons with previous works. Moreover, peer reviews are acompanied by responses from authors, who have carefully tried to address and clarify the reviewers questions. As both authors and reviewers are domain experts, responding to these inquiries necessitates a deep understanding of the paper and its broader research field. Consequently, we believe these questions are an excellent source for probing deep comprehension of research papers, contrasting with prior work that often targets shallow information-extraction or surface-level facts. However, not all such questions expressed in a review are useful. In addition they also need rewriting to stand alone as clear, self-contained queries suitable for a reading comprehension dataset. To ensure the quality and relevance of our dataset, we implement a human annotation process by domain experts, highlighted in Figure 2.

Our dataset features long-form questions and answer pairs, as shown in Table 1. It is diverse and also some questions require comprehension of figures, tables, equations, and references in addition to the paper text. Approximately 11% of the questions necessitate reasoning over at least one explicitly mentioned reference paper in addition to the candidate paper. We evaluate several open-source and proprietary large language models (LLMs) under various configurations (including retrieval-based setup and long-context reasoning) to benchmark their capabilties on this task. Our findings suggest that our dataset presents a significant challenge, as several large language models (LLMs) struggle to generate accurate factual answers across a variety of experimental setups. Our dataset, code, and model outputs to reproduce our results are available on the github repository [1].

2 Building the SciDQA Dataset

We present the pipeline for the collection of the SciDQA dataset and the preprocessing, manual filtering and rewriting steps involved. A schematic denoting the pipeline is presented in Figure 2. We present the various stages of data curation next.

2.1 Curation from OpenReview

We selected top-tier ML and DL venues, designated as A* rankings by ICORE Portal (CORE), with publicly accessible reviewer-author discussions on OpenReview (Appendix A). We curate 11400 papers from ICLR (2018-2022) and NeurIPS (2021-2022), with a major focus on including newer papers to decrease the risk of contamination with LLM pretraining datasets.

2.2 Processing the reviews

PDF to Text Conversion OpenReview portal hosts multiple submitted PDF versions of a submitted manuscript which are curated. Nougat (Blecher et al., 2023), a visual transformer model designed for scientific OCR tasks (details in Appendix A), is used for PDF to text conversion.

Regex Filtering OpenReview has nested discussions, i.e. authors and reviewers reply to messages, creating a time-stamp chain of discussion. We extract 18,658 reviewer-author discussions for 11,400 papers that contain questions and answers, by regex pattern matching (details in Appendix A).

LLM-based QA Extraction Next, we extract explicit questions that reviewers asked the authors from the reviews. For QA extraction, we utilized the PaLM text-bison-001 API (Google, 2023) to extract specific question-answer pairs within the reviewer-author discussions.[2] Initial attempts to extract questions and answers using non-LLM methods faced challenges, as authors and reviewers employ various patterns for posing questions and answers, making it difficult to comprehensively

[1] https://github.com/yale-nlp/SciDQA

[2] We chose to use PaLM because it consistently delivered high-quality extractions and offered an available API, capable of handling up to 60 requests per minute.

Dataset	Curation	Size	Source	Question length	Answer length	Multiple Docs	% Short Answers
QASA (2023)	Manual	1,554	Full-Text	15.86	44.95	×	1.61%
QASPER (2021)	Manual	5,089	Title/Abstract	9.33	18.19	×	39.94%
Covid-QA (2020)	Manual	2,019	Full-Text	10.61	15.79	×	32.64%
ScholarlyRead† (2020)	Synthetic	10,000	Abstract	NA	NA	×	NA
BioRead (2018)	Synthetic	16.4M	Full-Text	42.90	1.92	×	98.70%
BioMRC (2020)	Synthetic	700,000	Title/Abstract	16.01	1.73	×	99.38%
PubMedQA (2019)							
Annotated	Manual	1,000	Title/Abstract	14.42	43.23	×	0%*
Unlabeled	Synthetic	61,249	Title/Abstract	14.98	45.88	×	0%*
Artificial	Synthetic	211,269	Title/Abstract	16.35	40.97	×	0%*
SciDQA (Ours)	Hybrid	2,937	Full-Text	23.92	104.67	✓	1.74%

Table 1: Comparison of the related datasets. †ScholarlyRead dataset is unavailable publicly, hence we skip its statistics. *PubMedQA features two types of answers: a long answer, which is the last sentence of the abstract, and a short answer, which is yes/no. Here, we report statistics of long answers as 100% short answers are less than 5 words.

cover all instances. Through this approach, we extracted 26,085 question-answer pairs. Details of the prompts used are presented in the Appendix A.

2.3 Human Expert Annotation and Editing

In initial investigations, we found that many of the extracted questions are *not* useful and they would need additional revisions to be appropriate for a QA dataset. Therefore, to ensure the quality of the QA pairs in the SciDQA dataset, we employed an extensive manual annotation process by domain experts.[3] This included determining and keeping only the most relevant questions, rewriting both questions and answers, and editing references in the QA pairs. We briefly discuss annotation and editing stages.

Relevance Annotation This task selects information-seeking questions, whose answers are identifiable within the research paper text, from a set of synthetically generated QA pairs. Questions referencing figures, tables, equations, specific sections, or lines, and inquiries requiring data from multiple papers were categorized as relevant. Conversely, questions asking for edits, summaries, or subjective judgments about the paper's quality, or those based on the authors' personal experiences, were classified as irrelevant. To expedite the annotation process, we introduced an 'ambiguous' category for cases where the relevance of a question-answer pair was challenging to ascertain. Questions necessitating experimental validation for answers, and where it remained unclear whether the authors had conducted such experiments based on reviewer suggestion during reviewer-author discussion, were classified as ambiguous. We present a few samples for each category in Table 7 in Appendix A.

Two annotators, also the authors of this paper, annotated the dataset, starting with a common subset of 200 instances and achieving an 85% agreement rate. The disagreements were discussed and resolved, and the rest of the questions were annotated by a single annotator. In total, the annotators reviewed 7,000 instances, identifying 2,937 QA pairs as relevant, equivalent to a relevancy rate of approximately 41%. Additional details about the annotations are in Appendix A.1.

Decontextualizing Questions and Answers Originally, questions were directed towards the authors of the paper and authors provided answers from their perspective. We rewrote these QA pairs in the third-person point of view to make them universally applicable and to avoid biasing language models to generate answers in the first person when trained on SciDQA. This is also necessary for the models to understand that the question does not ask for their personal opinion, but is a factual question seeking information about the author's reasoning in the paper. We also add contextual information to the questions where the question is incomplete or incomprehensible without contextual information present in the review text. We present an example in Figure 3 showcasing scenarios where decontextualization and editing the narrative is necessary to comprehend the question. The perplexity of

[3] students with extensive experience in NLP and ML.

questions before and after rewriting, when evaluated with the GPT-2 model, exhibits a difference of 16.3 points, suggesting that decontextualization contributes to an enhancement in dataset quality.

Annotating the Source Document Certain conferences like NeurIPS and ICLR allow authors to submit revised manuscripts during the author-reviewer discussion period. For simplicity, we focus only on the initial submitted copy and the final camera-ready manuscript. For rejected papers, the last submitted manuscript is considered the final version, which may sometimes be identical to the initial submission. Establishing the source document between the initial and final manuscripts presents challenges, as author-reviewer discussions often result in added details like tables, figures, and text, making the camera-ready version a suitable source document. However, reviewers' questions may prompt authors to rewrite paper text to explicitly mention the answer, simplifying the dataset if the final version is used. We depict two such scenarios in Figure 5. To manage these variations, each question-answer pair is annotated with the version of the document used as the source, typically the initial or final version. If author responses indicate additions in a revision, the final version is marked as the source document. If no specific information is given, the initial version is defaulted as the source. This approach addresses potential ambiguities arising from updates in table, figure, and section numbers in the revised final manuscript.

Reference Editing Finally, to prevent language models from taking shortcuts by extracting answers based on reference text markers within the papers, we edited the references in the QA pairs, as shown in Figure 4. This process involved replacing specific reference markers with placeholders and providing a list of necessary references at the end of the question and the answer.

3 Dataset Details and Analysis

The SciDQA dataset comprises 2,937 question-answer pairs. We present the statistics of SciDQA in comparison to other related existing QA datasets in Table 1. Next, we discuss the diversity of answer sources, and fuzzy searching for answers, and the statistics of changes in initial and revised manuscripts.

Diversity of Answer Sources Our dataset features questions necessitating reasoning across multiple modalities beyond mere text, including figures, tables, equations, and both appendix and supplementary materials. This design ensures that comprehensive reasoning over the full-text of the paper is essential for answering the questions accurately. The statistics are presented in Table 2.

Information Source	% in Dataset
Tables	14.03%
Multiple documents	10.9%
Appendix and Supplementary	10.01%
Equations and Symbols	10.32%
Figures	6.98%

Table 2: Distribution of various modalities (text, figures, tables, equations, appendix, and supplementary) which are required to answer the questions in the dataset.

Fuzzy Search for Answers We search for answers in the research paper texts and find sections with at least 80% unigram overlap between answers and paragraphs. Such a high degree of overlap suggests that the text from the research papers is directly utilized as answers to questions, thereby simplifying the question-answering process to the identification of pertinent paragraphs. This implies a reduced necessity for reasoning or inferential thinking compared to scenarios where answers must be derived from an analysis of the text. Our findings reveal that only 25% of the answers in our dataset can be identified with an overlap exceeding 80%. By contrast, the QASA dataset (Lee et al., 2023), features 52% of answers that demonstrate more than 80% unigram overlap with the paper text, indicating a higher reliance on direct text retrieval for answering questions.

Edits in Initial and Revised Manuscripts We conducted an analysis of differences between PDF versions for each QA pair.[4] Our dataset of 576 unique papers shows that 66.3% vary in figure mentions, and 54.9% vary in table counts between initial and final manuscripts, highlighting the need to maintain separate versions.

4 Experimental Setup

We design four task configurations to evaluate the capabilities of LLMs in answering the questions in SciDQA. We experiment with open-source LLMs (Falcon (Almazrouei et al., 2023), Galactica (Taylor et al., 2022), Gemma (Team

[4]This is because authors often update their manuscripts in response to comments and questions by reviewers.

et al., 2024), Llama 2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Phi-2 (Javaheripi et al., 2023), Vicuna (Zheng et al., 2024), and Zephyr (Tunstall et al., 2023)) and two closed models Gemini (Google et al., 2023) and GPT-4 (Achiam et al., 2023).

Open-Domain Question Answering - Priming with Question (ODQA-PQ) Can LLMs generate the answer from the review text when primed with the question text without explicitly providing the paper? As the peer reviews are available publicly, there is a possibility that the LLMs might have seen the review during pretraining. Although the questions in our datasets have been significantly revised, they still originate from reviews. In this sense, it's conceivable that LLMs might have seen the reviews in their pretraining data and might be able to generate answers based solely on the question text, without any context from the associated research papers. To investigate this possibility, in the ODQA-PQ configuration, LLMs are presented with only the questions and instructions, without any information about the relevant research papers.

Open-Domain Question Answering - Priming with Question and Title/Abstract (ODQA-PTABS) In this setting, we provide the LLM with the question text, along with the Title/Abstract of the paper. The objective is to ascertain whether the inclusion of additional information, such as the paper's Title/Abstract, enhances the LLM's ability to accurately retrieve the correct answer. Unlike the Open-Domain Question Answering (ODQA) setting, it is not entirely infeasible to answer some questions with the information provided in the abstract. However, given that our dataset comprises questions that require complex reasoning, the answers to the majority of questions will not be found in the abstract alone.

Retrieval-Augmented Generation with LLMs (RAG) We follow a retrieval-augmented generation setup for this configuration. Research paper texts exceed the typical model context length with exception of few long-context models (which we will discuss in the next experimental setup). To accomodate processing such documents we employ a RAG setup, where we first retrieve the most relevant paragraphs to the question using a BM25 ranker, and subsequently input the top ranked paragraphs to the LLM, tasked with generating the response. The operational flow of this pipeline is depicted in Appendix Figure 8 and the chunking algorithm is presented in Appendix

Comprehending the Full-text using LLMs (CFT) In this experiment, LLMs are provided with the full-text of scientific papers and are tasked with answering a specific question. Given the extensive length of scientific texts, which exceeds the model's context length capacity, we divide the full-text into segments. Each segment, along with the question and instructions, is then presented to an LLM (referred to as base-LLM), which generates answers for each segment.

This setup produces multiple answer candidates for a single question, contingent on the number of passes required to present all chunks to the LLM. To distill these into a singular, optimal response, we introduce an answer selection phase. During this phase, the Gemini-pro model is prompted with the question and all answers generated by the base-LLM, with instructions to identify the most comprehensive response from the provided options. Details of this prompt are included in the Appendix B. We only segment paper's full-text when it exceeds the model's context length. For models with context length greater than the full-text, the base-LLM directly generates the answer, and the answer-selection phase is not required.

5 Results and Discussion

We use text generation metrics for evaluating the LLM generated answers, such as ROUGE score (Lin, 2004) (ROUGE-1, ROUGE-2, ROUGE-L as R-1, R-2, R-L resp.), BLEURT-20 (Pu et al., 2021) (abbreviated as BLRT), and BERTScore (Zhang* et al., 2020) (BERTScore F1 score as BS-F).

Instruction-tuned models perform better than their counterparts generally. The instruction-tuned counterparts of Gemma, Falcon, Llama 2, and Mistral perform better at retrieving the answers in ODQA-PQ and RAG setup.

Addition of Title/Abstract (ODQA-PTABS) degrades the performance for most LLMs in comparison to ODQA-PQ setup. Gemma IT, Falcon IT, Longchat 32k, Mistral, Vicuna (7B and 13B), Vicuna 16k 7B, and Llama 2 13B show a drop in performance by 7-8 points when Title/Abstract is added to the prompt. This decline could be attributed to the non-contiguous nature of the Title/Abstract with the review text, from which

Model	Prompt <INS, Q> R-1	R-2	R-L	BLRT	BS-F	Avg
			2-3 B			
Gemma (2024)	14.7	3.4	10.0	46.9	45.5	24.1
Gemma IT (2024)	19.7	4.2	13.1	41.1	50.0	25.6
Phi-2 (2023)	25.3	6.8	17.1	40.8	54.1	28.8
			6-7 B			
Falcon (2023)	7.9	1.3	6.8	48.0	38.0	20.4
Falcon IT (2023)	24.1	5.1	15.0	40.0	52.7	27.4
Galactica (2022)	6.3	1.0	4.9	45.4	40.0	19.5
Llama 2 (2023)	7.3	1.5	5.6	53.1	41.0	21.7
Llama 2 Chat (2023)	12.1	2.7	7.7	45.4	48.1	23.2
Longchat 32k (2023)	14.9	3.5	9.9	39.9	47.7	23.2
Mistral (2023)	11.1	2.4	8.0	49.2	43.2	22.8
Mistral IT (2023)	21.9	5.2	14.0	42.3	52.6	27.2
Vicuna (2024)	11.2	2.5	7.4	27.4	42.9	18.3
Vicuna 16k (2024)	14.1	3.2	9.1	40.3	47.5	22.8
Zephyr β (2023)	15.5	3.5	9.6	42.7	50.4	24.3
			13 B			
Llama 2 (2023)	8.1	1.6	6.0	54.3	41.0	22.2
Llama 2 Chat (2023)	12.6	2.8	7.8	45.0	48.9	23.4
Vicuna (2024)	15.1	3.4	9.5	40.6	48.8	23.5
Vicuna 16k (2024)	15.3	3.3	9.6	43.2	49.3	24.1
			70 B			
Llama2 (2023)	9.4	2.0	6.7	54.4	42.0	22.9
Llama2 Chat (2023)	13.4	2.9	8.4	44.7	49.0	23.7

Table 3: ODQA-PQ evaluates LLM's ability to recall the answer from reviews observed during pretraining.

Model	Prompt <INS, Q, Title+Abstract> R-1	R-2	R-L	BLRT	BS-F	Avg
			2-3 B			
Gemma (2024)	8.9	2.2	7.0	51.7	41.0	22.2
Gemma IT (2024)	13.9	2.7	11.1	18.8	39.5	17.2
Phi2 (2023)	15.5	4.0	10.6	44.7	47.9	24.5
			6-7 B			
Falcon (2023)	6.1	1.2	5.2	52.7	39.9	21.0
Falcon IT (2023)	13.8	3.0	9.3	22.8	42.4	18.3
Galactica (2022)	6.7	1.2	5.0	44.6	42.5	20.0
Llama 2 (2023)	7.6	1.6	5.8	53.8	40.4	21.8
Llama 2 Chat (2023)	18.9	2.5	13.7	35.9	48.8	24.0
Longchat 32k (2023)	2.0	0.4	1.5	9.3	34.1	9.5
Mistral (2023)	9.1	2.3	6.4	37.1	31.8	17.3
Mistral IT (2023)	24.5	6.9	16.0	39.2	53.0	27.9
Vicuna (2024)	2.6	0.5	1.9	13.7	34.3	10.6
Vicuna 16k (2024)	3.3	0.6	2.3	10.6	35.0	10.4
Zephyr β (2023)	19.8	4.7	12.1	40.6	51.9	25.8
			13 B			
Llama 2 (2023)	7.6	1.6	5.9	12.2	39.4	13.3
Llama 2 Chat (2023)	17.9	2.8	12.9	37.1	49.6	24.1
Vicuna (2024)	4.9	0.9	3.8	12.0	35.9	11.5
Vicuna 16k (2024)	18.9	4.5	11.9	36.5	49.7	24.3
			70 B			
Llama 2 (2023)	9.6	2.3	6.8	49.4	43.9	22.4
Llama 2 Chat (2023)	22.4	4.8	15.3	41.1	49.7	26.7

Table 4: ODQA-PTABS configuration evaluates if additional context helps in retrieving memorized answers.

the questions and answers are derived, leading to model confusion and negatively impacting its ability to recall answers. This highlights the sensitivity of these models to extraneous information.

Larger Chat-optimized Llama 2 models perform better when provided with more context. Model size increases in the Llama 2 model - 7B, 13B, and 70B variants — do not improve performance in ODQA-PQ setup, as no significant differences in average performance are observed. However, the Llama 2 Chat models at 13B and 70B sizes demonstrate improvements in metrics like ROUGE and BERTScore when Title/Abstract information is added (ODQA-PTABS setup). This suggests that the chat-optimized Llama 2 models are better at utilizing extended context to enhance information recall, an ability not seen with the standard Llama 2 model.

Scientific LLM Galactica performs poorly. Galactica (Taylor et al., 2022), which is the only LLM trained specifically on scientific texts (research papers, references, LaTeX, code, DNA sequences, and knowledge bases), performs poorly in comparison to most other LLMs. Galactica-generated answers were often incoherent, with repetitions, hallucinations, and noisy text. The Galactica model, initially exhibiting poor performance in ODQA-PQ setup, remains largely unchanged with the inclusion of Title/Abstract information (ODQA-PTABS) or paper chunks in RAG setup, suggesting the added details have minimal impact on its answer-recall ability. The improvement of roughly 10 points in average score in CFT setup thus can be attributed to the Gemini-Pro answer-selector.

Addition of paper chunks to prompt (RAG) improves performance over both open-domain question answering setups. An improved score is observed for all LLMs on the addition of relevant context from papers. Different-sized Llama 2 Chat models (7B, 13B, 70B), however, perform similarly, indicating that 7B performs similarly to 70B when provided with relevant context.

Gemma models produce noisy text in CFT setup. The CFT setup segments a full-text paper into chunks, from which the LLM consecutively generates answers, followed by an answer-selection

Model	Prompt <INS, Q> R-1	R-2	R-L	BLRT	BS-F	Avg
			2-3 B			
Gemma (2024)	13.7	4.1	10.9	44.5	39.4	22.5
Gemma IT (2024)	35.2	10.8	26.8	35.0	53.9	32.3
Phi-2 (2023)	9.1	1.9	5.8	33.5	36.7	17.4
			6-7 B			
Falcon (2023)	7.3	1.9	5.8	51.0	39.7	21.1
Falcon IT (2023)	27.0	7.5	19.0	42.8	50.9	29.4
Galactica (2022)	8.0	0.9	5.8	40.4	41.1	19.2
Llama2 (2023)	8.9	2.2	6.4	49.6	42.2	21.9
Llama2 Chat (2023)	35.0	9.5	26.0	35.9	53.7	32.0
Mistral (2023)	21.8	6.3	16.2	39.6	46.9	26.2
Mistral IT (2023)	33.7	10.9	23.7	39.2	54.8	32.5
Vicuna (2024)	31.3	9.0	22.5	35.3	52.6	30.1
Vicuna 16k (2024)	31.8	9.1	23.3	35.7	52.3	30.5
Zephyr Beta (2023)	24.8	6.9	15.4	40.7	53.7	28.3
			13 B			
Llama 2 (2023)	9.4	2.3	6.9	47.5	41.5	21.5
Llama 2 Chat (2023)	31.6	9.3	21.4	39.4	54.8	31.3
Vicuna (2024)	35.0	11.0	25.1	36.6	54.2	32.4
			70 B			
Llama 2 (2023)	17.2	4.8	12.4	44.7	44.6	24.7
Llama2 Chat (2023)	34.7	8.8	26.0	34.4	52.8	31.3

Table 5: RAG setup prompts the LLM with top-3 chunks extracted from the paper. RAG shows improvement in performance over ODQA-PQ and ODQA-PTABS both.

phase by the Gemini-pro model. The Gemma 2B and Gemma-Instruct 2B models perform the worst, frequently producing syntactically incorrect answers with repeated phrases, tokens, and symbols. Gemini-pro attempts to select a syntactically correct response from these candidates, but most candidates contain noisy text. The performance scores for the Gemma models in CFT are likely overestimated, as Gemini-pro often corrects formatting errors such as newlines and symbols in the Gemma-generated responses.

Answer-selection with Gemini-Pro contributes to similar performances across all LLMs in CFT setup. In the CFT setup, the average performance across all models remains relatively consistent, with significant variances observed primarily in ROUGE scores. However, comparable BERTScore suggests that while the wording may differ, the models generate answers with similar meanings. This uniformity in scores can largely be ascribed to the answer selection phase, wherein the Gemini-pro model is utilized to derive the final answers.

LLM Judge: We employ Prometheus (Kim et al., 2024) to assess LLM-generated answers on a scale of 1-5, focusing on the answer's syntactic correctness, relevance to the question, factuality, and comprehensiveness. The scores are presented in Appendix Table 8. GPT-4 achieves the highest average score of 4.05, followed by Llama 2 13B Chat with score 3.88, in CFT setup. All Vicuna models (7B and 13B both) have an average score in range 3.4-3.8. Similar to other metrics (ROUGE, BLEURT, and BERTScore), a degradation in average score is observed in ODQA-PTABS setup in comparison to ODQA-PQ setup for most LLMs.

Human Performance Estimation: Evaluating human performance on the SciDQA dataset is challenging due to the complex and domain-specific nature of its questions. To assess human proficiency, the authors compared human responses from 28 annotated QA pairs against those generated by GPT-4. Each QA pair included a question, a ground truth answer, an author-written answer, and a GPT-4 generated answer, with evaluations focusing on comprehensiveness, factuality, and clarity. Results showed that 32% of comparisons ended in a tie, indicating GPT-4's adequacy for simpler questions. Humans were preferred in 29% of cases, mainly due to factual inaccuracies in GPT-4 responses. GPT-4 outperformed humans in 21% of instances, typically in topics outside the authors' expertise. However, 18% of both answers were rejected as unsatisfactory, particularly for complex questions. Detailed performance metrics are available in the Appendix Table 9.

6 Related Work

Manually Curated Scientific QA Datasets: The QASPER dataset (Dasigi et al., 2021) involves NLP practitioners creating questions from paper titles/abstracts, with answers derived from full-texts by separate annotators. The QASA dataset (Lee et al., 2023) is generated by AI/ML practitioners and paper authors who formulate surface, testing, and deep questions. In contrast, the COVID-QA dataset (Möller et al., 2020) is crafted by 15 biomedical experts, who develop questions and annotate corresponding text as answers, focusing on COVID-19 research. QASPER has 40% questions answered in less than five words, while 30% of QASA QA pairs are sourced from only the introductions and abstracts, with 52% of answers showing high unigram overlap with the text, indicating easier retrieval. The ExpertQA dataset (Malaviya et al., 2024) features 2,177 questions across 32

Model	Prompt <INS, Q> R-1	R-2	R-L	BLRT	BS-F	Avg
2-3 B						
Gemma (2024)	27.4	7.4	21.9	24.0	44.1	25.0
Gemma IT (2024)	34.5	8.2	29.3	25.4	44.6	28.4
Phi2 (2023)	35.1	11.0	26.5	36.4	52.9	32.4
6-7 B						
Falcon (2023)	40.3	12.8	32.1	36.3	52.7	34.8
Falcon IT (2023)	30.9	8.4	22.3	38.2	51.8	30.3
Galactica (2022)	40.3	12.7	31.9	34.1	52.4	34.3
Llama 2 (2023)	44.1	14.6	34.0	36.3	55.2	36.8
Longchat 32k (2023)	39.8	11.3	30.3	34.3	53.5	33.8
Mistral (2023)	43.0	13.9	33.3	35.0	54.1	35.9
Mistral IT (2023)	36.6	10.9	27.1	34.5	53.3	32.5
Vicuna (2024)	29.8	8.7	20.0	38.9	54.3	30.3
Vicuna 16k (2024)	34.8	9.7	25.1	36.8	54.0	32.1
Zephyr β (2023)	35.0	9.6	25.4	35.8	54.0	32.0
13 B						
Llama 2 (2023)	42.2	13.9	32.0	36.5	55.0	35.9
Llama 2 Chat (2023)	26.8	7.8	17.7	41.8	53.9	29.6
Vicuna (2024)	31.1	9.0	21.2	38.3	54.2	30.8
Proprietary LLMs						
GPT-4 (2023)	28.8	8.3	17.6	40.5	55.7	30.2
Gemini Pro (2023)	26.3	8.4	17.9	31.7	53.7	27.6

Table 6: CFT evaluation - answer candidates are generated using base-LLM followed by Gemini-Pro answer selection.

fields, created by 524 experts to simulate complex, web-based information-seeking scenarios. BioASQ-QA dataset (Krithara et al., 2023; Tsatsaronis et al., 2015) involves expert-curated questions ranging from yes/no, factoid, list, and summary types, growing from 310 to 4,721 instances over ten years. Since 2016, BioASQ-QA has focused solely on titles and abstracts, reflecting the high effort in manual curation.

Synthetically Generated Scientific QA Datasets: BioRead (Pappas et al., 2018) and BioMRC (Pappas et al., 2020) are cloze-style biomedical MRC datasets that utilize text entities as answers, masking these entities in texts (passages in BioRead, abstracts in BioMRC) and forming questions from the last passage line or title. ScholarlyRead (Saikh et al., 2020) generates QA pairs by extracting noun phrases from abstracts and using a question-generation model. As shown in Table 1, these synthetically generated QA datasets generally feature shorter answers than ours. PubMedQA (Jin et al., 2019) starts with a labeled dataset where the title is a question and the last abstract line is the answer, creating 1000 instances with short answers (yes/no/maybe) annotated based on the abstract. Its synthetic counterpart uses syntax heuristics and modification rules to craft similar QA pairs.

Other datasets: The ARIES dataset (D'Arcy et al., 2023) compiles review comments and associated paper edits. Its synthetic subset uses a method similar to ours to extract comment-edit pairs based on textual overlap. Our dataset diverges by extracting questions from review comments using LLMs, not just from quoted responses but also from author rewrites. We employ human-expert annotation to refine questions and answers, avoiding reliance solely on textual overlap. This allows us to include high-quality queries involving tables, equations, and multi-paragraph reasoning. ARIES' use of GROBID versus markdown-formatted OpenReview comments results in mismatches of tables and equations.

SciDQA stands out among QA datasets as its questions are sourced directly from the peer review process, ensuring they are natural, evaluative, and of high quality due to the scientific discourse among domain experts. This sourcing guarantees that the questions require a deep understanding of the content, emphasizing depth as well as quality. Additionally, SciDQA's use of multi-document reasoning and well-formatted references provides an ideal testbed for evaluating LLMs.

7 Conclusion and Future Work

We curate SciDQA, dataset designed to challenge language models on the QA task aiming to evaluate their understanding of scientific papers. The dataset consists of 2937 QA pairs, and extracts QA asked by reviewers and answered by paper authors during reviewer-author discussion during manuscript review on OpenReview. Our multi-stage refinement pipeline annotates for quality, decontextualizes the QA pairs, edits references, and establishes the source document from different manuscript versions. Our dataset features questions necessitating reasoning across multiple modalities beyond mere text, including figures, tables, equations, appendix and supplementary materials. SciDQA also provides a testbed for evaluation of multi-document comprehension properties of LLMs. We evaluate the performance of several open-source and proprietary LLMs in generating the answer to questions after comprehending the research paper. Significant issues are observed with LLM-generated answers such as hallucinations, poorly formatted text, repetition of phrases and symbols, and infactual responses. We posit that SciDQA will serve as a useful resource to benchmark the performance of LLMs in scientific text comprehension.

8 Limitations

Multiple questions in our dataset necessitate comprehension and reasoning over multiple documents. The questions in the dataset often mention the reference text for previous works that need to be referred to for answering the question. However, in our experiments we do not search and include those documents for answer generation. Additionally, 7% questions are dependent on figures, but the Nougat parser does not extract images and only extracts the figure captions. We do not evaluate any visual or multimodal LLM. However, we extract figures for the specific figure-related questions using PDFFigures (Clark and Divvala, 2016), summarize it using GPT-4o and make that available. We also acknowledge that the evaluation of LLM-generated answers using automated metrics is not super reliable or efficient, and we employ it to bypass expensive manual annotation cost. Generating a meaningful peer review for different aspects (novelty, meaningful comparison, writing clarity, etc.) is a difficult task, and the dataset could be used to generate difficult questions from a manuscript. Our dataset does not have any judgment statements about paper acceptance/rejection. However, the questions dataset could still be used for training a question generator, and complex questions could be misused by reviewers as grounds for rejection. Finally, similar to other existing datasets, our dataset focuses on curating QA pairs from a specific domain (machine learning).

Acknowledgments

We extend our gratitude to Mike D'Arcy for providing feedback and valuable discussion on the paper. We also thank the anonymous reviewers and area chairs for their feedback. We are grateful for the compute support provided by the Microsoft Research's Accelerate Foundation Models Research (AFMR) program and Google's TRC program. Shruti is grateful for the support received through the Fulbright-Nehru Doctoral Research Fellowship to visit Yale University. Shruti is also supported by the Prime Minister's Research Fellowship (PMRF-1701251) awarded by the Government of India.

References

2008–2024. Grobid. https://github.com/kermitt2/grobid.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 143–152.

CORE. Icore conference portal. https://portal.core.edu.au/conf-ranks/. Accessed: 2024-04-15.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Google. 2023. Palm text-bison api. https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text.

Gemini Team Google, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pages 19036–19052. PMLR.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie (Yu-Yen) Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, F.Q. Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Michael Kinney, Aniket Kittur, Hyeonsu B Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Stuart Marsh, Tyler C. Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita R Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline M Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *ArXiv*, abs/2303.14334.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

OpenReview. 2023. Openreview. `https://openreview.net/about`. Accessed: 2024-04-15.

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. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Tanik Saikh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. ScholarlyRead: A new dataset for scientific article reading comprehension. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 5498–5504, Marseille, France. European Language Resources Association.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Curation of SciDQA- Data Pre-processing, Annotation and Editing

Curation from OpenReview We selected top-tier machine learning and deep learning venues, designated as A* rankings by ICORE Portal (CORE), with publicly accessible reviewer-author discussions on OpenReview. During the dataset compilation phase, NeurIPS, ICLR, ICML, and TMLR were the A* venues with available discussions. However, only discussions from ICML workshop papers and accepted papers from TMLR were accessible, with rejected papers from TMLR not being included. To ensure high quality, we excluded ICML workshop papers. Further, TMLR was also excluded to maintain diversity and avoid a narrow focus on only accepted papers. We curate 11400 papers from ICLR (2018-2022) and NeurIPS (2021-2022), with major focus to include newer papers to decrease the risk of contamination with LLM pretraining datasets.

PDF to Text Conversion OpenReview portal hosts the multiple versions of PDF files for papers submitted to ICLR and NeurIPS, which also includes the versions uploaded during the discussion phase. We downloaded the last manuscript submitted prior to the conference deadline, and refer to it as the initial version, as well as the final manuscript, known as the camera-ready version. In case of rejected manuscripts, the camera-ready version is not uploaded, and hence, we either take the latest version submitted during discussion with reviewers, or take the initially submitted manuscript. For converting PDFs to text, we employed Nougat (Blecher et al., 2023), a visual transformer model designed for the optical character recognition (OCR) task. Nougat parses research paper PDFs into markdown format and has been trained on a dataset comprising papers from arXiv, Pubmed Central, and the Industry Document Library. We opted for Nougat as it is the current state-of-the-art, showcasing superior performance in extracting tables, mathematical text (equations), and general text compared to GROBID (GRO, 2008–2024), another widely used OCR tool.

Regex Filtering OpenReview has nested discussions, i.e. authors and reviewers reply to corresponding messages, creating a time-stamp chain of discussion. Reviewers post the initial review message, generally consisting of paper summary, strengths and weakness, questions to authors, and a recommendation score. Segments of reviewer messages may be quoted in markdown or paraphrased by the authors in their replies to address specific content. Subsequently, reviewers may ask additional clarifying questions based on the authors' responses, or express satisfaction or dissatisfaction. There are instances where, despite the reviewers' questions, the authors do not provide responses. To extract nested discussions containing at least one question and answer, we employed regex pattern matching, searching for cues such as 'Question:', 'Q', etc. Using this method, we extracted 18,658 reviewer-author discussions for 11,400 papers that contained questions and answers. We use the following regex pattern to identify discussions that contain some questions:

Regex for Extraction

```
"que[ 0-9]*?[:-] .*[^\n]"
"Q[ 0-9]*?[:-] .*[^\n]"
"question[ 0-9]*?[:-] .*[^\n]"
"^> .*[^\n]"
```

LLM-based QA Extraction The prompt provided to PaLM text-bison-001 model to extract QA pairs is as follows:

Prompt for QA Extraction using PaLM

You are a helpful assistant. Read the following paragraph and find all question-answer pairs in it.

Author Response to Reviewer

Add 'Q:' before each question and 'A:' before answers. The question-answer pairs are:

A.1 Annotation details

The annotators achieved an 85% agreement rate in filtering the type of questions as relevant, irrelevant or ambiguous. Half of the disagreements pertained to the ambiguous category, with discrepancies arising from one annotator marking instances as 'ambiguous' to speed up annotation versus another favoring detailed assessment. In such cases, the annotation disagreement does not imply disagreement regarding inclusion of the instance in the dataset. The annotators resolved the remaining disagreements through discussion and refined the annotation guidelines to eliminate ambiguities before proceeding with the rest of the dataset.

The annotation process was facilitated by providing details such as the paper title, submission venue, area chair recommendations, and the extracted questions with their corresponding answers. To minimize the workload, questions from the same paper but different reviewers were assigned to the same annotator. Annotators were encouraged to consult the original review texts for additional context, enhancing the accuracy of their annotations.

We present scenarios depicting the requirement of editing QA pairs, and the references text to improve dataset quality in Figure 3 and Figure 4.

Source Document Annotation Scenarios depicting cases where initial vs revised manuscripts are appropriate for answering the reviewer questions are presented in Figure 5.

Evidence Extraction We extract evidential paragraphs, figures, tables, lines in paper text from the author responses. We also extract evidences for a smaller subset of the dataset automatically where there is high overlap between a paper section and the answer.

B Experiments

B.1 Experimental Setup

We present figures for the experimental setup of the following:

1. Open-Domain Question Answering - Priming with Questions (ODQA-PQ)
2. Open-Domain Question Answering - Priming with Question and Title/Abstract (ODQA-PTABS)
3. Retrieval Augmented Generation (RAG)
4. Comprehending the full-text (CFT)

We experimented with the parameters (temperature=0.1, 0.9, top_p=0.1, 0.5, 0.9) on a smaller subset of 20 QA pairs, and selected temperature=0.1 and top_p=0.9 after manually inspecting LLM answers. We carried out three runs initially, but upon observing no significant difference in performance, we reported the final numbers in the paper using a single run.

B.1.1 Chunk Creation Algorithm

Chunking for RAG Setup RAG setup ranks top-k chunks from the full-text which are then provided to the LLM to generate the answer. The chunking strategy is presented in Algorithm 1, and ensures that the individual chunks fit into the model context length. It also ensures that the collective top-k chunk lengths also fit the model context length, and sentences from different paper sections or paragraphs are not collated together in a single chunk. We found this setting to work better than naive chunking and truncating by paragraphs or sections.

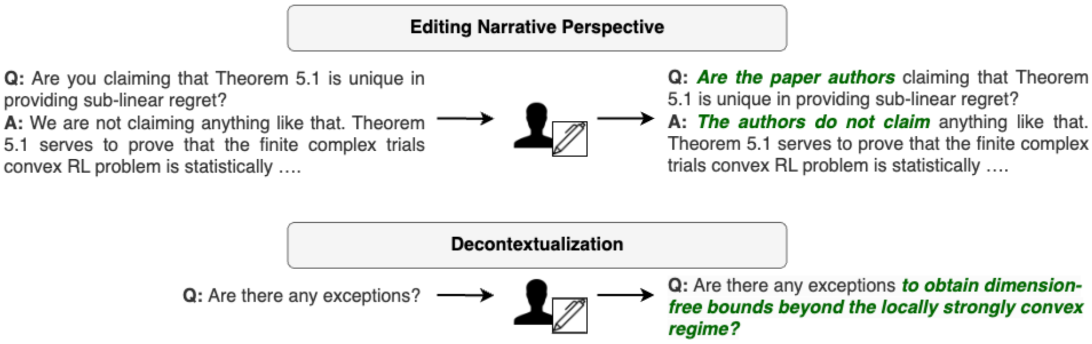


Figure 3: Rewriting QA pairs in a third-person narrative is crucial for models to recognize that questions seek factual answers based on the author's reasoning in the paper, rather than personal opinions. Furthermore, incorporating contextual information enhances the comprehension of questions that necessitate prior contextual knowledge for accurate interpretation.

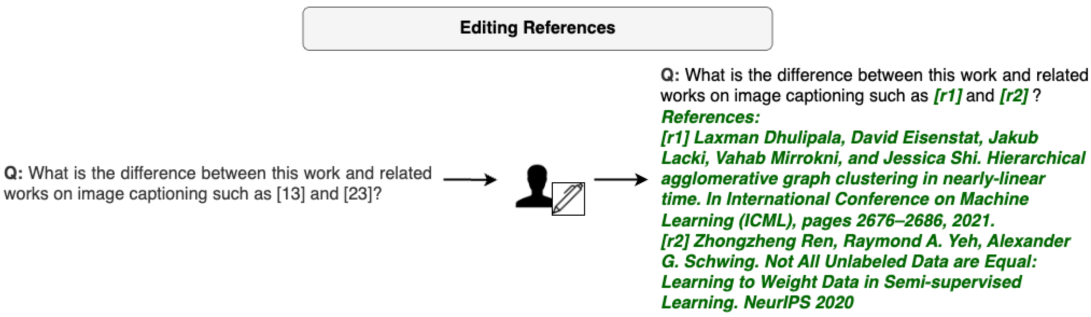
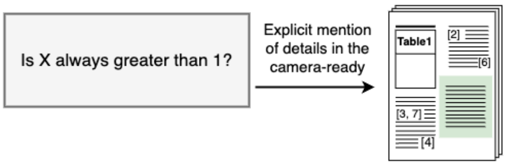


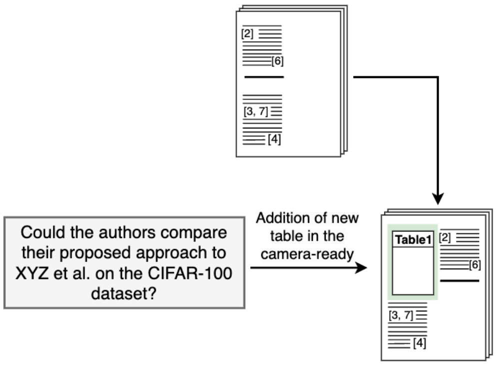
Figure 4: References in question and answer texts are uniformly renumbered (e.g., r1, r2, or 1, 2, or A, B) to preclude the LM from leveraging specific reference markers as shortcuts for answer retrieval. To facilitate accurate answer formulation by the LM, textual information pertaining to paper references is incorporated into questions, deterring reliance on mere reference numbers. Similarly, references in answers are renumbered and supplemented with the relevant reference text as necessary.

Relevant for SciDQA
Q: How is the expectation of TCE algorithm computed in Equation 18? A: The expectation is calculated with respect to the ... Q: In section 3.4.1, is it possible to apply ReMERT to non-episodic or continuing task? A: ReMERT might not provide proper weights to . . . .
Irrelevant for SciDQA
Q: Can the inversion method by Chen et al. 2022 be used to improve the latency? A: We believe that this may be possible, however it will require further analysis. Q: Can you correct the typos in Section 3.4? A: Yes, we will correct them in the revised version.
Ambiguous
Q: Can this inversion method be used in tandem with online filtering/smoothing (e.g. 4DVar, EnKF)? A: We believe that this may be possible, potentially leveraging ideas from Chen et al. [2021].
Q: Why don't the authors compare to PINNs? A: PINNs are typically employed to retrieve individual solutions, not learn distributions over data sets. When using them to solve the individual problems, inference is much slower since the network needs to be trained for each inferred solution. Iterative solvers seem like a better alternative in our setting.

Table 7: Categorization of questions for inclusion in the SciDQA dataset. Information-seeking questions, whose answers are ascertainable within the research paper text, from a collection of synthetically extracted question-answer pairs using PaLM text-bison-001 model are categorized as relevant, and added to the SciDQA dataset.



(a) Initial version is preferred as the revised copy explicitly mentions the answer.



(b) Revised version is preferred as the answer is originally absent.

Figure 5: We present scenarios where the initial and the revised manuscript versions are most appropriate for answering the reviewer's question. For each question in the dataset, we annotate the preferred manuscript version.

Chunking for FT Setup In the CFT setting, the chunk length is determined by the LLM's context length. If the model context length is N, we reserved 500 tokens for the instruction and the question, and utilized the rest N - 500 tokens for context. The chunking strategy is presented in Algorithm 2.

B.2 Answer Selection Prompt for Gemini

The prompt provided to Gemini-pro model to generate a single answer during the answer selection phase in CFT setup is as follows:

Prompt for Answer Selection using Gemini

You are provided with a question and some potential answers about a research paper submitted to a top-tier computer science conference in the domain of ML and DL. Your task is to select the best answer from the provided answer options, which comprehensively answers the question. Do not include any additional text other than the answer and select only one answer from the provided options.

Algorithm 1 Chunk Creation Algorithm for RAG

Input: Full-text document
Output: List of chunks
Split the full-text into paragraphs (demarcated by \n).
for each paragraph P **do**
    Split P into individual sentences (using the NLTK library).
    Initialize an empty list $chunks$
    for every 10 consecutive sentences in P **do**
        Join the sentences to build a chunk.
        Add the chunk to $chunks$
        Slide the window by nine sentences (i.e., keep a single overlapping sentence between consecutive chunks).
    end for
end for

Algorithm 2 Chunk Creation Algorithm for CFT

Split the full-text into paragraphs (demarcated by \n).
for each paragraph **do**
    if the length of paragraph is less than $N - 500$ **then**
        The entire paragraph is treated as a chunk
    else
        Split the paragraph into a list of sentences, say $S = [s_1, s_2, \ldots, s_n]$
        Initialize an empty $chunks\_list = [\,]$
        Initialize an empty string chunk $c =$ ""
        for sentence s in S **do**
            if $token\_count(c) + token\_count(s) < N - 500$ **then**
                Add sentence s to the chunk c
                continue
            else
                Add chunk c to the $chunks\_list$
                Reinitialize the empty chunk c
                Add sentence s to the chunk c
                continue
            end if
        end for
    end if
end for

	ODQA-PQ	ODQA-PTABS	RAG	CFT
	2-3 B			
Gemma (2024)	2.32	1.7	1.64	1.98
Gemma IT (2024)	3.35	1.34	2.2	1.88
Phi-2 (2023)	3.7	2.75	1.92	3.02
	6-7 B			
Falcon (2023)	1.23	1.3	1.37	2.62
Falcon IT (2023)	3.78	1.92	2.76	2.77
Galactica (2022)	1.43	1.52	1.37	2.8
Llama 2 (2023)	1.41	1.67	1.74	2.73
Llama 2 Chat (2023)	3.65	2.14	3.25	2.77
Mistral (2023)	2.37	2.23	2.16	2.6
Mistral IT (2023)	3.75	3.5	3.3	2.98
Vicuna (2024)	2.27	1.1	3.03	3.81
Vicuna 16k (2024)	3.18	1.19	2.83	3.36
Zephyr β (2023)	3.98	3.46	3.71	3.34
	13 B			
Llama 2 (2023)	1.4	1.5	1.74	2.82
Llama 2 Chat (2023)	3.91	2.23	3.55	3.88
Vicuna (2024)	3.44	1.19	3.17	3.78
	Proprietary LLMs			
GPT-4 (2023)	-	-	-	**4.05**
Gemini-Pro (2023)	-	-	-	2.89

Table 8: Average Prometheus Scores, where each answer is rated on a scale of 1-5 focusing on syntactic correctness, relevance to the question, factuality, and comprehensiveness. GPT-4 has the highest average score of 4.05.

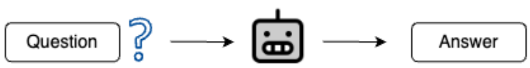


Figure 6: Priming LLMs with Questions (ODQA-PQ). This task evaluates the ability of LLM to recall the answer without any relevant context.

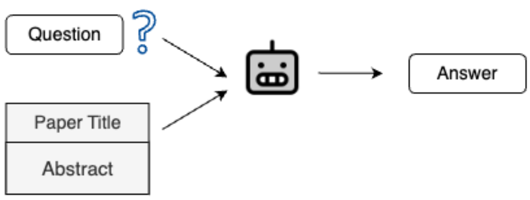


Figure 7: Open-Domain Question Answering - Priming with Question and Title/Abstract (ODQA-PTABS). This task evaluates the impact of additional context on LLM ability to recall the answer without reasoning about the question.

Preferred Answer	Tie	Human	GPT-4	None
GPT-4	32.5	30.4	37.0	34.6
Human	34.8	34.4	38.5	34.0

Table 9: Average scores of Human and GPT-4 generated answers on a subset of SciDQA dataset across instance categories. The average score (R-1, R-2, R-L, BLRT, BS-F) of human and GPT-4 generated answers are grouped by preference.

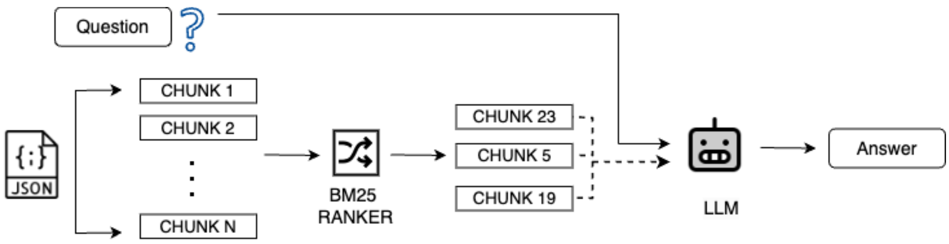


Figure 8: RAG setup ranks paper subsections based on their relevance to the question, and top-3 subsections are provided to the base-LLM, which generates the answer.

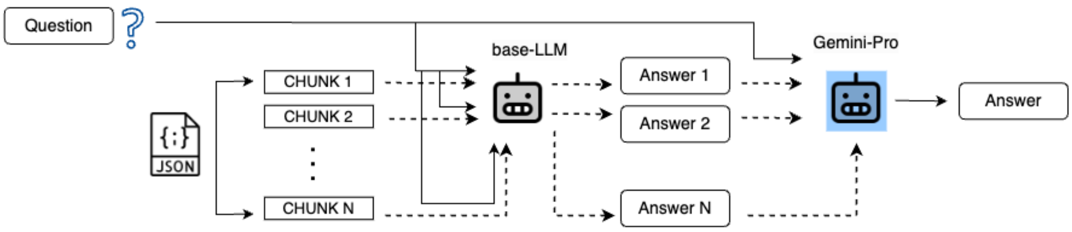


Figure 9: Comprehending the full-text (CFT) of the research paper by passing model context-length segments to the base-LLM and generating answers from each segment. Gemini-Pro selects the best answer among generated candidate answers.