cPAPERS: A Dataset of Situated and Multimodal Interactive Conversations in Scientific Papers

Anirudh Sundar* **Jin Xu*** **William Gay*** **Christopher Richardson*** **Larry Heck**
AI Virtual Assistant Lab
Georgia Institute of Technology
* Equal Contribution
{asundar34, jxu81, wgay7, crichardson8, larryheck}@gatech.edu

Abstract

An emerging area of research in situated and multimodal interactive conversations (SIMMC) includes interactions in scientific papers. Since scientific papers are primarily composed of text, equations, figures, and tables, SIMMC methods must be developed specifically for each component to support the depth of inquiry and interactions required by research scientists. This work introduces CONVERSATIONAL PAPERS (cPAPERS), a dataset of conversational question-answer pairs from reviews of academic papers grounded in these paper components and their associated references from scientific documents available on arXiv. We present a data collection strategy to collect these question-answer pairs from OpenReview and associate them with contextual information from LaTeX source files. Additionally, we present a series of baseline approaches utilizing Large Language Models (LLMs) in both zero-shot and fine-tuned configurations to address the cPAPERS dataset.

1 Introduction

Developing conversational assistants capable of situated and multimodal interactive conversations (SIMMC) over structured knowledge sources remains an open problem [33, 47, 27].

An emerging area of research within this domain is conversational interactions over scientific documents [41]. The number of scientific articles has grown dramatically over the past decade [1], making it difficult for scientists to find, read, understand, and connect advancements published by their fellow researchers. Scientists in the field of biomedicine, for example, publish over 1 million articles per year or a new article every 2 minutes on average [28]. Developing methods to understand scientific documents and assist researchers is an important problem, especially for the Natural Language Processing (NLP) community. Scientific documents present an interesting challenge in SIMMC-based methods since the content is frequently multimodal [6].

Besides textual paragraphs, researchers rely on various modalities to describe research methods. Figures convey information about concepts developed throughout the paper. Generally, figures represent many different types of information including most commonly graphical and image information. For example, figures explain model architectures and pipelines, summarize experimental results in graphical plots, and represent complex information such as gradient learning surfaces and feature maps. In addition, figures can be used to show various stages of image processing in computer vision and image generation papers.

Equations are crucial for grasping mathematical concepts in scientific texts but can be challenging to interpret. Equations frequently rely on notation introduced elsewhere in the document, often requiring readers to draw from the entire paper to understand the formulation. In addition, equations

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

are typically related to and in many cases help the reader understand other paper components such as figures and tables (and vice-versa).

Finally, tables typically have structures that convey semantic information. These include the text in the row and column headings, and text and often links in the table cells. These structures summarize multiple concepts (e.g., experimental results) in a human-readable form. Therefore, developing SIMMC methods for scientific papers must include a semantic understanding of these structures and table content.

To advance the development of conversational assistants capable of SIMMC in scientific papers, this paper introduces Conversational Papers (cPAPERS[1]), a dataset of conversations in English situated in equations (cPAPERS-EQNS), figures (cPAPERS-FIGS), and tabular information (cPAPERS-TBLS) for scientific texts. Question-answer pairs are sourced from reviews and rebuttals from OpenReview.[2] Textual grounding to answer the questions and answers are sourced from the corresponding scientific articles hosted on arXiv [3], an open-access repository of academic preprints.

The contributions of this work include:

- Introduction of the CONVERSATIONAL PAPERS (cPAPERS) dataset including three splits (EQNS, FIGS, TBLS)
- Development of a novel and scalable approach for collecting question-answer pairs from OpenReview and linking them with relevant contextual information from open access TEX sources on arXiv
- Creation of baseline LLM approaches that utilize weakly grounded multimodal context for dialogue responses

The rest of the paper is organized as follows: Section 2 discusses related work in language modeling and question-answering for equations, tables, and figures. Section 3 provides a description of cPAPERS and details the dataset collection process. Next, Section 4 describes baseline approaches to address this dataset with a series of experiments utilizing zero-shot prompting and parameter-efficient fine-tuning. Section 5 discusses the results, while Sections 6 and 7 address the conclusion and limitations, respectively.

2 Related Work

2.1 Language Modeling for Equations

The challenge of modeling mathematical equations has become an active research area in Natural Language Processing. [11] proposes an encoder-decoder approach to generate equations from math word problems. [53] presents an approach for the dual problem of generating math word problems consistent with equations. Prior work has also proposed learning representations from equations using the Transformer encoder [39], Tree-based encoders [54], and RNNs [56]. More recent work utilizing generative architectures includes MathGPT [44], an auto-regressive model based on GPT-2 [40] for various language+equation tasks, and Nougat [6], a Visual Transformer to parse academic documents to markup language. In contrast, cPAPERS is a dataset that addresses both grounded conversational question-answering and the modeling of mathematical equations. Answering the questions accurately requires understanding the expressions in the equations while also formulating an accurate response that addresses the task succinctly.

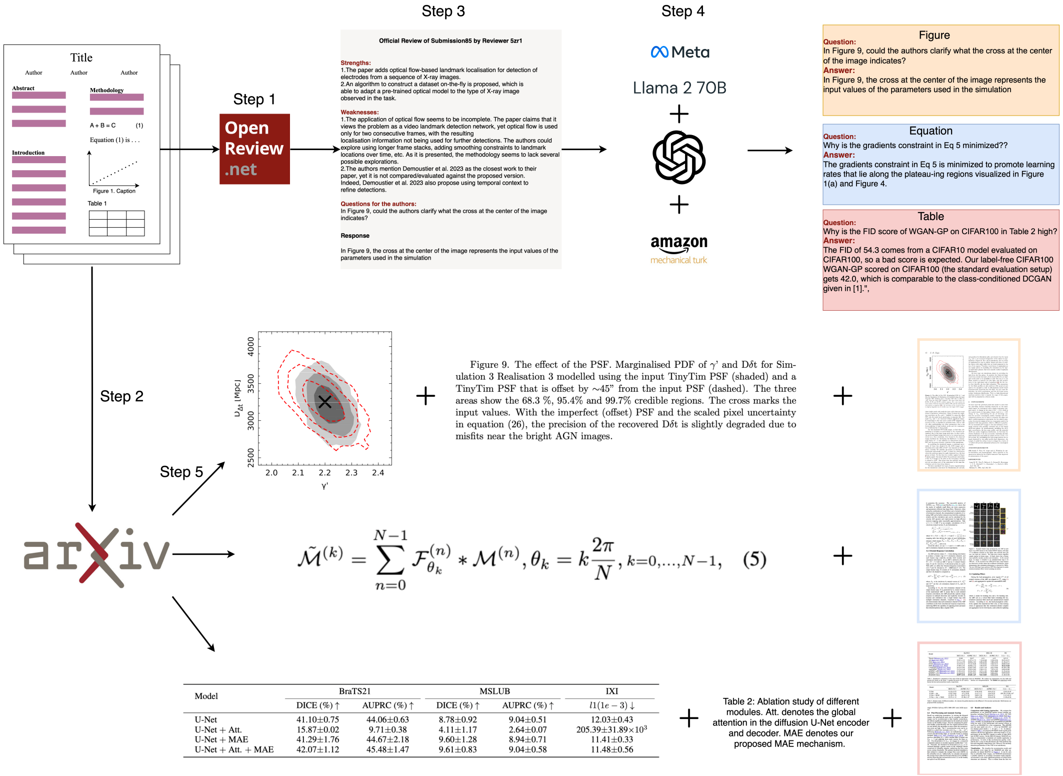
2.2 Question-Answering over Figures

Early datasets on question-answering over figures include VQA [2], Visual7W [63], VisDial [13], MANYMODALQA [21], and OK-VQA [32, 42]. More recent datasets addressing the problem of open-domain conversations include IGC [34], MOD [17], and Image-Chat [45]. However, the images

[1] The dataset is publicly available at `https://huggingface.co/datasets/avalab/cPAPERS`. The collection process, baseline models, and code are located at `https://github.com/avalab-gt/cPAPERS` under the GNU General Public License.

[2] `https://openreview.net/`

[3] `https://arxiv.org`



Official Review of Submission85 by Reviewer 5zr1

Strengths:
1.The paper adds optical flow-based landmark localisation for detection of electrodes from a sequence of X-ray images.
2.An algorithm to construct a dataset on-the-fly is proposed, which is able to adapt a pre-trained optical model to the type of X-ray image observed in the task.

Weaknesses:
1.The application of optical flow seems to be incomplete. The paper claims that it views the problem as a video landmark detection network, yet optical flow is used only for two consecutive frames, with the resulting localisation information not being used for further detections. The authors could explore using longer frame stacks, adding smoothing constraints to landmark locations over time, etc. As it is presented, the methodology seems to lack several possible explorations.
2.The authors mention Demoustier et al. 2023 as the closest work to their paper, yet it is not compared/evaluated against the proposed version. Indeed, Demoustier et al. 2023 also propose using temporal context to refine detections.

Questions for the authors:
In Figure 9, could the authors clarify what the cross at the center of the image indicates?

Response

In Figure 9, the cross at the center of the image represents the input values of the parameters used in the simulation

Figure
Question:
In Figure 9, could the authors clarify what the cross at the center of the image indicates?
Answer:
In Figure 9, the cross at the center of the image represents the input values of the parameters used in the simulation

Equation
Question:
Why is the gradients constraint in Eq 5 minimized??
Answer:
The gradients constraint in Eq 5 is minimized to promote learning rates that lie along the plateau-ing regions visualized in Figure 1(a) and Figure 4.

Table
Question:
Why is the FID score of WGAN-GP on CIFAR100 in Table 2 high?
Answer:
The FID of 54.3 comes from a CIFAR10 model evaluated on CIFAR100, so a bad score is expected. Our label-free CIFAR100 WGAN-GP scored on CIFAR100 (the standard evaluation setup) gets 42.0, which is comparable to the class-conditioned DCGAN given in [1].",

Figure 9. The effect of the PSF. Marginalised PDF of γ' and Dδt for Simulation 3 Realisation 3 modelled using the input TinyTim PSF (shaded) and a TinyTim PSF that is offset by ∼45" from the input PSF (dashed). The three areas show the 68.3 %, 95.4% and 99.7% credible regions. The cross marks the input values. With the imperfect (offset) PSF and the scaled pixel uncertainty in equation (26), the precision of the recovered Dδt is slightly degraded due to misfits near the bright AGN images.

$$\tilde{\mathcal{M}}^{(k)} = \sum_{n=0}^{N-1} \mathcal{F}_{\theta_k}^{(n)} * \mathcal{M}^{(n)}, \theta_k = k\frac{2\pi}{N}, {\scriptstyle k=0,\ldots,N-1}, \quad (5)$$

Model	BraTS21		MSLUB		IXI
	DICE (%) ↑	AUPRC (%) ↑	DICE (%) ↑	AUPRC (%) ↑	$l1(1e-3)$ ↓
U-Net	41.10±0.75	44.06±0.63	8.78±0.92	9.04±0.51	12.03±0.43
U-Net + Att.	15.87±0.02	9.71±0.38	4.11±1.17	2.64±0.07	205.39±31.89×10^3
U-Net + MAE	41.29±1.76	44.67±2.18	9.60±1.28	8.94±0.71	11.41±0.33
U-Net + Att. + MAE	42.07±1.12	45.48±1.47	9.61±0.83	9.04±0.58	11.48±0.56

Table 2: Ablation study of different modules. Att. denotes the global attention in the diffusion U-Net encoder and decoder. MAE denotes our proposed MAE mechanism.

Figure 1: Overview of the data collection approach. First, the reviews are obtained from OpenReview. The corresponding TEX source is obtained from arXiv (only if the paper is also posted to arXiv). The reviews and rebuttals are filtered using a combination of LLMs and Amazon Mechanical Turk to obtain question-answer pairs in the json format. The TEX source is parsed to obtain the associated equation, table, or figure to create a dataset of questions and answers paired with the corresponding LATEX.

in these datasets are collected from MS-COCO [30] or YFCC100M [51], where visual content targets commonly seen everyday objects as opposed to scientific documents where images target specific information relevant to the explanation of a concept.

Recent work has addressed some of the challenges associated with multimodal image+text tasks situated in scientific documents. [50] presents a dataset of charts from scientific papers and associated natural language captions summarizing the information present in the chart. [19] develops a method to link labels with relevant images in patents. [12] introduces a dataset of 150 computer science papers, ground truth labels for the locations of figures, tables, and captions, and an approach to automatically extract this information from PDFs. SCICAP [25] is a much larger dataset of 400,000 figures, their captions, and associated textual references from various scientific papers. The dataset is collected by scraping scientific preprints from arXiv. For each figure, they provide the associated caption and all paragraphs that mention the figure.

While these datasets address tasks related to multimodal content in scientific documents, they are not conversational in nature. In contrast, cPAPERS addresses the shortcomings of prior datasets as a multimodal, conversational dataset grounded in scientific documents. The questions and answers address specific contextual visual information grounded in scientific documents.

2.3 Tabular Question-Answering

Tabular question-answering addresses the problem of extractive QA grounded in the information contained in specific cells of a table. Prior tabular QA datasets include those collected from Wikipedia, such as WIKITABLEQUESTIONS [38], ManyModalQA [22], TABERT [57], NQ-Tables [23], FEVEROUS [4], FeTaQA [36], HYBRIDIALOGUE [35], and HiTab [10]. Other tabular datasets

	Equation			Table			Figure		
	train	dev	test	train	dev	test	train	dev	test
# Unique Papers	672	286	335	715	285	302	761	275	313
# QA Pairs	993	336	394	932	327	342	1052	297	357
# Tokens (average)									
	train	dev	test	train	dev	test	train	dev	test
Question	25	25	25	24	22	26	23	23	24
Answer	92	102	90	86	79	81	83	88	81
Contexts	10,232	11,851	12,288	2,981	2,746	2,610	433	400	431
References	7,323	8,413	9,517	1,757	1,645	1,427	366	375	323
Neighboring Contexts	1,144	1,153	1,084	994	1,043	925	-	-	-
Neighboring References	1,000	947	1,152	736	657	588	-	-	-

Table 1: Dataset Statistics

are constructed from financial reports - TAT-QA [62], FinQA [9], MultiHiertt [61], or arXiv - iTBLS [48].

Proposed approaches to address the tabular QA task include architectures based off of the Transformer encoder [57, 24, 8, 16, 31, 20, 55], decoder [18, 3, 58, 26, 59, 46], or both (encoder-decoder) [35, 14, 49, 43, 48].

In contrast to prior tabular datasets, cPAPERS-TBLS presents a new source of grounded questions and answers and to the best of our knowledge, is the first to be situated in tables from OpenReview. Additionally, the question-answer pairs are not factoids. Rather, they are conversational in nature.

3 The cPAPERS Dataset

3.1 Dataset Description

Conversational Papers (cPAPERS) is a dataset of conversations in English situated in scientific texts. cPAPERS consists of question-answer pairs pertaining to figures (cPAPERS-FIGS), equations (cPAPERS-EQNS), or tabular information (cPAPERS-TBLS) from scientific papers. These pairs are sourced from the official reviews and rebuttals of The Conference and Workshop on Neural Information Processing Systems (NeurIPS) and The International Conference on Learning Representations (ICLR) between 2020 and 2023, as available on OpenReview. cPAPERS comprises of 1723 question-answer pairs over equations, 1601 pairs over tables, and 1706 pairs over figures, totaling 5030 question-answer pairs spanning 2350 unique scientific papers. A detailed breakdown of the dataset is provided in Table 1.

Section 3.2 details the steps of collecting cPAPERS from official reviews on OpenReview, extracting question-answer pairs, and associating them with LaTeX source files from arXiv.

3.2 Dataset Collection

3.2.1 Step 1 - Extract Official Review and Comments

We leverage the OpenReview API to access official reviews and comments for each paper submission. These reviews, authored by conference reviewers, typically include a paper summary, strengths and weaknesses, specific questions, and limitations. Authors respond to these critiques in the comments, providing answers that address the reviewers' questions. This step involves downloading the official reviews and comments for all submissions.

3.2.2 Step 2 - Download LaTeX files

For each paper recorded in Step 1, we initiate a call to the arXiv API to verify the existence of the paper. If the paper is found, we proceed to download all associated LaTeX source files. If a paper cannot be located using the arXiv API, we exclude this submission from further processing.

3.2.3 Step 3 - Regex Filtering

As mentioned in Step 1, review-rebuttal pairs contain a summary of the paper and additional information that may not be related to equations, tables, or figures. Therefore, we use regular expressions to identify mentions of equations, tables, and figures within each review and comment pair. The following regular expressions are employed for this purpose:

```
Equation: re.compile(r’\b(?:Equation|Eq\.?)\s*\(?\s*\d+\)?\b’,re.IGNORECASE)
```

```
Table: re.compile(r’\bTable\s*\d+|\btables\b’,re.IGNORECASE)
```

```
Figure: re.compile(r’\b(?:Figure|fig)\.?\s*\d+\b’, re.IGNORECASE)
```

These expressions efficiently determine whether equations, tables, and figures are mentioned in either the review or the rebuttal, resulting in a separate list of review-rebuttal pairs each containing explicit references to an equation, table, or figure in the paper. It is possible for a review to contain references to multiple modalities.

3.2.4 Step 4 - Question-Answer Extraction

To extract question-answer pairs from the regex-filtered reviews and rebuttals, an LLM that is instruction fine-tuned to align with human preferences is prompted *in-context* with a single example. For this process, we use LLAMA-2-70B-CHAT-HF [52].

The prompt is as follows:

‘*Given this example Content, Question, and Answer, look through the Summary and Comments that are dedicated to the paper and extract Question and Answer pairs that specifically target a particular Figure/Equation/Table in the content. Questions may be under sections labeled similarly to “Weaknesses” or “Questions” or “Review” or “Clarity, Quality, Novelty And Reproducibility” in the comments, and Answers may be under sections labeled “Response to Reviewer” or “Rebuttal" or “Comments”.*’

System Prompt: *“You are a helpful question answer extracting assistant. Our goal is to extract Question and Answer pairs pertaining to a specific a Figure/Equation/Table in the content. Your response should be in the format of [Question] <question> [Answer] <answer> Do not add any other unnecessary content in your response.”*

Post-processing + JSON reformatting. GPT 3.5 [7] is used to reformat the question-answer pairs as JSON. The following system prompt is used.

Prompt: *You are a helpful assistant. Please find the Question-Answer pairs, and format it as a json ({“question_answers”: [{ “question”: “”, “answer”: “” },...]})*

For each question-answer pair, an additional regular expression is employed to extract the respective figure, equation, or table number following the keywords ‘Figure’, ‘Fig’, ‘Equation’, ‘Eq’, or ‘Table’.

```
pattern = r’\d+’
numbers = re.findall(pattern, input_string)
```

3.2.5 Step 5 - Crowdworker Cleanup

A significant portion of academic reviews and rebuttals on OpenReview pertain to clarification questions as well as those related to fixing typos. While the LLM processing removes most of the questions and answers addresing typos, to further ensure the quality of the dataset, we employ crowd workers. The crowd workers (recruited using Amazon Mechanical Turk) ascertain whether a question-answer pair about an equation, table, or figure is technical in nature or not. For each question-answer pair, two crowd workers provide feedback and only those question-answer pairs with a consensus are retained. Additional details of the crowdworker setup is provided in Appendix A.5.

3.2.6 Step 6 - Contextualizing QA Pairs

In addition to extracting question-answer pairs from OpenReview, cPAPERS associates these question-answer pairs with referring text from the LaTeX source on arXiv. Often, the text surrounding figures, equations, and tables in scientific documents provides additional relevant information. Regular expressions are employed to locate all instances of figures, equation, and tables in the LaTeX source using the \begin{figure}, \begin{equation}, and \begin{table} environments.

Equation: The equation is obtained by extracting content enclosed between \begin{equation} and \end{equation} tags. We provide all equations from the LaTeX source for a question-answer pair.

Table: The table is extracted from the content enclosed between \begin{tabular} and \end{tabular} tags. We provide all tables from the LaTeX source for a question-answer pair.

Figure: First, the content between \begin{figure} and \end{figure} tags is extracted. From this content, we obtain the caption (using \caption) and the figure path (using \includegraphics).

Additionally, for each modality, we provide context and references:

Context: The context represents the paragraph of text preceding and the paragraph of text following the equation, table, or figure referred to in the question-answer pair.

References: In addition, we report all referring text in a document associated with the equation, table, or figure in the question-answer pair. Within each environment, we search for a label (\label) and if it exists, we locate all references by searching for text using the label in LaTeX's \ref command.

Reviews of academic pre-prints provide a source for high-quality document-grounded questions and answers and success on the task requires contextual and multimodal understanding. However, one of the main challenges with matching questions and answers from OpenReview with their corresponding multimodal context from arXiv is that while reviewer and author references to equations, tables, and figures are purely ordinal (e.g. 'Equation 3', 'Figure 4', 'Table 2'), the LaTeX source does not provide any reliable approach to associate this ordinal position with the multimodal context in the final compiled document, specifically for equations and tables. Additionally, arXiv contains multiple versions of papers with no guarantees on consistency with the submission to OpenReview. As a result, the cPAPERS dataset reports both the position of the equation, table, or figure referred to in the question, and all equations and tables from that specific paper. Section 4 contains additional details on how this information is used in a baseline approach.

3.2.7 Step 7 - Additional Post-processing of Figures

Authors frequently utilize different graphical formats in their papers. To ensure consistency in our dataset, we convert all .pdf and .eps files to .png format using ImageMagick. [4]

4 Baseline Approaches

We experiment with zero-shot prompting and parameter-efficient fine-tuning for the baseline approaches. Zero-shot prompting involves querying pre-trained LLMs to answer the questions in the cPAPERS dataset without additional fine-tuning. In this approach, we use *off-the-shelf* Transformer decoder LLMs already pre-trained using a causal language modeling objective followed by supervised fine-tuning and reinforcement learning with human feedback to follow instructions. The LLM is not fine-tuned on cPAPERS and we do not provide examples in-context.

In addition, we experiment with parameter-efficient fine-tuning using QLoRA [15] to better align the model with the style of responses in the cPAPERS dataset. An LLM is fine-tuned independently for each modality (equation, table, or figure).

We serialize equations and tables using the corresponding LaTeX representations from the TeX source on arXiv. As previously indicated in Section 3.2.6, pairing questions and answers from OpenReview with LaTeX sources from arXiv presents a challenge with possible inconsistencies between the references in questions and answers to the corresponding sources in LaTeX. Therefore, we conduct question-answering experiments under two settings - utilizing all multimodal content, or utilizing a smaller subset of neighboring, weakly-grounded multimodal content. While using all multimodal content (equations or tables) ensures that the correct equation (or table) referred to by the question-answer pair is provided to the model, limitations on context length constrain the model from accurately

[4] https://imagemagick.org/

Modality	Setting	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT	BLEU
Question (Q)	-	**0.194**	**0.065**	**0.144**	0.240	**0.825**	**0.038**
Q+Equation	Neighboring	0.190	0.063	0.139	**0.245**	0.821	0.035
	All	0.170	0.056	0.123	0.218	0.740	0.035
Q+Context	Neighboring	0.186	0.063	0.137	0.237	0.809	0.036
	All	0.079	0.027	0.058	0.102	0.345	0.036
Q+References	Neighboring	0.176	0.061	0.129	0.223	0.764	0.036
	All	0.112	0.037	0.082	0.143	0.498	0.037

Table 2: Zero-shot language modeling on cPAPERS-EQNS test set with LLAMA-2-70B

using this information. Evidenced by prior work demonstrating the retrieval capabilities of generative LLMs when provided weakly-grounded context [49], we reconcile this situation by providing the model with a subset of equations or tables from the entire paper. First, we assign each equation or table with an ordinal position based on each instance of a `\begin{equation}` or `\begin{table*}` in the LaTeX source. Then, supposing the question-answer pair refers to equation$_i$, the LLM is provided a subset of equations: $\{\text{equation}_{i-1}, \text{equation}_i, \text{equation}_{i+1}\}$. We utilize one-sided grounding in the situation where grounding from either boundary does not exist.

5 Results

The following sections contain results of the baseline approaches on the development and test splits. We report results from automatic evaluation - ROUGE [29] [5], METEOR [5] [6], BERTScore [7] [60], and BLEU [8] [37]. ROUGE, METEOR, and BLEU respectively measure the recall, F1 score, and precision of n-grams between model generated and reference text. BERTScore leverages contextual embeddings to measure the semantic similarity of model generated responses to the ground truth.

Zero-shot. First, we experiment with zero-shot language modeling with LLAMA-2-70B [52] to answer questions on cPAPERS-EQNS, cPAPERS-TBLS, and cPAPERS-FIGS. The results from Tables 2 and 3 indicate that utilizing the neighboring equations/tables provides a significant improvement over using all equations/tables to answer questions in the dataset, evidenced by up to 2x improvement in ROUGE, METEOR, and BERTScores. Additionally, the results from Table 2 indicate that utilizing the question alone without neighboring textual context (paragraphs of text surrounding the equation) results in best ROUGE and METEOR scores on cPAPERS-EQNS, whereas utilizing referring text yields the best scores on these metrics in cPAPERS-TBLS when compared to using the question alone to generate answers (Table 3). We posit this is because referring text in the document is predominantly used to summarize tabular results.

Modality	Setting	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT	BLEU
Question (Q)	-	0.192	0.058	0.136	0.232	**0.832**	0.028
Q + Table	Neighboring	0.206	0.061	**0.145**	0.237	0.828	**0.031**
	All	0.176	0.052	0.123	0.199	0.697	0.030
Q + Context	Neighboring	0.202	0.062	0.142	0.241	0.828	**0.031**
	All	0.160	0.050	0.114	0.195	0.666	0.030
Q + References	Neighboring	**0.207**	**0.064**	0.144	**0.243**	0.829	**0.031**
	All	0.186	0.057	0.130	0.223	0.758	0.030

Table 3: Zero-shot language modeling on cPAPERS-TBLS test set with LLAMA-2-70B

Table 4 contains results for zero-shot language modeling on cPAPERS-FIGS. In this situation, utilizing context performs the best across metrics.

[5] https://huggingface.co/spaces/evaluate-metric/rouge

[6] https://huggingface.co/spaces/evaluate-metric/meteor

[7] https://huggingface.co/spaces/evaluate-metric/bertscore

[8] https://huggingface.co/spaces/evaluate-metric/bleu

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT	BLEU
Question (Q)	0.185	0.065	0.137	0.238	0.833	0.036
Q + Caption	0.200	0.074	0.149	0.248	0.837	0.039
Q + Context	**0.208**	**0.076**	**0.155**	**0.254**	**0.837**	**0.039**
Q + References	0.205	0.075	0.154	0.251	0.837	0.038

Table 4: Zero-shot language modeling on cPAPERS-FIGS test set with LLAMA-2-70B

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT	BLEU
Question (Q)	0.309	**0.138**	0.248	0.215	0.860	0.062
Q + Equation	**0.317**	0.134	**0.251**	**0.223**	**0.861**	0.070
Q + Context	0.297	0.126	0.235	0.221	0.817	0.077
Q + References	0.283	0.122	0.224	0.217	0.777	**0.081**

Table 5: Results of fine-tuning LLAMA-2-7B on cPAPERS-EQNS test set

Fine-tuning. We also report results by fine-tuning LLAMA-2-7B using QLoRA [15]. Motivated by the zero-shot experiments, we only conduct fine-tuning using the *neighboring* setting for cPAPERS-EQNS and -TBLS.

Results of fine-tuning on cPAPERS-EQNS are detailed in Table 5. Utilizing the question with the equation outperforms the question alone or using the question along with referring text or the context on ROUGE-1 and ROUGE-L scores, METEOR, and BERTScore.

In contrast, the results from finetuning cPAPERS-TBLS and cPAPERS-FIGS in Tables 6 and 7 indicate that using just the question results in the best performance across most metrics. We believe that this is a result of the nature of questions and answers in reviews where authors often utilize tables and figures to supplement reviewer questions. This is supported by Table 8. In the cPAPERS-EQNS split of the dataset, 55.9% of questions refer to an equation while 77.6% of answers refer to an equation. On cPAPERS-TBLS and cPAPERS-FIGS, this difference is exacerbated, with answers referring to a table or a figure more often and questions referring to them less often when compared to equations.

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT	BLEU
Question (Q)	**0.315**	**0.121**	**0.235**	0.218	**0.869**	0.059
Q + Table	0.293	0.107	0.218	0.212	0.820	0.065
Q + Context	0.294	0.106	0.216	**0.225**	0.838	**0.070**
Q + References	0.292	0.106	0.214	0.218	0.816	0.068

Table 6: Results of fine-tuning LLAMA-2-7B on cPAPERS-TBLS test set

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT	BLEU
Question (Q)	**0.329**	**0.155**	**0.269**	**0.250**	0.859	0.083
Q + Caption	0.322	0.147	0.260	0.246	**0.868**	0.083
Q + Context	0.321	0.140	0.256	**0.250**	0.858	**0.091**
Q + References	0.311	0.131	0.244	0.247	0.843	**0.091**

Table 7: Results of fine-tuning LLAMA-2-7B on cPAPERS-FIGS test set

6 Conclusion

This paper introduces the cPAPERS dataset, which addresses the shortcomings of prior datasets as a multimodal, conversational dataset grounded in scientific documents. Leveraging reviews of academic papers grounded in equations, figures, and tables, as well as their associated references from scientific documents available on arXiv, this dataset aims to further advance the development of conversational

Split	Question (%)	Answer (%)
cPAPERS-EQNS	55.94	77.65
cPAPERS-TBLS	46.47	79.51
cPAPERS-FIGS	54.16	80.48

Table 8: Comparison between the percentage of questions that contain a reference to an Equation, Table, or Figure versus the percentage of answers

assistants capable of situated and multimodal interactive conversation within scientific papers. We conduct a series of experiments with zero-shot prompting to benchmark state-of-the-art models performance and perform several parameter-efficient fine-tuning for the baseline approaches. The experimental results provide valuable insights and create opportunities for enhancing and innovating the development of future conversational AI assistant.

7 Discussion

This paper introduces an innovative and scalable approach for collecting a large dataset of situated and multimodal interactive conversations related to scientific papers. This approach has the potential to facilitate the collection of extensive question-answer pairs across a wide range of scientific fields with ease. Collecting such a large and diverse dataset would benefit the research community by providing opportunities to develop more advanced AI-based research assistants for scientific research.

There may be concerns that creating an AI assistant to help automate the scientific discovery process could diminish the role of human scientists. While this may be a possible use of this dataset, our intention is to develop an AI research assistant that amplifies human scientists, empowering them to achieve greater scientific discoveries more efficiently.

Limitations. The key limitation of this dataset is the presence of mismatched figures, tables, or equations across different versions of the manuscripts. The *.tex* files of the paper on arXiv often undergo multiple revisions, and comments on Open Review are typically specific to a particular version. Authors frequently make changes in response to reviewer comments, which may involve additions, removals, or reordering of figures, equations, or tables. This presents a potential mismatch between the figures, equations, or tables and the question-and-answer pairs, thereby introducing additional challenges for language modeling.

Acknowledgments and Disclosure of Funding

This work was supported by NSF IIS-2112633 and by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

[1] Monthly Submissions — arxiv.org. `https://arxiv.org/stats/monthly_submissions`. [Accessed 13-10-2024].

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual Question Answering. *arXiv:1505.00468 [cs]*, October 2016. arXiv: 1505.00468.

[3] Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore, December 2023. Association for Computational Linguistics.

[4] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In Rami Aly, Christos

Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[6] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural Optical Understanding for Academic Documents, August 2023. arXiv:2308.13418 [cs].

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[8] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, 2020.

[9] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[10] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. HiTab: A hierarchical table dataset for question answering and natural language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] Ting-Rui Chiang and Yun-Nung Chen. Semantically-aligned equation generation for solving and reasoning math word problems. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 2656–2668, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Christopher Clark and Santosh Divvala. Extracting Figures, Tables, and Captions from Computer Science Papers.

[13] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. *arXiv:1611.08669 [cs]*, August 2017. arXiv: 1611.08669.

[14] Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. PACIFIC: Towards proactive conversational question answering over tabular and textual data in finance. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6970–6984, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc., 2023.

[16] Julian Eisenschlos, Syrine Krichene, and Thomas Müller. Understanding tables with intermediate pre-training. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online, November 2020. Association for Computational Linguistics.

[17] Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Towards Expressive Communication with Internet Memes: A New Multimodal Conversation Dataset and Benchmark. *arXiv:2109.01839 [cs]*, September 2021. arXiv: 2109.01839.

[18] Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[19] Ming Gong, Xin Wei, Diane Oyen, Jian Wu, Martin Gryder, and Liping Yang. Recognizing Figure Labels in Patents. In *SDU@ AAAI*, 2021.

[20] Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4971–4983, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] Darryl Hannan, Akshay Jain, and Mohit Bansal. ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. *arXiv:2001.08034 [cs]*, January 2020. arXiv: 2001.08034.

[22] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886, 2020.

[23] Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. Open domain question answering over tables via dense retrieval. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online, June 2021. Association for Computational Linguistics.

[24] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online, July 2020. Association for Computational Linguistics.

[25] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[26] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. StructGPT: A general framework for large language model to reason over structured data. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore, December 2023. Association for Computational Linguistics.

[27] Satwik Kottur and Seungwhan Moon. Overview of situated and interactive multimodal conversations (SIMMC) 2.1 track at DSTC 11. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 235–241, Prague, Czech Republic, September 2023. Association for Computational Linguistics.

[28] Esther Landhuis. Scientific literature: Information overload. *Nature*, 535(7612):457–458, 2016.

[29] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[31] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*, 2021.

[32] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Seungwhan Moon, Satwik Kottur, Paul A. Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations. *CoRR*, abs/2006.01460, 2020.

[34] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. *arXiv:1701.08251 [cs]*, April 2017. arXiv: 1701.08251.

[35] Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[36] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318, 2002.

[38] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics.

[39] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv preprint arXiv:2105.00377*, 2021.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. A comparative study on the impact of model compression techniques on fairness in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada, July 2023. Association for Computational Linguistics.

[42] Benjamin Z Reichman, Anirudh Sundar, Christopher Richardson, Tamara Zubatiy, Prithwijit Chowdhury, Aaryan Shah, Jack Truxal, Micah Grimes, Dristi Shah, Woo Ju Chee, and others. Outside knowledge visual question answering version 2.0. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[43] Soumajyoti Sarkar and Leonard Lausen. Testing the limits of unified sequence to sequence llm pretraining on diverse table data tasks, 2023.

[44] Alexander Scarlatos and Andrew Lan. Tree-based representation and generation of natural and mathematical language. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3714–3730, Toronto, Canada, July 2023. Association for Computational Linguistics.

[45] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image Chat: Engaging Grounded Conversations. *arXiv:1811.00945 [cs]*, April 2020. arXiv: 1811.00945.

[46] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA, 2024. Association for Computing Machinery.

[47] Anirudh Sundar and Larry Heck. Multimodal conversational AI: A survey of datasets and approaches. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[48] Anirudh Sundar, Christopher Richardson, William Gay, and Larry Heck. itbls: A dataset of interactive conversations over tabular information. *arXiv preprint arXiv:2404.12580*, 2024.

[49] Anirudh S. Sundar and Larry Heck. cTBLS: Augmenting large language models with conversational tables. In Yun-Nung Chen and Abhinav Rastogi, editors, *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 59–70, Toronto, Canada, July 2023. Association for Computational Linguistics.

[50] Hao Tan, Chen-Tse Tsai, Yujie He, and Mohit Bansal. Scientific Chart Summarization: Datasets and Improved Text Modeling. In *SDU@AAAI*, 2022.

[51] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2):64–73, January 2016. arXiv: 1503.01817.

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[53] Zichao Wang, Andrew Lan, and Richard Baraniuk. Math word problem generation with mathematical consistency and problem context constraints. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5986–5999, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[54] Zichao Wang, Mengxue Zhang, Richard G Baraniuk, and Andrew S Lan. Scientific formula retrieval via tree embeddings. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1493–1503. IEEE, 2021.

[55] Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. TableFormer: Robust transformer modeling for table-text encoding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 528–537, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[56] Michihiro Yasunaga and John D Lafferty. Topiceq: A joint topic and mathematical equation model for scientific texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7394–7401, 2019.

[57] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online, July 2020. Association for Computational Linguistics.

[58] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, et al. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*, 2023.

[59] Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*, 2023.

[60] Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020.

[61] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[62] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, August 2021. Association for Computational Linguistics.

[63] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. *arXiv:1511.03416 [cs]*, April 2016. arXiv: 1511.03416.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See A.2
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported mean with standard errors
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [Yes] See A.3
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See A.2
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] As the data used in this study is publicly available online, explicit consent from individuals whose data is curated was not obtained. The data was accessed from publicly accessible sources, and no private or sensitive information was collected or utilized in this study.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The dataset used in this study consists of question-answer pairs regarding scientific papers and does not explicitly contain any personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See A.5
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See A.5

A Dataset Documentation

A.1 Dataset Description

cPAPERS is a dataset of conversational question-answer pairs from reviews of academic papers grounded in these paper components and their associated references from scientific documents available on arXiv. For a detailed description and intended uses, please refer to 1.

A.2 Dataset Accessibility

- The dataset is publicly available at `https://huggingface.co/datasets/avalab/cPAPERS`
- URL to Croissant metadata record for viewing and downloading by the reviewers: `https://huggingface.co/datasets/avalab/cPAPERS/blob/main/croissant.json`
- Code repository for collecting the dataset: `https://github.com/avalab-gt/cPAPERS`
- Code repository for reproducing the benchmark results: `https://github.com/avalab-gt/cPAPERS`

A.3 Hosting, licensing, and maintenance plan

The dataset is licensed under Creative Commons (CC), and the code is licensed under the GNU General Public License (GPL). We plan to host and maintain this dataset on HuggingFace.

A.4 Dataset Examples

Example question-answer pairs are provided in Tables 9 10 11, .

	Example
Question	"What does the symbol ˜ mean in Equation 1?"
Answer	"The symbol ˜ in Equation 1 represents “follows this distribution”. It means that the probability distribution of the context C is defined as the distribution of the random variable C."
Question	"Can you provide more information about what is meant by ’generative process in biological multi-agent trajectories’ in L27 and L83?"
Answer	"The generative process refers to Eq. (2), which is a conceptual equation representing the generative process in animal behaviors."
Question	"How does the DeepMoD method differ from what is written in/after Eq 3?"
Answer	"We add noise only to u, with the percentage being of the standard deviation of the dataset. Adding noise to x and t has not been studied to our knowledge and falls out of the scope of this paper."
Question	"How to do the adaptive attack based on Eq.(16)? Maximizing the loss in Eq.(16)?"
Answer	"By Maximizing the loss in Eq (16) using an iterative method such as PGD on the end-to-end model we attempt to maximize the loss to cause misclassification while minimizing the regret to avoid detection."
Question	"How does the proposed method handle the imputed reward?"
Answer	"The proposed method uses the imputed reward in the second part of Equation 1, which corresponds to the empirical risk of the combined dataset."

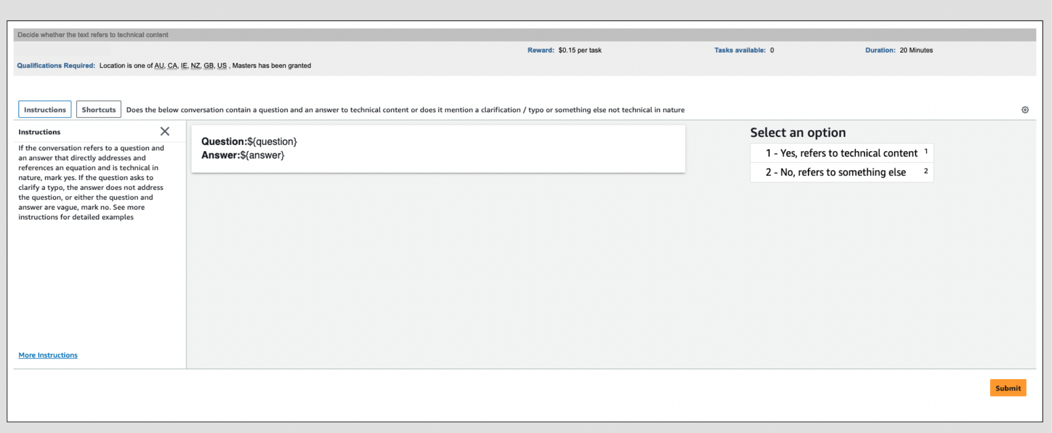
Table 9: Example QA Pairs in the cPAPERS-EQNS dataset

A.5 Crowdworker Instructions

A significant portion of academic reviews and rebuttals pertain to clarification questions and fixing typos. While the LLM processing removes most of these spurious questions and answers, to further ensure the quality of the dataset we employ crowdworkers from Amazon Mechanical Turk to ascertain whether a question-answer pair about an equation, table, or figure is technical in nature or asks to fix

	Example
Question	"What is the purpose of Table 2 in the paper?"
Answer	"Table 2 is used to provide a comparison of the computational complexity of the proposed approach with state-of-the-art methods."
Question	"Optimal number of clusters affected by the number of classes or similarity between classes?"
Answer	"The authors have addressed this concern by including a new experiment in Table 4 of the revised paper's appendix. The result shows that the optimal number of clusters is less affected by the number of classes but more affected by the similarity between classes."
Question	"Can you clarify the values represented in Table 1?"
Answer	"The values in Table 1 represent the number of evasions, which shows the attack strength. Therefore, the higher the value, the stronger the attack."
Question	"The experiments in table 1 do not seem to favor the proposed method much; softmax is better or similar. Can the authors explain why this might be the case?"
Answer	"The proposed method reduces to empirical risk minimization with a proper loss, and the seemingly trivial solution ϕ=Θ is often not optimal. This could explain why the proposed method might not perform as well as other methods in certain experiments. However, the authors hope that addressing concerns about the method's theoretical properties would be beneficial."
Question	"Does the first row of Table 2 correspond to the offline method?"
Answer	"Yes, the first row of Table 2 corresponds to the offline method."

Table 10: Example QA Pairs in the cPAPERS-TBLS dataset



Decide whether the text refers to technical content

Reward: $0.15 per task Tasks available: 0 Duration: 20 Minutes

Qualifications Required: Location is one of AU, CA, IE, NZ, GB, US , Masters has been granted

Does the below conversation contain a question and an answer to technical content or does it mention a clarification / typo or something else not technical in nature

Instructions

If the conversation refers to a question and an answer that directly addresses and references an equation and is technical in nature, mark yes. If the question asks to clarify a typo, the answer does not address the question, or either the question and answer are vague, mark no. See more instructions for detailed examples

Question:${question}
Answer:${answer}

Select an option

1 - Yes, refers to technical content 1
2 - No, refers to something else 2

Figure 2: Screenshot of crowdworker interface on Amazon Mechanical Turk

a typo. We recruit crowdworkers with the masters qualification from predominantly English-speaking countries: Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States of America. Crowdworkers are compensated USD 0.15 per Human Intelligence Task, with each task taking an average of one minute to complete. The crowdworker instructions and interface are provided in Figure 2. The total cost of cleaning up the dataset including fees paid to Amazon Mechanical Turk is USD 3215. Since the dataset is collected from OpenReview, we did not identify any risks to crowdworkers.

	Example
Question	"Why is there a gap between the proposed approach and the median approach in Fig. 5?"
Answer	"The gap is due to lower accuracy of the approximate median calculated by the bucketing scheme compared to a median method, albeit with faster epoch completion times."
Question	"What do experiments ensure us that the dependency is linear?"
Answer	"The linear dependency can be observed empirically in Figure 1. The paper will provide further experimental results in the updated version."
Question	"How do the different methods in Figure 5 have similar test errors, but the generalization bounds are so different?"
Answer	"The different methods in Figure 5 have similar test errors because they are all trained on the same dataset and have similar performance. However, the generalization bounds are different because they are computed using different methods. The proposed method uses a tighter bound that takes into account the structure of the ensemble, while the baselines use a looser bound that is based on an upper bound of the error rate."
Question	"What is the semantic meaning of “average episodic coverage” in Figure 5?"
Answer	"The semantic meaning of “average episodic coverage” in Figure 5 refers to the number of unique avatar positions. The authors have added a DIAYN baseline and a random agent baseline to provide context for how other methods fare."
Question	"In Figure 3, does the number of epochs mean the same thing for BAIL+ and MBAIL?"
Answer	"Yes, the number of epochs in Figure 3 means the same thing for BAIL+ and MBAIL. For MBAIL, it refers to the constant E in algorithm 2."

Table 11: Example QA Pairs in the cPAPERS-FIGS dataset

B Language Modeling Details

We randomly split the dataset into training (60%), dev set (20%), and test set (20%). Hyperparameters for parameter-efficient fine-tuning can be found in 12. Other model training details are listed in 13.

Parameter	Value
Rank	64
α	16
Dropout	0.1

Table 12: Hyperparameters for Parameter-efficient fine-tuning using QLoRA

Parameter	Value
Learning Rate	2e-4
Batch size	4
Warmup Schedule	Constant
Warmup Ratio	0.03
Epochs	5
Optimizer	`paged_adamw_32bit`[9]
Compute	8 Nvidia A40 GPUs

Table 13: Additional hyperparameters for fine-tuning experiments

B.1 Additional Model Performance

We conducted additional experiments to benchmark the baseline performance of state-of-the-art pre-trained LLMs in answering questions in the cPAPERS dataset without additional fine-tuning. In Tables 14, 15, and 16, we report the zero-shot performance of LLAMA-2-7B, LLAMA-2-70B, LLAMA-3-8B, and LLAMA-3-70B on the cPAPERS dataset.

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**	**BERTScore**
LLAMA-2-7B	0.189	0.060	0.136	0.232	0.823
LLAMA-2-70B	0.194	0.065	0.144	0.240	0.825
LLAMA-3-8B	0.139	0.047	0.107	0.161	0.768
LLAMA-3-70B	0.266	0.104	0.203	0.243	0.844

Table 14: Comparison of zero-shot performance across different models on the cPAPERS-EQNS test set

B.2 Impact of Temperatures on Zero-shot Language Modeling Results

We conducted additional experiments to evaluate the influence of temperature on the baseline performance. Temperature parameters were sampled from 0.0 to 1.0 at intervals of 0.1, and each experiment was repeated five times with five randomly generated seeds. Average scores and standard errors across metrics were computed, and the results are presented in tables 17, 18, and 19.

B.3 Impact of Temperatures on Fine-tuning Language Modeling Results

We conducted supplementary experiments to evaluate the influence of temperature on the fine-tuned model. Temperature parameters were sampled from 0.0 to 1.0 at intervals of 0.1, and each experiment was repeated five times with five randomly generated seeds. Average scores and standard errors across metrics were computed, and the results are presented in tables 20, 21, and 20.

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
LLAMA-2-7B	0.190	0.054	0.131	0.227	0.830
LLAMA-2-70B	0.192	0.058	0.136	0.232	0.832
LLAMA-3-8B	0.132	0.039	0.096	0.147	0.763
LLAMA-3-70B	0.256	0.086	0.187	0.217	0.850

Table 15: Comparison of zero-shot performance across different models on the cPAPERS-TBLS test set

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
LLAMA-2-7B	0.187	0.061	0.136	0.237	0.831
LLAMA-2-70B	0.185	0.065	0.137	0.238	0.833
LLAMA-3-8B	0.126	0.045	0.100	0.174	0.784
LLAMA-3-70B	0.282	0.119	0.218	0.256	0.853

Table 16: Comparison of zero-shot performance across different models on the cPAPERS-FIGS test set

Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
0.0	0.194 ± 0.000	0.065 ± 0.000	0.144 ± 0.000	0.240 ± 0.000	0.825 ± 0.000
0.1	0.193 ± 0.001	0.064 ± 0.000	0.143 ± 0.000	0.238 ± 0.000	0.825 ± 0.000
0.3	0.194 ± 0.001	0.065 ± 0.001	0.143 ± 0.001	0.240 ± 0.000	0.825 ± 0.000
0.5	0.194 ± 0.001	0.065 ± 0.000	0.142 ± 0.000	0.240 ± 0.000	0.825 ± 0.000
0.7	0.193 ± 0.001	0.065 ± 0.001	0.142 ± 0.001	0.239 ± 0.000	0.825 ± 0.000
0.9	0.193 ± 0.001	0.064 ± 0.000	0.141 ± 0.001	0.240 ± 0.001	0.825 ± 0.000

Table 17: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperature on the cPAPERS-EQNS test set.

Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
0.0	0.192 ± 0.000	0.058 ± 0.000	0.136 ± 0.000	0.232 ± 0.000	0.832 ± 0.000
0.1	0.192 ± 0.001	0.058 ± 0.000	0.136 ± 0.001	0.231 ± 0.000	0.832 ± 0.000
0.3	0.191 ± 0.001	0.058 ± 0.000	0.136 ± 0.000	0.232 ± 0.000	0.832 ± 0.000
0.5	0.192 ± 0.000	0.059 ± 0.000	0.137 ± 0.000	0.233 ± 0.001	0.832 ± 0.000
0.7	0.192 ± 0.001	0.058 ± 0.001	0.137 ± 0.001	0.233 ± 0.000	0.832 ± 0.000
0.9	0.191 ± 0.001	0.057 ± 0.001	0.136 ± 0.001	0.232 ± 0.000	0.832 ± 0.001

Table 18: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperature on the cPAPERS-TBLS test set.

Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
0.0	0.185 ± 0.000	0.065 ± 0.000	0.137 ± 0.000	0.238 ± 0.000	0.833 ± 0.000
0.1	0.188 ± 0.001	0.066 ± 0.000	0.140 ± 0.000	0.241 ± 0.001	0.834 ± 0.000
0.3	0.189 ± 0.001	0.067 ± 0.001	0.141 ± 0.000	0.241 ± 0.001	0.834 ± 0.000
0.5	0.191 ± 0.001	0.068 ± 0.001	0.143 ± 0.001	0.242 ± 0.001	0.835 ± 0.000
0.7	0.190 ± 0.001	0.067 ± 0.000	0.142 ± 0.000	0.241 ± 0.001	0.834 ± 0.000
0.9	0.190 ± 0.001	0.066 ± 0.001	0.141 ± 0.000	0.240 ± 0.000	0.834 ± 0.000

Table 19: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperature on the cPAPERS-FIGS test set.

Modality	Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
Question (Q)	0.0	0.309 ± 0.000	0.138 ± 0.000	0.248 ± 0.000	0.215 ± 0.000	0.860 ± 0.000
	0.1	0.309 ± 0.001	0.139 ± 0.001	0.249 ± 0.000	0.216 ± 0.000	0.860 ± 0.001
	0.3	0.306 ± 0.001	0.135 ± 0.000	0.245 ± 0.001	0.213 ± 0.000	0.858 ± 0.002
	0.5	0.301 ± 0.001	0.132 ± 0.001	0.242 ± 0.000	0.212 ± 0.001	0.860 ± 0.001
	0.7	0.297 ± 0.001	0.125 ± 0.001	0.236 ± 0.001	0.210 ± 0.001	0.858 ± 0.001
	0.9	0.285 ± 0.002	0.113 ± 0.001	0.223 ± 0.001	0.202 ± 0.002	0.856 ± 0.001
Q + Equation	0.0	0.317 ± 0.000	0.134 ± 0.000	0.251 ± 0.000	0.223 ± 0.000	0.861 ± 0.000
	0.1	0.313 ± 0.001	0.133 ± 0.001	0.250 ± 0.001	0.220 ± 0.001	0.860 ± 0.001
	0.3	0.310 ± 0.001	0.130 ± 0.001	0.245 ± 0.001	0.221 ± 0.001	0.859 ± 0.002
	0.5	0.305 ± 0.002	0.124 ± 0.001	0.239 ± 0.002	0.217 ± 0.002	0.857 ± 0.001
	0.7	0.297 ± 0.002	0.116 ± 0.001	0.231 ± 0.001	0.213 ± 0.002	0.857 ± 0.001
	0.9	0.281 ± 0.002	0.102 ± 0.002	0.215 ± 0.001	0.206 ± 0.002	0.850 ± 0.001
Q + Context	0.0	0.297 ± 0.000	0.126 ± 0.000	0.235 ± 0.000	0.221 ± 0.000	0.817 ± 0.000
	0.1	0.297 ± 0.001	0.126 ± 0.001	0.235 ± 0.000	0.220 ± 0.001	0.820 ± 0.001
	0.3	0.297 ± 0.002	0.123 ± 0.001	0.234 ± 0.001	0.222 ± 0.001	0.824 ± 0.001
	0.5	0.297 ± 0.002	0.119 ± 0.001	0.232 ± 0.002	0.221 ± 0.001	0.832 ± 0.002
	0.7	0.287 ± 0.001	0.108 ± 0.002	0.219 ± 0.003	0.218 ± 0.001	0.832 ± 0.002
	0.9	0.270 ± 0.002	0.089 ± 0.003	0.200 ± 0.004	0.207 ± 0.002	0.826 ± 0.001
Q + References	0.0	0.283 ± 0.000	0.122 ± 0.000	0.224 ± 0.000	0.217 ± 0.000	0.777 ± 0.000
	0.1	0.286 ± 0.001	0.123 ± 0.001	0.226 ± 0.001	0.222 ± 0.001	0.785 ± 0.002
	0.3	0.283 ± 0.001	0.118 ± 0.001	0.222 ± 0.001	0.221 ± 0.001	0.786 ± 0.003
	0.5	0.281 ± 0.002	0.115 ± 0.001	0.219 ± 0.001	0.221 ± 0.002	0.785 ± 0.003
	0.7	0.271 ± 0.001	0.106 ± 0.001	0.208 ± 0.002	0.216 ± 0.002	0.781 ± 0.002
	0.9	0.258 ± 0.001	0.093 ± 0.001	0.193 ± 0.002	0.209 ± 0.002	0.770 ± 0.002

Table 20: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperatures on the cPAPERS-EQNS test set.

Modality	Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
Question (Q)	0.0	0.315 ± 0.000	0.121 ± 0.000	0.235 ± 0.000	0.218 ± 0.000	0.869 ± 0.000
	0.1	0.316 ± 0.001	0.122 ± 0.001	0.234 ± 0.000	0.220 ± 0.001	0.869 ± 0.001
	0.3	0.313 ± 0.001	0.120 ± 0.001	0.232 ± 0.001	0.220 ± 0.001	0.869 ± 0.001
	0.5	0.309 ± 0.001	0.116 ± 0.001	0.229 ± 0.001	0.216 ± 0.001	0.868 ± 0.000
	0.7	0.302 ± 0.002	0.108 ± 0.002	0.221 ± 0.001	0.213 ± 0.001	0.868 ± 0.001
	0.9	0.288 ± 0.003	0.095 ± 0.002	0.208 ± 0.002	0.205 ± 0.003	0.865 ± 0.000
Q + Table	0.0	0.293 ± 0.000	0.107 ± 0.000	0.218 ± 0.000	0.212 ± 0.000	0.820 ± 0.000
	0.1	0.288 ± 0.002	0.105 ± 0.001	0.214 ± 0.001	0.211 ± 0.001	0.815 ± 0.003
	0.3	0.289 ± 0.002	0.103 ± 0.001	0.214 ± 0.001	0.215 ± 0.002	0.837 ± 0.003
	0.5	0.286 ± 0.001	0.096 ± 0.001	0.207 ± 0.001	0.216 ± 0.001	0.844 ± 0.003
	0.7	0.277 ± 0.002	0.085 ± 0.001	0.196 ± 0.001	0.212 ± 0.004	0.844 ± 0.003
	0.9	0.249 ± 0.002	0.064 ± 0.001	0.170 ± 0.001	0.200 ± 0.004	0.833 ± 0.004
Q + Context	0.0	0.294 ± 0.000	0.106 ± 0.000	0.216 ± 0.000	0.225 ± 0.000	0.838 ± 0.000
	0.1	0.301 ± 0.001	0.108 ± 0.000	0.219 ± 0.000	0.227 ± 0.000	0.846 ± 0.002
	0.3	0.301 ± 0.001	0.108 ± 0.001	0.218 ± 0.001	0.227 ± 0.001	0.846 ± 0.001
	0.5	0.297 ± 0.002	0.104 ± 0.001	0.214 ± 0.002	0.226 ± 0.002	0.848 ± 0.002
	0.7	0.287 ± 0.002	0.092 ± 0.001	0.203 ± 0.002	0.218 ± 0.003	0.845 ± 0.002
	0.9	0.272 ± 0.002	0.079 ± 0.001	0.187 ± 0.002	0.213 ± 0.003	0.842 ± 0.001
Q + Reference	0.0	0.292 ± 0.000	0.106 ± 0.000	0.214 ± 0.000	0.218 ± 0.000	0.816 ± 0.000
	0.1	0.297 ± 0.001	0.108 ± 0.001	0.217 ± 0.001	0.222 ± 0.001	0.830 ± 0.002
	0.3	0.298 ± 0.002	0.105 ± 0.001	0.217 ± 0.001	0.224 ± 0.001	0.842 ± 0.002
	0.5	0.294 ± 0.002	0.101 ± 0.001	0.212 ± 0.002	0.221 ± 0.001	0.843 ± 0.003
	0.7	0.288 ± 0.001	0.092 ± 0.001	0.205 ± 0.001	0.219 ± 0.001	0.847 ± 0.001
	0.9	0.272 ± 0.002	0.081 ± 0.001	0.190 ± 0.001	0.212 ± 0.002	0.837 ± 0.002

Table 21: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperatures on the cPAPERS-TBLS test set.

Modality	**Temp**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**	**BERTScore**
Question (Q)	0.0	0.329 ± 0.000	0.155 ± 0.000	0.269 ± 0.000	0.250 ± 0.000	0.859 ± 0.000
	0.1	0.333 ± 0.001	0.157 ± 0.001	0.272 ± 0.001	0.252 ± 0.001	0.859 ± 0.002
	0.3	0.329 ± 0.002	0.153 ± 0.001	0.267 ± 0.001	0.249 ± 0.001	0.860 ± 0.003
	0.5	0.324 ± 0.002	0.145 ± 0.002	0.259 ± 0.002	0.244 ± 0.002	0.861 ± 0.002
	0.7	0.315 ± 0.002	0.134 ± 0.002	0.248 ± 0.002	0.242 ± 0.002	0.862 ± 0.001
	0.9	0.291 ± 0.001	0.113 ± 0.002	0.223 ± 0.001	0.232 ± 0.002	0.857 ± 0.003
Question + Figure	0.0	0.322 ± 0.000	0.147 ± 0.000	0.260 ± 0.000	0.246 ± 0.000	0.868 ± 0.000
	0.1	0.325 ± 0.001	0.149 ± 0.001	0.262 ± 0.001	0.246 ± 0.001	0.865 ± 0.001
	0.3	0.322 ± 0.002	0.146 ± 0.001	0.258 ± 0.002	0.245 ± 0.002	0.865 ± 0.001
	0.5	0.318 ± 0.001	0.139 ± 0.001	0.251 ± 0.001	0.242 ± 0.001	0.865 ± 0.001
	0.7	0.306 ± 0.001	0.125 ± 0.001	0.239 ± 0.001	0.238 ± 0.002	0.863 ± 0.001
	0.9	0.281 ± 0.001	0.102 ± 0.002	0.212 ± 0.001	0.226 ± 0.001	0.852 ± 0.003
Context	0.0	0.321 ± 0.000	0.140 ± 0.000	0.256 ± 0.000	0.250 ± 0.000	0.858 ± 0.000
	0.1	0.317 ± 0.002	0.138 ± 0.001	0.254 ± 0.001	0.248 ± 0.001	0.853 ± 0.002
	0.3	0.323 ± 0.002	0.138 ± 0.001	0.255 ± 0.001	0.252 ± 0.001	0.861 ± 0.001
	0.5	0.318 ± 0.001	0.132 ± 0.002	0.248 ± 0.002	0.250 ± 0.002	0.863 ± 0.001
	0.7	0.307 ± 0.001	0.122 ± 0.002	0.236 ± 0.002	0.242 ± 0.002	0.861 ± 0.001
	0.9	0.278 ± 0.001	0.097 ± 0.002	0.209 ± 0.002	0.225 ± 0.002	0.851 ± 0.001
Reference	0.0	0.311 ± 0.000	0.131 ± 0.000	0.244 ± 0.000	0.247 ± 0.000	0.843 ± 0.000
	0.1	0.314 ± 0.002	0.134 ± 0.001	0.248 ± 0.001	0.250 ± 0.001	0.848 ± 0.003
	0.3	0.310 ± 0.002	0.130 ± 0.001	0.244 ± 0.001	0.249 ± 0.002	0.845 ± 0.002
	0.5	0.308 ± 0.002	0.125 ± 0.001	0.240 ± 0.001	0.249 ± 0.002	0.857 ± 0.003
	0.7	0.296 ± 0.001	0.113 ± 0.001	0.226 ± 0.001	0.241 ± 0.001	0.854 ± 0.002
	0.9	0.269 ± 0.001	0.090 ± 0.001	0.198 ± 0.002	0.226 ± 0.002	0.843 ± 0.001

Table 22: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperatures on the cPAPERS-FIGS test set.