SciMRC: Multi-perspective Scientific Machine Reading Comprehension

Xiao Zhang[123*], **Heqi Zheng**[4*], **Yuxiang Nie**[5*], **Heyan Huang**[123+], **Xian-Ling Mao**[123]
[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
[3]Southeast Academy of Information Technology, Beijing Institute of Technology
[4]State Grid Smart Grid Research Institute Co., Ltd.
[5]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
xiaoz987@gmail.com, hikizheng@gmail.com, ynieae@connect.ust.hk
{hhy63,maoxl}@bit.edu.cn

Abstract

Scientific Machine Reading Comprehension (SMRC) aims to facilitate the understanding of scientific texts through human-machine interactions. While existing dataset has significantly contributed to this field, it predominantly focus on single-perspective question-answer pairs, thereby overlooking the inherent variation in comprehension levels among different readers. To address this limitation, we introduce a novel multi-perspective scientific machine reading comprehension dataset, SciMRC, which incorporates perspectives from beginners, students, and experts. Our dataset comprises 741 scientific papers and 6,057 question-answer pairs, with 3,306, 1,800, and 951 pairs corresponding to beginners, students, and experts respectively. Extensive experiments conducted on SciMRC using pre-trained models underscore the importance of considering diverse perspectives in SMRC and highlight the challenging nature of our scientific machine comprehension tasks.

Keywords: Corpus, Question Answering, Natural Language Generation

1. Introduction

Scientific machine reading comprehension (SMRC) aims to understand scientific texts through interactions with humans by given questions. The ability of machines to understand and make sense of scientific texts is crucial for many applications such as scientific research (Cachola et al., 2020; Beltagy et al., 2019; Marie et al., 2021), education (de la Chica et al., 2008; Bianchi and Giorcelli, 2019) and industry (Zulfiqar et al., 2018; Bruches et al., 2022; Erera et al., 2019). With the increasing amount of scientific literature being produced, the need (Wadden et al., 2020; Sadat and Caragea, 2022; Dasigi et al., 2021) for machines to understand these texts is becoming more pressing.

While the dataset presented by Dasigi et al. (2021) has contributed significantly to this field by focusing on full-text scientific machine reading comprehension, it has predominantly concentrated on a specific aspect: enhancing machine reading comprehension (MRC) models for extracting information from scientific papers in response to questions. However, it has inadvertently disregarded a pivotal element, the inherent variation in comprehension levels among readers when digesting the same text. This dataset is constructed solely from question-answer pairs generated by annotators with NLP backgrounds, thus excluding a broader range of viewpoints, especially those of novices and domain experts. This exclusion of diverse perspectives is a critical limitation, as each perspective is associated with unique challenges, necessitating distinct depths of understanding. Addressing this limitation

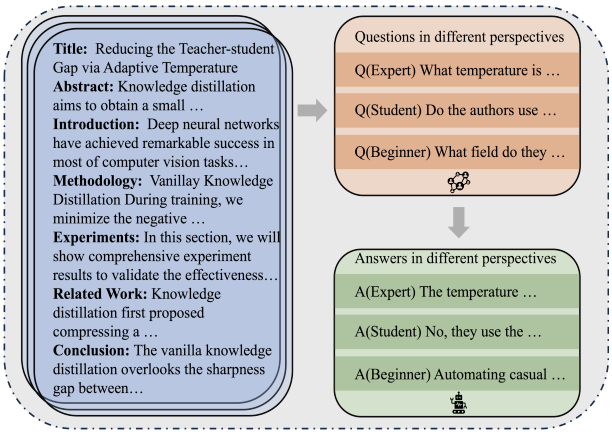

Figure 1: An illustration of an instance from our SciMRC dataset. Here, questions are annotated from three distinct perspectives - beginners, students, and experts. The objective is for the machine to provide appropriate responses that cater to the unique comprehension levels and needs of each user group.

* Equal contribution.
\+ Corresponding author.

LREC-COLING 2024, pages 14418–14428
20-25 May, 2024.

Question	Perspective	%
How does the number of hierarchical levels pre-defined in real scenarios? The two experiments used L(m) = 2, will that be better to set it as a tuned parameter?	Experts	15.7
What are the advantages of their model compared with the existing MRC systems?	Students	29.7
How is the data for training collected?	Beginners	54.6

Question	Answer	Type	%
What task leaderboard does the paper describe?	The commonsense reasoning task leaderboard of the AI2 WinoGrande Challenge.	Extractive	56.5
What is the function of the modality attention module in this paper?	It selectively chooses modalities to extract primary context from, maximizing information gain and suppressing irrelevant contexts from each modality.	Generative	25.7
Do the authors use a new dataset for training?	No.	Yes|No	10.0
Why is "sparse linear system[s]" preferred here instead of general linear and nonlinear ones?	N/A	Unanswerable	7.8

Table 1: Examples of questions, answers in different types sampled from SciMRC. % are relative frequencies of the corresponding type over all examples in SciMRC

and advancing the field of SciMRC requires a more holistic approach that incorporates and embraces these varied viewpoints.

To tackle the above problem, we introduce a pioneering contribution in the form of SMRC, a multi-perspective scientific paper machine reading comprehension dataset. SciMRC[1] comprises perspectives representing beginners, students, and experts. We adopt a multi-perspective annotation strategy to curate these perspectives. The beginner perspective is annotated by non-expert annotators, the student perspective is guided by summarization knowledge, and the expert perspective is based on open-review expert reviews. By adopting this multi-perspective approach, we effectively capture the full spectrum of comprehension levels and expertise in scientific machine reading comprehension. To emphasize the distinct characteristics of each perspective, we classify the questions into 28 categories based on the specific issues of concern to experts, students and beginners in academic institutions.

Our SciMRC dataset boasts a substantial collection of 741 scientific papers and a total of 6,057 question-answer pairs. These are distributed across the three perspectives as follows: 3,306 pairs for beginners, 1,800 pairs for students, and 951 pairs for experts. Our contribution goes beyond mere dataset creation; it also entails a comprehensive exploration of the impact of perspectives on scientific machine reading comprehension. Through extensive experiments involving various pre-trained models, we highlight the significance of considering different perspectives within SciMRC. Furthermore, our research underscores the inherent challenges posed by these multiple perspectives, making it clear that SciMRC is a more intricate task than previously acknowledged.

In summary, our contributions are summarized as follows:

- We present SciMRC, an innovative multi-perspective scientific machine reading comprehension dataset, encompassing perspectives from beginners, students, and experts.
- Through extensive experiments on SciMRC using pre-trained models, we underscore the imperative of acknowledging and considering the diverse perspectives in SciMRC. Furthermore, we illuminate the complex and challenging nature of the machine comprehension task in the context of SMRC.

2. SciMRC

In this section, we provide a comprehensive overview of the multi-perspective annotation strategy employed in constructing SciMRC.

2.1. Data Preparation

To assemble our dataset, we draw from multiple reputable sources, including S2ORC (Lo et al., 2020), QASPER (Dasigi et al., 2021), and openreview. These sources provide a diverse collection of scientific papers, which are processed to obtain pure textual content through a parser. Our dataset comprises a substantial 3,000 papers, ensuring a wide representation of scientific topics. To capture the varying interests and comprehension levels of readers, we take a proactive approach. We design a comprehensive questionnaire that is distributed early in the dataset creation process. This questionnaire is targeted at experts and students in relevant fields, beginners in academic institutes, soliciting their insights on key points within scientific papers and their top five concerns when

[1]Our dataset is publicly available at github.com/Yottaxx/SciMRC-Multi-Perspective

Type	Paper	Figure/Table	Question	Answer				Evidence
PERSPECTIVE	Avg Paper Length	Avg Figure/Table Number	Avg Question Length	Avg Answer Length	Yes|No	Generative	Extractive	Unanswerable	Avg Evidence Sentence Number
BEGINNERS	3725.6	5.32	10.0	17.2	331	754	2220	1	1.39
STUDENTS			9.8	11.7	266	340	1194	0	1.08
EXPERTS			22.4	95.9	5	467	8	471	4.56
ALL			11.0	21.8	602	1561	3422	472	1.56

Table 2: Representative features from SciMRC categorized by different perspectives.

1	Methods	2	Experimental results	3	Dataset/code	4	Experimental settings
5	Model architecture	6	Experimental analysis	7	Related work	8	Background
9	Baseline	10	Motivation	11	Contribution	12	Innovation point
13	Future work	14	Research field	15	Research background	16	Evaluation Metric
17	Insufficiency of previous work	18	Experimental conditions	19	Assumption	20	Model transferability
21	Limitations	22	Innovation	23	Thesis Outline	24	Landing Application
25	Case Study	26	Publication Details	27	Complexity	28	Whether to propose a new task

Table 3: All the 28 question categories in SciMRC

engaging with such documents. By collating and summarizing the responses, we distill 780 readers' questions into 28 distinct categories, which form a valuable foundation for the annotation process.

Questions Collection Our dataset's question collection phase is pivotal, as it involves the systematic generation and categorization of questions from different perspectives. Annotators are tasked with reading materials according to the specific perspective criteria assigned to them. Their responsibility is to either craft questions or extract them from the materials, subsequently mapping each question to one of the 28 predefined categories.

Answers Collection Equally significant is the answers collection stage. Here, annotators are entrusted with the critical task of locating pertinent answers and supporting evidence for the previously annotated questions. The nature of the answers dictates categorization, as they are assigned to either EXTRACTION, GENERATIVE, or YES/NO based on the content. Should no supporting evidence be found within the paper, the answer type is categorized as UNANSWERABLE.

2.2. Multi-perspective Annotation Strategy

To encapsulate the wide array of reader perspectives and levels of understanding, we categorize our dataset into three distinct reader categories: Beginner, Student, and Expert perspectives. To ensure the integrity of our data collection, we partner with a professional labeling company[2], followed by meticulous quality checks by graduate students with expertise in the relevant academic domains.

Beginner's Perspective Beginners, often with limited prior exposure to academic papers and the field, are captured through this perspective. Annotators without domain-specific knowledge craft questions and seek answers within the paper's full text, figures, and tables.

Student's Perspective Students, equipped with foundational knowledge of the academic field, possess specific viewpoints on the purpose, methodology, findings, and significance of the paper. To capture this perspective, we leverage the FacetSum (Meng et al., 2021) summarization dataset. This model transforms the paper's full text into four key aspects: purpose, method, finding, and value. Annotators verify the accuracy of the generated abstracts and formulate questions based on them, subsequently searching for answers within the paper.

Expert's Perspective Experts in the field bring a wealth of insights and detailed opinions to scientific papers. To incorporate this perspective, we acquire related reviews from open-review. Annotators extract questions and answers from reviewer comments and author responses. Questions that delve beyond the paper's content and require background knowledge are categorized as UNANSWERABLE. These questions offer valuable insights into the model's ability to address questions beyond the paper's scope.

[2]Compensation for annotation was provided on a per HIT basis, varying from $0.50 to $2.00 for each QA pair, depending on the perspective.

Figure 2: The distribution of question categories

Ensure the Quality of Annotation At each stage of the annotation process, we implement rigorous quality checks to maintain the integrity, reliability, and accuracy of our dataset. Our commitment to data quality ensures that SciMRC is a robust resource for the NLP community and aligns closely with the objectives laid out in the introduction. This multi-perspective dataset holds the potential to catalyze advancements in scientific machine reading comprehension, fostering a more inclusive and comprehensive understanding of the field.

3. Analysis of SciMRC

The SciMRC dataset is a comprehensive collection of 741 scientific papers, accompanied by 6,057 question-answer pairs. These pairs are meticulously categorized into three distinct perspectives: beginners, students, and experts. The dataset is further divided into training, validation, and test sets, following a 7:1:2 ratio, ensuring a balanced distribution for model training and evaluation. A detailed breakdown of the dataset's key features, categorized by the different perspectives, is presented in Table 2 and Figure 2. In the following subsections, we delve deeper into the analysis of these features.

3.1. Question Categories

In this subsection, we delve into the diversity and complexity of question categories within the SciMRC dataset. We have identified and categorized 28 distinct question categories from a sample of 780 questions[3]. These categories are designed to cater to the varying perspectives and levels of understanding of scientific papers. For instance, as shown in Figure 2, the expert perspective predominantly focuses on 'Methods' (36.3%), 'Experimental settings' (15.6%), and 'Experimental analysis' (10.0%). On the other hand, the student perspective shows a similar inclination towards 'Methods' (29.1%) but also gives considerable attention to 'Model architecture' (13.4%) and 'Dataset/code' (10.2%). The beginner perspective, while also prioritizing 'Methods' (16.8%), shows a more balanced distribution across 'Experimental settings' (12.5%) and 'Dataset/code' (12.0%). This analysis reveals that while 'Methods' is a common area of interest across all perspectives, the emphasis on other aspects varies. Experts tend to delve deeper into experimental details, students show a keen interest in model architectures, and beginners seek a more balanced understanding across different aspects. This diversity in question categories underscores the complexity of the SciMRC dataset and highlights the need for models to cater to a wide range of comprehension levels and expertise.

3.2. Evidence Selection

Our analysis extends to the selection of evidence by annotators. The comprehensiveness and complexity of questions are implicitly reflected in the number of sentences in annotated evidence. Table 2 illustrates the average number of sentences in evidence for different perspectives. Experts' questions require more comprehensive evidence, with an average of approximately 4.56 sentences per evidence. Beginners and students' perspectives generally involve less extensive evidence, with an average of about 1.39 and 1.08 sentences per evidence, respectively. These findings underscore the varying demands for evidence based on the perspective of the questioner.

3.3. Answer Types

The distribution of answer types across the three perspectives is presented in Table 2. Notable observations include Students and beginners share a similar distribution of answer types, with approximately 2/3 of answers being of the EXTRACTIVE type. Experts exhibit an increased proportion of

[3]The details of categories are illustrated in Table 3.

"Unanswerable" questions and generative answers , reflecting their questions about experiments and motivations that may not be explicitly mentioned in papers. Answers provided by experts tend to be longer in length, with an average of approximately 96 words per answer, compared to other perspectives.

4. Modeling

In order to accommodate the diverse answer types present in SciMRC, we employ text-to-text transformers, specifically T5 (Raffel et al., 2020) and LED (Beltagy et al., 2020), to model the scientific machine reading comprehension task.

4.1. Text-to-Text Transformer

The Text-to-Text Transformer (T5) (Raffel et al., 2020) is a model based on the Transformer (Vaswani et al., 2017) encoder-decoder architecture. It reframes all text processing tasks, such as question answering and classification, as a "text-to-text" problem. Each fine-tuning task is prefixed with a specific task identifier.

4.2. Longformer Encoder Decoder

The Longformer Encoder Decoder (LED) (Beltagy et al., 2020) is a specialized Transformer encoder-decoder model that integrates both local and global attention patterns. It substitutes the original self-attention mechanism with several sliding window attention patterns. This modification allows LED to scale its self-attention computation linearly with the input size, as opposed to the quadratic computation of full text self-attention.

4.3. Training objective

Given a context, a question, and the final answer a, which is composed of variable-length tokens x_i, the probabilities over the tokens are calculated as follows:

$$p(a) = \prod_{1}^{m} p(x_i|x_{<i}, f_e; \theta), \tag{1}$$

where θ donates the trainable parameters of our model. The training objective is computed as illustrated as following:

$$\mathcal{L}_{oss} = -\sum_{i=1}^{M} log\, p(x_i|x_{<i}, f_w; \theta), \tag{2}$$

4.4. Metrics

Given the presence of YES|NO answer types in SciMRC, we use both Rouge-L and accuracy as automatic proxies for the correctness of answers.

Rouge-L As our dataset contains many generative answers, we evaluate our reading comprehension task using Rouge-L (Lin, 2004) [4], a widely used metric in language generation evaluation. Rouge-L calculates the length of the longest common sequence (LCS) between the predicted answer (Prediction) and the golden answer (Golden) to compute the final score as follows:

$$\mathcal{R}_{LCS} = \frac{LCS(Prediction, Golden)}{len(Golden)} \tag{3}$$

$$\mathcal{P}_{LCS} = \frac{LCS(Prediction, Golden)}{len(Prediction)} \tag{4}$$

$$\mathcal{F}_{LCS} = \frac{(1+\beta^2)\mathcal{R}_{LCS}\mathcal{P}_{LCS}}{\mathcal{R}_{LCS} + \beta^2\mathcal{P}_{LCS}} \tag{5}$$

Accuracy For YES|NO questions, we use accuracy as the evaluation metric. For "yes" questions, an answer that begins with "yes" is considered correct. Conversely, for "no" questions, answers that do not begin with "yes" are deemed correct.

5. Experiments

We conducted a series of experiments to investigate the challenges posed by multi-perspective problems in SciMRC. The results, as shown in Table 4, demonstrate the performance of various models under different training settings. We used T5 and LED as our backbone models and trained them using different combinations of perspectives from beginners, students, and experts. We also examined the performance of different answer formats in SciMRC on YES|NO, GENERATIVE, EXTRACTIVE, and UNANSWERABLE subsets, as shown in Table 5.

The experimental settings considered are as follows:

- **B-formed**: The model is trained using the beginner's perspective training data.
- **S-formed**: The model is trained using the student's perspective training data.
- **E-formed**: The model is trained using the expert's perspective training data.
- **BS-formed**: The model is trained using both beginner's and student's perspective training data.
- **BE-formed**: The model is trained using both beginner's and expert's perspective training data.

[4]The implementation we used is from huggingface (https://huggingface.co/).

Model	Dev				Test			
	Beginner	Student	Expert	Overall	Beginner	Student	Expert	Overall
T5-B	24.77±0.27	39.50±0.72	11.29±0.45	26.47±0.23	25.60±0.34	41.51±0.44	11.39±0.20	29.03±0.13
T5-S	24.32±0.11	45.08±0.98	10.64±0.28	27.80±0.40	24.68±0.31	44.16±0.59	9.97±0.13	29.16±0.33
T5-E	9.74±0.16	17.03±0.57	**14.76**±0.20	13.11±0.18	11.12±0.21	18.19±0.66	**15.66**±0.19	13.79±0.34
T5-SE	23.50±0.33	45.17±0.28	13.34±0.07	28.09±0.08	24.26±0.29	43.34±0.84	13.97±0.43	29.12±0.21
T5-BE	25.18±0.49	39.65±0.47	13.64±0.71	27.21±0.46	25.94±0.02	41.50±0.82	14.68±0.61	29.61±0.30
T5-SB	26.06±0.54	45.93±0.87	11.24±0.16	29.12±0.41	**26.46**±0.15	**46.27**±0.22	10.92±0.09	30.99±0.21
T5-BSE	**26.87**±0.89	**46.86**±0.72	13.60±0.24	30.25±0.70	26.34±0.49	45.28±0.24	14.53±0.28	31.05±0.23
Per-T5	**26.87**±0.89	**46.86**±0.72	**14.76**±0.20	**30.45**±0.60	26.34±0.49	45.28±0.24	**15.66**±0.19	**31.18**±0.23
LED-B	25.51±1.61	32.1±2.06	10.02±0.29	24.25±1.33	25.02±1.21	33.15±0.67	11.36±0.23	26.15±0.93
LED-S	22.55±2.26	41.93±6.73	9.16±0.5	25.77±3.11	21.64±1.95	44.05±5.12	9.69±0.47	27.47±2.72
LED-E	6.80±0.49	7.75±0.47	14.62±0.32	8.76±0.45	7.22±0.53	7.98±0.75	**14.77**±0.11	8.27±0.53
LED-SE	24.57±0.35	45.16±1.03	13.29±0.34	28.55±0.48	23.79±0.27	47.42±1.23	12.52±0.48	30.00±0.14
LED-BE	27.40±0.94	32.62±1.15	13.02±0.69	25.89±0.64	25.80±0.95	33.81±1.03	12.68±0.15	26.90±0.88
LED-SB	31.07±0.07	**49.67**±0.27	9.07±0.3	32.17±0.25	29.78±0.4	**47.93**±0.42	10.41±0.39	33.37±0.41
LED-BSE	**32.24**±0.43	47.04±2.38	12.85±0.75	32.04±0.87	**29.80**±0.55	46.39±0.34	12.54±0.08	33.20±0.36
Per-LED	**32.24**±0.43	47.04±2.38	**14.62**±0.32	**32.47**±0.69	**29.80**±0.55	46.39±0.34	**14.77**±0.11	**33.44**±0.34

Table 4: The performance of answer generation on the validation and test set with different eval settings. The average results with standard deviation on 3 random seeds are reported.

Model	Dev				Test			
	Yes|No	Generative	Extractive	Unanswerable	Yes|No	Generative	Extractive	Unanswerable
T5-BSE	81.61±1.63	25.67±0.12	30.27±0.84	13.04±0.33	**83.61**±0.67	23.16±0.11	29.58±0.24	14.86±0.30
Per-T5	81.61±1.63	26.14±0.30	30.38±0.93	14.23±0.22	**83.61**±0.67	23.39±0.07	29.61±0.21	**16.11**±0.03
LED-BSE	**82.18**±0.82	25.56±0.81	34.78±1.43	12.79±0.57	78.69±1.34	24.81±0.36	33.54±0.36	13.17±0.29
Per-LED	**82.18**±0.82	**26.32**±0.79	**34.83**±1.44	**14.53**±0.11	78.69±1.34	**25.66**±0.40	**33.57**±0.33	13.99±0.23

Table 5: The performance of different answer types on the validation set and test set. The average results with standard deviation on 3 random seeds are reported.

- **SE-formed**: The model is trained using both student's and expert's perspective training data.
- **BSE-formed**: The model is trained using all training data.
- **Per-formed**: This setting involves the use of BSE-formed model and E-formed model in known question perspectives settings. The E-formed model is used for evaluating the expert perspective, while the BSE-formed model is used for evaluating student and beginner perspectives.

5.1. Results

Multi-Perspective We evaluated the performance of the models on different perspectives: beginner, student, expert, and overall. The evaluation was conducted on both a development set and a test set, with Rouge-L scores separated by beginner, student, expert, and overall perspectives. As shown in Table 4, the perspective-formed models, Per-T5, and Per-LED, achieved the highest Rouge-L scores on both the development and test sets. Specifically, Per-LED achieved the highest performance under the assumption of known perspectives, with a score of 33.44 on the test set. The E-formed models performed best for the expert perspective, with an overall mean performance of 15.66 and 14.77 on the test set, respectively. In general, the models performed better on the student and beginner perspectives compared to the expert perspective, indicating that the expert perspective is more challenging in SciMRC.

Multi-Type of Answer We compared the performance of several models (T5-BSE, Per-T5, LED-BSE, Per-LED) on SciMRC across four categories of answer types: YES|NO, GENERATIVE, EXTRACTIVE, and UNANSWERABLE. The results, presented in Table 5, show that T5-BSE and Per-T5achieved the highest performance on the YES|NO answer type on the test set, with scores of 83.61 each. For the GENERATIVE answer type, Per-T5performed worse than Per-LED on both the dev and test sets. However, Per-LED achieved the best performance on the EXTRACTIVE answer type on the test set. For the UNANSWERABLE answer type, Per-T5 achieved the best performance on the test set.

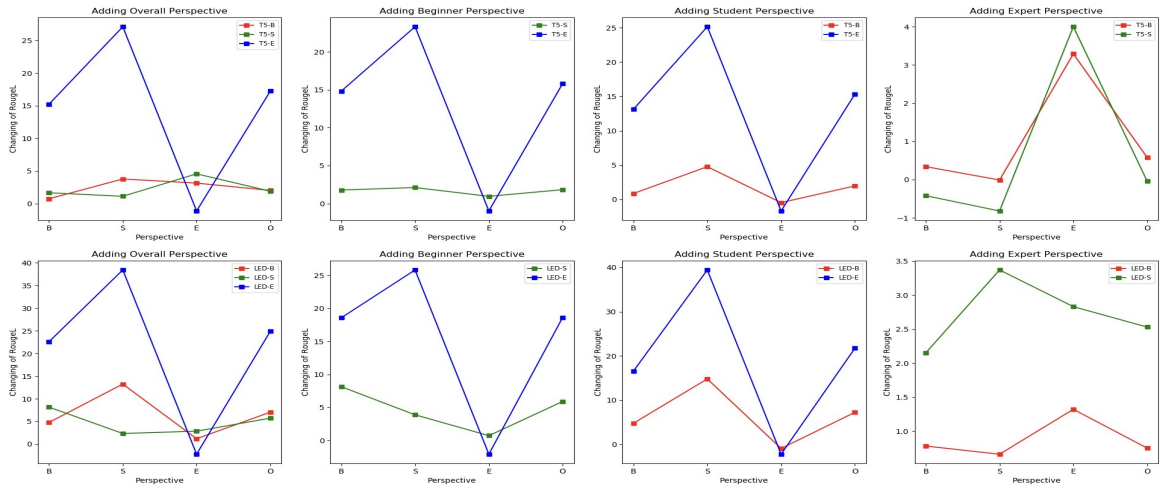

Figure 3: Based on the data of either perspective of S, B and E, we further add other perspective data, and then observe the change of Rouge-L scores on test set to quantity the influence of performance.

Model	Beginner	Student	Expert	Overall
Per-T5	26.34	45.28	**15.66**	31.66
Per-LED	**29.80**	**46.39**	14.77	**33.44**
GPT3.5-turbo	21.25	26.53	13.47	22.10

Table 6: Comparison of experimental results for various models on test set of SciMRC

5.2. Analysis

Performance Correlation with Perspectives As depicted in Table 4, the performance of the models is intimately tied to the perspectives. To further investigate this, we analyzed the influence of different perspective settings on the test set. The results indicate that incorporating the beginner perspective into the model enhances the overall performance, as evidenced by the superior performance of SB-formed and BSE-formed models compared to S-formed and SE-formed models. Similarly, the inclusion of the student perspective also positively impacts the overall performance, as demonstrated by the higher performance of B-formed and BE-formed models. The expert perspective also appears to have a positive effect on the overall performance. The T5-E model, which solely includes the expert perspective, achieves the highest mean performance for the expert perspective with a score of 15.66 on the test set. Furthermore, models that incorporate the expert perspective, such as SE-formed, BE-formed, SB-formed, and BSE-formed, all exhibit higher overall mean performance compared to models that exclude the expert perspective, such as S-formed, B-formed, and SB-formed.

Interplay Among Perspectives Figure 3 illustrates the interplay among perspectives. The fusion of student and beginner perspectives can enhance the model's comprehension ability across all perspectives. Further inclusion of expert perspective data can boost the understanding ability of the expert perspective, albeit with a slight reduction in performance for other perspectives. Moreover, models exhibit superior performance on student and beginner perspectives compared to the expert perspective, indicating that the expert perspective is more challenging than the student and beginner perspectives in SciMRC. This is evidenced by the lower performance scores of the models in the expert perspective. The expert perspective demands a higher level of understanding and subject matter knowledge, as the questions are posed by experienced reviewers with a deeper grasp of the topic. This necessitates a more complex understanding and more external subject matter knowledge compared to the student and beginner perspectives, making the expert perspective more challenging for the models to predict.

Do Answer Types Reflect Question Difficulty? As shown in Table 5, the scores for the UNANSWERABLE type are significantly lower than those for other answer types. This is because the answers of the UNANSWERABLE type require external knowledge of a specific academic topic and cannot be directly reasoned from the context. Additionally, LED-based models outperform T5-based models for the EXTRACTIVE type answers, with a score of 33.57 for Per-LED, underscoring the importance of context length. LED-based models utilize a 4k tokens context, while T5-based models only use a 512 tokens context. The experimental results also suggest that current models face challenges with EXTRACTIVE, GENERATIVE, and UNANSWERABLE answer types, especially the UNANSWERABLE type, demonstrating the lack of inner academic knowledge in current pre-trained models.

Title	Question	Prediction	Golden	Perspective
Stochastic Answer Networks for SQuAD 2.0	What data sets did the author use?	SQuAD 2.0 dataset, a new MRC dataset which is a combination of Stanford Question Answering Dataset (SQuAD) 1.0 and additional UNANSWERABLE question-answer pairs.	They evaluate their system on SQuAD 2.0 dataset, a new MRC dataset which is a combination of Stanford Question Answering Dataset (SQuAD) 1.0 and additional UNANSWERABLE question-answer pairs.	Beginner
From Textual Information Sources to Linked Data in the Agatha Project	What other related research path do the authors have explored?	The authors have already explored other related research path, namely author profiling BIBREF2, aggression identification BIBREF3 and hate-speech detection BIBREF4 over social media, plus statute law retrieval and entailment for Japanese.	Author profiling, aggression identification and hate-speech detection over social media, plus statute law retrieval and entailment for Japanese.	Student
Edge Partition Modulated Graph Convolutional Networks	As the authors imply, it is becoming more important to take into account real-world limitations when deploying MARL (e.g., communication failure). Could you discuss some relevance to a 2019 ICLR paper called SchedNet by Kim et al.? (which, to the best of my knowledge, is one of the earliest MARL works addressing real-life communications constraints.	We agree with Kim et al. that it is becoming more important to take into account real-world limitations when deploying MARL (e.g., communication failure). Therefore, we proposed the imputation assisted multiagent reinforcement learning (IA-MARL) that can achieve the performance of MARL without missing training data when sufficient imputation accuracy is supported.	In Kim et al (2019), the authors proposed SchedNet for the environment having limited communication channels between agents "during execution", but assumed that the training data always exists during training. Note that the communication between the agents during the execution is to encourage coordination among the agents. On the other hand, in our work, the training data collected from agents can be missed, which affects "training of agents". To summarize, there are two differences between SchedNet and IA-MARL: 1) SchedNet considers limited communication during execution and IA-MARL considers data missing for the training 2) the number of messages in ScheNet is limited while IA-MARL cannot obtain the missed data. We will cite and discuss it in the related works.	Expert

Table 7: The predictions of Per-T5 and Per-LED on the test set of SciMRC.

Evaluating the Efficacy of GPT in SciMRC The application of Generative Pre-trained Transformer (GPT) models (Brown et al., 2020; Ouyang et al., 2022) in the realm of scientific machine reading comprehension (SciMRC) warrants a thorough investigation. Given the complexity and diversity of perspectives inherent in SciMRC, it is crucial to assess whether GPT models can effectively comprehend and respond to scientific texts across different levels of understanding. Our research delves into this question, evaluating the zero-shot performance of GPT3.5-turbo on our novel multi-perspective dataset, SciMRC. Our findings, as presented in Table 6, reveal that while GPT3.5-turbo exhibit some degree of proficiency in this task, their performance varies significantly across different perspectives. Specifically, the GPT3.5-turbo model achieved a relatively lower overall score compared to other pre-trained models, indicating that the task of SciMRC may pose unique challenges that cannot be directly addressed by GPT models.

5.3. Case Study

Table 7 illustrates the predictions made by Per-T5 and Per-LED . The Per-LED model, which exhibits superior performance on the test set for student and beginner perspectives, is used to predict data from these perspectives. Conversely, the Per-T5 model, which demonstrates the highest performance on the test set for expert perspectives, is employed to predict data from this perspective. As depicted in Table 7, the predictions for the student and beginner perspectives indicate the models' robust comprehension abilities. However, the predictions for the expert perspective reveal that the comprehension requirements for experts are more complex than those for beginners and students. Although Per-T5 shows potential in comprehending the context at the expert level, its performance is still limited. To reason through QA pairs from the expert perspective, machines require not only a background in research but also a deep understanding of the specific research area.

6. Related work

Information Seeking MRC Datasets MRC tasks can be regarded as information-seeking work, especially when reasoning from a question to an answer needs a complex information-seeking strategy. For example, SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017) and HotpotQA (Yang et al., 2018) need to seek important information related to a question. However, these datasets are mainly extractive, which constrains the flexibility and diversity of the QA pairs. Like QASPER the type of answers in SciMRC is diverse. It makes SciMRC closer to real-world settings. Furthermore, the questions of general domain MRC datasets including WikiQA (Yang et al., 2015), Natural Questions (Kwiatkowski et al., 2019) and IIRC (Ferguson et al., 2020) are mainly based on common sense and context.

MRC Datasets in Academic Domain There are some MRC datasets targeting academic domains, where domain-specific knowledge is critical to tackling these problems. PubmedQA (Jin et al., 2019) is a biomedical MRC dataset, where the context is abstract, the question is the corresponding title and the answer can only be YES|NO|MAYBE. BioMRC (Pappas et al., 2020) focuses on cloze-

style MRC tasks in the biomedical domain. But these studies only conduct title and abstract as the context. As far as we know, there is only one study (Dasigi et al., 2021) focused on full-text scientific machine reading comprehension. QASPER (Dasigi et al., 2021) takes an entire paper as the context to do question answering, where the answer can be in various forms, including EXTRACTIVE, ABSTRACTIVE, YES|NO and UNANSWERABLE. However, QASPER (Dasigi et al., 2021) mainly takes the annotators as the single source of supervision, while multiple perspectives of annotations among different levels of researchers are needed in academic research works.

7. Conclusion

In this study, we introduced SciMRC, a novel multi-perspective scientific machine reading comprehension dataset. SciMRC incorporates diverse perspectives of readers, including beginners, students, and experts. Our extensive experimental results highlight the inherent relationships and differences among these perspectives, underscoring the importance of perspective analysis in scientific machine reading comprehension.

Ethical Considerations

In creating SciMRC, we utilized papers authored by other researchers. To respect copyright laws, we restricted ourselves to arXiv papers released under a CC-BY-* license. Furthermore, we ensured that the annotators were compensated well above the local minimum wage, and we took care to exclude any personal information from our dataset.

Limitations

Our current research on SciMRC is limited to three perspectives: beginner, student, and expert. However, in reality, the understanding of scientific papers can be further nuanced. For instance, even within the expert perspective, there can be distinctions between senior and junior experts. Future work may aim to refine and expand the range of perspectives considered in SciMRC.

Acknowledgements

The work is supported by National Key R&D Plan (No. 2020AAA0106600), MIIT Program (CEIEC-2022-ZM02-0247), National Natural Science Foundation of China (No.62172039, U21B2009 and 62276110).

Bibliographical References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv e-prints*, pages arXiv–2004.

Nicola Bianchi and Michela Giorcelli. 2019. Scientific education and innovation: From technical diplomas to university stem degrees. *ERN: Microeconometric Studies of Education Markets (Topic)*.

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

Elena Bruches, Olga Tikhobaeva, Yana Dementyeva, and Tatiana Batura. 2022. TERMinator: A system for scientific texts processing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3420–3426, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

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.

Sebastian de la Chica, Faisal Ahmad, James H. Martin, and Tamara R. Sumner. 2008. Pedagogically useful extractive summaries for science education. In *International Conference on Computational Linguistics*.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. 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): System Demonstrations*, pages 211–216, Hong Kong, China. Association for Computational Linguistics.

James Ferguson, Matt Gardner, Tushar Khot, and Pradeep Dasigi. 2020. Iirc: A dataset of incomplete information reading comprehension questions. In *Conference on Empirical Methods in Natural Language Processing*.

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.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

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.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Yi Yang, Wen tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Sonia Zulfiqar, Muhammad Farooq Wahab, Muhammad Ilyas Sarwar, and Ingo Lieberwirth. 2018. Is machine translation a reliable tool for reading german scientific databases and research articles? *Journal of chemical information and modeling*, 58 11:2214–2223.