MESAQA: A Dataset for Multi-Span Contextual and Evidence-Grounded Question Answering

Jui-I Wang[1], Hen-Hsen Huang[2], Hsin-Hsi Chen[1,3]
[1]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[2]Academia Sinica, Taiwan
[3]AI Research Center (AINTU), National Taiwan University, Taiwan
rywang@nlg.csie.ntu.edu.tw,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

We introduce MESAQA, a novel dataset focusing on multi-span contextual understanding question answering (QA). Unlike traditional single-span QA systems, questions in our dataset consider information from multiple spans within the context document. MESAQA supports evidence-grounded QA, demanding the model's capability of answer generation and multi-evidence identification. Our automated dataset creation method leverages the MASH-QA dataset and large language models (LLMs) to ensure that each Q/A pair requires considering all selected spans. Experimental results show that current models struggle with multi-span contextual QA, underscoring the need for new approaches. Our dataset sets a benchmark for this emerging QA paradigm, promoting research in complex information retrieval and synthesis.

1 Introduction

Question answering (QA) systems have made significant strides with the advent of large language models (LLMs) such as ChatGPT (OpenAI, 2023), excelling at extracting answers from single spans within context documents. However, real-world scenarios often require synthesizing information from multiple spans. On the other hand, evidence retrieval is crucial for providing transparent and verifiable answers (Zhou et al., 2023; Yao et al., 2023). Comparing with traditional QA models focusing on extracting relevant spans, the approach identifying and linking multiple pieces of evidence across a document enhances the accuracy, interpretability, and trustworthiness of the usage of LLM.

Existing datasets like SQuAD (Rajpurkar et al., 2018) have been instrumental in advancing QA capabilities, but they predominantly feature single-span questions, not fully representing the complexity of real-world information-seeking tasks.

To address this limitation, our MESAQA designed to challenge QA models with tasks that necessitate multi-span contextual understanding and evidence-grounded reasoning. Each instance is a triple (Q, A, C), where the question (Q) can only be accurately answered by considering multiple segments of the context (C). To guarantee the multi-span requirement, we propose a logically strict LLMs-based approach to create such a dataset. This requirement challenges models to perform sophisticated information retrieval and integration, providing a robust benchmark for developing and evaluating current QA capabilities in multi-span contextual and evidence-grounded QA.

MESAQA[1], Multi-Evidence-Span Abstractive QA, introduces tasks demanding deeper comprehension and complex reasoning, such as in fields like healthcare and law, where accurate and explainable answers are critical. Our work contributes significantly in three ways. First, we present a challenging dataset based on MASH-QA (Zhu et al., 2020), creating through a rigorous method to generate answers, which transcends traditional extractive QA tasks by requiring multi-span information synthesis within a document. Second, our dataset can be utilized in various QA tasks, including natural QA, extractive QA, and evidence-grounded QA, depending on the user's needs. This versatility makes our dataset a valuable resource for a wide range of research applications. Finally, experimental results reveal that evidence retrieval prior to answer generation enhances LLM performance while providing traceable response rationales, thereby augmenting interpretability and trustworthiness of QA models.

2 Related Work

Question Answering Datasets In the early development of QA datasets, research such as TriviaQA (Joshi et al., 2017) and SearchQA (Dunn et al., 2017), utilized a single span of text as the answer, limiting the applicability for tasks that re-

[1]https://github.com/reiiwang/MESAQA.git

Proceedings of the 31st International Conference on Computational Linguistics, pages 10891–10901
January 19–24, 2025.

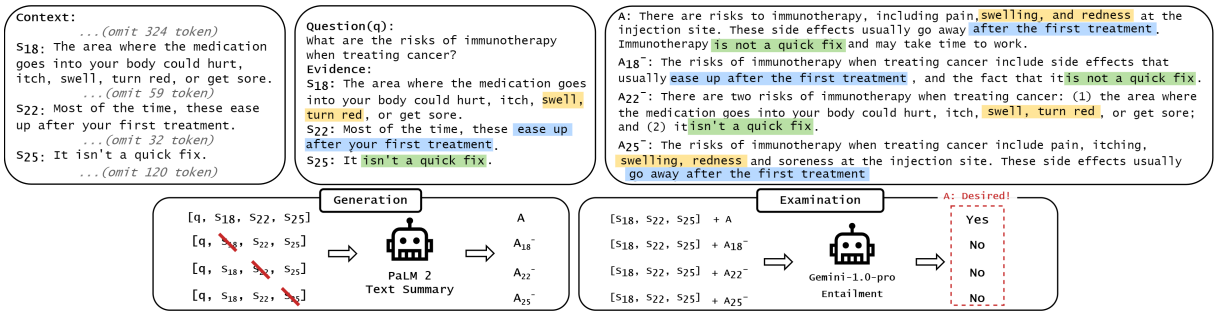

Figure 1: The process of data construction. Example with context length n=33, and evidence length m=3.

quire multiple pieces of information. To address this challenge, datasets like MS-MRC (Bajaj et al., 2018), MultiSpanQA (Li et al., 2022), and MA-MRC (Yue et al., 2023) were introduced, in which the answer contains multiple spans from the context to provide more comprehensive information. However, these datasets still rely on short spans, and the semantics of the answers are usually not continuous, making them less suitable for real-world scenarios. Recent researches on automated QA dataset generation address the resource-intensive nature of manual production. Notable frameworks include LIQUID (Lee et al., 2023), which iteratively filters incorrect answers based on confidence scores, operating at the entity level instead of the passage level; FABRICATOR (Golde et al., 2023), an open-source library for dataset generation; and SciQAG (Wan et al., 2024), which employs LLMs to generate and evaluate scientific QA pairs.

Evidence Retrieval Huo et al. (2023) integrated retrieval and verification processes, using LLMs to assess answer correctness based on retrieved content. Henning et al. (2023) introduced the WHERE dataset with annotated evidence, utilizing LLM for evidence retrieval and question answerability assessment, but this dataset is not publicly available. Our MESAQA dataset pushes the boundaries by requiring the synthesis of information from multiple spans, emphasizing evidence retrieval for accurate and verifiable answers.

3 Dataset Construction

To create the MESAQA dataset, we propose a novel approach by leveraging MASH-QA, a well-known extraction-based QA dataset, as the foundational material. In MASH-QA, each instance comprises a triple (Q, C, E), where Q is a question, $C = \{s_1, s_2, s_3 \ldots, s_n\}$ represents the context document containing n sentences, and E is a subset of C that forms the extractive-based answer to Q. For our abstractive-based QA dataset, we construct an abstraction answer A to Q for each instance in MASH-QA and consider E as the evidence extracted from C that supports the abstractive answer A.

As shown in Figure 1, for the question "What are the risks of immunotherapy when treating cancer?", the original answer in MASH-QA comprises the three sentences s_{18}, s_{22}, and s_{25} from C. In our dataset, we aim to generate an abstractive answer like "There are risks to immunotherapy, including pain, swelling, and redness at the injection site. These side effects usually go away after the first treatment. Immunotherapy is not a quick fix and may take time to work." Formally, the created answer A must satisfy two criteria:

- **Comprehensive Integration of Evidence**: A should integrate essential information from all three evidence sentences. Namely, the information from s_{18}, s_{22}, and s_{25} is indispensable to construct A to answer Q.
- **Alignment with Evidence Content**: Conversely, A must strictly adhere to the content provided in the evidence (E). Every aspect of A must find direct support within E, ensuring coherence and relevance.

To achieve this goal, we propose a novel dataset construction approach with the following steps:

1. **Preprocessing**: MASH-QA features answers ranging from consecutive to non-consecutive sentences and varies widely in context length and number of answer spans. To ensure multiple evidence spans, we exclude QA pairs with single evidence sentences ($|E| \geq 2$).
2. **Candidate Question and Answer Generation**: For each instance (Q, C, E) in MASH-QA, we utilize a large language model (LLM)

to generate an abstractive answer A to Q based on E. This can be denoted as:

$$A \leftarrow \text{LLM}_{QA}(Q, E)$$

3. **Validation for Multi-Span Necessity**: To ensure that the question Q requires the consideration of all spans in E for the correct answer A, we create a modified set of spans $E_i' = E \setminus \{s_i\}$ by excluding one span (sentence) s_i from E. We then ask the LLM to answer Q given E_i' instead of E:

$$A_i' \leftarrow \text{LLM}_{QA}(Q, E_i')$$

We observed that some MASH-QA instances exhibit high similarity within a set of extractive answers, which is not desirable for our multi-span dataset. Therefore, We calculated the similarity between A_i' and A using ROUGE score (Lin, 2004). If A_i' is found to be similar to A, take it as correct. The LLMs can still correctly answer to Q given E_i', i.e., $A_i' \approx A$, we discard this instance, as it indicates that Q can be answered without considering all the information in E. In other words, the information in s_i is redundant and can be inferred from E_i', the rest of evidence spans. This step ensures that the remaining instances in the dataset truly require multi-span information for answering Q. Formally, we keep the instance (Q, A, C) if and only if:

$$\forall s_i \in S,\ \text{LLM}_{QA}(Q, E \setminus \{s_i\}) \not\approx A$$

4. **Entailment Verification**: We examine if A is fully supported by E by identifying whether A entails each sentence in E. If A includes all information from every evidence span E without adding extraneous information, A maintains consistency with the original expert-curated answers. As textual entailment (TE) is a classical task of natural language inference (Williams et al., 2018), we instruct an LLM_{TE} to determine the entailment.

We instruct PaLM2 (Anil et al., 2023) as LLM_{QA} and Gemini-1.0-pro (Gemini Team, 2023) as LLM_{TE}. For semantic evaluation, we establish criteria to ensure that generated answers entail the original answer sentence spans by using Gemini-1.0-pro. It will judge either "yes" or "no" to indicate whether a PaLM2-generated answer is followed criteria. This rigorous validation ensures the quality of our dataset, setting a new standard for multi-span and evidence-grounded QA tasks.

		Consec.	Nonconsec.
Number of QA Pairs		4,479	1,704
Evidence	# sentences	3.21±1.15	3.73±1.70
	Δ sentences	2.21±1.15	16.03±10.42
	Δ tokens	20.81±21.44	277.46±209.30
Answer	# tokens	63.25±28.28	68.84±30.99

Table 1: Statistics of our MESAQA dataset.

Our dataset comprises 6.1k instances, each composed of quadruples (Q, C, A, E). The mean context length is 809 tokens (41 sentences) per context. Table 1 presents separate statistical analyses for these types of evidences. Δ in sentences and Δ in tokens denote the mean sentence and token count, respectively, between the first and last span of E. For nonconsecutive evidence, LLMs must extract evidence from a broad range of the context, increasing the complexity of the retrieval process. The distribution of evidence number is presented in Appendix A.

To assess the feasibility of our dataset creation, we conducted a human validation study involving three medical experts who analyzed 100 randomly selected instances, each comprising a triplet (Q, E, A) and four accompanying questions. Instructions with clearly defined options were provided to ensure consistency. The results showed high quality across multiple dimensions: 80% of instances cover all essential information in E and 17.5% covering most of E, with only negligible or overlapping spans excluded but judged as correct. On the other hand, 90% of instances demonstrated direct support of every aspect of A by E, with the remaining including minor additional information without altering the main concepts of E. Furthermore, 95% of instances are fluent, and 92% of instances are judged entirely correct. These findings affirm the effectiveness of our dataset creation methodology in producing comprehensive, accurate, and well-supported instances. A detailed description of the annotation procedure and results is available in Appendix B.

4 Experiments and Analysis

4.1 Task Design

Task 1: Evidence Retrieval. Mirroring the approach used in MASH-QA, where an attention-based model was employed, we instruct the LLMs to extract evidence (E), which is a collection of

Model	Avg. EM	Avg. Recall	Avg. F1
Mixtral 8x7B	14.38%	63.63%	57.89%
Llama3 70B	17.75%	64.12%	**61.16%**
GPT-4o-mini	12.92%	58.53%	57.14%
Gemini-1.0-pro	**19.18%**	60.87%	59.42%
Gemini-1.5-flash	16.42%	**66.23%**	60.33%

Table 2: Experimental Results of Task 1 (Evidence Retrieval)

sentences from the context (C) that addresses the question (Q). The process utilizes a template comprising C and Q. C is segmented into sentences, each assigned an index. LLMs are prompted to return a list of these indices.

Task 2: Evidence-Grounded Question Answering. Similar to conventional natural QAs, LLMs generate an answer text. The LLMs are provided with a question-context pair (Q, C) and instructed to extract pertinent evidence from C and generate a free-form abstractive answer (A). We explore if the sequence of evidence extraction and answer generation influences performance by experimenting with two prompting orders: Answer-first and Evidence-first.

4.2 Experimental Setup

To assess the performance of Mixtral-8x7B (Jiang et al., 2024), Gemini-1.0-pro, Gemini-1.5-Flash (Gemini Team, 2023), Llama3-70B (Meta, 2024), and GPT-4o-mini (OpenAI, 2023), we proceed with a zero-shot evaluation.

We sample 5 seeds of 1,000 data points each and average the results. For evidence retrieval, we report the average Exact Match (Avg. EM), average Recall score (Avg. Recall), and average F1-score (Avg. F1). For Task 2, we prompt Gemini-1.0-pro, Gemini-1.5-flash, and GPT-4o-mini to output a score evaluating the correctness and entailment of the answers, which judge the performance of abstractive answers (Zheng et al., 2023), we report the average score of LLM's output (LLM score). Model version and detailed prompts used in the experiments are presented in Appendix C.

4.3 Results

Table 2 presents the performance of various LLMs on Task 1. The random baseline exhibits a high Avg. Recall score of 48% and a poor Avg. F1 score of 13% due to the selection of numerous false spans. Gemini-1.0-pro achieved the highest Avg. EM score at 19.18%, Gemini-1.5-flash recorded the highest Avg. Recall at 66.23%, and Llama3 70B attained the highest Avg. F1 score at 61.16%.

For Task 2, the LLMs were instructed to provide both an answer and their supporting evidence spans in two distinct orders: **Answer-first**, where the model generates a free-form answer followed by supporting evidence, and **Evidence-first**, where the model selects relevant sentences as evidence before generating an answer based on this evidence set. The results are reported in Table 3.

In the Answer-first generation order, Llama3-70B achieved the highest scores in Avg. EM and Avg. F1, while the Avg. Recall was 0.2% lower than Mixtral-8x7B. For the abstractive answer, Mixtral-8x7B reached the highest score over all model's evaluation. Conversely, in the Evidence-first generation order, Mixtral-8x7B achieved the maximum Avg. Recall score of 64.79% and Llama3-70B reach maximum Avg. F1 score of 61.50%. Evidence-first consistently outperforms Answer-first across all models and metrics, with particularly significant improvements in the performance of abstractive answers. This highlights the importance of grounding responses in retrieved evidence.

Adhering to chain-of-thought reasoning (Wei et al., 2023), providing evidence first is a more rational approach for LLMs. This order ensures grounds answers in factual evidence, mitigating potential verification bias that could occur if answers are generated first and then supported with selectively chosen evidence. For users who seek natural answers with evidence, this order enhances credibility and trust. It allows users to independently assess the information quality, increasing the reliability and transparency of the LLM's output.

4.4 Discussion

We calculated the Spearman correlation coefficient to understand the relationship between evidence retrieval and answer quality. Spearman's rank correlation coefficient is a non-parametric measure of correlation appropriate for continuous variables, ranging from -1 for a perfect negative correlation to 1 for a perfect positive correlation, with 0 indicating no correlation. For evidence retrieval, we used Recall as a representative metric and calculated its correlation coefficient with the score of the abstractive answer. The average Spearman's Rho for Gemini-1.0-pro ranged from 0.52 to 0.55, while for Llama3-70B, it ranged from 0.31 to 0.39, and for Mixtral-8x7B, Gemini-1.5-flash, GPT-4o-mini, it ranged from 0.4 to 0.5. All p-values were below

Model	Generation Order	Avg. EM	Avg. Recall	Avg. F1	GPT-4o-mini score	Gemini-1.0-pro score	Gemini-1.5-flash score
Mixtral 8x7B	Answer first	13.08%	60.12%	55.92%	64.23%	51.78%	53.72%
	Evidence first	**13.18%**	**64.79%**	**57.35%**	**82.63%**	**62.91%**	**58.81%**
Llama3 70B	Answer first	16.68%	59.87%	59.87%	43.24%	42.56%	43.48%
	Evidence first	**18.16%**	**64.32%**	**61.50%**	**53.35%**	**47.07%**	**46.88%**
GPT-4o-mini	Answer first	13.00%	56.74%	56.42%	60.89%	50.09%	51.87%
	Evidence first	**13.76%**	**60.69%**	**57.53%**	**76.15%**	**58.83%**	**56.71%**
Gemini-1.0-pro	Answer first	14.94%	54.22%	56.16%	47.12%	48.81%	47.66%
	Evidence first	**18.23%**	**63.01%**	**59.82%**	**79.42%**	**61.49%**	**58.64%**
Gemini-1.5-flash	Answer first	16.43%	58.50%	58.68%	51.99%	44.67%	45.77%
	Evidence first	**18.22%**	**63.01%**	**59.82%**	**81.63%**	**61.49%**	**59.12%**

Table 3: Experimental results of Task 2 (evidence-grounded QA). In the "Generation Order" column, "Answer first" means the LLMs first generate an answer and then find supporting evidence. "Evidence first" means the LLMs first find relevant evidence sentences and then generate answers based on those evidences.

0.001, indicating a statistically significant correlation between evidence retrieval and abstractive answer quality.

To investigate potential biases or impacts arising between evaluatees and evaluators, we analyzed the rankings of evaluatees under various LLM evaluation metrics using the "Evidence First" setting. Rankings derived from the Gemini-1.0-pro score and GPT-4o-mini score exhibited a similar order: Mixtral-8x7B (highest), followed by Gemini-1.5-flash, Gemini-1.0-pro, GPT-4o-mini, and Llama3-70B (lowest). A slight deviation in ranking was observed with the Gemini-1.5-flash score, where the top three evaluatees were ranked as follows: Gemini-1.5-flash (highest), Mixtral-8x7B, and Gemini-1.0-pro. This minor inconsistency may be attributed to the comparable performance of these evaluatees under different LLM evaluators. Despite this variation, the overall ranking trends remained consistent, suggesting that utilizing LLMs as evaluators offers meaningful reference value. We also infer that the results were not significantly influenced by model self-preference.

5 Conclusions

This work presents a novel dataset, MESAQA, which not only challenges existing QA models but also opens new avenues for research in multi-span reasoning and evidence-grounded QA. By fostering the development of models that can integrate and reason over multiple pieces of evidence, we move closer to creating AI systems that are both accurate and trustworthy in their responses. Future research can leverage advanced LLMs to further enhance evidence retrieval and answer generation capabilities, thereby improving the reliability and effectiveness of QA systems.

6 Limitations

This paper presents a novel dataset constructed using a rigorous methodology. While the current work provides a solid foundation, we acknowledge several directions for future improvement. The proposed dataset, MESAQA, leverages MASH-QA as a raw material to expand the application of the question-answering dataset. However, the main field of MASH-QA is healthcare and medical-related. Despite this domain-specific focus, the method presented in MESAQA has the potential for broader applications. The main concept is ensuring that questions can only be correctly answered when considering multiple evidence spans, which can be generalizable beyond the healthcare domain. While we trust the expert curation process mentioned in MASH-QA, we acknowledge that some questions may raise concerns about the breadth or inclusiveness of the provided evidence inherited from MASH-QA. Furthermore, manually verifying randomly selected samples suggests that hallucinations and the inner knowledge of a large language model's internal knowledge base influence only a relatively small percentage of outputs. However, it is not completely guaranteed that the LLM-generated answer will be flawless.

Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grants NSTC 112-2634-F-002-005-, and Ministry of Education (MOE), Taiwan, under grants NTU-113L900901.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *Preprint*, arXiv:1704.05179.

Google Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805.

Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. Fabricator: An open source toolkit for generating labeled training data with teacher LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–11, Singapore. Association for Computational Linguistics.

Sophie Henning, Talita Anthonio, Wei Zhou, Heike Adel, Mohsen Mesgar, and Annemarie Friedrich. 2023. Is the answer in the text? challenging ChatGPT with evidence retrieval from instructive text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14229–14241, Singapore. Association for Computational Linguistics.

Siqing Huo, Negar Arabzadeh, and Charles L. A. Clarke. 2023. Retrieving supporting evidence for llms generated answers. *Preprint*, arXiv:2306.13781.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

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

Seongyun Lee, Hyunjae Kim, and Jaewoo Kang. 2023. Liquid: A framework for list question answering dataset generation. *Preprint*, arXiv:2302.01691.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, United States. Association for Computational Linguistics.

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

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *ArXiv*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *Preprint*, arXiv:2405.09939.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Feng Yao, Jingyuan Zhang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Yun Liu, and Weixing Shen. 2023. Unsupervised legal evidence retrieval via contrastive learning with approximate aggregated positive. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4783–4791.

Zhiang Yue, Jingping Liu, Cong Zhang, Chao Wang, Haiyun Jiang, Yue Zhang, Xianyang Tian, Zhedong Cen, Yanghua Xiao, and Tong Ruan. 2023. Mamrc: A multi-answer machine reading comprehension dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2144–2148, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In

Advances in Neural Information Processing Systems, volume 36, pages 46595–46623.

Yuxuan Zhou, Ziyu Jin, Meiwei Li, Miao Li, Xien Liu, Xinxin You, and Ji Wu. 2023. THiFLY research at SemEval-2023 task 7: A multi-granularity system for CTR-based textual entailment and evidence retrieval. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1681–1690, Toronto, Canada. Association for Computational Linguistics.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Distribution of Evidence Number

Table 4 shows the distribution of evidence spans across instances, with most instances having 2 or 3 spans (1,944 and 1,945, respectively). The number of instances decreases as spans increase, with 471 instances having 6 or more spans.

# span	2	3	4	5	6+
# instance	1,944	1,945	1,164	659	471

Table 4: Distribution of evidence span.

B Dataset Quality Annotation

B.1 Recruitment of Human Annotation

Given that most MASH-QA instances are healthcare-related, we recruited three annotators who are medical experts. One of these experts has completed post-graduate year training.

B.2 Instructions Provided

Before the annotation, each expert is provided with the guidelines below.

- Your task is to review samples from this dataset. We consider the Evidence as the ground truth for the Question, and the part to be evaluated is the Answer. For each sample, we will investigate four aspects.
- Each question has four options. Please click on "Option definition" below the options to view the definition of each option, and choose the most appropriate option after reading.
- The sample description does not include the context. If you want to know more about the context, you can click on "source context" below.

During the annotation, they would need to answer the four aspect questions below.

1. Question 1: Does the Answer integrate essential information from all evidence sentences? (namely, does the answer cover all important concepts in evidence?
2. Question 2: Does the Answer contain only information mentioned in the Evidence? (In other words, is there any information that goes beyond the Evidence?)
3. Question 3: Is the expression of the Answer fluent and coherent?
4. Question 4: Is the Answer correct?

Table 5 presents the average evaluation results of three annotators for 100 samples. Each column in the table represents the proportion of the 100 samples labeled as the corresponding option. We label every option definition below every instance, making sure they are clarified with the criteria of each option. The option definitions of each question are shown in Tables 6, 7, 8, and 9. Figure 2 shows the interface for Sample 27.

Option	Q1	Q2	Q3	Q4
1	79.80	89.90	95.40	92.80
2	19.30	9.20	3.60	5.00
3	0.90	0.90	1.00	2.20
4	0.00	0.00	0.00	0.00

Table 5: Summary of average annotation results across all questions (Q1–Q4) presented as percentages (%).

C Experimental Configurations

The models used in this paper including Mixtral-8x7B (`mixtral-8x7b-32768`), Gemini-1.0-pro (`gemini-1.0-pro`), Gemini-1.5-Flash (`gemini-1.5-flash`), Llama3-70B (`llama3-70b-8192`), and GPT-4o-mini (`gpt-4o-mini-2024-07-18`). Prompt templates for experiments and LLM evaluations are shown in Figures 3, 4, and 5.

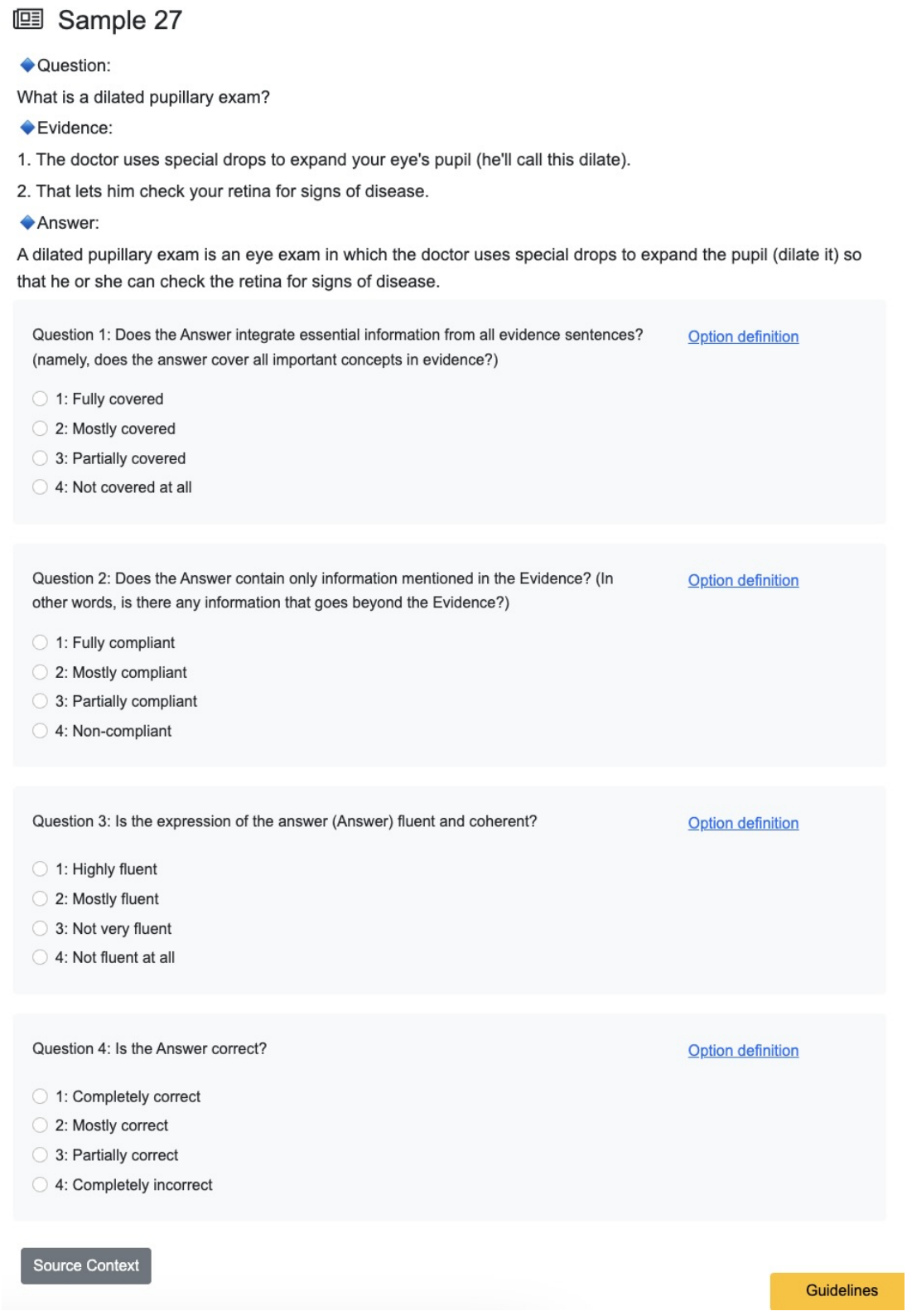

Figure 2: The above example shows Sample 27 of the annotation interface. Within this interface, annotators can interact with several key elements. They can click on option definitions to view the standards for each option. A "Source Context" button is available to access the contextual document. Additionally, annotators can click on "Guidelines" to review the annotation guidelines. These interactive elements are designed to assist annotators in their tasks.

Q1	**Does the Answer integrate essential information from all evidence sentences? (namely, does the answer cover all important concepts in evidence?)**
1 Fully covered	The answer's content encompasses every significant piece of information mentioned in the evidence. This information may be rewritten, condensed, or summarized in the answer, but the meaning remains unchanged.
2 Mostly covered	The answer's content encompasses the important information from most segments of the evidence, but some may be omitted. The omitted segments contain secondary or irrelevant information that can be disregarded.
3 Partially covered	Only a portion of the important information is covered in the answer. Most segments of the evidence are not mentioned in the answer, indicating that significant segments have been overlooked.
4 Not covered at all	The answer does not cover any of the important information mentioned in any sentence of the evidence.

Table 6: Question 1 and option definition.

Q2	**Does the Answer contain only information mentioned in the Evidence? (In other words, is there any information that goes beyond the Evidence?)**
1 Fully compliant	The answer strictly adheres to the information provided in the evidence, without any additional information beyond the evidence. In other words, all information mentioned in the answer can be found in the evidence. [Note: This includes answers that use different wording but are semantically equivalent, or phrases added to improve the fluency of the response, which are not considered as going beyond the evidence]
2 Mostly compliant	The answer largely follows the information in the evidence, but may include some supplementary information or explanations. These additions do not alter the main meaning of the evidence, but may provide some background or details, and these details are accurate.
3 Partially compliant	The answer is only partially based on the information from the evidence. A significant portion of the content consists of additional information or explanations added by the LLM. These additions may be extensions of the evidence or unrelated information, and the accuracy of these extensions cannot be guaranteed.
4 Non-compliant	The answer is not based on the information from the evidence at all. The content consists entirely of additional information, erroneous information, or content unrelated to the evidence.

Table 7: Question 2 and option definition.

Q3	**Is the expression of the answer (Answer) fluent and coherent?**
1 Highly fluent	The answer is expressed fluently and coherently.
2 Mostly fluent	The answer is generally fluent, but may contain a few grammatical errors or unnatural expressions, which do not affect overall comprehension.
3 Not very fluent	The answer is not very fluent, with several grammatical errors or unnatural expressions that affect comprehension.
4 Not fluent at all	The answer is not fluent, containing grammatical errors and unnatural expressions, making it difficult to understand.

Table 8: Question 3 and option definition.

Q4	**Is the Answer correct?**
1 Completely correct	Independent of other questions, assuming the Evidence is the Ground truth, the information in the answer is correct and consistent with the evidence or known facts. Or, based on your knowledge background, the information in the answer is correct.
2 Mostly correct	Independent of other questions, assuming the Evidence is the Ground truth, most of the information in the answer is correct, but there may be some minor errors or inaccuracies.
3 Partially correct	Independent of other questions, assuming the Evidence is the Ground truth, only a portion of the information in the answer is correct, with most of the content being incorrect.
4 Completely incorrect	Independent of other questions, assuming the Evidence is the Ground truth, all information in the answer is incorrect, with no part being correct.

Table 9: Question 4 and option definition.

You are a helpful assistant who are good at answering healthcare question.

I want you to search for the answer and its support evidences to the question below.
First I will give you, context, which is the sentences split from the context with the format of "index: sentence", followed by the question, then you need to reply me with two things:
1. the answer to the question
2. the index of evidence spans that can support your answer to the question.
(You have to find more than one evidence sentence.)

context: `{context}`
question: `{question}`

You should reply me with the following format:
Answer: <your answer>
Evidence sentences: [index1, index2, index3...]

Figure 3: Prompt template for tasking LLMs with generation prioritizes presenting the answer first (Answer first).

You are a helpful assistant who are good at answering healthcare question.

I will give you, context, which is the sentences split from the context with the format of "index: sentence", followed by the question, then you need to reply me with two things:
First, find the index of sentences that is the answer-related to the question, namely the evidence sentences. (You have to find more than one evidence sentence.)
Second, base on the evidence sentences you chose, give me the abstractive answer to the question.

context: `{context}`
question: `{question}`

You should reply me with the following format:
Evidence sentences: [index1,index2,index3...]
Answer: <your answer>

Figure 4: Prompt template for tasking LLMs with generation prioritizes presenting the evidence first (Evidence first).

Compare the following prediction to the real answer:

LLM's answer: `{LLM's answer}`
Real answer: `{real answer}`

Evaluation Instructions:
- Carefully read both the LLM's answer and the real answer.
- Assess the LLM's answer based on two primary criteria:
    1. Entailment: Does the LLM's answer logically imply or entail the information in the real answer?
    2. Correctness: Is the information in the LLM's answer factually correct when compared to the real answer?

Scoring guidelines:
- 0: The LLM's answer is empty, or contains information that directly contradicts the real answer.
- 1 - 25: The LLM's answer has major factual errors or fails to entail any significant part of the real answer.
- 25 - 50: The LLM's answer is partially correct and entails some aspects of the real answer, but has significant omissions or inaccuracies.
- 50- 75: The LLM's answer is mostly correct and entails a good portion of the real answer, with only minor errors or omissions.
- 75 - 99: The LLM's answer is highly accurate and entails nearly all of the real answer, with very minor imperfections.
- 100: The LLM's answer is entirely correct and fully entails the real answer.

Important considerations:
- Prioritize logical entailment: The LLM's answer should imply the information in the real answer, even if it's not stated in exactly the same way.
- A shorter LLM's answer can still score high if it correctly entails the key information from the real answer.
- However, if the LLM's answer is so brief that it fails to entail significant portions of the real answer, reduce the score accordingly.
- Any factually incorrect information should significantly lower the score, even if other parts are correct.
Reply with the following format:
Score: <score>

Figure 5: Template prompts for leveraging LLMs as judges.