SdQuAD: A Benchmark Question Answering Dataset for Low-resource Sindhi Language

Wazir Ali[†], **Muhammad Rafay**[‡], **Nadia Ali**[‡], **Amar Rehman**[‡]
[†]Department of Data Science
Quaid-e-Awam University of Engineering, Science & Technology, 67450 Nawabshah, Pakistan
aliwazirjam@gmail.com
[‡]Department of Artificial Intelligence
The Aror University of Art, Architecture, Design & Heritage, Rohri, 65170, Sukkur, Pakistan

Abstract

Question answering (QA) datasets are crucial for developing and evaluating monolingual and multilingual language models, yet low-resource languages like Sindhi lack open-source QA resources. We introduce SdQuAD, a novel open-source textual QA dataset for the low-resource Sindhi language, comprising more than 14K QA pairs curated and annotated by native speakers using the Label Studio. Sourced from diverse domains, including news, history, science, geography, business, and tourism, SdQuAD supports both extractive and abstractive QA tasks while capturing Sindhi's linguistic diversity. We assess annotation quality using span-level agreement and evaluate extractive performance with Exact Match (EM), F1 score, and a TF-IDF baseline. Additionally, we fine-tune mBERT, XLM-RoBERTa, and mT5 models on SdQuAD, benchmarking their performance to demonstrate the dataset's utility.

Keywords: Textual Question Answering, Sindhi Language, Extractive Methods

1. Introduction

Question answering is a fundamental task in natural language processing (NLP), enabling systems to provide precise and contextually relevant responses to user queries across diverse domains. The development of QA datasets, including SQuAD (Rajpurkar et al., 2016, 2018), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019), has been instrumental in advancing semantic parsing (Sarker et al., 2019), reading comprehension (Gómez-Adorno et al., 2013), and open-domain reasoning (Chen and Yih, 2020). These datasets have facilitated the creation of models capable of handling fact-based, multi-hop, and conversational queries. However, a significant gap remains for low-resource languages, as most QA resources are designed for high-resource languages, particularly English. This disparity limits access to NLP technologies for diverse linguistic communities, where the scarcity of data presents a major barrier to model development and evaluation.

To the best of our knowledge, Sindhi is represented only minimally in IndicQuest (Rohera et al., 2024), which contains just 200 QA pairs. Apart from IndicQuest, there are no publicly available monolingual or multilingual QA datasets for Sindhi. To address this gap, we introduce SdQuAD, a novel textual QA dataset for the Sindhi language, classified as low-resource. SdQuAD is a general, multi-domain dataset comprising over 14K QA pairs collected from diverse sources, including business-related content, general science articles, textbooks, news articles, and materials related to geography, tourism, and history. Unlike domain-specific datasets, SdQuAD covers multiple domains to capture the linguistic and cultural richness of Sindhi, enabling models to generalize across various topics and question types. Each question is paired with a context passage and a human-verified answer, ensuring high-quality annotations suitable for both extractive and abstractive QA tasks and reflecting real-world query patterns in a low-resource setting.

Accurate benchmarking requires the use of well-established and flexible evaluation metrics for QA systems. For extractive QA, where responses are short and span-based, traditional metrics such as Exact Match (EM) and token-level F1 score remain standard (Wang et al., 2024). In addition, lexical similarity measures such as TF-IDF provide further evaluation based on contextual representations (Dai et al., 2024). More recent approaches leverage encoder-only models, including Multilingual BERT (mBERT) and XLM-RoBERTa, as well as encoder–decoder models such as mT5 (Yue et al., 2025; Arif et al., 2024), which demonstrate stronger alignment with human judgments in both extractive and generative QA tasks.

This paper presents the creation and evaluation benchmarks of SdQuAD, highlighting its value as a resource for the NLP community. By spanning multiple domains, the dataset enriches QA resources for Sindhi and contributes to the broader goal of promoting linguistic diversity. The rest of the article is organized as follows: related work is presented in Section 2; the methodology for creating SdQuAD is outlined in Section 3; Section 4 presents the results, provides analysis, and discusses the dataset's potential to bridge the gap in low-resource language processing; finally, Section 5 concludes the paper.

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 55–61 May 11, 2026.

2. Related Work

This section presents existing work on QA datasets, primarily organized by their main focus and methodology, along with their creation processes, scale, and key characteristics.

Early QA datasets often leveraged structured knowledge bases such as Freebase to generate or collect questions, emphasizing semantic parsing and fact retrieval. WebQuestions (Berant et al., 2013) consists of 5.8K questions collected via the Google Suggest API, with answers sourced from Freebase. The dataset reflects real-world user queries and requires models to map natural language to knowledge base entities and relations; it has been widely used for benchmarking semantic parsing systems. Similarly, SimpleQuestions (Bordes et al., 2015) comprises 108K questions generated from Freebase triples, where each question corresponds to a single subject–relation–object fact, emphasizing a simple factoid QA approach. The 300K fact-based QA dataset further scales this approach by generating 300K questions from Freebase triples using recurrent neural networks (Serban et al., 2016). This large-scale, automatically generated corpus highlights the potential of neural methods for creating diverse QA pairs; however, it relies on synthetic questions tied to knowledge base facts. The GraphQuestions dataset (Su et al., 2016) introduces 5.1K questions generated from Freebase subgraphs using a semi-automated approach, incorporating characteristics such as structural complexity and paraphrasing.

The LC-QuAD dataset (Trivedi et al., 2017) provides 5K questions derived from SPARQL queries over DBpedia, offering syntactic variation and supporting multi-hop reasoning. This work was extended in ComplexWebQuestions (Talmor and Berant, 2018), which contains 34.6K questions created semi-automatically from SPARQL queries, where multi-hop reasoning is evaluated by combining sub-questions. FreebaseQA (Jiang et al., 2019) collects more than 28.3K questions from various websites such as TriviaQA, which are then matched to Freebase triples and verified by annotators. It reflects open-domain, knowledge-based QA with linguistically diverse, human-composed questions. CFQ (Keysers et al., 2020) is a benchmark dataset designed to evaluate a model's ability to handle unseen compositional structures; it includes more than 239K automatically generated questions derived from Freebase.

In the context of low-resource languages, several QA datasets have been developed to address the scarcity of annotated resources. The UQA dataset (Arif et al., 2024) was recently introduced for QA tasks in the low-resource Urdu language. Another dataset, UQuAD (Kazi and Khoja, 2024), is a large-scale Urdu QA dataset designed for extractive reading comprehension. In the Sindhi language, although there is growing interest in NLP, most existing resources focus on part-of-speech tagging (Ali et al., 2021b), named entity recognition (Jumani et al., 2018; Ali et al., 2020), and sentiment analysis (Barakzai et al., 2022; Ali et al., 2021a). To the best of our knowledge, IndicQuest (Rohera et al., 2024) is currently the only dataset that includes Sindhi, consisting of 200 question–answer pairs, with only a small portion in the language. In contrast, the proposed dataset will be publicly available and specifically developed for the Persian–Arabic Sindhi QA task.

3. Methodology

The development of the proposed SdQuAD[1] dataset for the Sindhi language involved a multi-stage process, including data collection, annotation using the Label Studio platform, quality assessment, baseline evaluation, and fine-tuning of encoder-only and encoder–decoder models. This methodology ensures the reliability and diversity of the dataset. In this section, we discuss each step involved in the construction of the SdQuAD dataset.

3.1. Data Collection

Sindhi is a low-resource language, and sufficient textual data for QA tasks is not readily available online. To address this limitation, we collected data through web scraping, gathering textual content from a variety of sources. The dataset spans multiple domains, including news from the Awami Awaz[2] Sindhi newspaper, historical content from books[3]; and Science[4] related material, geography[5], business news from the Associated Press of Pakistan[6], and tourism-related stories and books[7]. This diversity was included to capture a broader range of linguistic styles and topics.

- **News:** We collected the news data from a couple of newspapers, primarily from the Awami-Awaz newspaper, as well as from other

[1]The SdQuAD dataset is available at https://huggingface.co/datasets/Aliwj/SdQuAD

[2]https://awamiawaz.pk/category/national

[3]https://books.sindhsalamat.com

[4]https://lib.sindh.org/kitaab/detail/general-science-vol-2

[5]https://lib.sindh.org/kitaab

[6]https://sindhi.app.com.pk/sindhi/category/international/page/521/

[7]https://lib.sindh.org/kitaab/detail/around-the-world

Question (Sindhi & English)	Context (Sindhi & English)	Answer (Sindhi &English)
سنڌ جو راڄڌاني ڪهڙو آهي؟ What is the capital of Sindh?	سنڌ پاڪستان جو هڪ صوبو آهي جنهن جو راڄڌاني ڪراچي آهي. Sindh is a province of Pakistan whose capital is Karachi.	ڪراچي Karachi
موهن جو درڙو ڪهڑي تهذيب سان تعلق رکي ٿو؟ Which civilization is Mohenjo-Daro associated with?	موهن جو درڙو وادي سنڌ جي تهذيب جو هڪ قديم شهر آهي جيڪو 5000 سال پراڻو آهي؟ Mohenjo-Daro is an ancient city of the Indus Valley Civilization, dating back 5,000 years.	وادي سنڌ جي تهذيب Indus Valley Civilization
سنڌ ۾ سياحت لاءِ مشهور جاءِ ڪهڑي آهي؟ What is a famous tourist spot in Sindh?	گورکهه هيل اسٽيشن سنڌ ۾ سياحت لاءِ مشهور آهي جتي خوبصورت منظر آهن. Gorakh Hill Station is famous for tourism in Sindh, offering beautiful views.	گورکهه هيل اسٽيشن Gorakh Hill Station

Table 1: An example of the SdQuAD dataset with English Translation. Sindhi sentences/words in the Persio-Arabic script with their corresponding English translations. In this article, the term Sindhi refers specifically to the Persio-Arabic script, which is the most widely used and for which the majority of resources are available. However, Sindhi is also written in other scripts, including Devanagari and Roman.

sources covering current events, politics, and society.

- **Business:** The business- and commerce-related content was scraped from business-oriented websites and forums, consisting of questions related to economic affairs and market trends.
- **General Science:** Science-related books and guides are readily available with question–answer pairs, which were collected by the annotators. The data were then transformed into the required format. The content was extracted from Sindhi textbooks, including higher and secondary-level science curricula, covering subjects such as physics, biology, and chemistry.
- **Geography:** We collected descriptive texts and questions from geography books and various educational resources covering demographics and regional studies. Afterwards, the annotators created the question–answer pairs.
- **Tourism:** The text was scraped from multiple web resources, including official websites of the Sindh Tourism Department and travel blogs. After collecting tourism-related text, the annotators created the question–answer pairs.
- **History:** Historical narratives were taken from books available on the Sindh Salamat website. These books are publicly available without any copyright restrictions. After extracting the text from the books, question–answer pairs were created by the annotators using Label Studio.

In total, this process produced more than 20,000 raw documents. We then applied filtering and deduplication to improve data quality, and removed out-of-vocabulary content, such as English words, to maintain relevance and consistency in Sindhi. The text was further normalized for Sindhi language standards, and the cleaned data was stored in JSON format for the annotation stage.

Domain	Question-Answer pairs
Business	1695
Science	5080
News articles	2132
Geography	2153
Tourism	1886
History	1619
Total	**14565**

Table 2: The distribution of QA pairs across different domains in the SdQuAD dataset

3.2. Annotation

The crawled raw text was converted into question–answer pairs through a structured manual annotation process in Label Studio[8]. Three native Sindhi speakers, with backgrounds in annotation, linguistics, and relevant domains, carried out this task by selecting context passages—typically 50 to 150 tokens long—and creating questions directly from them. They then annotated the corresponding answers by marking exact spans for extrac-

[8] https://labelstud.io/templates/question_answering

tive tasks or writing paraphrased responses for abstractive ones. The distribution of QA pairs across various domains is shown in Table 2. This native-driven use of Label Studio ensured both linguistic accuracy and contextual relevance in the resulting SdQuAD dataset. An example of the dataset structure, including the question, context, answer, and English translation, is provided in Table 3.

3.3. Quality Assessment

We evaluated inter-annotator agreement using span-level metrics to ensure the annotation reliability of SdQuAD dataset. We use Exact Match (EM) and F1-score, defined as follows:

$$EM = \begin{cases} 1, & \text{if predicted span} = \text{gold span}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The EM metric measures whether the predicted span exactly matches the gold span, while the F1-score balances precision and recall to account for partial overlaps:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (2)$$

where precision is the fraction of predicted tokens that appear in the gold answer, and recall is the fraction of gold tokens that appear in the predicted answer.

3.4. Baseline Evaluation

We established a baseline for extractive QA using Term Frequency–Inverse Document Frequency (TF-IDF), which ranks candidate spans based on their similarity to the question. The TF-IDF weights terms as follows:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{|D|}{|\{d' \in D : t \in d'\}| + 1}\right). \quad (3)$$

here, t denotes a term, d represents a document, and D refers to the SdQuAD dataset. We vectorized the input using the scikit-learn library and used a custom Sindhi tokenizer.

3.5. Model Fine-Tuning and Evaluation

To evaluate the proposed SdQuAD dataset for training robust QA models, we fine-tuned multilingual architectures, including mBERT (Pires et al., 2019), mT5 (Xue et al., 2021), and XLM-RoBERTa (Conneau et al., 2020), using the Hugging Face Transformers library on a Google Colab T4 GPU. The training configuration employed a learning rate of $2e^{-4}$ for all models, a batch size of 16, and five training epochs.

We do not train a new tokenizer for BERT (mBERT-base), XLM-RoBERTa (XLM-RoBERTa-base), or mT5 (mT5-base). Instead, we fine-tune these models on our SdQuAD dataset. The tokenizers are pretrained on more than 100 languages, including Sindhi script patterns; therefore, these models already provide effective subword segmentation for Sindhi text.

4. Results and Analysis

The Table 4 shows the span-level inter-annotator agreement for the SdQuAD dataset. In order to assess the consistency in annotation, we report average pairwise-agreement metrics across annotator pairs. The Average F1-overlap of 43.49 represents the mean pairwise token-level overlap between annotated answer spans. This metric captures partial agreement by measuring how much the selected spans overlap, making it suitable for extractive QA tasks where boundaries may slightly differ. The EM score of 39.20 reflects strict span-level agreement, counting only cases where annotators selected identical start and end positions for the answer span. The Precision of 46.35 and Recall 38.52 are averaged pairwise span-level metrics, measuring how accurately one annotator's selected span matches of another's. The corresponding F1-score of 42.13 is the harmonic mean of precision and recall, summarizing overall agreement while balancing span over-selection and under-selection. The moderate agreement scores indicate reasonable alignment among annotators while reflecting the inherent difficulty of span selection in extractive QA tasks. Table 4 shows the train-test split for the baseline experiments .

Metric	Score
Average F1-overlap	43.49
Exact Match (EM)	39.20
Precision	46.35
Recall	38.52
F1-score	42.13

Table 3: Baseline retrieval and Exact Match scores.

Dataset Split	Size
Train set	12,052
Test set	3,013

Table 4: Train and test split for the baseline as well as multilingual transformer models.

Moreover, the TF-IDF-based baseline results shown in Table 4 provide a lexical retrieval bench-

mark for the SdQuAD dataset. This approach relies solely on term-frequency matching without any contextual understanding. The model achieves an average F1-overlap of 59.46, indicating a moderate token-level match between the predicted and gold answer spans. However, the Exact Match (EM) score of 46.29 is considerably lower, reflecting the difficulty in identifying precise span boundaries. This gap between F1-overlap and EM suggests that, although TF-IDF often retrieves text containing relevant keywords, it struggles to extract the exact answer spans accurately.

Metric	Score
Average F1-overlap	59.46
Exact Match	46.29

Table 5: TF-IDF retrieval baseline results on the proposed SdQuAD dataset.

Furthermore, Table 4 presents the performance of mBERT, XLM-RoBERTa, and mT5 on the SdQuAD dataset using the two primary evaluation metrics: Exact Match (EM) and F1-score. Among these models, mT5 achieves the highest performance, with an F1-score of 81.47 and an EM of 74.58. XLM-RoBERTa follows, with an F1-score of 79.28 and an EM of 68.52. In contrast, mBERT records lower scores, with an F1-score of 64.89 and an EM of 49.31, indicating comparatively weaker performance in identifying precise answer spans in Sindhi. The superior performance of mT5 can be attributed to its sequence-to-sequence architecture and large-scale multilingual pretraining, which enhance its ability to model contextual relationships. XLM-RoBERTa also produces competitive results due to its multilingual representations, although its lower EM suggests less precise span boundary detection compared to mT5. In summary, these results demonstrate that transformer-based models significantly outperform traditional baselines in both EM (exact span matching) and F1-score (partial span matching).

Model	F1-score	EM
mBERT	64.89	49.31
XLM-RoBERTa	79.28	68.52
mT5	**81.47**	**74.58**

Table 6: Performance comparison of encoder-only models (mBERT-base and XLM-RoBERTa-base) and the encoder–decoder model mT5-base on the SdQuAD dataset. Boldface values indicate the best performance, achieved by mT5.

5. Conclusion

In this article, we introduced SdQuAD, a benchmark dataset for Sindhi extractive question answering with high-quality span-level annotations. The evaluation results show that multilingual transformer models substantially outperform the TF-IDF baseline. mT5 achieves the best performance, yielding an F1-score of 81.47 and an EM of 74.58, followed by XLM-RoBERTa with an F1-score of 79.28 and an EM of 68.52, while mBERT obtains comparatively lower scores, with an F1-score of 64.89 and an EM of 49.31. The TF-IDF baseline, with an Average F1-overlap of 59.46 and an EM of 46.29, further confirms the limitations of lexical matching methods for precise span extraction. These findings establish SdQuAD as a reliable benchmark for Sindhi QA and a valuable resource for future research.

6. Ethical Statement

All data used to create QA pairs for developing SdQuAD was collected from publicly available sources, including news articles, textbooks, and other open Sindhi-language resources. The collection process followed ethical research practices, with no use of private, personal, or sensitive information at any stage. The dataset is composed entirely of publicly accessible and educational content, ensuring compliance with data protection and ethical standards.

7. Bibliographical References

Wazir Ali, Naveed Ali, Yong Dai, Jay Kumar, Saifullah Tumrani, and Zenglin Xu. 2021a. Creating and evaluating resources for sentiment analysis in the low-resource language: Sindhi. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 188–194, Online. Association for Computational Linguistics.

Wazir Ali, Junyu Lu, and Zenglin Xu. 2020. SiNER: A large dataset for Sindhi named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2953–2961, Marseille, France. European Language Resources Association.

Wazir Ali, Zenglin Xu, and Jay Kumar. 2021b. SiPOS: A benchmark dataset for Sindhi part-of-speech tagging. In *Proceedings of the Student Research Workshop Associated with RANLP*, pages 22–30, Online. INCOMA Ltd.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024. UQA: Corpus for Urdu question answering. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 17237–17244, Torino, Italia. ELRA and ICCL.

Fahama Barakzai, Sania Bhatti, and Salahuddin Saddar. 2022. Sentiment analysis of sindhi news articles using deep learning. In *Proceedings of the 17th International Conference on Computer Sciences and Information Technologies, CSIT, Lviv, Ukraine, November 10-12*, pages 26–31. IEEE.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, October 18-21, Grand Hyatt Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL , Online, July 5-10*, pages 8440–8451. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, and Biaoyan Fang. 2024. A critical look at meta-evaluating summarisation evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 14795–14808.

Helena Gómez-Adorno, David Pinto, and Darnes Vilariño Ayala. 2013. A question answering system for reading comprehension tests. In *Proceedings of the Pattern Recognition - 5th Mexican Conference, MCPR, Querétaro, Mexico, June 26-29.*, Lecture Notes in Computer Science, pages 354–363. Springer.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume-1 (Long and Short Papers)*, pages 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

A.K. Jumani, M.A. Memon, F.H. Khoso, A.A. Sanjrani, and S. Soomro. 2018. Named entity recognition system for Sindhi language. In *Proceedings of the Emerging Technologies in Computing*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, Cham.

Samreen Kazi and Shakeel Ahmed Khoja. 2024. Context-aware question answering in Urdu. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing, ICNLSP, Trento, Italy, October 19-20*, pages 233–242. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *Proceedings of the 8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30*. OpenReview.net.

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 Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, Texas, USA, November 1-4*, pages 2383–2392. The Association for Computational Linguistics.

Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. 2024. L3Cube-IndicQuest: A benchmark question answering dataset for evaluating knowledge of LLMs in Indic context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 982–988, Tokyo, Japan. Tokyo University of Foreign Studies.

Jaydeb Sarker, Mustain Billah, and Md. Al Mamun. 2019. Textual question answering for semantic parsing in natural language processing. In *Proceedings of the 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–5.

Iulian Vlad Serban, Alberto García-Durán, Çaglar Gülçehre, Sungjin Ahn, Sarath Chandar, Aaron C. Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, August 7-12, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for QA evaluation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, Volume-1 (Long Papers)*, pages 641–651. Association for Computational Linguistics.

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. In *Proceedings of the Semantic Web - ISWC - 16th International Semantic Web Conference, Vienna, Austria, October 21-25*, Lecture Notes in Computer Science, pages 210–218. Springer.

Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024. Revisiting automated evaluation for long-form table question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Miami, FL, USA, November 12-16*, pages 14696–14706. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Online, June 6-11*, pages 483–498. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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, Brussels, Belgium. Association for Computational Linguistics.

Tan Yue, Rui Mao, Xuzhao Shi, Shuo Zhan, Zuhao Yang, and Dongyan Zhao. 2025. QAEval: Mixture of evaluators for question-answering task evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, Vienna, Austria, July 27 - August 01*, pages 14717–14730. Association for Computational Linguistics.