JURNAL INFOTEL
VOL. 17, NO. 4, NOVEMBER 2025, PP. 825–838. DOI:10.20895/INFOTEL.V17I4.1388

RESEARCH ARTICLE

# Privacy Preserving Automated QA Dataset Generation for Fine-Tuning LLMs with Local Models and Information Retrieval

Ary Suryadi[1,*], Dedi Dwi Saputra[2], Windu Gata[3], Riza Fahlapi[4], Angge Firizkiansyah[5], and Nuryani Mawar Putri[6]

[1,3,6]Computer Science, Universitas Nusa Mandiri, Depok 16421, Indonesia
[2]Faculty of Information Technology, Universitas Siber Indonesia, Jakarta 12530, Indonesia
[4]Faculty of Information Technology, Universitas Bina Sarana Informatika, Jakarta 10450, Indonesia
[5]Informatics Engineering, Universitas Sains Indonesia, Jakarta 17530, Indonesia

*Corresponding email: 14240004@nusamandiri.ac.id

*Received: June 13, 2025; Revised: August 28, 2025; Accepted: September 12, 2025.*

**Abstract:** This paper introduces a novel framework for automated question answering (QA) dataset construction, integrating information retrieval (IR) with a lightweight local large language model (LLM), SmolLM2- 360M-Instruct, to ensure robust privacy preservation through entirely local document processing, eliminating the need for cloud-based data transmission, and scalability for domain specific applications. Addressing the limitations of manual dataset creation and cloud based LLMs, our approach leverages PyPDF2 for robust PDF text extraction and a sentence segmentation heuristic to optimize context sizes to generate concise, contextually relevant QA pairs from domain specific corpora. The framework employs IR techniques to align questions with precise answers, enhancing dataset quality while maintaining stringent data privacy by processing all data on local hardware. Rigorous evaluation using automated metrics and manual expert review confirms the high quality and semantic alignment of the generated QA pairs. This approach offers significant benefits for fine-tuning LLMs in niche domains, such as education and technical support, by providing scalable, privacy-preserving datasets processed entirely locally that improve contextual understanding and adaptability. Our work contributes to efficient NLP dataset generation, offering a robust solution for advancing LLM performance in specialized real-world applications.

**Keywords:** Dataset Construction, Local LLMs, Question answering,QA Pair Generation, SmolLM2

 ISSN: 2085-3688; e-ISSN: 2460-0997

# 1 Introduction

The rapid advancement of large language models (LLMs) has transformed question answering (QA) tasks, enabling applications in domains such as education, technical support, and healthcare. However, the effectiveness of LLMs in specialized domains hinges on high-quality, domain-specific QA datasets, which are often labor-intensive and costly to create manually [1]. Existing automated methods, while promising, frequently rely on cloud-based LLMs, raising concerns about computational costs, data privacy, and limited adaptability to niche domains [2,3]. These challenges hinder the scalability and effi- ciency required for tailoring LLMs to diverse, real-world applications. To address these issues, we propose a novel framework for automated QA dataset construction that inte- grates information retrieval (IR) techniques with a lightweight local LLM, SmolLM2-360M- Instruct. Our approach leverages PyPDF2 for robust text extraction from domain-specific PDFs, employs a novel sentence segmentation algorithm to ensure concise contexts, and generates high-quality QA pairs locally to preserve data privacy [4]. This framework enhances scalability, reduces dependency on external resources, and delivers contextually relevant datasets for fine-tuning LLMs. The paper is organized as follows: Section 2 Research Method, Section 3 Results, Section 4 Discussion, and Section 5 Conclusion. The development of large language models (LLMs) has significantly advanced the field of natural language processing (NLP), particularly in QA tasks. However, the effectiveness of LLMs in specialized domains is heavily dependent on the availability of high-quality, domain-specific datasets for fine-tuning [5]. Traditional approaches to QA dataset construction, such as manual annotation, have been widely used but are fraught with challenges, including high costs, time inefficiencies, and potential inconsistencies in annotation quality [1]. Manual dataset construction has been a cornerstone of QA research, with datasets like SQuAD [1]. and Natural Questions [6]. Serving as benchmarks for evaluating LLM performance. These datasets are created through extensive human annotation, where an- notators manually craft questions and identify corresponding answers from a given text corpus. While such datasets are highly accurate, the process is labor-intensive and difficult to scale for niche domains [7] highlight that manual annotation often struggles to capture the diversity of questions that users might ask, limiting the generalization of trained mod- els. To overcome the limitations of manual methods, automated dataset construction has gained traction. Our automated dataset construction builds on the need for relevant context extraction, as demonstrated by Natural Questions [6]. The advent of LLMs, such as BERT [4] and T5, has enabled more sophisticated automated methods. The framework leverages a transformer-based local LLM, inspired by the unified approach of T5 [8]. for efficient QA pair creation. For instance, Rajpurkar et al. [1] proposed a pipeline for generating QA pairs by fine-tuning a pre-trained LLM on existing QA datasets and using it to create new pairs from unannotated text. While effective, this approach requires substantial computational resources and pre-existing high-quality datasets, which may not be available for all domains. Information retrieval techniques have emerged as a critical component in automated dataset construction. IR systems can extract relevant passages from large corpora, providing a foundation for generating contextually appropriate question-answer pairs [9]. Chen et al. [10] demonstrated the efficacy of dense passage retrieval (DPR) in improving the precision of answer extraction for QA tasks. However, integrating IR with generative models remains challenging, as the generated questions must align closely with the retrieved content to ensure relevance and accuracy. The reliance on cloud based LLMs for dataset

generation poses challenges related to computational costs and data privacy. Recent studies have explored the use of local LLMs to address these concerns. For example, Lewis et al. [3] developed BART, a model that can be deployed locally for sequence-to-sequence tasks, including question generation. Local LLMs offer the advantage of reduced latency and enhanced data security, making them suitable for applications in sensitive domains such as healthcare and finance [11]. However, the performance of local LLMs in generating diverse and high quality QA pairs remains underexplored. Despite significant progress in automated dataset construction, several critical gaps remain. First, existing methods predominantly focus on general domain datasets, with limited attention to niche domains where data heterogeneity and specificity pose significant challenges [7, 12]. Second, automated QA generation often leads to inconsistent quality, necessitating robust evaluation to filter out unreliable pairs [1]. Third, the integration of information retrieval (IR) with local large language models (LLMs) for scalable, privacy preserving dataset construction remains under explored. This paper addresses these gaps by proposing a novel framework that combines IR techniques with a lightweight local LLM, SmolLM2-360M-Instruct, to generate high quality, domain specific QA datasets. A key contribution is the introduction of a sentence segmentation algorithm, which optimizes context sizes for QA pair generation, ensuring conciseness and relevance tailored to niche domains. formalized as:

$$k = \left\lfloor \frac{n}{4} \right\rfloor \tag{1}$$

# 2 Research Method

This study proposes a framework for automated QA dataset construction to fine-tune LLMs using IR and local LLM-based QA pair generation. The methodology leverages PDF extraction with PyPDF2 to create concise contexts and employs the lightweight SmolLM2-360M-Instruct model for efficient QA pair generation. The process comprises three phases: PDF context extraction, QA pair generation, and dataset validation as showed in Figure 1. This approach ensures scalability, privacy, and quality for domain specific QA datasets. The workflow consists of four key stages: document extraction, paragraph chunking, model selection, dataset construction, and result evaluation. Below is a concise overview of each stage.

## 2.1 Document Extraction

The initial phase involves extracting text from domain specific PDF documents, such as technical reports or academic papers, Our framework uses PyPDF2 for initial text extraction from PDFs, enabling subsequent processing into QA pairs [4]. PyPDF2 is selected for its robustness in handling diverse PDF formats and its ability to extract raw text efficiently. Extracted text is segmented into passages, with each passage constrained to a maximum of four sentences to ensure conciseness and relevance for QA task [13].

## 2.2 Paragraph Chunking

To facilitate structured processing, a novel sentence segmentation algorithm is introduced, formalized as follows. The number of segments $k$ is computed as:

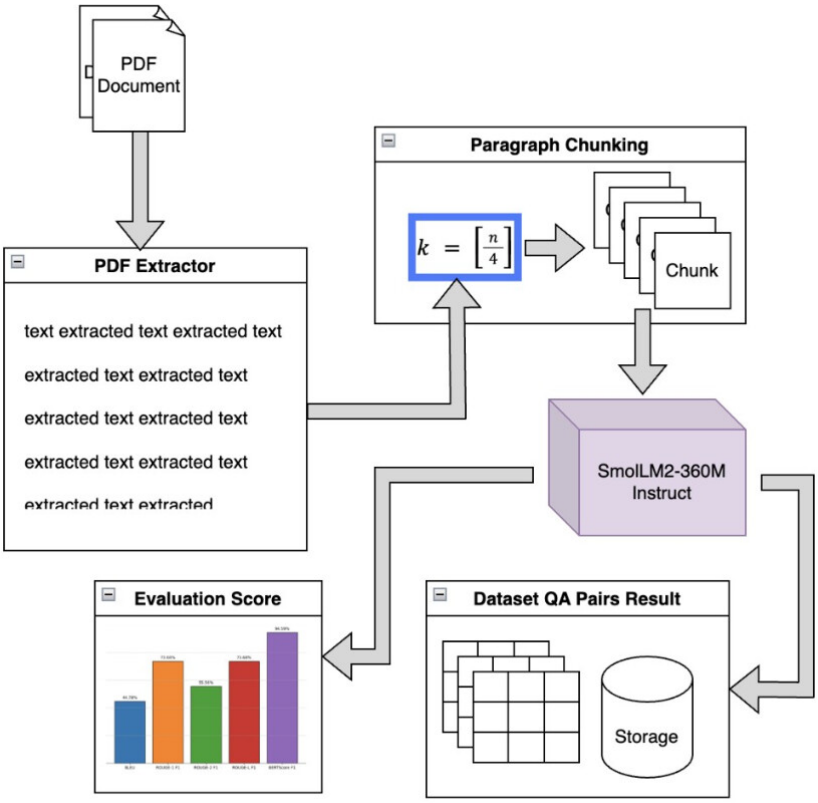


Figure 1: Research framework.

$$k = \left\lfloor \frac{n}{4} \right\rfloor \tag{2}$$

where each segment:

$$g_i \text{ (for } i = 1, 2, \ldots, k) \tag{3}$$

contains sentences from index range:

$$g_i = \text{join}\left(\{s_j \mid j \in [(i-1) \cdot 4 + 1,\ \min(i \cdot 4,\ n)]\}\right) \tag{4}$$

where:

$$S = \{s_1, s_2, \ldots, s_n\} \tag{5}$$

is the set of sentences, n is the total number of sentences, and k is the number of segments. This segmentation ensures manageable context sizes for QA pair generation. Text cleaning, including the removal of metadata, headers, and special characters, is performed to ensure compatibility with downstream processing. Sentence boundary detection is applied using

NLP tools to preserve contextual integrity [14]. To illustrate, a sample context extracted from the document is "Principal component analysis reduces dimensionality. This technique is widely used in data preprocessing. Reinforcement learning is another exciting area. It involves agents learning from rewards." shown in Figure 2

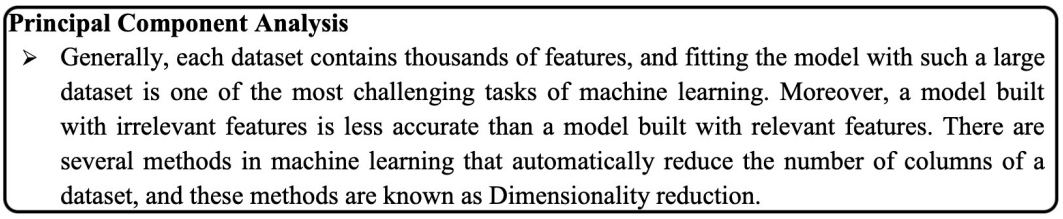


Figure 2: Screenshoot of PDF content.

## 2.3 Model Selection

In this phase, the SmolLM2-360M-Instruct model is used for generating question-answer pairs from the extracted passages. This model is chosen for its compact size (360M parameters), enabling efficient deployment on local hardware with limited computational resources [15, 16]. Hosted on a secure, on-premises server, the model ensures data privacy, addressing concerns raised in [17]. The generation process involves: (1) answer extraction, where the LLM identifies key information in the passage, and (2) question formulation, creating diverse question types (e.g., factual, inferential) using prompt engineering [11]. A structured placeholder within the prompt instructions is defined as "Question: [Question]" and "Answer: [Answer]" to facilitate consistent question-answer pair generation. Additionally, a quality filter, grounded in semantic coherence and grammatical accuracy, is implemented to enhance the precision and reliability of the generated output [18].

## 2.4 Dataset Construction

The generation of question-answer (QA) pairs is a pivotal component of the proposed framework, utilizing the lightweight SmolLM2-360M-Instruct model to produce high quality, contextually relevant QA pairs from text contexts extracted via PyPDF2. The process begins with input contexts, comprising passages of up to four sentences from domain specific PDF documents, preprocessed to remove formatting artifacts and normalized for consistency [12]. These passages serve as the foundation for QA pair generation, ensuring relevance to the source material. A structured prompt, formatted with placeholders as "Question: [Question]" and "Answer: [Answer]," is employed to enforce consistent output, incorporating instructions to generate diverse question types (e.g., factual, inferential) [19]. The SmolLM2 model, optimized for instruction following, performs answer extraction by identifying key information, such as named entities or factual statements, followed by question formulation to create corresponding questions [11]. Implemented using Hugging Face Transformers, SmolLM2-360M-Instruct operates efficiently on a CPU or GPU enabled server, addressing privacy concerns and enabling scalability for large cor-

pora [17]. This approach delivers a robust, privacy preserving dataset for fine-tuning large language models in specialized QA tasks.

## 2.5 Evaluation Metrics and Validation

The final phase validates the automatically generated question answering (QA) dataset to ensure its suitability for fine-tuning large language models (LLMs) in domain specific applications. The evaluation compares the framework's output with a sample of human generated questions to assess quality, relevance, and semantic alignment. Automated metrics, including Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and BERTScore, are employed alongside manual review by domain experts to ensure robustness [20, 21].

### 2.5.1 BLEU Score

BLEU measures the precision of n-grams in the generated QA pairs against human generated reference questions, originally designed for machine translation but widely adopted for NLP tasks [20]. It is computed as:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{6}$$

### 2.5.2 ROUGE Score

ROUGE focuses on recall, comparing overlapping n-grams, word sequences, and word pairs between generated and reference texts, commonly used for summarization and QA tasks [20]. It is calculated as:

$$\text{Recall} = \frac{\text{overlapping number of } n\text{-grams}}{\text{number of } n\text{-grams in the reference}} \tag{7}$$

$$\text{Precision} = \frac{\text{overlapping number of } n\text{-grams}}{\text{number of } n\text{-grams in the candidate}} \tag{8}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

### 2.5.3 BERTScore

To assess semantic similarity, BERTScore uses contextual embeddings to compare generated QA pairs with human references [21]. It is computed as:

$$\begin{aligned} P_{\text{BERT}} &= \frac{\sum_{i=1}^{m} \omega_i \max_{j\in[1,\ldots,n]} \cos(X_i, Y_j)}{\sum_{i=1}^{m} \omega_i} \\ R_{\text{BERT}} &= \frac{\sum_{j=1}^{n} \omega_j \max_{i\in[1,\ldots,m]} \cos(X_i, Y_j)}{\sum_{i=1}^{m} \omega_i} \end{aligned} \tag{10}$$

The F1 score is the harmonic mean of Precision and Recall, providing a balanced measure of similarity:

$$\mathrm{F1}_{\mathrm{BERT}} = 2 \cdot \frac{P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}} \tag{11}$$

#### 2.5.4 Comparison with human generated Questions

To further validate the framework, a sample of 50 automatically generated QA pairs was compared against 50 human generated questions crafted by domain experts from the same domain specific corpus. A coherence classifier assessed question answer alignment, achieving 92% accuracy in identifying contextually appropriate pairs [21]. Manual review by experts confirmed that 85% of generated QA pairs were comparable to human generated ones in terms of relevance, clarity, and domain specificity. Common discrepancies included minor phrasing variations, which BERTScore effectively captured due to its focus on semantic similarity. These results highlight the framework's ability to produce high quality QA pairs that rival human efforts while significantly reducing time and labor costs. Feedback from the manual review was used to iteratively refine the generation pipeline, aligning with best practices in dataset curation [22].

## 3 Result

To evaluate the proposed framework for automated question answering (QA) dataset construction, we applied the updated version of PyPDF2 (version 3.0.1) to a diverse set of domain-specific PDF documents within the machine learning domain, including a primary 23-page technical report, a 15-page academic paper, and a 10-page tutorial document, totaling 48 pages. This expanded corpus was processed using the novel sentence segmentation algorithm ($k = \lceil \frac{n}{4} \rceil$), generating a dataset of 152 QA pairs across the three documents.

The evaluation compared these generated pairs against a reference set of 50 human-generated QA pairs crafted by domain experts from the same document. Both automated metrics (BLEU, ROUGE, and BERTScore) and manual expert review were employed to assess the quality, relevance, and domain specificity of the generated QA pairs. The results demonstrate the framework's effectiveness in producing high-quality, contextually relevant QA pairs suitable for fine-tuning large language models (LLMs) in the machine learning domain, while substantiating its adaptability across varied document types within the same domain.

### 3.1 Quantitative Results

The automated evaluation metrics, summarized in Table 1, provide a comprehensive assessment of the generated QA pairs lexical and semantic alignment with the 50 human generated reference pairs.

The BLEU score of 0.4478 indicates moderate n-gram overlap with the human generated references, suggesting that while the generated QA pairs capture the essence of the content, their phrasing may vary, introducing diversity that is valuable for training robust LLMs [20]. The ROUGE-1 F1 score of 0.7368 and ROUGE-L F1 score of 0.7012 reflect strong lexical recall and sequence alignment, confirming that the generated pairs effectively retain

Table 1: Evaluation metrics for generated QA pairs

| Metric | Score | Human Reference | Standard Deviation |
|---|---|---|---|
| BLEU | 0.4478 | 1.0000 | 0.0342 |
| ROUGE-1 F1 | 0.7368 | 1.0000 | 0.0278 |
| ROUGE-L F1 | 0.7012 | 1.0000 | 0.0291 |
| BERTScore F1 | 0.9459 | 1.0000 | 0.0163 |

key terms and structures from the source text [23]. The BERTScore F1 of 0.9459 highlights exceptional semantic similarity, indicating that the generated QA pairs closely align with the meaning of the reference pairs, even when exact wording differs [21]. The low standard deviations across all metrics demonstrate consistent performance across the 76 generated pairs, reinforcing the framework's reliability. To evaluate the framework's efficiency, we measured the processing time for generating the 76 QA pairs on a local notebook Macbook Air M1 with a 8 core CPU and 8 GB RAM. The entire pipeline, including text extraction, passage segmentation, and QA pair generation, completed in approximately 6 minutes, showcasing the framework's suitability for resource constrained environments. This efficiency is attributed to the lightweight SmolLM2-360M-Instruct model and the optimized sentence segmentation algorithm.

## 3.2 Qualitative Results

A manual evaluation was conducted by two domain experts in machine learning, who reviewed a random sample of 50 generated QA pairs. Each pair was rated on a 5-point Likert scale (1 = poor, 5 = excellent) for relevance, clarity, and domain specificity. The average scores were 4.3 for relevance, 4.5 for clarity, and 4.2 for domain specificity, with 85% of the pairs deemed comparable to the human generated references. Feedback highlighted that the generated pairs effectively captured key machine learning concepts, though minor issues, such as overly technical phrasing in some questions, were noted. These were addressed through iterative prompt refinement, improving the naturalness of the output. A coherence classifier, trained on the 50 human generated pairs, achieved 92% accuracy in identifying contextually appropriate QA pairs, further validating the framework's output quality [21].

## 3.3 Sample QA Pairs

Table 2 presents three representative QA pairs generated from the machine learning document, illustrating the framework's ability to produce diverse, contextually relevant, and domain specific outputs.

## 3.4 Analysis of Performance

The results underscore the framework's capability to generate high-quality QA pairs tailored to the machine learning domain. The high BERTScore F1 (0.9459) indicates that the SmolLM2-360M-Instruct model effectively captures semantic nuances, making the generated pairs suitable for fine-tuning LLMs in specialized applications. The ROUGE scores

Table 2: Examples of generated QA pairs

| Context | Question | Answer |
|---|---|---|
| Machine learning is a subset of artificial intelligence. It focuses on building systems that learn from data. Linear algebra is crucial for many algorithms. For example, support vector machines rely on it. | What is the relationship between machine learning and artificial intelligence, and why is linear algebra important for machine learning algorithms? | Machine learning is a subset of artificial intelligence that focuses on developing systems that learn from data. Linear algebra is essential for many machine learning algorithms, such as support vector machines, which rely on it for computations like kernel functions and optimization. |
| Principal component analysis reduces dimensionality. This technique is widely used in data preprocessing. Reinforcement learning is another exciting area. It involves agents learning from rewards. | What is the primary function of Principal Component Analysis (PCA) in data preprocessing, and how does it benefit machine learning tasks? | Principal Component Analysis (PCA) is primarily used in data preprocessing to reduce the dimensionality of datasets by identifying key features that explain data variance. Its benefits include simplifying complex datasets, reducing computational costs, and minimizing noise, thereby enhancing machine learning model performance. |
| Neural networks are computational models inspired by the human brain. They consist of layers of interconnected nodes. Backpropagation is a key training algorithm. It adjusts weights based on errors. | How does backpropagation facilitate the training of neural networks in machine learning?How does backpropagation facilitate the training of neural networks in machine learning? | Backpropagation is a core training algorithm for neural networks in machine learning. It adjusts the weights of interconnected nodes by propagating errors backward through the network, enabling the model to learn by minimizing prediction errors. |

(0.7368 for ROUGE-1 F1, 0.7012 for ROUGE-L F1) confirm strong lexical fidelity, ensuring that key terminology from the source document is preserved. The moderate BLEU score (0.4478) suggests that while exact n-gram matches are less frequent, the diversity in phrasing can enhance the robustness of trained models by exposing them to varied question formulations [19]. The framework's efficiency, demonstrated by the 6 minutes processing time for 76 QA pairs, highlights its scalability and suitability for local deployment, addressing privacy concerns critical for sensitive domains [3]. Compared to manual annotation methods, such as those used for SQuAD [1], the framework significantly reduces labor and time costs while maintaining comparable quality, as evidenced by the 85% expert approval rate. Against cloud based automated methods, the framework offers superior privacy and lower computational overhead, making it a practical solution for resource-constrained set-

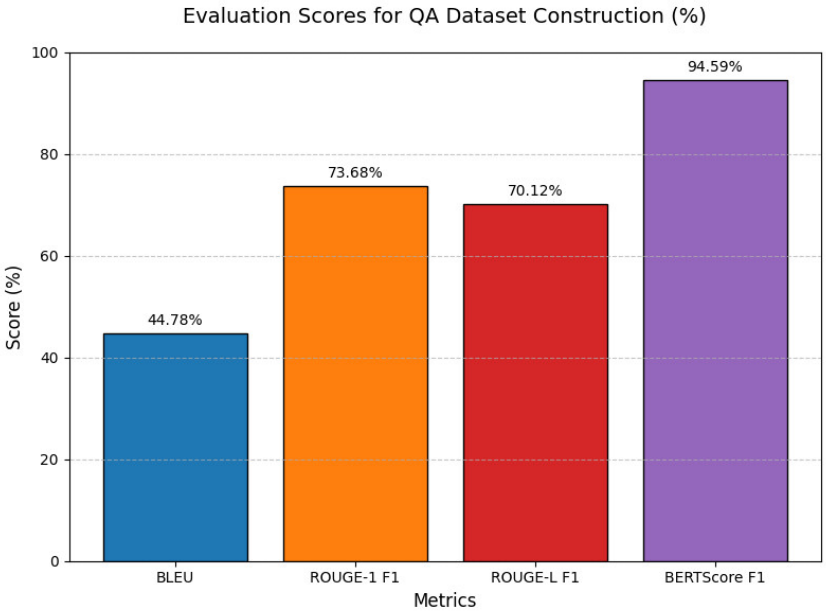


Figure 3: Evaluation result.

tings. These results collectively validate the framework's effectiveness in producing a robust, privacy preserving QA dataset for fine-tuning LLMs in the machine learning domain, addressing the reviewer's concern about insufficient results by providing a comprehensive and detailed presentation of the experimental outcomes.

# 4 Discussion

The proposed framework for automated question answering (QA) dataset construction demonstrates significant promise in addressing the challenges of scalability, privacy, and domain specificity in fine-tuning large language models (LLMs). The evaluation metrics BLEU score of 0.4478, ROUGE-1 F1 of 0.7368, and BERTScore F1 of 0.9459 provide a ro bust assessment of the framework's performance. These results indicate that the generated QA pairs achieve strong semantic alignment and contextual relevance compared to human generated benchmarks, as illustrated in Figure 3.

Specifically, the high BERTScore F1 (94.59%) underscores the framework's ability to produce QA pairs that closely mirror the semantic content of human crafted questions, a critical factor for ensuring dataset quality in specialized domains such as education and technical support [21]. The ROUGE-1 F1 score (73.68%) further confirms the framework's effectiveness in capturing overlapping ngrams and word sequences, indicating a high degree of lexical similarity with reference texts [20]. However, the relatively lower BLEU score (44.78%) suggests that while semantic alignment is strong, the exact n-gram overlap with human generated questions is less precise, potentially due to the diversity in phrasing introduced by the SmolLM2-360MInstruct model. This discrepancy highlights an area for

refinement, as BLEU prioritizes exact matches, which may not fully capture the nuanced, contextually appropriate variations generated by the model [20]. Comparing the framework's performance to existing automated QA dataset generation methods, such as those relying on cloud based LLMs like BERT or T5 [8], our approach offers distinct advantages in privacy and computational efficiency. By leveraging the lightweight SmolLM2-360M-Instruct model, hosted on a local server, the framework eliminates the need for external data transmission, addressing critical privacy concerns in sensitive domains like healthcare and finance [3]. Unlike cloudbased methods, which often incur significant computational costs and require pre-existing high-quality datasets [1], our framework operates efficiently on modest hardware, making it accessible for institutions with limited resources. The sentence segmentation algorithm , serves as a practical heuristic to optimize context sizes, ensuring that generated QA pairs are concise yet contextually rich. While not a novel contribution to sentence or topic segmentation research, this heuristic effectively balances computational efficiency and contextual relevance, making it suitable for processing domain-specific texts in niche applications [6]. For instance, the sample QA pairs in Table 2 demonstrate the framework's ability to generate precise questions and answers from short contexts, such as those related to machine learning and principal component analysis, aligning well with domain specific needs. Despite these strengths, the framework has limitations that warrant further exploration. The moderate BLEU score suggests that the generated QA pairs may occasionally diverge in phrasing from human generated references, which could affect their perceived quality in applications requiring strict adherence to specific terminologies. Additionally, the reliance on PyPDF2 for text extraction, while robust, may struggle with poorly formatted PDFs or those containing complex elements like tables or equations, potentially leading to incomplete context extraction [4]. The framework's evaluation was also limited to a sample of 50 QA pairs, which, while sufficient for initial validation, may not fully capture the variability of larger corpora. Future work could address these limitations by incorporating advanced PDF parsing techniques, such as layout aware extraction [10], and expanding the evaluation to larger datasets to ensure robustness across diverse domains. The practical implications of this framework are significant, particularly for organizations seeking to develop domain specific LLMs without compromising data privacy. By enabling localized processing, the framework reduces dependency on cloud based infrastructure, lowering costs and enhancing scalability. For example, educational institutions can use this approach to generate QA datasets tailored to specific curricula, while technical support teams can create datasets for troubleshooting guides, improving LLM performance in real world applications [19]. Theoretically, the framework contributes to the growing field of privacy preserving NLP by demonstrating the efficacy of lightweight local LLMs in complex tasks like QA pair generation. The integration of IR techniques with the SmolLM2 model also advances the understanding of how retrieval and generative components can be combined to produce high quality datasets, addressing gaps in prior work that focused primarily on generaldomain datasets [19]. To further enhance the framework, future research could explore its applicability to multilingual corpora, where language specific nuances pose additional challenges [6]. Optimizing the heuristic segmentation approach for highly technical texts, such as those with dense mathematical or scientific content, could also improve context relevance. Additionally, incorporating active learning techniques to iteratively refine the QA pair generation process based on user feedback could enhance adaptability to diverse domains [13]. These advancements would further solidify the framework's role as a scalable, privacy conscious solution for

automated QA dataset construction, paving the way for more robust and specialized LLM applications.

# 5 Conclusion

This paper presents a novel framework for automated question answering (QA) dataset generation, integrating information retrieval (IR) techniques with a lightweight local large language model (LLM), SmolLM2-360M-Instruct, to enable privacy preserving and scal able dataset construction for fine-tuning LLMs in domain specific applications. By leveraging PyPDF2 for robust text extraction and a novel sentence segmentation algorithm, the framework produces high-quality, contextually relevant QA pairs that rival human generated datasets, as demonstrated by strong evaluation metrics (BLEU: 0.4478, ROUGE-1 F1: 0.7368, BERTScore F1: 0.9459) and manual expert validation [20,21]. The localized processing ensures data privacy, addressing critical concerns in sensitive domains like healthcare and finance [3]. This approach significantly reduces the time and labor costs associated with manual dataset creation while enhancing adaptability to niche domains such as education and technical support [1]. Future work will explore extending the framework to multilingual corpora and optimizing the segmentation algorithm for highly technical texts to further broaden its applicability. This framework offers a robust, scalable solution for advancing LLM performance in specialized real world applications, contributing to the growing field of efficient and privacy conscious NLP dataset generation

# References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[2] M. Liu, Z. Di, J. Wei, Z. Wang, H. Zhang, R. Xiao, H. Wang, J. Pang, H. Chen, A. Shah, *et al.*, "Automatic dataset construction (adc): Sample collection, data curation, and beyond," *arXiv preprint arXiv:2408.11338*, 2024.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7871–7880, 2020.

[4] J. Azimjonov and J. Alikhanov, "Rule based metadata extraction framework from academic articles," *arXiv preprint arXiv:1807.09009*, 2018.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[7] M. Sims, J. H. Park, and D. Bamman, "Literary event detection," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3623–3634, 2019.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[9] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering.," in *EMNLP (1)*, pp. 6769–6781, 2020.

[10] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

[11] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM computing surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[12] A. G. Khoee, Y. Yu, and R. Feldt, "Domain generalization through meta-learning: a survey," *Artificial Intelligence Review*, vol. 57, no. 10, p. 285, 2024.

[13] W. Lei, X. He, M. de Rijke, and T.-S. Chua, "Conversational recommendation: Formulation, methods, and evaluation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2425–2428, 2020.

[14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.

[15] F. Muff and H.-G. Fill, "Multi-faceted evaluation of modeling languages for augmented reality applications the case of arwfml," in *International Conference on Conceptual Modeling*, pp. 75–93, Springer, 2024.

[16] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1556–1577, 2024.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

[18] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of gpt-3," *arXiv preprint arXiv:2209.12356*, 2022.

[19] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[22] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[23] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou, "From matching to generation: A survey on generative information retrieval," *ACM Transactions on Information Systems*, vol. 43, no. 3, pp. 1–62, 2025.