LLMQuoter: Enhancing RAG Capabilities Through Efficient Quote Extraction from Large Contexts

Yuri Façanha Bezerra[a] and Li Weigang[b]
TransLab, Department of Computer Science, University of Brasilia, Brasilia, Federal District, Brazil

Keywords: Knowledge Distillation, Large Language Models, LLM Reasoning, Low-Rank Adaptation, Retrieval-Augmented Generation.

Abstract: We introduce LLMQuoter, a lightweight, distillation-based model designed to enhance Retrieval-Augmented Generation (RAG) by extracting the most relevant textual evidence for downstream reasoning tasks. Built on the LLaMA-3B architecture and fine-tuned with Low-Rank Adaptation (LoRA) on a 15,000-sample subset of HotpotQA, LLMQuoter adopts a "quote-first-then-answer" strategy, efficiently identifying key quotes before passing curated snippets to reasoning models. This workflow reduces cognitive overhead and outperforms full-context approaches like Retrieval-Augmented Fine-Tuning (RAFT), achieving over 20-point accuracy gains across both small and large language models. By leveraging knowledge distillation from a high-performing teacher model, LLMQuoter achieves competitive results in a resource-efficient fine-tuning setup. It democratizes advanced RAG capabilities, delivering significant performance improvements without requiring extensive model retraining. Our results highlight the potential of distilled quote-based reasoning to streamline complex workflows, offering a scalable and practical solution for researchers and practitioners alike.

1 INTRODUCTION

Large Language Models (LLMs) have revolutionized natural language processing, exhibiting robust performance across a wide range of tasks such as open-domain question answering, summarization, and conversational AI (Lin et al., 2024; Jin et al., 2024; An et al., 2024). Yet, as model sizes grow, so do their computational demands, creating inefficiencies—particularly in tasks requiring complex reasoning or retrieval from large contexts. Retrieval-Augmented Generation (RAG) has emerged as a popular solution, integrating external knowledge sources so models can dynamically access relevant information without extensive retraining (Mirzadeh et al., 2024; Hu et al., 2024). However, smaller models still struggle to maintain coherent reasoning over extensive or noisy contexts, highlighting a persistent gap in efficiency and accuracy.

Knowledge distillation addresses these challenges by transferring capabilities from high-capacity teacher models to smaller students, preserving advanced features like multi-step reasoning and factual consistency while reducing computational overhead (Fu et al., 2024; Gogate et al., 2024). Distilled student models can leverage split-step reasoning, domain-specific fine-tuning, and self-correction mechanisms to tackle intricate tasks, improving both inference efficiency and overall performance (Yao et al., 2024; Zhang et al., 2024b).

Within the realm of retrieval-augmented approaches, RAFT (Retrieval-Augmented Fine-Tuning) exemplifies how "quote while thinking" strategies can bridge the gap between retrieval and generation (Zhang et al., 2024a; Di Oliveira et al., 2024). By training the model to reason, quote relevant passages, and answer in one sequence, RAFT demonstrates that targeted fine-tuning can enhance context-aware responses (see Figure 1). Nevertheless, even well-crafted frameworks like RAFT encounter difficulties when smaller LLMs face large documents or complex multi-step reasoning (Zhang et al., 2024b; Chen et al., 2024).

To address these limitations, we propose *LLMQuoter*, a lightweight model that adopts a "quote-first-then-answer" strategy. Rather than reasoning over an entire context, LLMQuoter identifies and retrieves the most pertinent excerpts, which are subsequently handed off to downstream models. This

[a] https://orcid.org/0009-0001-8294-7163
[b] https://orcid.org/0000-0003-1826-1850

Bezerra, Y. F. and Weigang, L.
LLMQuoter: Enhancing RAG Capabilities Through Efficient Quote Extraction from Large Contexts.
DOI: 10.5220/0013358700003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1335-1342
ISBN: 978-989-758-737-5; ISSN: 2184-433X

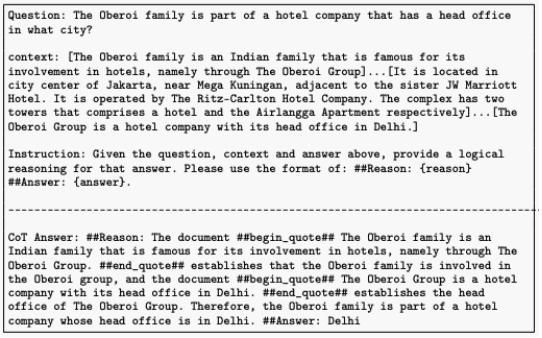

Figure 1: RAFT inference example(Zhang et al., 2024a).

decouples retrieval from reasoning, reducing the cognitive load and enabling both large and small models to achieve higher accuracy with less computational cost. By building on knowledge distillation and leveraging Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tune a LLaMA-3B model, LLMQuoter streamlines RAG pipelines, surpassing full-context approaches like RAFT in efficiency and scalability.

We evaluate LLMQuoter within the DSPy framework (Khattab et al., 2023), using a 15,000-sample subset of the HotpotQA dataset (Yang et al., 2018), a benchmark commonly employed for retrieval-augmented generation (RAG). Empirical results reveal that LLMQuoter excels in accuracy, all while remaining computationally lightweight. Through a two-phase workflow—quote retrieval followed by reasoning—LLMQuoter democratizes access to advanced RAG solutions, offering a scalable alternative for researchers and practitioners constrained by computational resources.

This workflow achieves over 20-point accuracy gains compared to full-context approaches like RAFT, demonstrating significant improvements across both small and large language models. Leveraging knowledge distillation from high-performing teacher models, LLMQuoter delivers competitive results with resource-efficient fine-tuning, eliminating the need for extensive retraining. The approach highlights the scalability of distilled quote-based reasoning, providing a practical and efficient solution for RAG workflows.

2 METHODOLOGY

With the goal of developing an efficient language model for extracting relevant quotes from contexts to properly answer questions about it, this section details the methodology employed in training and evaluating the distilled LLM. The process involves leveraging a high-performing LLM for dataset creation, fine-tuning a smaller LLM, and validating the approach with task-specific metrics.

We begin with a formalization of the distillation problem in Section 2.1, followed by an overview of the fine-tuning process in Section 2.2. Finally, the evaluation framework and metrics used to validate the model's performance are described, along with a simple approach to demonstrate the benefits of extracting relevant quotes instead of using the large content itself.

2.1 Problem Formalization

Let us consider a dataset of text samples, denoted by $D = \{(C,Q,A)\}$, where:

- C: a large text context.
- Q: a specific question.
- A: the expected answer.

The task is to train a model capable of extracting relevant quotes from C that support A in response to Q.

To achieve this, we employ a distillation process in which a large LLM generates high-quality training data, and a smaller LLM is fine-tuned on this dataset to efficiently replicate the behavior of the larger model.

2.2 LLM Distillation

The dataset creation process can be formalized as follows: Given a high-performance language model f_{high}, such as ChatGPT or Gemini, the task is to extract quotes $\mathcal{R}$ from a context C that directly support an answer A in response to a question Q. Formally, this process can be represented as:

$$f_{\text{high}} : (Q,A,C) \rightarrow \mathcal{R}$$

For each data point (Q,A,C), the high-performance model f_{high} generates the set of quotes $\mathcal{R}$, which serve as the ground truth:

$$\mathcal{D}_{\text{gold}} = \{(Q,A,C,\mathcal{R}) \mid \mathcal{R} = f_{\text{high}}(Q,A,C)\}$$

The result is a high-quality dataset $\mathcal{D}_{\text{gold}}$, consisting of tuples $(Q,A,C,\mathcal{R})$, where $\mathcal{R}$ represents the relevant quotes extracted by f_{high}. This dataset is then used to train and evaluate the smaller distilled model f_{small}.

2.3 Fine-Tuning LLM with LoRA

The smaller model f_{small} is fine-tuned on the $\mathcal{D}_{\text{gold}}$ dataset using Low-Rank Adaptation (LoRA) for task-specific learning in the extraction of relevant quotes. The fine-tuning process is defined as:

$$f_{\text{small}} : (Q,C) \rightarrow \mathcal{R}$$

where Q represents the question, C is the textual context, and $\mathcal{R}$ is the set of relevant quotes generated by the fine-tuned model. The training process is described in the following steps:

1. **Input:** Data from the $\mathcal{D}_{\text{gold}}$ dataset in the form of tuples (Q, C), where Q is the question, C is the textual context.
2. **Output:** The fine-tuned model f_{small} is optimized to predict $\mathcal{R}$, replicating the behavior of the larger model f_{high}, **but without knowing the answer**.

2.4 Evaluation Framework and Metrics

The model's performance is evaluated using the DSpy framework, which computes task-specific metrics tailored to LLM outputs. Precision and recall are redefined for the quote extraction task using an LLM Judge to assess semantic relevance between model predictions and ground truth.

Precision measures the proportion of predicted quotes (R_{model}) that align semantically with the golden answers (R_{gold}), defined as:

$$P = \frac{\sum_{r \in R_{\text{model}}} \text{Judge}(r, R_{\text{gold}})}{|R_{\text{model}}|}$$

where R_{model} is the set of quotes predicted by the model, R_{gold} is the set of golden answers, and $\text{Judge}(r, R_{\text{gold}})$ is a scoring function returning values from 0 (no match) to 1 (perfect match).

Recall quantifies the proportion of golden answers (R_{gold}) captured by the model's predictions (R_{model}), defined as:

$$R = \frac{\sum_{r \in R_{\text{gold}}} \text{Judge}(r, R_{\text{model}})}{|R_{\text{gold}}|}$$

F1-score balances precision and recall and is defined as:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

DSpy-Assisted Validation with LLM Judge: The DSpy framework incorporates large language models (LLMs) as automated evaluators, enabling robust and interpretable metric calculations. This flexibility allows DSpy to integrate a wide range of LLMs, referred to here as the LLM Judge. This variation of precision and recall, tailored for LLM-generated outputs and supported by the LLM Judge's semantic judgment, ensures a nuanced evaluation of the quote extraction model. The integration of DSpy and the Judge provides a systematic, interpretable, and robust framework for assessing and iteratively improving model performance.

2.5 Proving the Benefit of Using Quotes

Let f_{base} represent base models without any fine-tuning to establish a baseline for comparison. Two experimental setups are defined to demonstrate the advantage of using relevant quotes $\mathcal{R}$ instead of the full context C:

1. Providing only the gold quotes $\mathcal{R}$ from $\mathcal{D}_{\text{gold}}$ to the base models f_{base} to answer the questions:
$$f_{\text{base}} : (Q, \mathcal{R}_{\text{gold}}) \rightarrow A_{\text{base}}$$
2. Providing the full context C instead of the quotes $\mathcal{R}$ to the same base models f_{base} to answer the questions:
$$f_{\text{base}} : (Q, C) \rightarrow A_{\text{base}}$$

For both setups, Q represents the question, $\mathcal{R}_{\text{gold}}$ is the set of gold quotes extracted from the $\mathcal{D}_{\text{gold}}$ dataset, C is the entire context, and A_{base} is the base models answers.

The accuracy of the answers produced by f_{base} is measured using Semantic Accuracy (S_{acc}), which evaluates the alignment between the model-generated answers A_{base} and the expected answers A_{gold}. Semantic Accuracy is defined as:

$$S_{\text{acc}} = \frac{\sum_{a \in A_{\text{base}}} \text{Judge}(a, A_{\text{gold}})}{|A_{\text{gold}}|}$$

where $\text{Judge}(a, A_{\text{gold}})$ is a semantic similarity function scoring the alignment between a model-generated answer a and the ground truth A_{gold}, with scores ranging from 0 (no match) to 1 (perfect match).

3 EXPERIMENTS

This section describes the experimental setup used to analyze the performance of the proposed methodology. It begins with details of the datasets used for training and evaluation, followed by an explanation of the training configurations, including hyper-parameters and computational resources. An overview of the entire process, from data distillation to evaluation, is illustrated in Figure 2. Finally, the experiments designed to validate the effectiveness of using relevant quotes instead of full context are presented (Figure 3 illustrates the process). The code utilized in this work is available on GitHub[1]. Concrete examples of the experimental results can be found in the appendix for further clarification.

[1] https://github.com/yurifacanha/LLMQuoter

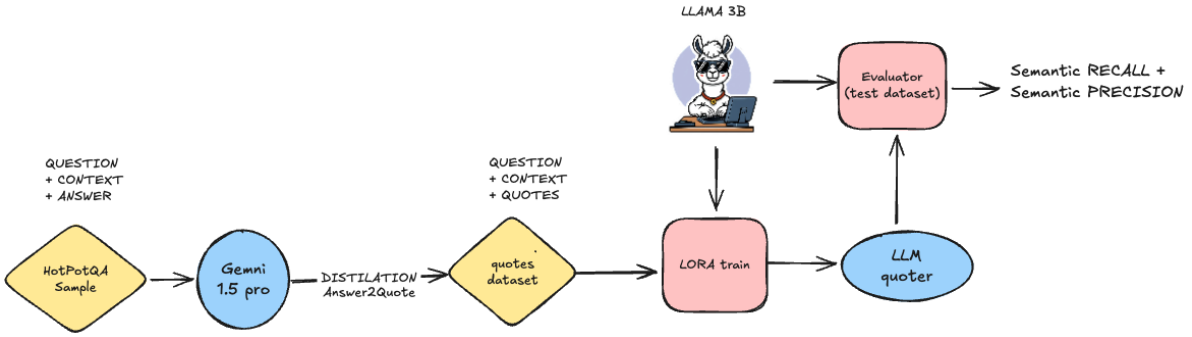

Figure 2: The LLMQuoter diagram.

3.1 Datasets

Our method was evaluated on the HotpotQA dataset (Yang et al., 2018), an open-domain question-answering benchmark derived from Wikipedia, with a focus on common knowledge topics such as movies, sports, and general trivia. The dataset consists of three columns: **question**, **context**, and **answer**, where each sample pairs a question with a large textual context and its corresponding answer.

Due to resource constraints, a random subset of 15,000 samples was selected from the original dataset to serve as the basis for applying the distillation process. From this subset, 600 samples were set aside for evaluation purposes, forming the test set. This test set was used to measure the model's performance during the evaluation phase and to validate the benefit of using extracted quotes as opposed to the entire context for answering questions. The remaining 14,400 samples were utilized for training and validation during the distillation and fine-tuning steps.

3.2 Data Distillation

The distillation process was performed using **Gemini Pro 1.5** as the high-performance model (f_{high}) and LangChain as the framework for managing the pipeline. The process involved generating relevant quotes for each sample in both the training and test datasets by leveraging the capabilities of Gemini Pro 1.5.

Gemini Pro 1.5, as one of the most powerful models available today, was tasked with extracting quotes directly supporting the answer to each question. Given the model's advanced performance and ability to generate high-quality answers, it is reasonable to assume that the resulting dataset represents an excellent "gold" standard for the task of quote extraction.

After this step, the dataset was finalized, augmented with a new column containing the extracted quotes ($\mathcal{R}$). This enriched dataset, now comprising question (Q), context (C), and quotes ($\mathcal{R}$), served as the foundation for training and evaluating the smaller f_{small} model.

3.3 Fine-Tuning Process

The fine-tuning process was applied to the smaller LLM, LLAMA 3.2 3B, using the Low-Rank Adaptation (LoRA) technique to optimize the model for the quote extraction task. LLAMA 3.2 3B was chosen as the base model due to its balance between computational efficiency and task-specific adaptability. The fine-tuning process was completed over a single epoch, ensuring efficient adaptation without overfitting.

The fine-tuning process was conducted on a NVIDIA A100-SXM4-40GB GPU, with a maximum memory capacity of 39.564 GB. The specific resource utilization and training parameters are summarized below:

Table 1: Summary of Fine-Tuning Configuration and Resource Usage.

Configuration/Metric	Value
Memory Usage	3.56GB(peak)
Training Memory	1.06GB(peak)
Batch Configuration	Batch size: 2
Gradient accumulation steps	4
Total effective batch size	8
Training Steps	60
Trainable Parameters	24M aprox
Training Time	5 minutes

This setup highlights the efficiency of the LoRA approach in adapting a compact model like LLAMA 3.2 3B for specific tasks with minimal resource usage and rapid training over just one epoch (see Table 1).

3.4 Evaluation and Proving the Benefits

The evaluation of the extracted quotes was performed using the DSpy framework in conjunction with OpenAI GPT-4.0. GPT-4.0 was selected as it operates outside the scope of the training data and methods, is recognized as one of the top reasoning models, and remains unbiased regarding the problem context. By leveraging these tools, the metrics defined in the methodology section were concretely implemented and materialized for evaluating the system's performance in a structured and measurable way.

To validate the benefit of using quotes instead of the full context, comparisons were performed across several base models (f_{base}), including **LLAMA 3.2:1B, LLAMA 3.2:3B, GPT-3.5 Turbo**. These models were evaluated in two configurations: using extracted quotes $\mathcal{R}$ and using the full context C. The accuracy of the answers produced by these models was assessed to determine the effectiveness of the quote extraction approach. GPT-4.0 was chosen as the external LLM Judge again to compute Semantic Accuracy (S_{acc}).

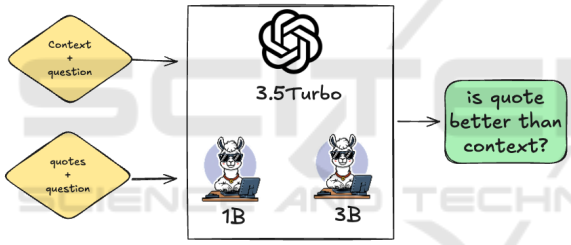

Figure 3: Context X context process.

4 RESULTS AND DISCUSSION

This section presents the experimental results obtained by evaluating the quote extraction model (quoter) and validating the benefit of using quotes over full context in open-domain question-answering tasks. The results demonstrate the effectiveness of the proposed method in improving the performance of both small and large language models in RAG (retrieval-augmented generation) scenarios.

4.1 Evaluation of the Quoter Model

The performance of the quoter model was evaluated using the metrics described in Section 3.3. The recall, precision, and F1-score were measured both before and after fine-tuning the smaller LLM using the LoRA approach. The results are summarized in Table 2.

Table 2: Performance of the Quoter Model Before and After Fine-Tuning.

Metric	Before	After
Recall	48.3%	68.0%(**+19.7%**)
Precision	43.6%	71.0%(**+27.4%**)
F1-Score	41.3%	69.1%(**+27.8%**)

The results show significant improvements in all three metrics after fine-tuning the quoter model. The F1-score increased from 41.3% to 69.1%, demonstrating the quoter's ability to accurately identify relevant quotes with low computational resources and a compact model.

4.2 Benefits of Using Quotes over Full Context

To validate the benefit of using quotes instead of full context, a comparison was performed using original models without any training. Both the gold quotes and the full context were provided as inputs to different models: LLAMA 1B, LLAMA 3B, and GPT-3.5 Turbo. The accuracy of the answers generated by each model in these two configurations is summarized in Table 3.

Table 3: Comparison of Accuracy Between Using Full Context and Quotes.

Model	Context	Quotes
LLAMA 1B	24.4%	62.2% (**+37.8%**)
LLAMA 3B	57.7%	83.0% (**+25.3%**)
GPT-3.5 Turbo	75.8%	88.5% (**+12.7%**)

The results highlight a clear improvement in accuracy when using gold quotes compared to full context. For instance, LLAMA 1B achieved an accuracy of 62.2% with quotes versus 24.4% with full context, and GPT-3.5 Turbo achieved 88.5% with quotes versus 75.8% with full context. These findings indicate that providing a good quoter model can significantly enhance the performance of both small and large language models in RAG scenarios.

4.3 Discussion

The results validate the hypothesis that using extracted quotes instead of full context significantly improves model performance in open-domain question-answering tasks. This finding aligns with the original RAFT approach, which involves reasoning and answering directly over the full context. However, our experiments demonstrate that separating the tasks—first extracting quotes with a simple quoter and then reasoning over the concise data—can lead

to comparable or better outcomes with lower computational overhead.

Table 4: Comparison of RAFT and Full Context Results on LLaMA2-7B over HotPotQA dataset.

Method	Accuracy
LLaMA2-7B + Full Context	26.43%
RAFT (LLaMA2-7B)	35.28%

To provide context, RAFT was tested with LLaMA2-7B over the full dataset, achieving an accuracy of 35.28% when reasoning over both context and question simultaneously. Using the same model (LLaMA2-7B) with only the full context reduced performance to 26.43% (see Table 4). While our experiments used a random sample of 15,000 rows from the HotpotQA dataset due to resource constraints, the results are promising. For instance, even with a lightweight 3B quoter model fine-tuned with minimal resources on Colab, the quote-based approach significantly boosted accuracy for various downstream models.

The comparison highlights that the quoter technique is a promising alternative. By offloading the task of quote extraction to a small and efficient model, we can streamline the reasoning process for larger models, avoiding the pitfalls of over-reasoning. The ”divide and conquer” strategy allows each model to focus on its strength: smaller models specialize in targeted preprocessing, while larger models excel in reasoning over concise, relevant data.

While our study only utilized a subset of the HotpotQA dataset, the results suggest that the quoter technique offers a scalable and efficient solution for enhancing retrieval-augmented generation (RAG) pipelines. Notably, the models used with the extracted quotes were not fine-tuned to reason better, yet still achieved significant improvements in accuracy. This highlights the power of the quoter approach in simplifying the reasoning task by reducing the cognitive load on base models, allowing even non-optimized models to perform effectively.

This approach could serve as a viable alternative to RAFT in scenarios with limited resources, demonstrating that a well-trained quoter can democratize access to high-performing NLP solutions. By offloading the preprocessing task of identifying relevant information, the quoter enables base models to focus their reasoning capabilities on concise, relevant data rather than processing large and noisy contexts.

5 CONCLUSIONS AND FUTURE WORK

This study demonstrates the effectiveness of data distillation and lightweight training for enhancing Retrieval-Augmented Generation (RAG) systems. By leveraging a high-performing teacher model to distill relevant quotes and fine-tuning a compact model, we achieved significant improvements in model performance. The fine-tuning process required minimal resources, with just 5 minutes of training on an NVIDIA A100 GPU, yet delivered robust results.

The experiments validate that an efficient quoter model can substantially enhance RAG performance by reducing the cognitive load on the reasoning process. By focusing the model's efforts on the answer rather than processing and reasoning over large contexts, we eliminate the need for extensive training while improving accuracy. This approach aligns with the principle of "divide and conquer," where the reasoning task is simplified and made more manageable for even small models. Ultimately, our results demonstrate that high-quality quote extraction can democratize access to high-performing RAG capabilities across a range of computational constraints.

While this work has established a strong foundation for quote-based RAG, several avenues for future research remain open:

- **Expanded Datasets:** Test the approach on diverse datasets across various domains and complexities to ensure broader applicability and robustness.
- **Reinforcement Learning:** Utilize techniques like Proximal Policy Optimization (PPO) or Direct Preference Optimization (DPO) to enhance quote extraction and reasoning.
- **Larger Models:** Explore scalability by training larger models, such as an 8B parameter LLAMA, to assess the impact of size on performance.
- **Prompt Engineering:** Develop advanced prompts to optimize extraction and reasoning, improving system accuracy and efficiency.
- **Extended Applications:** Adapt the methodology for memory-augmented systems to efficiently retrieve and manage information from extensive external knowledge bases.

By exploring these directions, we aim to further refine the quote-based RAG pipeline and expand its applicability to broader NLP tasks, offering scalable and resource-efficient solutions for both research and real-world scenarios.

REFERENCES

An, S., Ma, Z., Lin, Z., Zheng, N., and Lou, J.-G. (2024). Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Chen, X., Wang, L., Wu, W., Tang, Q., and Liu, Y. (2024). Honest ai: Fine-tuning" small" language models to say" i don't know", and reducing hallucination in rag. *arXiv preprint arXiv:2410.09699*.

Di Oliveira, V., Bezerra, Y. F., Weigang, L., Brom, P. C., and Celestino, V. R. (2024). Slim-raft: A novel fine-tuning approach to improve cross-linguistic performance for mercosur common nomenclature. In *WEBIST*.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2024). Generative context distillation. *arXiv preprint arXiv:2411.15927*.

Gogate, N. et al. (2024). Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *TechRxiv*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, S., Tu, Y., Han, X., et al. (2024). Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Jin, H., Han, X., Yang, J., Jiang, Z., Liu, Z., Chang, C.-Y., Chen, H., and Hu, X. (2024). Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., and Potts, C. (2023). Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Lin, B., Zhang, C., Peng, T., Zhao, H., Xiao, W., Sun, M., Liu, A., Zhang, Z., Li, L., Qiu, X., et al. (2024). Infinite-llm: Efficient llm service for long context with distattention and distributed kvcache. *arXiv preprint arXiv:2401.02669*.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yao, B., Zhang, Y., Li, Q., and Qin, J. (2024). Is sarcasm detection a step-by-step reasoning process in large language models? *arXiv preprint arXiv:2407.12725*.

Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., and Gonzalez, J. E. (2024a). Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Zhang, Y., Khalifa, M., Logeswaran, L., et al. (2024b). Small language models need strong verifiers to self-correct reasoning. *arXiv preprint arXiv:2404.17140*.

APPENDIX

This section presents examples of inferences drawn from the experiments.

A.1 Distillation

The input(Q,C,A):

```
"""
Instruction: Given the question,
the context
and the expected answer bellow,
provide relevant quotes from the
context that support the answer.
your answer must be just the
quotes, not the entire context.
format:
##begin_quote##quote##end_quote##
for each quote.
do not add anything else other
than the quotes.
Your turn:
Question: Unlike Xuzhou, where is Rugao
under the adminstration of?
Context: Rugao () is a county-level city
under the administration of Nantong,
Jiangsu province, China, located
in [about 200 words...] Shanghai.
Answer: Nantong
Quotes:
""")
```

And the generated quotes(R):

```
"""
##begin_quote## Rugao () is a county-level
city under the administration of
Nantong ##end_quote##
"""
```

A.2 Quote Train Sample

The input(Q,C):

```
"""
Instruction: Given the question
and the context
provide relevant quotes from the
context that support the answer.
your answer must be just the
quotes, not the entire context.
format:
##begin_quote##quote##end_quote##
for each quote.
do not add anything else other than
the quotes.
Question: What authority manages
the regional
passenger train service that
runs through
the same junction as
West Amesbury Branch Railroad?
```

```
Context: Rapido was the brand name
for the Canadian National Railway's
(CN) express passenger train service
in the Quebec City{Windsor Corridor.
"Rapido" service was introduced on
[..about 100 lines hide..]
The rail service may be replaced if the
line is closed due to rail maintenance,
a breakdown of a train, a rail accident,
strike action, or if the rail service
is not economically viable.
Quotes:
"""
```

The **desired output(R)**:

```
"""
##begin_quote## The Downeaster is a 145 mi
regional passenger train service, managed
by the Northern New England Passenger Rail
Authority (NNEPRA, created by the State of
Maine), and operated by Amtrak. ##end_quote##
##begin_quote## The West Amesbury Branch
Railroad was a railroad that once led form
Newton Junction, New Hampshire to Merrimac,
Massachusetts.  Although the railroad does
not exist, the Amtrak "Downeaster" line
now passes through the Newton Junction
station, which is now a pizza restaurant,
and most of the railroad, is now a gravel
walking trail. ##end_quote##
"""
```

Table 5 demonstrates an example of model performance before and after training in a test sample for the question: "Which film was produced first, Emil and the Detectives or Toy Story 2?"

A.3 Comparison: Quote x Context

An example illustrating the performance comparison between using full context and extracted quotes.

Question:

```
"""
Which Walt Disney Pictures
film was created first,
Finding Dory or The Wild Country?
"""
```

Context: A '5086' characters context about Disney and Pixar films.

Quotes:

```
"""
##begin_quote## The Wild Country is a 1970
American adventure film produced by Walt
Disney Pictures and directed by Robert Totten.
##end_quote##
##begin_quote## Finding Nemo is a 2003
American computer-animated
family film produced
by Pixar Animation Studios and released
by Walt Disney Pictures. ##end_quote##
"""
```

Table 5: Golden quote , before and after train quotes inferences from a test sample.

Gold	**##begin quote##** Emil and the Detectives is a 1964 film directed by Peter Tewksbury based on the novel by German author Erich Kästner. **##end quote##** **##begin quote##** Toy Story 2 is a 1999 American computer-animated comedy film produced by Pixar Animation Studios for Walt Disney Pictures. **##end quote##**
Before	**##begin quote##** "A Bug's Life", "Monsters, Inc.", "Finding Nemo", "Cars", "The Incredibles", "Ratatouille" **##end quote##** **##begin quote##** "Toy Story 3" (2010) is the third installment in Pixar's "Toy Story" series, and the sequel to 1999's "Toy Story 2". **##end quote##**
After	**##begin quote##** Emil and the Detectives is a 1964 film directed by Peter Tewksbury based on the novel by German author Erich Kästner. **##end quote##** **##begin quote##** Toy Story 2 is a 1999 American computer-animated comedy film produced by Pixar Animation Studios for Walt Disney Pictures. **##end quote##**

A detailed comparison of the model's performance, based on answers generated using either the provided context or the golden quotes, is presented in Table 6.

Table 6: Comparison of Q/A results: context vs. quotes.

Model	Context Answer	Quotes Answer
gpt3.5-turbo	Finding Nemo was created first.	The Wild Country
llama3.2:1b	Finding Dory is created first.	The Wild Country
llama3.2:3b	Finding Dory is created first.	The Wild Country