Self Supervised LLM Customizer(SSLC): Customizing LLMs on Unlabeled Data to Enhance Contextual Question Answering

Raveendra Hegde
Amazon.com
Bengaluru, India
raveendh@amazon.com

Saurabh Sharma
Amazon.com
Bengaluru, India
sharsar@amazon.com

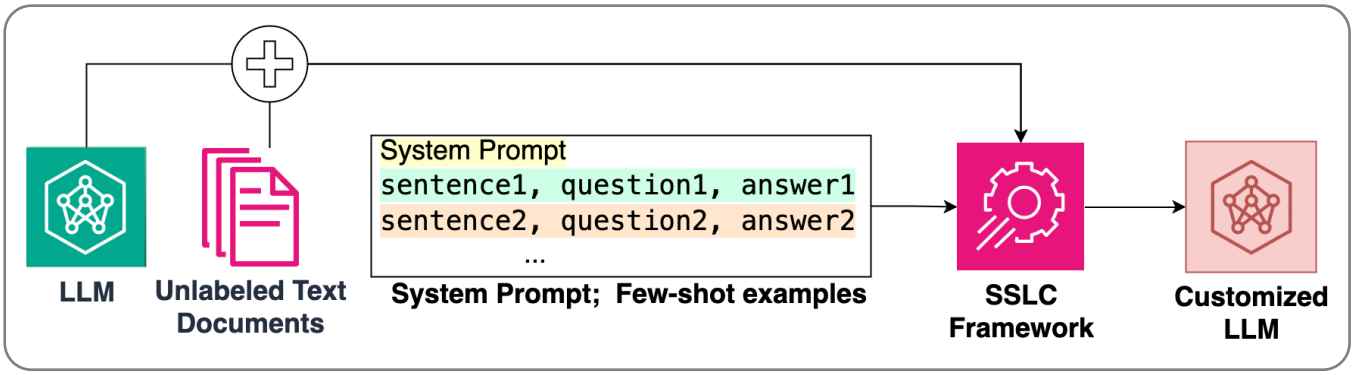

Figure 1: SSLC framework high level architecture

Abstract

While we can customize large language models (LLMs) on specific domains by finetuning using the domain specific labeled data, performance of the customized models is highly dependent on the quality of the labeled data. Obtaining high-quality labeled data for custom domains often requires considerable human effort and associated costs. However, in many cases, unlabeled data is readily available at little or no cost. Existing methods either rely on continued pre-training or use general purpose models trained for synthesis. But, continued pre-training necessitates vast amounts of data and adversely affects instruction tuned models. On the other hand, general purpose synthesis models might not capture the nuances of custom data. We present a framework (SSLC) for customizing LLMs using unlabeled text to enhance contextual question answering on custom data. Our approach employs few-shot synthesis using an instruction-tuned model, curates the synthesized data using a LLM response scorer and finetunes the model on this synthesized data. We demonstrate that the approach significantly improves contextual question answering performance compared to the baselines. It outperforms baselines in 75% (9/12) of the experiments as evidenced by both quantitative and qualitative metrics. On an average, it outperforms un-customized models by 19.3 percentage points and state-of-the-art approach by 4.4 percentage points in human evaluation(proxy) accuracy.

AIMLSystems 2024, October 08–11, 2024, Baton Rouge, USA

ACM ISBN 979-8-4007-1161-9…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS Concepts

• **Computing methodologies** → **Natural language generation**; *Learning from implicit feedback*.

Keywords

Language generation, Self Supervised, LLM

ACM Reference Format:
Raveendra Hegde and Saurabh Sharma. 2024. Self Supervised LLM Customizer(SSLC): Customizing LLMs on Unlabeled Data to Enhance Contextual Question Answering . In *Proceedings of 4th International Conference AI-ML Systems 2024 (AIMLSystems 2024).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models (LLMs) demonstrate remarkable capabilities in various natural language processing tasks, including contextual question answering. Contextual question answering is a crucial task that involves comprehending a given context and providing accurate answers to questions related to the context. With the rise of Retrieval Augmented Generation(RAG) use cases [17], contextual question answering has gained a lot of traction. While LLMs have shown promising results on benchmark datasets in contextual question answering task, their performance can degrade when applied to custom datasets with domain-specific data distributions and idiosyncrasies. This challenge arises due to the fact that LLMs are typically trained on large, general-purpose datasets, which may not capture the nuances of custom datasets.

Researchers have identified that continued pre-training [14], [6], [28], [19] is an effective approach to customize LLMs on custom data. In this approach, the large language model undergoes additional pre-training on custom data. But it requires a substantial amount of pre-training data and can potentially harm the performance of instruction-tuned models, as shown by [22].

To address these limitations, researchers have employed domain-specific instruction-tuning, as demonstrated by [23], [13] and [8]. This approach involves finetuning LLMs on domain specific instruction tuning dataset. However, this method often requires a considerable amount of high-quality labeled instruction data, which can be costly and time consuming to obtain.

In an effort to overcome this challenge, [22] has proposed training dedicated general-purpose models for synthesizing labeled data from unlabeled text. While these models have shown promising results outperforming previous approaches, the synthesized instructions might not fully capture the nuances of the custom data.

In this work, we propose **S**elf **S**upervised **LLM** **C**ustomizer (**SSLC**), a simple yet effective framework for customizing LLMs for contextual question answering on custom unlabeled text. The framework uses unlabeled text and handful of user provided examples to customize an instruction-tuned LLM. It combines existing techniques like few-shot synthesis [5], use of a state-of-the-art LLM response scorer Cappy [24] for curating high quality synthesis data to automatically generate contextual question answering instruction-data. Subsequently, the framework customizes the model through instruction-tuning [30] on the generated data.

We assess the effectiveness of the proposed approach by evaluating it on popular large language models: Tiny-Llama [29], Llama2-7B [25] and Mistral-7B [16]. We conduct the evaluation across three diverse datasets from SQuADShifts [21]: Reddit, New York Times (NYT), Wikipedia(Wiki) and AdversarialQA [4]. We compare the performance of our approach against two baselines: the un-customized model and the state-of-the-art approach Bonito[22]. The results demonstrate that our proposed approach achieves significant improvements over both the baselines, highlighting its effectiveness in customizing LLMs for contextual question answering on custom data.

In summary, our main contributions are:

- We introduce a simple yet effective framework for customizing LLMs for contextual question answering on custom unlabeled text documents.
- Through comprehensive evaluations across various models and datasets, we demonstrate that the proposed approach outperforms baselines, even when employing the same model for synthesis in a self-synthesis setting.
- Our empirical findings suggest that a smaller model can be a more suitable candidate for customization compared to certain larger models.
- We showcase the versatility of our framework by demonstrating its ability to enhance the customization of weaker models by leveraging the synthesis capabilities of more powerful models.

2 Problem Formulation

Let $\mathcal{M}_{pt}$ denote the pre-trained large language model (LLM), which has been trained on a large general-purpose dataset. The performance of $\mathcal{M}_{pt}$ can be sub-optimal for a given language modelling task, when applied to domain-specific datasets since there was no labelled domain-task specific datasets during pretraining leading to mismatch in data distributions.

To address this challenge, we aim to customize $\mathcal{M}_{pt}$ on an unlabelled domain-specific dataset $\mathcal{D}_{custom}$ using a self-supervised framework. Let $\mathcal{M}_{ft}$ denote the customized model obtained by finetuning $\mathcal{M}_{pt}$ on $\mathcal{D}_{custom}$, and a set of task instructions $\mathcal{I}$.

The objective is to learn the updated model parameters θ_{ft} of $\mathcal{M}_{ft}$ such that the model's performance on the domain-specific task (e.g., contextual question answering) is significantly improved compared to the pre-trained model $\mathcal{M}_{pt}$. Formally, we aim to solve the following optimization problem:

$$\theta_{ft} = \underset{\theta}{\operatorname{argmin}}(\mathcal{L}(\theta; (\mathcal{D}_{custom}, \mathcal{I}))) \tag{1}$$

$$\text{s.t.} \quad \mathcal{L}(\theta_{ft}; (\mathcal{D}_{custom}, \mathcal{I})) \ll \mathcal{L}(\theta_{pt}; (\mathcal{D}_{custom}, \mathcal{I})) \tag{2}$$

where $\mathcal{L}(\theta; \mathcal{D})$ is the task-specific loss function (e.g., cross-entropy loss for question answering) evaluated on the dataset $\mathcal{D}$, task instructions $\mathcal{I}$, and θ_{pt} are the parameters of the pre-trained model $\mathcal{M}_{pt}$.

The key challenge lies in the fact that $\mathcal{D}_{custom}$ is an unlabeled dataset, and obtaining high-quality labeled data for finetuning can be costly and time-consuming. To overcome this limitation, we propose a self-supervised framework that can effectively customize $\mathcal{M}_{pt}$ using only the unlabeled data in $\mathcal{D}_{custom}$.

3 Methodology

In this section, we first provide a high-level overview of the framework(SSLC) followed by a detailed explanation of its architecture, components and the underlying algorithms.

Figure 1 illustrates the high-level architecture of the SSLC framework. The framework accepts an instruction-tuned LLM, unlabeled text documents, and a system prompt along with few-shot synthesis examples as input. It produces a customized LLM with enhanced contextual question answering capability on the documents. The input model serves two purposes: 1) to synthesize question-answer pairs, which are subsequently used for instruction-tuning, and 2) to produce the customized LLM through instruction-tuning on the synthesized instruction-data. Optionally, SSLC can also accept a different model for synthesis step. Figure 2 illustrates the detailed architecture of SSLC framework, its components and the customization process. The framework has the following components:

(1) Text Segmentation
(2) Sentence Extraction
(3) Synthesis
(4) Curation
(5) Finetuning

Text Segmentation: The framework accepts large, unlabeled text documents as input, which need to be divided into smaller units, as LLMs only support a limited context length. Additionally, real-world datasets are often huge unstructured documents, unlike the short contexts found in publicly available datasets. Our approach tackles

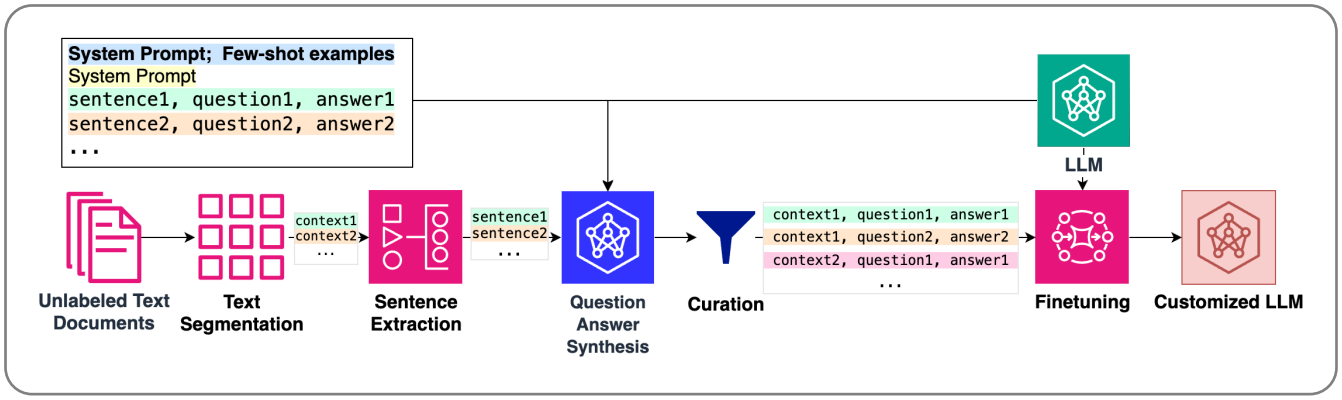

Figure 2: SSLC framework detailed architecture

this issue through a straightforward text segmentation technique. The segmentation process involves sequentially parsing the text and creating segments composed of consecutive sentences that adhere to a predefined maximum word length constraint. To maintain the semantic and syntactic integrity of the segments, the segmentation process stops at nearest complete sentence within the word length constraint, ensuring that segments contain only whole sentences. The extracted segments serve as the context for the customization process.

Sentence Extraction: While it is possible to pass a segmented context to LLM for synthesizing question-answer pairs, the synthesized questions and answers only cover portions of the context, leaving other parts unrepresented. To overcome this limitation and ensure comprehensive coverage of the entire context, the framework employs a technique that involves extracting individual sentences from the given context. These extracted sentences can then be passed to the LLM to synthesize question-answer pairs that collectively represent the whole context, avoiding any bias.

Question-Answer Synthesis: Once the sentences are extracted, the framework needs to derive question-answer pairs, which can be used as instruction-data for the customization process. This component accepts an instruction-tuned LLM and uses few-shot synthesis technique to generate the question-answer pairs. The LLM can be either the same model which is the candidate for customization or any other instruction-tuned model. For each sentence, this component constructs a prompt by combining the system prompt, few-shot examples, and the sentence itself. This prompt is then passed to instruct the LLM for synthesizing question-answer pairs. The resulting context, sentences, questions, and answers are then compiled to form a synthetic dataset, which can be used for finetuning the model.

The algorithms for segmentation, sentence extraction, and synthesis are detailed in Algorithm 1.

Curation: The synthesis process may produce inferior samples due to trivial, semantically meaningless input sentences or hallucinations of the LLM employed for the task. In order to overcome this problem, the framework includes a component which curates high quality data. The curation component leverages a specialized model

Algorithm 1 Synthesize question-answer

```
procedure SynthesizeQA(llm, dcouments, prompt_components)
    synthesized_data ← [ ]
    for d in documents do
        contexts ← extract_contexts(d)
        sys_prompt, exmpls ← prompt_components
        for c in contexts do
            sentences ← extract_sentences(c)
            for s in sentences do
                prompt ← format_prompt(sys_prompt, exmpls, s)
                q, a ← generate_question_answer(llm, prompt)
                synthesized_data.append((c, s, q, a))
            end for
        end for
    end for
    return synthesized_data
end procedure
```

Cappy, proposed by [24], which is trained to score the correctness of language model responses. The Cappy model accepts an input prompt, corresponding language model response, and generates a score between 0 and 1, with higher scores indicating better response quality. The curation process iterates through the synthesized sentence-question-answer triplets and constructs a prompt-response pair for each triplet in the format expected by the Cappy model. The Cappy model then evaluates the response quality by scoring the prompt-response pair. If the score falls below a predefined threshold, the corresponding synthesized question-answer pair is discarded from the dataset. The algorithms for curation are detailed in Algorithm 2.

Finetuning: To enhance the model's ability to answer contextual questions on the custom dataset, the framework instruction-tunes the model using the curated high-quality instruction-data obtained from the previous step. The instruction-tuning technique employed by the framework is supervised finetuning, a form of auto-regressive finetuning. The context and the question are combined with a domain-specific question-answering prompt provided by the user, forming the instruction part of the finetuning input, while the answer serves as the target response. The finetuning process trains a quantized Low-Rank Adaptation (QLoRA) adapter [[15] [10]]on this

Algorithm 2 Curate synthesized question-answer data

```
procedure CURATE(threshold,synthesized_data)
    curated_data ← [ ]
    for context,sentence,q,a in synthesized_data do
        prompt,response ← format_prompt(sentence,q,a)
        score ← get_response_score(prompt,response)
        if score >= threshold then
            curated_data.append((context, q, a))
        end if
    end for
    return curated_data
end procedure
```

instruction-response dataset using huggingface TRL [26] library. The resulting customized adapter can then be loaded on top of the base model for generating superior quality answers to contextual questions on the dataset.

4 Experiments

We evaluate our approach on TinyLlama-1.1B, Llama2-7B, and Mistral-7B models, which are top performing models in their category, using 3 datasets from SQuADShifts: Reddit, NYTimes(NYT), Wikipedia(Wiki) and AdversarialQA dataset. We employ both quantitative and qualitative evaluation metrics, and consider vanilla models as well as state-of-the-art Bonito synthesis approach as baselines. We assess the proposed approach under two distinct setups: 1) employing the same model as the customization candidate for the synthesis process (self-synthesis) and 2) utilizing the best-performing synthesis model among those evaluated for the synthesis task. These evaluations provide empirical evidence that the framework performs well across different models and datasets.

4.1 Datasets and evaluation metrics

SQuADShifts [21] is a standard set of datasets for contextual question answering on custom text, with a distribution different from the general text used for LLM pre-training and the finetuning. Similarly, AdversarialQA dataset [4] contains reading comprehension related questions and answers designed to challenge models by including adversarial examples that are difficult to answer correctly. As these datasets are not directly usable for the problem our research aims to address, we transform them accordingly. These datasets are labeled collections containing identifiers, titles, contexts, questions, and ground truth answers, where the title column points to the source document. Table 1 shows the statistics of these datasets. We group the contexts based on title, remove duplicates, and concatenate them to create unlabeled text documents. We keep 10% of these documents as a held-out set, whose ground truth question-answer pairs serve as the test set for evaluation. This ensures that the model does not encounter the test set data during customization finetuning. Next we discuss metrics employed for evaluating experiments on the test set.

While human assessment remains the gold standard for evaluating machine-generated responses, we can also employ quantitative NLP evaluation metrics. As the core objective of the evaluation process is to determine the similarity between the generated answer and the ground truth or reference answer, we use ROUGE [20], METEOR [3] and BERTScore [31], which are widely used NLP evaluation metrics.

Table 1: Dataset statistics

Dataset	No. of samples
AdversarialQA	40,117
Reddit	15,804
NYT	16,846
Wiki	17,598

Human Evaluation (Proxy): As full-fledged human evaluation is costly and time-consuming, we employ a quasi-human evaluation approach by utilizing the Claude [2] model as a proxy evaluator. This technique of leveraging LLMs for assessing NLP tasks has gained popularity, as demonstrated by [32], [27], [12], [11] and [18]. Following this trend, we employ a simple evaluation prompt to compare the generated answer with the ground-truth answer. The accuracy, which is the percentage of generated answers that match the ground-truth answers, serves as the quantitative metric for the proxy human evaluation.

4.2 Experiment setup

We conduct experiments in four distinct setups for every combination of dataset and model to illustrate that our proposed approach generates a customized model that outperforms the baselines. The remaining parts of the section details the setups used for experiments.

Base (Un-customized): In this setup, we employ a zero-shot technique with the base un-customized model, for generating answers to the contextual questions in the test set. The zero-shot method employs a custom prompt specifically tailored to the dataset.

Bonito: Bonito is a state-of-the-art model for synthesizing instruction-data from unlabeled text. As a general purpose synthesis model, it can generate instruction-data for various tasks. In this setup, we utilize it to generate instruction-data for contextual question answering task and finetune the un-customized model to produce a customized model, as suggested by the author [22]. As the synthesis model doesn't allow customization of instruction prompts in the synthesized data, we use a common prompt found in the synthesized data while generating the answers for the evaluation.

SSLC (Self-Synthesis): In this setup, we utilize a self-synthesis technique where the customization candidate model itself is used to synthesize instruction-data for customization. The customization is carried out using the SSLC framework as described in the Methodology section.

SSLC (Best-Synthesis): This setup is exactly same as the SSLC (Self-Synthesis) except that we substitute the synthesis model with the Mistral-7B model, which is the best among the evaluated models.

To facilitate a fair comparison of metrics across different setups and remove any bias, we maintain a consistent number of tokens employed for finetuning across models, within each experiment.

The prompts utilized for the experiments and evaluation are provided in the Appendix C. Additionally, all the configurations, including the LoRA configuration and finetuning hyper-parameters, are detailed in the Appendix B.

4.3 Results and discussion

SSLC with self-synthesis Table 2 presents the outcomes of 12 experiments carried out across the models and the datasets. It lists four quantitative metrics pertaining to NLP: ROUGE-2, ROUGE-L, METEOR and BERTScore-F1 along with one qualitative metric: human proxy accuracy(Human Eval), which utilizes Claude [2] as human proxy.

As we employ multiple metrics, there are cases where certain variants perform better in terms of certain metrics. We consider a variant to be better than other variants only if it outperforms across the majority of the metrics.

Overall our proposed approach outperformed in **75% (9/12)** of the experiments, across all the evaluation metrics, demonstrating its superior performance. In terms of human evaluation(proxy) accuracy, it outperformed the closest baseline by **4.4 percentage points**. For more detailed results please refer Appendix A.

Comparison with Base(un-customized): The proposed approach outperforms in 83%(10/12) of the experiments when compared to the baseline on majority of the metrics. On an average, the proposed approach yielded substantial improvements over the baseline, with a 155% increase in ROUGE-2, a 149% increase in ROUGE-L and a 32% increase in METEOR. While the percentage improvements in BERTScore may not be meaningful due to its non-linear scale, the proposed approach outperformed the baseline in terms of BERT-F1 in 100%(12/12) of the experiments, indicating its effectiveness. In terms of human evaluation(proxy) accuracy, the proposed approach outperformed in **92%(11/12)** of the experiments with an impressive **19.3 percentage points** improvement on an average, compared to the vanilla model baseline.

Comparison with Bonito: The proposed approach outperforms in **75% (9/12)** of the experiments across all the evaluation metrics when compared to the Bonito baseline. On an average it achieved remarkable improvements, with a **42%** increase in ROUGE-2, a **29%** increase in ROUGE-L, and a **45%** increase in METEOR. Furthermore, the proposed approach matched or outperformed the baseline in terms of BERT-F1 in 75% (9/12) of the experiments. It improved human evaluation(proxy) accuracy by a notable **4.4 percentage points** on an average compared to the Bonito baseline.

SSLC with best synthesis: The results of the self-synthesis experiments indicate that TinyLlama, despite being significantly smaller than Llama2 (1.1B vs. 7B parameters), outperformed Llama2. To better understand this phenomenon, we customized both the models using the synthesized data from the Mistral-7B model (best among the evaluated), ensuring any differences in synthesis capabilities are eliminated. Even then, TinyLlama continued to outperform Llama2, as shown in the Table 3. We attribute its superior performance to larger pre-training token size (3T vs. 2T) and higher-quality training data, though this needs further investigation to rule out other potential factors, such as prompt sensitivity and finetuning hyper-parameters.

As expected, the TinyLlama model exhibited superior performance when employing the Mistral-7B model for synthesis, demonstrating that the framework can be used to enhance capabilities of a weaker model by leveraging a more powerful model.

4.4 Ablations

We dissect and evaluate the contribution of the core element of SSLC, namely curation. Curation, that involves filtering high-quality synthesized data, is a crucial component in the framework. We conducted an ablation study to compare the results with and without this component. The findings, presented in the Table 4, clearly demonstrate that the SSLC variants *with curation* consistently enhance the customization performance, outperforming the scenarios with *no curation*.

5 Conclusion

In this work, we proposed a simple yet effective framework for customizing large language models to improve contextual question answering on custom unlabeled text, which outperforms the baselines.

Through extensive experiments across multiple models and datasets, we demonstrated that our approach with self-synthesis consistently outperforms un-customized base models and state-of-the-art customization technique(Bonito). The proposed approach outperformed baselines in 75% (9/12) of experiments in terms of ROUGE, METEOR, BERTScore and human evaluation(proxy) accuracy metrics. Key results include 42% increase in ROUGE-2, a 29% increase in ROUGE-L, and a 45% increase in METEOR and a 4.4 percentage point increase in human evaluation(proxy) accuracy over the closest baseline, on an average. The customized models improved the human evaluation(proxy) accuracy by 19.3 percentage points higher over un-customized model on an average.

Additionally, our empirical findings suggest that a smaller model can be a more suitable candidate for customization compared to some larger models. We also demonstrated that this framework can be used to enhance a weaker model's performance by leveraging a more powerful synthesis model.

The simplicity and strong empirical results make this framework attractive for customizing large language models to improve contextual question answering on custom domains without labeled data.

5.1 Limitations

While the proposed approach for contextual question answering has demonstrated promising results, there are several limitations that should be acknowledged. Currently, the framework only supports sentence-level synthesis, which may not be effective for complex question answering scenarios requiring knowledge from multiple sentences within a broader context. Additionally, the current framework is tailored specifically for contextual question answering and does not support customization for other NLP tasks, such as summarization or named entity recognition.

It is important to note that these limitations do not diminish the significance of the proposed approach but rather highlight areas for future exploration and improvement. Addressing these limitations

Table 2: Results of experiments across models and datasets.
Variants: Base ⇒ Base un-customized model | Bonito ⇒ Bonito synthesis and finetuning | SSLC ⇒ SSLC with self-synthesis

Model	Dataset	Variant	ROUGE-2↑	ROUGE-L↑	METEOR↑	BERT-F1↑	Human Eval↑ (Proxy)
TinyLlama (1.1B)	AdversarialQA	Base	0.06	0.11	0.19	0.84	14.74
		Bonito	0.08	0.19	0.14	**0.88**	15.62
		SSLC	**0.11**	**0.21**	**0.23**	0.87	**21.17**
	Reddit	Base	0.07	0.13	0.23	0.84	22.08
		Bonito	0.25	**0.52**	0.38	**0.92**	**55.14**
		SSLC	**0.26**	0.46	**0.43**	0.91	45.98
	NYT	Base	0.12	0.20	0.33	0.86	32.77
		Bonito	0.31	0.53	0.40	0.87	55.38
		SSLC	**0.34**	**0.54**	**0.49**	**0.92**	**56.47**
	Wiki	Base	0.15	0.22	0.37	0.86	28.43
		Bonito	0.19	0.33	0.26	0.87	30.04
		SSLC	**0.35**	**0.51**	**0.49**	**0.92**	**49.94**
Llama2 (7B)	AdversarialQA	Base	0.05	0.09	**0.17**	**0.84**	**6.93**
		Bonito	0.03	0.07	0.07	0.82	5.42
		SSLC	**0.06**	**0.12**	0.12	**0.84**	**6.93**
	Reddit	Base	0.06	0.13	**0.21**	0.84	11.70
		Bonito	0.09	0.18	0.14	**0.85**	13.33
		SSLC	**0.11**	**0.21**	0.19	**0.85**	**15.16**
	NYT	Base	0.10	0.16	**0.28**	0.82	17.17
		Bonito	0.11	0.20	0.15	0.85	14.51
		SSLC	**0.17**	**0.27**	0.25	**0.86**	**18.46**
	Wiki	Base	0.12	0.17	**0.32**	0.85	13.35
		Bonito	0.08	0.13	0.12	0.83	11.62
		SSLC	**0.19**	**0.31**	0.29	**0.87**	**19.28**
Mistral (7B)	AdversarialQA	Base	0.06	0.12	0.22	0.84	18.72
		Bonito	0.17	0.37	0.30	0.90	33.98
		SSLC	**0.21**	**0.41**	**0.36**	**0.91**	**40.64**
	Reddit	Base	0.08	0.16	0.28	0.85	24.72
		Bonito	**0.39**	**0.72**	0.55	**0.95**	**76.09**
		SSLC	0.38	0.68	**0.56**	0.94	72.84
	NYT	Base	0.18	0.28	0.43	0.88	37.41
		Bonito	**0.47**	**0.74**	0.58	**0.95**	**76.70**
		SSLC	**0.47**	0.71	**0.61**	**0.95**	74.33
	Wiki	Base	0.17	0.25	0.41	0.87	33.75
		Bonito	0.38	0.54	0.52	0.92	53.03
		SSLC	**0.46**	**0.69**	**0.60**	**0.94**	**72.56**

could lead to a more robust and versatile framework capable of handling complex question answering scenarios, supporting multiple natural language processing tasks, and enabling more comprehensive customization options.

5.2 Future work

This research opens up several promising avenues for future exploration. One potential direction is to extend the proposed framework to handle more complex question-answering scenarios that require synthesizing information from multiple sentences and broader contextual understanding. Additionally, it would be valuable to evaluate the framework's performance on the latest language models, such as Microsoft Phi-3[1] and Llama-3, which have demonstrated superior capabilities compared to previous models. Finally, extending the framework's applicability to other natural language processing

Table 3: SSLC performance with synthesized data from Mistral [16]

Dataset	Model	ROUGE-2↑	ROUGE-L↑	METEOR↑	BERT-F1↑	Human Eval↑ (Proxy)
AdversarialQA	Llama2	0.07	0.13	0.13	**0.84**	7.01
	TinyLlama	**0.12**	**0.26**	**0.23**	0.83	**23.37**
Reddit	Llama2	0.12	0.23	0.21	0.86	16.78
	TinyLlama	**0.30**	**0.55**	**0.44**	**0.92**	**57.68**
NYT	Llama2	0.15	0.26	0.24	0.86	19.35
	TinyLlama	**0.40**	**0.61**	**0.53**	**0.94**	**63.87**
Wiki	Llama2	0.23	0.34	0.35	0.87	21.38
	TinyLlama	**0.37**	**0.57**	**0.49**	**0.92**	**58.59**

Table 4: Comparison of SSLC with and without curation step using TinyLlama 1.1 B model

Dataset	Variant	ROUGE-2↑	ROUGE-L↑	METEOR↑	BERT-F1↑	Human Eval↑ (Proxy)
AdversarialQA	No Curation	0.09	0.19	0.22	**0.87**	17.43
	With Curation	**0.11**	**0.21**	**0.23**	**0.87**	**21.17**
Reddit	No Curation	0.25	0.44	0.42	0.90	44.56
	With Curation	**0.26**	**0.46**	**0.43**	**0.91**	**45.58**
NYT	No Curation	0.32	0.51	0.48	**0.92**	54.19
	With Curation	**0.34**	**0.54**	**0.49**	**0.92**	**56.47**
Wiki	No Curation	0.32	0.48	0.48	0.91	44.00
	With Curation	**0.35**	**0.51**	**0.49**	**0.92**	**49.94**

tasks, such as text summarization and named entity recognition, could broaden its impact and utility.

Even for use cases which can leverage cloud-hosted proprietary LLMs, the framework can be utilized to customize smaller LLMs for specific tasks, thereby reducing the usage of proprietary models, resulting in cost reductions and latency improvements.

References

[1] Marah Abdin et al. 2024. Phi-3 technical report: a highly capable language model locally on your phone. (2024). arXiv: 2404.14219 [cs.CL].

[2] Anthropic. 2024. Claude sonnet: a language model for conversational ai. Accessed: 2024-07-05. (2024). https://www.anthropic.com/news/claude-3-family.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, (Eds.) Association for Computational Linguistics, Ann Arbor, Michigan, (June 2005), 65–72. https://aclanthology.org/W05-0909.

[4] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8, 662–678. Mark Johnson, Brian Roark, and Ani Nenkova, (Eds.) DOI: 10.1162/tacl_a_00338.

[5] Tom Brown et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

[6] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: open-source legal large language model with integrated external knowledge bases. (2023). arXiv: 2306.16092 [cs.CL].

[7] Tri Dao. 2023. Flashattention-2: faster attention with better parallelism and work partitioning. (2023). arXiv: 2307.08691 [cs.LG].

[8] Cheng Deng et al. 2023. K2: a foundation language model for geoscience knowledge understanding and utilization. (2023). arXiv: 2306.05064 [cs.CL].

[9] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. (2022). arXiv: 2110.02861 [cs.LG].

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*. A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, (Eds.) Vol. 36. Curran Associates, Inc., 10088–10115. https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

[11] Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How reliable are automatic evaluation methods for instruction-tuned llms? (2024). arXiv: 2402.10770 [cs.CL].

[12] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: evaluate as you desire. (2023). arXiv: 2302.04166 [cs.CL].

[13] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, (Eds.) Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, (Dec. 2022), 505–525. DOI: 10.18653/v1/2022.emnlp-main.33.

[14] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, (Eds.) Association for Computational Linguistics, Online, (July 2020), 8342–8360. DOI: 10.18653/v1/2020.acl-main.740.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: low-rank adaptation of large language models. *CoRR*, abs/2106.09685. https://arxiv.org/abs/2106.09685 arXiv: 2106.09685.

[16] Albert Q. Jiang et al. 2023. Mistral 7b. (2023). arXiv: 2310.06825 [cs.CL].

[17] Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

[18] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. [n. d.] Alpacaeval: an automatic evaluator of instruction-following models. ().

[19] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. (2023). arXiv: 2303.14070 [cs.CL].

[20] Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, (July 2004), 74–81. https://aclanthology.org/W04-1013.

[21] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. (2020). arXiv: 2004.14444 [cs.LG].

[22] Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. arXiv:2402.18334 [cs.CL].

[23] Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, (Eds.) Association for Computational Linguistics, Seattle, United States, (July 2022), 112–128. DOI: 10.18653/v1/2022.findings-naacl.10.

[24] Bowen Tan, Yun Zhu, Lijuan Liu, Eric Xing, Zhiting Hu, and Jindong Chen. 2023. Cappy: outperforming and boosting large multi-task lms with a small scorer. In *Advances in Neural Information Processing Systems*. A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, (Eds.) Vol. 36. Curran Associates, Inc., 58875–58889. https://proceedings.neurips.cc/paper_files/paper/2023/file/b860c0c546f4a3a786f9c9468228c99f-Paper-Conference.pdf.

[25] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. (2023). arXiv: 2307.09288 [cs.CL].

[26] [SW] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert, TRL: Transformer Reinforcement Learning version 0.2.1. URL: https://github.com/huggingface/trl.

[27] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*. Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, (Eds.) Association for Computational Linguistics, Singapore, (Dec. 2023), 1–11. DOI: 10.18653/v1/2023.newsum-1.1.

[28] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: a large language model for finance. (2023). arXiv: 2303.17564 [cs.LG].

[29] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: an open-source small language model. (2024). arXiv: 2401.02385 [cs.CL].

[30] Shengyu Zhang et al. 2024. Instruction tuning for large language models: a survey. (2024). arXiv: 2308.10792 [cs.CL].

[31] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: evaluating text generation with bert. (2020). arXiv: 1904.09675 [cs.CL].

[32] Lianmin Zheng et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*. A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, (Eds.) Vol. 36. Curran Associates, Inc., 46595–46623. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

A Detailed results of experiments

A.1 SSLC with self synthesis

Table 5: Results of experiments across models and datasets.
Variant: Base ⇒ Base un-customized model | Bonito ⇒ Bonito synthesis and finetuning | SSLC(SS) ⇒ SSLC with self synthesis
Metrics: R-1 ⇒ Rouge1 | R-2 ⇒ Rouge2 | R-L ⇒ RougeL | M ⇒ METEOR | B-P ⇒ Bert Prceision | B-R ⇒ BERT Recall | B-F1 ⇒ BERT F1 | HP-A ⇒ Human proxy accuracy with Claude as human proxy

Model	Dataset	Variant	R-1↑	R-2↑	R-L↑	M↑	B-P↑	B-R↑	B-F1↑	HP-A↑
TinyLlama (1.1B)	AdversarialQA	Base	0.11	0.06	0.11	0.19	0.82	0.87	0.84	14.74
		Bonito	0.19	0.08	0.19	0.14	**0.88**	0.88	**0.88**	15.62
		SSLC(SS)	**0.22**	**0.11**	**0.21**	**0.23**	0.86	**0.89**	**0.87**	**21.17**
	Reddit	Base	0.14	0.07	0.13	0.23	0.82	0.87	0.84	22.08
		Bonito	**0.52**	0.25	**0.52**	0.38	**0.92**	**0.92**	**0.92**	**55.14**
		SSLC(SS)	0.46	**0.26**	0.46	**0.43**	0.90	**0.92**	0.91	45.98
	NYT	Base	0.21	0.12	0.20	0.33	0.84	0.89	0.86	32.77
		Bonito	0.53	0.31	0.53	0.40	0.87	0.87	0.87	55.38
		SSLC(SS)	**0.54**	**0.34**	**0.54**	**0.49**	**0.92**	**0.93**	**0.92**	**56.47**
	Wiki	Base	0.22	0.15	0.22	0.37	0.83	0.90	0.86	28.43
		Bonito	0.33	0.19	0.33	0.26	0.87	0.88	0.87	30.04
		SSLC(SS)	**0.52**	**0.35**	**0.51**	**0.49**	**0.91**	**0.92**	**0.92**	**49.94**
Llama2 (7B)	AdversarialQA	Base	0.09	**0.05**	0.09	**0.17**	0.81	**0.87**	**0.84**	**6.93**
		Bonito	0.07	0.03	0.07	0.07	0.80	0.85	0.82	5.42
		SSLC(SS)	**0.12**	**0.06**	**0.12**	0.12	**0.82**	0.86	**0.84**	**6.93**
	Reddit	Base	0.13	0.06	0.13	**0.21**	0.81	**0.87**	0.84	11.70
		Bonito	0.18	0.09	0.18	0.14	**0.83**	**0.87**	**0.85**	13.33
		SSLC(SS)	**0.21**	**0.11**	**0.21**	0.19	**0.83**	**0.87**	**0.85**	**15.16**
	NYT	Base	0.16	0.09	0.16	**0.28**	0.82	**0.88**	0.85	**17.18**
		Bonito	0.20	0.11	0.20	0.15	0.83	0.87	0.85	14.51
		SSLC(SS)	**0.27**	**0.17**	**0.27**	0.25	**0.84**	**0.89**	**0.86**	18.46
	Wiki	Base	0.18	0.12	0.17	**0.32**	0.82	**0.89**	0.85	13.35
		Bonito	0.13	0.08	0.13	0.12	0.81	0.85	0.83	11.62
		SSLC(SS)	**0.31**	**0.19**	**0.31**	0.29	**0.85**	0.88	**0.87**	**19.28**
Mistral (7B)	AdversarialQA	Base	0.12	0.06	0.12	0.22	0.82	0.88	0.84	18.72
		Bonito	0.37	0.17	0.37	0.30	**0.90**	0.91	0.90	33.98
		SSLC(SS)	**0.42**	**0.21**	**0.41**	**0.36**	**0.90**	**0.91**	**0.91**	**40.64**
	Reddit	Base	0.17	0.08	0.16	0.28	0.83	0.88	0.85	24.72
		Bonito	**0.72**	**0.39**	**0.72**	0.55	**0.95**	**0.95**	**0.95**	**76.09**
		SSLC(SS)	0.68	0.38	0.68	**0.56**	0.94	0.94	0.94	72.84
	NYT	Base	0.29	0.18	0.28	0.43	0.86	**0.90**	0.88	37.41
		Bonito	**0.74**	**0.47**	**0.74**	0.58	**0.95**	**0.95**	**0.95**	**76.70**
		SSLC(SS)	0.71	**0.47**	0.71	**0.61**	0.94	**0.95**	**0.95**	74.33
	Wiki	Base	0.26	0.17	0.25	0.41	0.84	0.91	0.87	33.75
		Bonito	0.54	0.38	0.54	0.52	0.91	0.92	0.92	53.03
		SSLC(SS)	**0.70**	**0.46**	**0.69**	**0.60**	**0.94**	**0.94**	**0.94**	**72.56**

A.2 SSLC - Mistral synthesis results

Table 6: SSLC performance with synthesized data from Mistral[16]

Dataset	Model	R-1 ↑	R-2 ↑	R-L ↑	M ↑	B-P ↑	B-R ↑	B-F1 ↑	HP-A↑
AdversarialQA	Llama2	0.14	0.07	0.14	0.13	0.82	0.86	0.84	7.01
	TinyLlama	**0.26**	**0.12**	**0.26**	**0.23**	**0.83**	**0.84**	**0.83**	**23.37**
Reddit	Llama2	0.23	0.12	0.23	0.21	0.84	0.88	0.86	16.78
	TinyLlama	**0.55**	**0.30**	**0.55**	**0.44**	**0.92**	**0.93**	**0.92**	**57.68**
NYT	Llama2	0.26	0.15	0.26	0.24	0.84	0.88	0.86	19.35
	TinyLlama	**0.61**	**0.40**	**0.61**	**0.53**	**0.93**	**0.94**	**0.94**	**63.87**
Wiki	Llama2	0.34	0.23	0.34	0.35	0.86	0.89	0.87	21.38
	TinyLlama	**0.58**	**0.37**	**0.57**	**0.49**	**0.93**	**0.92**	**0.92**	**58.59**

A.3 SSLC - Effect of curation

Table 7: Comparison of the framework with and without curation step using TinyLlama 1.1 B model

Dataset	Variant	R-1↑	R-2↑	R-L↑	M↑	B-P↑	B-R↑	B-F1↑	HP-A↑
AdversarialQA	Without Curation	0.20	0.09	0.19	0.22	0.85	0.88	**0.87**	17.43
	With Curation	**0.22**	**0.11**	**0.21**	**0.23**	**0.86**	**0.89**	**0.87**	**21.17**
Reddit	Without Curation	0.44	0.25	0.44	0.42	0.89	**0.91**	0.90	44.56
	With Curation	**0.46**	**0.26**	**0.46**	**0.43**	**0.92**	**0.91**	**0.91**	**45.58**
NYT	Without Curation	0.52	0.32	0.51	0.48	0.91	**0.93**	**0.92**	54.19
	With Curation	**0.54**	**0.34**	**0.54**	**0.49**	**0.92**	**0.93**	**0.92**	**56.47**
Wiki	Without Curation	0.48	0.32	0.48	0.48	0.90	**0.92**	0.91	44.00
	With Curation	**0.52**	**0.35**	**0.51**	**0.49**	**0.91**	**0.92**	**0.92**	**49.94**

B Experiment configuration and hyper-parameters

Hyper-parameter	Value
Hold out set(test set) fraction	10%
Segmentation word count limit	768
Synthesizer decoding strategy	greedy
Synthesizer max_new_tokens	512
Model quantisation	bitsandbytes
Model quantisation bits	4
Model weights data type	bfloat16
Attention optimization	flash_attention_2([7]
Cappy scorer threshold for curation	0.4
Finetuning LoRA alpha	256
Finetuning LoRA rank	1024
Finetuning LoRA dropout	0.1
Finetuning LoRA max_seq_length	2048
Finetuning learning_rate	1e-4
Finetuning optimizer	adamw_bnb_8bit [9]

C Prompts

Usage	Prompt
Synthesis system prompt	You are a teacher who prepares questions for students based on a piece of text. You will be given a piece of text and you will generate questions and answer for the given text
System prompt for instruction tuning and question answering on AdversarialQA dataset	You are a wikipedia author. You will be given a snippet of a wiki in triple quotes. Answer the user query at the end using the snippet.
System prompt for instruction tuning and question answering on Reddit dataset	You are a social media analyst. You will be given a snippet of a social media conversation in triple quotes. Answer the user query at the end using the conversation.
System prompt for instruction tuning and question answering on NYT dataset	You are a news analyst. You will be given a snippet of a news article in triple quotes. Answer the user query at the end using the snippet
System prompt for instruction tuning and question answering on Wiki dataset	You are a wikipedia author. You will be given a snippet of a wiki in triple quotes. Answer the user query at the end using the snippet.
Prompt for question answering with finetuned model on Bonito synthesized data	Given the following context: """{context}"""answer the following question:{question}
Claude Human Evaluation(Proxy) prompt	I'm going to give you two pieces of text. Your job is to compare them to check if they match approximately. Provide the result in single word. Output "MATCH" only if they match upto 80%, "NOMATCH" otherwise. <text1>{groundtruth_answer}</text1> <text2>{generated_answer}</text2>